Submitted to the Annals of Applied Statistics
arXiv: stat.ME/0000000

SEMIPARAMETRIC MODELING OF AUTONOMOUS NONLINEAR DYNAMICAL SYSTEMS WITH APPLICATION TO PLANT GROWTH

By Debashis Paul* and Jie Peng† and Prabir Burman‡

University of California, Davis

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 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.

*Supported by the NSF grants DMS-0806128 and DMR-1035468
†Supported by the NSF grants DMS-0806128 and DMS-1001256
‡Supported by the NSF grant DMS-0907622
AMS 2000 subject classifications: Primary 62G08; secondary 62P10
Keywords and phrases: Autonomous dynamical systems, cross-validation, growth displacement rate, Levenberg-Marquardt method, semiparametric modeling

1. Introduction. Continuous time dynamical systems arise, among other places, in modeling certain biological processes. This includes classical examples from population biology like the *Lotka-Voltera equations* for describing prey-predator dynamics (Perthame, 2007), or subject-specific processes like the progression of infectious diseases in individuals (Nowak and May, 2000). Most of the existing approaches estimate the dynamical system by assuming known functional forms of the system. Moreover, many of them aim at estimating individual dynamics for one subject. However, in many scientific studies, there is a need to model the dynamical system nonparametrically due to insufficient knowledge of the problem at hand. In addition, there could be an interest to know the dynamics of a certain process at a population level in order to answer various scientific questions. Thus, in this paper, we propose a new method to bridge the gap and tackle these challenges.

To motivate the model, we first briefly discuss a study on plant growth. There is a lot of research aiming to understand the effect of environmental conditions on the growth in plant. For example, root growth in plants is highly sensitive to environmental factors such as temperature, water deficit or nutrients (Schurr et al., 2006; Walter et al., 2002). 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 (Figure 1: left panel). 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 (Figure 1: right panel). The growth displacement rate is defined as the rate of displacement of a particle placed along the root and thus it is a function of distance from the root apex. By its definition, growth displacement rate characterizes the relationship between the growth trajectory and its derivative (with respect to time). Therefore, it is the gradient function in the corresponding dynamical system. In this study, there is a need to understand the dynamics at the population level, while accounting for subject-specific variations, in order to compare the growth displacement rates under two different water conditions.

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. Moreover, for a given sample curve, 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 that we just mentioned, 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. Each marker corresponds to one displacement trajectory (also referred to as growth trajectory/curve), and the number of measurements varies from two to seventeen, with measurement times varying across trajectories. (See Section 5 for a more detailed description.)

We first give a brief overview of the existing literature on fitting smooth deterministic dynamical systems in continuous time. A large number of physical, chemical or biological processes are modeled through systems of parametric differential equations (Ljung and Glad, 1994; Perthame, 2007; Strogatz, 2001). For example, Ramsay et al. (2007) consider modeling a continuously stirred tank reactor and propose a method called parameter cascading for model fitting. Zhu and Wu (2007) adopt a state space approach for estimating the dynamics of cell-virus interactions in an AIDS clinical trial. Ramsay and Silverman (2002, 2005) consider fitting dynamical systems given by systems of linear differential equations where the coefficients of the differential operator may be time varying. They propose methods for estimating these (linear) differential operators based on principal differential analysis when the data are recorded at dense and regular time points. Poyton et al. (2006) also use principal differential analysis approach to fit dynamical systems. Chen and Wu (2008a,b) propose to estimate parametric differential equations with known functional forms and time-dependent parameters through a two-stage approach where the first stage involves estimation of the sample trajectories and their derivatives by nonparametric smoothing. Brunel (2008) gives a comprehensive theoretical analysis of such an approach. Cao et al. (2008) propose a method for fitting nonlinear dynamical systems using splines with predetermined knots for describing the gradient function. This involves knowing the functional form of the differential equation and does not include any subject-specific effects. Wu and Ding (1999) and Wu et al. (1998) propose using nonlinear least squares procedure for fitting parametric differential equations that take into account subject-specific effects.

For the problems that we address in this paper, measurements are taken on a sparse set of points for each sample curve so that estimation of individual sample trajectory or its derivative based on nonparametric smoothing is error-prone and results in a loss of information. 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 at the population level than the individual dynamics of each subject. For example, in the plant study described above, we are interested in comparing the growth displacement rates under two different experimental conditions. On the other hand, we are not so interested

in the individual 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 the model parameters can have inflated variability. In this paper, we propose a semi-parametric approach for modeling dynamical systems which incorporates subject-specific effects while combining information across different subjects. A nonparametric model is often essential because of insufficient knowledge about the problem to suggest a reasonable parametric form of the dynamical system. In addition, if realistic parametric models can be proposed then the nonparametric fit can be used for diagnostics of lack of fit, for example by employing a distance measure between the parametric and nonparametric fits and studying its sampling variability. We propose an estimation procedure that combines nonlinear optimization techniques with a numerical ODE (ordinary differential equation) 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 show by simulation studies that the proposed approach can efficiently estimate the baseline dynamics with noisy and sparsely measured sample curves. Finally, we apply the proposed method to the plant data and compare the estimated growth displacement rates under the two experimental conditions and discuss some scientific implications of the results.

To the best of our knowledge, modeling and fitting dynamical systems nonparametrically while also allowing for subject-specific effects is new in the literature. In particular, our model differs from traditional nonlinear mixed effects models previously employed for fitting differential equations, which are almost exclusively parametric (Wu et al., 1998; Wu and Ding, 1999; Guedj et al., 2007; Li et al., 2002). For example, Guedj et al. (2007) consider a nonlinear state-space model where the state variable follows a parametric differential equation with subject-specific effects, and the parameters are estimated through a maximum likelihood approach. In contrast, for the model proposed in this paper, the form of the gradient function g is not assumed to be known and it is approximated in a sequence of bases with growing dimension. Note that, this gives rise to a sequence of parametric models with increasing complexity, and one needs to adopt a model selection procedure to select an appropriate model, as is typical in nonparametric function estimation. The theoretical derivations in Paul et al. (2009) also show that the problem of estimating the gradient function g nonparametrically is intrinsically different from that under a parametric nonlinear mixed-effects model.

The rest of the paper is organized as follows. In Section 2, we describe the proposed model. In Sections 3, we discuss the model fitting and model selection procedures. In Section 4, we conduct simulation studies to illustrate finite sample performance of the proposed method and compare the proposed method with a two-stage procedure. In Section 5, we apply this method to the plant data. Section 6 has a brief discussion. More details and additional simulation results are reported in the supplementary material (Paul et al., 2011).

2. Model. In this section, we describe a class of autonomous dynamical systems that is suitable for modeling the problems discussed in 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 implies that $X(t) = a + \int_0^t f(X(u))du$, where $a = X(0)$ is the initial condition. In an autonomous system, the dynamics, which is characterized by f, depends on time t only through the "state" $X(t)$. This type of systems arises in various scientific studies such as modeling prey-predator dynamics, virus dynamics, or epidemiology (Perthame, 2007).

In this paper, we consider the following class of autonomous dynamical systems:

$$X'_{il}(t) = g_i(X_{il}(t)), \quad l = 1, \cdots, N_i, \quad i = 1, \ldots, n, \tag{2.1}$$

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, where $N_i \geq 1$ is the number of 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 and there are multiple markers for each plant. We assume that all the curves associated with the same subject follow the same dynamics and are only differentiated by different initial conditions. These are described by the functions $\{g_i(\cdot)\}_{i=1}^n$. In this paper, we model $\{g_i(\cdot)\}_{i=1}^n$ as:

$$g_i(\cdot) = e^{\theta_i} g(\cdot), \quad i = 1, \ldots, n, \tag{2.2}$$

where

(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 to as the *gradient function*.

(2) θ_i's reflect subject-specific effects in these systems. The mean of θ_i's is assumed to be zero to impose identifiability.

Note that, one may view the trajectories for each plant as multivariate functional data. However, here for each subject, the different trajectories correspond to different initial conditions of the same ODE describing the system. This means that given the initial condition and the subject-specific scaling parameter θ_i, the corresponding trajectory is completely determined by the underlying dynamical system and the only source of randomness is from measurement errors.

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. Secondly, the scale parameter e^{θ_i} provides a subject-specific tuning of the dynamics. This is motivated by the fact that, for a large class of problems, the variations of the dynamics in a population are in the scale of the rate of change rather than in the shape of the gradient function. For example, for the plant data, by examining the scatter plot of empirical derivatives versus empirical fits (Figure 2, for more details, see Section 5), we observe an excessive variability towards the end which reflects plant-specific scaling effects. Moreover, the above model is also flexible in incorporating time-independent covariates, say z_i, for example by expressing the scaling factor as $e^{\eta^T z_i}$ for some parameter η. In this paper, our primary goal is to estimate the gradient function g nonparametrically.

Assuming the gradient function g to be smooth means that it can be well-approximated by a basis representation approach:

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

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.

3. Model Fitting.

3.1. *Estimation procedure.* In this section, we propose an estimation procedure based on sparsely observed noisy data. Specifically, we assume that the observations are given by

$$Y_{ilj} = X_{il}(t_{ilj}) + \varepsilon_{ilj}, \quad j = 1, \ldots, m_{il}, \tag{3.1}$$

where $0 \le t_{il1} < \cdots < t_{ilm_{il}} \le 1$ are the measurement times for the lth curve of the ith subject, and $\{\varepsilon_{ilj}\}$ are independently and identically distributed noise with mean zero and variance $\sigma_\varepsilon^2 > 0$. For model fitting with such data, we adopt an iterative updating procedure which imposes regularization on the estimates of $\boldsymbol{\theta}$ and $\boldsymbol{a}$. 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),

$$\sum_{i=1}^{n}\sum_{l=1}^{N_i}\sum_{j=1}^{m_{il}}[Y_{ilj} - \widetilde{X}_{il}(t_{ilj}; a_{il}, \theta_i, \boldsymbol{\beta})]^2 + \lambda_1 \sum_{i=1}^{n}\sum_{l=1}^{N_i}(a_{il} - \alpha)^2 + \lambda_2 \sum_{i=1}^{n}\theta_i^2, \tag{3.2}$$

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 et al., 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 (3.2) 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, since the expression in (3.2) can also be viewed as a regularized ℓ_2 loss with penalties on the variability of $\boldsymbol{\theta}$ and $\boldsymbol{a}$.

In many problems, there are natural constraints on the gradient function g. Some of these constraints can be expressed in the form of quadratic constraints in certain derivatives of g. Thus, to add flexibility to our estimation procedure, we allow for incorporating penalties of the form: $\boldsymbol{\beta}^T\mathbf{B}\boldsymbol{\beta}$ for an

$M \times M$ positive semi-definite matrix $\mathbf{B}$ in the loss function. Consequently, the modified objective function becomes

$$L(\boldsymbol{a}, \boldsymbol{\theta}, \boldsymbol{\beta}) := \sum_{i=1}^{n} \sum_{l=1}^{N_i} \sum_{j=1}^{m_{il}} \ell_{ilj}(a_{il}, \theta_i, \boldsymbol{\beta}) + \boldsymbol{\beta}^T \mathbf{B} \boldsymbol{\beta}. \tag{3.3}$$

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

$$(\widehat{\boldsymbol{a}}, \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\beta}}) := \arg \min_{\boldsymbol{a}, \boldsymbol{\theta}, \boldsymbol{\beta}} L(\boldsymbol{a}, \boldsymbol{\theta}, \boldsymbol{\beta}). \tag{3.4}$$

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 the most efficient. In contrast, in a two-stage approach, the trajectories and their derivatives are first obtained via pre-smoothing (see e.g. Chen and Wu (2008a,b); Varah (1982)), and then they are used in a nonparametric regression framework to derive an estimate of g. This is inefficient since estimation errors introduced in the pre-smoothing step effectively cause a loss of information. Indeed, simulation studies carried out in Section 4 and the supplementary material (Paul et al., 2011) show that two-stage estimators suffer from significant biases in estimating the gradient function g. Alternative ways of estimating g include using the reproducing kernel Hilbert space framework (Gu, 2002), and controlling the degree of smoothness of the fitted g by tuning a roughness penalty.

In the following, we propose a numerical procedure for solving (3.4) 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 can be implemented using a numerical ODE solver, such as the Runge-Kutta method (Tenenbaum and Pollard, 1985).
- Minimize (3.3) with respect to $(\boldsymbol{a}, \boldsymbol{\theta}, \boldsymbol{\beta})$. This amounts to a nonlinear regression problem. It can be carried out using either a specialized nonlinear least squares solver, like the Levenberg-Marquardt method (Nocedal and Wright, 2006); or a general optimization procedure, such as the Newton-Raphson algorithm.

The above fitting procedure bears similarity to the local, or gradient-based, methods discussed by Li et al. (2002), Guedj et al. (2007) and Miao et al. (2009) even though their works focus on parametric ODEs. The main distinction of the proposed framework and those of Li et al. (2002) and Guedj et al.

(2007) lies in that, for the current setting, the complexity of the model is allowed to grow with increasing sample size and one eventually needs to adopt a model selection procedure to select an appropriate model (as is done in this paper). From purely a model-fitting point of view, nonlinear mixed-effects (NLME) model-based estimation procedures may be used in principle to fit each of these parametric submodels. The work of Ke and Wang (2001) on semiparametric mixed-effects model fitting also shares some common computational challenges with our model. However, unlike in Ke and Wang (2001), in our case, the likelihood for the nonparametric component (i.e., the gradient function) is not available in closed form.

We now briefly describe an optimization procedure based on the idea of the Levenberg-Marquardt method by linearization of $\{\widetilde{X}_{il}(\cdot)\}$ with respect to a_{il}, θ_i, and $\boldsymbol{\beta}$. 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 their current values and then update them by a least squares fitting, while keeping the other two sets of parameters fixed at the current values. This process is repeated until convergence.

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

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

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

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

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

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

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 $\{\widetilde{X}_{il}(\cdot)\}'s$ and their gradients (as well as Hessians) can be easily evaluated on a fine grid by using numerical ODE solvers such as the fourth order Runge-Kutta method (see Paul et al.

(2011) for details). Since given the trajectories, their gradients satisfy linear differential equations, the solution may also be obtained explicitly (see Appendix). The equation for updating $\boldsymbol{\beta}$, while keeping $\boldsymbol{a}^*$ and $\boldsymbol{\theta}^*$ fixed, is

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

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 decreasing 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

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

After updating θ_i's, we re-center the current estimates such that their mean is set to be zero. This also helps in stabilizing the algorithm. The equation for updating a_{il}, while keeping θ_i and $\boldsymbol{\beta}$ fixed at θ_i^*, $\boldsymbol{\beta}^*$ is:

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

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. Note that on convergence, $\widehat{\alpha} := \alpha^*$ provides an estimate of α. The initial estimates can be conveniently chosen. For example, $a_{il}^{ini} = Y_{il1}$ and $\theta_i^{ini} \equiv 0$.

This procedure is quite stable and robust to the initial parameter estimates. However, it converges slowly in the neighborhood of the minima of the objective function as it is a first order procedure. On contrary, the Newton-Raphson algorithm has a fast convergence when starting from estimates that are already near the minima. Thus, in practice, one could first use the above approach (referred to as the Levenberg-Marquardt step hereafter) to obtain a reasonable estimate and then use the Newton-Raphson algorithm to expedite the search of the minima. The derivation of the Newton-Raphson algorithm is rather standard and thus is omitted. If the true gradient function g has high complexity, and/or if either the θ_i's or the noise are highly variable, the Newton-Raphson algorithm may be unstable, particularly when the initial conditions $\boldsymbol{a} = \{X_{il}(0)\}$ are also estimated. Under such situations, we recommend using a (relatively) large number of Levenberg-Marquardt steps, followed by a one-step Newton-Raphson update.

Note that, 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.

In this paper, we simply set $\lambda_{3j} = \lambda_3^0/j$ 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 loss function (3.3). 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 viewpoint, λ_1 and λ_2 are determined by the variances σ_ε^2, σ_a^2 and σ_θ^2 through $\lambda_1 = \sigma_\varepsilon^2/\sigma_a^2$ and $\lambda_2 = \sigma_\varepsilon^2/\sigma_\theta^2$. 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),

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

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. Instead, if we take the penalized loss function viewpoint, we can simply treat λ_1 and λ_2 as fixed regularization parameters which can be chosen by model selection criteria (see Section 3.3). Henceforth, we refer to the method as `adaptive` if λ_1 and λ_2 are updated after each iteration, and as `non-adaptive` if they are kept fixed throughout the optimization.

3.2. *Standard error of the estimates.* It is important to obtain the standard error of the estimated gradient function. Since it is typically not possible to obtain an estimate of the bias for a nonparametric procedure, we ignore the bias term and use the best projection of true g in the model space as the surrogate center (this is the standard practice in nonparametric literature). Thus equivalently, we provide an estimate of the asymptotic variance of $\widehat{\boldsymbol{\beta}}$. Based on the asymptotic analysis presented in Paul et al. (2009), we derive the following estimate:

$$V(\widehat{\beta}) := \widehat{E}[(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T] = \widehat{\sigma}_\varepsilon^2 \mathbf{W}_n, \tag{3.5}$$

with $\mathbf{W}_n = (\mathbf{A}_n + \mathbf{B} - \mathbf{C}_n^T(\mathbf{D}_n + \lambda_2 I_n)^{-1}\mathbf{C}_n)^{-1}$, where, I_n is the $n \times n$ identity matrix, $\mathbf{A}_n = \mathcal{G}_{\beta\beta}(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\beta}})$, $\mathbf{C}_n = \mathcal{G}_{\theta\beta}(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\beta}})$, $\mathbf{D}_n = \mathcal{G}_{\theta\theta}(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\beta}})$; where

$$\mathcal{G}_{\beta\beta}(\boldsymbol{\theta}, \boldsymbol{\beta}) := \sum_{i=1}^{n}\sum_{l=1}^{N_i}\sum_{j=1}^{m_{il}}\left(\frac{\partial X_{il}}{\partial \boldsymbol{\beta}}(t_{ilj}; a_{il}, \theta_i, \boldsymbol{\beta})\right)\left(\frac{\partial X_{il}}{\partial \boldsymbol{\beta}}(t_{ilj}; a_{il}, \theta_i, \boldsymbol{\beta})\right)^T;$$

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

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

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

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

Note that, the matrices $\mathbf{A}_n$, $\mathbf{C}_n$ and $\mathbf{D}_n$ are obtained as byproducts of the estimation procedure. An estimate of the standard error of $\widehat{g}(x)$ for x in the domain of $\{\phi_{k,M}\}_{k=1}^{M}$, is therefore given by

$$\widehat{\mathrm{SE}}(\widehat{g}(x)) = \left[\boldsymbol{\phi}_M(x)^T V(\widehat{\boldsymbol{\beta}}) \boldsymbol{\phi}_M(x) \right]^{1/2} \tag{3.6}$$

where $\boldsymbol{\phi}_M(x) := (\phi_{1,M}(x), \ldots, \phi_{M,M}(x))^T$ and $V(\widehat{\boldsymbol{\beta}})$ is as in (3.5). Note that, in the given asymptotic framework, we treat θ_i's as random effects and the initial conditions $\{a_{il}\}$ are assumed to be known. In deriving (3.5), we have ignored the correlation structure between θ_i and the gradient of the objective function with respect to θ_i, which yields a slightly conservative (i.e., upwardly biased) estimate of the standard error. Obtaining the asymptotic standard error estimates when the initial conditions $\{a_{il}\}$ are estimated from the data is beyond the scope of this paper.

As an alternative way of estimating the standard error, one may also use bootstrap where we resample the sample trajectories corresponding to each subject, in order to retain the overall structure of the model. The corresponding bootstrap estimates, though simple to obtain, are computationally expensive and we do not pursue this in this paper.

3.3. *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, penalty parameters, etc. In the literature, AIC/BIC/AICc criteria have been proposed for model selection of parametric dynamical systems, see for example Miao et al. (2009). 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 can be defined as

$$CV := \sum_{i=1}^{n} \sum_{l=1}^{N_i} \sum_{j=1}^{m_{il}} \ell_{ilj}^{cv}(\widehat{a}_{il}^{(-il)}, \widehat{\theta}_i^{(-il)}, \widehat{\boldsymbol{\beta}}^{(-il)}) \tag{3.7}$$

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 lth curve of the ith subject; 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}; and $\ell_{ilj}^{cv}(a_{il}, \theta_i, \boldsymbol{\beta}) :=$

$(Y_{ilj} - \widetilde{X}_{il}(t_{ilj}; a_{il}, \theta_i, \boldsymbol{\beta}))^2$ is the prediction error loss. When the initial conditions $a_{il} = X_{il}(0)$ are i.i.d. random variables and are known (and thus we set $\widehat{a}_{il}^{(-il)} = a_{il}$), the leave-one-curve-out CV score gives an asymptotically unbiased estimator of the prediction error. Calculating CV score (3.7) is computationally very demanding because one needs to obtain $\widehat{\theta}_i^{(-il)}$ and $\widehat{\boldsymbol{\beta}}^{(-il)}$ for every pair of (i, l). Therefore, we propose to approximate $\widehat{\theta}_i^{(-il)}$ and $\widehat{\boldsymbol{\beta}}^{(-il)}$ through a first order Taylor expansion around the estimates $\widehat{\theta}_i, \widehat{\boldsymbol{\beta}}$ based on the full data. We then obtain an approximation of $\widehat{a}_{il}^{(-il)}$ by minimizing the corresponding criterion with the approximations of $\widehat{\theta}_i^{(-il)}$ and $\widehat{\boldsymbol{\beta}}^{(-il)}$ imputed. Consequently we derive an approximate CV score $\widetilde{CV}$ by plugging these approximations in (3.7), which is computationally inexpensive since all the quantities involved in computing $\widetilde{CV}$ are byproducts of the Newton-Raphson step used in model fitting. This approximation scheme is similar to the one taken in Peng and Paul (2009) under the context of functional principal component analysis, which itself is motivated by the work of Burman (1990). Detailed derivations are given in the Appendix.

4. Simulation. In this section, we conduct a simulation study to demonstrate the effectiveness of the proposed estimation and model selection procedures. Since we apply our method to study the plant growth dynamics in Section 5, we consider a simulation setting that partly mimics that data set. 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 (to ensure positivity as well as to study model robustness), 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 (2.1), (2.2) and (3.1). For all settings, 50 independent data sets are used to evaluate the performance of the proposed procedure. The sample trajectories are evaluated using the 4th order Runge-Kutta method (as described in

Paul et al. (2011)) on an equally spaced grid with grid spacings $h = 0.0005$.

In the estimation procedure, we consider cubic B-spline basis functions with knots at $0.1 + j/M$, $j = 1, \ldots, M$, to model g, where M varies from 2 to 6. Note that, here $M = 4$ corresponds to the true gradient function. The Levenberg-Marqardt step is chosen to be `non-adaptive`, and the Newton-Raphson step is chosen to be `adaptive` (see Section 3.1 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 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 in the performance between the `moderate case` and the `sparse` case. On the other hand, when $\boldsymbol{a}$ are estimated, the advantages of having more measurements become more prominent. We also conduct further simulation studies (results not reported in details here) to check the effect of increasing the noise level, as well as the dispersion of the initial conditions $\boldsymbol{a}$. When $\boldsymbol{a}$ are known, even with $\sigma_\varepsilon = 0.05$, the convergence is almost unaffected, and in about 75% of the cases the true model ($M = 4$) is selected. Increasing σ_a to 0.1 does affect convergence, especially for larger M. But under this setting, even with $\sigma_\varepsilon = 0.05$ the true model converges in 90% cases and is selected to be the best in more than 75% cases. When $\boldsymbol{a}$ are estimated, the convergence deteriorates more obliviously under increased noise levels.

In Figure 4, we have a graphical 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 curve), the pointwise mean of $\widehat{g}$ (broken curve), and 2.5% and 97.5% pointwise quantiles (dotted 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, mainly due to a scarcity of data in that region. Indeed, the larger MISE of the estimator of g when initial conditions are estimated mainly results from the larger MISE on the domain of g where there is essentially no observed data. Due to the extrapolation effect, no method without using true true initial conditions is expected to work well on such a domain, especially under a nonparametric setting. This point is illustrated in more detail later in this section (cf. Table 3), as well as in the supplementary material (Paul et al. (2011), Section S3). Overall, as can be seen from these tables and figures, the proposed estimation and model selection procedures perform effectively.

To evaluate the accuracy of the pointwise standard error estimator given in (3.6), in Figure 6, we plotted the average of the estimate (blue curve) over 50 independent data sets and the ± 2 standard error bands of the estimates (broken red curves) based on the same 50 independent data sets under the true model ($M = 4$) when $\boldsymbol{a}$ is known. The pointwise standard errors are also computed empirically from the converged replicates (black curve) among the 50 simulation runs. We observe that although being somewhat conservative, (3.6) gives a quite satisfactory estimate of the pointwise standard error of $\widehat{g}$.

We also compare the performance of the proposed procedure with a two-stage approach. Following Chen and Wu (2008a), in the first stage, each individual trajectory $X_{il}(\cdot)$ and its derivative $X'_{il}(\cdot)$ are estimated by local linear and local quadratic smoothing, respectively. The bandwidths are chosen by cross validation. In the second stage, two different methods for estimating g are considered with $\{\widehat{X}'_{il}(t)\}$ as response and $\{\widehat{X}_{il}(t)\}$ as predictor: (i) a least squares regression fit of the basis coefficients using the true model; (ii) a local quadratic smoothing. A more detailed description of the two-stage approach and more simulation studies are given in the supplementary material (Paul et al. (2011), Section S2).

In Table 3, we report the integrated squared errors of the two-stage estimators as well as those of the hierarchical likelihood estimators (under the model selected by $\widetilde{CV}$) for the **sparse** case. While reporting the risk of the estimators, we divide the domain of x into three regions: $[-0.5, 0.2]$, $(0.2, 1]$ and $(1, 1.5]$. In this simulation, even though the true gradient function g has support effectively on $[-0.5, 1.5]$, the observed measurements Y_{ilj}'s are almost entirely confined in the region $(0.2, 1]$. Due to the extrapolation effect, methods without using the true initial conditions are expected to perform (relatively) poorly in the domains where there is no data. Thus, we divide the domain into different regions for more informative comparisons across

methods. We also plot the pointwise mean and median and pointwise 95% bands around the mean for the two-stage estimators of g in Figure 5. These results show that the two two-stage estimators are highly biased and variable. The one using the true model in the second stage has better behavior in the regions where there is no data, compared to the fully nonparametric estimator. However, the level of bias and variability is much higher than the proposed estimator on all three regions. Another important observation is that, for the hierarchical likelihood estimator, the median of integrated squared errors over the data domain $(0.2, 1]$ are comparable for the cases when the initial conditions $\boldsymbol{a}$ is known and when $\boldsymbol{a}$ is estimated.

To further compare these two approaches, we conduct another simulation study where all θ_i's are taken to be zero (equivalently, $\sigma_\theta = 0$), so that there is no subject-specific variability. For this simulation, we also consider a sampling design, referred to as "very dense", in which the number of measurements per curve is Uniform$[60, 100]$ so that the first stage estimates of the two-stage methods are more accurate. The number of subjects is chosen to be $n = 10$ and there is only one curve per subject (i.e., $N_i \equiv 1$). The results (reported in Table S5-5 in Paul et al. (2011)) show that the proposed method again gives better estimates and it is much less biased (even when the initial conditions are estimated). The mean integrated squared error over the data domain $(0.2, 1]$ of the hierarchical likelihood estimator, when $\boldsymbol{a}$ is estimated, is much smaller than that of the two-stage method, even when the true model is used in the second stage. For a more detailed comparison of the two approaches, see Section S2 of (Paul et al., 2011).

Moreover, we also do simulations when the true gradient function g is more complex and does not belong to the model space . The overall picture for the performance of the proposed estimation and model selection procedures, as well as the comparison with the two-stage methods, is consistent with the results presented here. See Section S3 of (Paul et al., 2011) for details.

Finally we comment on the computational time and the rate of convergence of the proposed procedure. These depend on several factors, especially, the model complexity and bias, noise level and criteria for convergence. Typically, the convergence is faster when $\boldsymbol{a}$ is treated as known, as opposed to when it is estimated from the data. For the simulation study presented here, under the true model ($M = 4$), with $\boldsymbol{a}$ known, convergence is generally achieved in about 30 to 40 Levenberg-Marquardt steps and often in only 2 to 3 Newton-Raphson steps. The number of Levenberg-Marquardt steps required for convergence almost doubles when $\boldsymbol{a}$ is estimated. For biased models (including those presented in Section S3 of (Paul et al., 2011)), the convergence often takes more steps (up to 150 Levenberg-Marquardt steps

and several Newton-Raphson steps). The computational times for the simulation study presented in this section are summarized in Table 4. These computations were carried out on a 64-bit Linux machine with Intel Core 2 Quad processors running at 3.2 GHz and with 8 GB RAM.

5. 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. One goal of this study is to investigate the effect of water stress on growth displacement rate within the meristem of the primary root of maize seedlings. Note that, meristem is the tissue in plants consisting of undifferentiated cells and found in zones of the plant where growth can take place. The growth displacement rate is defined as the rate of displacement of a particle placed along the root and it should not 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). Growth displacement rate is important to infer the cell division rate – the local rate of formation of cells – that is not directly observable in a changing population of dividing cells. 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.

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 meristem region of the root, where the measurements are taken, is shown in Figure 1 (left panel). 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 period while the plants were growing. The measurement process is shown schematically in the right panel of Figure 1. In Figure 3, 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. Note that, measurement times are different for these two plants. Also, measurements were only taken in the meristem. Thus whenever a marker grew

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 that in Figure 3 some growth trajectories were cut short. More sophisticated data acquisition techniques are described in Walter et al. (2002) and Basu et al. (2007), where the proposed method is also potentially applicable.

Many studies in plant science such as Silk (1994), Sacks et al. (1997), Fraser et al. (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 also 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. In addition, the form of the gradient function g is not known to the plant scientists, only its behavior at root cap junction and at some later stage of growth are known (Silk, 1994). Figure 2, the scatter plot of empirical derivatives versus empirical fits in the treatment group, indicates that there is an increase in the growth displacement rate starting from a zero rate at the root cap junction, 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 considerable scientific interest. These boundary behaviors also imply that a linear ODE model is obviously not appropriate. In addition, 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. Moreover, there is some controversy among plant scientists about the possible existence of a "growth bump" in the middle of the meristem. Taking all these features into consideration, the semi-parametric model proposed in this paper is appropriate for investigating the scientific questions associated with this study, in particular, comparing the baseline growth displacement rates between the treatment and control groups. Notice that, in order for the proposed estimation method to give accurate estimate of the gradient function, we need only that the measurements on the state variable x is dense in its domain, and that the measurement errors are independent across time. These are satisfied for the plant data since, even though each trajectory is recorded at a relatively small number of time points, there is a fairly large number of trajectories for each plant, corresponding to the different initial conditions. Note that, for each plant, the number of mea-

surements is indeed the sum total of all the measurements for its different trajectories. Moreover, the proposed method combines information across different plants (subjects), which allows one to fit the model reasonably well even with relatively few measurements per subject.

Now 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 (3.1), where m_{il} is the number of measurements taken for the ith plant at its lth 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. Note that, the constraint at the root cap junction corresponds to $g(0) = 0 = g'(0)$, which is imposed by simply omitting the constant and linear terms in the spline basis. The flatness of g at a (unknown) distance away from the root cap junction means that $g'(x) = 0$ for $x \geq A$ for some constant $A > 0$. In order to impose this, as part of the objective function (3.3) we use

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

where $\boldsymbol{\phi} = (\phi_{1,M}, \ldots, \phi_{M,M})^T$ 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 3.3. Moreover, since the initial conditions (marker positions) $\{a_{il}\}$ are chosen according to some fixed experimental design (though measured with errors), it is not appropriate to shrink their estimates towards a fixed number. Hence, we set $\lambda_1 = 0$ in the loss function (3.3).

Before fitting the proposed model, 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 to this method as `stepwise-regression`. A similar method is employed by Sacks et al. (1997). 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 moderately large. Comparisons given later (Figure 10) demonstrate a clear superiority of the proposed method over this simple approach.

We fit the proposed model to the control group and the treatment group separately. For the control group, we first use the procedure described in Section 3.1 with g represented in cubic B-splines with M (varying from 2 to 12) equally spaced knots. At this stage, we set $\boldsymbol{\beta}^{ini} = 1_M$, $\boldsymbol{\theta}^{ini} = 0_n$, $\boldsymbol{a}^{ini} = (X_{il}(t_{il1}) : l = 1, \ldots, N_i)_{i=1}^n$. The criterion based on the approximate CV score (equation (**??**) in the Appendix) selects the model with $M = 9$ basis functions. 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 a candidate set of knots. We use 28 equally spaced candidate knots on the interval $[0.5, 14]$ and use the fitted values $\{\widehat{\theta}_i^{(0)}\}_{i=1}^{10}$ from the previous B-spline fit. The AIC criterion selects a model with 10 knots among these 28 candidate knots, plus the quadratic term. We then consider various submodels with knots chosen from this set of selected knots and fit them again using the proposed estimation procedure. The approximate CV scores for a number of different submodels are reported in Table 5. 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. Also note that, the model selected by `stepwise-regression` has a larger CV score than those of the models reported in Table 5. A similar procedure is applied to the treatment group. It turns out that the model with knot sequence $(3.0, 3.5, 7.5)$, which is also selected by `stepwise-regression`, has considerably smaller CV score compared to all other candidate models, and hence we only report the CV scores under this model in Table 5 with various choices of (A, λ_R). It shows that $(A, \lambda_R) = (7, 10^3)$ has the smallest approximate CV score.

Figure 7 shows the estimated gradient functions $\widehat{g}$ under the selected models for the control and treatment groups, respectively. Apart from $\widehat{g}$, we also plot the estimated pointwise two-standard error bands using (3.6). The fact that the bands are generally non-overlapping except for a small region clearly indicates that the baseline growth displacement rates for the

control and treatment groups are different. The plot also shows that there is no growth bump for either group. 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 8 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. 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 signifies the major purpose and advantage of nonparametric modeling, which is to provide insights and to suggest candidate parametric models for further studies.

In order to check how our method performs in terms of estimating individual sample trajectories, we solved the differential equation model for each plant i with fitted values of $X_{il}(0)$, θ_i and g. Figure 9 shows the fitted (under the selected model) and observed trajectories for three plants each from the control and the treatment groups. As can be seen from this figure, although there are subject-specific variabilities in the fits, the overall shapes of the trajectories are captured fairly well. Figure 10 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 proposed procedure based on minimizing the objective function (3.3) 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, 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 10) 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 at least for practical purposes. Nevertheless, 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 aspect is discussed briefly in Section 6.

6. Discussion. The model and the fitting procedures presented in this paper are quite flexible and effective in terms of modeling autonomous dynamical systems nonparametrically when the data are from a number of subjects and when the underlying population level dynamics is of interest. When applying the proposed method to the plant growth data, we obtain results that are scientifically sensible. For 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, a fact also supported by the estimates which turn out to be nonnegative and the simulation results where the resulting estimators of g are always nonnegative for the `moderate` and/or "$\boldsymbol{a}$ known" cases. In general, if g is strictly positive (strictly negative) over the domain of interest, then we can model the logarithm of g (respectively, $-g$) by basis representation.

The proposed approach is flexible in terms of incorporating various constraints on the dynamics and is able to capture features of the dynamical system which are not known to us *a priori*. It can also be extended to incorporate covariate effects, as well as to model nonautonomous systems which are currently under investigation. Even though in this paper we use the plant growth data as an illustration, the proposed framework is potentially useful to many other studies with similar types of data, where estimating the underlying dynamical system is of interest. For example, the data set collected as part of the Multicenter AIDS Cohort Study (Kaslow et al., 1987; Diggle et al., 2002) can be used to study the dynamics of the CD4+ counts. Investigating the dynamics of CD4+ counts at a population level, while also

taking into account individual effects, is of great importance to understand the progression of AIDS. This data set consists of 2376 measurements of CD4+ cell counts against time since seroconversion (time when HIV becomes detectable) for 369 infected men enrolled in the study. In this data set, each patient is a subject and there is one sample curve associated with each subject which reflects CD4+ counts over time. Moreover, each curve is only observed at a few time points and the set of measurement times is different across patients. The estimation procedure proposed in this paper can be adjusted appropriately to deal with such scenarios more effectively. Specifically, in order to deal with a large number of random effects, instead of the hierarchical likelihood approach we can adopt a marginal maximum likelihood approach. These are topics of our ongoing research.

Appendix.

Gradient of the sample trajectories. Note that, $X_{il}(\cdot)$ satisfies

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

Differentiating (A-1) with respect to the parameters, we have

$$\begin{aligned}
X_{il}^{a_{il}}(t) := \frac{\partial X_{il}(t)}{\partial a_{il}} &= 1 + \int_0^t \frac{\partial X_{il}(s)}{\partial a_{il}} e^{\theta_i} \sum_{k=1}^{M} \beta_k \phi'_{k,M}(X_{il}(s))ds \\
X_{il}^{\theta_i}(t) := \frac{\partial X_{il}(t)}{\partial \theta_i} &= \int_0^t \left[\frac{\partial X_{il}(s)}{\partial \theta_i} e^{\theta_i} \sum_{k=1}^{M} \beta_k \phi'_{k,M}(X_{il}(s)) \right. \\
&\qquad \left. + e^{\theta_i} \sum_{k=1}^{M} \beta_k \phi_{k,M}(X_{il}(s)) \right] ds \\
X_{il}^{\beta_r}(t) := \frac{\partial X_{il}(t)}{\partial \beta_r} &= \int_0^t \left[\frac{\partial X_{il}(s)}{\partial \beta_r} e^{\theta_i} \sum_{k=1}^{M} \beta_k \phi'_{k,M}(X_{il}(s)) + e^{\theta_i} \phi_{r,M}(X_{il}(s)) \right] ds
\end{aligned}$$

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

$$\frac{d}{dt} X_{il}^{a_{il}}(t) = X_{il}^{a_{il}}(t) e^{\theta_i} \sum_{k=1}^{M} \beta_k \phi'_{k,M}(X_i(t)), \quad X_{il}^{a_{il}}(0) = 1,$$

$$\frac{d}{dt} X_{il}^{\theta_i}(t) = X_{il}^{\theta_i}(t) e^{\theta_i} \sum_{k=1}^{M} \beta_k \phi'_{k,M}(X_{il}(t)) + e^{\theta_i} \sum_{k=1}^{M} \beta_k \phi_{k,M}(X_{il}(t)), \quad X_{il}^{\theta_i}(0) = 0,$$

$$\frac{d}{dt}X_{il}^{\beta_r}(t) = X_{il}^{\beta_r}(t)e^{\theta_i}\sum_{k=1}^{M}\beta_k\phi'_{k,M}(X_{il}(t)) + e^{\theta_i}\phi_{r,M}(X_{il}(t)), \quad X_{il}^{\beta_r}(0) = 0.$$

If the a_{il}'s are positive and the function $g_{\boldsymbol{\beta}} := \sum_{k=1}^{M}\beta_k\phi_{k,M}$ is positive on the domain of a_{il}'s, then the trajectories $X_{il}(t)$ are nondecreasing in t. In this case, and more generally, whenever the solutions exist on the time interval $[0,1]$ and $g_{\boldsymbol{\beta}}$ is continuously differentiable the gradients of the trajectories can be solved explicitly:

$$X_{il}^{a_{il}}(t) = \frac{g_{\boldsymbol{\beta}}(X_{il}(t))}{g_{\boldsymbol{\beta}}(X_{il}(0))}\,, \tag{A-2}$$

$$X_{il}^{\theta_i}(t) = e^{\theta_i}t g_{\boldsymbol{\beta}}(X_{il}(t))\,, \tag{A-3}$$

$$X_{il}^{\beta_r}(t) = g_{\boldsymbol{\beta}}(X_{il}(t))\int_{X_{il}(0)}^{X_{il}(t)}\frac{\phi_{r,M}(x)}{(g_{\boldsymbol{\beta}}(x))^2}dx\,. \tag{A-4}$$

We verify (A-4). Proofs (A-2) and (A-3) are similar. We can express

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

Derivation of $\widetilde{CV}$. Observe that, when evaluated at the estimate $\widehat{\boldsymbol{a}}$, $\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\beta}}$ based on the full data,

$$\frac{\partial}{\partial\theta_i}\left(\sum_{l,j}\ell_{ilj}^{cv}\right)+2\lambda_2\theta_i = 0, \quad i = 1,\cdots,n; \quad \frac{\partial}{\partial\boldsymbol{\beta}}\left(\sum_{i,l,j}\ell_{ilj}^{cv}\right)+2\mathbf{B}\boldsymbol{\beta} = 0. \tag{A-5}$$

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

$$\frac{\partial}{\partial\theta_i}\left(\sum_{l^*,j:l^*\neq l}\ell_{il^*j}^{cv}\right) + 2\lambda_2\theta_i = 0; \quad \frac{\partial}{\partial\boldsymbol{\beta}}\left(\sum_{i^*,l^*,j:(i^*,l^*)\neq(i,l)}\ell_{i^*l^*j}^{cv}\right) + 2\mathbf{B}\boldsymbol{\beta} = 0. \tag{A-6}$$

Expanding the left hand side of (A-6) around $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\theta}}$, and using (A-5), we obtain the following first order approximations:

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

In the above, the gradients and Hessians of ℓ_{ilj}^{cv} 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. Hence, there is almost no additional computational cost to obtain these approximations. Now for $i = 1, \ldots, n; l = 1, \cdots, N_i$, define

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

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

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

TABLE 1
Convergence and model selection based on 50 *independent replicates.*

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

* all the numbers are multiplied by 100

TABLE 3
Comparison of estimation accuracy of **two-stage estimators** *(either local quadratic smoothing or parametric regression using true model in the second stage) with* **hierarchical likelihood estimators** *(for the* **selected model***, among models with* $M = 2, \ldots, 6$ *B-spline basis functions) under the "sparse" case.*

Two-stage estimator

Method in stage 2	**bandwidths in stage I**	**Summary statistics**	$x \in [-0.5, 0.2]$	$x \in (0.2, 1]$	$x \in (1, 1.5]$
Local quadratic smoothing	optimal bandwidths	Mean(ISE($\widehat{g}$))	3.8×10^7	20.177	7.3×10^6
		Median(ISE($\widehat{g}$))	$\mathbf{4.1}\times10^5$	**2.398**	$\mathbf{1.8}\times10^3$
		(s.d.(ISE($\widehat{g}$)))	(2.3×10^8)	(330.146)	(5.1×10^7)
Regression (true model)	optimal bandwidths	Mean(ISE($\widehat{g}$))	27.592	28.492	0.063
		Median(ISE($\widehat{g}$))	**3.812**	**2.094**	**0.004**
		(s.d.(ISE($\widehat{g}$)))	(423.283)	(565.281)	(1.249)

Hierarchical likelihood estimator

Method	**Summary statistics**	$x \in [-0.5, 0.2]$	$x \in (0.2, 1]$	$x \in (1, 1.5]$
$\boldsymbol{a}$ known	Mean(ISE($\widehat{g}$))	0.006	0.083	0.001
	Median(ISE($\widehat{g}$))	**0.003**	**0.041**	**0.000**
	(s.d.(ISE($\widehat{g}$)))	(0.009)	(0.106)	(0.002)
$\boldsymbol{a}$ estimated	Mean(ISE($\widehat{g}$))	0.710	0.195	0.007
	Median(ISE($\widehat{g}$))	**0.025**	**0.054**	**0.000**
	(s.d.(ISE($\widehat{g}$)))	(4.751)	(0.789)	(0.048)

TABLE 4

Computational cost for the simulation study in Section 4. Reported quantities are the average time in seconds and standard deviations (within brackets) over 50 replicates (including the ones without convergence).

Model (M)		2	3	4	5	6
moderate	$\boldsymbol{a}$ known	11.40	20.34	28.14	41.51	42.29
		(0.24)	(0.68)	(0.73)	(1.53)	(2.39)
	$\boldsymbol{a}$ estimated	21.22	89.20	44.23	56.34	69.89
		(1.18)	(18.38)	(4.54)	(9.33)	(23.05)
sparse	$\boldsymbol{a}$ known	11.50	20.35	28.25	41.53	42.57
		(0.33)	(0.63)	(0.80)	(1.58)	(3.05)
	$\boldsymbol{a}$ estimated	24.01	93.58	47.06	68.57	89.57
		(1.59)	(17.02)	(11.55)	(25.08)	(38.75)

TABLE 5

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

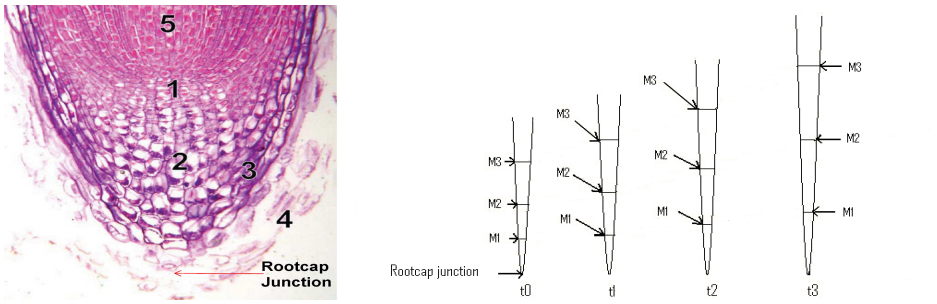

FIG 1. *Root tip. Left Panel: image of root tip with meristem*: 1 - meristem; 4 - root cap; 5 - elongation zone; Right Panel: an illustration of the root tip with the displacements of three markers $M1$, $M2$, $M3$ indicated at times t_0, t_1, t_2, t_3. (*From wikipedia)*

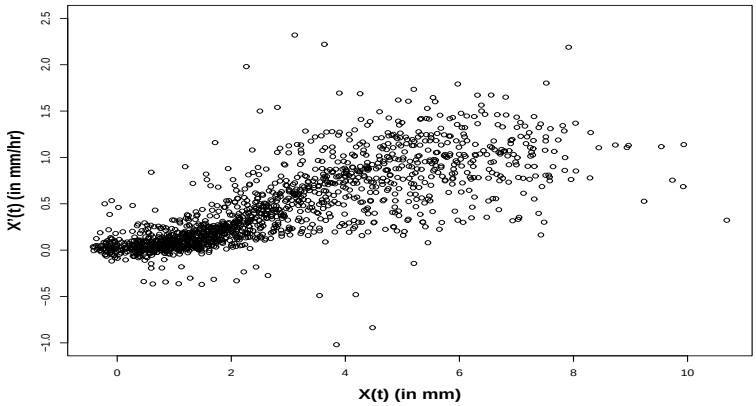

FIG 2. *Empirical derivatives $\widehat{X}'(t)$ against empirical fits $\widehat{X}(t)$ for the treatment group in the plant growth data.*

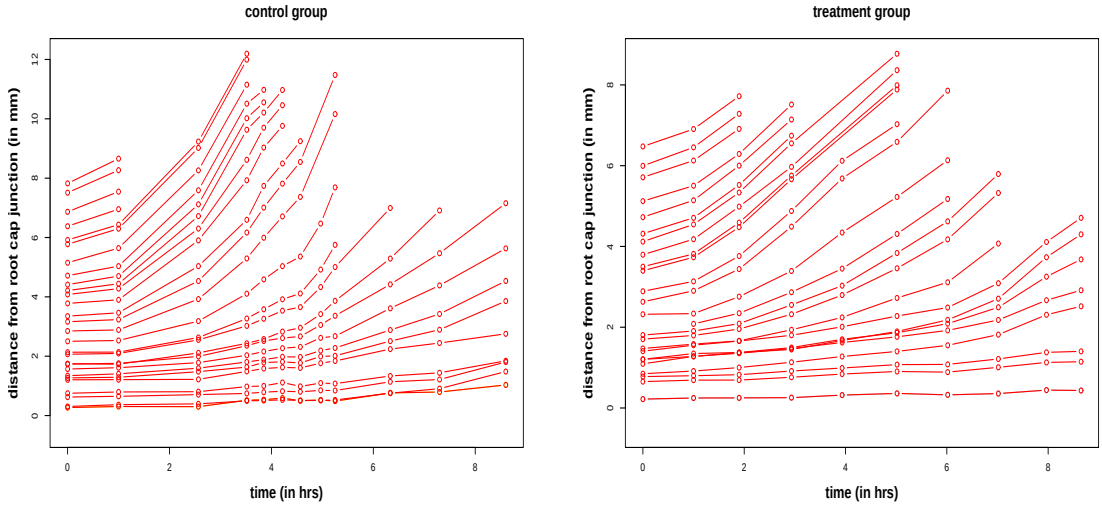

FIG 3. *Growth trajectories for plant data. Left panel: a plant in the control group; Right panel: a plant in the treatment group*

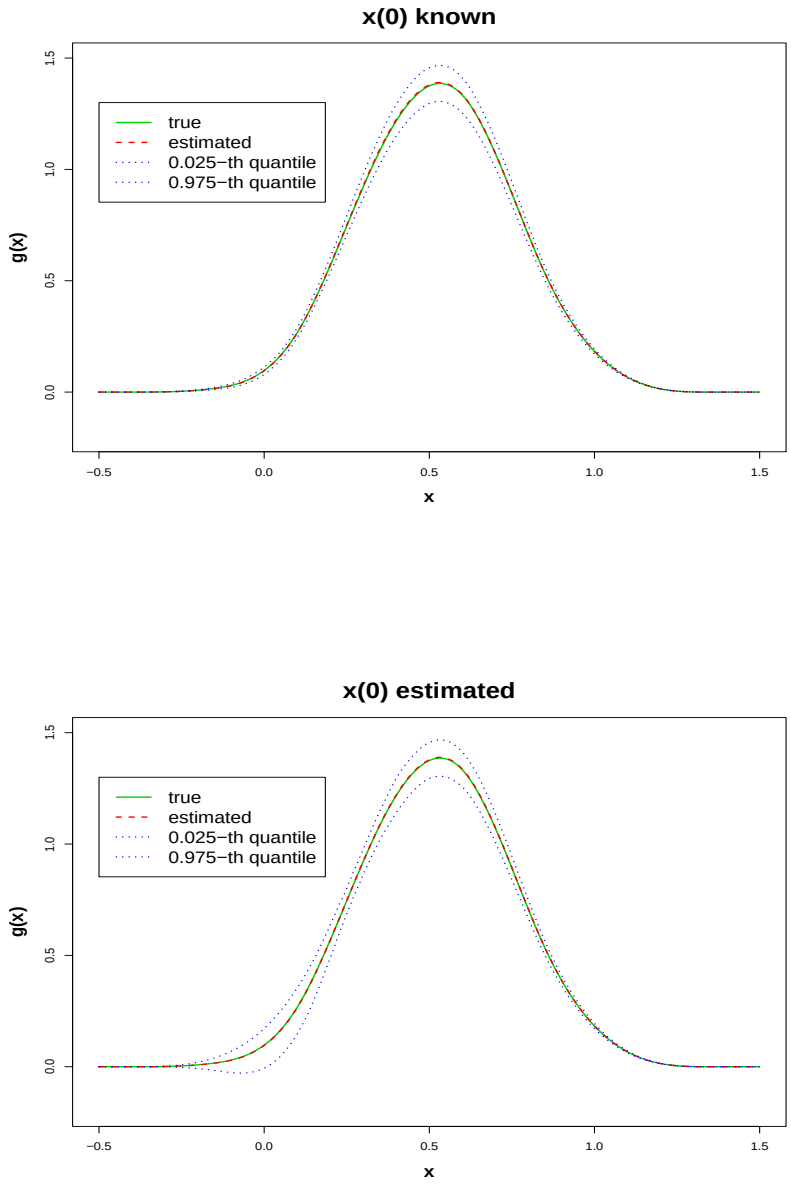

FIG 4. *True and fitted gradient function g by hierarchical likelihood approach for the* `sparse` *case. The true model (with $M = 4$ B-spline basis functions with equally spaced knots) is used in fitting. Top panel: initial conditions $\boldsymbol{a}$ are known; Bottom panel: initial conditions $\boldsymbol{a}$ are estimated.*

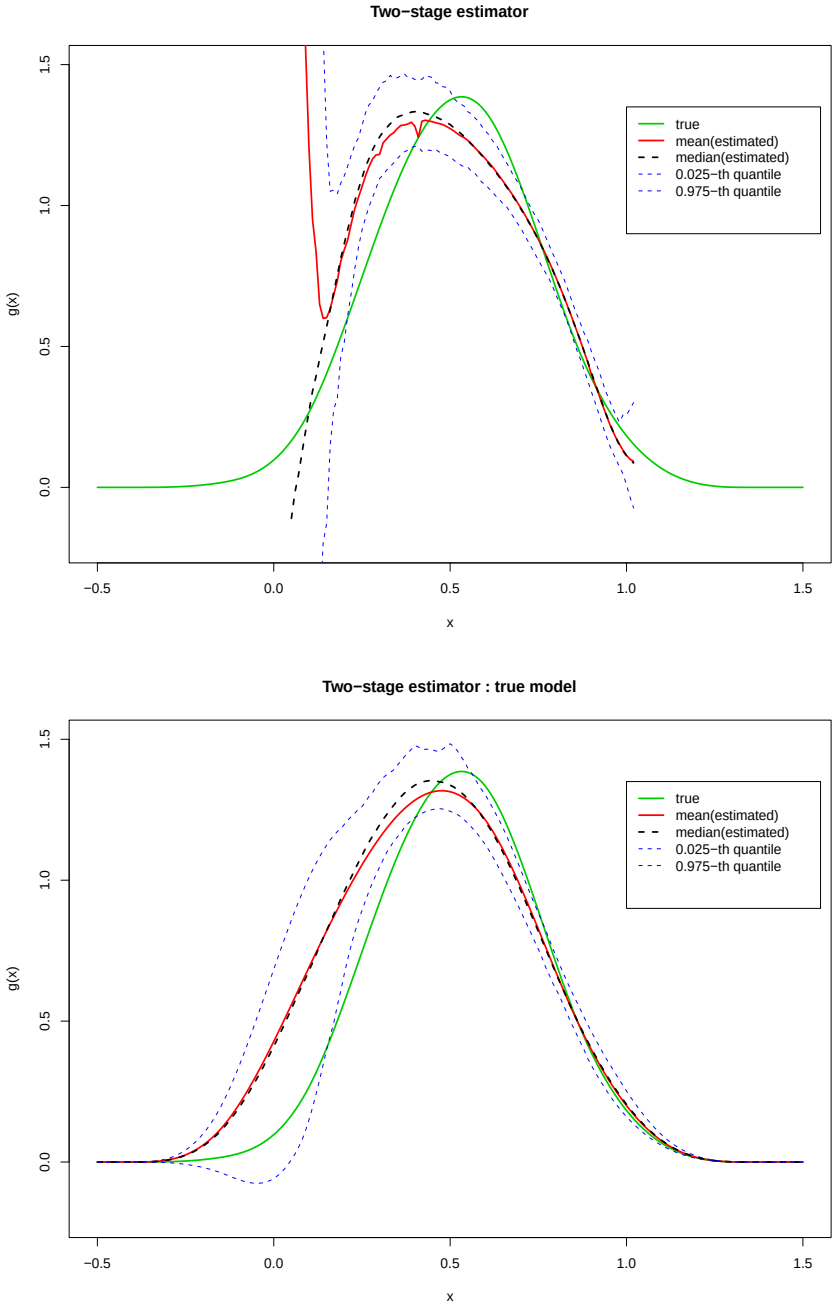

FIG 5. *True and fitted gradient function g by two-stage approach for the* **sparse** *case. Top panel: the second stage uses* **local quadratic smoothing**. *Bottom panel: the second stage uses* **regression under the true model**.

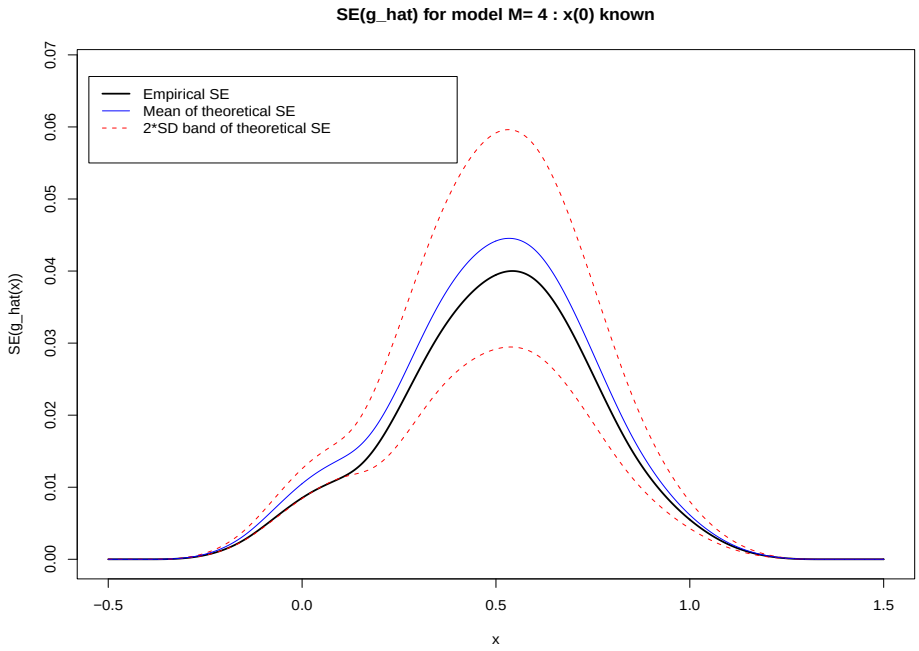

FIG 6. *Standard error estimates for the simulation study in Section 4. Pointwise standard error of $\widehat{g}$ for the* `sparse` *case with initial conditions* $\boldsymbol{a}$ *known. The true model (with* $M = 4$ *B-spline basis functions) is used in fitting. Solid black curve: pointwise standard error computed from 50 replicates; Solid blue curve: averaged pointwise standard error estimates from (3.6) (based on 50 replicates); Broken red curve: 2 standard deviations bands for the estimated pointwise standard error (based on 50 replicates).*

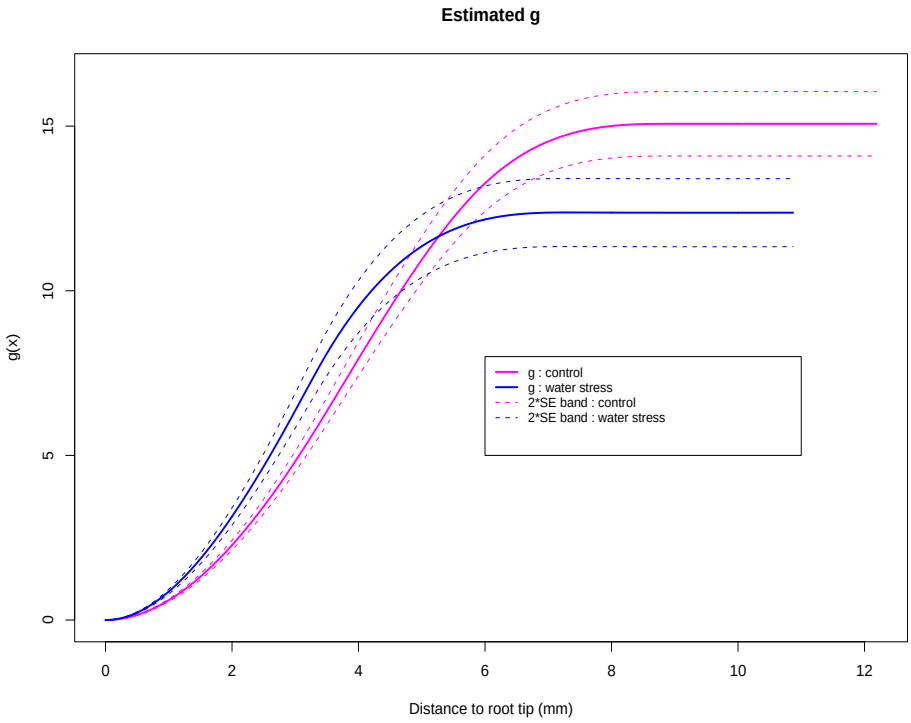

FIG 7. *Fitted gradient function $\widehat{g}$, and pointwise 2 standard error bands under the selected models for control and treatment groups.*

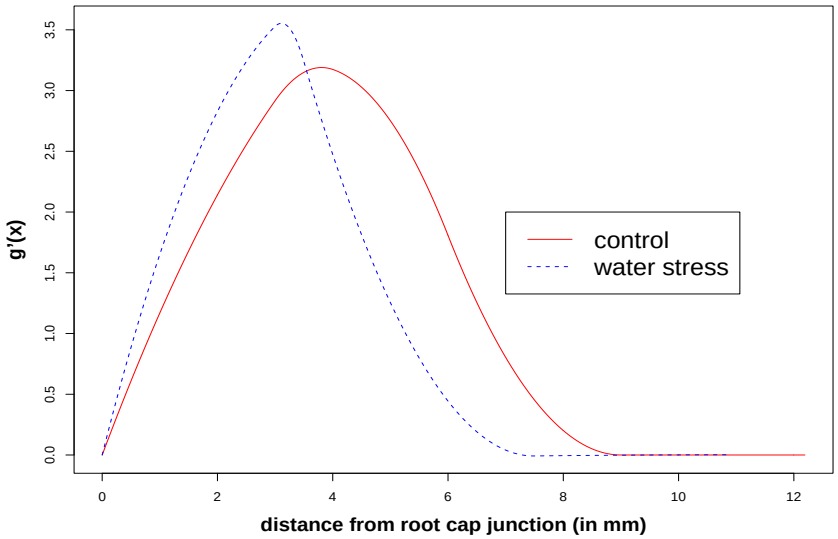

FIG 8. *Fitted relative elemental growth rate (REGR) under the selected models for control and treatment groups, respectively. The REGR is computed by differentiating the estimated gradient function g.*

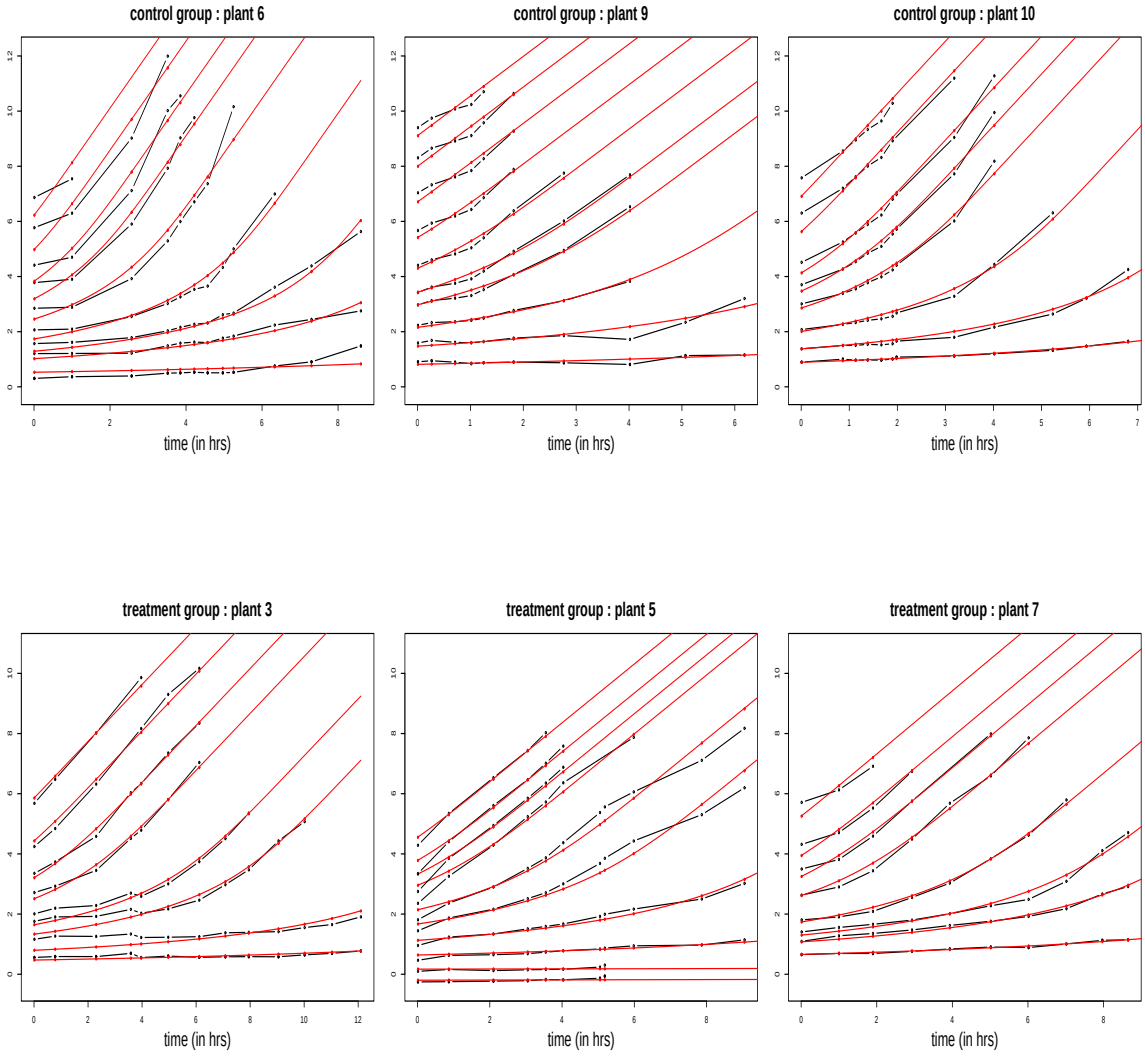

FIG 9. *Observed (black) and fitted (red) trajectories (under the selected models) for the plant data. Every third trajectory of each plant is plotted. Top panel: (from left) plant # 6, 9, 10 in the control group; Bottom panel: (from left) plant # 3, 5, 7 in the treatment (water stress) group.*

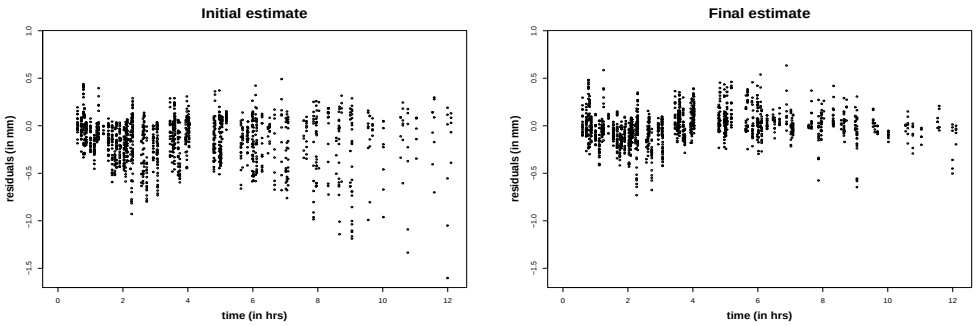

FIG 10. *Residual versus time plots for the treatment group. Left panel: fit by* `stepwise-regression`*; Right panel: fit by the proposed method based on maximizing the hierarchical likelihood.*

Acknowledgement. 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.

SUPPLEMENTARY MATERIAL

Supplement to "Semiparametric modeling of autonomous nonlinear dynamical systems with application to plant growth" (doi: ???http://lib.stat.cmu.edu/aoas/???/???; .pdf). The supplementary materials provide additional details on the computational schemes. It also contains further simulation studies elucidating the performance of the proposed estimators under scenarios not covered in the main article.

References.

Basu, P., Pal, A., Lynch, J. P. and Brown, K. M. (1998). A novel image-analysis technique for kinematic study of growth and curvature. *Plant Physiology* **145**, 305–316.

Burman, P. (1990). Estimation of generalized additive models. *Journal of Multivariate Analysis* **32**, 230–255.

Brunel, N. J.-B. (2008). Parameter estimation of ODE's via nonparametric estimators. *Electronic Journal of Statistics* **2**, 1242-1267.

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.

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

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 the American Statistical Association* **103**, 369–384.

Diggle, P. J., Heagerty, P., Liang, K.-Y., and Zeger, S. L. (2002). *Analysis of Longitudinal Data (2nd Edition)*. Oxford University Press.

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.

Gu, C. (2002). *Smoothing Spline ANOVA Models*. Springer.

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

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.

Kaslow, R. A., Ostrow, D. G., Detels, R., Phair, J. P., Polk, B. F. and Rinaldo, C. R. (1987). The Multicenter AIDS Cohort Study: rationale, organization, and selected characteristics of the participants. *American Journal of Epidemiology* **126(2)**, 310-318.

Ke, C. and Wang, Y. (2001) Semiparametric nonlinear mixed effects models and their applications. *Journal of the American Statistical Association* **96**, 1272–1281.

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

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.

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

Miao, H., Dykes, C., Demeter, L. M. and Wu, H. (2009). Differential equation modeling of HIV viral fitness experiments : model identification, model selection, and multimodel inference. *Biometrics* **65**, 292–300.

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

Nowak, M. A. and May, R. (2000). *Virus Dynamics : Mathematical Principles of Immunology and Virology*. Oxford University Press.

Paul, D., Peng, J. and Burman, P. (2009). Semiparametric modeling of autonomous nonlinear dynamical systems with applications. *Technical report*, (`http://arxiv.org/PS_cache/arxiv/pdf/0906/0906.3501v1.pdf`).

Paul, D., Peng, J. and Burman, P. (2011). Supplement to "Semiparametric modeling of autonomous nonlinear dynamical systems with application to plant growth".

Peng, J. and Paul, D. (2009). A geometric approach to maximum likelihood estimation of the functional principal components from sparse longitudinal data. *Journal of Computational and Graphical Statistics* **18**, 995-1015.

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

Pinheiro, J. C. and Bates, D. M. (2000). *Mixed-Effects Models in S and S-PLUS*. Springer.

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.

Ramsay, J. and Silverman, B. W. (2002). *Applied Functional Data Analysis*. Springer.

Ramsay, J. and Silverman, B. W. (2005). *Functional Data Analysis, 2nd Edtion*. Springer.

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.

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.

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.

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

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

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

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

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

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.

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

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.

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.

Department of Statistics
University of California
Davis, CA 95616
E-mail: debashis@wald.ucdavis.edu
jie@wald.ucdavis.edu
burman@wald.ucdavis.edu

SUPPLEMENT TO "SEMIPARAMETRIC MODELING OF AUTONOMOUS NONLINEAR DYNAMICAL SYSTEMS WITH APPLICATION TO PLANT GROWTH"

By Debashis Paul and Jie Peng and Prabir Burman

University of California, Davis

S1. Runge-Kutta method. 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:

$$\frac{d}{dt} f(t) = G(t, f(t), \eta_2), \quad f(0) = \eta_1, \quad t \in [0, 1]. \tag{S1-1}$$

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 $0 \leq m < [1/h]$), 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,

$$f(t, \boldsymbol{\eta}) = \eta_1 + \int_0^t G(s, f(s, \boldsymbol{\eta}), \eta_2) ds, \quad t \geq 0. \tag{S1-2}$$

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

$$\frac{\partial}{\partial\eta_1}f(t,\boldsymbol{\eta}) = 1 + \int_0^t \frac{\partial}{\partial\eta_1}f(s,\boldsymbol{\eta})G_f(s,f(s,\boldsymbol{\eta}),\eta_2)ds, \tag{S1-3}$$

$$\frac{\partial}{\partial\eta_2}f(t,\boldsymbol{\eta}) = \int_0^t \left[\frac{\partial}{\partial\eta_2}f(s,\boldsymbol{\eta})G_f(s,f(s,\boldsymbol{\eta}),\eta_2) + G_\eta(s,f(s,\boldsymbol{\eta}),\eta_2)\right] ds, \tag{S1-4}$$

where G_f and G_η denote the partial derivatives of G with respect to its second and third arguments, respectively. In equations (S1-3) and (S1-4), 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 (S1-2), and then using that approximate solution in place of $f(\cdot,\boldsymbol{\eta})$ in equations (S1-3) and (S1-4) 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: $\{0,2h,4h,\cdots\}$.

S2. Two-stage method for estimating the gradient function. In this section, we discuss two-stage procedures and compare them with the proposed hierarchical likelihood method. Chen and Wu (2008a,b) consider the problem of estimating the ODE model:

$$\frac{d}{dt}X(t) = F(X(t),\beta(t)) \tag{S2-1}$$

where F is a known function and $\beta(t)$ is a time-varying parameter. The observed data are of the form

$$Y_i = X(t_i) + \varepsilon_i, \quad i=1,\ldots,n,$$

where $t_1,\ldots,t_n$ are observational times. The model that Chen and Wu (2008a,b) consider is different from the model given by (1) and (2) in the paper in

that: (i) their model assumes a known from of the gradient function, except for a time varying parameter $\beta(t)$; (ii) the measurements are for one trajectory only; and (iii) model (S2-1) does not include any subject-specific effect. Thus, the two-stage estimation method suggested by Chen and Wu (2008a,b) is not directly applicable to the problems studied in this paper. However, in principle, this estimation procedure can be extended to deal with the current setting as described below.

The two-stage procedure involves first obtaining estimates of individual sample trajectories $X_{il}(t)$ and their derivatives $X'_{il}(t)$, and then using a regression approach (parametric or nonparametric) to estimate the gradient function. For the first step of the two-stage procedure, Chen and Wu (2008a) suggest using local linear and local quadratic smoothing, respectively. For the cases studies in this paper, this amounts to, for each pair (i,l), $\widehat{X}_{il}(t) := b_0(t)$ where

$$(b_0(t), b_1(t)) = \arg\min_{b_0,b_1} \sum_j K\left(\frac{t-t_{ilj}}{h_0}\right)(Y_{ilj} - b_0 - b_1(t-t_{ilj}))^2;$$

and $\widehat{X}'_{il}(t) := \widetilde{b}_1(t)$, where

$$\begin{aligned} & (\widetilde{b}_0(t), \widetilde{b}_1(t), \widetilde{b}_2(t)) \\ = \; & \arg\min_{b_0,b_1,b_2} \sum_j K\left(\frac{t-t_{ilj}}{h_1}\right)(Y_{ilj} - b_0 - b_1(t-t_{ilj}) - b_2(t-t_{ilj})^2)^2, \end{aligned}$$

where $K(\cdot)$ is a nonnegative kernel, and $h_0, h_1 > 0$ are the bandwidths. For the following simulation studies, we use the R package `locpol` to perform the local polynomial smoothing, and try both Gaussian and Epanechnikov kernels. It turns out that the estimates using the Gaussian kernel are more robust for sparse data and hence we report those results only.

If the form of g is unknown (which is the case of primary interest of this paper), in the second stage of the two-stage approach, one can adopt a nonparametric approach for estimating g, e.g., use nonparametric regression of $\widehat{X}'_{il}(t)$ on $\widehat{X}_{il}(t)$ through a local quadratic smoothing, i.e., $\widehat{g}(x) = a_0(x)$ where $(a_0(x), a_1(x), a_2(x))$ is the minimizer of

(S2-2)

$$\sum_i \sum_l \sum_j K\left(\frac{x - \widehat{X}_{il}(t_{ilj})}{\widetilde{h}}\right)(\widehat{X}'_{il}(t_{ilj}) - a_0 - a_1(x - \widehat{X}_{il}(t_{ilj})) - a_2(x - \widehat{X}'_{il}(t_{ilj}))^2)$$

where $\widetilde{h} > 0$ is the bandwidth. If the gradient function g is parameterized, e.g., $g = g_{\boldsymbol{\beta}} = \sum_{k=1}^M \beta_k \phi_k$ where $\phi_k(\cdot)$ are known basis functions, the pa-

rameter $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_M)^T$ can be estimated by nonlinear regression, i.e.,

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_i \sum_l \sum_j (\widehat{X}'_{il}(t_{ilj}) - g_{\boldsymbol{\beta}}(\widehat{X}_{il}(t_{ilj})))^2. \tag{S2-3}$$

Then $\widehat{g}_{\boldsymbol{\beta}} := \sum_{k=1}^M \widehat{\beta}_k \phi_k$ is an analog of the *two-step local hybrid estimator* of g proposed by Chen and Wu (2008a). Note that, this way, the second two-stage procedure eliminates any possible source of model bias from the fit and the problem becomes parametric. The reason to study such a procedure is to better understand the behavior of the two-stage method, especially the effect of the first stage smoothing on the estimate of the gradient function g. In the simulation studies, we consider both versions of the two-stage procedure, i.e., in the second stage, when g is estimated nonrparametrically by (S2-2), and when g is estimated parametrically by (S2-3).

S2.1. *Comparison of two-stage methods with hierarchical likelihood method.* As we have pointed out, the models studied in Chen and Wu (2008a,b) do not involve any random effects. Thus, they are most appropriately applicable to the case when the gradient function is the same across subjects. Hence, we carry out simulations where $\theta_i \equiv 0$ (cf. Tables S5-4 and S5-5) in order to make a fair comparison with the two-stage methods, even though this case falls outside of the regime that we are most interested in in this paper. We also compare the two-stage procedure with the proposed method when $\sigma_\theta = 0.1$ (cf. Table S5-3 which is the same as Table 3 of the main text). Under this setting, the two-stage smoother is likely to have some bias contribution from the fact that the functions $e^{\theta_i} g$ vary across subjects. This bias is indeed fairly small since $\mathbb{E}(e^{\theta_i}) = \exp(\sigma_\theta^2/2) \approx 1.005$. Therefore, we believe that both settings give a fair comparison.

As mentioned in the main text of the paper, while reporting the risk of the estimators, we divide the domain of x into three regions, $[-0.5, 0.2]$, $(0.2, 1]$ and $(1, 1.5]$. Under our simulation setting, even though the true gradient function g has support effectively on $[-0.5, 1.5]$, the observed measurements of Y_{ilj}'s are almost entirely confined in the region $(0.2, 1]$ (hereafter, referred to as the *data domain*). Due to the extrapolation effect, one would not expect any estimator that does not use the true initial conditions to perform very well in the domains where there is no data. Thus, we divide the domain into different regions to gain a better understanding of the performance of each method, as well as for a more informative comparison across different methods.

For all simulations carried out in this subsection, the true gradient function is the same as that used in Section 4 of the paper, i.e., g is represented

in $M_* = 4$ cubic B-spline basis functions with knots at $(0.35, 0.6, 0.85, 1.1)$ and the vector of basis coefficients $\boldsymbol{\beta} = (0.1, 1.2, 1.6, 0.4)^T$. Moreover, except for the standard deviation σ_θ for the random scale parameter, all other parameters are kept the same as in Section 4 of the paper. For example, the number of subjects is $n = 10$ and the number of trajectories (replicates) per subject is $N_i \equiv 20$. Moreover, we consider two settings based on the number of measurements m_{il} per trajectory: **sparse** – m_{il} i.i.d. Uniform$[3, 8]$; and **moderate** – m_{il} i.i.d. Uniform$[5, 20]$. Two different values of σ_θ, viz., 0 and 0.1, are considered. The model space for search consists of functions represented by $M = 2, \ldots, 6$ cubic B-spline basis functions with knots at $0.1 + j/M$, for $j = 1, \ldots, M$.

The main findings and some discussions are given below, with numerical and graphical summaries given in Table S5-3 (for $\sigma_\theta = 0.1$) and Table S5-4 (for $\sigma_\theta = 0$); and Figures S5-1, S5-2 (for $\sigma_\theta = 0.1$) and Figures S5-3 to S5-8 (for $\sigma_\theta = 0$).

- The proposed hierarchial likelihood method gives considerably more accurate estimates than the two-stage procedure.
- The choice of bandwidths in the presmoothing stage of the two-stage method plays a big role. Following the prescription in Chen and Wu (2008a), we consider a bandwidth selected by cross validation for each of the sample curves, which is henceforth referred to as the *optimal bandwidth*. This method results in high bias, which does not go away even if the number of measurements per curve becomes very large (see Table S5-5 for the simulation described in the next subsection). In contrast, the proposed method shows much less bias. In addition, compared to the proposed method, the two-stage approach shows much more variability, even within the data domain $(0.2, 1]$ and when $\sigma_\theta = 0$ (see e.g., Table S5-4). These hold irrespective of whether the second stage is through local quadratic smoothing, or by regression using the true model. Although, for the latter scenario, the two-stage estimates of g are smoother and behave better near the boundaries.
- Brunel (2008) considers a two-stage estimation scheme under a general parametric ODE setting whereby the parameter is estimated by a nonlinear regression approach, after first estimating the trajectory $X(t)$ and its derivative $X'(t)$. Based on the analysis in that paper, it is anticipated that the optimal choice of bandwidth in terms of smoothing individual sample trajectories and their derivatives leads to a bias when estimating the gradient function g in the second stage. This bias is likely to decrease if smaller bandwidths are used in the first stage, however at the expense of increased variability. Keeping this in mind,

we also consider two-stage methods where the bandwidths in the first stage of smoothing are "optimal" within a restricted range (here, 0.05 to 0.15) compared to the "optimal" bandwidths over the entire range of possible values. The resulting estimates of g do have slightly smaller bias, but with an increased variability, which is reflected in the wider 95% pointwise percentile bands (e.g., compare Figure S5-3 to Figure S5-5 and compare Figure S5-4 to Figure S5-6). Based on both theoretical analysis and numerical results, it appears that it is very difficult to achieve an optimal bias-variance trade-off by choosing bandwidths for two-stage procedures.

- The two-stage procedures involve first estimating the trajectories and their derivatives and then estimating the dynamical system based on these pre-smoothed curves, whereas the proposed approach not only combines information across all the trajectories but also estimates the dynamics and the trajectories simultaneously which is crucial when only sparse measurements are available. With only relatively sparse measurements for each curve, the estimated trajectories obtained by pre-smoothing contain a substantial amount of bias. Thus, smoothing every curve and then using the estimated trajectory of these smoothed curves to solve the differential equations (as is done in the two-stage procedure) will lead to biases in the parameter estimates. This is confirmed by the simulation results here.

S2.2. *Comparison of two-stage methods and hierarchical likelihood method when there is one curve per subject with dense measurements.* In this subsection, we consider a simulation setting where m_{il} are i.i.d. Uniform$[60, 100]$ (referred to as "very dense"). Moreover, the number of subjects is $n = 10$ with only one curve per subject (i.e., $N_i \equiv 1$) and $\sigma_\theta = 0$, i.e., there is no subject specific variability. The true gradient function g is the same as previous simulations and the model space for search is also the same as that in the previous subsection. The purpose of this simulation is to examine the two-stage method when the first stage smoothing can be done very accurately (here due to the large number of measurements per curve). The results are summarized in Table S5-5 and Figures S5-9 to S5-12. The hierarchical likelihood method still gives much more accurate estimates and it is less biased (whether the initial conditions are estimated or not). Indeed, even for this simulation setting, the bias of the two-stage methods does not go away even if bandwidths smaller than the "optimal bandwidth" selected by cross validation are used in the first stage and when the true model is used in the second stage.

S3. Additional simulations. In order to further explore the proposed estimation and model selection procedures, we carry out two additional sets of simulation studies. In the first simulation, the gradient function g is not in the model space for search and there is a fairly large model bias. For the second simulation, the true g has an interpretable parametric form.

S3.1. *When the true g is not in the model space.* In this simulation study, the true gradient function g is expressed as a linear combination of 10 B-spline basis functions with knots at $0.1 + 0.1j$ for $j = 1, \ldots, 10$. The corresponding basis coefficients are $\boldsymbol{\beta} = (0.1, 0.5, 1.5, 2.0, 1.3, 0.8, 0.5, 1.5, 0.6, 0.4)^T$. While estimating g by the proposed procedure, we consider the models with $M = 2, \ldots, 8$ cubic B-spline basis functions with equally spaced knots placed at $0.1 + j/M$ for $j = 1, \ldots, M$. Note that, the true model is not included in this model space and thus there is a model bias. Indeed, the gradient function g has two bumps (Figure S5-14) and cannot be approximated very well by any of the models considered, with the model with $M = 7$ B-splines giving the best approximation. We thus refer to this case as the "challenging case". We again consider two settings for the number of measurements m_{il} per trajectory: **sparse** – m_{il} i.i.d. Uniform$[3, 8]$; and **moderate** – m_{il} i.i.d. Uniform$[5, 20]$. All other settings are the same as the simulation considered in Section 4 of the paper, e.g., $\sigma_\theta = 0.1, n = 10, N_i \equiv 20$.

As can be seen from Tables S5-7 (sparse case) and S5-8 (moderate case), as well as Figures S5-14 (spare case with known initial conditions) and S5-15 (sparse case with estimated initial conditions), if we restrict our attention to the data domain (i.e. $(0.2, 1]$), the proposed hierarchical likelihood estimator performs comparably whether the initial conditions are estimated or not. Only in the region $[-0.5, 0.2]$, the estimator with initial conditions estimated shows excessive variability compared to the estimator using true initial conditions. This is attributed to the effect of extrapolation, due to which we do not expect any method which does not use the true initial conditions to work very well in regions without data. Note that similar effects are not observed on region $(1, 1.5]$ even though there is also no data on this domain. This is due to properties of initial value problems whereby information up to 1 suffices to give a reasonable estimate on a right side neighborhood of 1. All these are also observed in simulations performed in Section S2.

We also perform the two-stage methods under this simulation setting. Again, the two-stage estimators are highly biased and variable compared to the proposed hierarchical likelihood estimator on all three regions. This has been observed throughout in all simulations considered in the paper and Supplementary Materials.

S3.2. *When g has a shape similar to the real data.* In this section, we carry out a simulation study where the true gradient function g has a parametric form that mimics the shape of the estimated gradient function for the plant data (Section 5 of the paper). Specifically, g is chosen to be a logistic function:

$$g(x) = \frac{c}{1 + a\exp(-bx)}\ , \qquad x \in \mathbb{R}$$

where $a, b, c > 0$. In the simulation, we set $a = 10$, $b = 7$ and $c = 1$. The simulated data are fitted using the cubic spline basis $\{x, x^2, x^3\} \cup \{(x - k_l)_+^3 : l = 1, \ldots, M\}$, where $\{k_l\}_{l=1}^M$ is a knot sequence (which will be selected based on the data). Note that we intentionally omit the constant term in the basis in order to mimic the real data application. This actually leads to a model mis-specification since the true gradient function g for this simulation setting is not zero at $x = 0$. Therefore, this is another instance in which the true function g is not contained in the model space. In this case, boundary constraints are also imposed at both ends of the domain (i.e., for small as well as large x) to stabilize the fit and get good estimates. The true g and estimated g, starting from a rather crude initial estimate, are shown in Figure S5-13 where the knot sequence is selected to be $(-0.40, 0.10, 0.38, 0.66, 0.94, 1.22)$ and the boundary constraints are $g'(x) = 0$ for $x \leq -2$ and $x \geq 1$. The result indicates that with some knowledge of the boundaries of g, through a reasonable choice of boundary constraints, a fairly low dimensional nonparametric model can effectively estimate the gradient functions g that are smooth and monotone increasing.

S4. Estimation errors for the fitted model in real data application. We report the sum of squared errors in estimating the trajectories corresponding to different plants in the control and treatment groups in Tables S5-1 and S5-2. The initial conditions are estimated from the data, and the estimated θ_i's and g are used to generate the fitted trajectories (shown in Figure 9 of the paper).

TABLE S5-1
Plant-specific sum of squared error (SSE) of the trajectory estimates : control group

plant ID	1	2	3	4	5	6	7	8	9	10	Total
SSE	1.560	6.580	1.233	2.213	1.416	22.462	1.947	1.871	5.732	7.590	52.604

TABLE S5-2
Plant-specific sum of squared error (SSE) of the trajectory estimates : treatment (water stress) group

plant ID	1	2	3	4	5	6	7	8	9	Total
SSE	1.805	6.987	7.519	8.019	9.341	11.460	3.579	11.844	3.944	64.497

S5. Tables and Figures.

TABLE S5-3

"Sparse" measurements with $\sigma_\theta = 0.1$. Estimation accuracy for **two-stage estimators** *(both local quadratic smoothing and parametric regression using the true model in the second stage) and* **hierarchical likelihood estimators** *(for the selected model based on $\widetilde{CV}$, among models with $M = 2, \ldots, 6$ B-spline basis functions with equally spaced knots). True g is represented in $M_* = 4$ cubic B-spline basis functions with knots at $(0.35, 0.6, 0.85, 1.1)$ and $\boldsymbol{\beta} = (0.1, 1.2, 1.6, 0.4)^T$.*

Two-stage estimator					
Method in stage 2	**bandwidths in stage I**	**Summary statistics**	$x \in [-0.5, 0.2]$	$x \in (0.2, 1]$	$x \in (1, 1.5]$
Local quadratic smoothing	optimal bandwidths	Mean(ISE($\widehat{g}$))	3.8×10^7	20.177	7.3×10^6
		Median(ISE($\widehat{g}$))	$\mathbf{4.1}\times10^5$	**2.398**	$\mathbf{1.8}\times10^3$
		(s.d.(ISE($\widehat{g}$)))	(2.3×10^8)	(330.146)	(5.1×10^7)
Regression (true model)	optimal bandwidths	Mean(ISE($\widehat{g}$))	27.592	28.492	0.063
		Median(ISE($\widehat{g}$))	**3.812**	**2.094**	**0.004**
		(s.d.(ISE($\widehat{g}$)))	(423.283)	(565.281)	(1.249)
Hierarchical likelihood estimator					
Method		**Summary statistics**	$x \in [-0.5, 0.2]$	$x \in (0.2, 1]$	$x \in (1, 1.5]$
$\boldsymbol{a}$ known		Mean(ISE($\widehat{g}$))	0.006	0.083	0.001
		Median(ISE($\widehat{g}$))	**0.003**	**0.041**	**0.000**
		(s.d.(ISE($\widehat{g}$)))	(0.009)	(0.106)	(0.002)
$\boldsymbol{a}$ estimated		Mean(ISE($\widehat{g}$))	0.710	0.195	0.007
		Median(ISE($\widehat{g}$))	**0.025**	**0.054**	**0.000**
		(s.d.(ISE($\widehat{g}$)))	(4.751)	(0.789)	(0.048)

TABLE S5-4

"Sparse" measurements with $\sigma_\theta = 0$. Estimation accuracy for **two-stage estimators** *(both local quadratic smoothing and parametric regression using the true model in the second stage) and* **hierarchical likelihood estimators** *(for the selected model, among models with $M = 2, \ldots, 6$ cubic B-spline basis functions). True g is represented in $M_* = 4$ cubic B-spline basis functions with knots at $(0.35, 0.6, 0.85, 1.1)$ and $\boldsymbol{\beta} = (0.1, 1.2, 1.6, 0.4)^T$.*

Two-stage estimator					
Method in stage 2	**bandwidths in stage 1**	**Summary statistics**	$x \in [-0.5, 0.2]$	$x \in (0.2, 1]$	$x \in (1, 1.5]$
Local quadratic smoothing	optimal bandwidths	Mean(ISE($\widehat{g}$))	7.6×10^7	3.525	720209.984
		Median(ISE($\widehat{g}$))	$\mathbf{5.3}\times10^5$	**2.428**	$\mathbf{3.9}\times10^4$
		(s.d.(ISE($\widehat{g}$)))	(4.8×10^8)	(10.683)	(6.5×10^7)
	bandwidths in [0.05,0.15]	Mean(ISE($\widehat{g}$))	3.8×10^7	132.645	8.5×10^6
		Median(ISE($\widehat{g}$))	$\mathbf{9.3}\times10^4$	**0.912**	**858.182**
		(s.d.(ISE($\widehat{g}$)))	(2.9×10^8)	(2.4×10^4)	(7.0×10^7)
Regression (true model)	optimal bandwidths	Mean(ISE($\widehat{g}$))	9.739	3.015	0.007
		Median(ISE($\widehat{g}$))	**3.999**	**2.098**	**0.002**
		(s.d.(ISE($\widehat{g}$)))	(101.000)	(13.686)	(0.037)
	bandwidths in [0.05,0.15]	Mean(ISE($\widehat{g}$))	39.369	30.299	0.325
		Median(ISE($\widehat{g}$))	**0.866**	**0.684**	**0.003**
		(s.d.(ISE($\widehat{g}$)))	(440.981)	(459.817)	(5.852)
Hierarchical likelihood estimator					
Method		**Summary statistics**	$x \in [-0.5, 0.2]$	$x \in (0.2, 1]$	$x \in (1, 1.5]$
$\boldsymbol{a}$ known		Mean(ISE($\widehat{g}$))	0.0036	0.0032	0.0009
		Median(ISE($\widehat{g}$))	**0.0010**	**0.0014**	**0.0000**
		(s.d.(ISE($\widehat{g}$)))	(0.0070)	(0.0050)	(0.0021)
$\boldsymbol{a}$ estimated		Mean(ISE($\widehat{g}$))	0.669	0.085	0.003
		Median(ISE($\widehat{g}$))	**0.636**	**0.083**	**0.003**
		(s.d.(ISE($\widehat{g}$)))	(0.341)	(0.035)	(0.003)

TABLE S5-5

"Very dense" measurements with $\sigma_\theta = 0$. Estimation accuracy for **two-stage estimators** *(both local quadratic smoothing and parametric regression using the true model in the second stage) and* **hierarchical likelihood estimators** *(for the selected model, among models with $M = 2, \ldots, 6$ B-spline basis functions). True g is represented in $M_* = 4$ cubic B-spline basis functions with knots at $(0.35, 0.6, 0.85, 1.1)$ and $\boldsymbol{\beta} = (0.1, 1.2, 1.6, 0.4)^T$.*

Two-stage estimator					
Method in stage 2	**bandwidths in stage I**	**Summary statistics**	$x \in [-0.5, 0.2]$	$x \in (0.2, 1]$	$x \in (1, 1.5]$
Local quadratic smoothing	optimal bandwidths	Mean(ISE($\widehat{g}$))	3.3×10^7	0.626	3.5×10^6
		Median(ISE($\widehat{g}$))	$\mathbf{1.2}\times10^6$	**0.292**	$\mathbf{3.2}\times10^5$
		(s.d.(ISE($\widehat{g}$)))	(1.4×10^8)	(4.699)	(1.0×10^7)
	bandwidths in [0.05,0.15]	Mean(ISE($\widehat{g}$))	2.1×10^7	0.357	4.7×10^6
		Median(ISE($\widehat{g}$))	$\mathbf{8.9}\times10^5$	**0.291**	$\mathbf{2.3}\times10^5$
		(s.d.(ISE($\widehat{g}$)))	(7.7×10^7)	(0.538)	(2.0×10^7)
	bandwidths in [0.01,0.1]	Mean(ISE($\widehat{g}$))	4.4×10^7	0.722	1.8×10^7
		Median(ISE($\widehat{g}$))	$\mathbf{1.2}\times10^6$	**0.279**	$\mathbf{3.6}\times10^5$
		(s.d.(ISE($\widehat{g}$)))	(1.7×10^8)	(5.077)	(1.4×10^8)
Regression (true model)	optimal bandwidths	Mean(ISE($\widehat{g}$))	0.218	0.190	0.003
		Median(ISE($\widehat{g}$))	**0.097**	**0.182**	**0.003**
		(s.d.(ISE($\widehat{g}$)))	(0.328)	(0.058)	(0.001)
	bandwidths in [0.05,0.15]	Mean(ISE($\widehat{g}$))	0.206	0.198	0.003
		Median(ISE($\widehat{g}$))	**0.089**	**0.189**	**0.003**
		(s.d.(ISE($\widehat{g}$)))	(0.314)	(0.057)	(0.001)
	bandwidths in [0.01,0.1]	Mean(ISE($\widehat{g}$))	0.226	0.179	0.003
		Median(ISE($\widehat{g}$))	**0.097**	**0.172**	**0.002**
		(s.d.(ISE($\widehat{g}$)))	(0.344)	(0.057)	(0.001)
Hierarchical likelihood estimator					
Method		**Summary statistics**	$x \in [-0.5, 0.2]$	$x \in (0.2, 1]$	$x \in (1, 1.5]$
$\boldsymbol{a}$ known		Mean(ISE($\widehat{g}$))	0.0194	0.0051	0.0008
		Median(ISE($\widehat{g}$))	**0.0027**	**0.0023**	**0.0000**
		(s.d.(ISE($\widehat{g}$)))	(0.0693)	(0.0076)	(0.0019)
$\boldsymbol{a}$ estimated		Mean(ISE($\widehat{g}$))	0.1638	0.0182	0.0015
		Median(ISE($\widehat{g}$))	**0.0349**	**0.0092**	**0.0000**
		(s.d.(ISE($\widehat{g}$)))	(0.5989)	(0.0398)	(0.0028)

TABLE S5-6
"Challenging" case with $\sigma_\theta = 0.1$. Convergence and model selection based on 50 independent replicates.

		$\boldsymbol{a}$ known							$\boldsymbol{a}$ estimated						
	Model	2	3	4	5	6	7	8	2	3	4	5	6	7	8
`moderate`	# converged	50	50	47	37	47	49	49	50	50	43	5	44	49	46
	# selected	0	1	0	0	0	49	0	0	1	0	0	6	32	11
`sparse`	# converged	50	50	50	50	50	50	49	50	50	50	21	49	50	46
	# selected	0	0	0	0	5	45	0	0	0	0	0	8	34	8

TABLE S5-7
"Challenging" case with "sparse" measurements and $\sigma_\theta = 0.1$. Estimation accuracy for hierarchical likelihood-based estimator (for the selected model) and two-stage nonparametric estimator.

Method	**Summary statistics**	$x \in [-0.5, 0.2]$	$x \in (0.2, 1]$	$x \in (1, 1.5]$
Hierarchical likelihood : $\boldsymbol{a}$ known	Mean(ISE($\widehat{g}$))	0.057	0.344	0.007
	Median(ISE($\widehat{g}$))	**0.035**	**0.317**	**0.007**
	(s.d.(ISE($\widehat{g}$)))	(0.061)	(0.108)	(0.003)
Hierarchical likelihood : $\boldsymbol{a}$ estimated	Mean(ISE($\widehat{g}$))	1.945	0.514	0.007
	Median(ISE($\widehat{g}$))	**0.882**	**0.513**	**0.007**
	(s.d.(ISE($\widehat{g}$)))	(2.628)	(0.201)	(0.003)
Two-stage : optimal bandwidths	Mean(ISE($\widehat{g}$))	1.8×10^8	13.125	8.5×10^6
	Median(ISE($\widehat{g}$))	$\mathbf{5.2}\times10^5$	**6.480**	**266.527**
	(s.d.(ISE($\widehat{g}$)))	(6.2×10^7)	(45.498)	(4.5×10^7)

TABLE S5-8
"Challenging" case with "moderate" measurements and $\sigma_\theta = 0.1$. Estimation accuracy for hierarchical likelihood-based estimator (for the selected model) and two-stage nonparametric estimator.

Method	**Summary statistics**	$x \in [-0.5, 0.2]$	$x \in (0.2, 1]$	$x \in (1, 1.5]$
Hierarchical likelihood : $\boldsymbol{a}$ known	Mean(ISE($\widehat{g}$))	0.214	0.464	0.010
	Median(ISE($\widehat{g}$))	**0.028**	**0.288**	**0.007**
	(s.d.(ISE($\widehat{g}$)))	(1.252)	(1.038)	(0.019)
Hierarchical likelihood : $\boldsymbol{a}$ estimated	Mean(ISE($\widehat{g}$))	2.193	0.632	0.011
	Median(ISE($\widehat{g}$))	**0.419**	**0.403**	**0.007**
	(s.d.(ISE($\widehat{g}$)))	(7.277)	(1.160)	(0.028)
Two-stage : optimal bandwidths	Mean(ISE($\widehat{g}$))	4.3×10^7	6.002	6.2×10^5
	Median(ISE($\widehat{g}$))	$\mathbf{5.9}\times10^6$	**6.011**	**736.174**
	(s.d.(ISE($\widehat{g}$)))	(1.3×10^8)	(0.430)	(3.7×10^6)

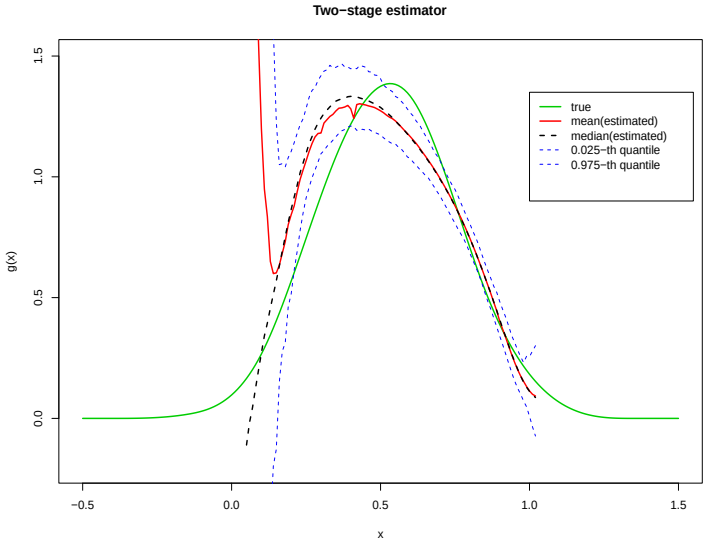

FIG S5-1. **Sparse** *case with* $\sigma_\theta = 0.1$*: true gradient function* g *and the* **Two-stage estimator** *where the second stage uses* **local quadratic smoothing** *and the first stage uses optimal bandwidths.*

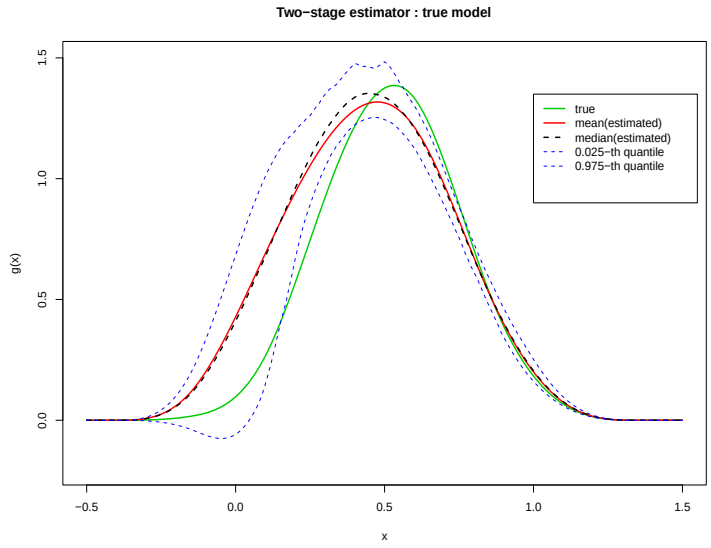

FIG S5-2. **Sparse** *case with* $\sigma_\theta = 0.1$*: true gradient function* g *and the* **Two-stage estimator** *where the second stage uses* **regression under the true model** *and the first stage uses optimal bandwidths.*

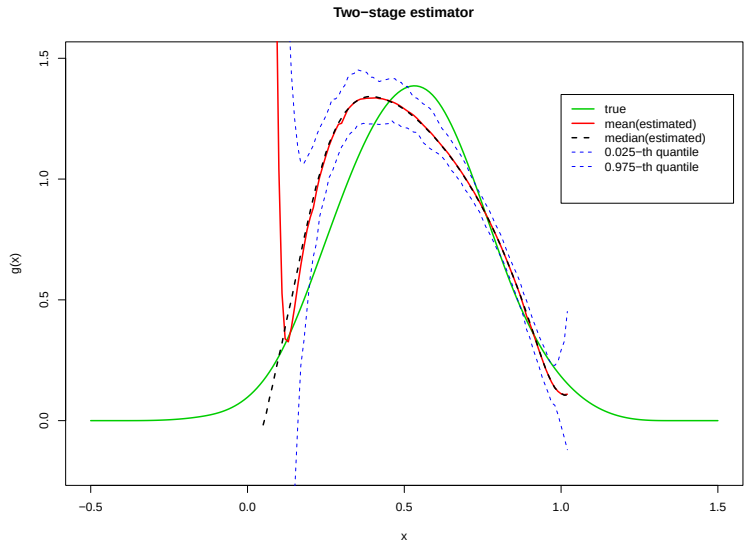

FIG S5-3. **Sparse** *case with* $\sigma_\theta = 0$*: true gradient function* g *and the* **Two-stage estimator** *where the second stage uses* **local quadratic smoothing** *and the first stage uses optimal bandwidths.*

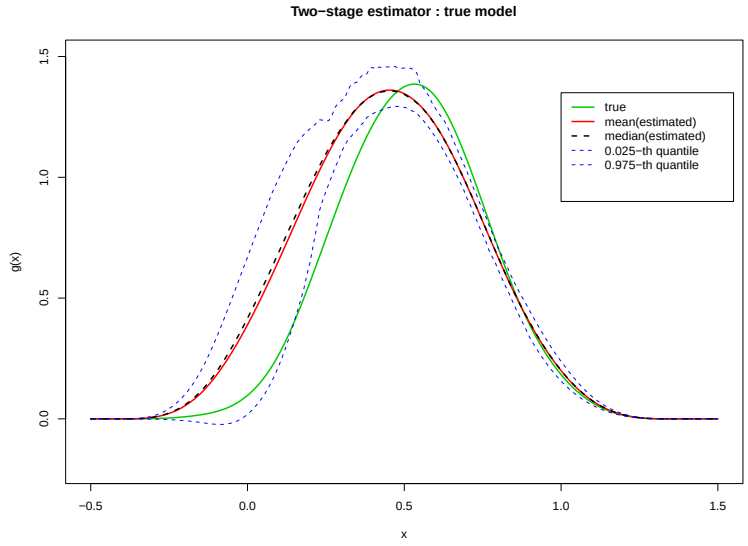

FIG S5-4. **Sparse** *case with* $\sigma_\theta = 0$*: true gradient function* g *and the* **Two-stage estimator** *where the second stage uses* **regression under the true model** *and the first stage uses optimal bandwidths.*

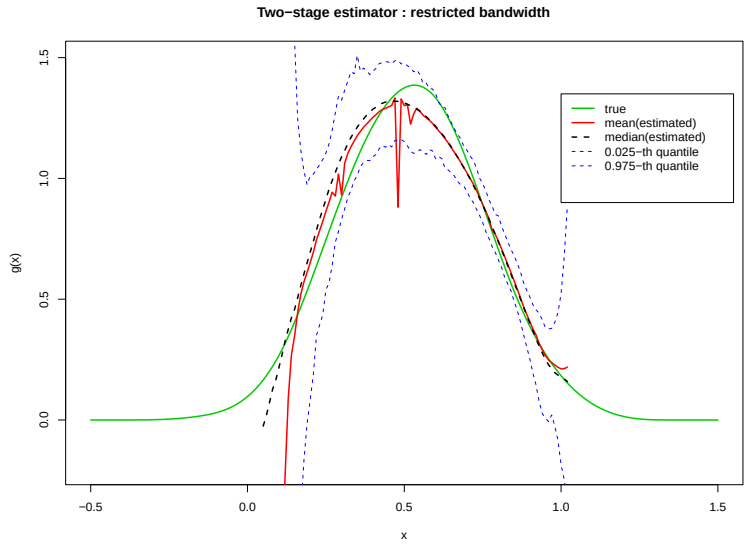

FIG S5-5. **Sparse** *case with* $\sigma_\theta = 0$*: true gradient function* g *and the* **Two-stage estimator** *where the second stage uses* **local quadratic smoothing** *and the first stage bandwidths are restricted to* $[0.05, 0.15]$.

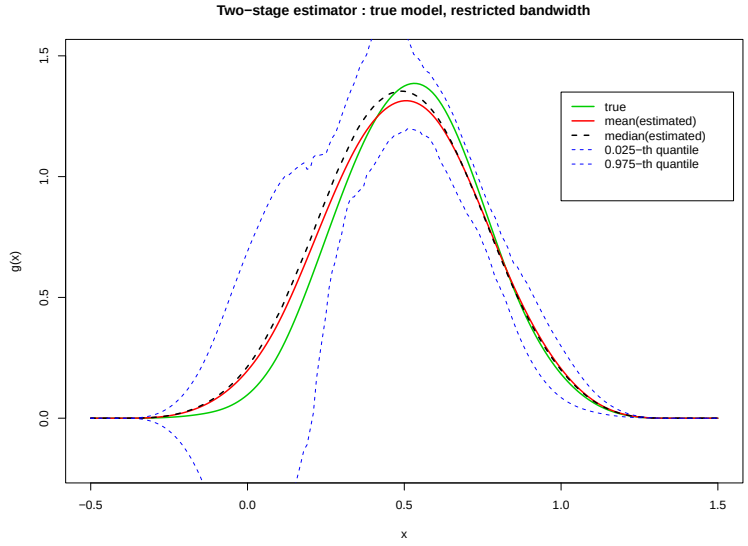

FIG S5-6. **Sparse** *case with* $\sigma_\theta = 0$*: true gradient function* g *and the* **Two-stage estimator** *where the second stage uses* **regression under the true model** *and the first stage bandwidths are restricted to* $[0.05, 0.15]$.

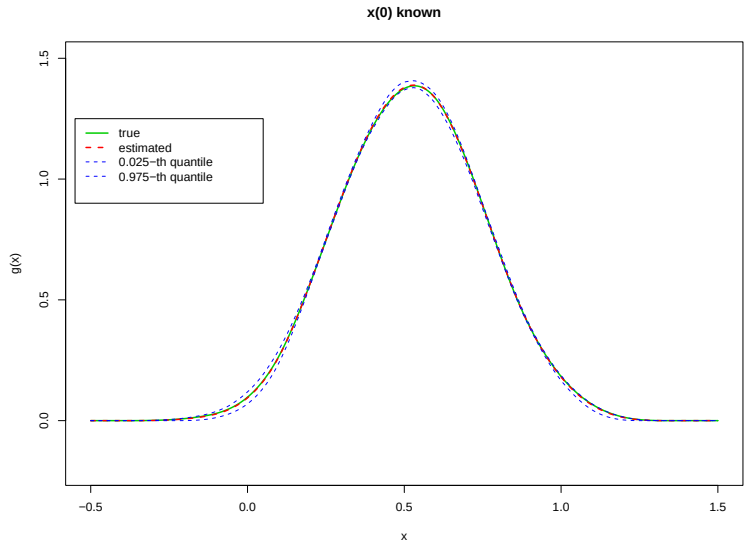

FIG S5-7. **Sparse** *case with* $\sigma_\theta = 0$*: true gradient function* g *and estimates for the selected model using* **hierarchical likelihood** *method with* $\boldsymbol{a}$ **known**, *where the candidate models involve* $M = 2, \ldots, 6$ *B-spline basis functions.*

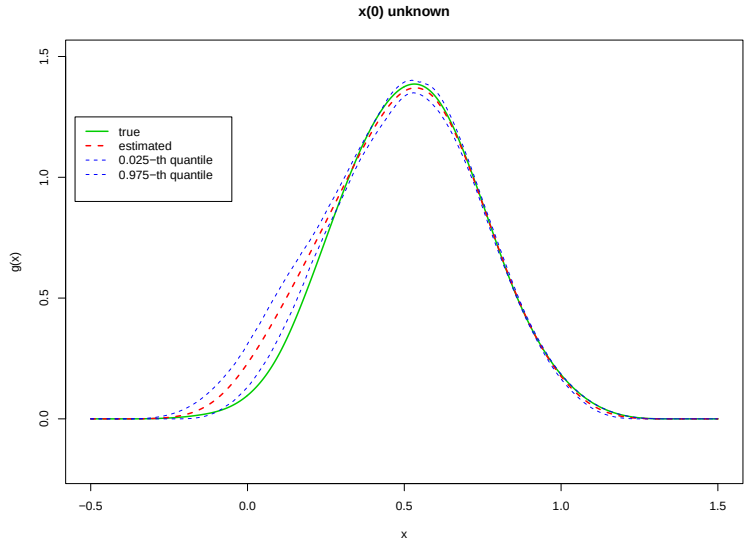

FIG S5-8. **Sparse** *case with* $\sigma_\theta = 0$*: true gradient function* g *and estimates for the selected model using* **hierarchical likelihood** *method with* $\boldsymbol{a}$ **estimated**, *where the candidate models involve* $M = 2, \ldots, 6$ *B-spline basis functions.*

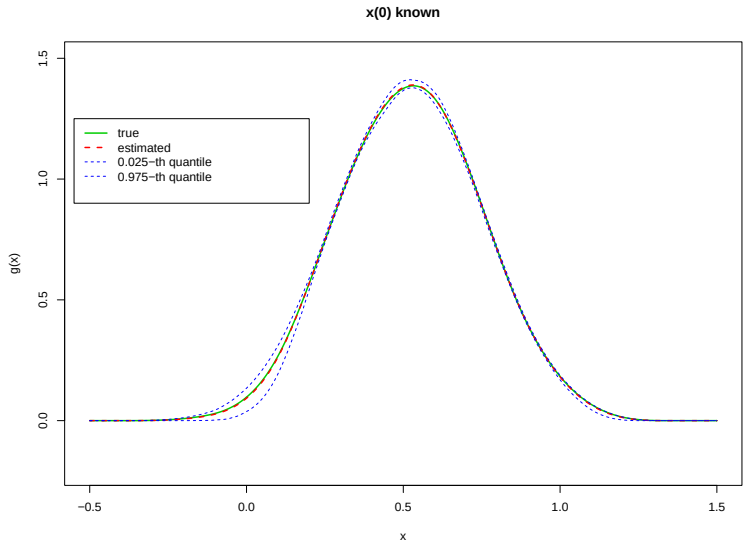

FIG S5-9. **Very dense** *case with* $\sigma_\theta = 0$*: true gradient function* g *and and estimates for the selected model using* **hierarchical likelihood** *method with* $\boldsymbol{a}$ **known**, *where the candidate models involve* $M = 2, \ldots, 6$ *B-spline basis functions.*

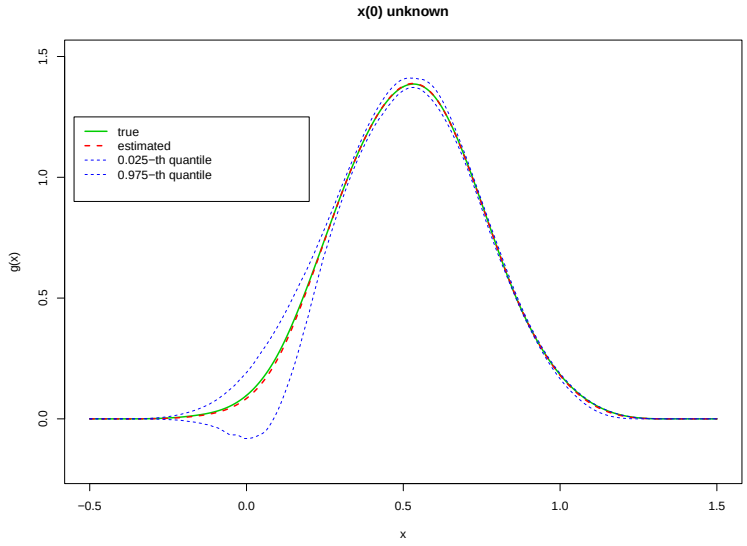

FIG S5-10. **Very dense** *case with* $\sigma_\theta = 0$*: true gradient function* g *and and estimates for the selected model using* **hierarchical likelihood** *method with* $\boldsymbol{a}$ **estimated**, *where the candidate models involve* $M = 2, \ldots, 6$ *B-spline basis functions.*

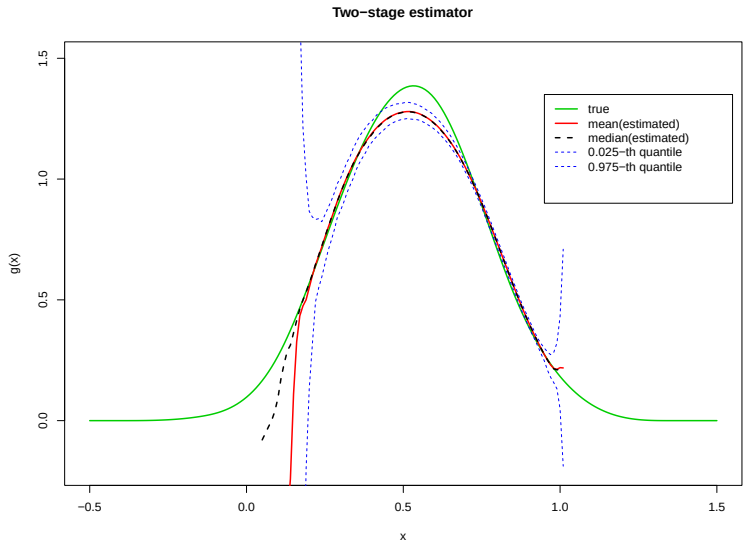

FIG S5-11. **Very dense** *case with* $\sigma_\theta = 0$*: true gradient function* g *and and* **Two-stage method** *where the second stage uses* **local quadratic smoothing** *and the first stage uses optimal bandwidths.*

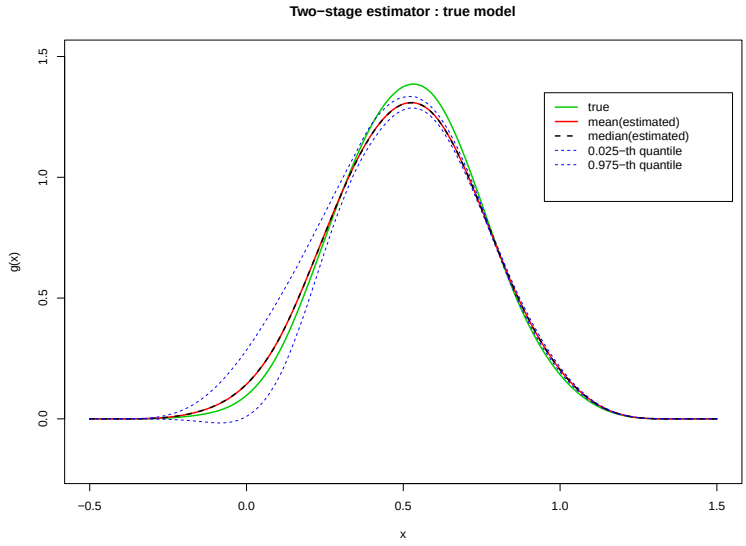

FIG S5-12. **Very dense** *case with* $\sigma_\theta = 0$*: true gradient function* g *and and* **Two-stage method** *where the second stage uses* **regression under the true model** *and the first stage uses optimal bandwidths.*

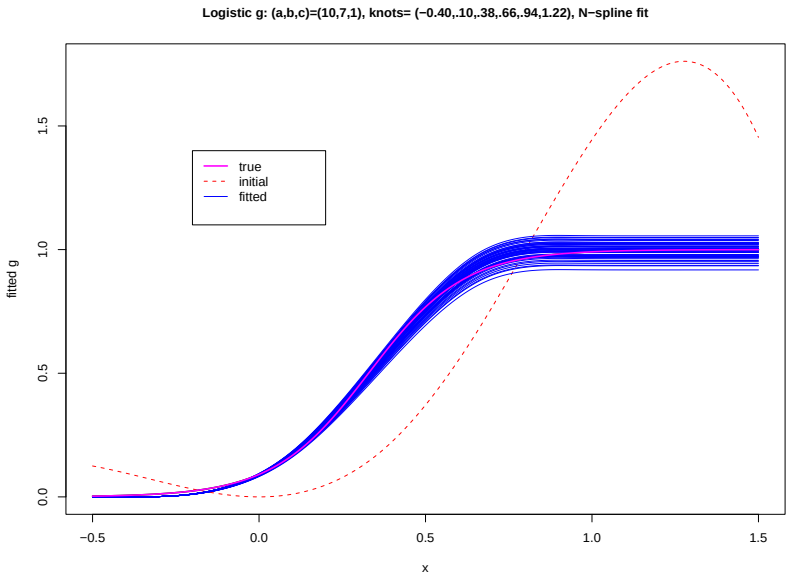

FIG S5-13. *Logistic g : the blue curves correspond to $\widehat{g}$ for 50 replicates with the boundary constraints $g'(x) = 0$ for $x \leq -2$ and $x \geq 1$. The magenta curve is the true g (logistic function with $a = 10$, $b = 7$ and $c = 1$) and the red curve is the initial estimate of g.*

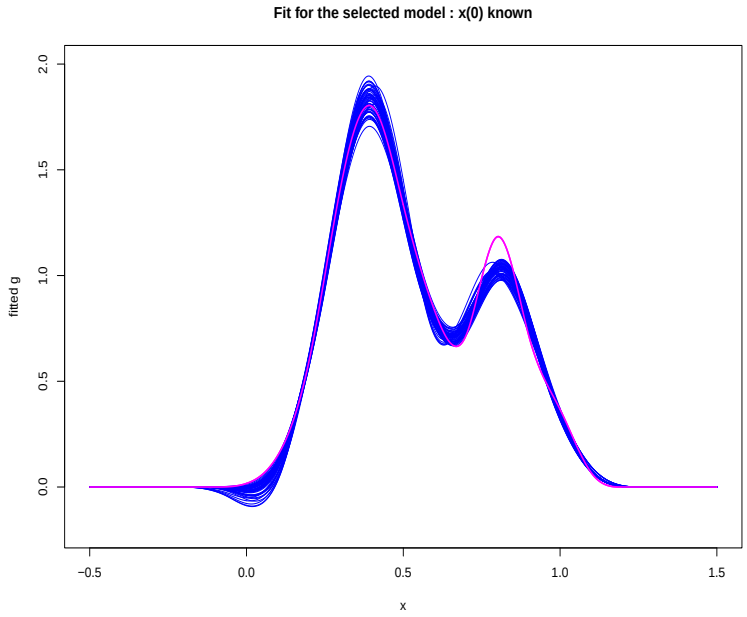

FIG S5-14. **Challenging** *case with sparse measurements and* $\sigma_\theta = 0.1$ *: true gradient function g (magenta curve) and estimates across* 50 *replicates (blue curves) for the selected model using* **hierarchical likelihood** *method with* $\boldsymbol{a}$ **known**.

References.

Brunel, N. J.-B. (2008). Parameter estimation of ODE's via nonparametric estimators. *Electronic Journal of Statistics* **2**, 1242-1267.

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

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 the American Statistical Association* **103**, 369–384.

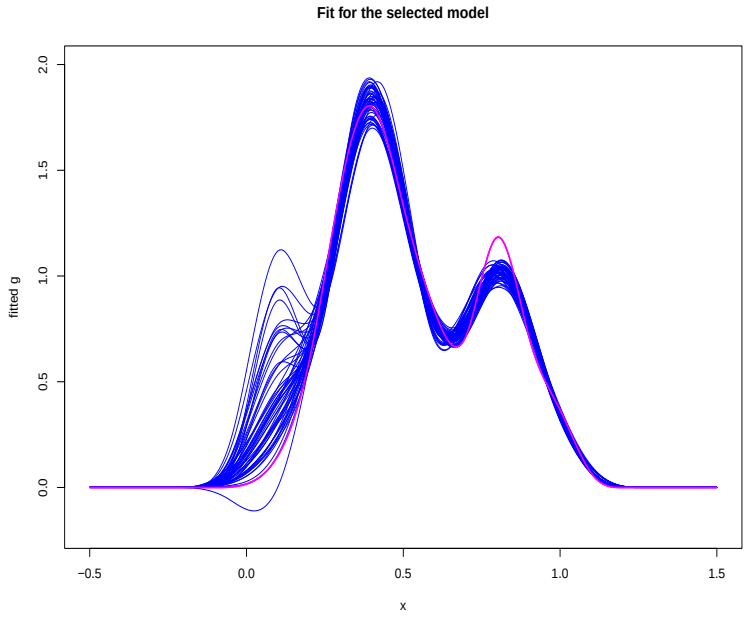

FIG S5-15. **Challenging** *case with sparse measurements and* $\sigma_\theta = 0.1$ *: true gradient function* g *(magenta curve) and estimates across* 50 *replicates (blue curves) for the selected model using* **hierarchical likelihood** *method with* $\boldsymbol{a}$ **estimated**.