Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA

Parameter Estimation of Partial Differential Equation Models

Xiaolei XUN, Jiguo CAO, Bani MALLICK, Arnab MAITY, and Raymond J. CARROLL

Partial differential equation (PDE) models are commonly used to model complex dynamic systems in applied sciences such as biology and finance. The forms of these PDE models are usually proposed by experts based on their prior knowledge and understanding of the dynamic system. Parameters in PDE models often have interesting scientific interpretations, but their values are often unknown and need to be estimated from the measurements of the dynamic system in the presence of measurement errors. Most PDEs used in practice have no analytic solutions, and can only be solved with numerical methods. Currently, methods for estimating PDE parameters require repeatedly solving PDEs numerically under thousands of candidate parameter values, and thus the computational load is high. In this article, we propose two methods to estimate parameters in PDE models: a parameter cascading method and a Bayesian approach. In both methods, the underlying dynamic process modeled with the PDE model is represented via basis function expansion. For the parameter cascading method, we develop two nested levels of optimization to estimate the PDE parameters. For the Bayesian method, we develop a joint model for data and the PDE and develop a novel hierarchical model allowing us to employ Markov chain Monte Carlo (MCMC) techniques to make posterior inference. Simulation studies show that the Bayesian method and parameter cascading method are comparable, and both outperform other available methods in terms of estimation accuracy. The two methods are demonstrated by estimating parameters in a PDE model from long-range infrared light detection and ranging data. Supplementary materials for this article are available online.

KEY WORDS: Asymptotic theory; Basis function expansion; Bayesian method; Differential equations; Measurement error; Parameter cascading.

1. INTRODUCTION

Differential equations are important tools in modeling dynamic processes and are widely used in many areas. The forward problem of solving equations or simulating state variables for given parameters that define the differential equation models has been studied extensively by mathematicians. However, the inverse problem of estimating parameters based on observed error-prone state variables has a relatively sparse statistical literature, and this is especially the case for partial differential equation (PDE) models. There is growing interest in developing efficient estimation methods for such problems.

Various statistical methods have been developed to estimate parameters in ordinary differential equation (ODE) models. There is a series of work in the study of HIV dynamics to understand the pathogenesis of HIV infection. For example, Ho et al. (1995) and Wei et al. (1995) used standard nonlinear least squares regression methods, while Wu, Ding, and DeGruttola (1998) and Wu and Ding (1999) proposed a mixed-effects model approach. Refer to Wu (2005) for a comprehensive review of these methods. Furthermore, Putter et al. (2002); Huang and Wu (2006); and Huang, Liu, and Wu (2006) proposed hierarchical Bayesian approaches for this problem. These methods require repeatedly solving ODE models numerically, which could be time consuming. Ramsay (1996) proposed a data reduction technique in functional data analysis, which involved solving for coefficients of linear differential operators, see Poyton et al. (2006) for an example of application. Li et al. (2002) studied a pharmacokinetic model and proposed a semiparametric approach for estimating time-varying coefficients in an ODE model. Ramsay et al. (2007) proposed a generalized smoothing approach, based on profile likelihood ideas, which they named parameter cascading, for estimating constant parameters in ODE models. Cao, Wang, and Xu (2011) proposed robust estimation for ODE models when data have outliers. Cao, Huang, and Wu (2012) proposed a parameter cascading method to estimate time-varying parameters in ODE models. These methods estimate parameters by optimizing certain criteria. In the optimization procedure, using gradient-based optimization techniques may have the parameter estimates converge to a local minima, otherwise global optimization is computationally intensive.

Another strategy to estimate parameters of ODE is the two-stage method, which in the first stage estimates the function and its derivatives from noisy observations using data smoothing methods without considering differential equation models, and then in the second stage estimates of ODE parameters are obtained by least squares. Liang and Wu (2008) developed a two-stage method for a general first-order ODE model, using local polynomial regression in the first stage, and established asymptotic properties of the estimator. Similarly, Chen and Wu

Xiaolei Xun is Senior Biometrician, Beijing Novartis Pharma Co. Ltd., Pudong New District, Shanghai 201203, China (E-mail: *xiaolei.xun@novartis.com*). Jiguo Cao is Associate Professor, Department of Statistics & Actuarial Science, Simon Fraser University, 8888 University Drive, Burnaby, BC V5A1S6, Canada (E-mail: *jiguo_cao@sfu.ca*). Bani Mallick is University Distinguished Professor, Department of Statistics, Texas A&M University, 3143 TAMU, College Station, TX 77843-3143 (E-mail: *bmallick@stat.tamu.edu*). Arnab Maity is Assistant Professor, Department of Statistics, North Carolina State University, Raleigh, NC 27695 (E-mail: *amaity@ncsu.edu*). Raymond J. Carroll is University Distinguished Professor, Department of Statistics, Texas A&M University, 3143 TAMU, College Station, TX 77843-3143 (E-mail: *carroll@stat.tamu.edu*). The research of Mallick, Carroll, and Xun was supported by grants from the National Cancer Institute (R37-CA057030) and the National Science Foundation DMS (Division of Mathematical Sciences) grant 0914951. This publication is based in part on work supported by the Award Number KUS-CI-016-04, made by King Abdullah University of Science and Technology (KAUST). Cao's research is supported by a discovery grant (PIN: 328256) from the Natural Science and Engineering Research Council of Canada (NSERC). Maity's research was performed while visiting the Department of Statistics, Texas A&M University, and was partially supported by the Award Number R00ES017744 from the National Institute of Environmental Health Sciences.

(2008) developed local estimation for time-varying coefficients. The two-stage methods are easy to implement; however, they might not be statistically efficient because derivatives cannot be estimated accurately from noisy data, especially higher-order derivatives.

As for PDEs, there are two main approaches. The first is similar to the two-stage method in Liang and Wu (2008). For example, Bar, Hegger, and Kantz (1999) modeled unknown PDEs using multivariate polynomials of sufficiently high order, and the best fit was chosen by minimizing the least squares error of the polynomial approximation. Based on the estimated functions, the PDE parameters were estimated using least squares (Muller and Timmer 2004). The issues of noise level and data resolution were addressed extensively in this approach. See also Parlitz and Merkwirth (2000) and Voss et al. (1999) for more examples. The second approach uses numerical solutions of PDEs, thus circumventing derivative estimation. For example, Muller and Timmer (2002) solved the target least-squares type minimization problem using an extended multiple shooting method. The main idea was to solve initial value problems in subintervals and integrate the segments with additional continuity constraints. Global minima can be reached in this algorithm, but it requires careful parameterization of the initial condition, and the computational cost is high.

In this article, we consider a multidimensional dynamic process, $g(\mathbf{x})$, where $\mathbf{x} = (x_1, \ldots, x_p)^{\mathrm{T}} \in \mathbb{R}^p$ is a multidimensional argument. Suppose this dynamic process can be modeled with a PDE model

$$\mathcal{F}\left(\mathbf{x}, g, \frac{\partial g}{\partial x_1}, \ldots, \frac{\partial g}{\partial x_p}, \frac{\partial^2 g}{\partial x_1 \partial x_1}, \ldots, \frac{\partial^2 g}{\partial x_1 \partial x_p}, \ldots; \boldsymbol{\theta}\right) = 0, \tag{1}$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)^{\mathrm{T}}$ is the parameter vector of primary interest, and the left-hand side of (1) has a parametric form in $g(\mathbf{x})$ and its partial derivatives. In practice, we do not observe $g(\mathbf{x})$ but instead observe its surrogate $Y(\mathbf{x})$. We assume that $g(\mathbf{x})$ is observed over a meshgrid with measurement errors so that for $i = 1, \ldots, n$, we observe data $(Y_i, \mathbf{x}_i)$ satisfying

$$Y_i = g(\mathbf{x}_i) + \epsilon_i,$$

where ϵ_i, $i = 1, \ldots, n$, are independent and identically distributed measurement errors and are assumed here to follow a Gaussian distribution with mean zero and variance σ_ϵ^2. Our goal is to estimate the unknown $\boldsymbol{\theta}$ in the PDE model (1) from noisy data and to quantify the uncertainty of the estimates.

As mentioned before, a straightforward two-stage strategy, though easy to implement, has difficulty in estimating derivatives of the dynamic process accurately, leading to biased estimates of the PDE parameter. We propose two joint modeling schemes: (a) a parameter cascading or penalized profile likelihood approach and (b) a fully Bayesian treatment. We conjecture that joint modeling approaches are more statistically efficient than a two-stage method, a conjecture that is borne out in our simulations. For the parameter cascading approach, we make two crucial contributions besides the extension to multivariate splines. First, we develop an asymptotic theory for the model fit, along with a new approximate covariance matrix that includes the smoothing parameters. Second, we propose a new criterion for the smoothing parameter selection, which is shown to outperform available criteria used in ODE parameter estimation. Because of the nature of the penalization in the parameter cascading approach, there is no obvious direct "Bayesianization" of it. Instead, we develop a new hierarchical model for the PDE. At the first stage of the hierarchy, the unknown function is related to the data. At the next stage, the PDE induces a prior on the parameters, which is very different from the penalty used in the parameter cascading algorithm. This PDE restricted prior is new in the Bayesian literature. Further, we allow multiple smoothing parameters and perform Bayesian model mixing to obtain the whole uncertainty distribution of the smoothing parameters. Our Markov chain Monte Carlo (MCMC)-based method is of course also very different from the parameter cascading method where we jointly draw parameters rather than using conditional optimization.

The main idea of our two methods is to represent the unknown dynamic process via a nonparametric function while using the PDE model to regularize the fit. In both methods, the nonparametric function is expressed as a linear combination of B-spline basis functions. In the parameter cascading method, this nonparametric function is estimated using penalized least squares, where a penalty term is defined to incorporate the PDE model. This penalizes the infidelity of the nonparametric function to the PDE model so that the nonparametric function is forced to better represent the dynamic process modeled by the PDE. In the Bayesian method, the PDE model information is coded in the prior distribution. We recognize that there is no exact solution by substituting the nonparametric function into the PDE model (1). This PDE modeling error is then modeled as a random process, hence inducing a constraint on the basis function coefficients. We also introduce in the prior an explicit penalty on the smoothness of the nonparametric function. Our two methods avoid direct estimation of the derivative of the dynamic process, which can be obtained easily as a linear combination of the derivatives of the basis functions, and also avoid specifying boundary conditions.

In principle, the proposed methods are applicable to all PDEs, thus having potentially wide applications. As quick examples of PDEs, the heat equation and wave equation are among the most famous ones. The heat equation, also known as the diffusion equation, describes the evolution in time of the heat distribution or chemical concentration in a given region and is defined as $\partial g(\mathbf{x}, t)/\partial t - \theta \sum_{i=1}^{p} \partial^2 g(\mathbf{x}, t)/\partial x_i^2 = 0$. The wave equation is a simplified model for description of waves, such as sound waves, light waves, and water waves, and is defined as $\partial^2 g(\mathbf{x}, t)/\partial t^2 = \theta^2 \sum_{i=1}^{p} \partial^2 g(\mathbf{x}, t)/\partial x_i^2$. More examples of famous PDEs are the Laplace equation, the transport equation, and the beam equation. See Evans (1998) for a detailed introduction to PDEs.

For illustration, we will do specific calculations based on our empirical example of long-range infrared light detection and ranging (LIDAR) data described in Section 5 and also used in our simulations in Section 4. There we propose a PDE model for received signal $g(t, z)$ over time t and range z given as

$$\partial g(t, z)/\partial t - \theta_D \partial^2 g(t, z)/\partial z^2 - \theta_S \partial g(t, z)/\partial z - \theta_A g(t, z) = 0. \tag{2}$$

The PDE model (2) is a linear PDE of parabolic type in one space dimension and is also called a (one-dimensional) linear reaction-convection-diffusion equation. If $g(t, z)$ were observable, (2)

has a closed-form solution, obtained by separating variables, but the solution is the sum of an infinite sequence. Such a solution requires a high computational load to evaluate the solution over a meshgrid of moderate size.

The rest of the article is organized as follows. The parameter cascading method is introduced in Section 2, and the asymptotic properties of the proposed estimator are established. In Section 3, we introduce the Bayesian framework and explain how to make posterior inference using an MCMC technique. Simulation studies are presented in Section 4 to evaluate the finite sample performance of our two methods in comparison with a two-stage method. In Section 5, we illustrate the methods using LIDAR data. Finally, we conclude with some remarks in Section 6.

2. PARAMETER CASCADING METHOD

2.1 Basis Function Approximation

When solving PDEs, it is possible to obtain a unique, explicit formula for certain specific examples, such as the wave equation. However, most PDEs used in practice have no explicit solutions and can only be solved by numeric methods such as finite difference method (Morton and Mayers 2005) and finite element method (Brenner and Scott 2010). Instead of repeatedly solving PDEs numerically for thousands of candidate parameters, which is computationally expensive, we represent the dynamic process, $g(\mathbf{x})$, modeled in (1), by a nonparametric function, which can be expressed as a linear combination of basis functions

$$g(\mathbf{x}) = \sum_{k=1}^{K} b_k(\mathbf{x})\beta_k = \mathbf{b}^{\mathrm{T}}(\mathbf{x})\boldsymbol{\beta}, \tag{3}$$

where $\mathbf{b}(\mathbf{x}) = \{b_1(\mathbf{x}), \ldots, b_K(\mathbf{x})\}^{\mathrm{T}}$ is the vector of basis functions and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)^{\mathrm{T}}$ is the vector of basis coefficients.

We choose B-splines as basis functions in all simulations and applications in this article, since B-splines are nonzero only in short subintervals, a feature called the compact support property (de Boor 2001), which is useful for efficient computation and numerical stability, compared with other basis (e.g., truncated power series basis). The B-spline basis functions are defined with their order, the number, and locations of knots. Some work has been aimed at automatic knot placement and selection. Many of the feasible frequentist methods, for example, Friedman and Silverman (1989) and Stone et al. (1997), are based on stepwise regression. A Bayesian framework is also available, see Denison, Mallick, and Smith (1997) for example. Despite good performance, knot selection procedures are highly computationally intensive. To avoid the complicated knot selection problem, we use a large enough number of knots to make sure the basis functions are sufficiently flexible to approximate the dynamic process. To prevent the nonparametric function overfitting the data, one penalty term will be defined with the PDE model in the next subsection to penalize the roughness of the nonparametric function.

The PDE model (1) can be expressed using the same set of B-spline basis functions by substituting (3) into model (1) as follows

$$\mathcal{F}[\mathbf{x}, \mathbf{b}^{\mathrm{T}}(\mathbf{x})\boldsymbol{\beta}, \{\partial\mathbf{b}(\mathbf{x})/\partial x_1\}^{\mathrm{T}}\boldsymbol{\beta}, \ldots; \boldsymbol{\theta}] = 0.$$

In the special case of linear PDEs, the above expression is also linear in $\boldsymbol{\beta}$, which can be expressed as

$$\begin{aligned}&\mathcal{F}[\mathbf{x}, \mathbf{b}^{\mathrm{T}}(\mathbf{x})\boldsymbol{\beta}, \{\partial\mathbf{b}(\mathbf{x})/\partial x_1\}^{\mathrm{T}}\boldsymbol{\beta}, \ldots; \boldsymbol{\theta}]\\ &\quad = \mathbf{f}^{\mathrm{T}}\{\mathbf{b}(\mathbf{x}), \partial\mathbf{b}(\mathbf{x})/\partial x_1, \ldots; \boldsymbol{\theta}\}\boldsymbol{\beta} = 0,\end{aligned} \tag{4}$$

where $\mathbf{f}\{\mathbf{b}(\mathbf{x}), \partial\mathbf{b}(\mathbf{x})/\partial x_1, \ldots; \boldsymbol{\theta}\}$ is a linear function of the basis functions and their derivatives. In the following, we denote $\mathcal{F}\{\mathbf{x}, g(\mathbf{x}), \ldots; \boldsymbol{\theta}\}$ by the short-hand notation $\mathcal{F}\{g(\mathbf{x}); \boldsymbol{\theta}\}$ and $\mathbf{f}\{\mathbf{b}(\mathbf{x}), \partial\mathbf{b}(\mathbf{x})/\partial x_1, \ldots; \boldsymbol{\theta}\}$ by $\mathbf{f}(\mathbf{x}; \boldsymbol{\theta})$. For the PDE example (2), the form of $\mathbf{f}(\mathbf{x}; \boldsymbol{\theta})$ is given in Appendix A.1.

2.2 Estimating $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$

Following Section 2.1, the dynamic process, $g(\mathbf{x})$, is expressed as a linear combination of basis functions. It is natural to estimate the basis function coefficients, $\boldsymbol{\beta}$, using penalized splines (Ruppert, Wand, and Carroll 2003; Eilers and Marx 2010). If we were simply interested in estimating $g(\cdot) = \mathbf{b}^{\mathrm{T}}(\cdot)\boldsymbol{\beta}$, then we would use the usual penalty $\lambda\boldsymbol{\beta}^{\mathrm{T}}\mathbf{P}^{\mathrm{T}}\mathbf{P}\boldsymbol{\beta}$, where λ is a penalty parameter and $\mathbf{P}$ is a matrix performing differencing on adjacent elements of $\boldsymbol{\beta}$ (Eilers and Marx 2010). Such a penalty does penalize to achieve smoothness of the estimated function; however, it is not in fidelity with (1). Instead, for fixed $\boldsymbol{\theta}$, we define the roughness penalty as $\int[\mathcal{F}\{g(\mathbf{x}); \boldsymbol{\theta}\}]^2 d\mathbf{x}$. This penalty incorporates the PDE model, containing derivatives involved in the model. As a result, the penalty is able to regularize the spline fit. It also shows fidelity to the PDE model, that is, smaller value indicates more fidelity of the spline approximation to the PDE. Hence, we propose to estimate the coefficients, $\boldsymbol{\beta}$, for fixed $\boldsymbol{\theta}$ by minimizing the penalized least squares

$$J(\boldsymbol{\beta}|\boldsymbol{\theta}) = \sum_{i=1}^{n}\{Y_i - g(\mathbf{x}_i)\}^2 + \lambda\int[\mathcal{F}\{g(\mathbf{x}); \boldsymbol{\theta}\}]^2 d\mathbf{x}. \tag{5}$$

The integration in (5) can be approximated numerically by numerical integration methods. Burden and Douglas (2010) suggested that a composite Simpson's rule provided an adequate approximation, a suggestion that we use. See Appendix B.1 in the online supplementary materials for details.

The PDE parameter $\boldsymbol{\theta}$ is then estimated using a higher level of optimization. Denote the estimate of the spline coefficients by $\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta})$, which is considered as a function of $\boldsymbol{\theta}$. Define $\widehat{g}(\mathbf{x}, \boldsymbol{\theta}) = \mathbf{b}^{\mathrm{T}}(\mathbf{x})\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta})$. Because the estimator $\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta})$ is already regularized, we propose to estimate $\boldsymbol{\theta}$ by minimizing the least squares measure of fit

$$H(\boldsymbol{\theta}) = \sum_{i=1}^{n}\{Y_i - \widehat{g}(\mathbf{x}_i, \boldsymbol{\theta})\}^2 = \sum_{i=1}^{n}\{Y_i - \mathbf{b}^{\mathrm{T}}(\mathbf{x}_i)\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta})\}^2. \tag{6}$$

For a general nonlinear PDE model, the function $\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta})$ might have no closed form, and the estimate is thus obtained numerically. This lower level of optimization for fixed $\boldsymbol{\theta}$ is embedded inside the optimization of $\boldsymbol{\theta}$. The objective functions $J(\boldsymbol{\beta}|\boldsymbol{\theta})$ and $H(\boldsymbol{\theta})$ are minimized iteratively until convergence to a solution. In some cases, the optimization can be accelerated and made more stable by providing the gradient, whose analytic form, by the chain rule, is $\partial H(\boldsymbol{\theta})/\partial\boldsymbol{\theta} = \{\partial\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\}^{\mathrm{T}} \times \partial H(\boldsymbol{\theta})/\partial\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta})$. Although $\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta})$ does not have an explicit expression, the implicit function theorem can be applied to find the analytic form of the first-order derivative of $\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ required in the

above gradient. Because $\widehat{\boldsymbol{\beta}}$ is the minimizer of $J(\boldsymbol{\beta}|\boldsymbol{\theta})$, we have $\partial J(\boldsymbol{\beta}|\boldsymbol{\theta})/\partial\boldsymbol{\beta}|_{\widehat{\boldsymbol{\beta}}} = 0$. By taking the total derivative with respect to $\boldsymbol{\theta}$ on the left-hand side and assuming $\partial^2 J(\boldsymbol{\beta}|\boldsymbol{\theta})/\partial\boldsymbol{\beta}^{\mathrm{T}}\partial\boldsymbol{\beta}|_{\widehat{\boldsymbol{\beta}}}$ is nonsingular, the analytic expression of the first-order derivative of $\widehat{\boldsymbol{\beta}}$ is

$$\frac{\partial\widehat{\boldsymbol{\beta}}}{\partial\boldsymbol{\theta}} = -\left(\frac{\partial^2 J}{\partial\boldsymbol{\beta}^{\mathrm{T}}\partial\boldsymbol{\beta}}\bigg|_{\widehat{\boldsymbol{\beta}}}\right)^{-1}\left(\frac{\partial^2 J}{\partial\boldsymbol{\theta}^{\mathrm{T}}\partial\boldsymbol{\beta}}\bigg|_{\widehat{\boldsymbol{\beta}}}\right).$$

When the PDE model (1) is linear, $\widehat{\boldsymbol{\beta}}$ has a close form and the algorithm can be stated as follows. By substituting in (3) and (4), the lower level criterion (5) becomes

$$J(\boldsymbol{\beta}|\boldsymbol{\theta}) = \sum_{i=1}^{n}\{Y_i - \mathbf{b}^{\mathrm{T}}(\mathbf{x}_i)\boldsymbol{\beta}\}^2 + \lambda\int\boldsymbol{\beta}^{\mathrm{T}}\mathbf{f}(\mathbf{x};\boldsymbol{\theta})\mathbf{f}^{\mathrm{T}}(\mathbf{x};\boldsymbol{\theta})\boldsymbol{\beta}d\mathbf{x}.$$

Let $\mathbf{B}$ be the $n\times K$ basis matrix with ith row $\mathbf{b}^{\mathrm{T}}(\mathbf{x}_i)$, and define $\mathbf{Y} = (Y_1,\ldots,Y_n)^{\mathrm{T}}$, and the $K\times K$ penalty matrix $\mathbf{R}(\boldsymbol{\theta}) = \int\mathbf{f}(\mathbf{x};\boldsymbol{\theta})\mathbf{f}^{\mathrm{T}}(\mathbf{x};\boldsymbol{\theta})d\mathbf{x}$. See Appendix B.1 in the online supplementary materials for calculation of $\mathbf{R}(\boldsymbol{\theta})$ for the PDE example (2). Then the penalized least squares criterion (5) can be expressed in the matrix notation

$$J(\boldsymbol{\beta}|\boldsymbol{\theta}) = (\mathbf{Y}-\mathbf{B}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{Y}-\mathbf{B}\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}^{\mathrm{T}}\mathbf{R}(\boldsymbol{\theta})\boldsymbol{\beta}, \tag{7}$$

which is a quadratic function of $\boldsymbol{\beta}$. By minimizing the above penalized least squares criterion, the estimate for $\boldsymbol{\beta}$, for fixed $\boldsymbol{\theta}$, can be obtained in a close form as

$$\widehat{\boldsymbol{\beta}}(\boldsymbol{\theta}) = \{\mathbf{B}^{\mathrm{T}}\mathbf{B} + \lambda\mathbf{R}(\boldsymbol{\theta})\}^{-1}\mathbf{B}^{\mathrm{T}}\mathbf{Y}. \tag{8}$$

Then by substituting in (8), (6) becomes

$$H(\boldsymbol{\theta}) = \|\mathbf{Y} - \mathbf{B}\{\mathbf{B}^{\mathrm{T}}\mathbf{B} + \lambda\mathbf{R}(\boldsymbol{\theta})\}^{-1}\mathbf{B}^{\mathrm{T}}\mathbf{Y}\|^2. \tag{9}$$

To summarize, when estimating parameters in linear PDE models, we minimize criterion (9) to obtain an estimate, $\widehat{\boldsymbol{\theta}}$, for parameters in linear PDE models. The estimated basis coefficients, $\widehat{\boldsymbol{\beta}}$, are obtained by substituting $\widehat{\boldsymbol{\theta}}$ into (8).

2.3 Smoothing Parameter Selection

Our ultimate goal is to obtain an estimate for the PDE parameter $\boldsymbol{\theta}$ such that the solution of the PDE is close to the observed data. For any given value of the smoothing parameter, λ, we obtain the PDE parameter estimate, $\widehat{\boldsymbol{\theta}}$, and the basis coefficient estimate, $\widehat{\boldsymbol{\beta}}(\widehat{\boldsymbol{\theta}})$. Both can be treated as functions of λ, which are denoted as $\widehat{\boldsymbol{\theta}}(\lambda)$ and $\widehat{\boldsymbol{\beta}}\{\widehat{\boldsymbol{\theta}}(\lambda),\lambda\}$. Define $e_i(\lambda) = Y_i - \widehat{g}\{\mathbf{x}_i,\widehat{\boldsymbol{\theta}}(\lambda),\lambda\}$ and $\eta_i(\lambda) = \mathcal{F}\{\widehat{g}(\mathbf{x}_i);\widehat{\boldsymbol{\theta}}(\lambda)\}$, the latter of which is $\widehat{\mathbf{f}}^{\mathrm{T}}\{\mathbf{x}_i;\widehat{\boldsymbol{\theta}}(\lambda)\}\widehat{\boldsymbol{\beta}}\{\widehat{\boldsymbol{\theta}}(\lambda),\lambda\}$ for linear PDE models. Fidelity to the PDE can be measured by $\sum_{i=1}^n\eta_i^2(\lambda)$, while fidelity to the data can be measured by $\sum_{i=1}^n e_i^2(\lambda)$. Clearly, minimizing just $\sum_{i=1}^n e_i^2(\lambda)$ leads to $\lambda = 0$ and gives far too undersmoothed data fits, while simultaneously not taking the PDE into account. On the other hand, our experience shows that minimizing $\sum_{i=1}^n\eta_i^2(\lambda)$ always results in the largest candidate value for λ.

Hence, we propose the following criterion, which considers data fitting and PDE model fitting simultaneously. To choose an optimal λ, we minimize

$$G(\lambda) = \sum_{i=1}^{n} e_i^2(\lambda) + \sum_{i=1}^{n}\eta_i^2(\lambda).$$

The first summation term in $G(\lambda)$, which measures the fit of the estimated dynamic process to the data, tends to choose a small value of the smoothing parameter. The second summation term in $G(\lambda)$, which measures the fidelity of the estimated dynamic process to the PDE model, tends to choose a large value of the smoothing parameter. Adding these two terms together allows a choice of the value for the smoothing parameter that makes the best trade-off between fitting to data and fidelity to the PDE model. For the sake of completeness, we tried cross-validation and generalized cross-validation to estimate the smoothing parameter. The result was to greatly undersmooth the function fit, while leading to biased and quite variable estimates of the PDE parameters.

2.4 Limit Distribution and Variance Estimation of Parameters

We analyze the limiting distribution of $\widehat{\boldsymbol{\theta}}$ following the same line of argument as in Yu and Ruppert (2002), under Assumptions 1–4 in Appendix A.2. As in their work, we assume that the spline approximation is exact so that $g(\mathbf{x}) = \mathbf{b}^{\mathrm{T}}(\mathbf{x})\boldsymbol{\beta}_0$ for a unique $\boldsymbol{\beta}_0 = \boldsymbol{\beta}_0(\boldsymbol{\theta}_0)$, our Assumption 2. Let $\boldsymbol{\theta}_0$ be the true value of $\boldsymbol{\theta}$, and define $\widetilde{\lambda} = \lambda/n$, $\mathbf{S}_n = n^{-1}\sum_{i=1}^n\mathbf{b}(\mathbf{x}_i)\mathbf{b}^{\mathrm{T}}(\mathbf{x}_i)$, $\mathbf{G}_n(\boldsymbol{\theta}) = \mathbf{S}_n + \widetilde{\lambda}\mathbf{R}(\boldsymbol{\theta})$, $\mathbf{R}_{j\theta}(\boldsymbol{\theta}) = \partial\mathbf{R}(\boldsymbol{\theta})/\partial\theta_j$, $\widehat{\mathcal{V}}_j = \mathbf{R}(\widehat{\boldsymbol{\theta}})\mathbf{G}_n^{-1}(\widehat{\boldsymbol{\theta}})\mathbf{R}_{j\theta}(\widehat{\boldsymbol{\theta}})$ and $\widehat{\mathcal{W}}_j = \widehat{\mathcal{V}}_j + \widehat{\mathcal{V}}_j^{\mathrm{T}}$. Define $\boldsymbol{\Lambda}_n(\boldsymbol{\theta})$ to have (j,k)th element

$$\Lambda_{n,jk}(\boldsymbol{\theta}_0) = \boldsymbol{\beta}_0^{\mathrm{T}}(\boldsymbol{\theta}_0)\mathbf{R}_{j\theta}^{\mathrm{T}}(\boldsymbol{\theta}_0)\mathbf{G}_n^{-1}(\boldsymbol{\theta}_0)\mathbf{S}_n\mathbf{G}_n^{-1}(\boldsymbol{\theta}_0)\mathbf{R}_{k\theta}(\boldsymbol{\theta}_0)\boldsymbol{\beta}_0(\boldsymbol{\theta}_0).$$

Define $\boldsymbol{\Sigma}_{n,\mathrm{prop}} = \boldsymbol{\Lambda}_n^{-1}(\boldsymbol{\theta}_0)\mathcal{C}_n(\boldsymbol{\theta}_0)\{\boldsymbol{\Lambda}_n^{-1}(\boldsymbol{\theta}_0)\}^{\mathrm{T}}$, where $\mathcal{C}_n(\boldsymbol{\theta}_0)$ has (j,k)th element $\mathcal{C}_{n,jk}(\boldsymbol{\theta}_0) = \sigma_\epsilon^2\boldsymbol{\beta}_n^{\mathrm{T}}(\boldsymbol{\theta}_0)\mathcal{W}_j\mathbf{G}_n^{-1}(\boldsymbol{\theta}_0)S_n\mathbf{G}_n^{-1}(\boldsymbol{\theta}_0)\mathcal{W}_k\boldsymbol{\beta}_n(\boldsymbol{\theta}_0)$. Let $\boldsymbol{\Sigma}_{n,\mathrm{prop}}^{-1/2}$ be the inverse of the symmetric square root of $\boldsymbol{\Sigma}_{n,\mathrm{prop}}$.

Following the same basic outline of Yu and Ruppert (2002), and essentially their assumptions, although the technical details are considerably different, we show in Appendix A.2 that under Assumptions 1–4 stated there, and assuming homoscedasticity,

$$n^{1/2}\boldsymbol{\Sigma}_{n,\mathrm{prop}}^{-1/2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \to \mathrm{Normal}(0,\mathbf{I}). \tag{10}$$

Estimating $\boldsymbol{\Sigma}_{n,\mathrm{prop}}$ is easy by replacing $\boldsymbol{\theta}_0$ by $\widehat{\boldsymbol{\theta}}$ and $\boldsymbol{\beta}_0$ by $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}(\widehat{\boldsymbol{\theta}})$, and estimating σ_ϵ^2 by fitting a standard spline regression and then forming the residual variance. In the case of heteroscedastic errors, the term $\sigma_\epsilon^2\mathbf{S}_n$ in $\mathcal{C}_{n,jk}(\boldsymbol{\theta}_0)$ can be replaced by its consistent estimate $(n-p)^{-1}\sum_{i=1}^n\mathbf{b}(\mathbf{x}_i)\mathbf{b}^{\mathrm{T}}(\mathbf{x}_i)\{Y_i - \mathbf{b}^{\mathrm{T}}(\mathbf{x}_i)\widehat{\boldsymbol{\beta}}\}^2$, where p is the number of parameters in the B-spline. We use this sandwich-type method in our numerical work.

3. BAYESIAN ESTIMATION AND INFERENCE

3.1 Basic Methodology

In this section, we introduce a Bayesian approach for estimating parameters in PDE models. In this Bayesian approach, the dynamic process modeled by the PDE model is represented by a linear combination of B-spline basis functions, which is estimated with Bayesian P-splines. The coefficients of the basis functions are regularized through the prior, which contains the PDE model information. Therefore, data fitting and PDE fitting are incorporated into a joint model. As described in the paragraph after Equation (1), our approach is

not a direct "Bayesianization" of the methodology described in Section 2.

We use the same notation as before. With the basis function representation given in (3), the basis function model for data fitting is $Y_i = \mathbf{b}^{\mathrm{T}}(\mathbf{x}_i)\boldsymbol{\beta} + \epsilon_i$, where the ϵ_i are independent and identically distributed measurement errors and are assumed to follow a Gaussian distribution with mean zero and variance σ_ϵ^2. The basis functions are chosen with the same rule introduced in the previous section.

In conventional Bayesian P-splines, which will be introduced in Section 3.2, the penalty term penalizes the smoothness of the estimated function. Rather than using a single optimal smoothing parameter as in frequentist methods, the Bayesian approach performs model mixing with respect to this quantity. In other words, many different spline models provide plausible representations of the data, and the Bayesian approach treats such model uncertainty through the prior distribution of the smoothing parameter.

In our problem, we know further that the underlying function satisfies a given PDE model. Naturally, this information should be coded into the prior distribution to regularize the fit. Because we recognize that there may be no basis function representation that exactly satisfies the PDE model (1), for the purposes of Bayesian computation, we will treat the approximation error as random, and the PDE modeling errors are

$$\mathcal{F}\{\mathbf{b}^{\mathrm{T}}(\mathbf{x}_i)\boldsymbol{\beta};\boldsymbol{\theta}\} = \zeta(\mathbf{x}_i), \tag{11}$$

where the random modeling errors, $\zeta(\mathbf{x}_i)$, are assumed to be independent and identically distributed with a prior distribution Normal$(0, \gamma_0^{-1})$, where the precision parameter, γ_0, should be large enough so that the approximation error in solving (1) with a basis function representation is small. Similarly, instead of using a single optimal value for the precision parameter, γ_0, a prior distribution is assigned to γ_0. The modeling error distribution assumption (11) and a roughness penalty constraint together induce a prior distribution on the basis function coefficients $\boldsymbol{\beta}$. The choice of roughness penalty depends on the dimension of $\mathbf{x}$. For simplicity, we state the Bayesian approach with the specific penalty shown in Section 3.2. The prior distribution of $\boldsymbol{\beta}$ is

$$[\boldsymbol{\beta}|\boldsymbol{\theta}, \gamma_0, \gamma_1, \gamma_2] \propto (\gamma_0\gamma_1\gamma_2)^{K/2} \exp\{-\gamma_0\boldsymbol{\zeta}^{\mathrm{T}}(\boldsymbol{\beta}, \boldsymbol{\theta})\boldsymbol{\zeta}(\boldsymbol{\beta}, \boldsymbol{\theta})/2 - \boldsymbol{\beta}^{\mathrm{T}}(\gamma_1 H_1 + \gamma_2 H_2 + \gamma_1\gamma_2 H_3)\boldsymbol{\beta}/2\}, \tag{12}$$

where, as before, K denotes the number of basis functions, γ_0 is the precision parameter, $\boldsymbol{\zeta}(\boldsymbol{\beta}, \boldsymbol{\theta}) = [\mathcal{F}\{\mathbf{b}^{\mathrm{T}}(\mathbf{x}_1)\boldsymbol{\beta};\boldsymbol{\theta}\}, \ldots, \mathcal{F}\{\mathbf{b}^{\mathrm{T}}(\mathbf{x}_n)\boldsymbol{\beta};\boldsymbol{\theta}\}]^{\mathrm{T}}$, γ_1 and γ_2 control the amount of penalty on smoothness, and the penalty matrices H_1, H_2, H_3 are the same as in the usual Bayesian P-splines, given in (14). We assume conjugate priors for σ_ϵ^2 and γ_ℓ as $\sigma_\epsilon^2 \sim \mathrm{IG}(a_\epsilon, b_\epsilon)$, $\gamma_\ell \sim \mathrm{Gamma}(a_\ell, b_\ell)$, for $\ell = 0, 1, 2$, where IG(a, b) denotes the inverse-gamma distribution with mean $(a-1)^{-1}b$. For the PDE parameter, $\boldsymbol{\theta}$, we assign a Normal$(\mathbf{0}, \sigma_\theta^2\mathbf{I})$ prior, with variance large enough to remain noninformative.

Denote $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \gamma_2)^{\mathrm{T}}$ and $\boldsymbol{\phi} = (\boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{\beta}, \sigma_\epsilon^2)^{\mathrm{T}}$. Based on the above model and prior specification, the joint posterior distribution of all unknown parameters is

$$\begin{aligned}[\boldsymbol{\phi}|\mathbf{Y}] \propto & \prod_{\ell=0}^{2} \gamma_\ell^{a_\ell + K/2 - 1} \left(\sigma_\epsilon^2\right)^{-(a_\epsilon + n/2) - 1} \\ & \exp\left\{ - b_\epsilon/\sigma_\epsilon^2 - \sum_{\ell=0}^{2} b_\ell\gamma_\ell - \boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{\theta}/\left(2\sigma_\theta^2\right) \right\} \\ & \exp\{-\gamma_0\boldsymbol{\zeta}^{\mathrm{T}}(\boldsymbol{\beta}, \boldsymbol{\theta})\boldsymbol{\zeta}(\boldsymbol{\beta}, \boldsymbol{\theta})/2 - \boldsymbol{\beta}^{\mathrm{T}}(\gamma_1 H_1 + \gamma_2 H_2 \\ & + \gamma_1\gamma_2 H_3)\boldsymbol{\beta}/2 - \left(2\sigma_\epsilon^2\right)^{-1}(\mathbf{Y} - \mathbf{B}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{Y} - \mathbf{B}\boldsymbol{\beta})\}.\end{aligned} \tag{13}$$

The posterior distribution (13) is not analytically tractable, hence we use an MCMC-based computation method (Gilks, Richardson, and Spiegelhalter 1996) or more precisely Gibbs sampling (Gelfand and Smith 1990) to simulate the parameters from the posterior distribution. To implement the Gibbs sampler, we need the full conditional distributions of all unknown parameters. Due to the choice of conjugate priors, the full conditional distributions of σ_ϵ^2 and γ_ℓ's are easily obtained as inverse-gamma and gamma distributions, respectively. The full conditional distributions of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ are not of standard form, and hence, we employ Metropolis–Hastings algorithm to sample them.

In the special case of a linear PDE, simplifications arise. With approximation (4), the PDE modeling errors are represented as $\zeta(\mathbf{x}_i) = \mathbf{f}^{\mathrm{T}}(\mathbf{x}_i;\boldsymbol{\theta})\boldsymbol{\beta}$, for $i = 1, \ldots, n$. Define the matrix $\mathbf{F}(\boldsymbol{\theta}) = \{\mathbf{f}(\mathbf{x}_1;\boldsymbol{\theta}), \ldots, \mathbf{f}(\mathbf{x}_n;\boldsymbol{\theta})\}^{\mathrm{T}}$. Then the prior distribution of $\boldsymbol{\beta}$ given in (12) becomes

$$\begin{aligned}[\boldsymbol{\beta}|\boldsymbol{\theta}, \gamma_0, \gamma_1, \gamma_2] \propto (\gamma_0\gamma_1\gamma_2)^{K/2} \exp[-\boldsymbol{\beta}^{\mathrm{T}}\{\gamma_0\mathbf{F}^{\mathrm{T}}(\boldsymbol{\theta})\mathbf{F}(\boldsymbol{\theta}) \\ + \gamma_1 H_1 + \gamma_2 H_2 + \gamma_1\gamma_2 H_3\}\boldsymbol{\beta}/2],\end{aligned}$$

where the exponent is quadratic in $\boldsymbol{\beta}$. Then the joint posterior distribution of all unknown parameters given in (13) becomes

$$\begin{aligned}[\boldsymbol{\phi}|\mathbf{Y}] \propto & \prod_{\ell=0}^{2} \gamma_\ell^{a_\ell + K/2 - 1} \left(\sigma_\epsilon^2\right)^{-(a_\epsilon + n/2) - 1} \\ & \exp\left\{ - b_\epsilon/\sigma_\epsilon^2 - \sum_{\ell=0}^{2} b_\ell\gamma_\ell - \boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{\theta}/\left(2\sigma_\theta^2\right) \right\} \\ & \exp[-\boldsymbol{\beta}^{\mathrm{T}}\{\gamma_0\mathbf{F}^{\mathrm{T}}(\boldsymbol{\theta})\mathbf{F}(\boldsymbol{\theta}) + \gamma_1 H_1 + \gamma_2 H_2 \\ & + \gamma_1\gamma_2 H_3\}\boldsymbol{\beta}/2 - \left(2\sigma_\epsilon^2\right)^{-1}(\mathbf{Y} - \mathbf{B}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{Y} - \mathbf{B}\boldsymbol{\beta})].\end{aligned}$$

Under linear PDE models, the full conditional of $\boldsymbol{\beta}$ is easily seen to be a Normal distribution. This reduces the computational cost significantly compared with sampling under nonlinear cases, because the length of the vector $\boldsymbol{\beta}$ increases quickly as dimension increases. Computational details of both nonlinear and linear PDEs are shown in Appendix A.3.

3.2 Bayesian P-Splines

Here we describe briefly the implementation of Bayesian penalized splines, or P-splines. Eilers and Marx (2003) and Marx and Eilers (2005) dealt specifically with bivariate penalized B-splines. In the simulation studies and the application of this article, we use the bivariate B-spline basis, which is formed by the tensor product of one-dimensional B-spline basis.

Following Marx and Eilers (2005), we use the difference penalty to penalize the interaction of one-dimensional

coefficients as well as each dimension individually. Denote the number of basis functions in each dimension by k_ℓ, the one-dimensional basis function matrices by $\mathbf{B}_\ell$, and the m_ℓth order difference matrix of size $(k_\ell - m_\ell) \times k_\ell$ by $\mathbf{D}_\ell$, for $\ell = 1, 2$. The prior density of the basis function coefficient $\boldsymbol{\beta}$ of length $K = k_1 k_2$ is assumed to be $[\boldsymbol{\beta}|\gamma_1, \gamma_2] \propto (\gamma_1\gamma_2)^{K/2} \exp\{-\boldsymbol{\beta}^{\mathrm{T}}(\gamma_1 H_1 + \gamma_2 H_2 + \gamma_1\gamma_2 H_3)\boldsymbol{\beta}/2\}$, where γ_1 and γ_2 are hyperparameters, and the matrices are

$$H_1 = \mathbf{B}_1^{\mathrm{T}}\mathbf{B}_1 \otimes \mathbf{D}_2^{\mathrm{T}}\mathbf{D}_2;\ H_2 = \mathbf{D}_1^{\mathrm{T}}\mathbf{D}_1 \otimes \mathbf{B}_2^{\mathrm{T}}\mathbf{B}_2;\ H_3 = \mathbf{D}_1^{\mathrm{T}}\mathbf{D}_1 \otimes \mathbf{D}_2^{\mathrm{T}}\mathbf{D}_2. \tag{14}$$

When assuming conjugate prior distributions as $[\sigma_\epsilon^2] = \mathrm{IG}(a_\epsilon, b_\epsilon)$, $[\gamma_1] = \mathrm{Gamma}(a_1, b_1)$, and $[\gamma_2] = \mathrm{Gamma}(a_2, b_2)$, the posterior distribution can be derived easily and sampled using the Gibbs sampler. Although the prior distribution of $\boldsymbol{\beta}$ is improper, the posterior distribution is proper (Berry, Carroll, and Ruppert 2002).

4. SIMULATIONS

4.1 Background

In this section, the finite sample performances of our methods are investigated via Monte Carlo simulations, which are also compared with a two-stage method described below.

The two-stage method is constructed for PDE parameter estimation as follows. In the first stage, $g(\mathbf{x})$ and the partial derivatives of $g(\mathbf{x})$ are estimated by the multidimensional penalized signal regression (MPSR) method (Marx and Eilers 2005). Marx and Eilers (2005) showed that their MPSR method was competitive with other popular methods and had several advantages such as taking full advantage of the natural spatial information of the signals and being intuitive to understand and use. Let $\widehat{\boldsymbol{\beta}}$ denote the estimated coefficients of the basis functions in the first stage. In the second stage, we plug the estimated function and partial derivatives into the PDE model, $\mathcal{F}\{g(\mathbf{x}); \boldsymbol{\theta}\} = 0$, for each observation, that is, we calculate $\widehat{\mathcal{F}}\{\widehat{g}(\mathbf{x}_i); \boldsymbol{\theta}\}$ for $i = 1, \ldots, n$. Then, a least-squares type estimator for the PDE parameter, $\boldsymbol{\theta}$, is obtained by minimizing $J(\boldsymbol{\theta}) = \sum_{i=1}^{n} \widehat{\mathcal{F}}^2\{\widehat{g}(\mathbf{x}_i); \boldsymbol{\theta}\}$. For comparison purposes, the standard errors of two-stage estimates of the PDE parameters are estimated using a parametric bootstrap, which is implemented as follows. Let $\widehat{\boldsymbol{\theta}}$ denote the estimated PDE parameter using the two-stage method and $S(\mathbf{x}|\widehat{\boldsymbol{\theta}})$ denote the numerical solution of PDE (2) using $\widehat{\boldsymbol{\theta}}$ as the parameter value. New simulated data are generated by adding independent and identically distributed Gaussian noises with the same standard deviation as the data to the PDE solutions at every 1 time unit and every 1 range unit. The PDE parameter is then estimated from the simulated data using the two-stage method, and the PDE parameter estimate is denoted as $\widetilde{\boldsymbol{\theta}}^{(j)}$, where $j = 1, \ldots, 100$, is the index of replicates in the parametric bootstrap procedure. The experimental standard deviation of $\widetilde{\boldsymbol{\theta}}^{(j)}$ is set as the standard error of two-stage estimates.

4.2 Data-Generating Mechanism

The PDE model (2) is used to simulate data. The PDE model (2) is numerically solved by setting the true parameter values as $\theta_D = 1, \theta_S = 0.1$, and $\theta_A = 0.1$; the boundary condition as $g(t, 0) = 0$; and the initial condition as $g(0, z) = \{1 + 0.1 \times (20 - z)^2\}^{-1}$ over a meshgrid in the time domain $t \in [1, 20]$ and the range domain $z \in [1, 40]$. To obtain a precise numerical solution, we take a grid of size 0.0005 in the time domain and of size 0.001 in the range domain. The numerical solution is shown in Figure 1, together with cross-sectional views along time and range axes. Then the observed error-prone data are simulated by adding independent and identically distributed Gaussian noises with standard deviation $\sigma = 0.02$ to the PDE solutions at every 1 time unit and every 1 range unit. In other words, our data is on a 20-by-40 meshgrid in the domain $[1, 20] \times [1, 40]$. This value of σ is close to that of our data example in Section 5. To investigate the effect of data noise on the parameter estimation, we do another simulation study in which the simulated data are generated in the exact same setting except that the standard deviation of noises is set as $\sigma = 0.05$.

4.3 Performance of the Proposed Methods

The parameter cascading method, the Bayesian method, and the two-stage method were applied to estimate the three parameters in the PDE model (2) from the simulated data. The simulation is implemented with 1000 replicates. This section

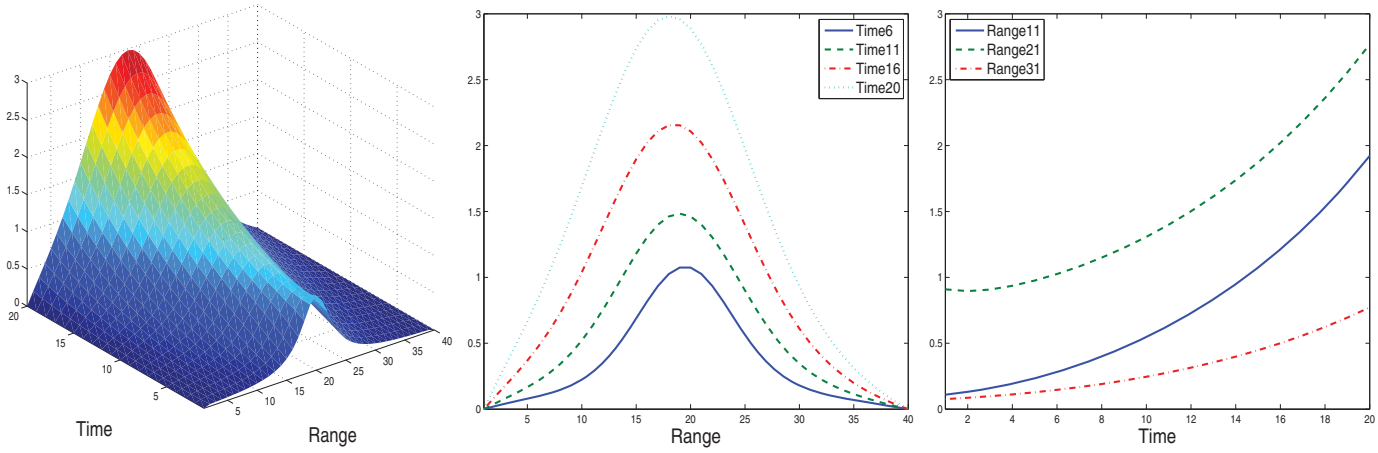

Figure 1. Snapshots of the numerical solution, $g(t, z)$, for the PDE model (2). Left: three-dimensional plot of the surface $g(t, z)$. Middle: plot of $g(t_i, z)$ for time values t_i over range, with $t_i = 6, 11, 16, 20$. Right: plot of $g(t, z_j)$ for range values z_j over time, with $z_j = 11, 21, 31$.

Table 1. The biases, standard deviations (SD), square roots of average squared errors (RASE) of the parameter estimates for the PDE model (2) using the Bayesian method (BM), the parameter cascading method (PC), and the two-stage method (TS) in the 1000 simulated datasets when the data noise has the standard deviation $\sigma = 0.02,\ 0.05$

Noise		$\sigma = 0.02$			$\sigma = 0.05$		
Parameters		θ_D	θ_S	θ_A	θ_D	θ_S	θ_A
True		1.0	0.1	0.1	1.0	0.1	0.1
Bias	BM	−16.5	−0.4	−0.2	−35.6	1.0	0.6
$\times 10^3$	PC	−29.7	−0.1	−0.3	−55.9	−0.2	−0.5
	TS	−225.2	−0.7	−1.8	−337.8	0.5	0.6
SD	BM	9.1	1.6	0.2	22.2	3.8	0.5
$\times 10^3$	PC	24.9	3.8	0.5	40.5	6.2	0.8
	TS	91.0	5.9	1.1	140.7	10.2	2.1
RASE	BM	18.81	1.66	0.27	42.0	3.9	0.8
$\times 10^3$	PC	38.96	3.75	0.54	69.1	6.2	1.0
	TS	243.21	5.91	20.66	365.9	10.2	2.2
CP	BM	93.9%	99.9%	98.8%	74.0%	97.8%	86.4%
	PC	84.3%	96.7%	94.9%	78.1%	96.5%	93.5%
	TS	41.8%	93.6%	72.1%	37.6%	94.0%	93.8%

NOTE: The actual coverage probabilities (CP) of nominal 95% credible/confidence intervals are also shown. The true parameter values are also given in the second row.

summarizes the performance of these three methods in this simulation study.

The PDE model (2) indicates that the second partial derivative with respect to z is continuously differentiable, and thus we choose quartic basis functions in the range domain. Therefore, for representing the dynamic process, $g(t, z)$, we use a tensor product of one-dimensional quartic B-splines to form the basis functions, with 5 and 17 equally spaced knots in time domain and range domain, respectively, in all three methods.

In the two-stage method for estimating PDE parameters, the Bayesian P-spline method is used to estimate the dynamic process and its derivatives by setting the hyperparameters defined in Section 3.1 as $a_\epsilon = b_\epsilon = a_1 = b_1 = a_2 = b_2 = 0.01$ and taking the third-order difference matrix to penalize the roughness of the second derivative in each dimension. In the Bayesian method for estimating PDE parameters, we take the same smoothness penalty as in the two-stage method, and the hyperparameters defined in Section 3 are set to be $a_\epsilon = b_\epsilon = a_\ell = b_\ell = 0.01$ for $\ell = 0, 1, 2$, and $\sigma_\theta^2 = 9$. In the MCMC sampling procedure, we collect every 5th sample after a burn-in stage of length 5000, until 3000 posterior samples are obtained.

We summarize the simulation results in Table 1, including the biases, standard deviations, square root of average squared errors, and coverage probabilities of 95% confidence intervals for each method. We see that the Bayesian method and the parameter cascading method are comparable, and both have smaller biases, standard deviations, and square root of average squared errors than the two-stage method. The improvement in θ_D is substantial, which is associated with the second partial derivative, $\partial^2 g(t, z)/\partial z^2$. This is consistent with our conjecture that the two-stage strategy is not statistically efficient because of the inaccurate estimation of derivatives, especially higher-order derivatives.

To validate numerically the proposed sandwich estimator of variance in the parameter cascading method, we applied a parametric bootstrap of size 200 to each of the same 1000 simulated datasets and obtained the bootstrap estimator for standard errors of parameter estimates in each of the 1000 datasets. Table 2 displays the means of sandwich and bootstrap standard error estimators, which are highly consistent to each other. Both are also close to the sample standard deviations of parameter estimates obtained from the same 1000 simulated datasets.

The modeling error for the PDE model (2) is estimated as $\widehat{\mathcal{F}}\{\widehat{g}(t, z); \widehat{\boldsymbol{\theta}}\} = \partial \widehat{g}(t, z)/\partial t - \widehat{\theta}_D \partial^2 \widehat{g}(t, z)/\partial z^2 - \widehat{\theta}_S \partial \widehat{g}(t, z)/\partial z - \widehat{\theta}_A \widehat{g}(t, z)$. To assess the accuracy of the estimated dynamic process, $\widehat{g}(t, z)$, and the estimated PDE

Table 2. Numerical validation of the proposed sandwich estimator in the parameter cascading method when the data noise has the standard deviation $\sigma = 0.02, 0.05$

Parameters			θ_D	θ_S	θ_A
$\sigma = 0.02$	$\widehat{\text{SE}}$	Mean of Sandwich Estimators	0.0246	0.00375	0.000467
		Mean of Bootstrap Estimators	0.0257	0.00374	0.000474
	Sample Standard Deviation		0.0249	0.00375	0.000465
$\sigma = 0.05$	$\widehat{\text{SE}}$	Mean of Sandwich Estimators	0.0392	0.00599	0.000783
		Mean of Bootstrap Estimators	0.0404	0.00597	0.000791
	Sample Standard Deviation		0.0405	0.00617	0.000795

NOTE: Under each scenario, the first two rows are means of 1000 sandwich and bootstrap standard error ($\widehat{\text{SE}}$) estimators obtained from the same 1000 simulated datasets, respectively; the last row is the sample standard deviation of 1000 parameter estimates obtained from the same 1000 simulated datasets.

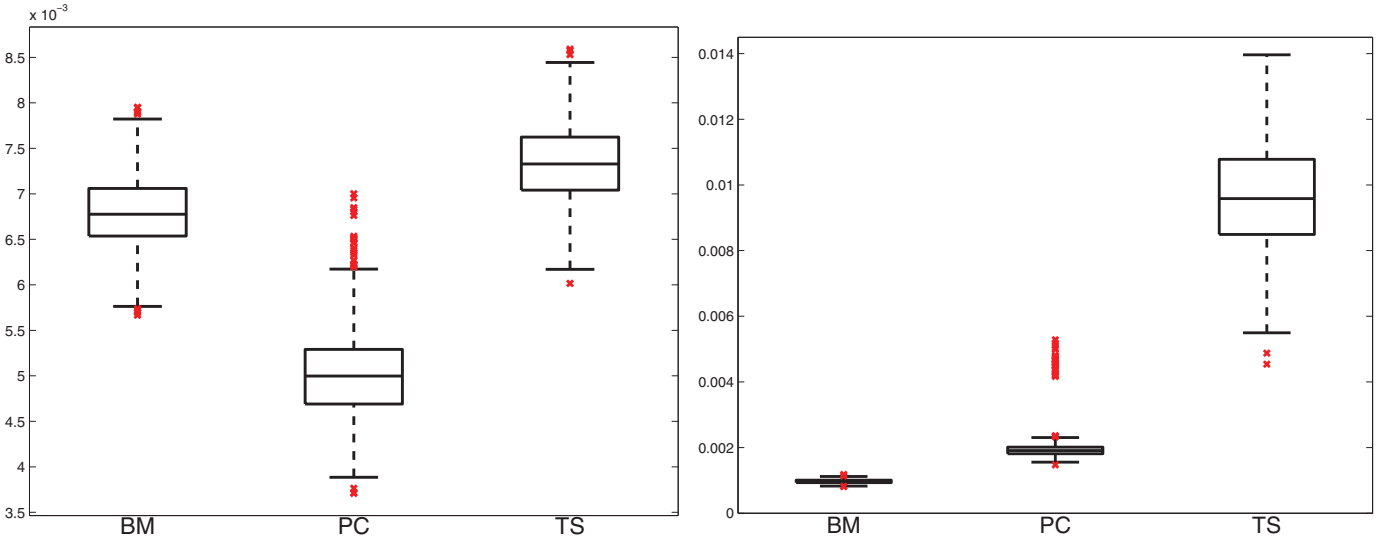

Figure 2. Boxplots of the square roots of average squared errors (RASE) for the estimated dynamic process, $\widehat{g}(t, z)$, and the PDE modeling errors, $\widehat{\mathcal{F}}\{\widehat{g}(t, z); \widehat{\boldsymbol{\theta}}\}$, using the Bayesian method (BM), the parameter cascading method (PC), and the two-stage method (TS) from 1000 datasets in the simulation study. Left: boxplots of RASE($\widehat{g}$), defined in (15), by all three methods. Right: boxplots of RASE($\widehat{\mathcal{F}}$), defined in (16), by all three methods. The online version of this figure is in color.

modeling errors, $\widehat{\mathcal{F}}\{\widehat{g}(t, z); \widehat{\boldsymbol{\theta}}\}$, we use the square root of the average squared errors (RASEs), which are defined as

$$\text{RASE}(\widehat{g}) = \left[m_{\text{tgrid}}^{-1} m_{\text{zgrid}}^{-1} \sum_{j=1}^{m_{\text{tgrid}}} \sum_{k=1}^{m_{\text{zgrid}}} \{\widehat{g}(t_j, z_k) - g(t_j, z_k)\}^2 \right]^{1/2}, \tag{15}$$

$$\text{RASE}(\widehat{\mathcal{F}}) = \left[m_{\text{tgrid}}^{-1} m_{\text{zgrid}}^{-1} \sum_{j=1}^{m_{\text{tgrid}}} \sum_{k=1}^{m_{\text{zgrid}}} \widehat{\mathcal{F}}^2\{\widehat{g}(t_j, z_k); \widehat{\boldsymbol{\theta}}\} \right]^{1/2}, \tag{16}$$

where m_{tgrid} and m_{zgrid} are the number of grid points in each dimension; t_j, z_k are grid points for $j = 1, \ldots, m_{\text{tgrid}}$; and $k = 1, \ldots, m_{\text{zgrid}}$. Figure 2 presents the boxplots of RASEs for the estimated dynamic process, $\widehat{g}(t, z)$, and PDE modeling errors, $\widehat{\mathcal{F}}\{\widehat{g}(t, z); \widehat{\boldsymbol{\theta}}\}$, from the simulated datasets. The Bayesian method and the parameter cascading method have much smaller RASEs for the estimated PDE modeling errors, $\widehat{\mathcal{F}}\{\widehat{g}(t, z); \widehat{\boldsymbol{\theta}}\}$, than the two-stage method because the two-stage method produces inaccurate estimation of derivatives, especially higher-order derivatives.

5. APPLICATION

5.1 Background and Illustration

We have access to a small subset of LIDAR data described by Warren et al. (2008; Warren, Vanderbeek, and Ahl 2009, 2010). A comic describing the LIDAR data is given in Figure 3. Our dataset consists of samples collected for 28 aerosol clouds, 14 of them being biological and the other 14 being nonbiological. Briefly, for each sample, there is a transmitted signal that is sent into the aerosol cloud at 19 laser wavelengths, and for $t = 1, \ldots, T$ time points. For each wavelength and time point, received LIDAR data were observed at equally spaced ranges $z = 1, \ldots, Z$. The experiment also included background data, that is, before the aerosol cloud was released, and the received data were then background corrected.

An example of the background-corrected received data for a single sample and a single wavelength are given in Figure 4. Data such as this are well described by the PDE model (2). This equation is a linear PDE of parabolic type in one space dimension and is also called a (one-dimensional) linear reaction-convection-diffusion equation. If we describe this equation as $g(t, z)$, the parameters θ_D, θ_S, and θ_A describe the diffusion rate, the drift rate/shift, and the reaction rate, respectively.

In fitting model (2) to the real data, we take $T = 20$ time points and $Z = 60$ range values so that the sample size n is $20 \times 60 = 1200$. To illustrate what happens with the data in Figure 4, the parameter cascading method, Bayesian method,

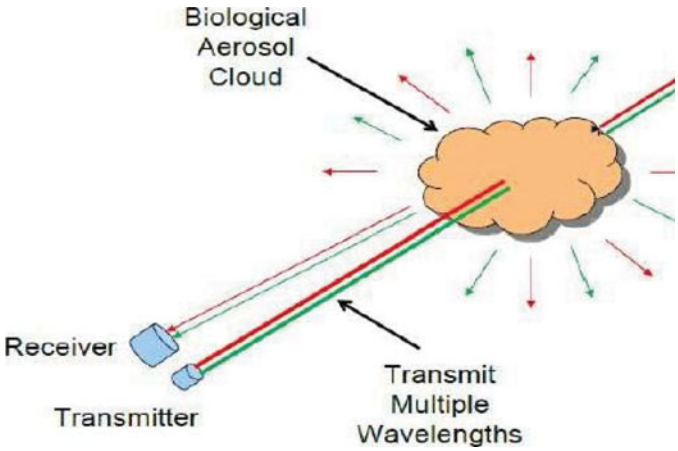

Figure 3. A comic describing the LIDAR data. A point source laser is transmitted into an aerosol cloud at multiple wavelengths and over multiple time points. There is scattering of the signal and reflected back to a receiver over multiple range values. See Figure 4 for an example of the received data over bursts and time for a single wavelength and a single sample. The online version of this figure is in color.

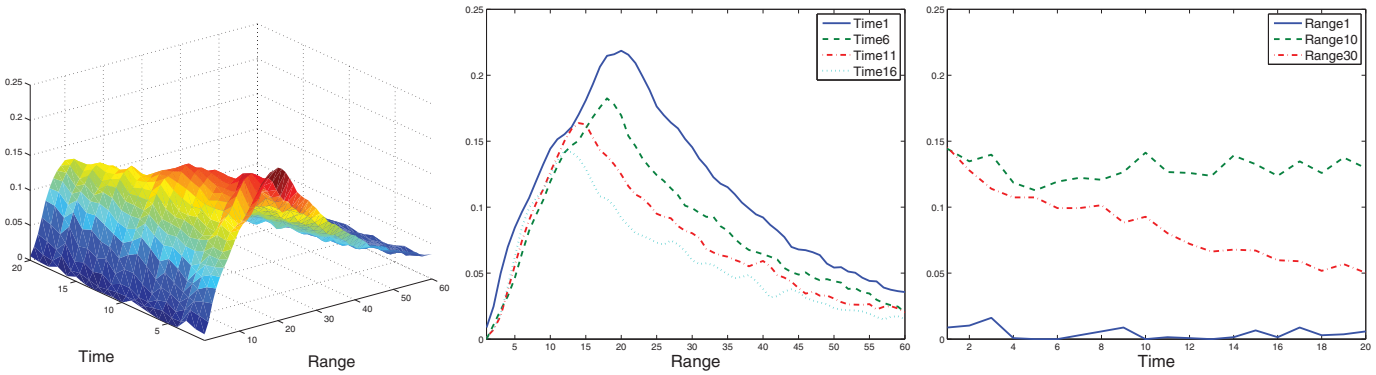

Figure 4. Snapshots of the empirical data. Left: three-dimensional plot of the received signal. Middle: the received signal at a few time values, $t_i = 1, 6, 11, 16$, over the range. Right: the received signal at a few range values, $z_j = 1, 10, 30$, over the time.

and the two-stage method were applied to estimate the three parameters in the PDE model (2) from the above LIDAR dataset. All three methods use bivariate quartic B-spline basis functions constructed with 5 inner knots in the time domain and 20 inner knots in the range domain.

Table 3 displays the estimates for the three parameters in the PDE model (2). While the three methods produce similar estimates for parameters θ_S and θ_A, the parameter cascading estimate and Bayesian estimate for θ_D are more consistent with each other than with the two-stage estimate. This phenomenon is consistent with what was seen in our simulations. Moreover, in this application, the three methods produce almost identical smooth curves, but not derivatives. This fact is also found in our simulation studies, where all three methods lead to similar estimates for the dynamic process, $g(t, z)$, but the two-stage method performs poorly for estimating its derivatives.

5.2 Differences Among the Types of Samples

To understand if there are differences between the received signals for biological and nonbiological samples, we performed the following simple analysis. For each sample, and for each wavelength, we fit the PDE model (2) to obtain estimates of $(\theta_D, \theta_S, \theta_A)$ and then performed t-tests to compare them across aerosol types. Strikingly, there was no evidence that the diffusion rate θ_D differed between the aerosol types at any wavelength, with a minimum p-value being of 0.12 across all wavelengths and both the parameter cascade and Bayesian methods. For the drift rate/shift θ_S, all but 1 wavelength had a p-value < 0.05 for both methods and multiple wavelengths reached Bonferroni significance. For the reaction rate θ_A, the results are somewhat intermediate. While for both methods, all but 1 wavelength had a p-value < 0.05, none reached Bonferroni significance. In summary, the differences between the two types of aerosol clouds are clearly expressed by the drift rate/shift, with some evidence of differences in the reaction rate, but no differences in the diffusion rate. In almost all cases, the drift rate is larger in the nonbiological samples, while the reaction rate is larger in the biological samples.

Table 3. Estimated parameters for the PDE model (2) from the LIDAR dataset using the Bayesian method (BM), the parameter cascading method (PC), and the two-stage method (TS)

		θ_D	θ_S	θ_A
Estimates	BM	−0.4470	0.2563	−0.0414
	PC	−0.3771	0.2492	−0.0407
	TS	−0.1165	0.2404	−0.0436

6. CONCLUDING REMARKS

Differential equation models are widely used to model dynamic processes in many fields such as engineering and biomedical sciences. The forward problem of solving equations or simulating state variables for given parameters that define the models has been extensively studied in the past. However, the inverse problem of estimating parameters based on observed state variables is relatively sparse in the statistical literature, and this is especially the case for PDE models.

We have proposed a parameter cascading method and a fully Bayesian treatment for this problem, which are compared with a two-stage method. The parameter cascading method and Bayesian method are joint estimation procedures that consider the data fitting and PDE fitting simultaneously. Our simulation studies show that the proposed two methods are more statistically efficient than a two-stage method, especially for parameters associated with higher-order derivatives. Basis function expansion plays an important role in our new methods, in the sense that it makes joint modeling possible and links together fidelity to the PDE model and fidelity to data through the coefficients of basis functions. A potential extension of this work would be to estimate time-varying parameters in PDE models from error-prone data.

APPENDIX

A.1 Calculation of $\mathbf{f}(\mathbf{x};\boldsymbol{\theta})$ and $\mathbf{F}(\boldsymbol{\theta})$

Here we show the form of $\mathbf{f}(\mathbf{x};\boldsymbol{\theta})$ and $\mathbf{F}(\boldsymbol{\theta})$ for the PDE example (2). The vector $\mathbf{f}(\mathbf{x};\boldsymbol{\theta})$ is a linear combination of basis functions and their derivatives involved in model (2). We have that $\mathbf{f}(\mathbf{x};\boldsymbol{\theta}) = \partial\mathbf{b}(\mathbf{x})/\partial t - \theta_D \partial^2\mathbf{b}(\mathbf{x})/\partial z^2 - \theta_S \partial\mathbf{b}(\mathbf{x})/\partial z - \theta_A \mathbf{b}(\mathbf{x})$. Similar to the basis function matrix $\mathbf{B} = \{\mathbf{b}(\mathbf{x}_1), \ldots, \mathbf{b}(\mathbf{x}_n)\}^{\mathrm{T}}$, we define the following $n \times K$ matrices consisting of derivatives of the basis functions

$$\mathbf{B}_t = \{\partial\mathbf{b}(\mathbf{x}_1)/\partial t, \ldots, \partial\mathbf{b}(\mathbf{x}_n)/\partial t\}^{\mathrm{T}},$$
$$\mathbf{B}_z = \{\partial\mathbf{b}(\mathbf{x}_1)/\partial z, \ldots, \partial\mathbf{b}(\mathbf{x}_n)/\partial z\}^{\mathrm{T}},$$
$$\mathbf{B}_{zz} = \{\partial^2\mathbf{b}(\mathbf{x}_1)/\partial z^2, \ldots, \partial^2\mathbf{b}(\mathbf{x}_n)/\partial z^2\}^{\mathrm{T}}.$$

Then the matrix $\mathbf{F}(\boldsymbol{\theta}) = \{\mathbf{f}(\mathbf{x}_1;\boldsymbol{\theta}), \ldots, \mathbf{f}(\mathbf{x}_n;\boldsymbol{\theta})\}^{\mathrm{T}} = \mathbf{B}_t - \theta_D \mathbf{B}_{zz} - \theta_S \mathbf{B}_z - \theta_A \mathbf{B}$.

A.2 Sketch of the Asymptotic Theory

A.2.1 Assumptions and Notation. Asymptotic theory for our estimators follows in a fashion very similar to that of Yu and Ruppert (2002). Let $\widetilde{\lambda} = \lambda/n$ denote the true value of $\boldsymbol{\theta}$ as $\boldsymbol{\theta}_0$ and define

$$\begin{aligned}
\mathbf{S}_n &= n^{-1}\sum_{i=1}^{n}\mathbf{b}(\mathbf{x}_i)\mathbf{b}^{\mathrm{T}}(\mathbf{x}_i);\\
\mathbf{G}_n(\boldsymbol{\theta}) &= \mathbf{S}_n + \widetilde{\lambda}\mathbf{R}(\boldsymbol{\theta});\\
\widehat{\boldsymbol{\beta}}_n(\boldsymbol{\theta}) &= \mathbf{G}_n^{-1}(\boldsymbol{\theta})n^{-1}\sum_{i=1}^{n}\mathbf{b}(\mathbf{x}_i)Y_i;\\
\boldsymbol{\beta}_n(\boldsymbol{\theta}) &= \mathbf{G}_n^{-1}(\boldsymbol{\theta})n^{-1}\sum_{i=1}^{n}\mathbf{b}(\mathbf{x}_i)g(\mathbf{x}_i);\\
\mathbf{R}_{j\theta}(\boldsymbol{\theta}) &= \frac{\partial\mathbf{R}(\boldsymbol{\theta})}{\partial\theta_j};\\
\boldsymbol{\Omega}_1 &= E(\mathbf{S}_n);\\
\boldsymbol{\Omega}_2(\boldsymbol{\theta}) &= \boldsymbol{\Omega}_1 + \widetilde{\lambda}\mathbf{R}(\boldsymbol{\theta}).
\end{aligned}$$

The parameter $\boldsymbol{\theta}$ is estimated by minimizing

$$\mathcal{L}_n(\boldsymbol{\theta}) = n^{-1}\sum_{i=1}^{n}\{Y_i - \mathbf{b}^{\mathrm{T}}(\mathbf{x}_i)\widehat{\boldsymbol{\beta}}_n(\boldsymbol{\theta})\}^2. \tag{A.1}$$

Assumption 1. The sequence $\widetilde{\lambda}$ is fixed and satisfies $\widetilde{\lambda} = o(n^{-1/2})$.

Assumption 2. The function $g(\mathbf{x}) = \mathbf{b}^{\mathrm{T}}(\mathbf{x})\boldsymbol{\beta}_0$ for a unique $\boldsymbol{\beta}_0$, that is, the spline approximation is exact, and hence $\boldsymbol{\beta}_n(\boldsymbol{\theta}_0) = \mathbf{G}_n^{-1}(\boldsymbol{\theta}_0)\mathbf{S}_n\boldsymbol{\beta}_0$.

Assumption 3. The parameter $\boldsymbol{\theta}_0$ is in the interior of a compact set and, for $j = 1, \ldots, n$, is the unique solution to $0 = \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{R}(\boldsymbol{\theta}_0)\{E(S_n)\}^{-1}\mathbf{R}_{j\theta}(\boldsymbol{\theta}_0)\boldsymbol{\beta}_0$.

Assumption 4. Assumptions (1)–(4) of Yu and Ruppert (2002) hold with their $m(v, \theta)$ being our $\mathbf{b}^{\mathrm{T}}(\mathbf{x})\boldsymbol{\beta}_n(\boldsymbol{\theta})$.

A.2.2 Characterization of the Solution to (A.1). Remember the matrix fact that for any nonsingular symmetric matrix $\mathbf{A}(z)$ for scalar z, $\partial\mathbf{A}^{-1}(z)/\partial z = -\mathbf{A}^{-1}(z)\{\partial\mathbf{A}(z)/\partial z\}\mathbf{A}^{-1}(z)$. This means that for $j = 1, \ldots, m$,

$$\begin{aligned}
\partial\widehat{\boldsymbol{\beta}}_n(\boldsymbol{\theta})/\partial\theta_j &= -\widetilde{\lambda}\mathbf{G}_n^{-1}(\boldsymbol{\theta})\mathbf{R}_{j\theta}(\boldsymbol{\theta})\mathbf{G}_n^{-1}(\boldsymbol{\theta})n^{-1}\sum_{i=1}^{n}\mathbf{b}(\mathbf{x}_i)Y_i\\
&= -\widetilde{\lambda}\mathbf{G}_n^{-1}(\boldsymbol{\theta})\mathbf{R}_{j\theta}(\boldsymbol{\theta})\widehat{\boldsymbol{\beta}}_n(\boldsymbol{\theta}).
\end{aligned}\tag{A.2}$$

Minimizing $\mathcal{L}_n(\boldsymbol{\theta})$ is equivalent to solving for $j = 1, \ldots, m$ for the system of equations

$$0 = n^{-1/2}\sum_{i=1}^{n}\{Y_i - \mathbf{b}^{\mathrm{T}}(\mathbf{x}_i)\widehat{\boldsymbol{\beta}}_n(\boldsymbol{\theta})\}\mathbf{b}^{\mathrm{T}}(\mathbf{x}_i)\{\partial\widehat{\boldsymbol{\beta}}_n(\boldsymbol{\theta})/\partial\theta_j\} = n^{-1/2}\sum_{i=1}^{n}\Psi_{ij}(\boldsymbol{\theta}),$$

where we define $\Psi_{ij}(\boldsymbol{\theta}) = \{Y_i - \mathbf{b}^{\mathrm{T}}(\mathbf{x}_i)\widehat{\boldsymbol{\beta}}_n(\boldsymbol{\theta})\}\mathbf{b}^{\mathrm{T}}(\mathbf{x}_i)\{\partial\widehat{\boldsymbol{\beta}}_n(\boldsymbol{\theta})\partial\theta_j\}$. From now on, we define the score for θ_j as $\mathcal{T}_{nj}(\boldsymbol{\theta}) = n^{-1/2}\sum_{i=1}^{n}\Psi_{ij}(\boldsymbol{\theta})$ and define $\mathcal{T}_n(\boldsymbol{\theta}) = \{\mathcal{T}_{n1}(\boldsymbol{\theta}), \ldots, \mathcal{T}_{nm}(\boldsymbol{\theta})\}^{\mathrm{T}}$.

There are some further simplifications of $\mathcal{T}_n(\boldsymbol{\theta})$. Because of (A.2),

$$\mathcal{T}_{nj}(\boldsymbol{\theta}) = -\widetilde{\lambda}n^{-1/2}\sum_{i=1}^{n}\{Y_i - \mathbf{b}^{\mathrm{T}}(\mathbf{x}_i)\widehat{\boldsymbol{\beta}}_n(\boldsymbol{\theta})\}\mathbf{b}^{\mathrm{T}}(\mathbf{x}_i)\mathbf{G}_n^{-1}(\boldsymbol{\theta})\mathbf{R}_{j\theta}(\boldsymbol{\theta})\widehat{\boldsymbol{\beta}}_n(\boldsymbol{\theta}).$$

However,

$$\begin{aligned}
n^{-1/2}\sum_{i=1}^{n}Y_i\mathbf{b}^{\mathrm{T}}(\mathbf{x}_i) &= n^{1/2}n^{-1}\sum_{i=1}^{n}Y_i\mathbf{b}^{\mathrm{T}}(\mathbf{x}_i)\mathbf{G}_n^{-1}(\boldsymbol{\theta})\mathbf{G}_n(\boldsymbol{\theta})\\
&= n^{1/2}\widehat{\boldsymbol{\beta}}^{\mathrm{T}}(\boldsymbol{\theta})\mathbf{G}_n(\boldsymbol{\theta});\\
n^{-1/2}\sum_{i=1}^{n}\mathbf{b}^{\mathrm{T}}(\mathbf{x}_i)\widehat{\boldsymbol{\beta}}_n(\boldsymbol{\theta})\mathbf{b}^{\mathrm{T}}(\mathbf{x}_i) &= n^{-1/2}\sum_{i=1}^{n}\widehat{\boldsymbol{\beta}}_n^{\mathrm{T}}(\boldsymbol{\theta})\mathbf{b}(\mathbf{x}_i)\mathbf{b}^{\mathrm{T}}(\mathbf{x}_i)\\
&= n^{1/2}\widehat{\boldsymbol{\beta}}_n^{\mathrm{T}}(\boldsymbol{\theta})\mathbf{S}_n.
\end{aligned}$$

Thus for any $\boldsymbol{\theta}$,

$$\begin{aligned}
\mathcal{T}_{nj}(\boldsymbol{\theta}) &= -\widetilde{\lambda}n^{1/2}\{\widehat{\boldsymbol{\beta}}_n^{\mathrm{T}}(\boldsymbol{\theta})\mathbf{G}_n(\boldsymbol{\theta}) - \widehat{\boldsymbol{\beta}}_n^{\mathrm{T}}(\boldsymbol{\theta})\mathbf{S}_n\}\mathbf{G}_n^{-1}(\boldsymbol{\theta})\mathbf{R}_{j\theta}(\boldsymbol{\theta})\widehat{\boldsymbol{\beta}}_n(\boldsymbol{\theta})\\
&= -\widetilde{\lambda}^2n^{1/2}\widehat{\boldsymbol{\beta}}_n^{\mathrm{T}}(\boldsymbol{\theta})\mathbf{R}(\boldsymbol{\theta})\mathbf{G}_n^{-1}(\boldsymbol{\theta})\mathbf{R}_{j\theta}(\boldsymbol{\theta})\widehat{\boldsymbol{\beta}}_n(\boldsymbol{\theta}).
\end{aligned}\tag{A.3}$$

Hence, $\widehat{\boldsymbol{\theta}}$ is the solution to the system of equations $0 = \widehat{\boldsymbol{\beta}}_n^{\mathrm{T}}(\boldsymbol{\theta})\mathbf{R}(\boldsymbol{\theta})\mathbf{G}_n^{-1}(\boldsymbol{\theta})\mathbf{R}_{j\theta}(\boldsymbol{\theta})\widehat{\boldsymbol{\beta}}_n(\boldsymbol{\theta})$.

A.2.3 Further Calculations. Yu and Ruppert showed that if $\widetilde{\lambda} \to 0$ as $n \to \infty$, then uniformly in $\boldsymbol{\theta}$, $\widehat{\boldsymbol{\beta}}_n(\boldsymbol{\theta}) = \boldsymbol{\beta}_0 + o_p(1)$ and that if $\widetilde{\lambda} = o(n^{-1/2})$ as $n \to \infty$, then $n^{1/2}\{\widehat{\boldsymbol{\beta}}_n(\boldsymbol{\theta}_0) - \boldsymbol{\beta}_0\} \to \mathrm{Normal}(0, \sigma_\epsilon^2\boldsymbol{\Omega}_1^{-1})$. Define the Hessian matrix as $\mathcal{M}_n(\boldsymbol{\theta}) = \partial\mathcal{T}_n(\boldsymbol{\theta})/\partial\boldsymbol{\theta}^{\mathrm{T}}$. Because of these facts and Assumption 3, it follows that $\widehat{\boldsymbol{\theta}} = \boldsymbol{\theta}_0 + o_p(1)$, that is, consistency. It then follows that

$$0 = \mathcal{T}_n(\widehat{\boldsymbol{\theta}}) = \mathcal{T}_n(\boldsymbol{\theta}_0) + n^{-1/2}\mathcal{M}_n(\boldsymbol{\theta}_*)n^{1/2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0),$$

where $\boldsymbol{\theta}_* = \boldsymbol{\theta}_0 + o_p(1)$ is between $\widehat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_0$, and hence that

$$n^{1/2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = -\{n^{-1/2}\mathcal{M}_n(\boldsymbol{\theta}_*)\}^{-1}\mathcal{T}_n(\boldsymbol{\theta}_0). \tag{A.4}$$

Define $\boldsymbol{\Lambda}_n(\boldsymbol{\theta})$ to have (j, k)th element

$$\Lambda_{n,jk}(\boldsymbol{\theta}_0) = \boldsymbol{\beta}_n^{\mathrm{T}}(\boldsymbol{\theta}_0)\mathbf{R}_{j\theta}^{\mathrm{T}}(\boldsymbol{\theta}_0)\mathbf{G}_n^{-1}(\boldsymbol{\theta}_0)\mathbf{S}_n\mathbf{G}_n^{-1}(\boldsymbol{\theta}_0)\mathbf{R}_{k\theta}(\boldsymbol{\theta}_0)\boldsymbol{\beta}_n(\boldsymbol{\theta}_0).$$

In what follows, as in Yu and Ruppert (2002), we continue to assume that $\widetilde{\lambda} = o(n^{-1/2})$. However, with a slight abuse of notation, we will write $\mathbf{G}_n(\boldsymbol{\theta}_0) \to \boldsymbol{\Omega}_2(\boldsymbol{\theta}_0)$ rather than $\mathbf{G}_n(\boldsymbol{\theta}_0) \to \boldsymbol{\Omega}_1$, because we have found that implementing the covariance matrix estimator for $\widehat{\boldsymbol{\theta}}$ is more accurate if this is retained: a similar calculation is done in Yu and Ruppert's section 3.2. Now using Assumption 3, we see that

$$\begin{aligned}
\mathcal{T}_{nj}(\boldsymbol{\theta}_0) &= -\widetilde{\lambda}^2n^{1/2}\widehat{\boldsymbol{\beta}}_n^{\mathrm{T}}(\boldsymbol{\theta}_0)\mathbf{R}(\boldsymbol{\theta}_0)\mathbf{G}_n^{-1}(\boldsymbol{\theta}_0)\mathbf{R}_{j\theta}(\boldsymbol{\theta}_0)\widehat{\boldsymbol{\beta}}_n(\boldsymbol{\theta}_0)\\
&= -\widetilde{\lambda}^2n^{1/2}\{\widehat{\boldsymbol{\beta}}_n(\boldsymbol{\theta}_0) - \boldsymbol{\beta}_n(\boldsymbol{\theta}_0)\}^{\mathrm{T}}\mathbf{R}(\boldsymbol{\theta}_0)\mathbf{G}_n^{-1}(\boldsymbol{\theta}_0)\mathbf{R}_{j\theta}(\boldsymbol{\theta}_0)\widehat{\boldsymbol{\beta}}_n(\boldsymbol{\theta}_0)\\
&\quad -\widetilde{\lambda}^2n^{1/2}\boldsymbol{\beta}_n^{\mathrm{T}}(\boldsymbol{\theta}_0)\mathbf{R}(\boldsymbol{\theta}_0)\mathbf{G}_n^{-1}(\boldsymbol{\theta}_0)\mathbf{R}_{j\theta}(\boldsymbol{\theta}_0)\{\widehat{\boldsymbol{\beta}}_n(\boldsymbol{\theta}_0) - \boldsymbol{\beta}_n(\boldsymbol{\theta}_0)\}.
\end{aligned}$$

Define $\mathcal{V}_j = \mathbf{R}(\boldsymbol{\theta}_0)\mathbf{G}_n^{-1}(\boldsymbol{\theta}_0)\mathbf{R}_{j\theta}(\boldsymbol{\theta}_0)$ and $\mathcal{W}_j = \mathcal{V}_j + \mathcal{V}_j^{\mathrm{T}}$. Then we have that

$$\mathcal{T}_{nj}(\boldsymbol{\theta}_0) = -\widetilde{\lambda}^2\boldsymbol{\beta}_n^{\mathrm{T}}(\boldsymbol{\theta}_0)\mathcal{W}_jn^{1/2}\{\widehat{\boldsymbol{\beta}}_n(\boldsymbol{\theta}_0) - \boldsymbol{\beta}_n(\boldsymbol{\theta}_0)\}. \tag{A.5}$$

Now recall that $\mathbf{S}_n \to \boldsymbol{\Omega}_1$ and $\mathbf{G}_n(\boldsymbol{\theta}_0) \to \boldsymbol{\Omega}_2(\boldsymbol{\theta}_0)$ in probability. Hence we have that

$$\begin{aligned}
n^{1/2}\{\widehat{\boldsymbol{\beta}}_n(\boldsymbol{\theta}_0) - \boldsymbol{\beta}_n(\boldsymbol{\theta}_0)\} &= \mathbf{G}_n^{-1}(\boldsymbol{\theta}_0)n^{-1/2}\sum_{i=1}^{n}\mathbf{b}(\mathbf{x}_i)\epsilon(\mathbf{x}_i)\\
&\to \mathrm{Normal}\left\{0, \sigma_\epsilon^2\boldsymbol{\Omega}_2^{-1}(\boldsymbol{\theta}_0)\boldsymbol{\Omega}_1\boldsymbol{\Omega}_2^{-1}(\boldsymbol{\theta}_0)\right\},
\end{aligned}$$

in distribution. So using (A.5), the (j, k)th element of the covariance matrix of $\mathcal{T}_n$ is given by

$$\begin{aligned}
&\mathrm{cov}(\mathcal{T}_{nj}, \mathcal{T}_{nk})\\
&\quad = \widetilde{\lambda}^4\sigma_\epsilon^2\boldsymbol{\beta}_n^{\mathrm{T}}(\boldsymbol{\theta}_0)\mathcal{W}_j\boldsymbol{\Omega}_2^{-1}(\boldsymbol{\theta}_0)\boldsymbol{\Omega}_1\boldsymbol{\Omega}_2^{-1}(\boldsymbol{\theta}_0)\mathcal{W}_k\boldsymbol{\beta}_n(\boldsymbol{\theta}_0)\{1 + o_p(1)\}.
\end{aligned}$$

We now analyze the term $n^{-1/2}\mathcal{M}_n(\boldsymbol{\theta}_*)$. Because of consistency of $\widehat{\boldsymbol{\theta}}$,

$$n^{-1/2}\mathcal{M}_n(\boldsymbol{\theta}_*) = n^{-1/2}\mathcal{M}_n(\boldsymbol{\theta}_0)\{1 + o_p(1)\}. \tag{A.6}$$

The (j,k)th element of $\mathcal{M}_n(\boldsymbol{\theta})$ is

$$\begin{aligned}\mathcal{M}_{n,jk}(\boldsymbol{\theta}) &= -n^{-1/2}\sum_{i=1}^{n}\frac{\partial\widehat{\boldsymbol{\beta}}_n^{\mathrm{T}}(\boldsymbol{\theta})}{\partial\theta_j}\mathbf{b}(\mathbf{x}_i)\mathbf{b}^{\mathrm{T}}(\mathbf{x}_i)\frac{\partial\widehat{\boldsymbol{\beta}}_n(\boldsymbol{\theta})}{\partial\theta_k}\\ &\quad + n^{-1/2}\sum_{i=1}^{n}\{Y_i-\mathbf{b}^{\mathrm{T}}(\mathbf{x}_i)\widehat{\boldsymbol{\beta}}_n(\boldsymbol{\theta})\}\mathbf{b}^{\mathrm{T}}(\mathbf{x}_i)\frac{\partial^2\widehat{\boldsymbol{\beta}}_n(\boldsymbol{\theta})}{\partial\theta_j\partial\theta_k}\\ &= \mathcal{M}_{n1,jk}(\boldsymbol{\theta})+\mathcal{M}_{n2,jk}(\boldsymbol{\theta}).\end{aligned}$$

We see that by (A.2),

$$\begin{aligned}n^{-1/2}\mathcal{M}_{n1,jk}(\boldsymbol{\theta}) &= -n^{-1}\sum_{i=1}^{n}\frac{\partial\widehat{\boldsymbol{\beta}}_n^{\mathrm{T}}(\boldsymbol{\theta})}{\partial\theta_j}\mathbf{b}(\mathbf{x}_i)\mathbf{b}^{\mathrm{T}}(\mathbf{x}_i)\frac{\partial\widehat{\boldsymbol{\beta}}_n(\boldsymbol{\theta})}{\partial\theta_k}\\ &= -n^{-1}\widetilde{\lambda}^2\sum_{i=1}^{n}\widehat{\boldsymbol{\beta}}_n^{\mathrm{T}}(\boldsymbol{\theta})\mathbf{R}_{j\theta}^{\mathrm{T}}(\boldsymbol{\theta})\mathbf{G}_n^{-1}(\boldsymbol{\theta})\mathbf{b}(\mathbf{x}_i)\mathbf{b}^{\mathrm{T}}(\mathbf{x}_i)\mathbf{G}_n^{-1}(\boldsymbol{\theta})\\ &\quad\times\mathbf{R}_{k\theta}(\boldsymbol{\theta})\widehat{\boldsymbol{\beta}}_n(\boldsymbol{\theta})\\ &= -\widetilde{\lambda}^2\widehat{\boldsymbol{\beta}}_n^{\mathrm{T}}(\boldsymbol{\theta})\mathbf{R}_{j\theta}^{\mathrm{T}}(\boldsymbol{\theta})\mathbf{G}_n^{-1}(\boldsymbol{\theta})\mathbf{S}_n^{\mathrm{T}}\mathbf{G}_n^{-1}(\boldsymbol{\theta})\mathbf{R}_{k\theta}(\boldsymbol{\theta})\widehat{\boldsymbol{\beta}}_n(\boldsymbol{\theta}).\end{aligned}$$

Now using the fact that $\widehat{\boldsymbol{\beta}}_n(\boldsymbol{\theta})=\boldsymbol{\beta}_n(\boldsymbol{\theta})+o_p(1)$ for any $\boldsymbol{\theta}$, and recalling the definition of $\Lambda_n(\boldsymbol{\theta})$, we have at $\boldsymbol{\theta}_0$ that

$$n^{-1/2}\mathcal{M}_{n1,jk}(\boldsymbol{\theta}_0) = -\widetilde{\lambda}^2\Lambda_{n,jk}(\boldsymbol{\theta}_0)\{1+o_p(1)\}.$$

Similarly for the remaining term of the Hessian matrix, we have

$$\begin{aligned}&n^{-1/2}\mathcal{M}_{n2,jk}(\boldsymbol{\theta}_0)\\ &= \left[n^{-1}\sum_{i=1}^{n}\{Y_i-\mathbf{b}^{\mathrm{T}}(\mathbf{x}_i)\boldsymbol{\beta}_n(\boldsymbol{\theta}_0)\}\mathbf{b}^{\mathrm{T}}(\mathbf{x}_i)\right]\frac{\partial^2\boldsymbol{\beta}_n(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}_{0j}\partial\boldsymbol{\theta}_{0k}}\{1+o_p(1)\}\\ &= n^{-1}\sum_{i=1}^{n}\epsilon(\mathbf{x}_i)\mathbf{b}^{\mathrm{T}}(\mathbf{x}_i)\frac{\partial^2\boldsymbol{\beta}_n(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}_{0j}\partial\boldsymbol{\theta}_{0k}}\{1+o_p(1)\}\\ &\quad+\left[n^{-1}\sum_{i=1}^{n}\{g(\mathbf{x}_i)-\mathbf{b}^{\mathrm{T}}(\mathbf{x}_i)\boldsymbol{\beta}(\boldsymbol{\theta}_0)\}\mathbf{b}^{\mathrm{T}}(\mathbf{x}_i)\right]\frac{\partial^2\boldsymbol{\beta}_n(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}_{0j}\partial\boldsymbol{\theta}_{0k}}\{1+o_p(1)\}.\end{aligned}$$

By Assumption 3, and since $\epsilon(\mathbf{x})$ has mean zero, we see that

$$n^{-1/2}\mathcal{M}_{n,jk}(\boldsymbol{\theta}_0) = -\widetilde{\lambda}^2\Lambda_{n,jk}(\boldsymbol{\theta}_0)\{1+o_p(1)\}. \tag{A.7}$$

Hence using (A.4) and (A.6), it follows that

$$n^{1/2}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0) = \boldsymbol{\Lambda}_n^{-1}(\boldsymbol{\theta}_0)\{\widetilde{\lambda}^{-2}\mathcal{T}_n(\boldsymbol{\theta}_0)\}+o_p(1). \tag{A.8}$$

Hence using (A.8), we obtain (10), but with $\boldsymbol{\Omega}_1$ and $\boldsymbol{\Omega}_2(\boldsymbol{\theta})$ replaced by their consistent estimates S_n and $\mathbf{G}_n(\boldsymbol{\theta})$.

A.3 Full Conditional Distributions

To sample from the posterior distribution (13) using Gibbs sampler, we need full conditional distributions of all the unknowns. Due to conjugacy, parameters σ_ϵ^2 and the γ terms have closed-form full conditionals. Define $\mathrm{SSE}=(\mathbf{Y}-\mathbf{B}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{Y}-\mathbf{B}\boldsymbol{\beta})$. If we define "rest" to mean conditional on everything else, we have

$$\begin{aligned}[\sigma_\epsilon^2|\mathrm{rest}] &\propto \left(\sigma_\epsilon^2\right)^{-(a_\epsilon+n/2)-1}\exp\left\{-(b_\epsilon+\mathrm{SSE}/2)/\sigma_\epsilon^2\right\}\\ &= \mathrm{IG}(a_\epsilon+n/2, b_\epsilon+\mathrm{SSE}/2),\\ [\gamma_0|\mathrm{rest}] &\propto \gamma_0^{a_0+K/2-1}\exp\{-b_0\gamma_0-\gamma_0\boldsymbol{\zeta}^{\mathrm{T}}(\boldsymbol{\beta},\boldsymbol{\theta})\boldsymbol{\zeta}(\boldsymbol{\beta},\boldsymbol{\theta})/2\}\\ &= \mathrm{Gamma}(a_0+K/2, b_0+\boldsymbol{\zeta}^{\mathrm{T}}(\boldsymbol{\beta},\boldsymbol{\theta})\boldsymbol{\zeta}(\boldsymbol{\beta},\boldsymbol{\theta})/2),\\ [\gamma_1|\mathrm{rest}] &\propto \gamma_1^{a_1+K/2-1}\exp\{-b_1\gamma_1-\boldsymbol{\beta}^{\mathrm{T}}(\gamma_1H_1+\gamma_1\gamma_2H_3)\boldsymbol{\beta}/2\}\\ &= \mathrm{Gamma}(a_1+K/2, b_1+\boldsymbol{\beta}^{\mathrm{T}}(H_1+\gamma_2H_3)\boldsymbol{\beta}/2),\\ [\gamma_2|\mathrm{rest}] &\propto \gamma_2^{a_2+K/2-1}\exp\{-b_2\gamma_2-\boldsymbol{\beta}^{\mathrm{T}}(\gamma_2H_2+\gamma_1\gamma_2H_3)\boldsymbol{\beta}/2\}\\ &= \mathrm{Gamma}(a_2+K/2, b_2+\boldsymbol{\beta}^{\mathrm{T}}(H_2+\gamma_1H_3)\boldsymbol{\beta}/2).\end{aligned}$$

The parameters $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ do not have closed-form full conditionals, which are instead

$$\begin{aligned}[\boldsymbol{\beta}|\mathrm{rest}] &\propto \exp\left\{-\boldsymbol{\beta}^{\mathrm{T}}\left(\sigma_\epsilon^{-2}\mathbf{B}^{\mathrm{T}}\mathbf{B}+\gamma_1H_1+\gamma_2H_2+\gamma_1\gamma_2H_3\right)\boldsymbol{\beta}/2\right.\\ &\quad\left.-\sigma_\epsilon^{-2}\boldsymbol{\beta}^{\mathrm{T}}\mathbf{B}^{\mathrm{T}}\mathbf{Y}-\gamma_0\boldsymbol{\zeta}^{\mathrm{T}}(\boldsymbol{\beta},\boldsymbol{\theta})\boldsymbol{\zeta}(\boldsymbol{\beta},\boldsymbol{\theta})/2\right\},\\ [\boldsymbol{\theta}|\mathrm{rest}] &\propto \exp\left\{-\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{\theta}/\left(2\sigma_\theta^2\right)-\gamma_0\boldsymbol{\zeta}^{\mathrm{T}}(\boldsymbol{\beta},\boldsymbol{\theta})\boldsymbol{\zeta}(\boldsymbol{\beta},\boldsymbol{\theta})/2\right\}.\end{aligned}$$

To draw samples from these full conditionals, a Metropolis–Hastings update within the Gibbs sampler is applied for each component of θ_i. The proposal distribution for the ith component is a normal distribution $\mathrm{Normal}(\theta_{i,\mathrm{curr}}, \sigma_{i,\mathrm{prop}})$, where the mean $\theta_{i,\mathrm{curr}}$ is the current value and the standard deviation $\sigma_{i,\mathrm{prop}}$ is a constant.

In the special case of a linear PDE, the model error is also linear in $\boldsymbol{\beta}$, represented by $\boldsymbol{\zeta}(\boldsymbol{\beta},\boldsymbol{\theta})=\mathbf{F}(\boldsymbol{\theta})\boldsymbol{\beta}$. Then the term $\boldsymbol{\zeta}^{\mathrm{T}}(\boldsymbol{\beta},\boldsymbol{\theta})\boldsymbol{\zeta}(\boldsymbol{\beta},\boldsymbol{\theta})$ is a quadratic function in $\boldsymbol{\beta}$. Define $\mathbf{H}=\mathbf{H}(\boldsymbol{\theta})=\gamma_0\mathbf{F}^{\mathrm{T}}(\boldsymbol{\theta})\mathbf{F}(\boldsymbol{\theta})+\gamma_1H_1+\gamma_2H_2+\gamma_1\gamma_2H_3$, and $\mathbf{D}=\{\mathbf{B}^{\mathrm{T}}\mathbf{B}+\sigma_\epsilon^2\mathbf{H}(\boldsymbol{\theta})\}^{-1}$. By completing the square in $[\boldsymbol{\beta}|\mathrm{rest}]$, the full conditional of $\boldsymbol{\beta}$ under linear PDE models is in the explicit form

$$\begin{aligned}[\boldsymbol{\beta}|\mathrm{rest}] &\propto \exp\left[-\left(2\sigma_\epsilon^2\right)^{-1}\left\{\boldsymbol{\beta}^{\mathrm{T}}\left(\mathbf{B}^{\mathrm{T}}\mathbf{B}+\sigma_\epsilon^2\mathbf{H}\right)\boldsymbol{\beta}-2\boldsymbol{\beta}^{\mathrm{T}}\mathbf{B}^{\mathrm{T}}\mathbf{Y}\right\}\right]\\ &= \mathrm{Normal}\left(\mathbf{D}\mathbf{B}^{\mathrm{T}}\mathbf{Y}, \sigma_\epsilon^2\mathbf{D}\right).\end{aligned}$$

SUPPLEMENTARY MATERIALS

Supplementary materials provide the technical details of calculating the penalty matrix $R(\theta)$ used in Equation (7) and the variance estimator for the PDE parameters given in Equation (10).

[Received November 2012. Revised March 2013.]

REFERENCES

Bar, M., Hegger, R., and Kantz, H. (1999), "Fitting Differential Equations to Space-Time Dynamics," *Physical Review E*, 59, 337–342. [1010]

Berry, S. M., Carroll, R. J., and Ruppert, D. (2002), "Bayesian Smoothing and Regression Splines for Measurement Error Problems," *Journal of the American Statistical Association*, 97, 160–169. [1014]

Brenner, S. C., and Scott, R. (2010), *The Mathematical Theory of Finite Element Methods*, New York: Springer. [1011]

Burden, R. L., and Douglas, F. J. (2010), *Numerical Analysis*, (9th ed.), Belmont, CA: Brooks/Cole. [1011]

Cao, J., Huang, J. Z., and Wu, H. (2012), "Penalized Nonlinear Least Squares Estimation of Time-Varying Parameters in Ordinary Differential Equations," *Journal of Computational and Graphical Statistics*, 21, 42–56. [1009]

Cao, J., Wang, L., and Xu, J. (2011), "Robust Estimation for Ordinary Differential Equation Models," *Biometrics*, 67, 1305–1313. [1009]

Chen, J., and Wu, H. (2008), "Efficient Local Estimation for Time-Varying Coeffients in Deterministic Dynamic Models With Applications to HIV-1 Dynamics," *Journal of the American Statistical Association*, 103, 369–384. [1010]

de Boor, C. (2001), *A Practical Guide to Splines* (Revised edition), Applied Mathematical Sciences 27, New York: Springer. [1011]

Denison, D. G. T., Mallick, B. K., and Smith, A. F. M. (1997), "Automatic Bayesian Curve Fitting," *Journal of the Royal Statistical Society*, Series B, 60, 333–350. [1011]

Eilers, P., and Marx, B. (2003), "Multidimensional Calibration With Temperature Interaction Using Two-Dimensional Penalized Signal Regression," *Chemometrics and Intelligent Laboratory Systems*, 66, 159–174. [1013]

——— (2010), "Splines, Knots and Penalties," *Wiley Interdisciplinary Reviews: Computational Statistics*, 2, 637–653. [1011]

Evans, L. C. (1998), *Partial Differential Equations*, Graduate Studies in Mathematics 19, Providence, RI: American Mathematical Society. [1010]

Friedman, J. H., and Silverman, B. W. (1989), "Flexible Parsimonious Smoothing and Additive Modeling," *Technometrics*, 31, 3–21. [1011]

Gelfand, A. E., and Smith, A. F. M. (1990), "Sampling-Based Approaches to Calculating Marginal Densities," *Journal of the American Statistical Association*, 85, 398–409. [1013]

Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. (1996), *Markov Chain Monte Carlo in Practice: Interdisciplinary Statistics*, London: Chapman & Hall. [1013]

Ho, D. D., Neumann, A. S., Perelson, A. S., Chen, W., Leonard, J. M., and Markowitz, M. (1995), "Rapid Turnover of Plasma Virions and CD4 Lymphocytes in HIV-1 Infection," *Nature*, 373, 123–126. [1009]

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. [1009]

Huang, Y., and Wu, H. (2006), "A Bayesian Approach for Estimating Antiviral Efficacy in HIV Dynamic Models," *Journal of Applied Statistics*, 33, 155–174. [1009]

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. [1009]

Liang, H., and Wu, H. (2008), "Parameter Estimation for Differential Equation Models Using a Framework of Measurement Error in Regression Models," *Journal of the American Statistical Association*, 103, 1570–1583. [1009,1010]

Marx, B., and Eilers, P. (2005), "Multidimensional Penalized Signal Regression," *Technometrics*, 47, 13–22. [1013,1014]

Morton, K. W., and Mayers, D. F. (2005), *Numerical Solution of Partial Differential Equations, An Introduction*, Cambridge: Cambridge University Press. [1011]

Muller, T., and Timmer, J. (2002), "Fitting Parameters in Partial Differential Equations From Partially Observed Noisy Data," *Physical Review,* D, 171, 1–7. [1010]

——— (2004), "Parameter Identification Techniques for Partial Differential Equations," *International Journal of Bifurcation and Chaos*, 14, 2053–2060. [1010]

Parlitz, U., and Merkwirth, C. (2000), "Prediction of Spatiotemporal Time Series Based on Reconstructed Local States," *Physical Review Letters*, 84, 1890–1893. [1010]

Poyton, A. A., Varziri, M. S., McAuley, K. B., McLellan, P. J., and Ramsay, J. O. (2006), "Parameter Estimation in Continuous-Time Dynamic Models Using Principal Differential Analysis," *Computer and Chemical Engineering*, 30, 698–708. [1009]

Putter, H., Heisterkamp, S. H., Lange, J. M. A., and De Wolf, F. (2002), "A Bayesian Approach to Parameter Estimation in HIV Dynamical Models," *Statistics in Medicine*, 21, 2199–2214. [1009]

Ramsay, J. O. (1996), "Principal Differential Analysis: Data Reduction by Differential Operators," *Journal of the Royal Statistical Society,* Series B, 58, 495–508. [1009]

Ramsay, J. O., Hooker, G., Campbell, D., and Cao, J. (2007), "Parameter Estimation for Differential Equations: A Generalized Smoothing Approach" (with discussion), *Journal of the Royal Statistical Society,* Series B, 69, 741–796. [1009]

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003), *Semiparametric Regression*, Cambridge: Cambridge University Press. [1011]

Stone, C. J., Hansen, M. H., Kooperberg, C., and Truong, Y. K. (1997), "Polynomial Splines and Their Tensor Products in Extended Linear Modeling," *The Annals of Statistics*, 25, 1371–1425. [1011]

Voss, H. U., Kolodner, P., Abel, M., and Kurths, J. (1999), "Amplitude Equations From Spatiotemporal Binary-Fluid Convection Data," *Physical Review Letters*, 83, 3422–3425. [1010]

Warren, R. E., Vanderbeek, R. G., and Ahl, J. L. (2009), "Detection and Classification of Atmospheric Aerosols Using Multi-Wavelength LWIR Lidar," in *Proceedings of SPIE*, 7304, 73040E. [1016]

——— (2010), "Estimation and Discrimination of Aerosols Using Multiple Wavelength LIWR Lidar," in *Proceedings of SPIE*, 7665, 766504-1. [1016]

Warren, R. E., Vanderbeek, R. G., Ben-David, A., and Ahl, J. L. (2008), "Simultaneous Estimation of Aerosol Cloud Concentration and Spectral Backscatter From Multiple-Wavelength Lidar Data," *Applied Optics*, 47, 4309–4320. [1016]

Wei, X., Ghosh, S. K., Taylor, M. E., Johnson, V. A., Emini, E. A., Deutsch, P., Lifson, J. D., Bonhoeer, S., Nowak, M. A., Hahn, B. H., Saag, M. S., and Shaw, G. M. (1995), "Viral Dynamics in Human Immunodeficiency Virus Type 1 Infection," *Nature*, 373, 117–123. [1009]

Wu, H. (2005), "Statistical Methods for HIV Dynamic Studies in AIDS Clinical Trials," *Statistical Methods in Medical Research*, 14, 171–192. [1009]

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. [1009]

Wu, H., Ding, A., and DeGruttola, V. (1998), "Estimation of HIV Dynamic Parameters," *Statistics in Medicine*, 17, 2463–2485. [1009]

Yu, Y., and Ruppert, D. (2002), "Penalized Spline Estimation for Partially Linear Single-Index Models," *Journal of the American Statistical Association*, 97, 1042–1054. [1012,1018]