Generalized Profiling Method and the Applications to Adaptive Penalized Smoothing, Generalized Semiparametric Additive Models and Estimating Differential Equations

Jiguo Cao

Department of Mathematics and Statistics
McGill University, Montreal

October, 2006

A thesis submitted to McGill University in partial fulfilment of the requirements of the degree of Doctor of Philosophy

Your file Votre référence
ISBN: 978-0-494-27759-1
Our file Notre référence
ISBN: 978-0-494-27759-1

Abstract:

Many statistical models involve three distinct groups of variables: local or nuisance parameters, global or structural parameters, and complexity parameters. In this thesis, we introduce the generalized profiling method to estimate these statistical models, which treats one group of parameters as an explicit or implicit function of other parameters. The dimensionality of the parameter space is reduced, and the optimization surface becomes smoother. The Newton-Raphson algorithm is applied to estimate these three distinct groups of parameters in three levels of optimization, with the gradients and Hessian matrices written out analytically by the Implicit Function Theorem if necessary and allowing for different criteria for each level of optimization. Moreover, variances of global parameters are estimated by the Delta method and include the variation coming from complexity parameters. We also propose three applications of the generalized profiling method.

First, penalized smoothing is extended by allowing for a functional smoothing parameter, which is adaptive to the geometry of the underlying curve, which is called *adaptive penalized smoothing*. In the first level of optimization, the smoothing coefficients are local parameters, estimated by minimizing sum of squared errors, conditional on the functional smoothing parameter. In the second level, the functional smoothing parameter is a complexity parameter, estimated by minimizing generalized cross-validation (GCV), treating the smoothing coefficients as explicit functions of the functional smoothing parameter. Adaptive penalized smoothing is shown to obtain better estimates for fitting functions and their derivatives.

Next, the generalized semiparametric additive models are estimated by three levels of optimization, allowing response variables in any kind of distribution. In the first level, the nonparametric functional parameters are nuisance parameters, estimated by maximizing the regularized likelihood function, conditional on the linear coefficients and the smoothing parameter. In the second level, the linear coefficients are structural parameters, estimated by maximizing the likelihood function with the nonparametric functional parameters treated as implicit functions of linear coefficients and the smoothing parameter. In the third level, the smoothing parameter is a complexity parameter, estimated by minimizing the approximated GCV with the linear coefficients treated as implicit functions of the smoothing parameter. This method is applied to estimate the generalized semiparametric additive model for the effect of air pollution on the public health.

Finally, parameters in differential equations (DE's) are estimated from noisy data with the generalized profiling method. In the first level of optimization, fitting functions are estimated to approximate DE solutions by penalized smoothing with the penalty term defined by DE's, fixing values of DE parameters. In the second level of optimization, DE parameters are estimated by weighted sum of squared errors, with the smoothing coefficients treated as an implicit function of DE parameters. The effects of the smoothing parameter on DE parameter estimates are explored and the optimization criteria for smoothing parameter selection are discussed. The method is applied to fit the predator-prey dynamic model to biological data, to estimate DE parameters in the HIV dynamic model from clinical trials, and to explore dynamic models for thermal decomposition of α- Pinene.

Résumé:

Plusieurs modèles statistiques comportent trois groupes distincts de paramètres: des paramètres locaux ou de nuisance, des paramètres globaux ou structurels, ainsi que des paramètres de complexité. Une méthode de profil généralisée est présentée dans cette thèse. Cette méthode traite un groupe de paramètres en tant que fonction explicite ou implicite des autres paramètres. La dimensionnalité de l'espace des paramètres est réduite et la surface d'optimisation devient plus lisse. L'algorithme de Newton-Raphson est utilisé pour estimer ces trois groupes distincts de paramètres en trois niveaux d'optimisation, avec les gradients et les matrices hessiennes obtenus analytiquement par le théorème de la fonction implicite lorque nécessaire, en permettant différents critères pour chaque niveau d'optimisation. De plus, les variances des paramètres globaux sont estimés par la méthode Delta et incluent la variation venant des paramètres de complexité. On présente trois applications de la méthode de profil généralisée.

D'abord, le lissage pénalisé est étendu par l'ajout d'un paramètre de lissage fonctionnel qui s'adapte à la géométrie de la courbe sous-jacente, que l'on appelle lissage pénalisé adaptif. Dans le premier niveau d'optimisation, les coefficients de lissage sont des paramètres locaux, estimés en minimisant la somme des carrés des erreurs, en conditionnant sur le paramètre de lissage fonctionnel. Au second niveau, le paramètre de lissage fonctionnel est un paramètre de complexité, estimé en minimisant la validation croisée généralisée (VCG), en se servant des coefficients de lissage comme des fonctions explicites du paramètre de lissage fonctionnel. On démontre que le lissage pénalisé adaptif obtient de meilleures valeurs estimées pour

les courbes de lissage et leurs dérivées.

Ensuite, des modèles additifs généralisés semiparamétriques sont estimés par trois niveaux d'optimisation, en permettant des variables de réponse de toutes sortes de lois. Au premier niveau, les paramètres fonctionnels nonparamétriques sont des paramètres de nuisance, estimés en maximisant la fonction de vraisemblance régularisée, en conditionnant sur les coefficients linéaires et le paramètre de lissage. Au deuxième niveau, les coefficients linéaires sont des paramètres structuraux, estimés par maximisation de la fonction de vraisemblance, en traitant les paramètres fonctionnels nonparamétriques comme des fonctions implicites des coefficients linéaires et du paramètre du lissage. Au troisième niveau, le paramètre de lissage est un paramètre de complexité, estimé par minimisation de la VCG en se servant des coefficients linéaires comme des fonctions implicites du paramètre de lissage. Cette méthode est utilisée pour estimer les modèles additifs généralisés semiparamétriques des effets de la pollution de l'air sur la santé publique.

Finalement, les paramètres d'équations différentielles (ÉD) sont estimés à partir de données bruyantes par la méthode de profil généralisée. Dans le premier niveau d'optimisation, des courbes de lissage sont estimées pour obtenir des solutions approximatives des ÉD par lissage pénalisé, la pénalité étant définie par les ÉD en donnant des valeurs fixes à leurs paramètres. Au second niveau d'optimisation, les paramètres des ÉD sont estimés par la minimisation de la somme pondérée des carrés des erreurs, en traitant les coefficients de lissage comme une fonction implicite des paramètres des ÉD. Les effets du paramètre de lissage sur l'estimation des paramètres des ÉD sont explorés et l'on présente une discussion des critères de sélection du paramètre de lissage. La méthode est appliquée pour

accomoder le modèle dynamique de prédateur-proie à des données biologiques, également à l'estimation des paramètres d'ÉD dans le modèle dynamique du VIH d'essais cliniques ainsi que pour explorer des modèles pour la décomposition thermique de l'α-pinène.

Original Contributions

The following lists the most important original contributions in this dissertation.

1. In Chapter 2, penalized smoothing is extended by allowing for a functional smoothing parameter, which is adaptive to the geometry of the underlying curve. The variance estimate for the functional smoothing parameter also include the variation coming from the fitting function.

2. In Chapter 3, the generalized semiparametric additive models are estimated by three levels of optimization, allowing response variables in any kind of distributions. The optimization criteria are based on the likelihood function, instead of the simple sum of squared errors. Estimates for variances of linear coefficients also include variation coming from smoothing parameters.

3. In Section 4.6, the criteria for smoothing parameter selection are introduced when penalized smoothing data with the penalty term defined by differen-

tial equations. When differential equations are nonlinear, the approximated generalized cross-validation is also derived.

4. In Section 4.7, initial values for DE components are estimated when we fit DE's to noisy data, and the effect of the smoothing parameter is discussed when we estimate DE parameters from noisy data with the generalized profiling method.

5. In Section 4.8, functional parameters in DE's are estimated from noisy data.

6. In Section 4.9, a predator-prey dynamic model is estimated to fit the biological data.

7. In Section 4.10, statistical inferences are obtained for a HIV dynamic model from clinical trials.

8. In Section 4.11, dynamic models are explored for thermal decomposition of α-Pinene.

Acknowledgments

I would like to express my deepest gratitude to my supervisor, Dr. James O. Ramsay. He introduced me to an exciting area, functional data analysis, and shared many ideas with me. Whenever I was stuck in my research, he could always find some solutions and keep me going forward. He also supported me in finance, research equipment and going to conferences. Because of him, the PhD study became easy and exciting. I felt the deep love from him, like a father loves his son. Whenever I left him for a while, he would appear in my dreams. Jim, no thanks are enough for what you have given to me, and I will miss you!

Jim's research team was also an important academic support for my PhD study. Especially, Giles Hooker gave me many thoughtful comments for my research and presentation skills. I will never forget the happy times when we cooked, skated and skied together. My PhD study would have been much harder without him. David Campbell was always so kind to me. Both of them proofread my thesis and gave me many comments. I am so lucky to have them as my friends

and coauthors.

Thank Dr. Tim Ramsay for his data and great comments on the work to estimate the generalized semiparametric additive models. Thank Dr. Chong Gu for his kind help on the approximated generalized cross-validation. Thank Dr. Gregor Fussmann, Dr. Hulin Wu, and Dr. Yangxin Huang for their data of good quality and kind support in the predator-prey project and HIV project.

I thank all professors in statistics of the Department of Mathematics and Statistics: William Anderson, Masoud Asgharian, Jose Correa, Russell Steele, George Styan, Alain Vandal, David Wolfson and Keith Worsley. All of them have given me excellent lectures and all kinds of help and support. I also would like to thank Dr. William Brown for his kindness and patience as my TA's course coordinator. All the statistics graduate students are so kind to me. Pierre-Jerome Bergeron translated the abstract into French. The administrative faculty always give me so much help, especially Carmen Baldonado and Gregory LeBaron. The facilities in the department were also wonderful, and no place can be better than the Math library for studying and research.

On June 7, 2003, I got married with the most beautiful, smart and gentle girl, Liangliang Wang. Her love and support is my largest wealth. Honey, I promise to love and take care of you in all my life.

I was born in a poor village of China. The chance was very small for people there to obtain a PhD degree abroad, but I did. In my 18-year study, I owe so many people my thanks. My mother and father deserve my largest thanks. They raised me and gave me the first push in my way of fighting. My sister and aunts

always give me a helping hand. I also thank my kind teachers, especially Baozheng Su, Yupeng Zhao, Tao Wang and Jufu Feng.

I will be your pride!

Jiguo Cao

June, 2006

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Functional Data

This thesis focuses on modeling observations distributed over time, space, or some other continuum. Ramsay and Silverman (2005) define the *resolution* of a set of data as "inversely related to the width of the narrowest event that can be estimated to our satisfaction", and suggest that the resolution of a set of data is a more useful concept than simply the number of observations taken. If data resolutions are relatively high, these kinds of data are called *functional data.*

Functional data don't have to be sampled uniformly, and points at which functions are observed can vary among multiple replications. For instance, Figure 1.1 shows the HIV virus levels for 42 patients measured before treatment, and in around 1, 2, 4, 8, 12, 16, 20 and 24 weeks after treatment. The time points

to measure the HIV virus are nonequally spaced and different across 42 patients. These data were collected by AIDS Clinical Trials Group, Acosta et al. (2004). The number of HIV viruses for each patient shows a different pattern. Some patients, such as Patient 42, have their number of virus decreasing all the time. But other patients, such as Patient 23, have their virus levels going down at the beginning and up after 4 weeks. The HIV virus level is a function of time, and we have 42 functional data in total.

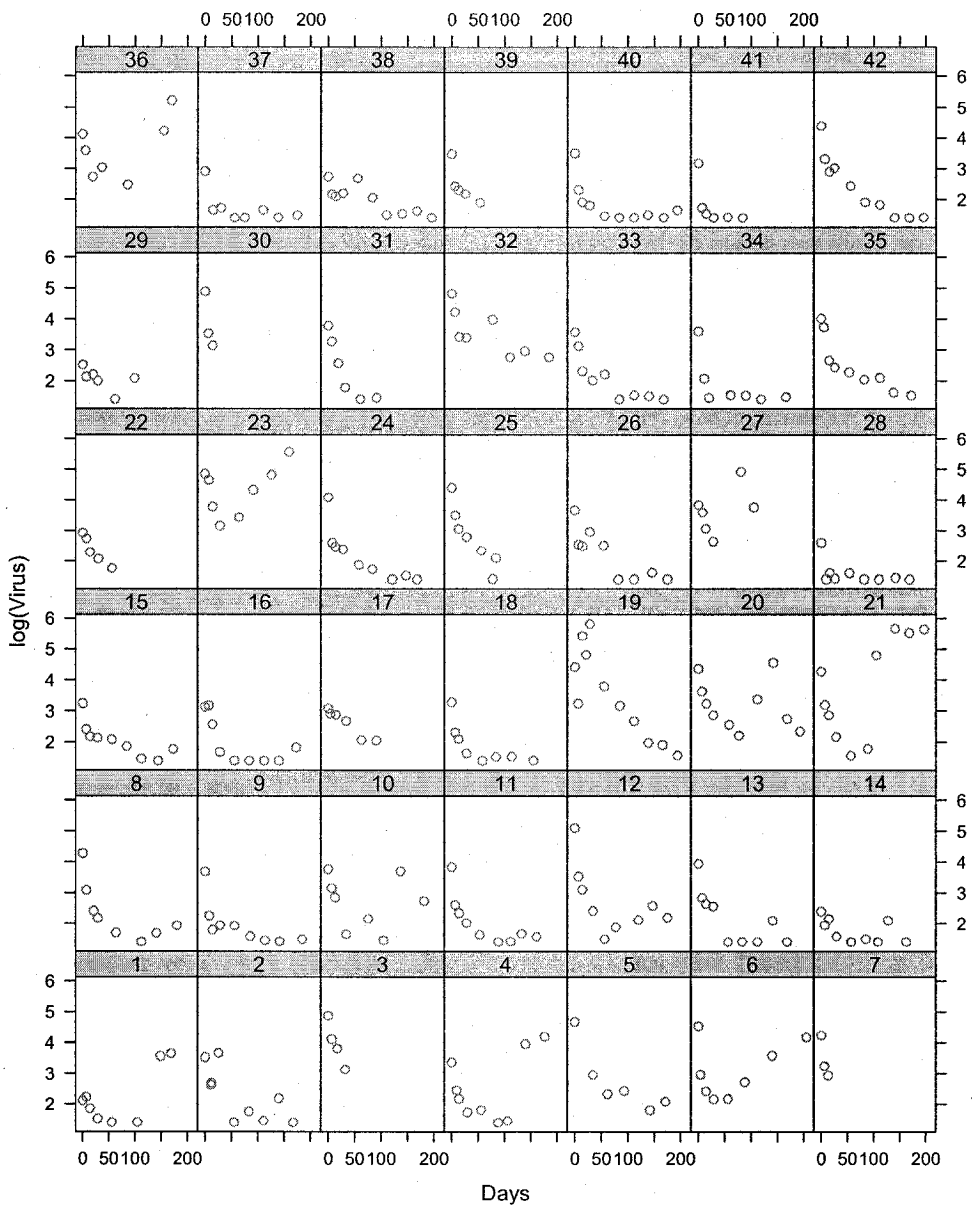

Figure 1.1: The number of free virus for 42 patients in logarithm scale.

In addition, functional data may be a single long record. For example, Figure 1.2 displays the daily counts of non-accidental deaths from 1987 to 1988 in Toronto, as well as the daily one-hour-maximum ozone. Both of them are also functional data.

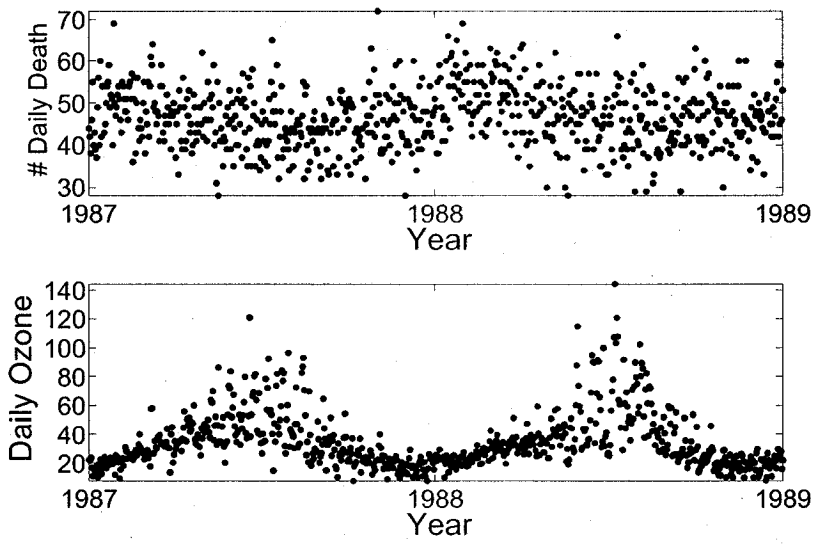

Figure 1.2: The top panel displays the daily count of non-accidental deaths from 1987 to 1988 in Toronto, and the bottom panel shows the associated daily one-hour-maximum ozone.

The objective of this thesis is to explore tools for functional data analysis (FDA) based on Ramsay and Silverman (2005). The central theme of FDA is the many uses of derivatives. We use the notation D for differentiation, for instance,

$$Dx(t) = \frac{dx(t)}{dt} \qquad \text{and} \qquad D^2x(t) = \frac{d^2x(t)}{dt^2}\,.$$

Representations of noisy observations in discrete time points as functions are often in forms of linear combinations of basis functions:

$$x(t) = c_1\phi_1(t) + c_2\phi_2(t) + \cdots + c_K\phi_K(t)\,, \tag{1.1}$$

where ϕ_i is the i-th basis function and c_i is the corresponding basis coefficient. The Fourier and spline basis systems are often used for periodic and non-periodic data, respectively. Smoothing and interpolation are two common tools to convert the discrete observations into functions. In the process, the dimensionality of data is reduced from the number of observations n per subject to the number of basis functions K used to represent functional data. The number of basis functions may be larger than the number of observations when the underlying functions are difficult to approximate because of sharp changes, discontinuity or other features. In this case, penalized smoothing can be applied to estimate fitting functions, which has a penalty term to control the roughness of estimated functions. The penalty term can be defined by some order of derivatives. Differential equations can also be applied to define the penalty term in penalized smoothing (Ramsay and Silverman 2005), leading to better estimates for fitting functions and their derivative. Penalized smoothing and differential equations are two key elements in this thesis.

It is useful to compare functional data with the kinds of data analyzed by more traditional methods, such as time series and longitudinal data analysis. Time series analysis usually requires observations to be stationary and time points between observations to be equally spaced. Differencing is widely used in time series analysis, but derivatives are the most popular elements in FDA. Compared with

longitudinal data analysis, functional data analysis requires more frequent observations and the time itself usually does not appear as an explicit covariate in functional models while some covariates and parameters can often be functions of time.

1.2 Local, Global and Complexity Parameters

In order to increase computational efficiency, most kinds of basis functions are zero except over short intervals, which is called the compact support property. Hence, each basis coefficient only controls the local behavior of the estimated function. These basis coefficients are therefore *local parameters*. Besides them, statistical models often involve *global parameters* and/or *complexity parameters*. The global parameters control the model everywhere. The complexity parameters are used in the roughness penalties and control the effective degrees of freedom of statistical models.

The distinction between local and global parameters was first discussed by Neyman and Scott (1948), where they were called local and global parameters, respectively. Let $\mathbf{X}_i$ be a (possibly multivariate) random variable, and the variables in the sequence $\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_n, \cdots$ be mutually independent. Parameters, $\theta_1, \theta_2, \cdots, \theta_m$ are *structural* or *global* if each appears in an infinite number of the probability laws of the observable random variables $\{\mathbf{X}_i\}_{i=1}^{\infty}$. A possibly infinite number of parameters, $\{\xi_k\}_{k=1}^{\infty}$, are *incidental* or *local* if each appears in a finite number of the probability laws of the observable random variables $\{\mathbf{X}_i\}_{i=1}^{\infty}$. In other words, the local parameters only capture the local variation and the num-

ber of them is large and increases with the size of data. On the other hand, the structural parameters are affected by the whole data and the number of them is small and fixed with the size of data. The local or incidental parameters can also be categorized as *nuisance parameters*, in the sense that they are required to construct statistical models but are not of direct interest. Local, global and complexity parameters are illustrated by three examples in the following subsections.

1.2.1 Adaptive Penalized Smoothing

The top panel of Figure 1.3 shows measurements of a property g of titanium changing with temperatures from 595 °C to 1075 °C, adapted from de Boor (2001). The measurement errors are small but not negligible. Because of the sharp peak, these data have become a standard challenge and have been used extensively as an example in nonparametric smoothing.

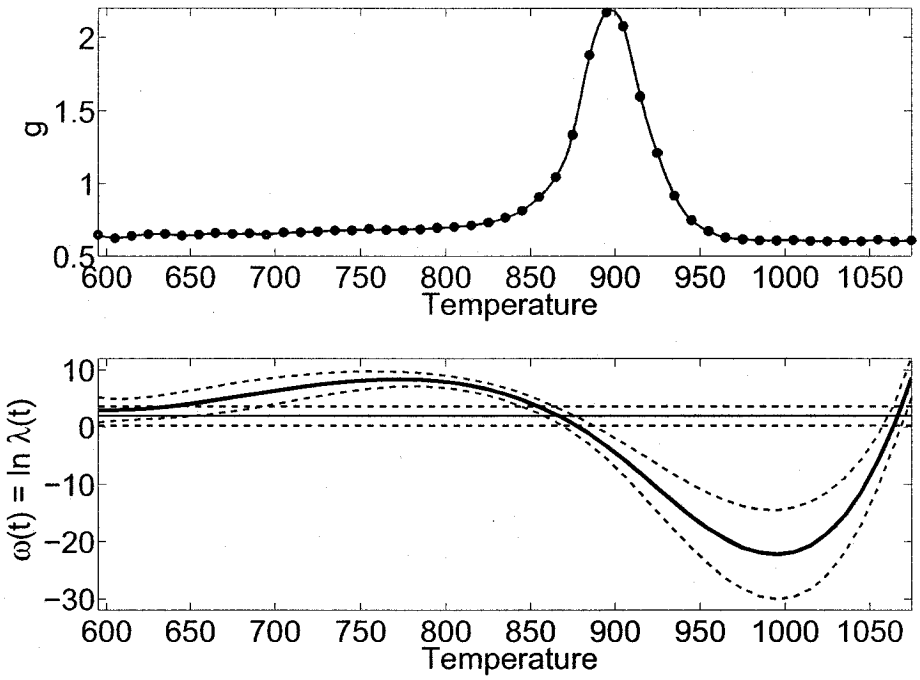

Figure 1.3: Top panel: The titanium heat data are smoothed by cubic B-splines defined by putting one knot at each observation using adaptive penalized smoothing. The dots are observations, and the solid line is the fitting function. Bottom panel: The optimal $\omega(t) = \ln \lambda(t)$ by minimizing GCV when it is a constant (thin solid line) or expanded by 5 cubic B-splines with a single interior knot at 900 (heavy solid line). The dashed curves define their 95% pointwise confidence bands.

There are two ways to estimate the fitting function from data. We can define basis functions by specifying the number of knots and their locations, and estimate basis coefficients by (weighted) least squares. Unfortunately, there is

no gold standard method to choose the optimal basis functions, and hence non-experienced users may find it difficult. On the other hand, users may put one knot on each observation, and avoid overfitting by defining a roughness penalty. The roughness penalty is often defined by the integrated squared second derivative of the fitting function. Let $x(t)$ be the fitting function in the form of (1.1), the coefficient vector $\mathbf{c} = (c_1, \cdots, c_K)$ can be estimated by minimizing the penalized sum of squared errors written as

$$H(\mathbf{c}|\lambda, \mathbf{y}) = \sum_{i=1}^{n} w_i[y_i - x(t_i)]^2 + \lambda \int [D^2 x(t)]^2 dt\,, \tag{1.2}$$

where w_i is the weight for the i-th observation y_i. The smoothing parameter λ measures the rate of exchange between fitting the data and varibility of the fitting function.

A better smooth can often be obtained by the latter method. As $\lambda \longrightarrow \infty$, the fitting function approaches the standard linear regression line, where the roughness penalty is 0. On the other hand, as $\lambda \longrightarrow 0$, the fitting function become more and more variable and eventually goes through all the observations. Hence, λ controls the complexity of fitting functions, and is a *complexity parameter*. The value of the smoothing parameter is chosen by generalized cross-validation or other criteria.

However, for titanium data shown in Figure 1.3, the underlying fitting function shows large variations over the range 850 °C to 950 °C, and is flat in the other intervals. This indicates that it may be more appropriate to allow different scales of roughness penalty according the geometry of the underlying function. For exam-

ple, we require a small value of the smoothing parameter λ over the range 850 °C to 950 °C and a large value of λ in other regions. In other words, λ should be a function of temperature, which is called as a *functional smoothing parameter.* Figure 1.3 shows the estimated log-transformed functional smoothing parameter $\lambda(t)$, which is exactly what we expect, but it is small in the region [950,1050], which is caused by the small measurement errors. This process is called *adaptive penalized smoothing.*

In adaptive penalized smoothing, the basis function coefficients are *local parameters*, and the functional smoothing parameter is a *complexity parameter.* Adaptive penalized smoothing can obtain good estimates of fitting functions and their derivatives. More details can be found in Chapter 2.

1.2.2 Generalized Semiparametric Additive Models

Generalized semiparametric additive models are widely used to explore the health effect of air pollution. The U.S. Environmental Protection Agency (EPA) periodically reviews the National Ambient Air Quality Standards for six air pollutants that protect the public's health, along with the updated statistical technology. In 2002, EPA delayed completion of the review documents because statisticians and epidemiologists found that the default settings in the *gam* function of the S-Plus software package (version 3.4) didn't assure that the back-fitting algorithm was convergent, and could overestimate effects of air pollution (Dominici, McDermott, Zeger, and Samet 2002). Moreover, Ramsay, Burnett, and Krewski (2003) showed that S-Plus also underestimated the variance of air pollution effects.

Figure 1.2 displays the daily counts of non-accidental deaths from 1987 to 1988 in Toronto, as well as the daily one-hour-maximum ozone. Let $\{y_j\}_{j=1}^n$ be daily counts of non-accidental deaths, x_j be the daily one-hour-maximum ozone, and j be the index of the day. If we assume y_j to have a Poisson distribution (possibly with over-dispersion), then the density function can be written as:

$$f(y_j) = \exp\{y_j\eta_j - e^{\eta_j} - \log(y_j!)\}\,.$$

The generalized semiparametric additive model for mean $\mu_j = \mathrm{E}(y_j)$ is

$$\eta_j = \log(\mu_j) = f(t_j) + \beta x_j\,, \tag{1.3}$$

where the functional parameter $f(t)$ takes account of the time effect on the response, which is represented by a linear combination of basis functions. The coefficients of basis functions are *local parameters*; the *global parameter* β represents the increase of the response associate with a unit increase in the amount of the covariate, allowing for the effects of the time trend. Moreover, a smoothing parameter λ is used to control the roughness penalty on $f(t)$. λ is therefore a *complexity parameter*. Chapter 3 introduces a method to estimate these three groups of parameters in three levels of optimization.

The variance of air pollution effects is usually estimated under the condition of a fixed value of the smoothing parameter λ. Therefore, the variance estimation potentially ignores the variation source coming from λ. Chapter 3 introduces a method to estimate the variance of global parameters unconditionally.

1.2.3 Estimating differential equations

Differential equations (DE's) are used to model the rate of change of a process defined over time, space, or some other continuum. They are widely used in engineering, biology, ecology, economics, neuroscience, and medicine. Recently DE's are also applied to model the dynamic behavior of gene expression (Jaeger et al. 2004). The oldest and most famous example is perhaps Newton's second law: $F = ma$, where a is the acceleration (the first derivative of the velocity or second derivative of the position), m is the mass, and F is the exogenous force. Newton's second law can also be written in the form of DE:

$$D^2x(t) = \frac{F}{m},$$

where $x(t)$ is the position function. This simple DE beautifully reveals the linear relationship between the acceleration and the force.

Estimating derivatives plays a central role in FDA. The traditional smoothing tools are often found to obtain unstable derivative estimates, especially at the boundaries. DE's are embraced in FDA because they explicitly describe the relationship between derivatives and functions.

For example, HIV dynamic models, usually in the forms of DE's, describe the rate of population change of uninfected cells, infected cells and virus as a function of their populations and interactions. They have contributed significantly to our understanding of HIV infection and the development of antiviral drug therapy. Huang, Liu, and Wu (2005) proposed a set of nonlinear DE's to characterize the

long-term HIV dynamics with antiretroviral therapy. Let U, I, and V be the number of uninfected cells, infected cells and free virus, respectively, their DE's are simplified as

$$\begin{aligned} \frac{d}{dt}U &= -\alpha \cdot U - \rho \cdot UV + \nu \\ \frac{d}{dt}I &= -\beta \cdot I + \rho \cdot UV \\ \frac{d}{dt}V &= -\gamma \cdot V + N \cdot \beta \cdot I\,, \end{aligned} \tag{1.4}$$

The first terms in the right sides of the three DE's take into account the death of uninfected and infected cells and the clearance of virus, respectively. Parameters α and β are the death rate of uninfected cells and infected cells, and γ is the clearance rate of free virus. The term $\rho \cdot UV$ characterizes the infection of uninfected cells by virus. This product term is based on the fact that the infection rate depends on not only the number of virus but also the number of uninfected cells. This makes sense if we assume that the more uninfected cells, the easier for the virus to "catch" an uninfected cell and infect it. Parameter ρ is the infection rate and ν is the rate at which new uninfected cells are created from sources within the body, such as the thymus. The term $N \cdot \beta \cdot I$ describes each infected cell as producing N new free virus during its life.

How can we estimate the six parameters in DE's (1.4) from data shown in Figure 1.1? This is called the system identification problem in engineering. The current methods to estimate parameters in DE's from noisy data are slow and unstable. There are few statistical techniques to conduct formal and rigorous interval estimates and inferences. Chapter 4 introduces one approach to obtain

statistical inferences for parameters defining DE's. DE solutions are estimated by a linear combination of basis functions, instead of solving DE's directly. This is implemented by penalized smoothing with the roughness penalty defined by DE's. The basis coefficients are *local parameters*. The parameters in DE's are *global parameters*. The smoothing parameter controls the trade-off between fitting data and satisfying DE's, and therefore is a *complexity parameter*.

1.3 Literature Review for Nuisance and Structural Parameter Estimations

It is difficult to obtain statistical inferences for structural parameters in the presence of many nuisance parameters. Among various methods of eliminating nuisance parameters, the most straightforward for Bayesian analysis is to obtain the marginal posterior distribution of the structural parameters by integrating the joint posterior distribution over the nuisance parameters (Gelman, Carlin, Stern, and Rubin 2004). But it is often difficult to find the closed form of marginal posterior distributions, and in this case Markov chain Monte Carlo (MCMC) is a popular method to obtain the samples for structural parameters. Another simulation method is to draw samples from the joint posterior distribution and then focus on values of structural parameters while ignore values of nuisance parameters. The drawback of this method is the intensive and inefficient computations that are required.

Profiling the likelihood is another standard approach to eliminate nuisance

parameters. In the following, we consider a statistical model for the vector of observations $\mathbf{y}$ with parameters $(\boldsymbol{\theta}, \mathbf{c})$, where $\boldsymbol{\theta}$ is the vector of structural parameters and $\mathbf{c}$ is the vector of nuisance parameters. Let $\hat{\mathbf{c}}_{\boldsymbol{\theta}}$ stand for the maximum likelihood estimate (MLE) of $\mathbf{c}$ for each fixed $\boldsymbol{\theta}$, the profile likelihood can be defined as

$$L_p(\boldsymbol{\theta}|\mathbf{y}) = L(\boldsymbol{\theta}, \hat{\mathbf{c}}_{\boldsymbol{\theta}}|\mathbf{y}) = \sup_{\mathbf{c}} L(\boldsymbol{\theta}, \mathbf{c}|\mathbf{y}), \tag{1.5}$$

and the optimal value for $\boldsymbol{\theta}$ is then obtained by maximizing $L_p(\boldsymbol{\theta}|\mathbf{y})$.

Let's consider the example of the Neyman-Scott problem. Let $y_{ij} \sim \texttt{Normal}(\mu_j, \sigma^2)$ for $i = 1, \cdots, n$; $j = 1, 2$ and $\mu_j \sim \texttt{Normal}(\mu_0, \sigma_0^2)$. Assuming that σ^2 and σ_0^2 are known, μ_0 is the structural parameter and μ_j's are nuisance parameters, we can write the negative log likelihood function up to a constant as

$$l(\mu_j, \mu_0|\mathbf{y}) = \frac{1}{\sigma^2}\sum_{j=1}^{2}\sum_{i=1}^{n}(y_{ij} - \mu_j)^2 + \frac{1}{\sigma_0^2}\sum_{j=1}^{2}(\mu_j - \mu_0)^2. \tag{1.6}$$

By minimizing the negative log likelihood $l(\mu_j, \mu_0|\mathbf{y})$ with the fixed value of μ_0, we obtain the estimate for μ_j as an explicit function of the structural parameter μ_0:

$$\hat{\mu}_j = \frac{\sigma^2}{\sigma^2 + n\sigma_0^2}\mu_0 + \frac{n\sigma_0^2}{\sigma^2 + n\sigma_0^2}y_{\cdot j}, \tag{1.7}$$

where $y_{\cdot j} = \sum_{i=1}^{n} y_{ij}/n$. Then by plugging $\hat{\mu}_j$ into the log likelihood (1.6), we obtain the profile log likelihood

$$l(\mu_0|\mathbf{y}) = \frac{1}{\sigma^2}\sum_{j=1}^{2}\sum_{i=1}^{n}(y_{ij} - \hat{\mu}_j)^2 + \frac{1}{\sigma_0^2}\sum_{j=1}^{2}(\hat{\mu}_j - \mu_0)^2. \tag{1.8}$$

By minimizing the profile log likelihood (1.8), we attain the estimate for μ_0 as

$$\hat{\mu}_0 = \frac{1}{n}\sum_{j=1}^{2}\sum_{i=1}^{n} y_{ij}\,. \tag{1.9}$$

But $\mathrm{E}(\hat{\mu}_0) = 2\mu_0$, so $\hat{\mu}_0$ is a biassed estimator for μ_0.

This result is not surprising if we realize that the profile likelihood is not a true likelihood. For example, let $\boldsymbol{\alpha}$ denote for a vector of all parameters in the likelihood function $L(\boldsymbol{\alpha}|\mathbf{y}) = \exp(l(\boldsymbol{\alpha}|\mathbf{y}))$. Then most log likelihood functions $l(\boldsymbol{\alpha}|\mathbf{y})$ satisfy

$$\mathrm{E}\left(\frac{\partial l}{\partial \boldsymbol{\alpha}}\right) = 0\,; \tag{1.10}$$

$$\mathrm{E}\left[\left(\frac{\partial l}{\partial \boldsymbol{\alpha}}\right)\left(\frac{\partial l}{\partial \boldsymbol{\alpha}}\right)^{T} + \frac{\partial^2 l}{\partial \boldsymbol{\alpha}\partial \boldsymbol{\alpha}^{T}}\right] = 0\,. \tag{1.11}$$

The function $\partial l/\partial \boldsymbol{\alpha}$ is called the score function. But Identities (1.10) and (1.11) do not hold for the profile likelihood functions, in general. The profiling estimate $(\hat{\boldsymbol{\theta}}, \hat{\mathbf{c}}_{\hat{\boldsymbol{\theta}}})$ is sometimes not equal to the joint MLE $(\hat{\boldsymbol{\theta}}, \hat{\mathbf{c}})$, and it can cause both the bias and incorrect standard error estimates, as shown in the above example. Therefore, several adjustments have been proposed for the profile likelihood.

Barndorff-Nielsen (1983) approximated the profile likelihood as follows:

$$L_{BN}(\boldsymbol{\theta}|\mathbf{y}) = \left|\frac{\partial \hat{\mathbf{c}}_{\boldsymbol{\theta}}}{\partial \hat{\mathbf{c}}}\right|^{-1} |T_{\mathbf{c},\mathbf{c}}(\boldsymbol{\theta}, \hat{\mathbf{c}}_{\boldsymbol{\theta}})|^{-1/2} L_p(\boldsymbol{\theta}|\mathbf{y})\,, \tag{1.12}$$

where $T_{\mathbf{c},\mathbf{c}}(\boldsymbol{\theta}, \mathbf{c}) = -\partial^2 l/\partial \mathbf{c}\partial \mathbf{c}^T$. Ferguson et al. (1991) and DiCiccio et al. (1996) showed that the biases of the score and information functions are of order $O(1/n)$,

that is

$$\mathrm{E}\left(\frac{\partial l}{\partial \boldsymbol{\alpha}}\right) = O(1/n)\,;$$

$$\mathrm{E}\left[\left(\frac{\partial l}{\partial \boldsymbol{\alpha}}\right)\left(\frac{\partial l}{\partial \boldsymbol{\alpha}}\right)^T + \frac{\partial^2 l}{\partial \boldsymbol{\alpha}\partial \boldsymbol{\alpha}^T}\right] = O(1/n)\,.$$

$L_{BN}(\boldsymbol{\theta}|\mathbf{y})$ is also invariant under transformations of parameters.

However, it is difficult to obtain $|\partial \hat{\mathbf{c}}_{\boldsymbol{\theta}}/\partial \hat{\mathbf{c}}|$ when calculating the modified profile likelihood function $L_{BN}(\boldsymbol{\theta}|\mathbf{y})$. There is an alternative expression for $L_{BN}(\boldsymbol{\theta}|\mathbf{y})$ that does not involve this term. Assuming a as an ancillary statistic such that $(\hat{\boldsymbol{\theta}}, \hat{\mathbf{c}}, a)$ is a minimal sufficient statistic, and $l(\boldsymbol{\theta}, \mathbf{c}|\mathbf{y})$ as the log-likelihood function that depends on the data only through $(\hat{\boldsymbol{\theta}}, \hat{\mathbf{c}}, a)$, Barndorff-Nielsen (1983) showed that

$$L_{BN}(\boldsymbol{\theta}|\mathbf{y}) = \left|\frac{\partial^2 l}{\partial \mathbf{c}\partial \hat{\mathbf{c}}}\right|^{-1} |T_{\mathbf{c},\mathbf{c}}(\boldsymbol{\theta}, \hat{\mathbf{c}}_{\boldsymbol{\theta}})|^{1/2} L_p(\boldsymbol{\theta}|\mathbf{y})\,. \tag{1.13}$$

But the alternative expression for $L_{BN}(\boldsymbol{\theta}|\mathbf{y})$ requires the specification of an ancillary statistic and the calculation of the sample space derivative $\partial^2 l/\partial \mathbf{c}\partial \hat{\mathbf{c}}$. Several approximations to $L_{BN}(\boldsymbol{\theta}|\mathbf{y})$ were proposed to simplify its evaluation by Severini (2000). These approximations do not require to calculate the sample space derivative and do not involve the ancillary statistics, either.

Cox and Reid (1987) proposed another adjustment to the profile likelihood function when the structural and nuisance parameters were orthogonal, that is, $\partial l/\partial \boldsymbol{\theta}$ and $\partial l/\partial \mathbf{c}$ were uncorrelated. Their modified profile likelihood function is

$$L_{CR}(\boldsymbol{\theta}) = |T_{\mathbf{c},\mathbf{c}}(\boldsymbol{\theta}, \hat{\mathbf{c}}_{\boldsymbol{\theta}})|^{-1/2} L_p(\boldsymbol{\theta}|\mathbf{y})\,. \tag{1.14}$$

Under the orthogonality of the structural and nuisance parameters, Liang (1987) showed that the bias of the score function was of order $O(1/n)$, but the information bias was not of order $O(1/n)$ (DiCiccio et al. 1996).

The modified profile likelihood function $L_{CR}(\boldsymbol{\theta})$ requires the orthogonality of the structural and nuisance parameters, but it is not always possible to find this parameterization. Moreover, it is not invariant under the parameter transformations.

1.4 The Generalized Profiling Method

We can generalize the profile likelihood method as follows. Besides the nuisance parameter vector $\mathbf{c}$ and the structural parameter vector $\boldsymbol{\theta}$, we assume that our statistical models also have another distinct group of parameters, the complexity parameter λ. Three levels of optimization are used to estimate these three groups of parameters. In the first level, the nuisance parameter vector $\mathbf{c}$ is estimated by optimizing a criterion $J(\mathbf{c}|\boldsymbol{\theta}, \lambda, \mathbf{y})$ for each fixed value of $\boldsymbol{\theta}$ and λ. $\mathbf{c}$ is eliminated from the parameter space by treating the estimate $\hat{\mathbf{c}}$ as an explicit or implicit function of $\boldsymbol{\theta}$ and λ. In the second level, the structural parameter vector $\boldsymbol{\theta}$ is estimated by optimizing a criterion $H(\boldsymbol{\theta}|\lambda, \mathbf{y})$ for each fixed value of λ. Thus $\boldsymbol{\theta}$ is removed from the parameter space by treating the estimate $\hat{\boldsymbol{\theta}}$ as an explicit or implicit function of λ. In the third level, the complexity parameter λ is the only parameter left in the model, and can be estimated by optimizing a criterion $F(\lambda|\mathbf{y})$ from data.

Each level can have a different optimization criterion. For example, the modified profile likelihood methods use the likelihood function as the optimization criterion in the first level and the modified profile likelihood function as the optimization criterion in the second level. The modified profile likelihood methods are special cases in the generalized profiling method in that they only have two levels of optimization without or fixing the complexity parameter and the optimization criteria are special.

The Newton-Raphson method is applied in each level of optimization, and the gradient and Hessian matrix can be obtained analytically, using the Implicit Function Theorem which is introduced in Section 2.2. So the optimization process converges quickly. Section 2.2 and Section 2.3 give more mathematical details about the analytical formulas of the gradient and Hessian matrix for general criteria in two levels of optimization. Chapter 3 introduces how to deal with three distinct groups of parameters in three levels of optimization.

After obtaining the complexity parameter estimate, we can go back to first estimate $\hat{\boldsymbol{\theta}}$ and then $\hat{\mathbf{c}}$ in two steps, since $\hat{\boldsymbol{\theta}}$ is a function of λ, and $\hat{\mathbf{c}}$ is a function of $\boldsymbol{\theta}$ and λ. It is also important to have the functional relationship among three groups of parameters. For example, when we use the Delta method (Casella and Berger 1990) to estimate the standard error of the structural parameter vector $\boldsymbol{\theta}$, we can calculate the full derivative of $\boldsymbol{\theta}$ with respect to data $\mathbf{y}$ as:

$$\frac{d\boldsymbol{\theta}}{d\mathbf{y}} = \frac{\partial\boldsymbol{\theta}}{\partial\mathbf{y}} + \frac{\partial\boldsymbol{\theta}}{\partial\lambda}\frac{\partial\lambda}{\partial\mathbf{y}}. \tag{1.15}$$

If λ is fixed, the second term in the right side of Equation (1.15) is 0, then the

standard error of $\boldsymbol{\theta}$ is underestimated.

It is also important to have a different criterion in each level of optimization to obtain the unbiased estimates. Otherwise, the profile likelihood method can lead to a biased estimate, as shown in the Neyman-Scott problem. This is also the initial motivation to propose the modified profile likelihood methods. Then another key question is coming up: how can we decide which criterion to use in each level of optimization? Generally, the first and second level can use the criteria proposed by modified profile likelihood methods, although those criteria are difficult to evaluate. The third level to estimate the complexity parameter λ can use any criteria for model selection, for example, Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), Deviance Information Criterion (DIC), Cross Validation (CV) and Generalized Cross Validation (GCV).

All the work in this thesis is based on the penalized nonparametric smoothing method. For this situation, different criteria from modified profile likelihood methods are used for the first and second level of optimizations. The idea is summarized here and is explained in detail in the following three chapters. The regularized likelihood function is the optimization criterion in the first level. The optimization criterion in the second level is just the likelihood function without the regularization term, because the estimated nuisance parameter vector $\hat{\mathbf{c}}$ already contains the regularizing information, and this information passes to the second level of optimization by treating $\hat{\mathbf{c}}$ as a function of $\boldsymbol{\theta}$ and λ.

Let us return to the example of Neyman-Scott problem as an illustration of the generalized profiling method. We set up the Neyman-Scott problem as

a data smoothing problem, as shown in Figure 1.4. y_{i1} are observations along with time points $t_{i1} = 1, 2, \cdots, n$. y_{i2} are observations along with time points $t_{i2} = n, n+1, \cdots, 2n-1$. The fitting function $\mu(t)$ is a linear combination of an order 1 B-spline basis system with an interior knot on the time point n. That is, $\mu(t)$ is a two-value step function, with one constant value μ_1 over the time interval $[1, n]$, and another constant value μ_2 over the time interval $[n, 2n-1]$. The negative log likelihood function $l(\mu_j, \mu_0|\mathbf{y})$ in (1.6) can also be written as the penalized sum of squared errors for penalized smoothing:

$$l(\mu_j, \mu_0|\mathbf{y}) = \frac{1}{\sigma^2}\sum_{j=1}^{2}\sum_{i=1}^{n}(y_{ij} - \mu(t_{ij}))^2 + \frac{1}{\sigma_0^2}\int(\mu(t) - \mu_0)^2 dt\,, \tag{1.16}$$

where the smoothing parameter $\lambda = \sigma^2/\sigma_0^2$.

We use the generalized profiling method to estimate the structural parameter μ_0 and the nuisance parameters μ_j's. Equation (1.16) is used as the criterion for the first level of optimization. By minimizing $l(\mu_j, \mu_0|\mathbf{y})$, we get the estimate $\hat{\mu}_j$ as an explicit function of the structural parameter μ_0:

$$\hat{\mu}_j = \frac{\sigma^2}{\sigma^2 + n\sigma_0^2}\mu_0 + \frac{n\sigma_0^2}{\sigma^2 + n\sigma_0^2}y_{\cdot j}\,, \tag{1.17}$$

where $y_{\cdot j} = \sum_{i=1}^{n} y_{ij}/n$. As explained above, the optimization criterion in the second level drops the roughness penalty term, which is written as:

$$H(\mu_0|\mathbf{y}) = \frac{1}{\sigma^2}\sum_{j=1}^{2}\sum_{i=1}^{n}(y_{ij} - \hat{\mu}_j)^2\,.$$

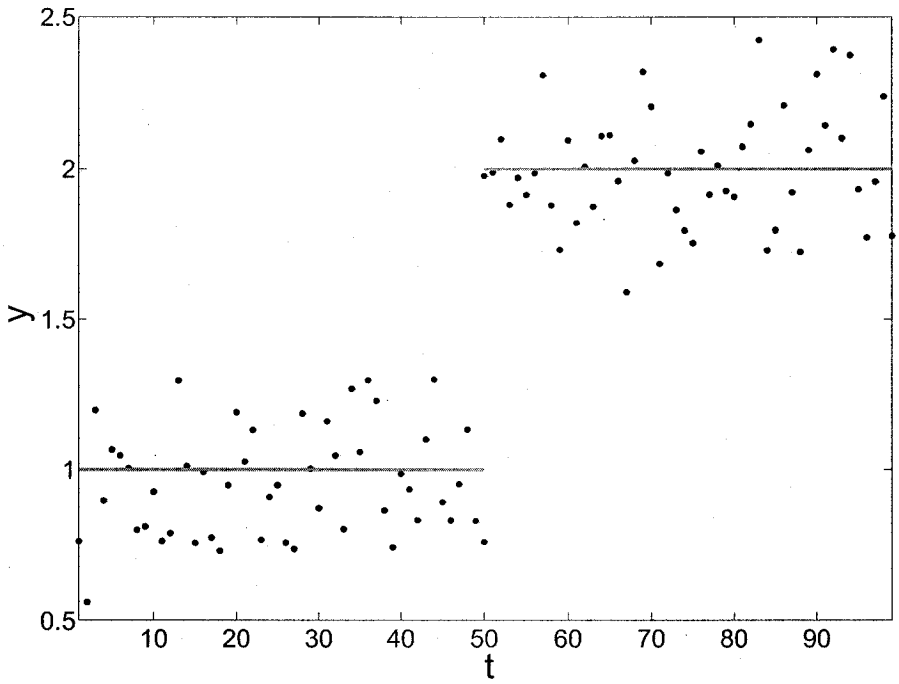

Figure 1.4: Illustration for the Neyman-Scott problem. The two clusters of black dots are normally distributed observations y_{ij}'s with the respective means μ_j's (red lines) and the same variance σ^2, $j = 1, 2$, $i = 1, \cdots, 50$. We assume that y_{i1}'s are observed at the time points $t_{i1} = 1, \cdots, 50$, and y_{i2}'s are observed at the time points $t_{i2} = 50, \cdots, 99$.

The optimal value for μ_0 by minimizing $H(\mu_0|\mathbf{y})$ is

$$\hat{\mu}_0 = \frac{1}{2n} \sum_{j=1}^{2} \sum_{i=1}^{n} y_{ij} \,. \tag{1.18}$$

Plugging the formula for $\hat{\mu}_0$ into (1.7), the estimate for μ_j is written as:

$$\hat{\mu}_j = \frac{\frac{\sigma^2}{2n}\sum_{j=1}^{2}\sum_{i=1}^{n} y_{ij} + \sigma_0^2 \sum_{i=1}^{n} y_{ij}}{\sigma^2 + n\sigma_0^2}. \tag{1.19}$$

It can easily be shown that the estimates for both the structural parameter μ_0 and the nuisance parameters μ_j's are unbiased.

1.5 Outline of the Thesis

Chapter 2 reviews the literature about nonparametric regression, and introduces the point and interval estimations for global and local parameters with the generalized profiling method for the general case. Adaptive penalized smoothing is then introduced, which has a functional smoothing parameter, adaptive to the geometry property of underlying curves. We compare the adaptive penalized smoothing with non-adaptive penalized smoothing and investigate the effect of data noise, data resolution and basis systems on the adaptive penalized smoothing based on simulated data. The estimates for variances of functional smoothing parameter and fitting functions are also verified. Finally, adaptive penalized smoothing is applied to smooth titanium heat data and to estimate second derivatives of growth curves. Matrix calculations for adaptive penalized smoothing are shown in Appendix A.

Chapter 3 reviews the literature on estimating generalized semiparametric additive models and introduces how to estimate these models based on likelihood functions by the generalized profiling method, allowing the response variable in any distribution. Our method is then applied to estimate the effect of air pollution on

the public's health. Variances for global parameters are estimated unconditionally, including variation coming from complexity parameters. All the mathematical details are written in Appendix B. The generalized profiling method shown in this chapter is also easy to extend to estimate other statistical models involve three distinct groups of parameters by changing with appropriate criteria.

Chapter 4 reviews the literature about estimating DE's from noisy data and introduces how to estimate fitting curves by penalized smoothing with the penalty term defined by DE's, with the smoothing parameter selected by generalized cross-validation and Stein's unbiased risk estimate. We introduce how to estimate DE parameters from noisy data with the generalized profiling method, and discuss the effect and selection of the smoothing parameter. Our method is applied to fit the predator-prey DE's and the HIV DE's to true data and explore dynamic models for the thermal decomposition of α-Pinene.

Chapter 5 provides the summary of the work, and discusses unanswered questions and further directions in research.

Chapter 2

Adaptive Penalized Smoothing

Nonparametric regression, or smoothing, describes the flexible association between covariates and responses, and many competing methods have been proposed, including the kernel-based method and the spline smoothing. Let $(t_1, y_1), \cdots, (t_n, y_n)$ be a random sample, we consider the following statistical model:

$$y_i = x(t_i) + \epsilon_i \,,$$

where $x(t)$ is an unknown smooth function to be estimated, and ϵ_i is the measurement error on t_i with mean 0.

The remainder of this chapter is organized as follows. Section 2.1 reviews the literature on nonparametric regression. Section 2.2 introduces point estimations for global and local parameters with the generalized profiling method for the general case, with the interval estimations for the two groups of parameters given in Section

2.3. Section 2.4 introduces adaptive penalized smoothing, which has a functional smoothing parameter, adaptive to the geometry property of underlying curves. Section 2.5 compares adaptive penalized smoothing with non-adaptive penalized smoothing and investigates the effects of data noise, data resolution and choice of basis systems on the adaptive penalized smoothing based on the simulated data. The estimates for variances of functional smoothing parameters and fitting functions are also verified. The applications to titanium heat data and growth curves are shown in Section 2.6 and Section 2.7, respectively.

2.1 Literature Review for Nonparametric Regression

Without specifying the form of $x(t)$, the most intuitive idea is that the influence of observations on $x(t)$ decreases with their distance to t. Hence $x(t)$ can be estimated by the locally weighted average:

$$\hat{x}(t) = \sum_{i=1}^{n} w_i(t) Y_i \,, \tag{2.1}$$

where $w_i(t)$ is the local weight, satisfying $\sum_{i=1}^{n} w_i(t) = 1$. Nadaraya (1964) and Watson (1964) proposed that:

$$w_i(t) = \frac{K_h(T_i - t)}{\sum_{i=1}^{n} K_h(T_i - t)} \,,$$

where $K_h(\cdot) = K(\cdot/h)/h$, and the kernel function $K(\cdot)$ is a symmetric probability density. It is common to use Gaussian kernel $K(t) = (1/\sqrt{2\pi})\exp(-t^2/2)$ and the symmetric Beta family

$$K(t) = \frac{[(1-t^2)_+]^\gamma}{Beta(1/2, \gamma+1)}, \gamma = 0, 1, \cdots,$$

where the subscript $+$ denotes the positive part. The choices $\gamma = 0, 1, 2, 3$ correspond to the uniform, Epanechnikov, biweight, and triweight kernel functions, respectively. The bandwidth h is a nonnegative number controlling the size of the local neighborhood. Gasser and Müller (1979) gave another form of $w_i(t)$:

$$w_i(t) = \int_{s_{i-1}}^{s_i} K_h(u-t)du\,,$$

where $s_i = (T_i + T_{i+1})/2$, $T_0 = -\infty$ and $T_{n+1} = +\infty$.

Fan (1992) and Fan and Gijbels (1992) proposed local polynomial fitting by using Taylor's expansion for $x(t)$. Specially, when the polynomial is in order 1, the estimator is called a local linear regression smoother. This estimator can also be written in the form of (2.1) with $w_i = v_i/\sum_{i=1}^n v_i$ where

$$v_i = K_h(t_i - t)(S_{n2} - (t_i - t)S_{n1})\,;$$

$$S_{nj} = \sum_{i=1}^{n} K_h(t_i - t)(t_i - t)^j\,.$$

Fan (1992) summarized the pointwise asymptotic bias and variance of these three estimators in Table 2.1.

Table 2.1: Pointwise asymptotic bias and variance of kernel regression smoothers

Method	Bias	Variance
Nadaraya-Watson	$(x''(t) + \frac{2x'(t)f'(t)}{f(t)})b_n$	V_n
Gasser-Müller	$x''(t)b_n$	$1.5V_n$
Local linear	$x''(t)b_n$	V_n

Here, $b_n = \frac{1}{2}h^2 \int_{-\infty}^{+\infty} u^2 K(u)du$ and $V_n = \frac{\mathtt{Var}(Y|T=t)}{f_T(t)nh} \int_{-\infty}^{+\infty} K^2(u)du$.

A basis function system is a set of known functions $\{\phi_k(t)\}_{k=1}^{K_c}$ that are mathematically independent of each other and a linear combination of them can well approximate any functions. There are many good basis function systems. For instance, the Fourier basis system is usually used to approximate periodic functions.

Any piecewise smooth general function can be well approximated by the spline basis system, which is defined by a sequence of knots. de Boor (2001) shows how to improve the spline approximation ability and efficiency by knot selection. However, there are few gold standard methods that can select the optimal knot sequence automatically. Instead, we prefer to put at least one knot on each point with an observation, so that the basis function expansion is powerful enough to capture any amount of variation in the observed data. To prevent the estimated curve from overfitting the data, we require a roughness penalty in our optimization criterion.

Suppose the fitting function $x(t)$ can be approximated by a linear expansion

of K_c basis functions $\{\phi_k(t)\}_{k=1}^{K_c}$ as follows:

$$x(t) = \sum_{k}^{K_c} c_k \phi_k(t) = \mathbf{c}'\boldsymbol{\phi}(t)\,,$$

where $\boldsymbol{\phi}(t)$ is a vector of the basis functions and $\mathbf{c}$ is a vector of coefficients. The fitting criterion for penalized smoothing is given by

$$H(\mathbf{c}|\lambda, \mathbf{y}) = \sum_{i=1}^{n} w_i[Y_i - x(t_i)]^2 + \lambda \int [Lx(t)]^2 dt\,, \tag{2.2}$$

where w_i is the weight for the i-th observation y_i. For data with inconstant variance, w_i can be designed to be the reciprocal of the variance $\texttt{Var}(y_i)$. L is a linear differential operator of order m:

$$Lx(t) = \sum_{j=0}^{m-1} \beta_j(t) D^j x(t) + D^m x(t)\,.$$

All simulations and applications in this chapter use the second derivative to define the roughness penalty term, that is, $L = D^2$. Chapter 4 talks about how to use a general differential operator L to define the roughness penalty term, and estimate L from data.

Let $K_c \times K_c$ matrix $\mathbf{R} = \int [L\boldsymbol{\phi}(t)][L\boldsymbol{\phi}(t)]'dt$, and $\boldsymbol{\Phi}$ is an $n \times K_c$ matrix with the jk-th element $\Phi_{jk} = \phi_k(t_j)$. By minimizing $H(\mathbf{c}|\lambda, \mathbf{y})$, we can estimate the coefficient vector $\mathbf{c}$, which is written analytically as an explicit function of λ and $\mathbf{y}$:

$$\hat{\mathbf{c}}(\lambda, \mathbf{y}) = [\boldsymbol{\Phi}'\mathbf{W}\boldsymbol{\Phi} + \lambda\mathbf{R}]^{-1}\boldsymbol{\Phi}'\mathbf{W}\mathbf{y}\,, \tag{2.3}$$

where $\mathbf{W}$ is the weight matrix, which can be a diagonal matrix with the diagonal elements w_i, or allow for more complex covariance structures among residuals.

The optimal smoothing parameter $\hat{\lambda}$ can be chosen by minimizing the generalized cross-validation (GCV):

$$\texttt{GCV}(\lambda) = \left[\frac{n}{\texttt{dfe}(\lambda)}\right]\left[\frac{\texttt{SSE}(\lambda)}{\texttt{dfe}(\lambda)}\right], \tag{2.4}$$

where both the degrees of freedom measure $\texttt{dfe}(\lambda)$ and the sum of squared errors $\texttt{SSE}(\lambda)$ can be written in terms of the order n matrix $\mathbf{A}(\lambda) = \mathbf{\Phi}(\mathbf{\Phi}'\mathbf{W}\mathbf{\Phi} + \mathbf{R})^{-1}\mathbf{\Phi}'\mathbf{W}$:

$$\begin{aligned} \texttt{dfe}(\lambda) &= n - \texttt{Tr}[\mathbf{A}(\lambda)]; \\ \texttt{SSE}(\lambda) &= \mathbf{y}'[I - \mathbf{A}(\lambda)]'[I - \mathbf{A}(\lambda)]\mathbf{y}. \end{aligned}$$

Penalized smoothing has been found to produce better estimates of functions and their derivatives than the kernel-based methods, and Ramsay and Silverman (2005) show how to obtain better estimates for derivatives by penalized smoothing with penalty terms defined with differential operators.

As discussed in Chapter 1, the coefficient vector $\mathbf{c}$ is a local parameter, and the smoothing parameter λ is a complexity parameter. Parameter λ is also a global parameter in the sense that it controls the shape of the whole fitting function. The estimate $\hat{\mathbf{c}}$ is attained by minimizing the first level optimization criterion (2.2), conditional on λ. The smoothing parameter λ is then estimated by minimizing the optimization criterion GCV in the second level with $\hat{\mathbf{c}}$ treated as an explicit function

of λ. In the next two sections, we will derive the generalized profiling method for a general statistical model involving the local and global parameters, allowing that the local parameter is an explicit or implicit function of the global parameter.

2.2 Point Estimations for Global and Local Parameters

In this section, we outline how to estimate the local and global parameters with the generalized profiling method for the general case. That is, the two levels of optimization can apply to any criteria and the optimal local parameters can be explicit or implicit functions of global parameters. The generalized profiling method is also used in Chapters 3 and 4, except that Chapter 3 estimates three distinct groups of parameters in three levels of optimization.

Let $\boldsymbol{\theta}$ be a vector of global parameters, and $\mathbf{c}$ be a vector of local parameters. The statistical model is assumed not to involve the complexity parameter λ or has a fixed value of λ. Chapter 3 shows how to estimate λ in the third level of optimization. We assume that $\mathbf{c}$ can be uniquely estimated by optimizing the criterion $H(\mathbf{c}|\boldsymbol{\theta}, \mathbf{y})$ in the first level, conditional on $\boldsymbol{\theta}$ and $\mathbf{y}$. In this way, the estimated local parameter vector $\hat{\mathbf{c}}$ is defined as an explicit or implicit function of $\boldsymbol{\theta}$ and $\mathbf{y}$. Then we can estimate the global parameter vector $\boldsymbol{\theta}$ by optimizing the criterion $F(\hat{\mathbf{c}}(\boldsymbol{\theta}, \mathbf{y}), \boldsymbol{\theta}|\mathbf{y})$ in the second level, conditional on $\mathbf{y}$, where $\hat{\mathbf{c}}$ is removed from the parameter space as a function of $\boldsymbol{\theta}$. Thus the optimal global parameter vector $\hat{\boldsymbol{\theta}}$ is defined as an explicit or implicit function of $\mathbf{y}$. The functional relationship

between $\hat{\boldsymbol{\theta}}$ and $\mathbf{y}$ is used to estimate the variance of $\hat{\boldsymbol{\theta}}$, which is introduced in the next section. Here and below, all partial derivatives as well as total derivatives are assumed to be evaluated at $\hat{\mathbf{c}}$ and the optimal global parameter vector $\hat{\boldsymbol{\theta}}$.

The optimization of $F(\hat{\mathbf{c}}(\boldsymbol{\theta}, \mathbf{y}), \boldsymbol{\theta}|\mathbf{y})$ becomes much faster and more stable if we have the gradient

$$\frac{dF(\hat{\mathbf{c}}(\boldsymbol{\theta}, \mathbf{y}), \boldsymbol{\theta}|\mathbf{y})}{d\boldsymbol{\theta}} = \frac{\partial F(\hat{\mathbf{c}}(\boldsymbol{\theta}, \mathbf{y}), \boldsymbol{\theta}|\mathbf{y})}{\partial \boldsymbol{\theta}} + \frac{\partial F(\hat{\mathbf{c}}(\boldsymbol{\theta}, \mathbf{y}), \boldsymbol{\theta}|\mathbf{y})}{\partial \hat{\mathbf{c}}} \frac{\partial \hat{\mathbf{c}}}{\partial \boldsymbol{\theta}}, \tag{2.5}$$

where $dF(\hat{\mathbf{c}}(\boldsymbol{\theta}, \mathbf{y}), \boldsymbol{\theta}|\mathbf{y})/d\boldsymbol{\theta}$ is the total derivative of $F(\hat{\mathbf{c}}(\boldsymbol{\theta}, \mathbf{y}), \boldsymbol{\theta}|\mathbf{y})$ with respect to $\boldsymbol{\theta}$. Notice that the formula of $dF(\hat{\mathbf{c}}(\boldsymbol{\theta}, \mathbf{y}), \boldsymbol{\theta}|\mathbf{y})/d\boldsymbol{\theta}$ involves the term $\partial\hat{\mathbf{c}}/\partial\boldsymbol{\theta}$. If we can find the explicit function $\hat{\mathbf{c}}(\boldsymbol{\theta})$ by optimizing the criterion $H(\mathbf{c}|\boldsymbol{\theta}, \mathbf{y})$ in the first level, it is easy to calculate $\partial\hat{\mathbf{c}}/\partial\boldsymbol{\theta}$. But if not, the Implicit Function Theorem can be applied to find $\partial\hat{\mathbf{c}}/\partial\boldsymbol{\theta}$, which is shown below.

Implicit Function Theorem can be stated as follows. Let $\mathbf{x} = (x_1, \cdots, x_m)$, $\mathbf{y} = (y_1, \cdots, y_n)$, $\mathbf{a} = (a_1, \cdots, a_m)$, $\mathbf{b} = (b_1, \cdots, b_n)$, and $\mathbf{G}(\mathbf{x}, \mathbf{y}) = (G_1(\mathbf{x}, \mathbf{y}), \cdots, G_l(\mathbf{x}, \mathbf{y}))$. If $\mathbf{G}(\mathbf{a}, \mathbf{b}) = 0$ and $\mathbf{G}(\mathbf{x}, \mathbf{y})$ is continuously differentiable on some open disk with center $(\mathbf{a}, \mathbf{b})$ and $|D_{\mathbf{y}}\mathbf{G}(\mathbf{a}, \mathbf{b})| \neq 0$, then there exists an $h > 0$ and a unique function $\boldsymbol{\varphi}(\mathbf{x}) = (\varphi_1(\mathbf{x}), \cdots, \varphi_n(\mathbf{x}))$ defined for $|\mathbf{x} - \mathbf{a}| < h$ such that $\boldsymbol{\varphi}(\mathbf{a}) = \mathbf{b}$ and $\mathbf{G}(\mathbf{x}, \boldsymbol{\varphi}(\mathbf{x})) = 0$ for $|\mathbf{x} - \mathbf{a}| < h$. Moreover, on $|\mathbf{x} - \mathbf{a}| < h$, the function $\boldsymbol{\varphi}(\mathbf{x})$ is continuously differentiable and

$$\frac{d\boldsymbol{\varphi}(\mathbf{x})}{d\mathbf{x}} = -\left[\frac{d\mathbf{G}(\mathbf{x}, \boldsymbol{\varphi}(\mathbf{x}))}{d\mathbf{y}}\right]^{-1} \frac{d\mathbf{G}(\mathbf{x}, \boldsymbol{\varphi}(\mathbf{x}))}{d\mathbf{x}}.$$

Since the optimal local parameter vector $\hat{\mathbf{c}}$ satisfying $\partial H(\mathbf{c}|\boldsymbol{\theta}, \mathbf{y})/\partial\mathbf{c} = 0$, and

$\hat{\mathbf{c}}$ is a function of $\boldsymbol{\theta}$ and $\mathbf{y}$, we can take the $\boldsymbol{\theta}$-derivative on $\partial H(\mathbf{c}|\boldsymbol{\theta},\mathbf{y})/\partial \mathbf{c}|_{\hat{\mathbf{c}}} = 0$ as follows:

$$\frac{d}{d\boldsymbol{\theta}}\left(\frac{\partial H(\mathbf{c}|\boldsymbol{\theta},\mathbf{y})}{\partial \mathbf{c}}\bigg|_{\hat{\mathbf{c}}}\right) = \frac{\partial^2 H(\mathbf{c}|\boldsymbol{\theta},\mathbf{y})}{\partial \mathbf{c}\partial\boldsymbol{\theta}}\bigg|_{\hat{\mathbf{c}}} + \frac{\partial^2 H(\mathbf{c}|\boldsymbol{\theta},\mathbf{y})}{\partial \mathbf{c}^2}\bigg|_{\hat{\mathbf{c}}}\frac{\partial \hat{\mathbf{c}}}{\partial \boldsymbol{\theta}} = 0\,, \tag{2.6}$$

which holds since $\partial H(\mathbf{c}|\boldsymbol{\theta},\mathbf{y})/\partial \mathbf{c}|_{\hat{\mathbf{c}}}$ is a function of $\boldsymbol{\theta}$ that is identically 0. Assuming that $\left|\partial^2 H(\mathbf{c}|\boldsymbol{\theta},\mathbf{y})/\partial \mathbf{c}^2|_{\hat{\mathbf{c}}}\right| \neq 0$, from the Implicit Function Theorem we obtain

$$\frac{\partial \hat{\mathbf{c}}}{\partial \boldsymbol{\theta}} = -\left[\frac{\partial^2 H(\mathbf{c}|\boldsymbol{\theta},\mathbf{y})}{\partial \mathbf{c}^2}\bigg|_{\hat{\mathbf{c}}}\right]^{-1}\left[\frac{\partial^2 H(\mathbf{c}|\boldsymbol{\theta},\mathbf{y})}{\partial \mathbf{c}\partial\boldsymbol{\theta}}\bigg|_{\hat{\mathbf{c}}}\right]. \tag{2.7}$$

2.3 Interval Estimations for Global and Local Parameters

In this section, we derive the variances for global and local parameters with the Delta method. By treating local parameters as functions of global parameters, the variance of local parameters also include the variation coming from the global parameters.

The estimated global parameter vector $\hat{\boldsymbol{\theta}}$ satisfies $dF(\hat{\mathbf{c}}(\boldsymbol{\theta}),\boldsymbol{\theta},\mathbf{y})/d\boldsymbol{\theta} = 0$. By taking the $\mathbf{y}$-derivative on both sides of $dF(\hat{\mathbf{c}}(\boldsymbol{\theta},\mathbf{y}),\boldsymbol{\theta},\mathbf{y})/d\boldsymbol{\theta}|_{\hat{\boldsymbol{\theta}},\mathbf{y}} = 0$, we obtain:

$$\frac{d}{d\mathbf{y}}\left(\frac{dF(\hat{\mathbf{c}}(\boldsymbol{\theta},\mathbf{y}),\boldsymbol{\theta},\mathbf{y})}{d\boldsymbol{\theta}}\bigg|_{\hat{\boldsymbol{\theta}},\mathbf{y}}\right) = \frac{d^2F(\hat{\mathbf{c}}(\boldsymbol{\theta},\mathbf{y}),\boldsymbol{\theta},\mathbf{y})}{d\boldsymbol{\theta}d\mathbf{y}}\bigg|_{\hat{\boldsymbol{\theta}},\mathbf{y}} + \frac{d^2F(\hat{\mathbf{c}}(\boldsymbol{\theta},\mathbf{y}),\boldsymbol{\theta},\mathbf{y})}{d\boldsymbol{\theta}^2}\bigg|_{\hat{\boldsymbol{\theta}},\mathbf{y}}\frac{d\hat{\boldsymbol{\theta}}}{d\mathbf{y}} = 0\,, \tag{2.8}$$

where

$$\frac{d^2F}{d\boldsymbol{\theta}^2} = \frac{\partial^2 F}{\partial \boldsymbol{\theta}^2} + \frac{\partial^2 F}{\partial \hat{\mathbf{c}} \partial \boldsymbol{\theta}} \frac{\partial \hat{\mathbf{c}}}{\partial \boldsymbol{\theta}} + \left(\frac{\partial \hat{\mathbf{c}}}{\partial \boldsymbol{\theta}}\right)' \frac{\partial^2 F}{\partial \hat{\mathbf{c}}^2} \frac{\partial \hat{\mathbf{c}}}{\partial \boldsymbol{\theta}} + \frac{\partial F}{\partial \hat{\mathbf{c}}} \frac{\partial^2 \hat{\mathbf{c}}}{\partial \boldsymbol{\theta}^2}, \tag{2.9}$$

and

$$\frac{d^2F}{d\boldsymbol{\theta} d\mathbf{y}} = \frac{\partial^2 F}{\partial \boldsymbol{\theta} \partial \mathbf{y}} + \frac{\partial^2 F}{\partial \hat{\mathbf{c}} \partial \mathbf{y}} \frac{\partial \hat{\mathbf{c}}}{\partial \boldsymbol{\theta}} + \frac{\partial^2 F}{\partial \boldsymbol{\theta} \partial \hat{\mathbf{c}}} \frac{\partial \hat{\mathbf{c}}}{\partial \mathbf{y}} + \frac{\partial^2 F}{\partial \hat{\mathbf{c}}^2} \frac{\partial \hat{\mathbf{c}}}{\partial \mathbf{y}} \frac{\partial \hat{\mathbf{c}}}{\partial \boldsymbol{\theta}} + \frac{\partial F}{\partial \hat{\mathbf{c}}} \frac{\partial^2 \hat{\mathbf{c}}}{\partial \boldsymbol{\theta} \partial \mathbf{y}}. \tag{2.10}$$

Equations (2.8) holds since $\partial F/\partial \boldsymbol{\theta}|_{\hat{\boldsymbol{\theta}},\mathbf{y}}$ is a function of $\mathbf{y}$ that is identically 0. The formulas (2.9) and (2.10) for $d^2F/d\boldsymbol{\theta}^2$ and $d^2F/d\boldsymbol{\theta} d\mathbf{y}$ involve the terms $\partial \hat{\mathbf{c}}/\partial \mathbf{y}$, $\partial^2 \hat{\mathbf{c}}/\partial \boldsymbol{\theta}^2$ and $\partial^2 \hat{\mathbf{c}}/\partial \boldsymbol{\theta} \partial \mathbf{y}$. The calculations for them are given in Appendix A.

Solving Equation (2.8), we get the first derivative of $\hat{\boldsymbol{\theta}}$ with respect to $\mathbf{y}$:

$$\frac{d\hat{\boldsymbol{\theta}}}{d\mathbf{y}} = -\left[\frac{d^2F(\hat{\mathbf{c}}(\boldsymbol{\theta},\mathbf{y}),\boldsymbol{\theta},\mathbf{y})}{d\boldsymbol{\theta}^2}\bigg|_{\hat{\boldsymbol{\theta}},\mathbf{y}}\right]^{-1} \left[\frac{d^2F(\hat{\mathbf{c}}(\boldsymbol{\theta},\mathbf{y}),\boldsymbol{\theta},\mathbf{y})}{d\boldsymbol{\theta} d\mathbf{y}}\bigg|_{\hat{\boldsymbol{\theta}},\mathbf{y}}\right]. \tag{2.11}$$

Let $\boldsymbol{\mu} = \mathrm{E}(\mathbf{y})$, then using the first order Taylor expansion, we have

$$\hat{\boldsymbol{\theta}}(\mathbf{y}) \approx \hat{\boldsymbol{\theta}}(\boldsymbol{\mu}) + \frac{d\hat{\boldsymbol{\theta}}}{d\boldsymbol{\mu}}(\mathbf{y} - \boldsymbol{\mu}).$$

Consequently, the variance of $\hat{\boldsymbol{\theta}}(\mathbf{y})$ can be estimated by

$$\mathrm{Var}[\hat{\boldsymbol{\theta}}(\mathbf{y})] \approx \left[\frac{d\hat{\boldsymbol{\theta}}}{d\boldsymbol{\mu}}\right] \Sigma \left[\frac{d\hat{\boldsymbol{\theta}}}{d\boldsymbol{\mu}}\right]' \tag{2.12}$$

$$\approx \left[\frac{d\hat{\boldsymbol{\theta}}}{d\mathbf{y}}\right] \Sigma \left[\frac{d\hat{\boldsymbol{\theta}}}{d\mathbf{y}}\right]', \tag{2.13}$$

where Σ is the variance-covariance matrix for $\mathbf{y}$ and it can be estimated by:

$$\hat{\Sigma} = \frac{\mathtt{SSE}(\hat{\boldsymbol{\theta}})}{\mathtt{dfe}(\hat{\boldsymbol{\theta}})} \cdot \mathbf{I}. \tag{2.14}$$

Approximation (2.13) makes sense since

$$\mathtt{E}\left(\frac{d\hat{\boldsymbol{\theta}}}{d\boldsymbol{\mu}}\right) \approx \mathtt{E}\left(\frac{d\hat{\boldsymbol{\theta}}}{d\mathbf{y}}\right), \tag{2.15}$$

when $d^2\hat{\boldsymbol{\theta}}/d^2\boldsymbol{\mu}$ are bounded by a fixed number. Approximation (2.15) can be derived by taking expectation on both sides of the first order Taylor expansion for $d\hat{\boldsymbol{\theta}}/d\mathbf{y}$:

$$\frac{d\hat{\boldsymbol{\theta}}}{d\mathbf{y}} \approx \frac{d\hat{\boldsymbol{\theta}}}{d\boldsymbol{\mu}} + \frac{d^2\hat{\boldsymbol{\theta}}}{d^2\boldsymbol{\mu}}(\mathbf{y} - \boldsymbol{\mu}) \tag{2.16}$$

Similarly, the sampling variance of $\hat{\mathbf{c}}(\hat{\boldsymbol{\theta}}(\mathbf{y}), \mathbf{y})$ is attained by

$$\mathtt{Var}[\hat{\mathbf{c}}(\hat{\boldsymbol{\theta}}(\mathbf{y}), \mathbf{y})] \approx \left[\frac{d\hat{\mathbf{c}}}{d\mathbf{y}}\right] \Sigma \left[\frac{d\hat{\mathbf{c}}}{d\mathbf{y}}\right]', \tag{2.17}$$

where

$$\frac{d\hat{\mathbf{c}}}{d\mathbf{y}} = \frac{\partial\hat{\mathbf{c}}}{\partial\hat{\boldsymbol{\theta}}}\frac{d\hat{\boldsymbol{\theta}}}{d\mathbf{y}} + \frac{\partial\hat{\mathbf{c}}}{\partial\mathbf{y}}. \tag{2.18}$$

The method used to estimate the sampling variance of $\hat{\boldsymbol{\theta}}$ and $\hat{\mathbf{c}}$ is called *the Delta method* in this thesis. This method is also used elsewhere in this thesis to estimate the sampling variances of parameters. This definition is slightly different

from that given by Casella and Berger (1990), which is a generalization of the Central Limit Theorem.

If we don't consider the functional relationship between $\hat{\mathbf{c}}$ and $\hat{\boldsymbol{\theta}}$, the sampling variance of $\hat{\mathbf{c}}(\hat{\boldsymbol{\theta}}(\mathbf{y})|\mathbf{y})$ will then be underestimated by replacing the full derivative of $\hat{\mathbf{c}}$ with respect to $\mathbf{y}$ by the partial derivative of $\hat{\mathbf{c}}$ with respect to $\mathbf{y}$:

$$\mathtt{Var}[\hat{\mathbf{c}}|\hat{\boldsymbol{\theta}}, \mathbf{y}] \approx \left[\frac{\partial \hat{\mathbf{c}}}{\partial \mathbf{y}}\right] \boldsymbol{\Sigma} \left[\frac{\partial \hat{\mathbf{c}}}{\partial \mathbf{y}}\right]' . \tag{2.19}$$

We call $\mathtt{Var}[\hat{\mathbf{c}}|\hat{\boldsymbol{\theta}}, \mathbf{y}]$ *the conditional sampling variance* for $\hat{\mathbf{c}}$, because it ignores the uncertainty from the estimate $\hat{\boldsymbol{\theta}}$.

2.4 Introduction to Adaptive Penalized Smoothing

Quite often the underlying function $x(t)$ shows different scales of variation in different regions. In some regions, $x(t)$ may be almost linear, and thus we would require a very smooth fitting function and would use a large value of smoothing parameter λ for penalized smoothing. On the other hand, $x(t)$ may have sharp variations in other regions, and a more variable fitting function would be required, and λ would have to be small. When we penalize smooth data with a constant smoothing parameter λ estimated by optimizing GCV or other criteria, the fitting function is often found to undersmooth in the regions with low variations.

Hence, we express λ as a function of t, so that data are smoothed with

different scales of penalty in different regions, adaptive to the geometry property of the underlying curve. This process is called *adaptive penalized smoothing*, and $\lambda(t)$ is called the functional smoothing parameter (FSP). In contrast, when $\lambda(t)$ is a constant function, the process is then called *nonadaptive penalized smoothing*. In order to ensure a positive penalty term, the FSP $\lambda(t)$ is expressed as the exponential function of $\omega(t)$, written as a linear expansion of K_ω number of basis functions:

$$\lambda(t) = \exp[\omega(t)], \quad \text{where} \quad \omega(t) = \sum_{\ell}^{K_\omega} \theta_\ell \psi_\ell(t) = \boldsymbol{\theta}' \boldsymbol{\psi}(t), \tag{2.20}$$

where $\boldsymbol{\theta}$ is a vector of the FSP coefficients, and $\boldsymbol{\psi}(t)$ is a vector of the FSP basis functions. In the following, the notations have the same meanings as Section 2.1 if they are not mentioned. The fitting criterion for the adaptive penalized smoothing is written as follows:

$$H(\mathbf{c}|\lambda, \mathbf{y}) = \sum_{i=1}^{n} w_i [y_i - x(t_i)]^2 + \int \lambda(t) [Lx(t)]^2 \, dt. \tag{2.21}$$

By minimizing $H(\mathbf{c}|\lambda, \mathbf{y})$, we obtain the analytical expression for the optimal coefficient vector $\hat{\mathbf{c}}$ as an explicit function of λ and $\mathbf{y}$:

$$\hat{\mathbf{c}}(\lambda, \mathbf{y}) = [\boldsymbol{\Phi}' \mathbf{W} \boldsymbol{\Phi} + \mathbf{R}]^{-1} \boldsymbol{\Phi}' \mathbf{W} \mathbf{y}, \tag{2.22}$$

where order K_c matrix $\mathbf{R} = \int \lambda(t) [L\boldsymbol{\phi}(t)][L\boldsymbol{\phi}(t)]' dt$.

For the adaptive penalized smoothing, the coefficient vector $\mathbf{c}$ is the local parameter, and the FSP coefficient vector $\boldsymbol{\theta}$ is the complexity parameter. The complexity parameter space is now of dimension K_ω. The outer optimization cri-

terion is GCV, and we obtain the optimized FSP coefficient vector $\hat{\boldsymbol{\theta}}$ by minimizing GCV with respect to $\boldsymbol{\theta}$. The explicit expressions for gradient and Hessian matrix are given as (A.6) and (A.7) in Appendix A.

The sample variance of the FSP coefficient vector $\boldsymbol{\theta}$ is obtained by the Delta method:

$$\mathtt{Var}[\boldsymbol{\theta}(\mathbf{y})] \approx \left[\frac{d\boldsymbol{\theta}}{d\mathbf{y}}\right] \boldsymbol{\Sigma} \left[\frac{d\boldsymbol{\theta}}{d\mathbf{y}}\right]',$$

where $\boldsymbol{\Sigma}$ is the residual variance-covariance matrix. As shown in (2.11), if $F(\boldsymbol{\theta}|\mathbf{y})$ is the optimization criterion in the second level, $d\boldsymbol{\theta}/d\mathbf{y}$ requires the calculations of $d^2F/d^2\boldsymbol{\theta}$ and $d^2F/d\boldsymbol{\theta}d\mathbf{y}$. Specially, for adaptive penalized smoothing, GCV is the optimization criterion in the second level. Appendix A gives the analytic expressions for $d^2F/d^2\boldsymbol{\theta}$ and $d^2F/d\boldsymbol{\theta}d\mathbf{y}$ in (A.7) and (A.8). The Delta method is also applied to estimate the sampling variances of FSP and the fitting function, which are given in (A.9) and (A.10) in Appendix A.

2.5 Results for Adaptive Penalized Smoothing by Simulation

In this section, based on simulated data, we first compare the adaptive and non-adaptive penalized smoothing, and then explore the effects of data noise, data resolution and basis systems on the adaptive penalized smoothing. We also verify our estimates for variances of functional smoothing parameters and fitting func-

tions for adaptive penalized smoothing by simulation.

The simulated data are generated by adding Gaussian noise with a standard deviation (SD) of 5.0 to the function

$$\mu(t) = t^2/2 + 50\exp(-t^2/2) \tag{2.23}$$

over the interval $[-10, 10]$ (Figure 2.1). It is a good example for applying adaptive penalized smoothing since the curvature magnitude is 1.0 over most of the interval except over $[-3, 3]$ where it reaches 50.0. Results are reported for $\omega(t) = \ln(\lambda(t))$ defined as a constant and as a cubic B-spline basis expansion with 5 basis functions, which are defined by putting three knots at $(-10, 0, 10)$.

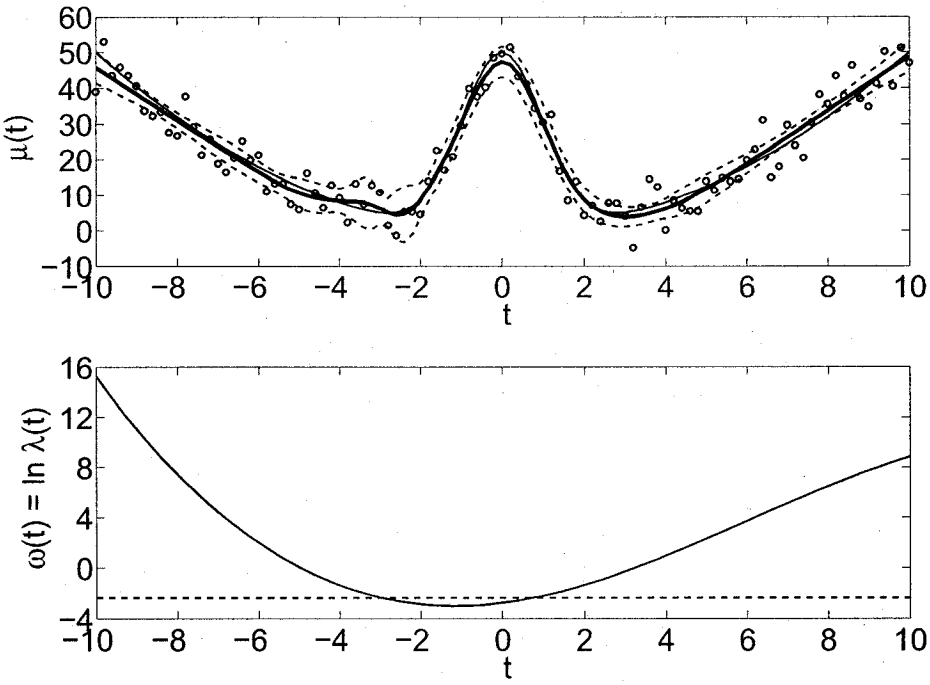

Figure 2.1: The top panel displays simulated data (circles) generated by adding Gaussian noise (SD = 5) to the proposed function $\mu(t) = t^2/2 + 50\exp(-t^2/2)$ with 101 equally spaced points. The heavy and thin solid lines are the adaptively estimated fitting function and the true curve, respectively, and the dashed lines are estimated 95% pointwise confidence bands for the estimated curve. The bottom panel contains estimates for $\omega(t) = \ln\lambda(t)$. The solid curve is defined by 5 cubic B-spline basis, and the dashed straight line is the estimate for constant ω.

The top panel in Figure 2.1 shows that the adaptive fitting function can well estimate the true function over all the region. The bottom panel shows that the

values of $\lambda(t)$ adapt to the curvature of the true function, by ranging from its lowest value of about 0.05 in the middle to 4×10^6 on the left boundary.

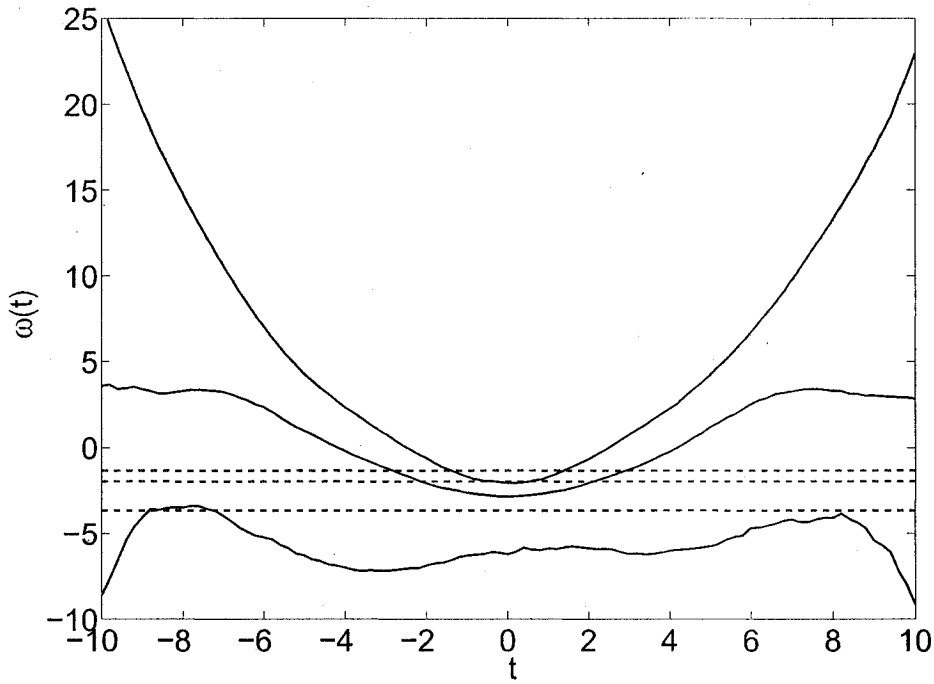

Figure 2.2: The solid lines are the 2.5%, 50%, and 97.5% pointwise quantiles of the estimated $\omega(t) = \ln \lambda(t)$ over 1000 simulated data sets, and the dashed lines represent the corresponding values when $\omega(t) = \ln \lambda(t)$ is a constant.

Figure 2.2 displays the empirical median and 95% confidence limits for estimates of $\omega(t)$ taken over 1000 simulation datasets for nonadaptive and adaptive penalized smoothing. The median constant function $\omega(t)$ is slightly larger in the region with large curvature and much smaller in the region with small cur-

vature, which means that the non-adaptive fitting functions are comparatively under-smoothed in the region with small curvature. However, the wide confidence limits on $\omega(t)$ indicate that the estimates of $\omega(t)$ are not stable.

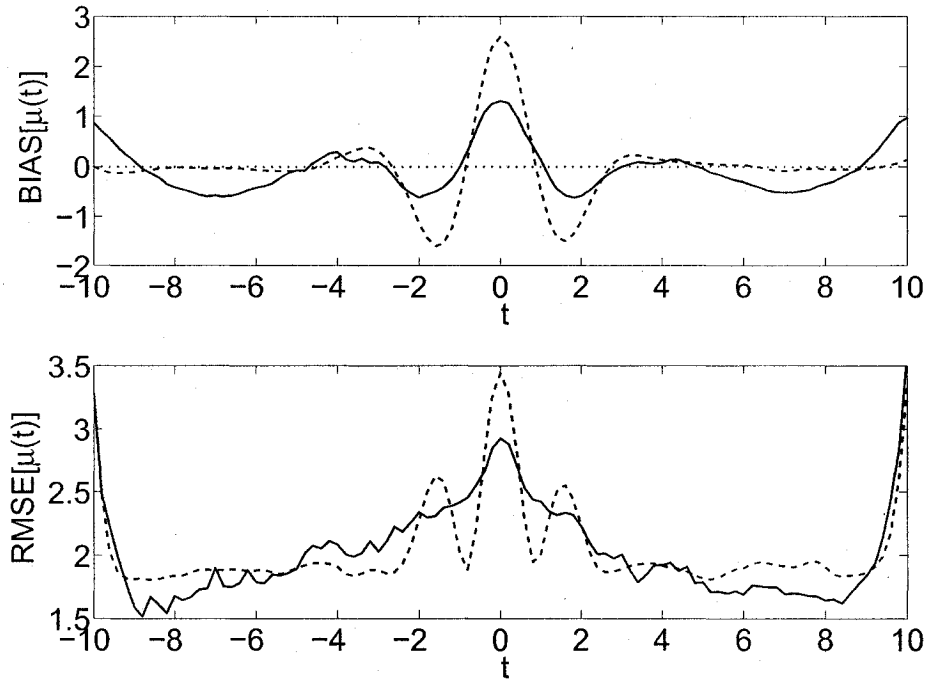

Figure 2.3: The solid line indicates the mean bias and RMSE of adaptive fitting functions over 1000 simulated data sets when $\omega(t)$ is expanded by 5 cubic spline basis. The corresponding results for non-adaptive penalized smoothing are shown as dashed lines.

Figure 2.3 displays the mean bias and root mean squared error (RMSE) for the fitting functions $x(t)$ estimated over 1000 simulations. The bias is much smaller

for the adaptive penalized smoothing in the region [-2, 2] where the curvature of $\mu(t)$ is large, but the bias for the adaptive penalized smoothing is larger at $t = \pm 7$ because the limited information available in these data leads to the unstable estimate of $\omega(t)$ and sometimes over-smoothing the data.

To investigate the effect of data noise, data resolution and flexibility of FSP $\lambda(t)$ on the adaptive penalized smoothing, we do 4 contrastive simulation experiments independently for 1000 times each. Data are simulated by adding Gaussian noise with a specified SD (shown in Table 2.2) to n equally spaced points in the proposed function $\mu(t) = t^2/2 + 50\exp(-t^2/2)$ over the interval [-10, 10]. The functional smoothing parameter $\omega(t) = \ln \lambda(t)$ is expanded by K_ω cubic B-splines with interior knots shown in Table 2.2.

Table 2.2: Settings for 4 contrastive simulation experiments in adaptive penalized smoothing. Data are simulated by adding Gaussian noise with a specified SD to n equally spaced points in the proposed function $\mu(t) = t^2/2+50\exp(-t^2/2)$ over the interval [-10, 10]. The functional smoothing parameter $\omega(t) = \ln \lambda(t)$ is expanded by K_ω cubic B-splines with the specified interior knots.

Setting	SD	n	Interior Knots	K_ω
Setting 1	5	101	0	5
Setting 2	10	101	0	5
Setting 3	5	51	0	5
Setting 4	5	101	-5,0,5	7

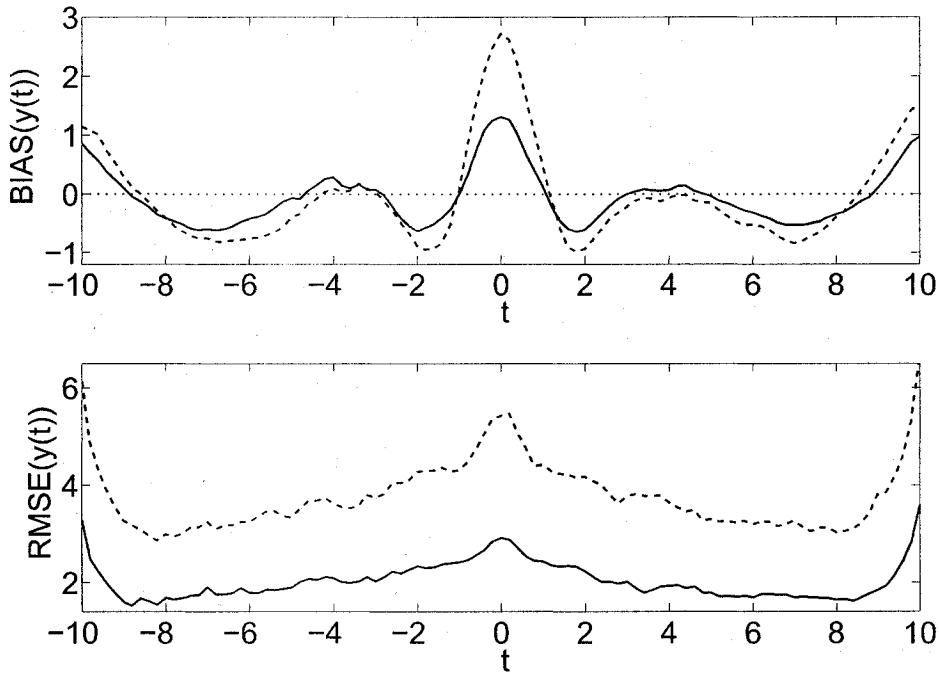

Figure 2.4: The bias and RMSE of adaptive fitting functions contrasting $SD = 5$ under Setting 1 (solid lines) and $SD = 10$ under Setting 2 (dashed lines).

Comparing the results under Setting 1 and Setting 2, we can find the effect of data noise on the adaptive penalized smoothing. Figure 2.4 shows that the bias and RMSE of adaptive fitting functions is smaller for simulated data with smaller noise, as we expect.

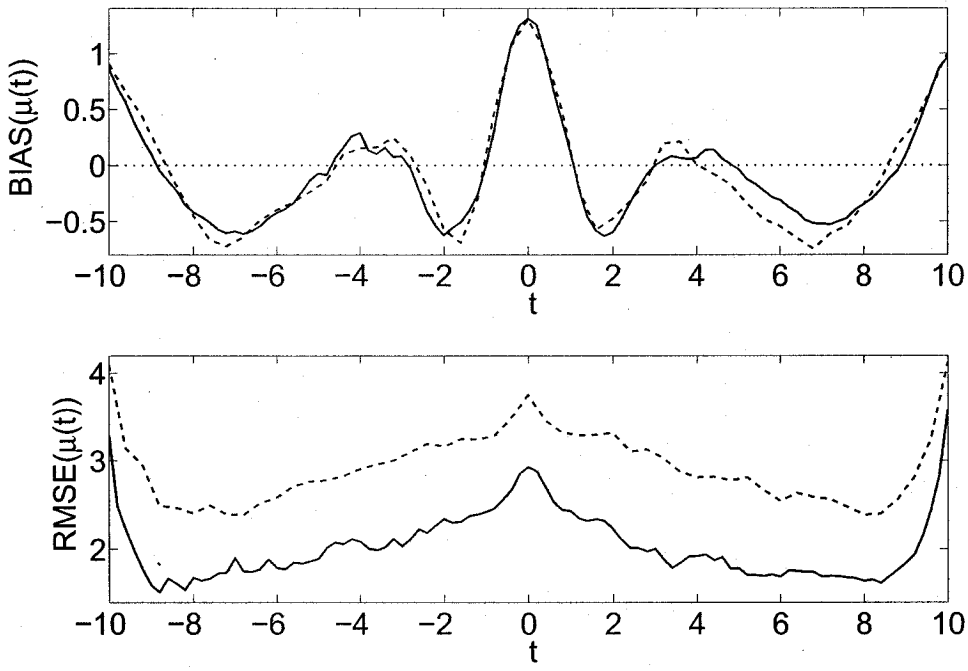

Figure 2.5: The bias and RMSE of adaptive fitting functions contrasting $n = 101$ under Setting 1 (solid lines) and $n = 51$ under Setting 3 (dashed lines).

The data resolution effect on the adaptive penalized smoothing can be investigated by comparing the results under Setting 1 and Setting 3. Figure 2.5 shows that RMSE of adaptive fitting functions becomes larger for sparse simulated data. But the bias of adaptive fitting functions is little affected by the data resolution.

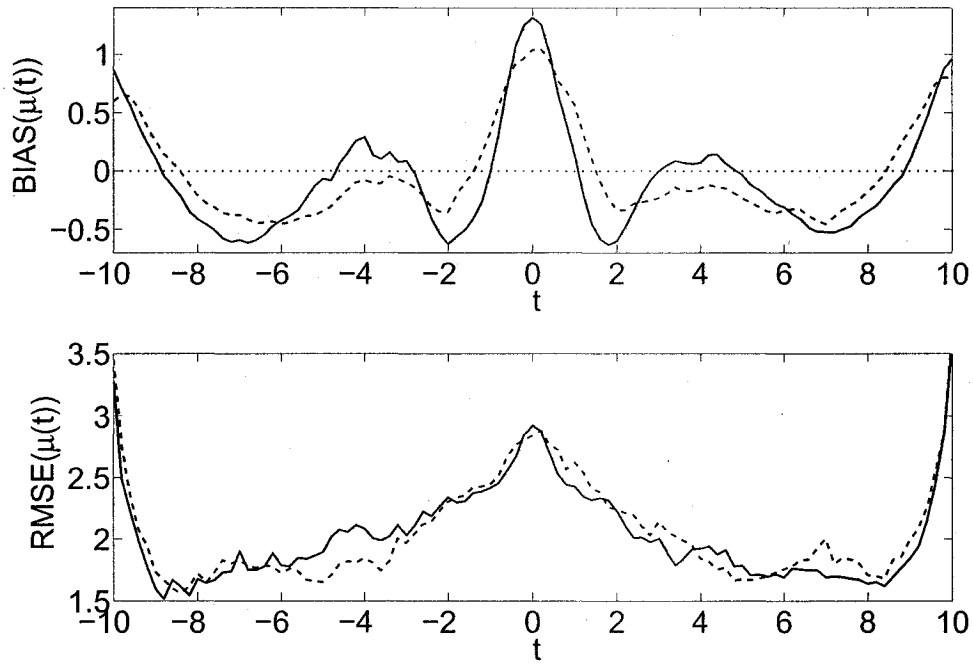

Figure 2.6: The bias and RMSE of adaptive fitting functions contrasting $K_\omega = 5$ under Setting 1 (solid lines) and $K_\omega = 7$ under Setting 4 (dashed lines).

Comparing the results under Setting 1 and Setting 4, we can find the effect of FSP variability on the adaptive penalized smoothing. Figure 2.6 shows that bias of adaptive fitting functions becomes smaller in region [-2, 2] with large curvature when the basis system has more flexibility, but larger in both sides because the information from the data is not enough to obtain a stable estimate for FSP. RMSE of adaptive fitting functions is slightly larger in most of the region when the basis system has more flexibility, which is also caused by the instability of the FSP

estimates.

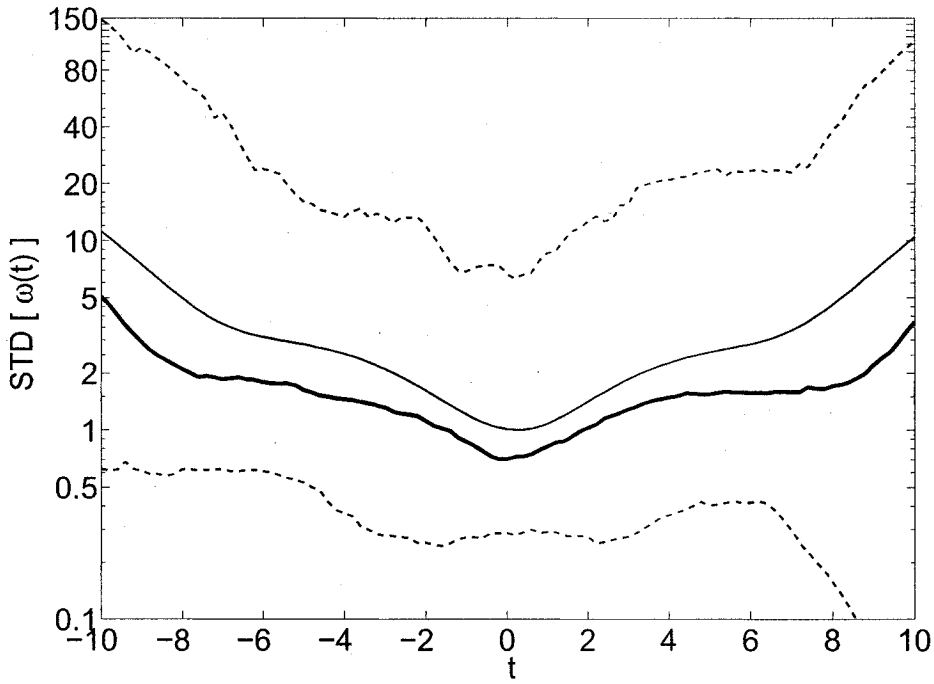

Figure 2.7: The heavy solid line is the median of the estimated standard deviations of $\omega(t)$ over 1000 simulated data sets, and the thin solid line is the empirical standard deviation of the estimates. The pointwise 95% confidence band for the estimated standard deviations of $\omega(t)$ is shown by the dashed lines. The y-axis is in log scale.

In the following, we estimate the standard deviations for the functional smoothing parameter and fitting functions from the simulated data generated under Setting 1 in Table 2.2. The standard deviation $\sigma_\omega(t)$ of the optimal smoothing

function $\hat{\omega}(t) = \ln \hat{\lambda}(t)$ is estimated by Equation (A.9) in the Appendix. Figure 2.7 shows the empirical SD is well within the pointwise 95% confidence band for the estimated $\hat{\sigma}_\omega(t)$ through the range, but the estimate is about 70% too low near $t = 0$.

Figure 2.8 shows the estimated standard deviation of the fitting function. The estimate is also satisfactory, although about 92% of the empirical value at $t = 0$. The usual practice of estimating the SD of the fitting function $\mu(t)$ conditioned on the estimated value of $\lambda(t)$ underestimates the SD of the fitting function more severely, about 70% at $t = 0$, as we explain before. We can also see the large gap between the empirical median and the 97.5% quantile, which means that minimizing the GCV criterion can sometimes give very bad estimates (Gu 2002).

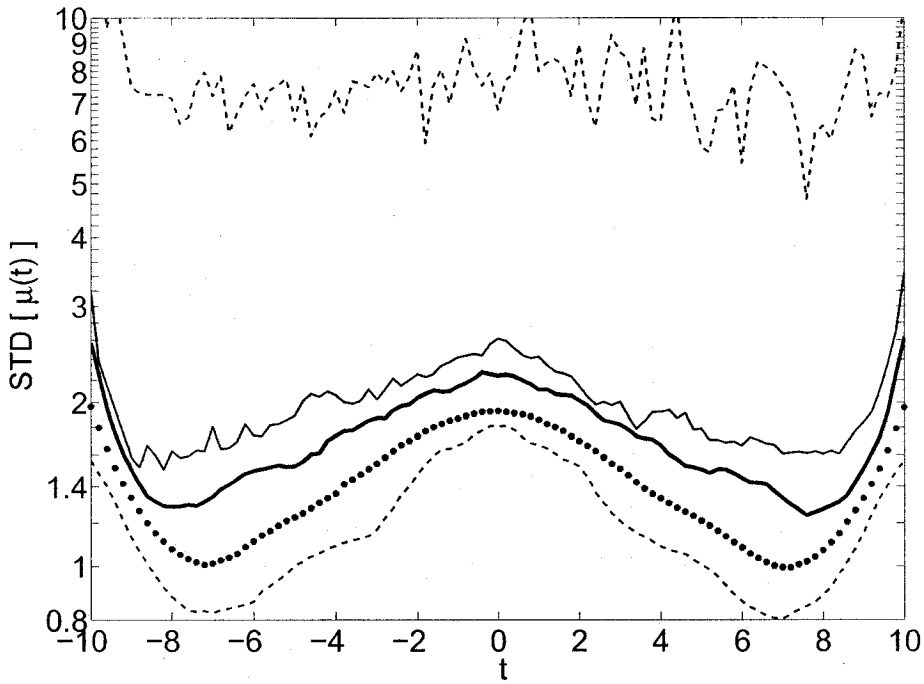

Figure 2.8: The heavy solid line is the median estimate of the standard deviation of $x(t)$, and the thin solid line is the experimental standard deviation of $x(t)$ computed over 1000 simulated samples. The 95% confidence pointwise confidence band for the estimate of the standard deviation of $x(t)$ is shown by the dashed lines. The dotted line is the median conditional estimate of the standard deviation of $x(t)$ that does not take into account the uncertainty in the estimate of $\lambda(t)$. The y-axis is in log scale.

2.6 Adaptive Penalized Smoothing the Titanium Heat Data

The top panel of Figure 2.9 shows measurements of a property g of titanium changing with the temperature from 595 °C to 1075 °C, adapted from de Boor (2001). The measurement errors are small but not negligible. Because of the sharp peak, this data has become a standard challenge and has been used extensively as a problem in nonparametric smoothing. It is appropriate to apply adaptive penalized smoothing to these data because of their different scale of variation over the region. The bottom panel of Figure 2.9 shows the logarithm of the functional smoothing parameter $\omega(t) = \ln \lambda(t)$. It is large in [575,850] and [1050,1075], where the underlying curve is almost a straight line with larger errors, and small in [850, 1050], where the underlying curve has a large curvature in [850, 950] and the observations have less errors in [950, 1050]. The constant λ is much larger in the regions with large variation, and thus the nonadaptive fitting function is oversmoothed there.

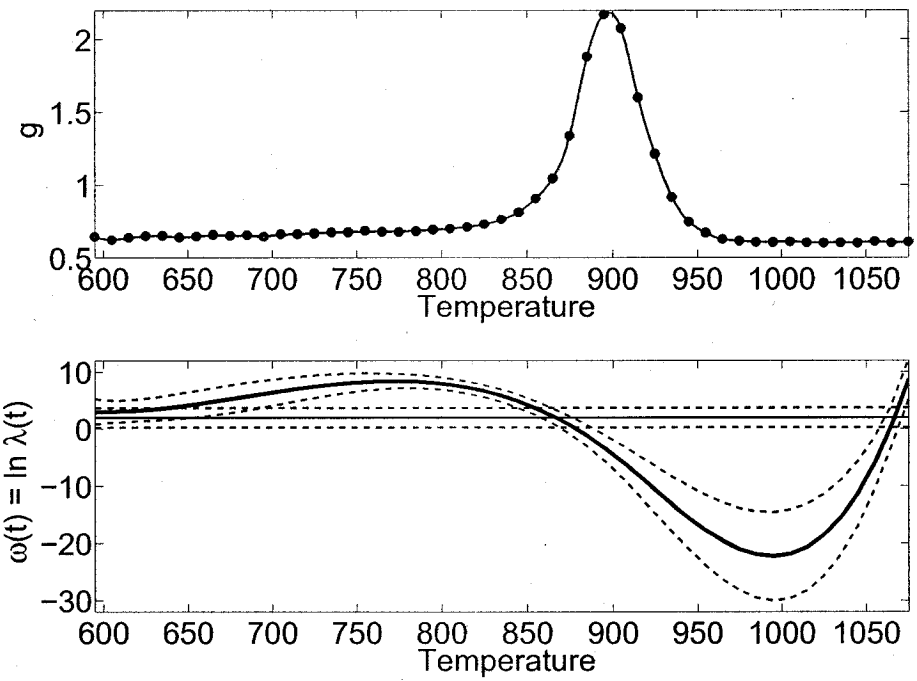

Figure 2.9: Top panel: The titanium heat data are smoothed by cubic B-splines defined by putting one knot at each observation using adaptive penalized smoothing. The dots are observations, and the solid line is the adaptive penalized fitting function. Bottom panel: The optimal $\omega(t) = \ln \lambda(t)$ by minimizing GCV when it is a constant (thin solid line) or expanded by 5 cubic B-splines with a single interior knot at 900 (heavy solid line). The dashed curves define their 95% pointwise confidence bands.

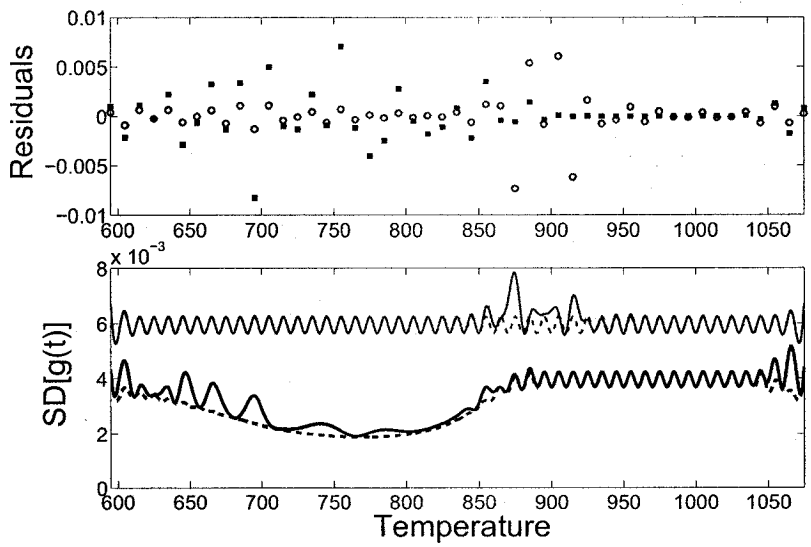

Figure 2.10: Top panel: The residuals of the smoothing splines when $\omega(t) = \ln \lambda(t)$ is a constant (circles) or expanded by 5 cubic B-splines with the interior knot on 800 (square dots); Bottom panel: The unconditionally estimated standard deviations of the smoothing splines $\hat{g}(t)$ when $\omega(t) = \ln \lambda(t)$ is a constant (thin solid line) or expanded by 5 cubic B-splines with the interior knot on 900 (heavy solid line). The dashed lines are the corresponding conditional estimates.

The estimated standard deviations of data are $4.2 * 10^{-3}$ in adaptive penalized smoothing and $6.8 * 10^{-3}$ in nonadaptive penalized smoothing. The top panel of Figure 2.10 shows the residuals for both non-adaptive and adaptive penalized smoothing. The non-adaptive penalized smoothing over-fits the data in the flat regions, and over-smooths the data in the region with large curvature, as we expect from $\omega(t)$ shown in Figure 2.9. The lower panel shows that the uncondi-

tionally estimated pointwise standard deviations of the adaptive fitting functions are substantially smaller than those for the nonadaptive fitting functions, and the corresponding conditionally estimates underestimate the pointwise standard deviations of the fitting functions.

2.7 Adaptive Penalized Smoothing Growth Curves

It is important to study human growth, and to understand how the body regulates its own growth, but it is exceedingly expensive to collect growth data over the entire growing period (Ramsay and Silverman 2005). Children must be brought into the laboratory at preassigned ages over about twenty years, requiring the long-term commitment of maintaining a growth laboratory and great dedication and persistence on the part of parents. The dropout rate is understandably high. Considerable training is also required to measure height accurately. Height also depends on many factors. For example, the spine compression causes height to diminish throughout the day. Infants must be measured lying down, and the measurements of their standing height shrink by about one centimeter. Fels Institute in Ohio has been collecting growth data since 1929, and is now measuring the third generation for some of its original cases (Roche 1991).

Much research has been done on the growth data analysis. The classic approach is to develop the parametric models to capture the growth features. For instance, Jolicoeur et al. (1992) proposed a parametric growth curve in the follow-

ing form:

$$h(t) = a\frac{\sum_{l=1}^{3}[b_l(t+e)]^{c_l}}{1+\sum_{l=1}^{3}[b_l(t+e)]^{c_l}}. \tag{2.24}$$

Bock and Thissen (1980) fitted Jolicoeur's model to the Fels growth data (Roche 1991) by estimating the eight parameters a, b_1, b_2, b_3, c_1, c_2, c_3 and e. Then variations of parameter estimates can be summarized by a multivariate normal distribution with mean and SD given in Table 2.3. The SD of measurement errors has also been estimated from the Fels growth data, displayed in Figure 2.11. We can see that the SD's of measurement errors are different throughout the growth period. The standard deviation is around 7 millimeters during infancy and about 5 millimeters after age six.

Table 2.3: Parameter estimates for Jolicoeur's growth model

Parameters	a	b_1	b_2	b_3	c_1	c_2	c_3	e
Mean	164.7	0.31	0.11	0.08	0.73	3.68	16.67	1.47
SD	5.9	0.04	0.0078	0.0058	0.059	0.22	0.74	0.32

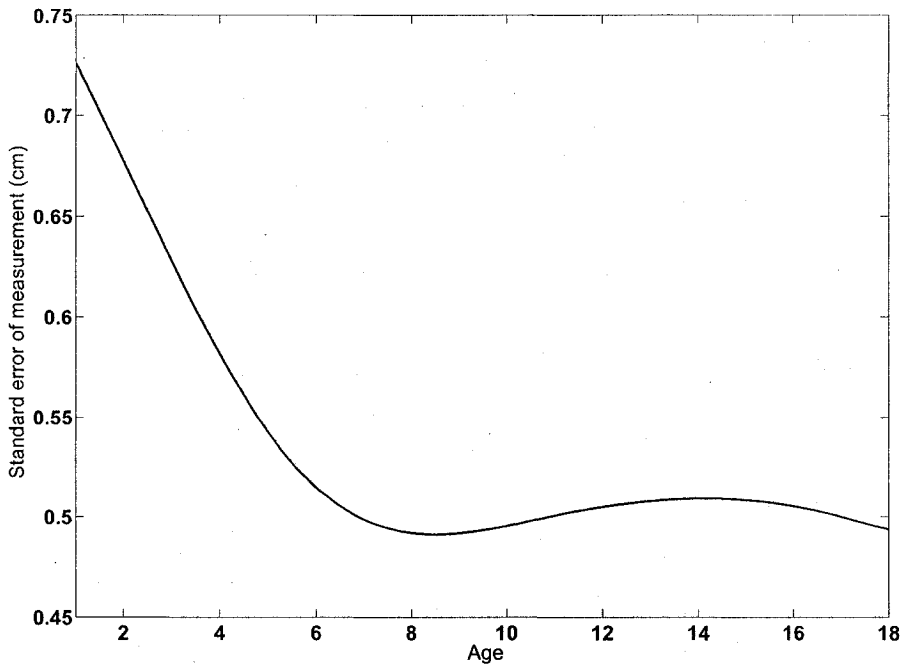

Figure 2.11: The SD's of measurement errors in height as a function of age.

Nonparametric smoothing methods have been applied to growth data, and have successfully detected new features missed by parametric models (Ramsay and Silverman 2005). The main interest in nonparametric smoothing of growth curves is to obtain good estimates for second derivatives of growth curves. In the following, we use adaptive penalized smoothing to estimate the second derivatives of the growth curves on the simulated data.

One thousand simulated vectors of the eight parameters values in Jolicoeur's

model (2.24) are sampled from the multivariate normal distribution of the eight parameters, taking their correlation into consideration. Then one thousand growth curves μ_i, $i = 1, \cdots, 1000$, are generated from Jolicoeur's model (2.24) with the simulated vectors of parameter values. The observations are attained by adding the Gaussian noise with nonconstant SD displayed in Figure 2.11 to the simulated growth curves. The sampling ages are the same as the Berkeley growth data (Tuddenham and Snyder 1954), four measurements between one and two years, one measurement between two and eight years, and biannually after that until eighteen years old. Order 6 B-splines are used as the basis functions to approximate the growth curves with one knot on each observation. We choose order six B-splines because the estimated second derivatives of the growth curves would be cubic splines, which are smooth enough with continuous second derivatives. The weight matrix $\mathbf{W}$ is diagonal with the diagonal entries being the reciprocals of the squares of the measurement error SD's shown in Figure 2.11. The functional parameter $\omega(t) = \ln(\lambda(t))$ is expanded by two distinct cubic splines with 3 and 7 equally spaced knots, respectively.

Figure 2.12 displays a typical result for adaptive penalized smoothing growth curves. Non-adaptive and adaptive fitting functions both approximate the true growth curve well, but non-adaptive penalized smoothing gives oscillated estimates for the first and second derivatives of the growth curves.

The quantiles of estimated functional smoothing parameters are shown in Figure 2.13. The estimated functional smoothing parameters are small at around 12, and large at both sides. This makes sense since the growth curves have a large curvature at around 12 and are very smooth at other ages. However, the wide

experimental pointwise 95% confidence interval also indicates that there are not enough information to obtain stable estimates for functional smoothing parameters $\lambda(t)$.

Figure 2.14 shows the bias and RMSE of estimates for second derivatives of growth curves conditional on each individual, for instance, the bias is defined as $\mathtt{E}(\hat{\ddot{\mu}}_i - \ddot{\mu}_i)$. The estimates for second derivatives of growth curves have very small bias by applying adaptive penalized smoothing, which are similar to nonadaptive penalized smoothing. However, RMSE of estimates for second derivatives of growth curves decreases by 30% if applying adaptive penalized smoothing instead of non-adaptive penalized smoothing over the region [10, 15]. This is the region where human growth becomes slow and then stops, and second derivatives of growth curves have a large curvature. Since the main interest in nonparametric smoothing of growth curves is to obtain good estimates for second derivatives of growth curves, adaptive penalized smoothing wins in this sense.

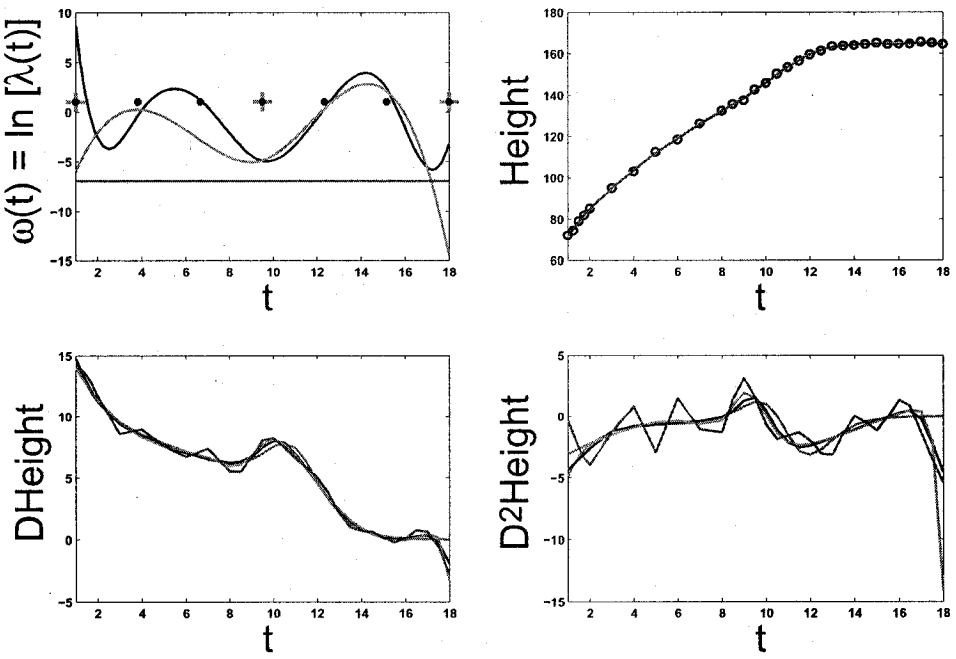

Figure 2.12: The red and black curves correspond to adaptive penalized smoothing when $\omega(t)$ are expanded by cubic B-splines with 3 and 7 equally spaced knots, indicated by red cross and black dots, respectively. The blue curves correspond to the non-adaptive penalized smoothing. The green curves are the true simulated growth curves and the derivatives.

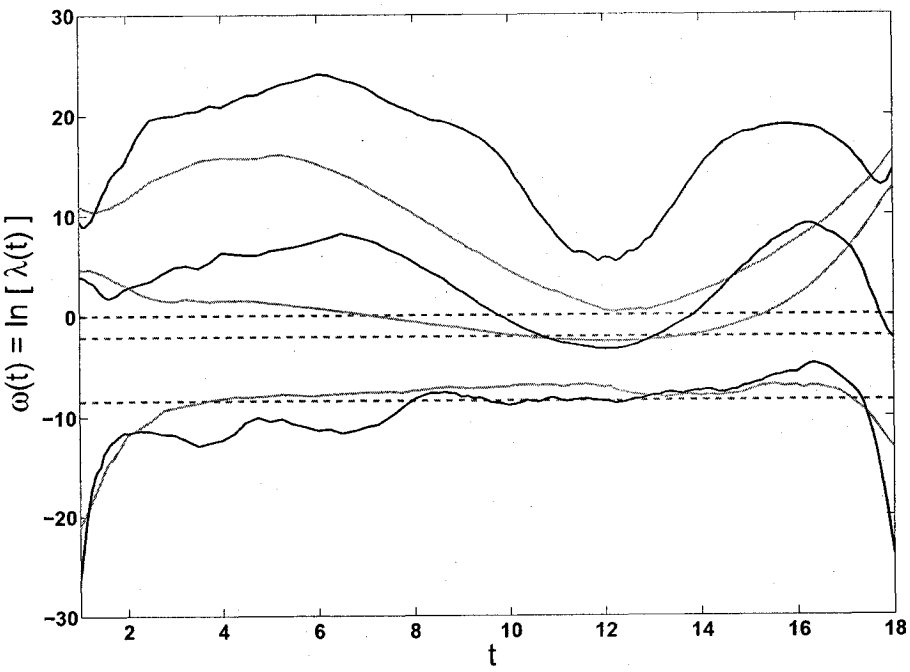

Figure 2.13: The 2.5%, 50%, and 97.5% quartiles of the estimated smoothing functions in 1000 experiments. The red and black curves correspond to adaptive penalized smoothing when $\omega(t)$ are expanded by cubic splines with 3 and 7 equally spaced knots, respectively. The blue curves correspond to the non-adaptive penalized smoothing.

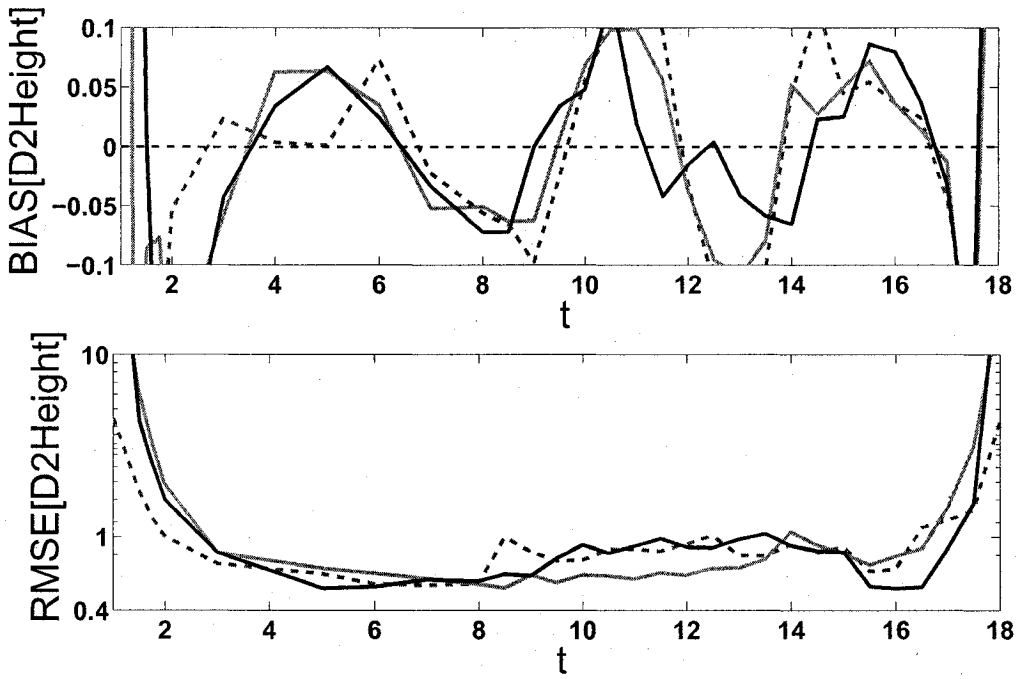

Figure 2.14: The bias and RMSE of the estimates for second derivatives of growth curves. The red and black solid curves correspond to adaptive penalized smoothing when $\omega(t)$ are expanded by cubic splines with 3 and 7 equally spaced knots, respectively. The blue dashed curves correspond to the non-adaptive penalized smoothing.

Chapter 3

Estimating the Generalized Semiparametric Additive Model

3.1 Literature Review on the Generalized Semiparametric Additive Model

Longitudinal data are repeated observations over time or space. Functional data are longitudinal data with the medium or high resolution. Many parametric models and statistical methods have been proposed to analyze longitudinal data (Diggle et al. 2002), which can provide the explanatory relationship between the response variable and the covariates. But they can sometimes be misspecified and fit the data poorly. On the other hand, it is hard to explain the exact relationship between the response variable and the covariates based on completely nonparametric

models. As a trade-off between parametric and nonparametric models, semiparametric additive models keep the flexibility of nonparametric models on confounding variables and explanatory parametric form on variables of interest.

Assuming functional data $\{y_j\}_{j=1}^n$ to be distributed with mean $\mu_j = E(y_j)$, we can write the generalized semiparametric additive model as follows:

$$\eta_j = g(\mu_j) = \sum_{i=1}^{P} f_i(Z_{ij}) + \sum_{k=1}^{Q} \beta_k X_{kj}\,, \tag{3.1}$$

where $g(\cdot)$ is the link function. For instance, $g(\cdot)$ can be a log function for Poisson distributed observations or the logistic function for the binomial distributed data. Variable X_k is of interest with the value X_{kj} on time t_j, Z_i is a confounding variable with the value Z_{ij} on time t_j, and the functional parameter $f_i(Z_i)$ is estimated in a nonparametric form. There are P functional parameters which we consider to be nuisance parameters, and the linear coefficient vector $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_Q)$ is the parameter of interest.

For example, the generalized semiparametric additive model for air pollution data (Ramsay 2005) can be written as follows:

$$\eta_j = \log(\mu_j) = f(D_j) + \beta P_j\,, \tag{3.2}$$

where μ_j's are expectations of daily counts of adverse health events, such as mortality and hospital admissions. Index j is for the day D_j, P_j is the amount of air pollution on day j, and the functional parameter $f(D_j)$ is a nuisance parameter that takes account of the time effect on the log-transformed response. The

structural parameter β is of interest, representing the increase of log-transformed response associated with a unit increase in the amount of air pollution, allowing for the effects of the time trend.

Zeger and Diggle (1994) proposed a back-fitting algorithm to estimate a nonparametric time trajectory $f(t)$ and parametric covariate effects $\boldsymbol{\beta}$. They estimated $f(t)$ with a kernel method and estimated $\boldsymbol{\beta}$ using weighted least squares by accounting for the within-cluster correlations. Lin and Carroll (2001) proposed generalized estimating equations to estimate the semiparametric generalized linear model for cluster data. They used kernel estimating equations to estimate the nonparametric functions and a profile-based estimating equation to estimate the linear coefficient vector $\boldsymbol{\beta}$. Lin and Ying (2001) integrated counting process techniques into estimating model (3.1) and proved that their estimate for $\boldsymbol{\beta}$ was $n^{1/2}-$consistent and asymptotically normal with a simple variance-covariance estimator. They simplified computations by choosing singleton nearest-neighbor smoothing technique. Fan and Li (2004) used local polynomial regression techniques to estimate the nonparametric functions and to simultaneously select significant variables. All the above authors used weighted least square (WLS) to estimate the linear coefficient vector $\boldsymbol{\beta}$. However, as we know, WLS is only valid for Gaussian-distributed data.

Severini and Staniswalis (1994) estimated model (3.1) using a quasi-likelihood function and developed aymptotic distributions for their estimators. They also generalized their method to the case with multivariate response. Liang et al. (1999) pointed out that the quasi-likelihood method would lead to biased estimates for both the nonparametric and parametric terms when measurement errors for covariates were ignored. Liang et al. (1999) estimated the linear coefficient vector $\boldsymbol{\beta}$

by least squares, taking into account the measurement errors of covariates. They also developed sandwich-type estimates for the standard errors of data. Lin and Carroll (2006) considered a wide class of semiparametric problems and proposed profile kernel and back-fitting estimation methods. They showed that profiling and back-fitting have identical limit distributions using kernel smoothing when maximizing the profile likelihood, and they suggested computing the gradients by numerical differentiation, and pointed out that this would be difficult to implement numerically.

One important application of generalized semiparametric additive models is the analysis of the health effect of air pollution. Model (3.2) is often used for this kind of analysis, in which the estimated regression coefficient β is small. The U.S. Environmental Protection Agency (EPA) periodically reviews the National Ambient Air Quality Standards for six air pollutants to protect the public's health. In 2002, EPA delayed completion of the review documents because statisticians and epidemiologist found that the default settings in the *gam* function of the S-Plus software package (version 3.4) didn't assure the convergence of the back-fitting algorithm, and could overestimate effects of air pollution (Dominici et al. 2002). Moreover, Ramsay et al. (2003) showed that S-Plus also underestimated variances of air pollution effects. Dominici et al. (2004) pointed out that the confounding bias could be removed by including the sufficient flexible smoothing functions of time. They also developed a closed-form estimate of the asymptotically exact variance of the linear coefficient β. However, Ramsay (2005) argued that the three assumptions for the smoothing basis in Dominici et al. (2004) were invalid. Ramsay (2005) also discussed two sources of bias: concurvity and model selection,

and demonstrated that the bootstrap couldn't correct conconvity-induced bias.

We develop a method to estimate the generalized semiparametric additive models based on the likelihood functions, working for arbitrarily distributed response variables. The nonparametric functions are estimated by penalized smoothing, with the smoothing parameter vector $\boldsymbol{\lambda}$ controlling the smoothness of the nonparametric functions. We use the generalized profiling method to estimate three distinct groups of parameters: the functional parameters $f_i(Z_i)$'s, the linear coefficient vector $\boldsymbol{\beta}$, and the smoothing parameter vector $\boldsymbol{\lambda}$ and their standard deviations. Each parameter can be multidimensional. The three levels of optimization procedures are conducted: first, the coefficient vector $\mathbf{c}$ is estimated, given $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$, by maximizing the regularized log likelihood function $J(\mathbf{c}|\boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{y})$. Hence, the optimal coefficient vector $\hat{\mathbf{c}}$ is a function of $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$. Next, the linear coefficient vector $\boldsymbol{\beta}$, given $\boldsymbol{\lambda}$, is estimated by maximizing the log likelihood function $H(\boldsymbol{\beta}|\boldsymbol{\lambda}, \mathbf{y})$. Therefore, the optimal linear coefficient vector $\hat{\boldsymbol{\beta}}$ is a function of $\boldsymbol{\lambda}$. Finally, the smoothing parameter vector is estimated by minimizing the criterion $F(\boldsymbol{\lambda}|\mathbf{y})$, which can be defined by any model selection methods.

The functional relationship between these three parameters are important. First, we can derive the unconditional standard deviation estimate of $\boldsymbol{\beta}$, which includes the uncertainty of $\hat{\boldsymbol{\lambda}}$, and thus we can solve the underestimation problem found by (Ramsay, Burnett, and Krewski 2003). Second, in each level of optimization, the gradient and Hessian matrix can be worked out analytically, which is essential for fast and stable computation.

Bates and Watts (1988) used a Newton-Raphson method to find the mini-

mum of the objective function, applying a local quadratic approximation to the objective function. Let $S(\boldsymbol{\theta})$ be the objective function and $\boldsymbol{\theta}^{(i)}$ be the parameter value at the i-th iteration, then the Newton-Raphson method updates the parameter value by

$$\boldsymbol{\theta}^{(i+1)} = \boldsymbol{\theta}^{(i)} - H^{-1}\mathbf{g},$$

where

$$\mathbf{g} = \frac{\partial S}{\partial \boldsymbol{\theta}}$$

is the gradient of $S(\boldsymbol{\theta})$ evaluated at $\boldsymbol{\theta}^{(i)}$, and

$$H = \frac{\partial S}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}$$

is the Hessian matrix of $S(\boldsymbol{\theta})$ evaluated at $\boldsymbol{\theta}^{(i)}$. In our generalized profiling method, the Newton-Raphson algorithm is used to do all three levels of optimization. The algorithm converges quickly and stably with the gradients and Hessian matrices worked out analytically.

A package to estimate the generalized semiparametric additive models with our method has been developed in the Matlab computing language, making use of functional data analysis software intended to compliment Ramsay and Silverman (2005). Users are only required to provide several derivatives of the log likelihood function with respect to $\mathbf{c}$ and $\boldsymbol{\beta}$.

The remainder of this chapter is organized as follows. Section 3.2 introduces how to estimate generalized semiparametric additive models by the generalized profiling method. All the mathematical details are written in Appendix B. Section

3.3 shows our estimates based on the air pollution data. The parametric bootstrap is applied to validate our estimates and to estimate the variance of linear coefficients. The generalized profiling method shown in this chapter is also easy to extend to estimate other statistical models involving three distinct groups of parameters by choosing appropriate criteria.

3.2 The Generalized Profiling Method

In this section we first write down the generalized semiparametric additive model in a simple form, and then introduce how to estimate the nonparametric functions, linear coefficients and smoothing parameters in three levels of optimization. Finally, we derive unconditional estimates for variances of linear coefficients.

The functional parameters $f_i(Z_i)$ are estimated by linear combinations of K_i B-spline basis functions:

$$f_i(Z_i) = \sum_{k=1}^{K_i} c_{ik}\phi_{ik}(Z_i) = \mathbf{c}_i'\boldsymbol{\phi}_i(Z_i)\,,$$

where $\mathbf{c}_i = (c_{i1}, \cdots, c_{iK_i})'$ and $\boldsymbol{\phi}_i(Z_i) = \Big(\phi_{i1}(Z_i), \cdots, \phi_{iK_i}(Z_i)\Big)'$. Let $\boldsymbol{\Phi}_i$ be an order $n \times K_i$ matrix with the j-th row $\boldsymbol{\phi}_i(Z_{ij})'$, then the generalized semiparametric additive model (3.1) can be written in the simple matrix form:

$$\boldsymbol{\eta} = g(\boldsymbol{\mu}) = \boldsymbol{\Phi}\mathbf{c} + \mathbf{X}\boldsymbol{\beta}\,, \tag{3.3}$$

where $\boldsymbol{\eta} = (\eta_1, \cdots, \eta_n)$, $\boldsymbol{\mu} = (\mu_1, \cdots, \mu_n)$, $\mathbf{c} = (\mathbf{c}_1', \cdots, \mathbf{c}_P')'$, $\boldsymbol{\Phi} = (\boldsymbol{\Phi}_1, \cdots, \boldsymbol{\Phi}_P)$ and $\mathbf{X}$ is an $n \times Q$ matrix with jk-th entry x_{kj}.

3.2.1 The First Optimization Level to Estimate Local Parameters

The optimization criterion in the first level is written as:

$$J(\mathbf{c}|\boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{y}) = -l(\mathbf{c}, \boldsymbol{\beta}|\mathbf{y}) + \sum_{i=1}^{P} \lambda_i \int [L_i f_i(Z_i)]^2 dZ_i \,, \tag{3.4}$$

where $l(\mathbf{c}, \boldsymbol{\beta}|\mathbf{y})$ is the log likelihood function. The second term in (3.4) penalizes the roughness of functional parameters, so a positive sign is used in front of it such that the optimal coefficient vector $\mathbf{c}$ can be estimated by minimizing $J(\mathbf{c}|\boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{y})$. L_i is a linear differential operator of order m:

$$L_i x(t) = \sum_{j=0}^{m-1} \alpha_j(t) D^j x(t) + D^m x(t) \,.$$

The penalty term $\int [L_i f_i(Z_i)]^2 dZ_i$ can be written as a quadratic function of the coefficient vector $\mathbf{c}_i$:

$$\int [L_i f_i(Z_i)]^2 dZ_i = \mathbf{c}_i' \mathbf{R}_i \mathbf{c}_i \,,$$

where $\mathbf{R}_i = \int [L_i \boldsymbol{\phi}_i(t)][L\boldsymbol{\phi}_i(t)]' dt$ is an order K_i matrix. Then the second term in (3.4) can be represented in the matrix form:

$$\sum_{i=1}^{P} \lambda_i \int [L_i f_i(Z_i)]^2 dZ_i = \mathbf{c}'\mathbf{R}\mathbf{c}\,,$$

where $\mathbf{c} = (\mathbf{c}_1', \cdots, \mathbf{c}_P')'$ and $\mathbf{R} = \text{diag}(\lambda_1\mathbf{R}_1, \cdots, \lambda_P\mathbf{R}_P)$. In order to attain a positive estimate for the smoothing parameter vector, we express $\boldsymbol{\lambda} = (\lambda_1, \cdots, \lambda_P)' = \exp(\boldsymbol{\theta})$, where $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_P)'$. All simulations and applications in this chapter use the second derivative to define the roughness penalty term, that is, $L = D^2$, but Ramsay and Silverman (2005) show how to obtain better estimates by penalized smoothing with penalty terms defined by differential operators. The first and second derivatives of $J(\mathbf{c}|\boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{y})$ with respect to $\mathbf{c}$ are given in (B.3) and (B.4), respectively.

For given values of $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$, the coefficient vector $\mathbf{c}$ can be estimated by minimizing the optimization criterion (3.4) in the first level, so that the estimated $\hat{\mathbf{c}}$ can be viewed as a function of $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$. However, there is no explicit form of this function except when observations are normally distributed. That is why least squares estimations are often used in many of the literature, instead of likelihood functions. Fortunately, we can write out any order derivatives of $\hat{\mathbf{c}}$ with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$ analytically using Implicit Function Theorem. The details are given in Appendix B.

3.2.2 The Second Optimization Level to Estimate Global Parameters

The optimization criterion in the second level is written as:

$$H(\boldsymbol{\beta}|\boldsymbol{\lambda}, \mathbf{y}) = -l(\hat{\mathbf{c}}(\boldsymbol{\beta}), \boldsymbol{\beta}|\mathbf{y}) . \tag{3.5}$$

The coefficient vector $\hat{\mathbf{c}}$ disappears in the log likelihood function, because it is now a function of $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$. As explained in Chapter 1, the optimization criterion in the second level does not include the penalty term any more, since $\hat{\mathbf{c}}$ itself already contains the regularization information, and this information is passed to the log likelihood function by treating $\hat{\mathbf{c}}$ as a function of $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$.

The first and second derivatives of $H(\boldsymbol{\beta}|\boldsymbol{\lambda}, \mathbf{y})$ with respect to $\boldsymbol{\beta}$ are given in (B.28) and (B.29), respectively.

The linear coefficient vector $\boldsymbol{\beta}$ can be estimated, given any value of $\boldsymbol{\lambda}$. Therefore, the estimator $\hat{\boldsymbol{\beta}}$ is a function of $\boldsymbol{\lambda}$. In most cases, this function is not explicit, but we can attain analytical forms of any order derivatives of $\hat{\boldsymbol{\beta}}$ with respect to $\boldsymbol{\lambda}$, as shown in Appendix B.

3.2.3 The Third Optimization Level to Estimate Complexity Parameters

The smoothing parameter vector $\boldsymbol{\lambda}$ is a complexity parameter, and controls the effective degrees of freedom of the generalized semiparametric additive models.

Efron (2004) reviewed the model selection methods and proposed some interesting new approaches. However, none of these methods leave analytic formulas for observations in any distributions. In the following, the response variable is assumed to come from an exponential family such that the approximated GCV (Gu and Xiang 2001) can be applied as the optimization criterion in the third level to estimate $\boldsymbol{\lambda}$. If we can find other model selection criteria in close forms for other distributed observations, our method can still be applied easily.

Moreover, we can also write out $d\boldsymbol{\lambda}/dy$ analytically, and use the Delta method to find the standard deviation for $\boldsymbol{\lambda}$. The estimated linear coefficient vector $\hat{\boldsymbol{\beta}}$ is a function of $\boldsymbol{\lambda}$, so the unconditional estimate for the standard deviation of $\hat{\boldsymbol{\beta}}$, $SD(\hat{\boldsymbol{\beta}})$, can be derived, which includes the deviation coming from $\hat{\boldsymbol{\lambda}}$. This solves the underestimation problem for $SD(\hat{\boldsymbol{\beta}})$, which is found by Ramsay et al. (2003).

Assuming that the observation Y_j is distributed in the exponential family, we can write down the probability density function:

$$f(Y_j) = \exp\{\frac{Y_j\eta_j - b(\eta_j)}{a(\phi)} + h(Y_j, \phi)\}\,, \tag{3.6}$$

where η_j has the same definition as (3.1), ϕ is a nuisance parameter, and $a(\phi)$ is called the dispersion parameter. From the standard exponential family theory, we know that $db(\eta_j)/d\eta_j = \mu_j = \mathtt{E}(Y_j)$.

Since $h(Y_j, \phi)$ is independent of η_j, the log likelihood function $l(\mathbf{c}, \boldsymbol{\beta}|\mathbf{y})$ can

be written as

$$l(\mathbf{c}, \boldsymbol{\beta}|\mathbf{y}) = \frac{1}{n}\sum_{j=1}^{n}\{Y_j\eta_j - b(\eta_j)\}\,. \tag{3.7}$$

up to an additive constant. Notice that the dispersion parameter $a(\phi)$ is absorbed into the smoothing parameter vector $\boldsymbol{\lambda}$ in the first level of optimization criterion (3.4).

When data were distributed in the exponential family, Xiang and Wahba (1996) proposed the generalized approximate cross-validation (GACV) score to choose the proper value of the smoothing parameter vector $\boldsymbol{\lambda}$. Gu and Xiang (2001) reported that the computation for the GACV score could be numerically unstable for large n, and proposed an alternative derivation of the GACV score, which was computationally stable for all sample sizes. This new GACV score is used as the optimization criterion in the third level:

$$F(\boldsymbol{\lambda}|\mathbf{y}) = -\frac{1}{n}\sum_{j=1}^{n}\{y_j\eta_j - b(\eta_j)\} + \frac{\alpha}{n}\frac{\mathtt{Tr}(\boldsymbol{\Phi}\mathbf{B}^{-1}\boldsymbol{\Phi}')}{n - \mathtt{Tr}\mathbf{A}}\sum_{j=1}^{n} y_j(y_j - \mu_j)\,, \tag{3.8}$$

where $\mathbf{B} = \boldsymbol{\Phi}'\mathbf{W}\boldsymbol{\Phi} + \mathbf{R}$, $\mathbf{A} = \boldsymbol{\Phi}\mathbf{B}^{-1}\boldsymbol{\Phi}'\mathbf{W}$, $\mathbf{W} = \mathrm{diag}(w_i)$ with $w_i = \partial^2 b(\eta_i)/\partial\eta_i^2$, and $\alpha \geq 1$ is a constant. Gu and Ma (2003) suggested α in the range of $1.2 \sim 1.4$ to prevent severe undersmoothing typically suffered by cross-validation methods, with little loss of general effectiveness.

A Newton-Raphson algorithm is applied to find the optimal smoothing parameter vector $\hat{\boldsymbol{\lambda}}$, and it converges quickly and stably with the analytic gradient and Hessian matrix given in (B.39) and (B.40), respectively.

3.2.4 Unconditional Variance Estimation for Global Parameters

The total derivative of $\boldsymbol{\beta}$ with respect to $\mathbf{y}$ is:

$$\frac{d\boldsymbol{\beta}}{d\mathbf{y}} = \frac{\partial\boldsymbol{\beta}}{\partial\boldsymbol{\theta}}\frac{d\boldsymbol{\theta}}{d\mathbf{y}} + \frac{\partial\boldsymbol{\beta}}{\partial\mathbf{y}}, \tag{3.9}$$

where $\boldsymbol{\theta} = \ln(\boldsymbol{\lambda})$. Derivatives $\frac{\partial\boldsymbol{\beta}}{\partial\boldsymbol{\theta}}$, $\frac{d\boldsymbol{\theta}}{d\mathbf{y}}$, and $\frac{\partial\boldsymbol{\beta}}{\partial\mathbf{y}}$ are given in (B.30), (B.41) and (B.32). By the Delta method, the unconditional variance-covariance matrix of the linear coefficient vector is estimated by:

$$\mathtt{Var}[\boldsymbol{\beta}(\mathbf{y})] = \left[\frac{d\boldsymbol{\beta}}{d\mathbf{y}}\right]\boldsymbol{\Sigma}\left[\frac{d\boldsymbol{\beta}}{d\mathbf{y}}\right]', \tag{3.10}$$

where $\boldsymbol{\Sigma}$ is the variance-covariance matrix of $\mathbf{y}$. We assume observations are independent, and estimate $\boldsymbol{\Sigma}$ by:

$$\hat{\boldsymbol{\Sigma}} = \mathrm{diag}\left[\frac{\mathbf{r}^T\mathbf{r}}{n - \mathtt{Tr}A}\frac{dg^{-1}(\mathbf{y})}{d\mathbf{y}}\left(\frac{dg^{-1}(\mathbf{y})}{d\mathbf{y}}\right)'\right], \tag{3.11}$$

where $g(\cdot)$ is the link function in the generalized semiparametric additive model (3.1) and the residual vector $\mathbf{r} = g(\mathbf{y}) - \boldsymbol{\Phi}\mathbf{c} - X\boldsymbol{\beta}$.

On the other hand, when we assume a fixed value of the smoothing parameter vector, the conditional variance-covariance matrix of the linear coefficient vector is estimated by

$$\mathtt{Var}[\boldsymbol{\beta}(\mathbf{y})|\boldsymbol{\lambda}] = \left[\frac{\partial\boldsymbol{\beta}}{\partial\mathbf{y}}\right]\boldsymbol{\Sigma}\left[\frac{\partial\boldsymbol{\beta}}{\partial\mathbf{y}}\right]'. \tag{3.12}$$

3.3 Parameter Estimates from Air Pollution Data

Figure 3.1 displays the daily counts of non-accidental deaths from 1987 to 1988 in Toronto, as well as the daily one-hour-maximum ozone, where ozone has the seasonal trend with large concentrations in summer. Our objective is to find whether the amount of daily ozone has any effect on mortality, allowing for a seasonal trend. In this section, we estimate the generalized semiparametric additive model for air pollution data and apply parametric bootstrap to validate our estimates.

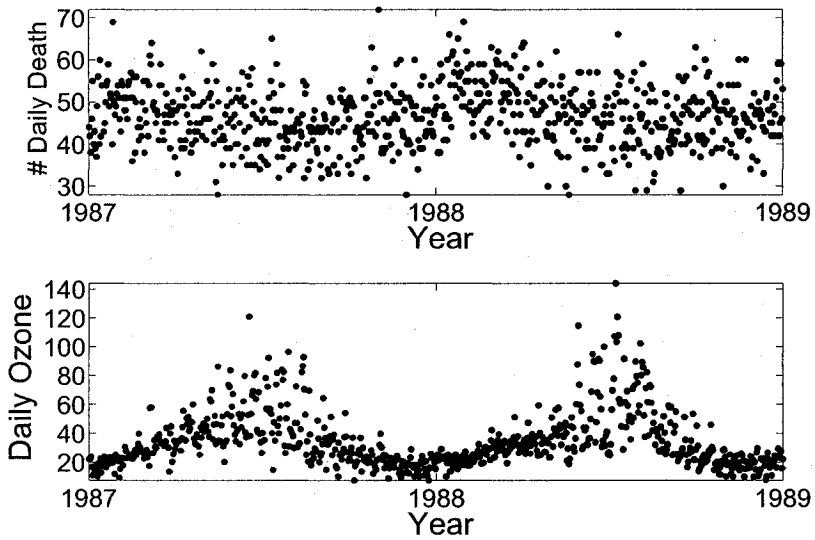

Figure 3.1: The top panel displays the daily count of non-accidental deaths from 1987 to 1988 in Toronto, and the bottom panel shows the associate daily one-hour-maximum ozone.

Let $\{y_j\}_{j=1}^n$ be daily counts of non-accidental deaths, x_j is the daily one-hour-maximum ozone, and j is the index of the day. We assume y_j to have a

Poisson distribution, possibly with over-dispersion, then the probability density function of y_j can be written in the form of (3.7) with $b(\eta_j) = e^{\eta_j}$, and (3.2) is the generalized semiparametric additive model for y_j.

3.3.1 Estimates for Local, Global and Complexity parameters

The estimated smoothing parameter is $\hat{\lambda} = 53.7$. The linear coefficient estimate $\hat{\beta} = 9.1 * 10^{-4}$, representing about a 0.09 percent increase in mortality associated with an unit increase of the daily one-hour-maximum ozone. The estimated non-parametric function $f(t)$ shows the seasonal trend, large in winter, as displayed in Figure 3.2. The corresponding expectation of daily counts of deaths also shows the similar seasonal trend, except that it is increased by the effect of Ozone in summer.

The estimated degrees of freedom $\mathtt{df} = \mathtt{Tr}A = 11$, and the estimated variance of daily death counts is shown in Figure 3.3. Comparing with the daily death counts, we conclude that the data have an overdispersed Poisson distribution. The estimated SD for λ is 26.3, and the estimated SD for β is $4.1 * 10^{-4}$. The 95% confidence interval for β is $[1.1, 17.2] * 10^{-4}$, which indicates that ozone has a significant effect on mortality. In the following, we validate our estimates by a parametric bootstrap.

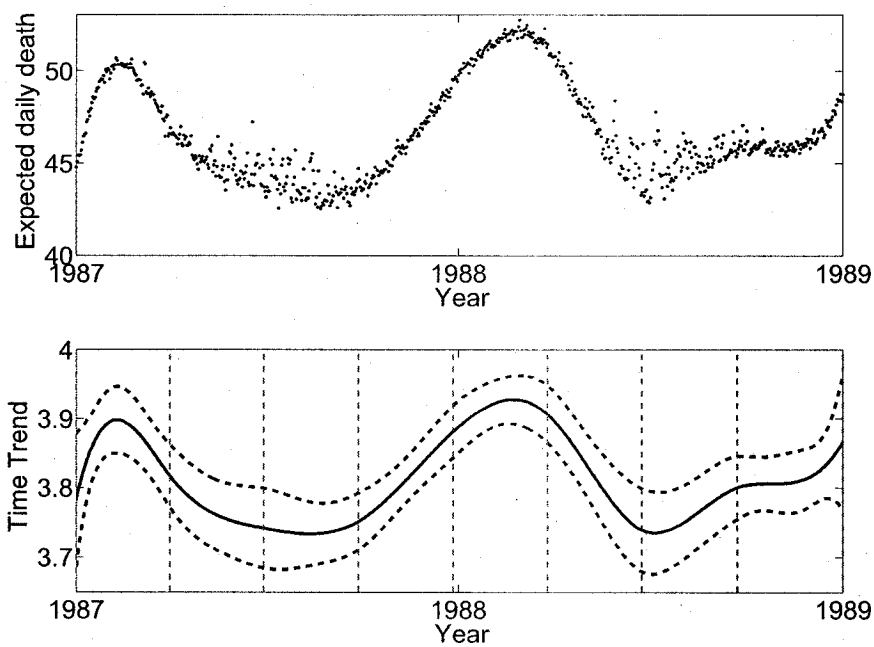

Figure 3.2: The unconditional estimated expectation of daily count of non-accidental deaths from 1987 to 1988 in Toronto (top panel). The bottom panel shows the estimated functional parameter $\hat{f}(t)$ with the 95% confidence band, which is expanded by cubic B-splines with the knots indicated by the blue dashed lines.

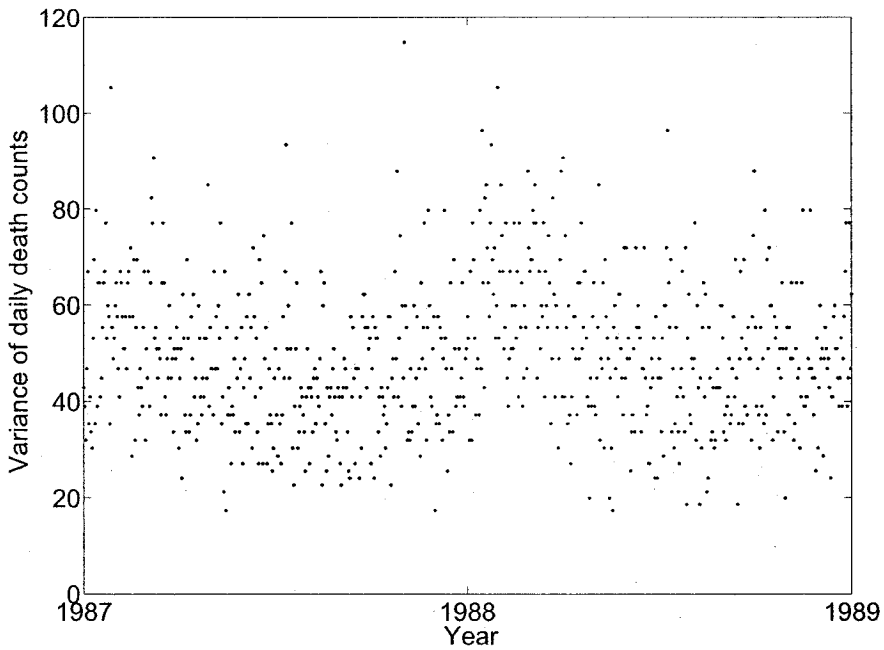

Figure 3.3: The estimated variance of daily death counts from 1987 to 1988 in Toronto.

3.3.2 Bootstrap Validation for Parameter Estimates

We use parametric bootstrap to validate our estimates for the generalized semi-parametric additive model. We generate 1000 sets of Poisson data $\{y_j\}_{j=1}^n$ with the mean $\hat{\mu}(t)$ estimated from the real data set, and figure 3.4 shows one typical data set. In the following, we estimate the smoothing parameter λ, the linear coefficient β and the functional parameter $f(t)$ from these data sets with the generalized profiling method. Figure 3.5 shows the bias and RMSE of estimated $\hat{\mu}(t)$ on the air

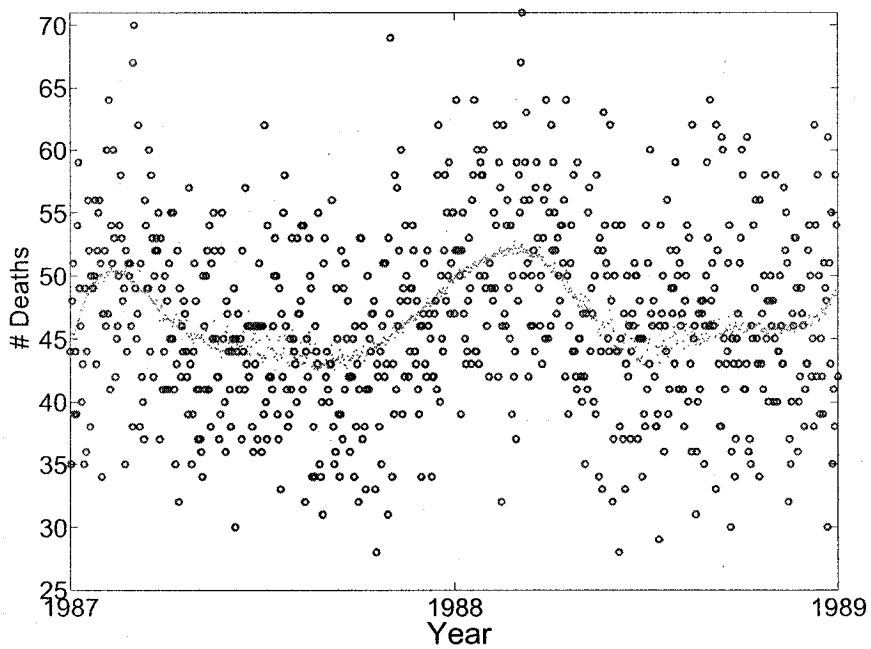

Figure 3.4: One set of simulated Poisson data (blue circles) with the mean $\hat{\mu}(t)$ estimated from the real data set (red dots).

pollution data, which are both small.

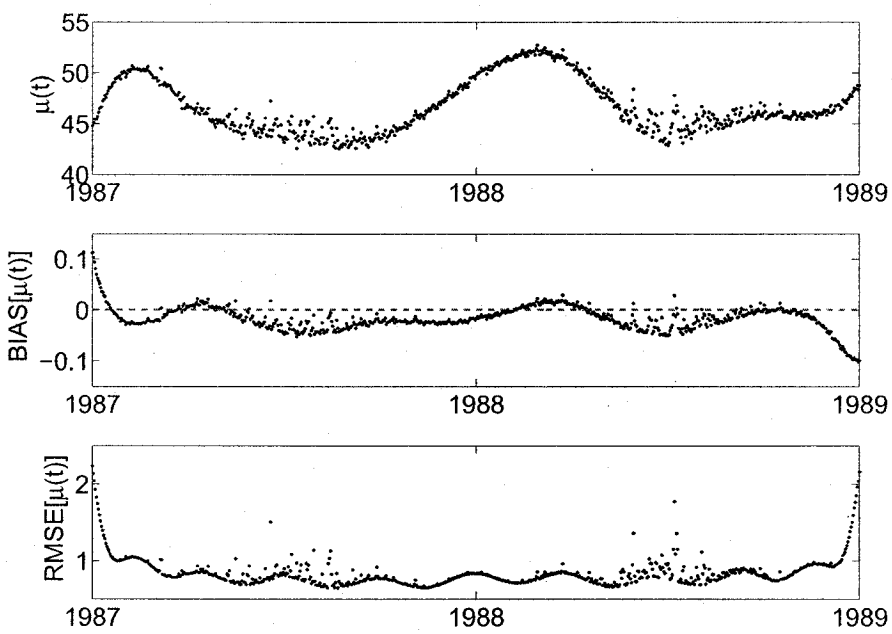

Figure 3.5: The bias and RMSE of estimated $\hat{\mu}(t)$ on the air pollution data.

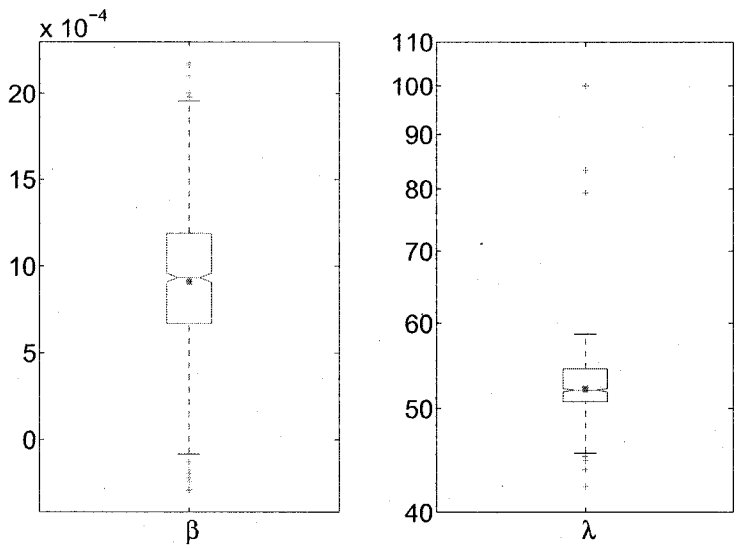

Figure 3.6: The boxplots for the estimated linear coefficient $\hat{\beta}$'s and smoothing parameters $\hat{\lambda}$'s. The red dots are the values of β and λ used to generate the simulated data sets, respectively.

The boxplot for the estimated smoothing parameter $\hat{\lambda}$ is shown in the right panel of Figure 3.6. The standard deviation for $\hat{\lambda}$ is 20.7, showing that GACV does not give stable estimates for λ. The boxplot for β is displayed in Figure 3.6. The bias of the estimated $\hat{\beta}$ is only 1% of the true value, and the SD of estimated $\hat{\beta}$ is $4.0 * 10^{-4}$.

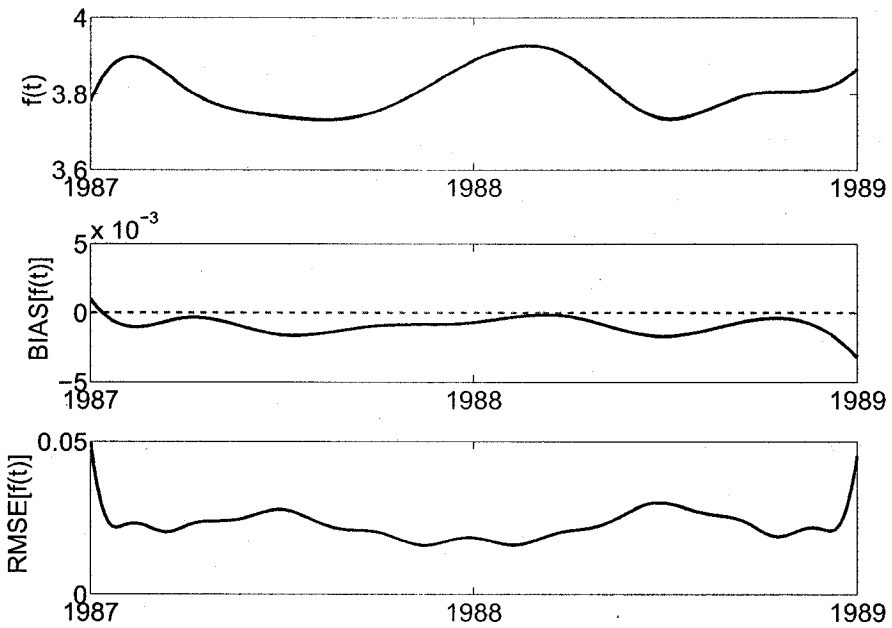

Figure 3.7: The bias and RMSE of estimated $\hat{f}(t)$ on the air pollution data.

Figure 3.7 displays the bias and RMSE of estimated $\hat{f}(t)$ on the air pollution data, which are both small.

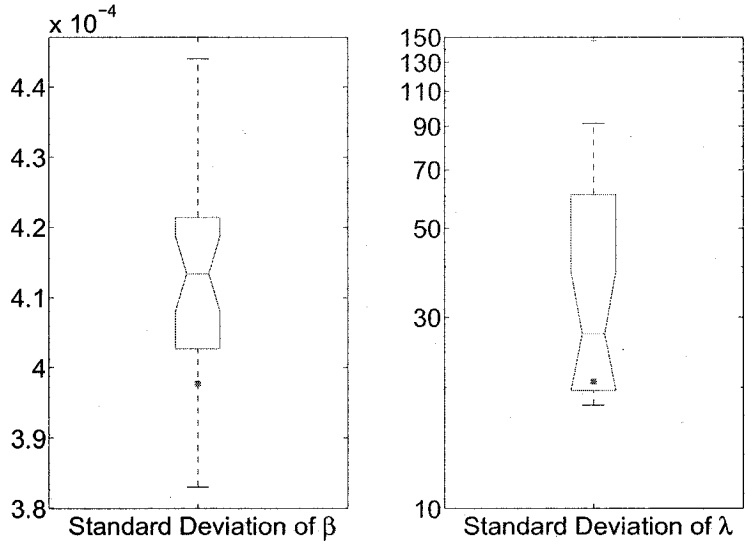

Figure 3.8: The boxplots for estimated SD's of the linear coefficient β and the smoothing parameter λ. The red dots are the experimental SD's of β and λ, respectively.

Figure 3.8 displays the boxplots for estimated SD's of the linear coefficient β and the smoothing parameter λ. The experimental SD's of β and λ are well in the 95% confidence intervals of estimated SD's. The median of the estimated SD for β is 4% larger than the experimental value, and the median of the estimated SD for λ is 32% larger than the experimental value.

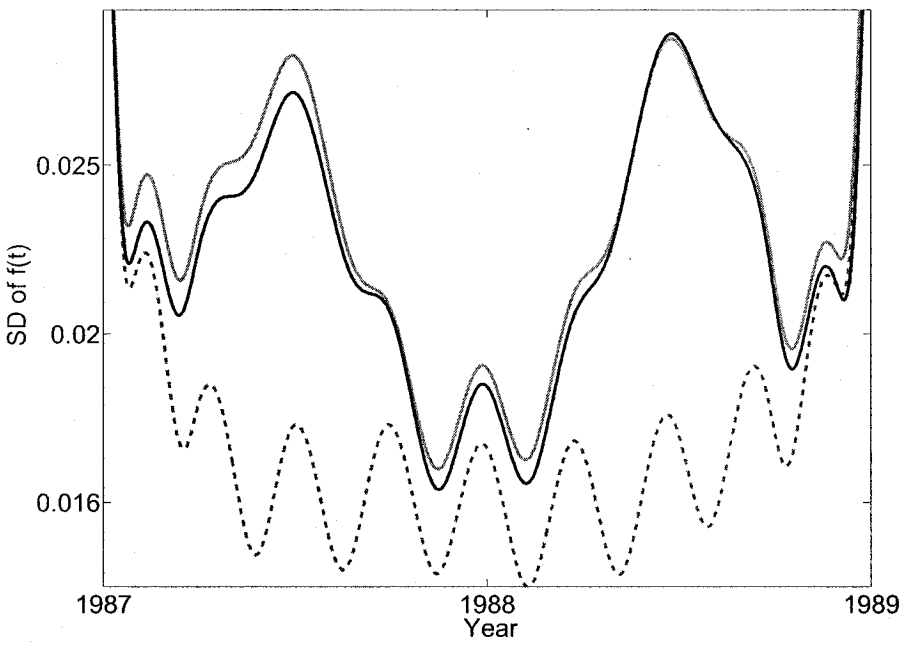

Figure 3.9: The estimated SD's of the nonparametric function $f(t)$. The blue line is the conditional estimate, ignoring the variance coming from β and λ, and the red line is the unconditional estimate. The blue line is the experimental value.

Figure 3.9 displays the estimated SD's of the nonparametric function $f(t)$. The unconditional estimate is well close the experimental value, but the conditional estimate underestimate the SD's, since it ignores the variance coming from β and λ.

Chapter 4

Estimating Differential Equations (DE's)

4.1 Introduction to Estimating DE's from Data

Differential equations (DE's) are used to model the rate of change of a process defined over time, space, or some other continuum. We can write down a general formulation for DE's as follows:

$$D\mathbf{x}(t) = \mathbf{f}(\mathbf{x}, t|\boldsymbol{\theta}), \tag{4.1}$$

where $\mathbf{x}$ is a vector of T components, which are functions varying over t, $D\mathbf{x}$ is the corresponding vector of first derivatives with respect to t, and $\boldsymbol{\theta}$ is a vector of parameters. Higher order DE's

$$D^p\mathbf{x}(t) = \mathbf{f}(\mathbf{x}, D\mathbf{x}(t), \cdots, D^{p-1}\mathbf{x}(t), t|\boldsymbol{\theta})$$

can be reduced to the first order DE's by adding some new DE's:

$$\begin{aligned} D\mathbf{x}(t) &= \mathbf{x}_1(t)\,, \\ D\mathbf{x}_1(t) &= \mathbf{x}_2(t)\,, \\ &\cdots \\ D\mathbf{x}_{p-1}(t) &= \mathbf{f}(\mathbf{x}, \mathbf{x}_1(t), \cdots, \mathbf{x}_{p-1}(t), t|\boldsymbol{\theta})\,. \end{aligned}$$

DE's are widely used in engineering, biology, ecology, economics, neuroscience, and medicine, and have recently been used to model the dynamic behavior of gene expression (Jaeger et al. 2004). The oldest and most famous example is perhaps Newton's second law: $F = ma$, where a is the acceleration (the first derivative of the velocity or second derivative of position), m is the mass, and F is the exogenous force. Newton's second law can also be written in the form of DE:

$$D^2 x(t) = \frac{F}{m}\,,$$

where $x(t)$ is the position function. This simple DE beautifully reveals the linear relationship between the acceleration and the force.

How can we fit dynamic models to data? This is called the system identification problem in engineering. In statistical terms, we assume the whole or part of T component vector $\mathbf{x}$ to be observed at n time points $t_1, \cdots, t_n$, and $\mathbf{x}$ to satisfy (4.1), and our objective is to obtain the statistical inference for $\boldsymbol{\theta}$.

If DE's can be solved analytically, it is easy to implement the parameter estimation, model fitting and verification (Bates and Watts 1988). Unfortunately,

very few real-world DE's can be solved analytically, and numerical approximation is almost always the only option in the large and realistic world of nonlinear DE's and non-stationary processes.

At the same time, the current numerical methods for solving DE's are much more highly developed for initial value problems where the only information required and used is the complete state of a system at the initial time point. But DE's often have to be fit to data available throughout a time period.

The current methods to estimate parameters in DE's from noisy data are slow and unstable. There are few statistical techniques to conduct formal and rigorous interval estimations and inferences. In this chapter we introduce an approach to obtain statistical inferences for parameters defining DE's, proposed by Ramsay, Hooker, Cao, and Campbell (2005). This method is based on the modified penalized smoothing and the generalized profiling method.

The remainder of this chapter is organized as follows. The literature about estimating DE's is reviewed in Section 4.2, and Section 4.3 reviews the literature about the predator-prey dynamic systems and displays one experimental predator-prey data set. Section 4.4 introduces a simple HIV dynamic model and data of the number of HIV virus for 42 patients. Section 4.5 introduces penalized smoothing of the data with the penalty term defined by DE's, and the smoothing parameter is optimized by generalized cross validation and Stein's unbiased risk estimate, as discussed in Section 4.6. Section 4.7 introduces how to estimate DE parameters from noisy data with the generalized profiling method, and discusses the effect and selection of smoothing parameters. Section 4.8 introduces how to estimate

functional parameters in DE's. The results of fitting the predator-prey DE's and the HIV DE's to real data are shown in Section 4.9 and 4.10, respectively. Section 4.11 explores dynamic models for the thermal decomposition of α-Pinene.

4.2 Literature Review for Estimating DE's from Data

The most commonly used method for identifying DE's from data is the nonlinear optimization procedure. DE's, given the specific parameter values and initial values of components, are solved with some numerical methods, such as Runge-Kutta methods. While most methods use sum of squared errors as the optimization criterion, other objective functions can also be computed to determine the goodness of fit, which can be likelihood functions, or fairly complex nonlinear functions that incorporate our assumptions about the general covariance structure of measurement errors. A nonlinear optimization method is then employed to update the parameter values and initial values of components. The Newton-Raphson algorithm can be applied here, which is introduced in Section 3.1. Supplying the gradient and the Hessian matrix can increase the efficiency and stability of this algorithm (Biegler, Damiano, and Blau 1986).

There are many drawbacks in the nonlinear optimization procedure. First, the computations are usually intensive, since DE's are repeatedly numerically solved when updating the parameter values and initial values of components. Second, initial values of components become additional parameters to estimate. Fi-

nally, this procedure relies heavily on the quality of the initial guess of parameter values and initial values of components, the algorithms can be easily trapped in local minima, and in some cases DE's may not even be solvable (Bock 1981).

Bock (1981) and Bock (1983) overcame the last problem by a multiple shooting method. The whole time interval of measurement is partitioned into segments. The nonlinear optimization procedure is applied over each segment with the different guessed initial values and the same parameter values. The trajectory is allowed to be discontinuous at the beginning of the optimizing iterations, but is forced to be continuous at the end. Timmer et al. (2000) exemplified this strategy on an experimental time series from a chaotic circuit and reconstructed accurately the observed attractor. The multiple shooting method has been applied in the parameter estimates in partial differential equations by Müller and Timmer (2004) and delay differential equations by Horbelt et al. (2002). However, the multiple shooting method increases the number of initial values to estimate, which increases the dimensionality of the parameter space linearly with the number of segments. The computational burden is also increased by solving DE's over each segment.

There can be many local minima when estimating DE parameters. The global optimal values of DE parameters can be found by simulated annealing when the fit surface has local minima in the nonlinear optimization procedure. But the intensive computation makes this method unreasonable for routine usage. For instance, Jaeger et al. (2004) reported that it took 10 2.4-Ghz Pentium P4 Xeon processors between 8 and 160 hours per optimization run. Esposito and Floudas (2000) proposed a deterministic global optimization approach to find the global optimized parameter values in differential-algebraic equations by generating a valid

convex underestimation of the original nonconvex fit surface.

When a large number of observations are available, Himmelblau et al. (1967) integrated (4.1) by numerical quadrature, converting DE's (4.1) to the system of linear equations

$$\mathbf{x}(t_k) - \mathbf{x}(t_0) = \int_{t_0}^{t_k} \mathbf{f}(\mathbf{x}, t|\boldsymbol{\theta})dt\,,$$

where t_k is the time points with observations. When the number of parameters is no more than the number of equations, this system can be solved by the simple least square method. However, the integral estimation is very sensitive when components change rapidly. This sensitivity becomes even worse when the initial component values are not accurately measured. This method also involves intensive computations. When only a small number of data points were available, Tang (1971) extended this method by estimating $\mathbf{x}(t)$ by natural cubic splines and obtaining the integrals analytically. Swartz and Bremermann (1975) improved this method by the global optimization technique, but required a long computation time. To improve the efficiency, they suggested transforming parameters such that their expected variances were same. Swartz and Bremermann (1975) also calculated variances of parameter estimates using the technique of Rosenbrock and Storey (1966).

de Boor and Swartz (1973) approximated solutions of nonlinear DE's with piecewise polynomial functions by collocation. They required the piecewise polynomial functions to satisfy DE's at the collocation sites and derived them by solving the sequence of linear collocation problems associated with Newton's method.

When all components $\mathbf{x}(t)$ in DE's are measured, an alternative approach

is to estimate the derivative vector $D\mathbf{x}(t)$ by smoothing observations. Then the system identification problem becomes much easier, and many routine statistical techniques can be applied, for example, functional linear models (Ramsay and Silverman 2005). The derivative vector $D\mathbf{x}(t)$ was estimated with the finite difference method by Voss et al. (1998). But Swartz and Bremermann (1975) pointed out that "small errors in the measured values of the state variables can produce large errors in numerical differentiation". Instead, Swartz and Bremermann (1975) estimated derivatives by smoothing data with polynomials. But when the data have a large amount of noise, it is very easy to overfit, that is, the fitting functions have a lot of unexpected ripples and the estimated derivatives are correspondingly too large. Varah (1982) decreased the computation work by smoothing data with B-splines. The B-spline basis functions are non-zero only over localized intervals, which is called "compact support" by de Boor (2001). Varah (1982) also overcome the overfitting problem by choosing the number and positions of knots using interactive graphics. There are two shortcomings for this approach. First, the estimate for the derivative vector $D\mathbf{x}(t)$ is still biased and unstable, especially at the boundaries (Ramsay and Silverman 2005). As a result, the parameter estimates are also biased. Next, in practice, it can often happen that some components are not observable, and hence there is no way to estimate the corresponding derivatives.

Benson (1979) developed a package PARFIT to estimate DE parameters when some components do not have observations available. This package allows users to guess the initial values of components and to select the derivative or integral fitting methods in an interactive manner. This package is especially useful when initial values of parameters are not good.

Gelman, Bois, and Jiang (1996) and Huang, Liu, and Wu (2005) used a Bayesian approach in which they proposed some informative priors for DE parameter vector $\boldsymbol{\theta}$. For given $\boldsymbol{\theta}$, equations (4.1) are solved numerically with solutions, say, $\mathbf{g}(\boldsymbol{\theta})$. The observations or their transformations are assumed to have a distribution with mean $\mathbf{g}(\boldsymbol{\theta})$. There are no closed forms for the posterior distributions without analytic DE solutions. Markov chain Monte Carlo (MCMC) is the common method for posterior simulations. The statistical inferences for $\boldsymbol{\theta}$ can then be obtained from the posterior samplings. The Bayesian method can also handle mixed effect models.

However, there are also many downsides to this Bayesian method. First, the computation burden is large, since DE's must be solved at each iteration with updated $\boldsymbol{\theta}$. Second, the initial values for the system components must also be treated as additional parameters. Furthermore, choosing a prior may be difficult since non-informative priors may lead to improper posteriors (Bates and Watts 1988). Finally, it can be difficult to get the simulation chains to converge and more advanced methods like tempering may be necessary to overcome bifurcations in DE's or multiple posterior modes.

Ramsay et al. (2005) proposed a method that was economical in computation time. DE's do not have to be solved, and hence the initial values of components are not needed. Their method can also work satisfactorily when some components are not observable. The idea is to smooth data with a linear combination of basis functions, penalized by its fidelity to DE's. A smoothing parameter λ reconciles the trade-off between fitting the data and fidelity to DE's. This process is called *the L-spline smoothing* by Ramsay and Silverman (2005). For any given parameter

vector $\boldsymbol{\theta}$, a coefficients vector $\mathbf{c}$ of basis functions can be estimated by the L-spline smoothing. The dimension of parameter space is reduced by treating the coefficient vector $\mathbf{c}$ as the implicit function of $\boldsymbol{\theta}$. The gradient and Hessian matrix for optimizing $\boldsymbol{\theta}$ is also calculated analytically using the Implicit Function Theorem. This is called the generalized profiling method, as introduced in Chapter 1 and 2. A byproduct of this method is that it can estimate initial values of missing components in the L-spline smoothing process.

4.3 Introduction for Predator-Prey Dynamic Models

Many organisms in the field and laboratory display fluctuations in population size that can be modeled mathematically by nonlinear interactions among species. These deterministic nonlinear mathematical models can help us to understand and predict the dynamics of interacting populations. In this section we review some of these models and show a set of experimental observations for one predator-prey dynamic system.

The Lotka-Volterra model is the pioneering and the simplest possible predator-prey dynamic model. Let H and P be the number of prey and predators per unit area or volume, respectively, then the Lotka-Volterra model is

$$\begin{aligned} \frac{dH}{dt} &= rH - aHP \\ \frac{dP}{dt} &= eaHP - dP, \end{aligned} \tag{4.2}$$

where r is the per-head rate of increase including the fraction dying from causes other than predation, and aH is the number of prey killed by predators per unit time. Each killed prey is converted to e new predators, and d is the death rate per predator, which is assumed independent of prey density. Many implicit assumptions are made in this model. For instance, it assumes the populations are large enough such that the state variables H and P can be regarded as continuous. The populations are "closed" and there is no input from outside. All parameters are constant, allowing no changes caused by seasonality, weather or other factors.

The Lotka-Volterra model assumes the prey population to grow exponentially in the absence of the predator, which is reasonable for low prey density. When the prey density is high, the prey's resource population is depressed, and the prey's feeding rate is decreased, and hence the prey's birth and death rate is also decreased. A simple way to model the density-dependence growth rate of the prey is to replace rH in the Lotka-Volterra model by a logistic form $rH(1-H/k)$, where k is a constant.

The functional response $g(H)$ describes how feed rate per predator changes with the prey density. In (4.2), $g(H) = aH$, that is, the functional response increases linearly with the prey density, which is called a type 1 response. Real predators cannot eat an unlimited amount of prey per unit time. When the prey density is high, this assumption is clearly not feasible. It takes some time, say T_h, for each predator to search, find and kill one prey (Holling 1959). Although the number of prey encountered per unit of search time is still aH, the fraction of time spent searching decreases when the prey density H increases. Therefore the maximum predation rate is $1/T_h$ prey per day. This yields a type 2 functional

response, $g(H)$, which is substituted for aH in (4.2). It is showed that all type 2 functional responses cause an unstable equilibrium regardless of the particular function form (Oaten and Murdoch 1975). A common form for type 2 functional responses is:

$$g(H) = \frac{aH}{1 + aT_h H}.$$

There are also the type 3 functional responses. One common form is:

$$g(H) = \frac{aH^2}{H^2 + k},$$

where k is a constant. Oaten and Murdoch (1975) showed that all type 3 functional responses led to the stable equilibrium when the prey density was low, regardless of the particular form.

Hassell (1978) pointed out that the interference between predators could reduce their searching efficiency. Beddington (1975) suggested the type 2 functional response should hence decrease with predator density:

$$g(H) = \frac{aH}{1 + aT_h H + kP}.$$

McNair (1987) considered a prey with a juvenile and adult stage, which differed in their vulnerability. Let A and I be the number of adult and juvenile

prey, respectively, then his model can be simplified as:

$$\begin{aligned}\frac{dI}{dt} &= rI - mI - a_I IP \\ \frac{dA}{dt} &= mI - a_A AP \\ \frac{dP}{dt} &= eP(a_A A + a_I I) - dP\,,\end{aligned}$$

where m is the maturation rate per juvenile prey, and a_I and a_A are the per-predator attack rate on adult and juvenile prey, respectively. McNair (1987) showed that the equilibrium tended to be stable when the difference in vulnerability increased, especially when the adult prey was less vulnerable ($a_A < a_I$).

Murdoch and Stewart-Oaten (1975) took into account that the prey in two patches had heterogeneous vulnerability and random migration between them. Let I and A be the number of prey in two patches with different attack rate a_I, and a_A, respectively, then the simplified version of his model is:

$$\begin{aligned}\frac{dI}{dt} &= rI + mA - mI - a_I IP \\ \frac{dA}{dt} &= rA + mI - mA - a_A AP \\ \frac{dP}{dt} &= eP(a_A A + a_I I) - dP\,,\end{aligned}$$

where r and m are the growth rate and migration rate, respectively.

The Lotka-Volterra model has been modified in many ways by considering other factors, such as time lags. A good review for these predator-prey models can be found in Murdoch, Briggs, and Nisbet (2003).

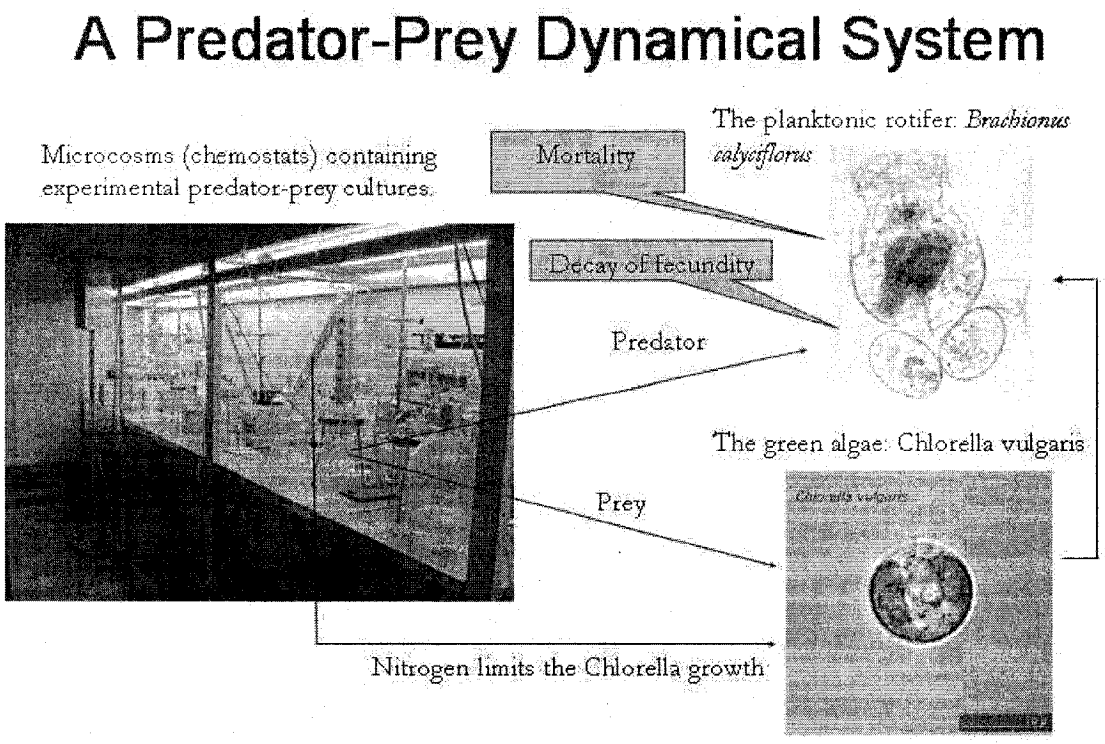

Figure 4.1: A diagram for a predator-prey dynamic system proposed by Fussmann et al. (2000).

Fussmann et al. (2000) studied the dynamic behavior of an aquatic laboratory community formed by two species. This is a predator-prey food chain (Figure 4.1), in which unicellular green algae, Chlorella vulagaris, are eaten by planktonic rotifers, Brachionus calyciflorus. Chlorella growth is also limited by nitrogen supply. In their experiment, Chlorella and Brachionus are deposited together in a

chemostats. Nitrogen continuously flows into the system with the concentration N_i at dilution rate δ, and all components are removed from the chemostats at the same rate δ. They provide a set of nonlinear differential equations to model the interactions between the planktonic rotifers, green algae, and the nitrogen resource. Let N, C, R, B be the concentrations of nitrogen, Chlorella, reproducing Brachionus, and total Brachionus, respectively. $F_C(N) = b_C N/(k_C + N)$, $F_B(C) = b_B C/(k_B + C)$ are two link functions, and ϵ, α, and m are the assimilation efficiency, the decay of fecundity, and the mortality of Brachionus, respectively. Their nonlinear DE's are

$$\begin{aligned}
\frac{dN}{dt} &= \delta(N_i - N) - F_C(N)C \\
\frac{dC}{dt} &= F_C(N)C - F_B(C)B/\epsilon - \delta C \\
\frac{dR}{dt} &= F_B(C)R - (\delta + m + \alpha)R \\
\frac{dB}{dt} &= F_B(C)R - (\delta + m)B\,.
\end{aligned} \tag{4.3}$$

Their model includes the mortality and decay of fecundity of Brachionus. They also introduce the nitrogen resource as a state variable, which can accurately model the uptake dynamics of the Chlorella population. But the concentrations of nitrogen and reproducing Brachionus are not measurable, and can be looked on as latent variables. This can bring some extra difficulty in estimating DE parameters from the experimental data, but one advantage of our method is to easily deal with missing variables, as discussed later.

Their model predicts correctly at a qualitative level three dynamic behaviors

of the experimental system. The predator and prey coexist at an equilibrium with the low nutrient supply (small δ or small N_i). Increasing N_i or δ switches the system to a limit cycle. The nitrogen input that is too low causes the extinction of the predator or both the predator and the prey.

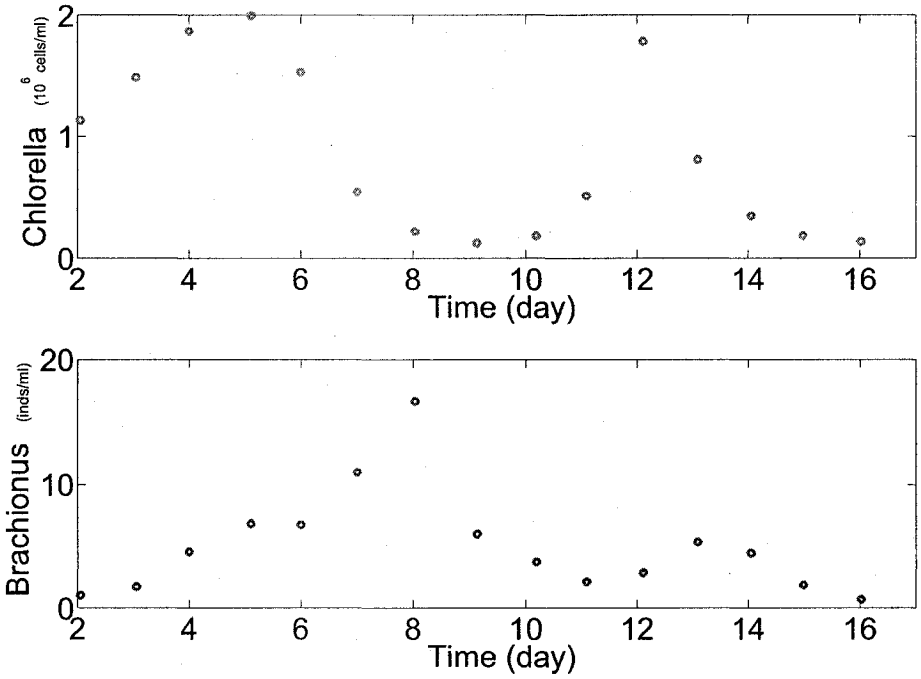

Figure 4.2: The concentration of Chlorella and Brachionus when the dilution rate $\delta = 0.68$ and the inflow Nitrogen concentration $N_i = 80$.

Fussmann has kindly offered us the data of the concentration of Chlorella and Brachionus under different experimental conditions, that is, with different values of δ and N_i. Figure 4.2 shows the oscillations of the Chlorella and Brachionus

populations when $\delta = 0.68$ and $N_i = 80$. Our goal is to estimate the parameter vector $\boldsymbol{\theta} = (\alpha, \epsilon, m, k_B, k_C, b_B, b_C)$ in Equations (4.3) from the noisy data. The results are shown in Section 4.9.

4.4 Introduction to an HIV Dynamic Model

HIV dynamic models, usually in the forms of DE's, describe the rate of population change of uninfected cells, infected cells and virus as a function of their populations and interactions. They have significantly contributed to our understanding of HIV infection and the development of antiviral drug therapy. Huang et al. (2005) proposed a set of nonlinear DE's to characterize the long-term HIV dynamics with antiretroviral therapy. Let U, I, and V be the number of uninfected cells, infected cells and free virus, respectively. Parameters α and β are the death rate of uninfected cells and infected cells, respectively, γ is the clearance rate of free virus, ρ is the infection rate, and ν is the rate at which uninfected cells are created from sources within the body, such as the thymus. Their DE's are simplified as follows:

$$\begin{aligned} \frac{d}{dt}U &= -\alpha \cdot U - \rho \cdot UV + \nu \\ \frac{d}{dt}I &= -\beta \cdot I + \rho \cdot UV \\ \frac{d}{dt}V &= -\gamma \cdot V + N \cdot \beta \cdot I\,. \end{aligned} \tag{4.4}$$

The first terms in the right sides of the three DE's take into account the death of uninfected and infected cells and the clearance of virus, respectively. The term

$\rho \cdot UV$ characterizes the infection of uninfected cells by the virus. This product term is based on the fact that the infection rate depends on not only the number of virus but also the number of uninfected cells. This makes sense because the more uninfected cells, the easier it is for the virus to infect an uninfected cell. The term $N \cdot \beta \cdot I$ quantifies the factor that each infected cell produces N new free virus during its life.

Figure 4.3 shows the HIV virus levels for 42 patients measured before treatment, and in around 1, 2, 4, 8, 12, 16, 20 and 24 weeks since treatment. These data are collected by AIDS Clinical Trials Group (Acosta et al. 2004). The detection limit of the viral load (HIV RNA copies) assay is 50 copies per ml blood. If it is below detectable, it is then imputed as 25 in the data set. The number of HIV virus for each patients shows different patterns. Some patients, such as Patient 42, have their number of virus decreasing all the time. But other patients, such as Patient 23, have their virus levels going down at the beginning and up after 4 weeks. The HIV virus level is a function of time, and we have 42 functional data $V_i(t)$, $i = 1, \cdots, 42$, in total. The two components in (4.4), the number of uninfected cells and infected cells are too noisy to be used for all patients. Our objective is to estimate the parameter vector $\boldsymbol{\theta} = (\alpha, \beta, \gamma, \rho, \nu, N)$ from the real data. The results are shown in Section 4.10.

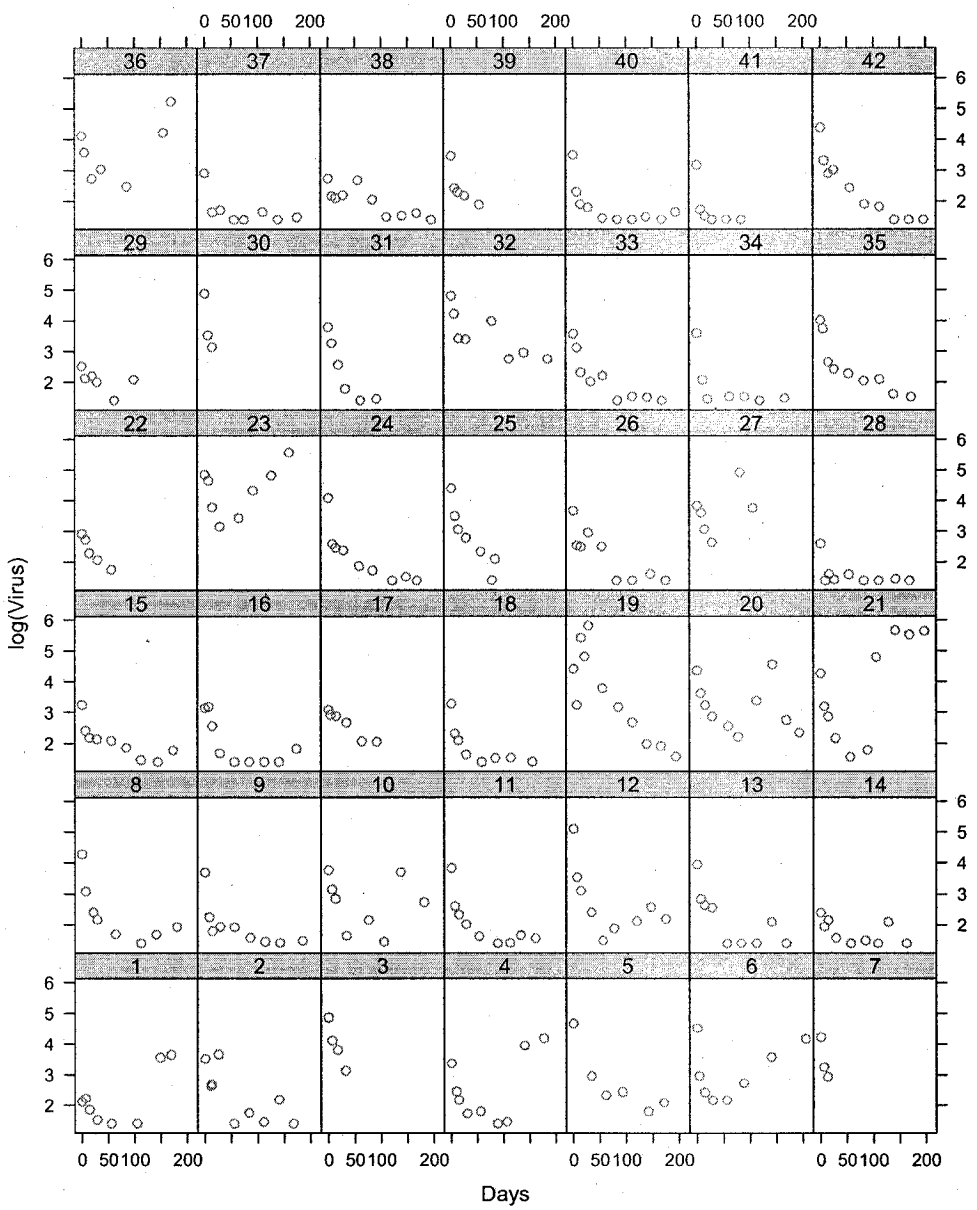

Figure 4.3: The number of free virus for 42 patients in logarithm scale.

4.5 Penalized Smoothing with the Penalty Defined by DE's

Section 2.1 mentions that the penalty term in penalized smoothing can be defined by DE's, leading to better estimates of fitting functions and their derivatives (Ramsay and Silverman 2005). This process is called *L-spline smoothing* by Gu (2002) and Ramsay and Silverman (2005), which is introduced in detail in this section.

Let $\mathbf{y} = (y(t_1), \cdots, y(t_n))$ be a vector of n observations, and the estimated fitting function be a linear expansion of K basis functions $\{\phi_k(t)\}_{k=1}^K$ as follows:

$$x(t) = \sum_k^K c_k \phi_k(t) = \mathbf{c}'\boldsymbol{\phi}(t)\,.$$

The basis system must have the capacity to approximate DE solutions, as well as derivatives involved in DE's. Most DE solutions have sharp features, such as peaks, valleys, high frequency oscillations and discontinuities in derivatives. The B-spline basis system can accommodate the discontinuities by assigning multiple knots to the critical locations (Ramsay and Silverman 2005). In practice, we can explore the DE solutions under initial estimates of parameters, and decide where we need to put many knots. Or we can begin with a very large number of equally spaced knots, and reduce knot density where appropriate. For instance, the cubic B-spline basis system with 400 equally spaced knots is found appropriate to approximate each component in the predator-prey DE's, because of the sharp change of the Chlorella concentration around the 12-th day (Figure 4.16).

The fitting function can be estimated by minimizing sum of squared errors (SSE), which can be written as:

$$\text{SSE} = \sum_{i=1}^{n} [y_i - x(t_i)]^2 .$$

To avoid over-fitting, nonparametric smoothing often requires a penalty term to penalize the roughness of the fitting function. For instance, in order to obtain a fitting function, the penalty term can be defined in term of the second derivative, that is,

$$\text{PEN}(x) = \int [D^2 x(t)]^2 dt .$$

When we require the estimated curve to satisfy a DE $Dx(t) = f(x|\boldsymbol{\theta})$, it is natural to define the penalty term with the differential operator $Lx(t) = Dx(t) - f(x|\boldsymbol{\theta})$:

$$\text{PEN}(x) = \int [Lx(t)]^2 dt , \tag{4.5}$$

and the fitting criterion to estimate the fitting function is given by

$$H(\mathbf{c}|\lambda, \mathbf{y}) = \sum_{i=1}^{n} [y(t_i) - x(t_i)]^2 + \lambda \int [Lx(t)]^2 dt . \tag{4.6}$$

When there are S DE's and M components observed, the fitting criterion

can be generalized to be:

$$
\begin{aligned}
H(\mathbf{c}|\lambda, \mathbf{y}) &= \sum_{j=1}^{M} \mathtt{SSE}_j + \sum_{d=1}^{S} \lambda_d \mathtt{PEN}_d \qquad (4.7)\\
&= \sum_{j=1}^{M} \omega_j \sum_{i=1}^{n} [y_j(t_i) - x_j(t_i)]^2 + \sum_{d=1}^{S} \lambda_d \omega_d \int [L_d \mathbf{x}(t)]^2 dt\,,
\end{aligned}
$$

where $y_j(t_i)$ is the observation for j-th component at t_i and $\mathbf{x}(t) = (x_1(t), \cdots, x_T(t))$ is a vector of fitting functions for the total T components. Sometimes T is larger than M, which means there are some unobservable components. The differential operator $L_d\mathbf{x}(t) = Dx_d(t) - f_d(\mathbf{x}|\boldsymbol{\theta})$ is defined by the d-th DE: $Dx_d(t) = f_d(\mathbf{x}|\boldsymbol{\theta})$. Parameter ω_j is the normalizing weight in order to keep different components having comparable scales for $\mathtt{SSE}_j$ and $\mathtt{PEN}_j$. In practice, ω_j can be the reciprocal of the initial value, $\omega_j = 1/x_j(0)$, or the reciprocal of variance of observations, $\omega_j = 1/\mathtt{Var}(x_j)$. When some components are not observable, ω_j can also be the reciprocal of variance of the initial estimate of the DE solution for the j-th component. The smoothing parameter λ_d controls the trade off between fitting to data and fidelity to DE's, and we discuss the selection of λ_d in the following section.

For simplicity of notation, we assume that the dynamic system is composed of one single component, i.e. $T = S = 1$. Let L be a homogenous linear differential operator of order m

$$
Lx(t) = \sum_{j=0}^{m-1} \beta_j(t) D^j x(t) + D^m x(t)\,,
$$

then we can minimize the fitting criterion $H(\mathbf{c}|\lambda, Y)$ and derive the analytical form

of the coefficient vector $\mathbf{c}$ as:

$$\hat{\mathbf{c}}(\lambda, Y) = [\mathbf{\Phi}'\mathbf{\Phi} + \lambda\mathbf{R}]^{-1}\mathbf{\Phi}'\mathbf{y}\,, \tag{4.8}$$

where $\mathbf{R} = \int [L\boldsymbol{\phi}(t)][L\boldsymbol{\phi}(t)]'dt$ is a $K \times K$ matrix, and $\mathbf{\Phi}$ is an $n \times K$ matrix with the jk-th element $\Phi_{jk} = \phi_k(t_j)$.

When L is a nonlinear differential operator, we have to approximate the penalty term (4.5) as

$$\texttt{PEN}(\mathbf{x}) \approx \sum_{q}^{Q} v_q [L(x(t_q))]^2\,, \tag{4.9}$$

where t_q is a quadrature point and v_q is the corresponding quadrature weight. Let ξ_ℓ be the unique knot location, the evaluation points t_q can be chosen by dividing each interval $[\xi_\ell, \xi_{\ell+1}]$ into the odd number of equal-sized intervals, say r, and the quadrature weight $v_q = [1, 4, 2, 4, \cdots, 2, 4, 1](\xi_{\ell+1} - \xi_\ell)/5$ from Simpson's rule. In our experience, the integrals can be satisfactorily approximated when $r = 5$. In practice, the total quadrature points and weights along with the corresponding basis function values can be saved at the beginning of the computation in order to save computation time. The speed of computation can be further improved by using the sparse matrix methods in Matlab if a B-spline basis system is used.

4.6 Optimizing Smoothing Parameter λ

From the fitting criterion (4.7), we can see that it is very important to choose a proper value for the smoothing parameters in L-spline smoothing. Figures 4.4,

4.5, and 4.6 show the fitting functions under different scales of smoothing parameters on the same simulated data generated by adding noise to the Predator-Prey DE's. When the smoothing parameter is too small, the fitting functions tend to be rough (Figure 4.4). On the other hand, the fitting function is far from data if the smoothing parameter is too large, since there is too much weight on the roughness penalty and the fitting function is forced to be very smooth. Figure 4.6 shows that there is a large difference between the fitting function and true curve (DE solutions) over the range [0, 5] when the smoothing parameter $\lambda = 10^8$. We can only obtain a good fitting function with a moderate smoothing parameter value. The fitting function shown in Figure 4.5 when $\lambda \approx 32$ can approximate the true curve almost exactly. In the following, generalized cross-validation (GCV) and Stein's unbiased risk estimate (SURE) are shown to be good criteria to find the optimal value of the smoothing parameter, which minimizes mean square errors (MSE) of fitting functions and true curves.

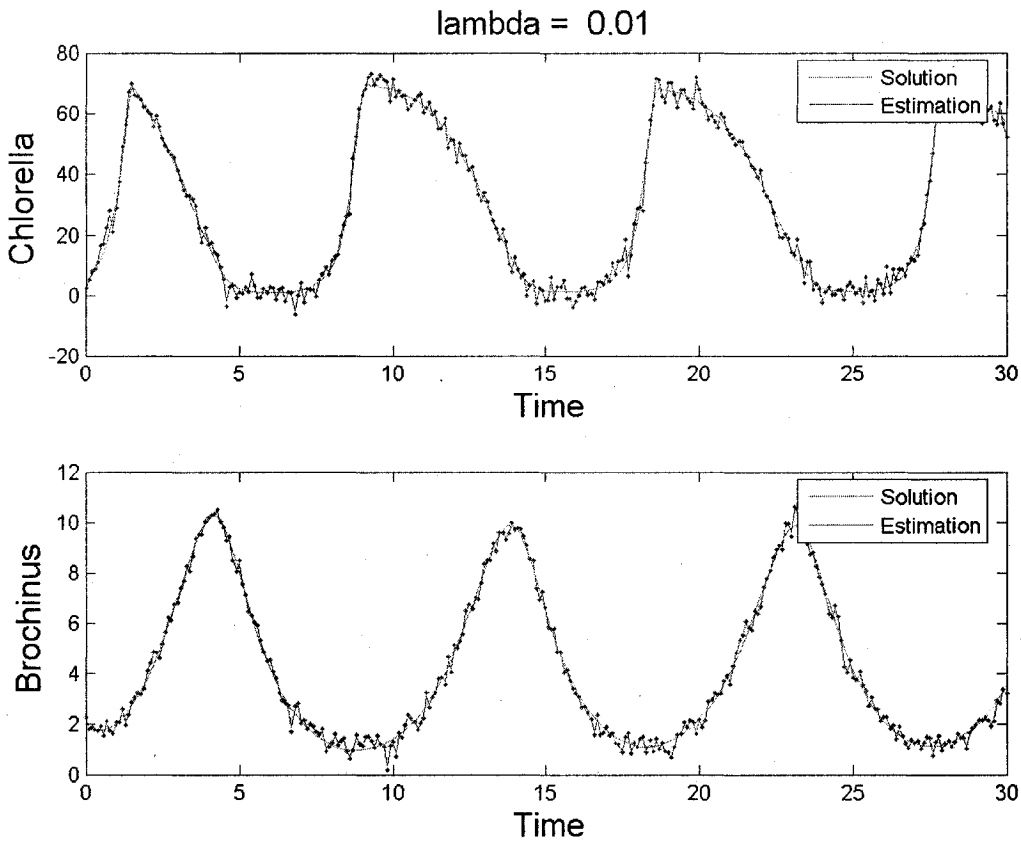

Figure 4.4: Smoothing data when the smoothing parameter $\lambda = 0.01$

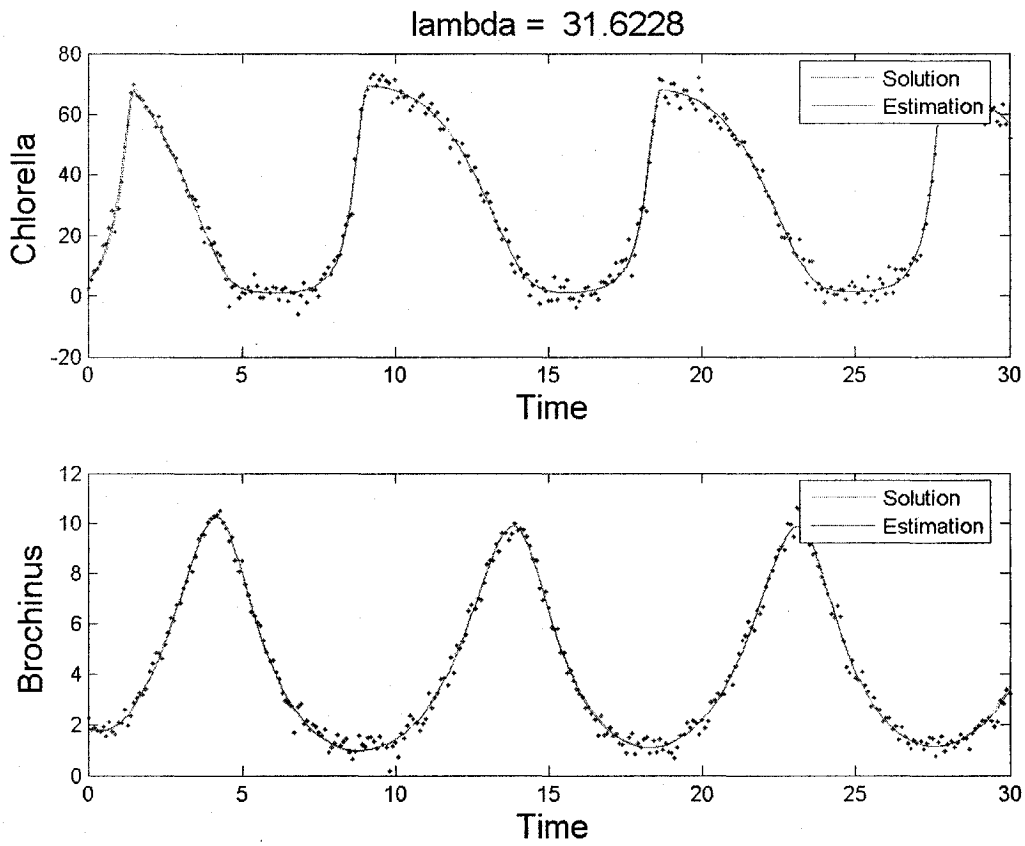

Figure 4.5: Smoothing data when the smoothing parameter $\lambda \approx 32$

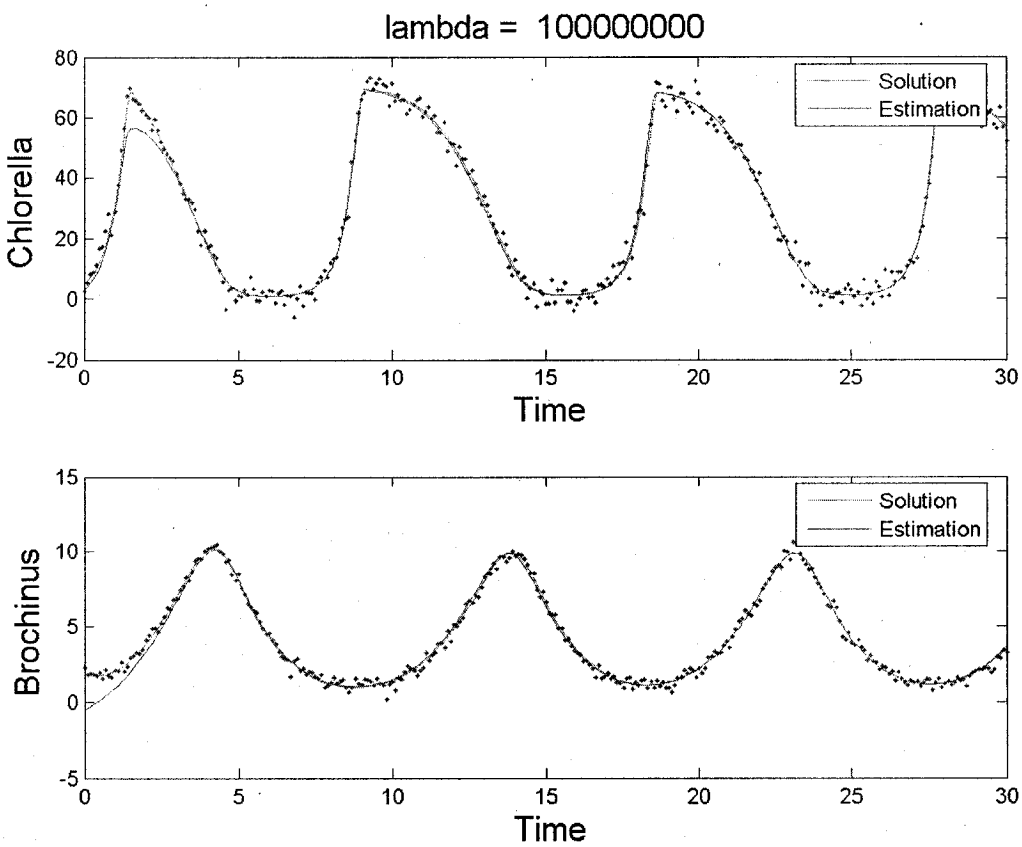

Figure 4.6: Smoothing data when the smoothing parameter $\lambda = 10^8$

4.6.1 Optimizing λ by Generalized Cross-Validation

When the differential operator L is linear, the optimal smoothing parameter $\hat{\lambda}$ is chosen by minimizing GCV, which can be written as follows:

$$\mathtt{GCV}(\lambda) = \left[\frac{n}{\mathtt{dfe}(\lambda)}\right]\left[\frac{\mathtt{SSE}(\lambda)}{\mathtt{dfe}(\lambda)}\right], \tag{4.10}$$

where degrees of freedom measure $dfe(\lambda)$ are

$$\mathtt{dfe}(\lambda) = n - tr[\mathbf{\Phi}(\mathbf{\Phi}'\mathbf{\Phi} + \lambda\mathbf{R})^{-1}\mathbf{\Phi}'].$$

Chapter 2 shows how to use the Newton-Raphson algorithm to find the optimal smoothing parameter $\hat{\lambda}$, where this is called nonadaptive penalized smoothing.

When the differential operator L is nonlinear, we can not get the expression for the coefficients $\mathbf{c}$ explicitly. However, GCV can be approximated by replacing $\mathbf{R}$ by

$$\hat{\mathbf{R}} = \int (\frac{\partial Lx(t)}{\partial \mathbf{c}})(\frac{\partial Lx(t)}{\partial \mathbf{c}})' dt. \tag{4.11}$$

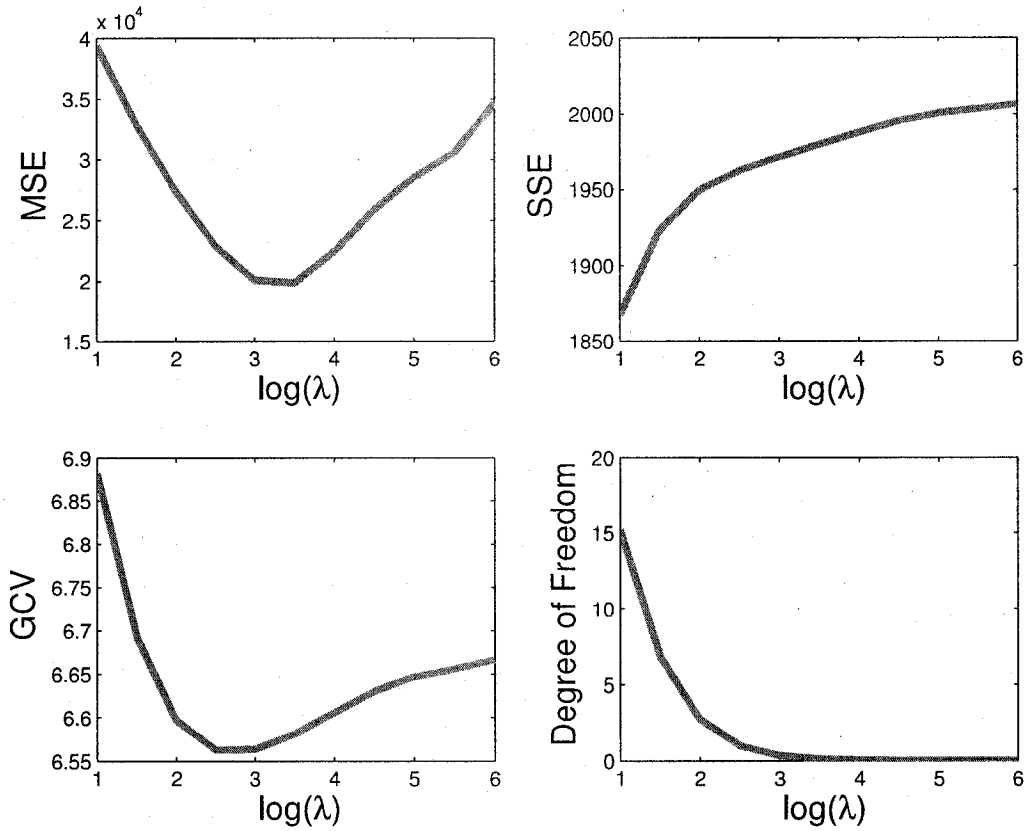

Figure 4.7: Choosing Smoothing Parameters by locally linearized GCV for Predator-Prey DE's. MSE is the sum squared errors between DE solutions (true curves) and fitting functions. SSE is the sum squared errors of the fitting functions.

Figure 4.7 shows simulation results on the nonlinear predator-prey DE's (4.3), with the simulated data sets shown in Figure 4.5. SSE is an increasing function of the smoothing parameter, because the small smoothing parameters put large weight on fitting data. MSE is the sum squared errors between DE solu-

tions (true curves) and fitting functions, which is minimized when the smoothing parameter λ is around 10^3. The locally linearized GCV calculated with (4.11) is also optimized at the approximate value of the smoothing parameter.

4.6.2 Optimizing λ by Minimizing Stein's Unbiased Risk Estimate

Whenever the differential operator L is linear or nonlinear, Stein's unbiased risk estimate (SURE) for total prediction error (Stein 1981) is convenient to use as the criterion for smoothing parameter selection. When observations are normally distributed, $\mathbf{y}(t_i) \sim N(\mathbf{x}(t_i), \sigma^2 I)$, SURE for total prediction error (TPE) is:

$$\text{TPE} = \text{SSE} + 2\sigma^2 \sum_{i=1}^{n} \frac{\partial \mathbf{x}(t_i)}{\partial \mathbf{y}(t_i)}. \tag{4.12}$$

According (A.2) in Appendix A, we can calculate the first derivative of the coefficient vector $\mathbf{c}$ with respect to the data vector $\mathbf{y}$ as:

$$\frac{\partial \mathbf{c}}{\partial \mathbf{y}} = -\left(\frac{\partial^2 H}{\partial \mathbf{c}^2}\right)^{-1} \frac{\partial^2 H}{\partial \mathbf{c} \partial \mathbf{y}},$$

where H is the fitting criterion (4.6). Then we can attain TPE (4.12) with the second term calculated by

$$\frac{\partial \mathbf{x}(t)}{\partial \mathbf{y}} = \boldsymbol{\phi}(t) \frac{\partial \mathbf{c}}{\partial \mathbf{y}}. \tag{4.13}$$

When we apply the generalized profiling method to estimate the variance of

DE parameters $\boldsymbol{\theta}$, $\partial \mathbf{c}/\partial \mathbf{y}$ has to be calculated first. Therefore, it is free to calculate TPE for smoothing parameter selections.

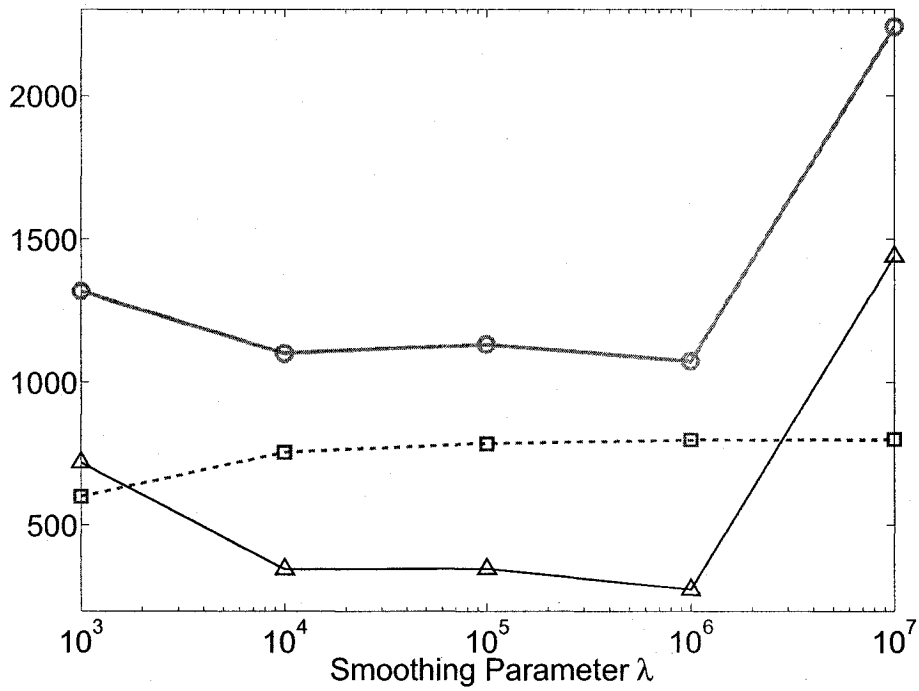

Figure 4.8: Stein's unbiased risk estimate for total prediction error (red circles) when smoothing HIV data with HIV DE's (4.4). The blue rectangles are SSE of fitting functions. The black triangles are their difference, or $2\sigma^2 \sum_{i=1}^{n} \frac{\partial \mathbf{x}(t_i)}{\partial \mathbf{y}(t_i)}$.

We smoothed HIV data (Figure 4.3) with HIV DE's (4.4) using different values of the smoothing parameter λ. The corresponding SURE for TPE is shown in Figure 4.8. SSE is an increasing function of the smoothing parameter, because the large smoothing parameters tend to put less weight on fitting data. TPE is

minimized when the smoothing parameter $\lambda = 10^6$, which is the optimal value to smooth HIV data.

4.7 Estimating DE's with Generalized Profiling Method

Section 4.5 and 4.6 show that the fitting functions can be estimated by the L-spline smoothing, and the smoothing parameter can be optimized by GCV or SURE. In the following, we introduce how to estimate the DE parameter vector $\boldsymbol{\theta}$ from noisy data. A byproduct is that we can estimate initial values for DE components, which is shown in Section 4.7.2. Section 4.7.3 explores the effect of smoothing parameter on DE parameter estimates and Section 4.7.4 discusses the smoothing parameter selection. When the coefficient vector is viewed as functions of DE parameters, the likelihood surface can become smooth, as discussed in Section 4.7.5. Section 4.7.6 investigates the effect of data noise, data resolution and flexibility of basis systems on DE parameter estimates.

Let $y_j(t_i)$ be the observation for the j-th component in the dynamic system at t_i, $i = 1, \cdots, n_j$ and $j = 1, \cdots, M$. All M components can be observed at different time points from each other, and $x_j(t)$ is the corresponding fitting function by L-spline smoothing for the j-th component, which is a linear expansion of K_j basis

functions $\{\phi_{ik}(t)\}_{k=1}^{K_j}$:

$$x_j(t) = \sum_{k}^{K_j} c_k \phi_{ik}(t) = \mathbf{c}_j' \boldsymbol{\phi}_j(t)\,,$$

where $\boldsymbol{\phi}_j(t)$ is a vector of basis functions for the j-th component, and $\mathbf{c}_j$ is the corresponding coefficient vector. The coefficient vector $\mathbf{c}$ is denoted as a vector of all M coefficient vectors, i.e., $\mathbf{c} = (\mathbf{c}_1', \cdots, \mathbf{c}_T')'$. For a fixed value of $\boldsymbol{\theta}$, when we penalized smooth data with the penalty term defined by DE's (4.1), the coefficient vector $\mathbf{c}$ can be estimated by minimizing the criterion $H(\mathbf{c}|\lambda, \mathbf{y})$ in (4.7). In other words, the coefficient vector $\mathbf{c}$ can be treated as a function of DE parameter vector $\boldsymbol{\theta}$. This function $\mathbf{c}(\boldsymbol{\theta})$ is explicit if the DE's (4.1) are linear, given by (4.8). When the DE's (4.1) are nonlinear, the function $\mathbf{c}(\boldsymbol{\theta})$ is implicit.

In both cases, we can obtain the estimate and sampling variance of the DE parameter vector $\boldsymbol{\theta}$ with the generalized profiling method introduced in Chapter 2. The coefficient vector $\mathbf{c}$ is the nuisance parameter, and the DE parameter vector $\boldsymbol{\theta}$ is the structural parameter. The inner optimization criterion is $H(\mathbf{c}|\lambda, \mathbf{y})$ defined in (4.7), and the outer optimization criterion is sum of squared errors for all observed components:

$$\begin{aligned} F(\boldsymbol{\theta}|\lambda, \mathbf{y}) &= \sum_{j=1}^{M} \mathrm{SSE}_j \\ &= \sum_{j=1}^{M} \omega_j \sum_{i=1}^{n} [y_j(t_i) - x_j(t_i)]^2\,, \end{aligned} \tag{4.14}$$

where the notations have the same definitions as (4.7). We call this method as Pro-

filing PDA, PDA being the abbreviation of principle differential analysis (Ramsay and Silverman 2005).

4.7.1 Estimating Initial Values of Components in DE's

Numerically solving DE's relies on initial values, which are the values of DE components at the first time point. A small change in initial values results in a large difference in the numerical DE solutions. However, observations in real life, including the observed initial values, usually have some measurement error, and it is dangerous to use the first observations as the initial values directly. Moreover, some components in DE's are not observable, in which case there is no way to observe the initial values for these components.

The byproduct of Profiling PDA is that we have fitting functions for all components after we derive the DE parameter estimate $\hat{\boldsymbol{\theta}}$. We can then estimate initial values by evaluating the fitting functions for all components at the first time point. We show that the DE solutions can fit data better with the estimated initial values for all components when we estimate parameters in the predator-prey DE's and HIV DE's.

4.7.2 Effect of Smoothing Parameter λ when Estimating DE's

The smoothing parameter λ controls the trade off between fitting to data and fidelity to DE's in the inner criterion (4.7), which implicitly controls the functional relationship between the coefficient vector $\mathbf{c}$ and the DE parameter vector $\boldsymbol{\theta}$. So the smoothing parameter also has a large effect on the DE parameter estimates. In the following, we explore the smoothing parameter effect on DE parameter estimates with simulation.

Each simulated data set is generated by adding Gaussian noise to the Predator-Prey DE solutions, with one typical simulated data sets shown in Figure 4.5. With parameters k_C, k_B, b_C, and b_B fixed, the other parameters ϵ, α, and m are estimated from 100 such simulated data sets when the smoothing parameter λ is 10, 10^2, 10^3, 10^4, 10^5 and 10^6. Each component is approximated by two contrastive B-splines basis, respectively. One B-spline basis system is generated by putting one knot on each time point with observations, which we call *Setting 1* in this and next sections. The other B-spline basis system is generated by doubling the number of knots of Setting 1, which we call *Setting 2* in this and next sections. The knots in both settings are equally spaced. The boxplot for estimates of α under Setting 1 is shown in Figure 4.9. A large smoothing parameter value, such as 10^6, leads to a large bias and small variance of estimated α's. On the other hand, the estimated α's have a small bias and large variance with a small smoothing parameter value, such as 10. Estimates of ϵ and m under Setting 1 show the same property as estimates of α.

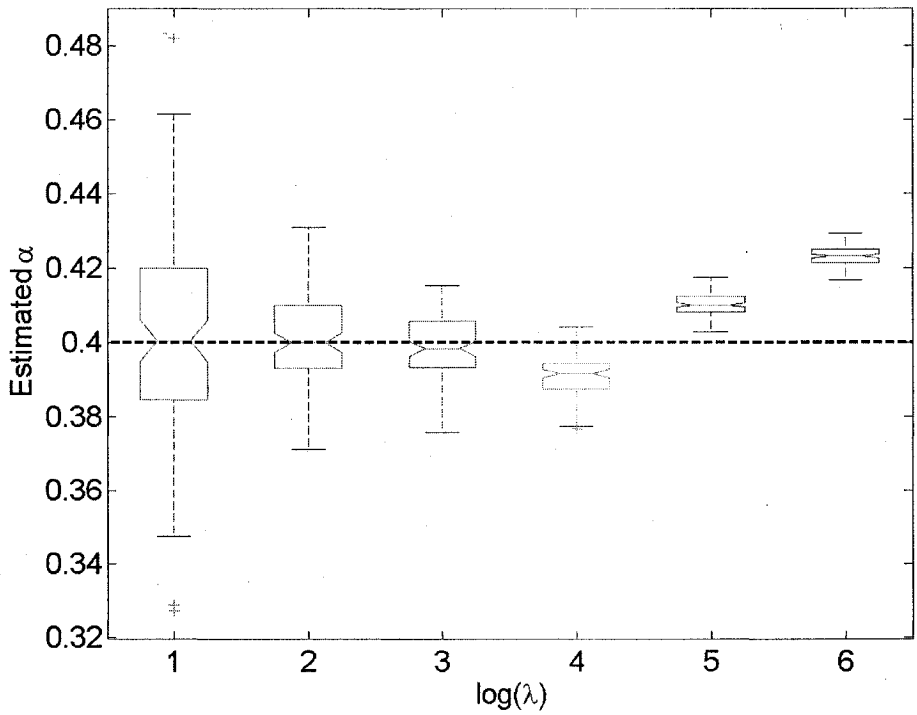

Figure 4.9: The boxplot of the estimated α's in the predator-prey DE's from simulated data under different smoothing parameters when each component is approximated by cubic B-splines with the same number of equally-spaced knots as the number of observations. The dashed line in the boxplot is the true value.

DE's often have solutions with high curvatures. As a result, the basis system sometimes does not have enough flexibility to approximate DE solutions satisfactorily. In particular, it cannot approximate the derivatives of DE solutions well. As a result, the estimate for the penalty term (4.5) in L-spline smoothing brings a

large bias. As the smoothing parameter becomes large, this kind of bias is magnified. This is one reason that we cannot choose a smoothing parameter that is too large. When the basis system is more flexible, it is more possible to approximate DE solutions well, and the optimal value of the smoothing parameter is larger. In this and next sections, we assume the optimal value of the smoothing parameter as the one minimizing MSE between DE parameter estimates and real parameter values.

For instance, we estimate the parameters ϵ, α, and m with the more flexible B-spline basis system under Setting 2, fixing the other parameters k_C, k_B, b_C, and b_B, when the smoothing parameter λ is 10, 10^2, 10^3, 10^4, 10^5 and 10^6. The boxplot for estimates of α under Setting 2 is shown in Figure 4.10. The bias becomes much smaller than before when the smoothing parameter λ is large.

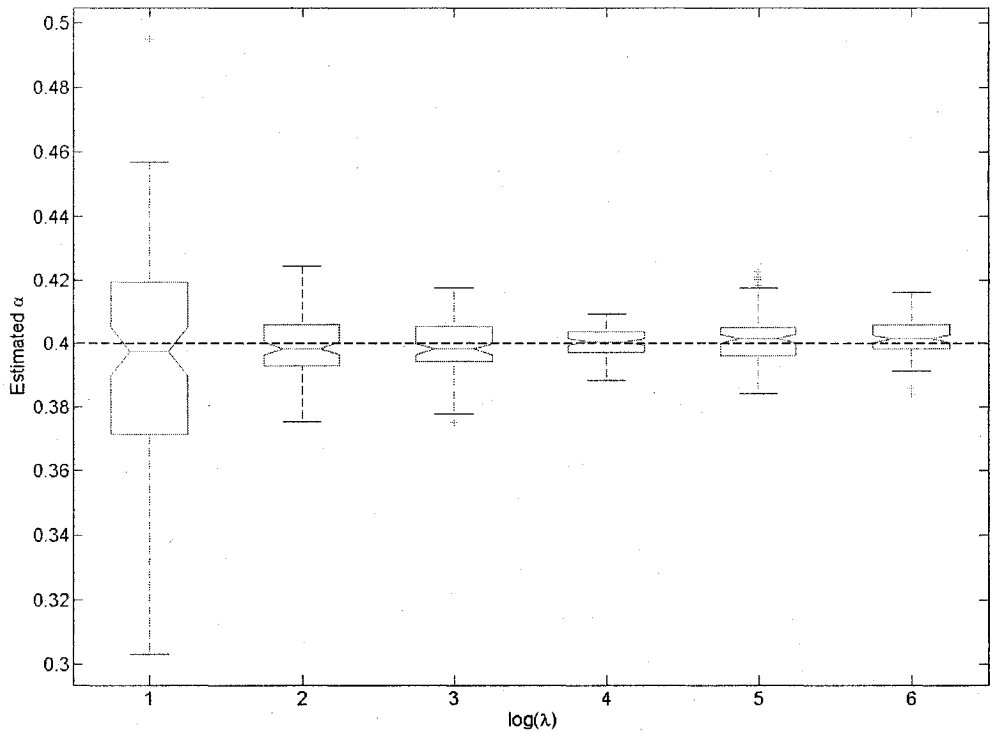

Figure 4.10: The boxplot of the estimated α's in the predator-prey DE's from simulated data under different smoothing parameters when each component is approximated by cubic B-splines with the number of equally-spaced knots doubling the number of observations. The dashed line in the boxplot is the true value.

4.7.3 Optimizing Smoothing Parameter λ when Estimating DE's

The previous section shows that the smoothing parameter controls the biases and variances of parameter estimates. The optimal smoothing parameter should be larger with a more flexible basis system. We first discuss where the biases of parameter estimates come from, and then show that GCV can give some clues to choose the optimal smoothing parameter. With a more powerful basis system, GCV also tends to choose a larger smoothing parameters.

We first define some notation. For the j-th component among the M observed components, let $x_j(t_i)$ be its observation at time t_i, and $x_j^S(t)$ is the corresponding fitting function by L-spline smoothing, and $x_j^D(t)$ is the corresponding DE solution with the estimated initial values $x_i^S(t_1)$. Then the biases of Profiling PDA parameter estimates come from replacing weighted mean squared errors between observations and DE solutions

$$\texttt{MSE}_{OD} = \frac{1}{nM}\sum_{j=1}^{M}\{\omega_j \sum_{i=1}^{n}[x_j(t_i) - x_j^D(t_i)]^2\} \tag{4.15}$$

by weighted mean squared errors between observations and fitting functions

$$\texttt{MSE}_{OS} = \frac{1}{nM}\sum_{j=1}^{M}\{\omega_j \sum_{i=1}^{n}[x_j(t_i) - x_j^S(t_i)]^2\}$$

in the outer optimization. The weighted mean squared errors between fitting

functions and DE solutions

$$\texttt{MSE}_{SD} = \frac{1}{nM}\sum_{j=1}^{M}\{\omega_j \sum_{i=1}^{n}[x_j^S(t_i) - x_j^D(t_i)]^2\}$$

is approximately their difference, i.e.,

$$\texttt{MSE}_{SD} \approx \texttt{MSE}_{OD} - \texttt{MSE}_{OS}\,.$$

Therefore, a good value of smoothing parameter λ with a neglectable $\texttt{MSE}_{SD}$ leads to small biases of parameter estimates.

In the rest of this section, we do the simulations for the predator-prey DE's on the simulated data set shown in Figure 4.5. When $\lambda > 10^4$, $\texttt{MSE}_{SD}$ can be neglected, and

$$\texttt{MSE}_{OD} \approx \texttt{MSE}_{OS}\,,$$

as shown in Figure 4.11. We have already shown that the smoothing parameter can be selected by minimizing GCV or SURE to obtain the minimum $\texttt{MSE}_{OS}$, which thus can also minimize the $\texttt{MSE}_{OD}$. Figure 4.12 shows that GCV is a good criterion to find the optimal smoothing parameter value that minimizes $\texttt{MSE}_{OD}$. Both of them show the similar pattern and are minimized at the same smoothing parameter values.

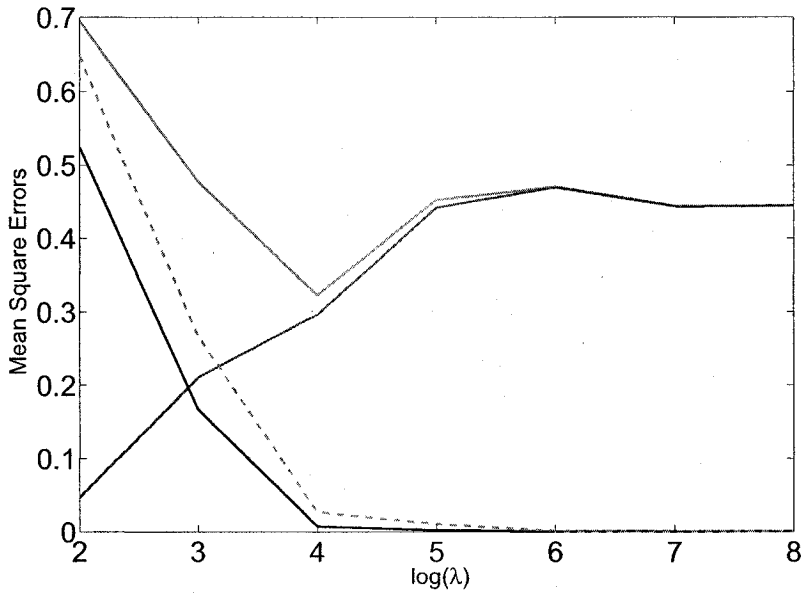

Figure 4.11: MSE_{OD} (red solid line), MSE_{OS} (blue dashed line), and MSE_{SD} (black solid line) curve changing with the log smoothing parameter in Profiling PDA estimates for the Predator-Prey Equations (4.3) from Fussmann's data. The red dashed line is the difference $\texttt{MSE}_{OD} - \texttt{MSE}_{OS}$. Each component is approximated by cubic B-splines with the number of equally-spaced knots same as the number of observations.

However, the optimal smoothing parameter minimizing $\texttt{MSE}_{OD}$ does not necessarily minimize MSE of DE parameter estimates, as shown in Figure 4.12. $\texttt{MSE}_{OD}$ is minimized when the smoothing parameter $\lambda = 10$, while the optimal value of the smoothing parameter to minimize MSE of DE parameter estimates is 10^3.

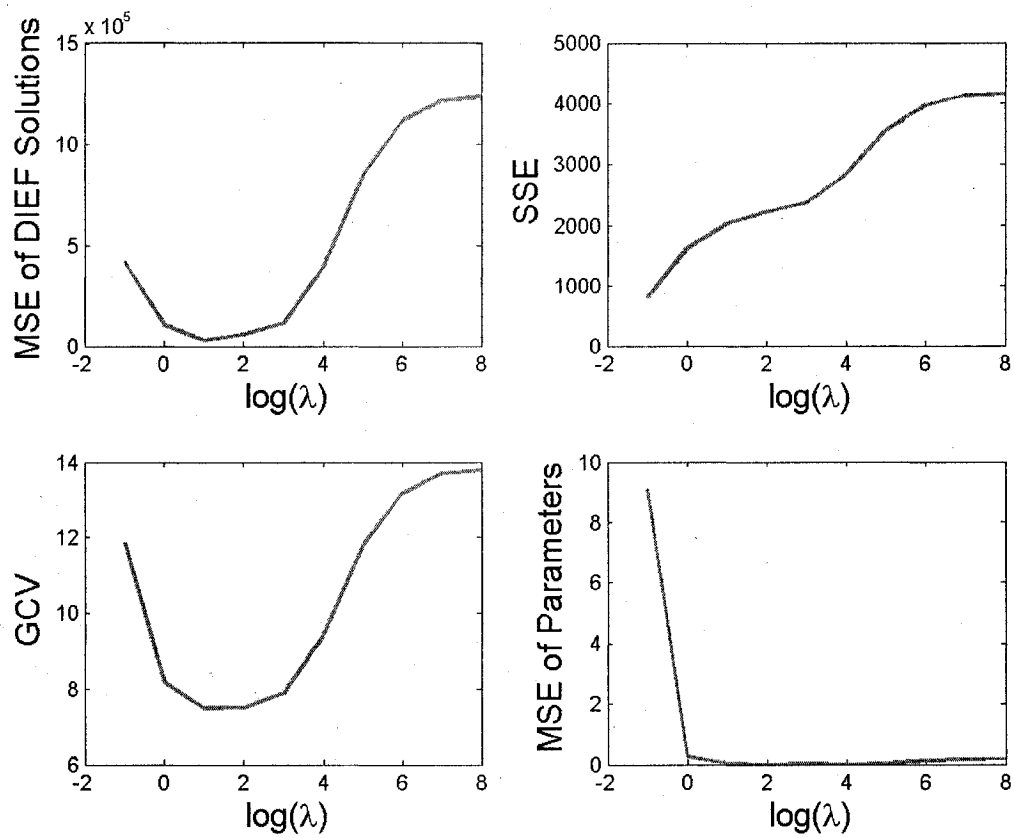

Figure 4.12: The left top panel displays MSE of solutions for the predator-prey DE's, the right panel shows SSE of fitting functions, the left bottom panel displays GCV and the right bottom panel displays MSE of DE parameter estimates in the predator-prey DE's from simulated data. Each component is approximated by cubic B-splines with the number of equally-spaced knots same as the number of observations.

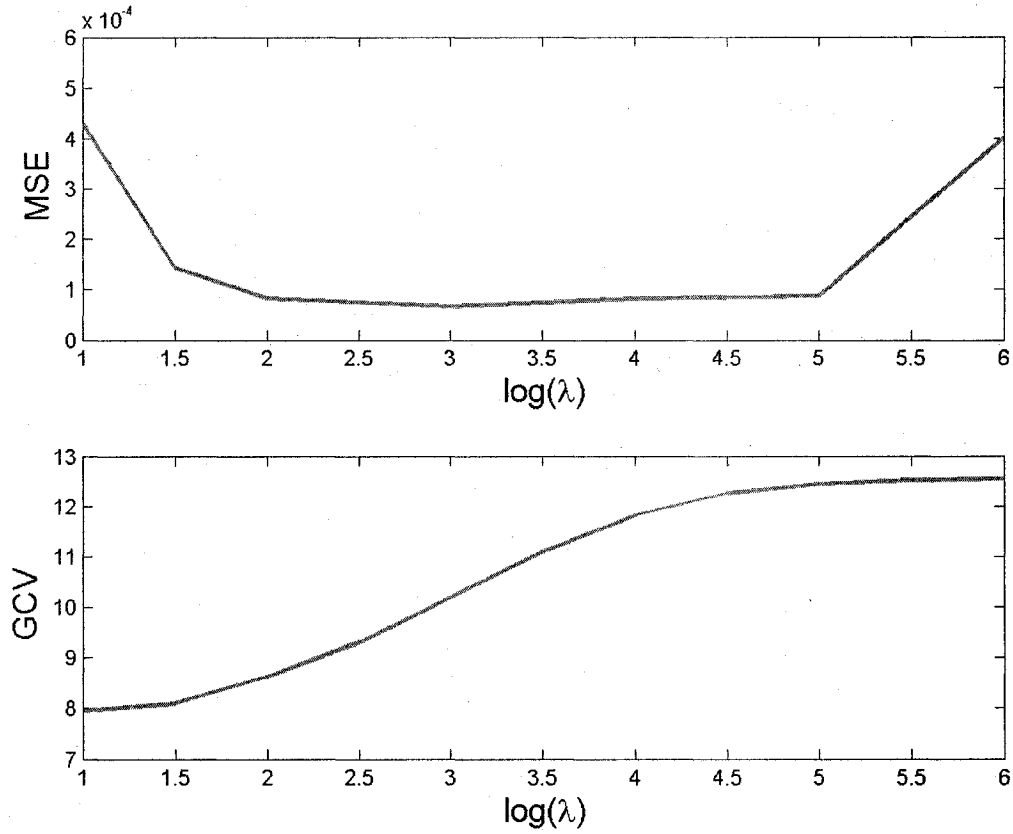

Figure 4.13: MSE of DE parameter estimates in the predator-prey DE's from simulated data and GCV under different values of smoothing parameters λ when each component is approximated by cubic B-splines with the number of equally-spaced knots same as the number of observations.

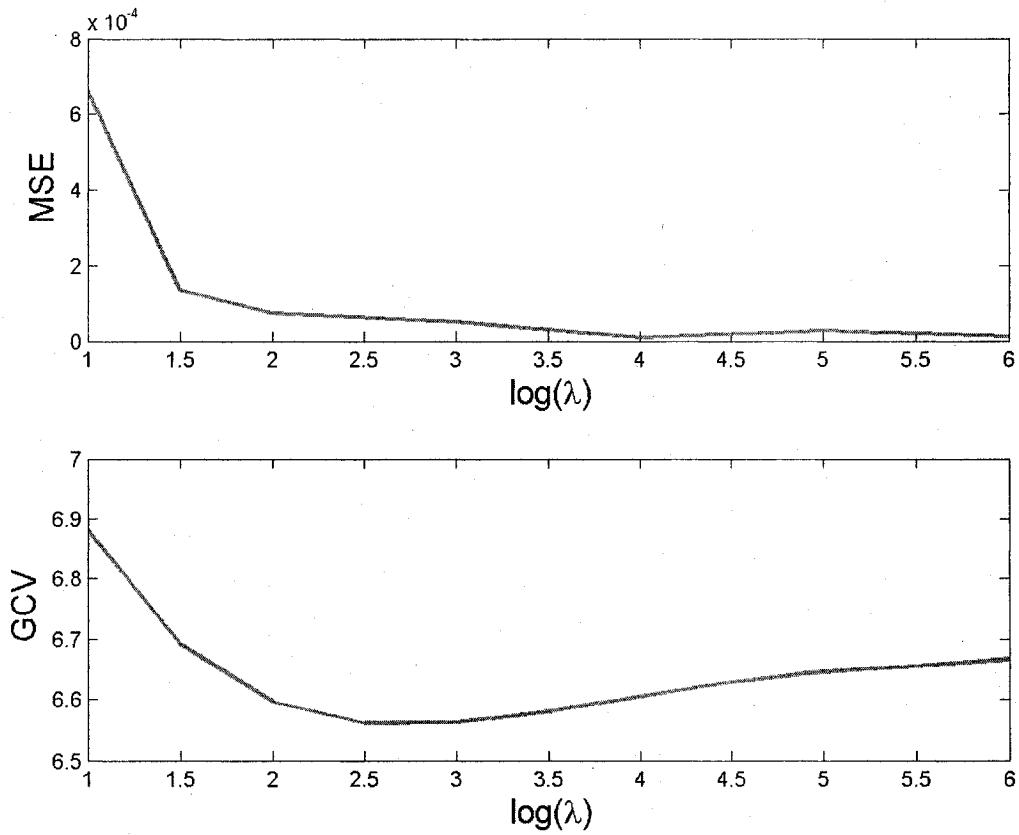

Figure 4.14: MSE of DE parameter estimates in the predator-prey DE's from simulated data and GCV under different values of smoothing parameters λ when each component is approximated by cubic B-splines with the number of equally-spaced knots doubling the number of observations.

Figure 4.13 displays MSE of DE parameter estimates and GCV under different smoothing parameters in Setting 1. GCV does not give the optimal value of the

smoothing parameter, however, we can see that GCV does give us some clues about the optimal smoothing parameter value in the sense that MSE is almost same when $1.5 \leq \lambda \leq 5$, and GCV is still not large when λ is around 1.5. In Setting 2, we double the number of knots of Setting 1, which implicitly increases the basis approximation ability. Figure 4.14 shows MSE of DE parameter estimates and GCV under different smoothing parameters in Setting 2. The optimal smoothing parameter value minimizing MSE of DE parameter estimates increases from 10^3 to 10^4 (Figure 4.14) when the basis system becomes more flexible. The corresponding optimal smoothing parameter value minimizing GCV also increases from 10 to $10^{2.5}$, although it still does not reach the optimal value minimizing MSE of DE parameter estimates, either.

4.7.4 Optimization Surface when Estimating DE's

We generate the simulated data by adding Gaussian noise with $SD_C = 3$, $SD_B = 0.3$ to Predator-Prey DE solutions for Chlorella and Brachionus, respectively, with the same sampling time points as the real data shown in Figure 4.2. The scale of noise is selected such that coefficients of variance of simulated data for Chlorella and Brachionus are around same. Figure 4.15 displays SSE surface of the fitting function to simulated noisy data when changing the values of parameters ϵ and α in (4.3) and fixing the values of the other parameters under three different values of the smoothing parameter. When the smoothing parameter is small, the SSE surface is flatter, which allows for finding the global minimum. So a small value of smoothing parameter leads to the small biases and large sampling variances of parameter estimates. When the smoothing parameter increases, the SSE surface is

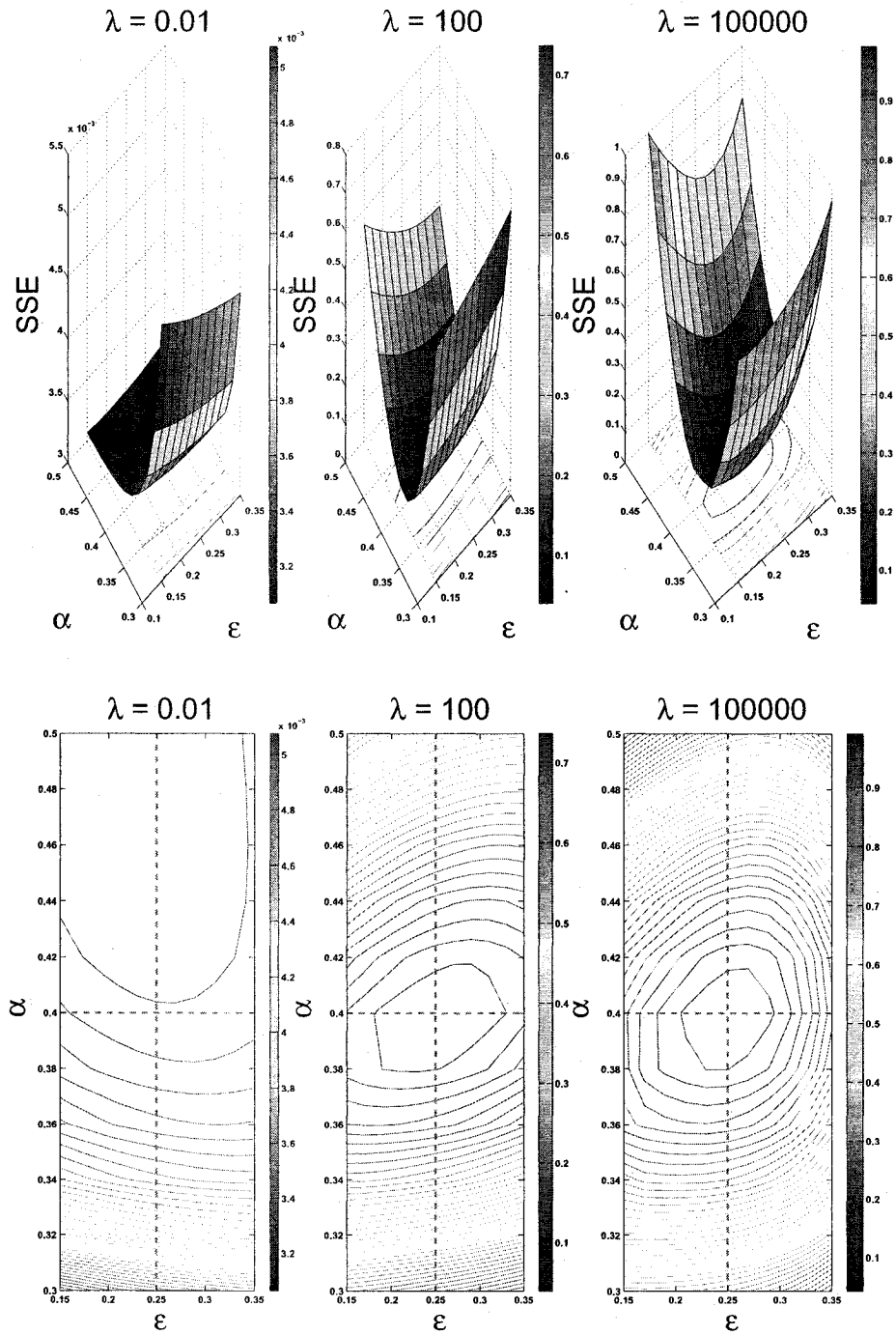

Figure 4.15: SSE surface of the spline fit to simulated noise data as DE parameters ϵ and α are varied and the others are fixed under three different scales of smoothing parameters. The bottom three graphs are the corresponding contours.

steeper. The SSE surface with a large smoothing parameter seems to be convex. However some other DE's have been found to have a rough SSE surfaces in Ramsay et al. (2005). This can also explain why large smoothing parameters lead to large biases and small sampling variances of parameter estimates. In practice, we can start with a small smoothing parameter value and obtain the DE parameter estimates. The obtained DE parameter estimates are updated by increasing the smoothing parameter, in order to find global optimal DE parameter estimates with small sampling variances.

4.7.5 Estimate DE's from Simulated Data

We estimate the parameter vector $\boldsymbol{\theta} = (\epsilon, \alpha, m, b_C, b_B, k_C, k_B)$ in (4.3) on 100 simulated data sets. The simulated data are generated by adding Gaussian noise with $SD_C = 3$, $SD_B = 0.3$ to Predator-Prey DE solutions for Chlorella and Brachionus, respectively, with two observations per day. The scale of noise is selected such that coefficients of variance of simulated data for Chlorella and Brachionus are around same. Figure 4.16 shows a typical set of simulated data. From this figure, we can see that the Chlorella solution has a very large curvature around the 12th day, which makes it challenging to estimate the correct curve. Moreover, the data resolution is small. For example we only have three observations in the interval [11.5, 12.5], which go through most of the range of Chlorella.

To investigate the effect of data noise, data resolution and flexibility of basis systems on parameter estimates, we set up 4 contrastive simulation experiments. Data are simulated by adding Gaussian noise with standard deviations SD_C, SD_B

to Predator-Prey DE solutions of Chlorella and Brachionus, respectively, with n observations per day. All four components are approximated by cubic B-splines with K equally spaced knots (Table 4.1). The smoothing parameter is chosen as $\lambda = 10^3$, which minimizes the locally linearized GCV as discussed in Section 4.6.1.

Table 4.1: Settings for 4 contrastive simulation experiments when estimating parameters in the predator-prey DE's.

Setting	SD_C	SD_B	n/Day	K
Setting 1	3	0.3	2	100
Setting 2	6	0.6	2	100
Setting 3	3	0.3	1	100
Setting 4	3	0.3	2	200

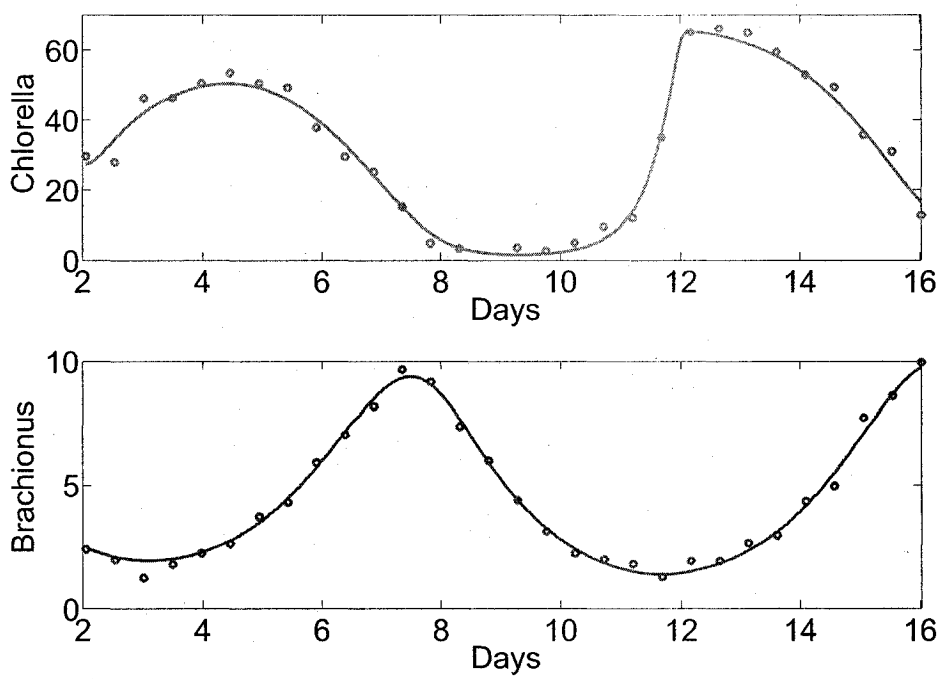

Figure 4.16: Simulated Data with two observations per day, generated by adding Gaussian noise with standard deviations $SD_C = 3$, $SD_B = 0.3$ to Predator-Prey DE solutions of Chlorella and Brachionus, respectively.

The experimental 95% confidence intervals, biases and SD's for the DE parameter vectors are shown in Table 4.2 under Setting 1. The 95% confidence intervals for ϵ, m and k_C include the true parameter values. The lower 95% confidence bound for m is negative, and the SD of m is relatively large with coefficient of variance (CV) around 50%. This is because m is in two additive terms and relatively undetermined. From our experiments on other DE's, we also find that

the additive relationship often causes parameters to be poorly identified. The estimates for α, b_C, b_B and k_B have small biases.

Table 4.2: Parameter estimates under Setting 1

Parameters	ϵ	α	m	b_C	b_B	k_C	k_B
True	0.25	0.4	0.055	3.3	2.25	4.3	15
Lower 95% bound	0.24	0.47	-0.031	3.4	2.27	4.1	13.5
Upper 95% bound	0.27	0.56	0.071	3.8	2.42	5.0	14.8
BIAS*100	0.78	11.8	-3.5	34	9.7	28	-81
SD*100	0.72	2.3	2.6	9.7	3.9	24	33

Table 4.3: Parameter estimates under Setting 2

Parameters	ϵ	α	m	b_C	b_B	k_C	K_B
True	0.25	0.4	0.055	3.3	2.25	4.3	15
Lower 95% bound	0.23	0.42	-0.052	3.3	2.20	3.8	13.0
Upper 95% bound	0.29	0.60	0.113	3.9	2.50	5.2	15.4
BIAS*100	0.86	11	-2.5	31	9.9	20	-79
SD*100	1.5	4.7	4.2	15	7.8	37	60

Table 4.4: Parameter estimates under Setting 3

Parameters	ϵ	α	m	b_C	b_B	k_C	k_B
True	0.25	0.4	0.055	3.3	2.25	4.3	15
Lower 95% bound	0.26	0.44	0.052	3.1	2.23	4.4	14.7
Upper 95% bound	0.29	0.61	0.098	3.4	2.55	5.5	15.0
BIAS*100	2.62	12.3	2.0	-4.5	15	63	-17
SD*100	0.94	4.5	1.2	7.3	8.0	27	9.0

Table 4.5: Parameter estimates under Setting 4

Parameters	ϵ	α	m	b_C	b_B	k_C	k_B
True	0.25	0.4	0.055	3.3	2.25	4.3	15
Lower 95% bound	0.24	0.37	0.013	3.13	2.19	3.84	14.6
Upper 95% bound	0.27	0.45	0.101	3.5	2.33	5.15	15.4
BIAS*100	0.29	0.84	0.18	0.023	1.07	19.7	0.92
SD*100	0.74	2.1	2.2	8.7	3.6	33	21

Table 4.3 shows the experimental 95% confidence intervals, biases and SD's for the DE parameter vectors under Setting 2. The 95% confidence intervals for ϵ, m, b_C, b_B, k_C, k_B include the true parameter values. The lower 95% confidence bound for m is still negative, as explained above. The estimates for α have small bias. Comparing the results under Setting 1 and Setting 2, we can investigate the

effect of data noise on parameter estimates. Figure 4.17 shows that the parameter estimates have similar medians, but their SD's double when the noise SD's double.

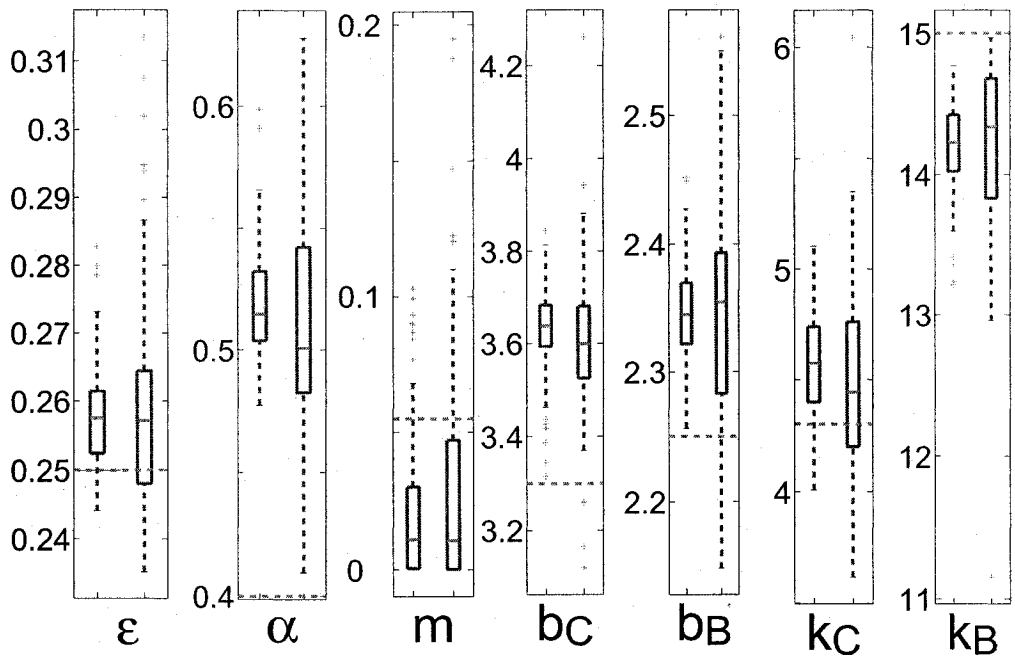

Figure 4.17: Boxplots for parameter estimates for Predator-Pray DE's. In each boxplot, the left corresponds to Setting 1 and the right corresponds Setting 2; The red dashed lines correspond to the true parameter values.

Table 4.4 shows the experimental 95% confidence intervals, biases and SD's for the DE parameter vectors under Setting 3. The 95% confidence intervals for m, b_C, b_B, k_C, k_B include the true parameter values. The confidence interval for m does not include any negative values. The estimates for ϵ and α have small biases.

Comparing the results under Setting 1 and Setting 3, we can investigate the data resolution effect on parameter estimates. Figure 4.18 shows that the parameter estimates have similar SD's, but their medians are very different with different data resolutions.

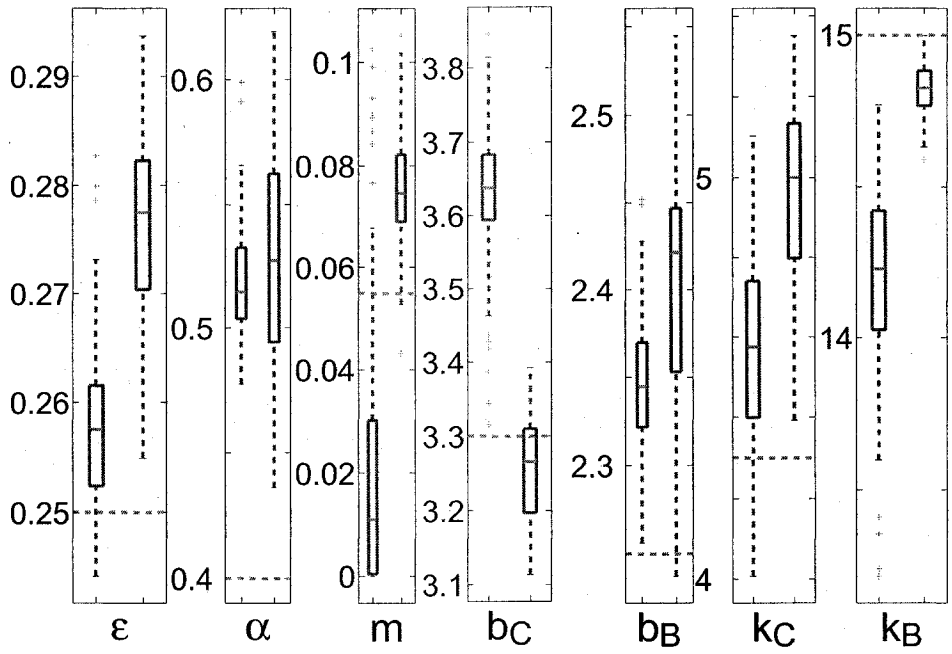

Figure 4.18: Boxplots for parameter estimates for Predator-Pray DE's. In each boxplot, the left corresponds to Setting 1 and the right corresponds to Setting 3; The red dashed lines correspond to the true parameter values.

The experimental 95% confidence intervals, biases and SD's for the DE parameter vectors under Setting 4 are shown in Table 4.5. The true parameter values

fall into the 95% confidence intervals. Comparing the results under Setting 1 and Setting 4, we can investigate the flexibility of basis system effects on parameter estimates. The biases of parameter estimates under Setting 4 are only 1% of those under Setting 1. Figure 4.19 shows that the parameter estimates have similar SD's. The medians of parameter estimates under Setting 4 are also very close to the true parameter values, but those under Setting 1 are far from the true parameter values.

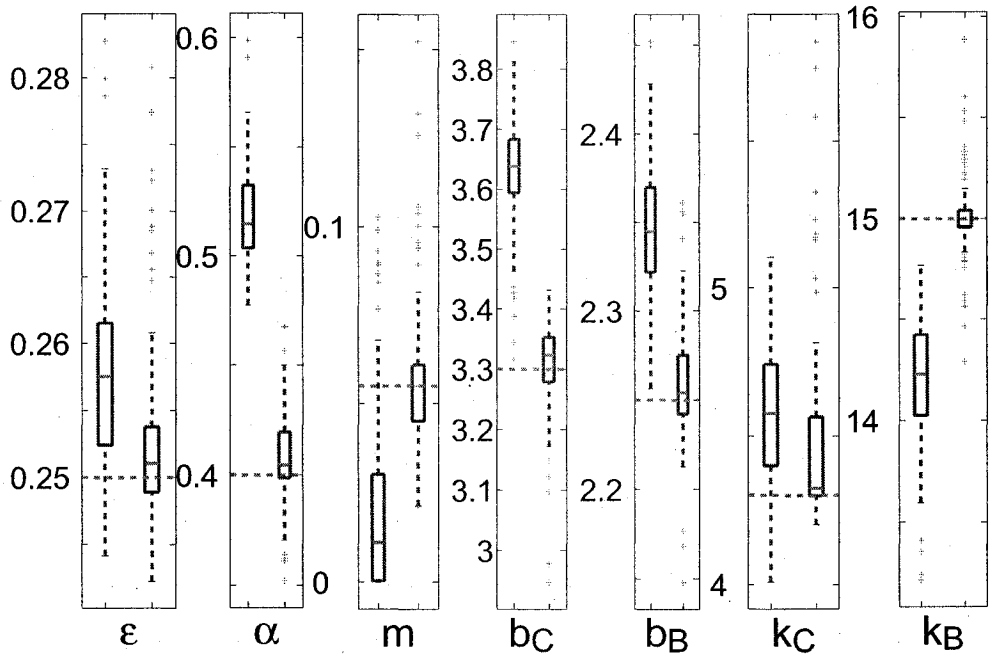

Figure 4.19: Boxplots for parameter estimates for Predator-Pray DE's. In each boxplot, the left corresponds to Setting 1 and the right corresponds Setting 4; The red dashed lines correspond to the true parameter values.

From our simulations, we can conclude that the data noise affects SD's of parameter estimates, and has little effect on biases of parameter estimates. Instead, the data resolution affects the biases, but has a small effect on SD's. It is very important for the B-spline basis system to have enough knots such that it is flexible enough to approximate DE solutions. Otherwise, it causes serious biases of parameter estimates.

4.8 Estimating Functional Parameters in DE's from Data

Some DE's have some functional parameters, that is, functions in term of time or some components in the DE's. For example, in the predator-prey DE's (4.3), the link functions $F_C(N) = b_C N/(k_C + N)$ and $F_B(C) = b_B C/(k_B + C)$ control the effect of nitrogen concentration on the rate of change of the Chlorella concentration, and the effect of the Chlorella concentration on the rate of change of the Brachionus concentration, respectively. But we are not sure whether the link functions should be specified in those forms. In the following, we estimate the link functions in the predator-prey DE's from data with the generalized profiling method. First, we explore the appropriate setting of the basis systems to expand both the DE components and functional parameters based on simulated data, and then estimate the link functions from real data. This is also a typical process when applying the generalized profiling method to estimate DE parameters from real data.

4.8.1 Estimating Functional Parameters from Simulated Data

Our simulated data are generated by adding Gaussian noise to the predator-prey DE (4.3) solutions for Chlorella and Brachionus, taking the same time points as the real data shown in Figure 4.2. In the following, the link functions in the predator-prey DE (4.3) that generate the simulated data are called the "true" link functions. The objective is to estimate the two link functions $F_C(N)$ and $F_B(C)$ and parameters ϵ, α and m from simulated data, which should be close to the true ones.

It is natural to express the two link functions as linear combinations of B-Spline basis functions, which can be written as

$$F_C(N) = \sum (c_i^1 \psi_i^1(N))$$

$$F_B(C) = \sum (c_i^2 \psi_i^2(C)),$$

where $\psi_i^1(N)$ and $\psi_i^2(C)$ are basis functions, and c_i^1 and c_i^2 are the corresponding coefficients, respectively. In order to investigate the effect of basis system on the parameter estimates, we do the following two experiments:

Setting 1: Each component in the predator-prey DE's (4.3) is expanded by the cubic B-spline basis with 400 equally spaced knots. The link function $F_C(N)$ is expanded by the cubic B-spline basis with interior knots 10, 20, 40, and 60 and the link function $F_B(C)$ is expanded by the cubic B-spline basis with

interior knots 20, 40, and 60.

Setting 2: Each component in the predator-prey DE's (4.3) is expanded by the cubic B-spline basis with 800 equally spaced knots. The link function $F_C(N)$ is expanded by the cubic B-spline basis with interior knots 10 and 40 and the link function $F_B(C)$ is expanded by the cubic B-spline basis with interior knots 20 and 60.

The estimated parameter values for ϵ, α and m are shown in Table 4.6. It is obvious that estimates under Setting 2 have little bias and are better than those under Setting 1. The estimated link function for $F_C(N)$ under Setting 1 have more variations than the true one (Figure 4.20). The estimated link functions for $F_C(N)$ and $F_B(C)$ under Setting 2 are almost the same as the true ones (Figure 4.21).

Table 4.6: The parameter estimates when estimating the link functions from simulated data

Prameters	ϵ	α	**m**
True	0.25	0.40	0.055
Setting 1	0.269	0.44	0.100
Setting 2	0.245	0.39	0.054

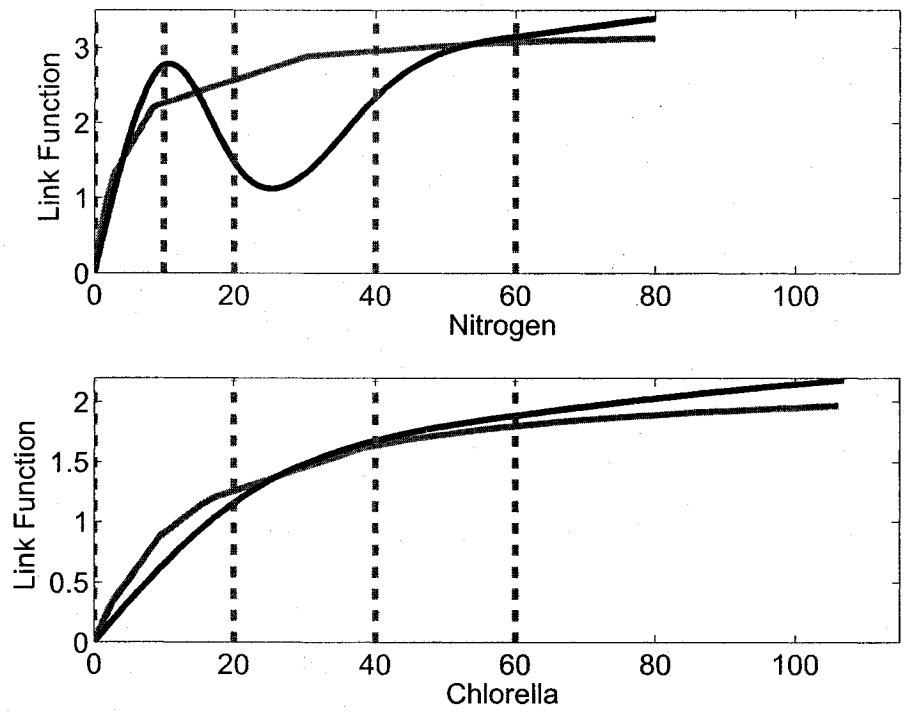

Figure 4.20: The estimated link functions for $F_C(N)$ (top) and $F_B(C)$ (bottom) under Setting 1. The blue lines are the Fussmann's original link function (true), and the black ones are the estimated link function from simulated data. The blue dashed lines indicate the interior knot locations for B-splines approximating the link functions.

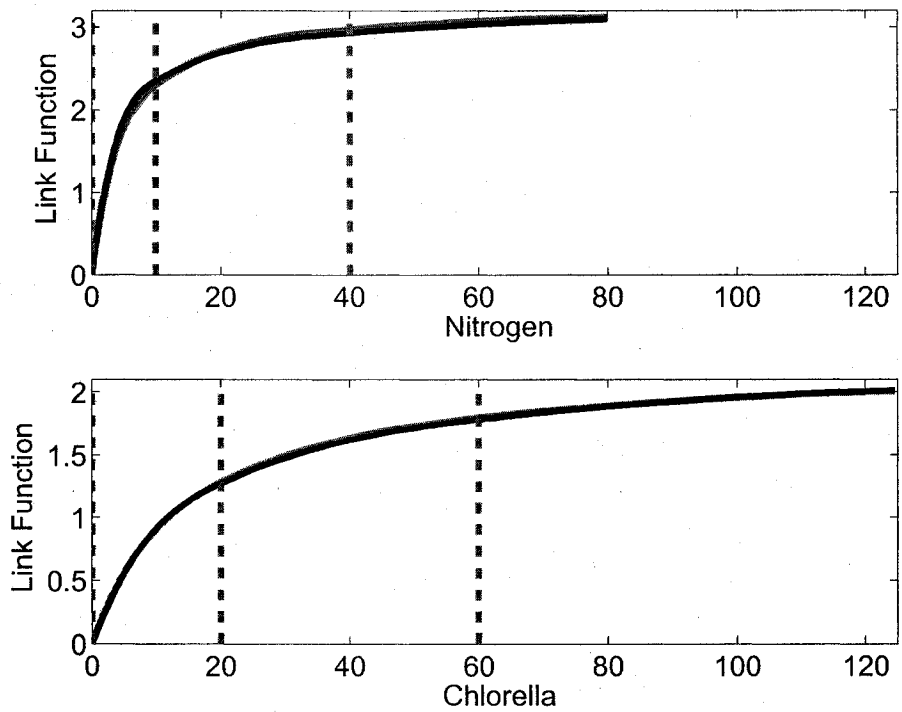

Figure 4.21: The estimated link functions for $F_C(N)$ (top) and $F_B(C)$ (bottom) under Setting 2. The blue lines are the Fussmann's original link function (true), and the black ones are the estimated link function from simulated data. The true and estimated link functions are almost on top of each other. The blue dashed lines indicate the interior knot locations for B-splines approximating the link functions.

4.8.2 Estimating Functional Parameters from Real Data

Setting 2 has been shown to be good to estimate the link functions $F_C(N)$, $F_B(C)$ and parameters ϵ, α and m in the simulations. We use this setting to do the same

task on the real data shown in Figure 4.2. The estimated parameter values for ϵ, α and m are shown in Table 4.7. Figure 4.22 displays the estimated link functions for $F_C(N)$ and $F_B(C)$, which have the same patterns as Fussmann's. The difference can be caused by different values of parameters in the link functions, so the forms of link functions proposed by Dr. Fussmann are verified to be appropriate. What we do next is to estimate the parameters b_C, b_B, K_C and K_B which define the link functions.

Table 4.7: The parameter estimates when estimating the link functions from real data

Prameters	ϵ	α	**m**
True	0.25	0.40	0.055
Setting 2	0.34	0.57	0.28

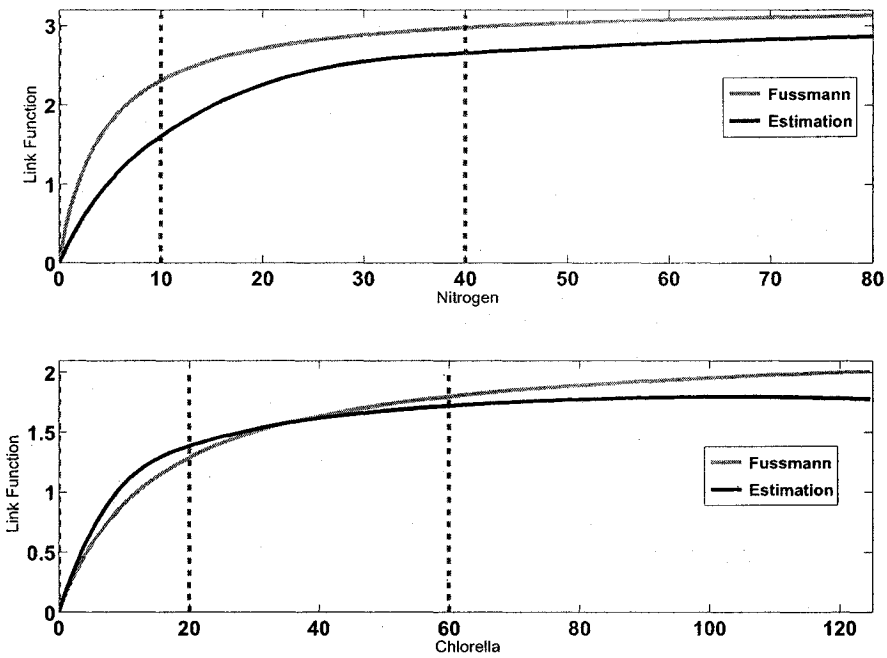

Figure 4.22: The estimated link functions for $F_C(N)$ (top) and $F_B(C)$ (bottom) under Setting 2 (right) from real data. The red lines are the Fussmann's original link function, and the black ones are the estimated link function from real data. The blue dashed lines indicate the interior knot locations for B-splines approximating the link functions.

4.9 Fitting a Predator-Prey Dynamic System to Biological Data

In the following three sections, we work on estimating DE parameters from real data. Section 4.3 discusses that the DE's proposed by Fussmann et al. (2000) predict correctly the dynamic behaviors of the experimental observations. However, the scales of the DE solutions are actually far from observations. In the following, we first rescale data by multiplying constants; This procedure is also biologically meaningful. We then show that DE solutions are much closer to observations with our estimated DE parameters and initial values of components.

4.9.1 Rescaling Observations for a Predator-Prey Dynamic System

Let $\mathbf{y} = (y(t_1), \cdots, y(t_n))$ be the functional data, and $x(t)$ be the corresponding DE solution, then we can rescale data $\mathbf{y}$ with a constant coefficient s by minimizing

$$H(s|\mathbf{y}) = \sum_{i=1}^{n} (sy(t_i) - x(t_i))^2 . \tag{4.16}$$

It is easy to get that

$$s = \frac{\sum_{i=1}^{n} [x(t_i)y(t_i)]}{\sum_{i=1}^{n} [y(t_i)]^2} . \tag{4.17}$$

The estimated scale parameters are 28 and 0.57 for Chlorella and Brachionus, respectively. In this predator-prey dynamic system, these two scale parameters can be interpreted as the amount of Nitrogen inside per individual Chlorella and Brachionus, respectively. The rescaled data are shown in Figure 4.23. Data are close to DE solution obtained with the original values of parameters. However, the rescaled Chlorella data are far from the DE solution for Chlorella on the boundaries. The DE solution for Brachionus does not show the same two modes as the corresponding data, either. In the following, we estimate DE parameters from the rescaled data.

4.9.2 Estimating Parameters in a Predator-Prey Dynamic System

Let M be the number of observed components (here $M = 2$), and n be the number of observations. If $x_j(t_i)$ is the observation for the j-th component at time t_i, and $\hat{x}_j(t_i)$ is the DE solution with the estimate $\hat{\boldsymbol{\theta}}$ for the j-th component at time t_i, then MSE is defined as

$$\text{MSE} = \frac{1}{nM} \sum_{j=1}^{M} \{\omega_j \sum_{i=1}^{n} [x_j(t_i) - \hat{x}_j(t_i)]^2\} .$$

EMSE has the same definition as MSE except that DE's are solved with the estimated initial values, which is attained by smoothing data using DE's with estimated parameters. We evaluate the goodness of fit in terms of MSE and EMSE.

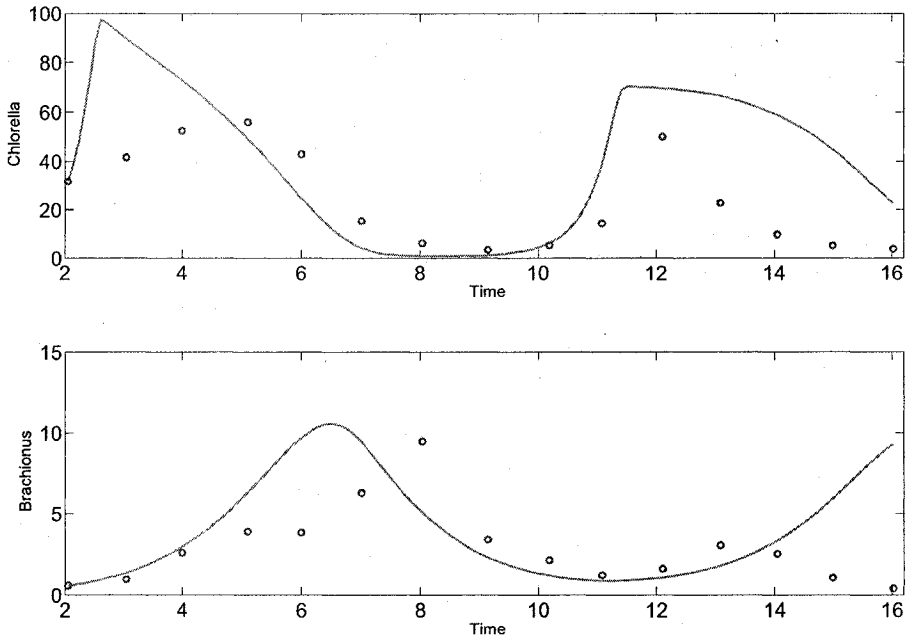

Figure 4.23: Fit data by simply rescaling data with a constant, where we multiply the observed number of Chlorella and Brachionus by 28.005 and 0.571, respectively. The red lines are the DE solutions of Chlorella and Brachionus; and the blue circles are rescaled data observed in the biological experience with the dilution rate $\delta = 0.68$.

Each component is expanded by the cubic B-spline with 400 equally spaced knots, and the smoothing parameter $\lambda = 10^5$. The parameter estimates are shown in Table 4.8, and MSE decreases by 33.5%. With the estimated DE parameter vector $\boldsymbol{\theta}$, and the first observations as the initial values for Chlorella and Brachionus, DE's (4.3) are solved with the solutions shown in Figure 4.24. The DE solution

for Chlorella is closer to rescaled data on the right side, but has little improvement on the left side. The DE solution for Brachionus shows the similar period of cycle as the rescaled Brachionus data.

Table 4.8: Parameter Estimates for Predator-Prey DE's

Parameters	ϵ	α	m	b_C	b_B	k_C	k_B	**MSE**	**EMSE**
Fussmann	0.25	0.40	0.055	3.3	2.25	4.3	15.0	1.96	1.29
Estimates	0.14	0.51	0.019	3.5	2.19	2.2	14.9	1.30	0.34

With the estimated initial values and parameter values, DE solutions can fit data much better (Figure 4.25). EMSE decreases by 65.6%. The DE solution for Chlorella can fit data very well over all the region. The DE solution for Brachionus also show the same pattern as the rescaled Brachionus data, although not well at day 8, which is suspected to be an outlier.

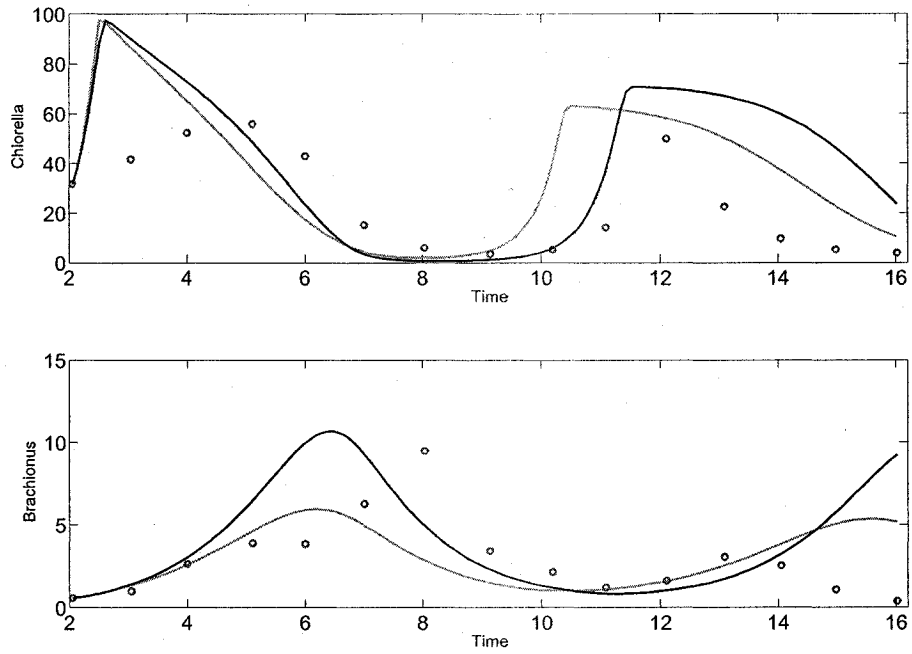

Figure 4.24: Solving the Predator-Prey DE's (4.3) with Fussmann's parameter values (Black solid line) and Profiling PDA parameter estimates (Red solid line), using the first observations as the initial values. Blue circles are the rescaled data. The smoothing parameter $\lambda = 10^5$ for Profiling PDA.

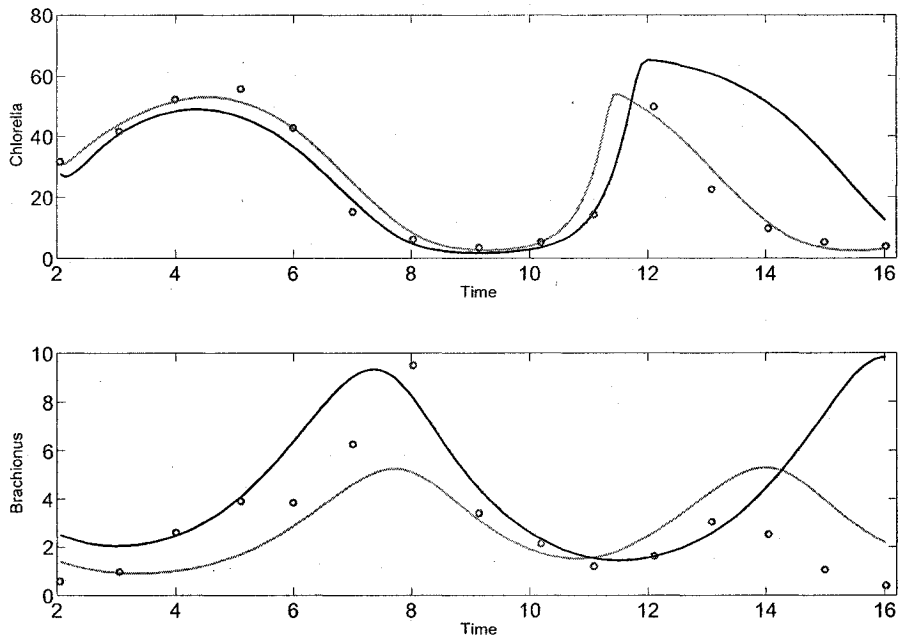

Figure 4.25: Solving the predator-prey DE's (4.3) with Fussmann's parameter values (Black solid line) and Profiling PDA parameter estimates (Red solid line), using the estimated initial values by resmoothing data using DE's with the corresponding parameter values. Blue circles are the rescaled data. The smoothing parameter $\lambda = 10^5$ for Profiling PDA.

4.10 Statistical Inference for a HIV Dynamic Model from Clinical Trials

Section 4.4 shows three simple DE's that model the rate of population change of uninfected cells, infected cells and virus. In this section, we show that solutions of HIV DE's with our estimated parameters and initial values are close to data.

We randomly select Subject 40 and estimate the parameter vector $\boldsymbol{\theta}$ in (4.4) from his observations, which are shown in Figure 4.26. One challenging problem is that the number of uninfected cells and infected cells are not measurable. In these circumstances, mathematicians tend to choose initial values for the unobserved components based on steady-state conditions (Figure 4.27). However, doing this yields DE solutions that are far from data. We smooth data with HIV DE's (4.4), and evaluate the fitting functions at the beginning of time points as the initial values for both observed and unobserved components. DE solutions with our estimated initial values can fit data better (Figure 4.27).

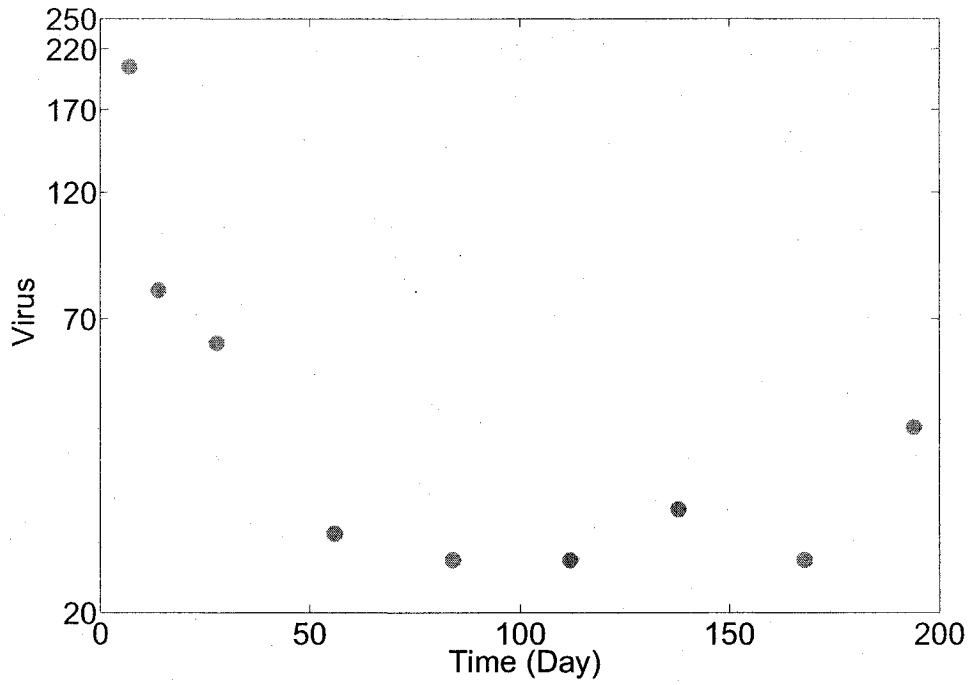

Figure 4.26: The number of free virus for Subject 40.

Moreover, we estimate the parameter vector $\boldsymbol{\theta}$ in the HIV DE's (4.4). Each component is approximated by B-splines with 160 equally spaced knots, and the smoothing parameter $\lambda = 10^3$. With the estimated parameter values (Table 4.9), DE solutions can fit the data very well (Figure 4.28).

Table 4.9 also shows the estimated SD's of parameters, which are relatively large, due to having 6 parameters estimated from 9 observations, and the degree of freedom is very small. This problem can be overcome by pooling data of 42 subjects together and estimating the fixed and random effects, which we call a

mixed dynamic model.

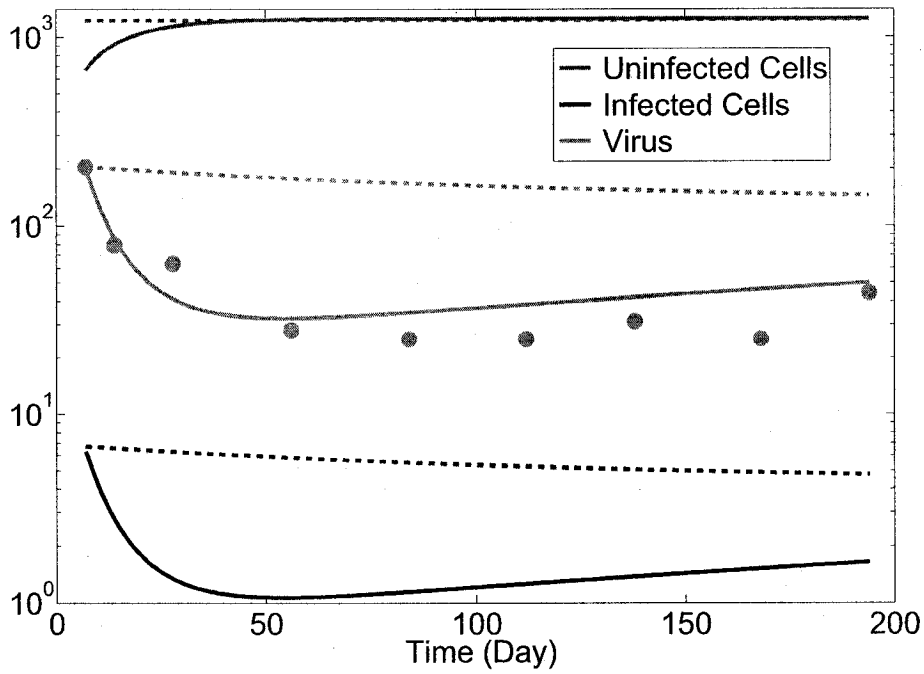

Figure 4.27: DE solutions with our estimated initial values (solid lines). The dashed lines are DE solutions with initial values estimated from the steady-state conditions.

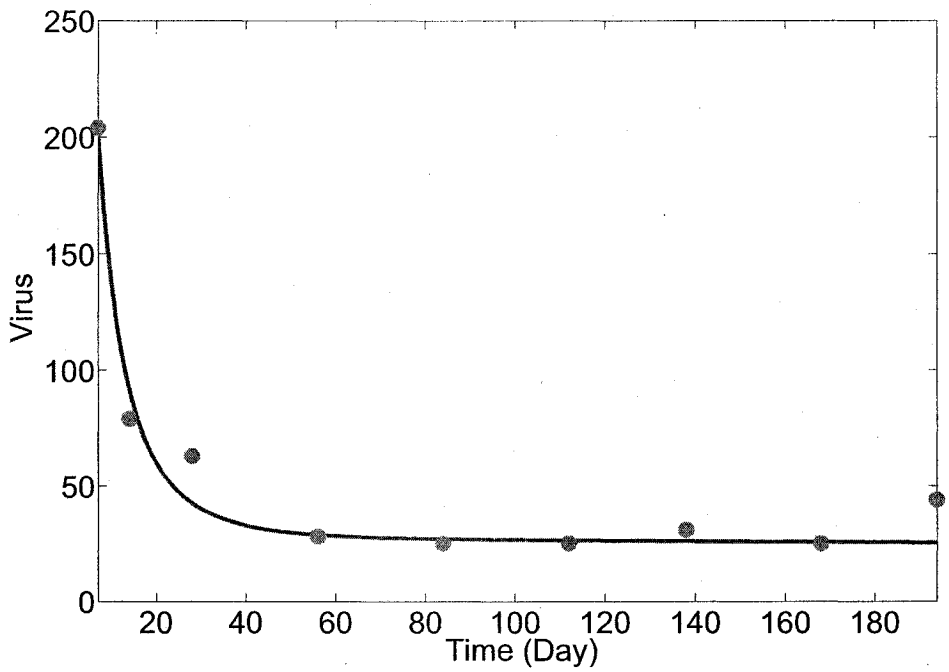

Figure 4.28: DE solutions with the estimated parameter and initial values (Table 4.9). The smoothing parameter $\lambda = 1000$. 160 equally spaced knots are used for each component.

Table 4.9: Parameter and initial values estimates for HIV DE's

Estimation	b	r_u	i	r_i	n	r_v	$U(0)$	$I(0)$	$V(0)$
Huang2005	100.0	0.080	9.9e-6	0.37	246	3.0	657.3	6.400	204.7
Profiling PDA	93.4	0.072	9.5e-6	0.40	244	3.0	659.3	6.397	204.4
Estimated $\hat{\sigma}$	89	0.035	4.7e-6	0.34	368	5.8	-	-	-

4.11 Dynamic Models for Thermal Decomposition of $\alpha-$ Pinene

The compound $\alpha-$ pinene is a component of turpentine, and is used in pharmaceutical and aroma-chemical products. Fuguitt and Hawkins (1945) and Fuguitt and Hawkins (1947) investigated the thermal decomposition of α-pinene when heating α-pinene in the liquid phase over the temperature range 189.5 °C – 285 °C. They found that the α-pinene first decomposed into dipentene and allo-ocimene simultaneously, and the allo-ocimene further decomposed into α-pyronene, β-pyronene and a dimer. They also reported the relative concentrations of α-pinene and four by-products at 8 time points under the temperature 189.5 °C and 285 °C. In this section, we explore several sets of DE's to model the thermal decomposition of α-pinene, and test the best model among them by fitting them to the real data.

Box et al. (1973) examined Fuguitt and Hawkins' papers, and pointed out that pyronene was not actually measured because of experimental difficulties and was imputed from the other concentrations under mass balance considerations, instead. So in the following, both the literature and ourselves treat the data of pyronene concentration as missing.

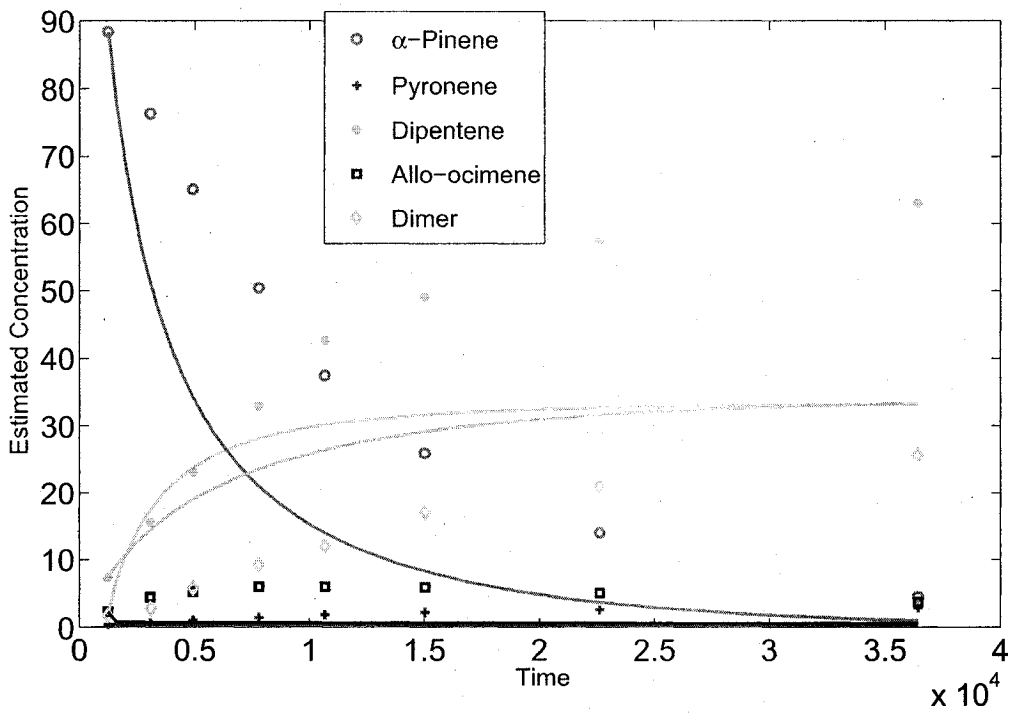

Figure 4.29: The solid curves are the solutions of α-Pinene DE's (4.18) with parameter values give by Stewart and Sorensen (1981); The points are data.

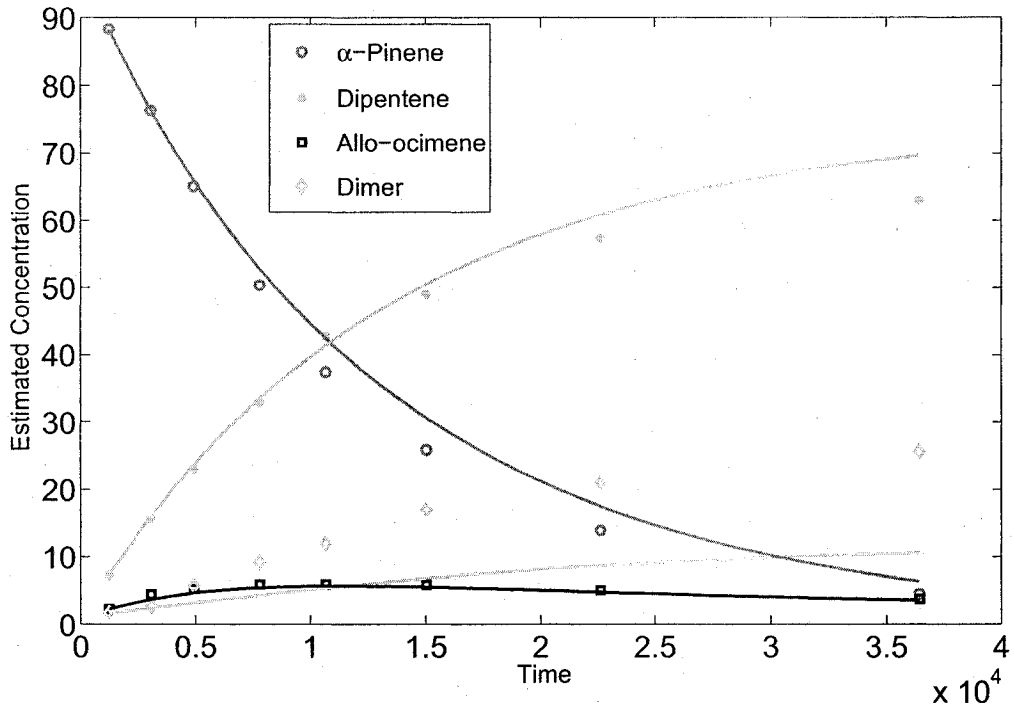

Figure 4.30: The solid curves are the solutions of α-Pinene DE's (4.18) with the new parameter estimates from Profiling PDA; The points are data.

Box et al. (1973) proposed a set of linear DE's to model the thermal isomerization of α-Pinene. But Bates and Watts (1988) showed that the residuals were not well behaved and had some trends after fitting DE's in Box et al. (1973) to data. Bates and Watts (1988) also pointed out that linear DE's are not flexible enough to fit data. Assuming f_i, $i = 1, \cdots, 5$ to be normalized weight percentage of α-pinene, α- and β- pyronene, dipentene, allo-ocimene, and a dimer, respectively,

Stewart and Sorensen (1981) gave a set of nonlinear DE's:

$$\begin{aligned}
\frac{df_1}{dt} &= -(\theta_1+\theta_2)f_1 - 2\theta_3 f_1^2 \\
\frac{df_2}{dt} &= -\theta_4 f_2 + \theta_5 f_4 \\
\frac{df_3}{dt} &= \theta_1 f_1 \\
\frac{df_4}{dt} &= \theta_2 f_1 + \theta_4 f_2 - \theta_5 f_4 - 2\theta_6 f_4^2 + 2\theta_7 f_5 \\
\frac{df_5}{dt} &= \theta_3 f_1^2 + \theta_6 f_4^2 - \theta_7 f_5 \,.
\end{aligned} \tag{4.18}$$

Stewart and Sorensen (1981) also derived the Bayesian estimation of parameters in nonlinear DE's. But these set of nonlinear DE's does not fit the data well with the parameter values they gave (Figure 4.29). Using their parameter estimates as the initial values, we estimate parameter values with the Profiling PDA method. Each component is approximated by B-spline with 160 equally spaced knots, and the smoothing parameter $\lambda = 10$. With our new parameter estimates, the DE solutions can fit the data well except for the dimer (Figure 4.30).

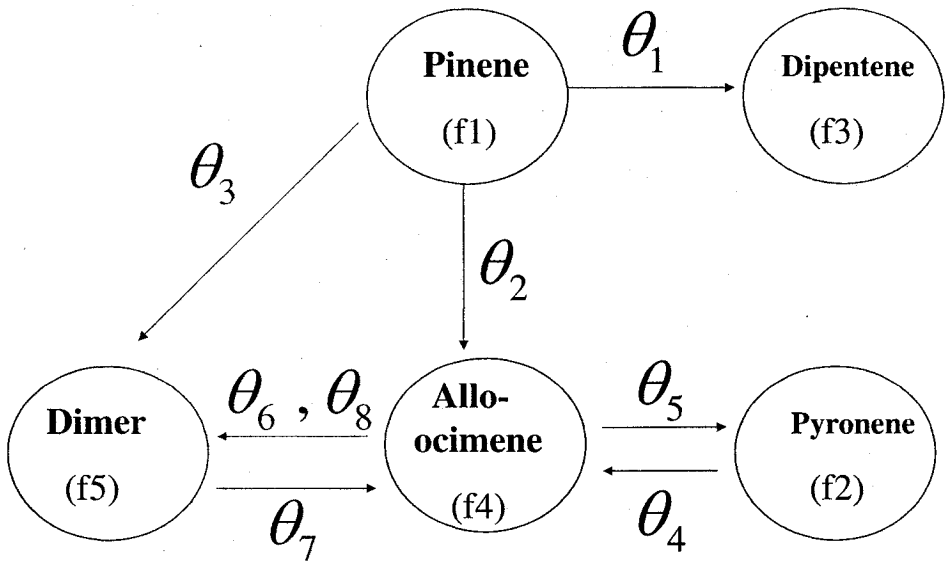

Figure 4.31: A system diagram for the α-Pinene DE's. The arrows represents a thermal decompositions.

We then combine these two sets of DE's in Box et al. (1973) and Stewart

and Sorensen (1981) to give the following DE's:

$$\begin{aligned}
\frac{df_1}{dt} &= -(\theta_1+\theta_2)f_1 - 2\theta_3 f_1^2 \\
\frac{df_2}{dt} &= -\theta_4 f_2 + \theta_5 f_4 \\
\frac{df_3}{dt} &= \theta_1 f_1 \\
\frac{df_4}{dt} &= \theta_2 f_1 + \theta_4 f_2 - \theta_5 f_4 - 2\theta_6 f_4^2 + 2\theta_7 f_5 \\
\frac{df_5}{dt} &= \theta_3 f_1^2 + \theta_6 f_4^2 - \theta_7 f_5 \,.
\end{aligned} \tag{4.19}$$

A system diagram for DE's (4.19) is shown in Figure 4.31, in which each arrow corresponds to one chemical reaction. The DE's in Box et al. (1973) correspond to $\theta_3 = 0$, $\theta_4 = 0$ and $\theta_6 = 0$. This means that they assume α-pinene does not decompose directly into dimer, α- and β- pyronene do not decompose into allo-ocimene, and the decomposition rate of allo-ocimene to dimer is linear with the percentage of allo-ocimene. The DE's in Stewart and Sorensen (1981) are equivalent to $\theta_8 = 0$. This means that they assume the decomposition rate of allo-ocimene to dimer is only quadratic with the percentage of allo-ocimenen.

Table 4.10: Parameter estimates for 4.19.

Parameters	θ_1	θ_2	θ_3	θ_4	θ_5	θ_6	θ_7	θ_8
Stewart(10^{-5})	5.83	2.88	0.156	14.1	8.04	210	2.50	0
Bates(10^{-5})	5.94	2.86	0	0	0.45	0	5.79	31.12
Profiling PDA(10^{-5})	5.938	2.92	−0.0001	20.0	9.07	−3.13	4.18	44.3
SD(10^{-9})	2.62	2.17	0.047	7.67	27.9	11.4	2.87	58.3

The parameter estimates for (4.19) are shown in Table 4.10, using the Profiling PDA method. Each component is approximated by B-spline with 160 equally spaced knots, and the smoothing parameter $\lambda = 10^5$. The estimated $\hat{\theta}_3$ is negative, but its value is negligible, compared with the scales of other parameter estimates. The estimated θ_6 is also negative, which make sense because θ_6 is the coefficient to the extra quadratic term $\theta_6 f_4^2$ besides the linear term $\theta_8 f_4$ for the decomposition of allo-ocimene into dimer. Parameters θ_4 and θ_6 obviously cannot be zero, which can explain why Bates and Watts (1988) found the linear DE's were not adequate to fit the data well. Parameter θ_8 shouldn't be zero, either, which can explain why we can not fit the data well with the nonlinear DE's (4.18).

We define MSE as a criterion to assess the fit of DE's to data:

$$\texttt{MSE} = \frac{1}{nm} \sum_{i=1,3,4,5} \sum_{j=1}^{n} [x_{ij} - \hat{x}_{ij}]^2 ,$$

where m is the number of components (here $m = 4$), and n is the number of observation; x_{ij} is the observation for component i at time t_j, and $\hat{x}_{ij}$ is the DE's solutions for component i at time t_j. MSE decreases by 74% with our estimated parameter values, compared with those in Bates and Watts (1988). If we estimated the initial values for all 5 components, MSE decreases by 8% further.

Figure 4.32 displays DE solutions with our estimated parameter values and initial values. The DE solutions are close to data, especially for allo-oeimene and dimer. The residuals of fit versus time and each component are shown in Figure 4.33, which display no obvious patterns.

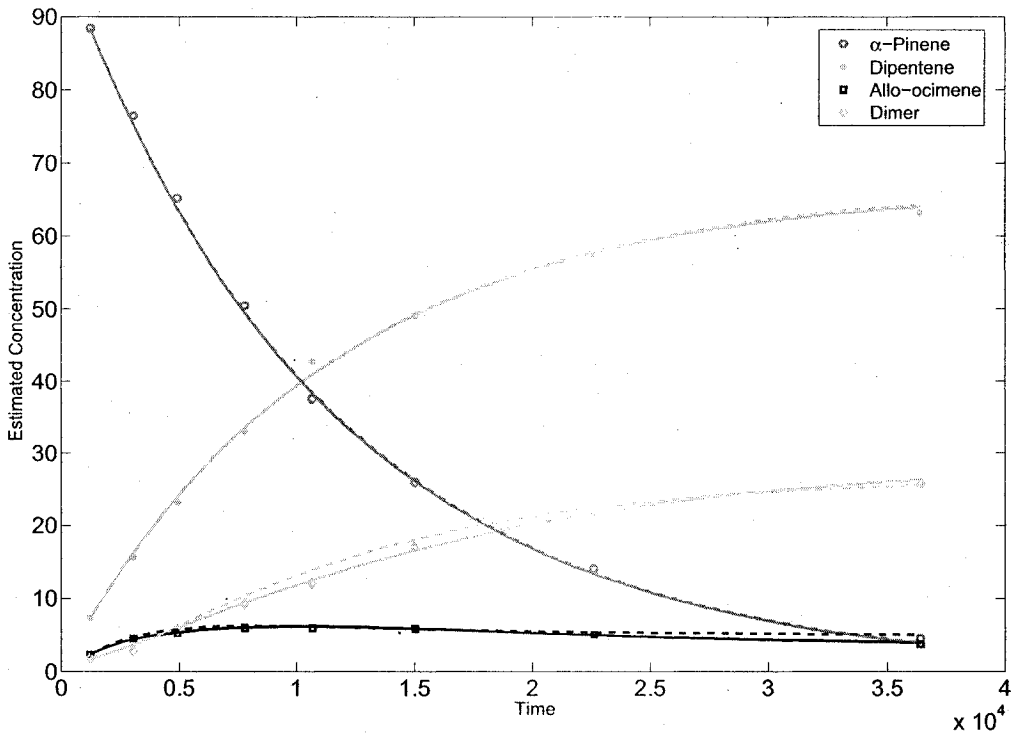

Figure 4.32: The solutions of α-Pinene DE's (4.19), using our estimated parameter values (solid lines) or Bates' estimates (dashed lines). The points are data

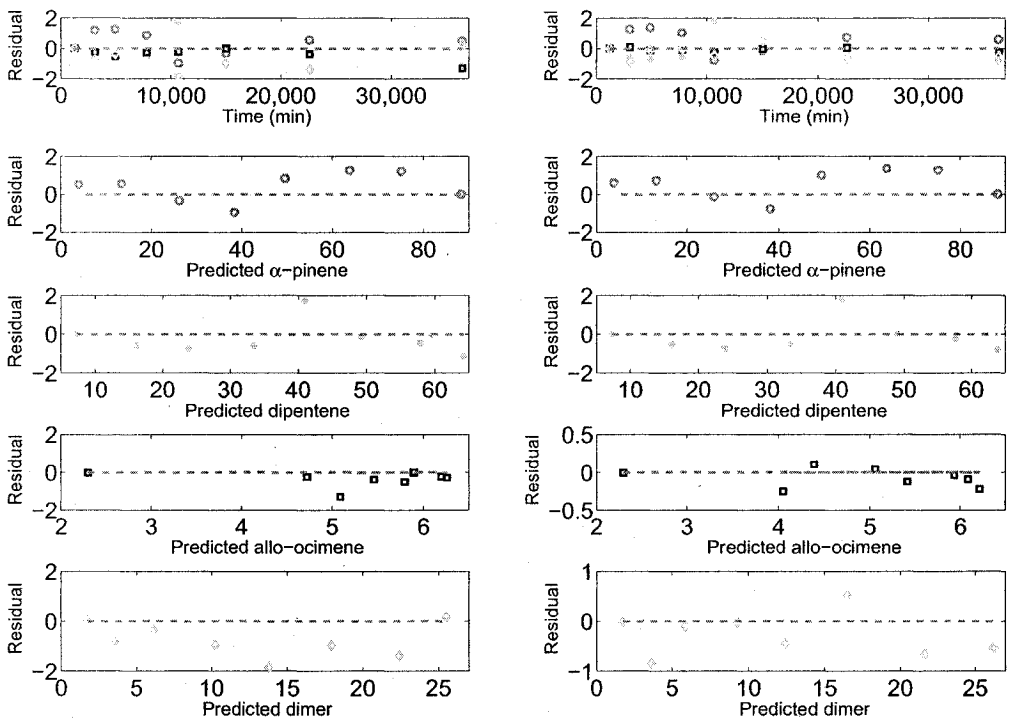

Figure 4.33: The residuals of data to the DE solutions using Bates' parameter estimates (left panel) and the Profiling PDA estimates (right panel). The top two graphs displays the two kinds of residuals of all four component versus time, and the second to the fifth lines of two graphs shows the two kinds of residuals for α-pinene, dipentenen, allo-ocimene, and dimer versus their respective predictions.

However, without any information for α- and β-pyronene, Equations (4.19) are unstable. For example, when θ_4 is 0, f_2 can change a lot with the different initial values for f_2. Moreover, although the calculated pyronene data are not

reliable, we do know that there is no pyronene at time 0, so we include this reliable information that the initial concentration of α-pinene to be 100% and the other four product to be 0% at time 0.

Table 4.11: Parameter estimates for 4.19 with 5 more observations.

Parameters	θ_1	θ_2	θ_3	θ_4	θ_5	θ_6	θ_7	θ_8
Stewart(10^{-5})	5.83	2.88	0.156	14.1	8.04	210	2.50	0
Bates(10^{-5})	5.94	2.86	0	0	0.45	0	5.79	31.12
Profiling PDA(10^{-5})	5.93	2.70	0.005	22.2	10.8	-1.96	3.29	34.1
SD(10^{-9})	2.62	2.16	0.047	7.67	27.5	11.4	2.87	58.3

Now we have 5 more data points, especially one observation for α- and β-pyronene, and DE solutions are more stable. We estimate the parameter values with the Profiling PDA method on this larger data set. Each component is approximated by a cubic B-spline with 160 equally spaced knots, and the smoothing parameter $\lambda = 10^6$. The parameter estimates are shown in Table 4.11. The parameter estimates are similar to those estimated from data without the 5 more observations, but MSE decreases 28% further from the best result before (Estimating DE parameters and initial values from 32 observations). The fit to data are shown in Figure 4.34. The DE solutions with our parameter estimates are closer to data, as we expected. The penalized fitting functions to the data are almost same with the DE solutions, which means that the B-splines are powerful enough to represent the solutions. This is also a good way to check if we have used enough knots for each components. Moreover, we can use the theoretical initial values directly (100 for α-pinene, and 0 for other components) to solve DE's (4.19) with the estimated parameter values. The residuals show no obvious patterns, either

(Figure 4.35).

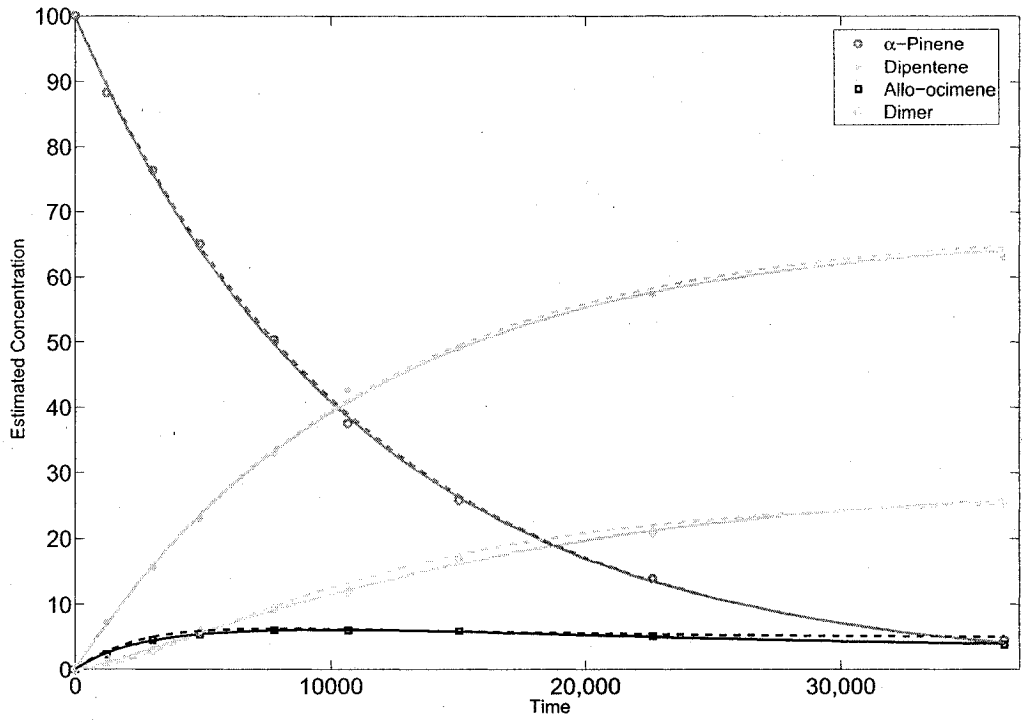

Figure 4.34: The solutions of α-Pinene DE's (4.19), using our estimated parameter values (solid lines) or Bates' estimates (dashed lines). The points are data, the dotted line is the penalized fitting functions using DE's, which are almost on top of the solutions.

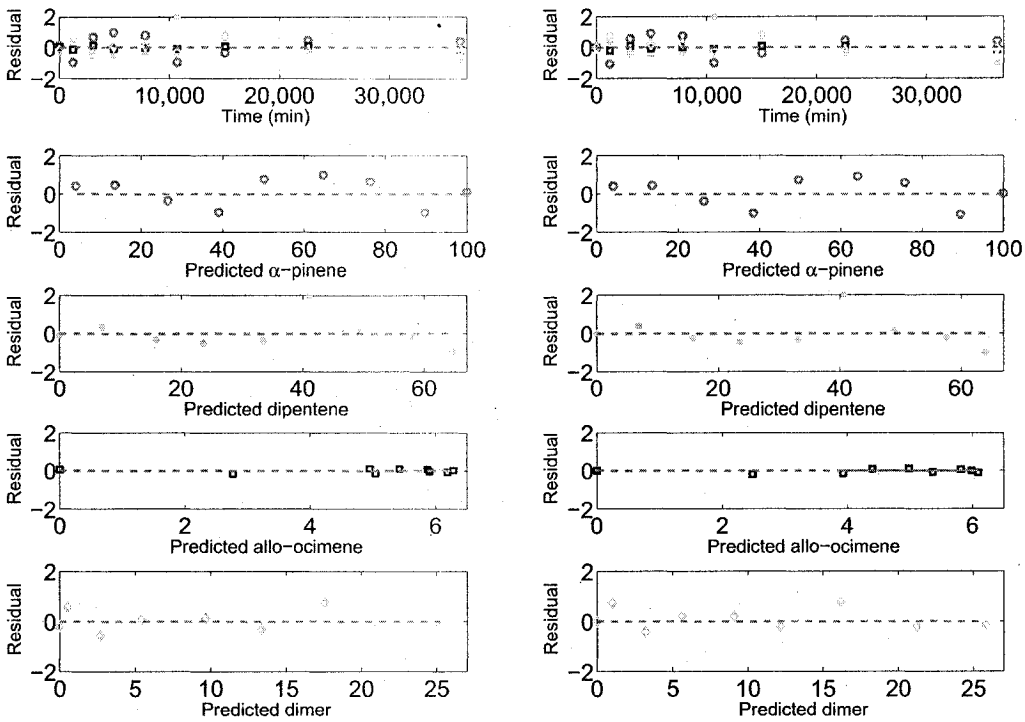

Figure 4.35: The residuals of data to the penalized smoothing splines (left panel) and the residuals of data to the solutions of α-Pinene DE's (4.19) (right panel). The top two graphs displays the two kinds of residuals of all four component versus time, and the second to the fifth lines of two graphs shows the two kinds of residuals for α-pinene, dipentenen, allo-ocimene, and dimer versus their respective predictions.

Chapter 5

Conclusions and Conjectures

This thesis explores tools for functional data analysis based on Ramsay and Silverman (2005). We introduce the generalized profiling method and three applications: adaptive penalized smoothing, estimating generalized semiparametric additive models, and fitting differential equations to noisy data. The generalized profiling method is an elegant way to estimate statistical models with local, global and complexity parameters. It also provides the unconditional estimates for variances of these three distinct groups of parameters.

Chapter 2 shows that adaptive penalized smoothing can estimate a functional smoothing parameter which is adaptive to the shape of the underlying curve. It is large where the underlying curve is almost linear, and small where the underlying curve has large curvatures. This is useful when the underlying curve has different scales of variation. The results from both simulated data and real data show that adaptive penalized smoothing can provide better estimates for fitting functions and

their derivatives than nonadaptive penalized smoothing. However, the estimates for the functional smoothing parameter are not stable when the function is not observed with sufficient resolution. When functional data with replications have similar shape, one promising solution is to pool the replicated functional data together to estimate one single functional smoothing parameter.

Chapter 3 shows that we can estimate generalized semiparametric additive models with response variables in any distributions based on their likelihood functions. Moreover, the unconditional estimates for variances of linear coefficients are derived, which include the variation coming from the smoothing parameter. However, The estimate for the smoothing parameters by minimizing the approximated GCV proposed by Gu and Xiang (2001) is not stable. We will try to propose or find some alternative criteria which can give a more stable estimate of the smoothing parameter.

In chapter 4, it is shown that DE's are good tools to model the dynamic behavior in medicine, biology and chemical engineering. Nonparametric curves and their derivatives can be well estimated by penalized smoothing with the penalty term defined by DE's, and this process is also called L-spline smoothing. The value of smoothing parameter can be selected by generalized cross-validation and Stein's unbiased risk estimate. When differential equations are nonlinear, the approximated generalized cross-validation is also derived.

Chapter 4 also shows that DE parameters can be estimated from noisy data with the generalized profiling method. DE's are not solved directly. Instead, a smoothing spline is estimated to approximate DE solutions by L-spline smoothing.

A byproduct of this method is that the initial values for DE components can also be estimated by L-spline smoothing. The functional parameters in DE's can also be estimated in term of linear combinations of basis functions. The data are close to the solutions of DE's, solved with the estimated DE parameters and the estimated initial values. Our method can also handle dynamic systems with some unmeasurable components. Three applications are demonstrated, which come from ecology, medicine and chemical engineering, respectively.

For the predator-prey dynamic system, we have succeeded in fitting DE's proposed by Fussmann et al. (2000) to their experiment observations. Dr. Fussmann also collected several sets of observations measured daily in one whole year. The computation is too intensive for the generalized profiling method to handle with these long term data. In this case, the multiple shooting method proposed by Bock (1983) will be promising if we combine it with the generalized profiling method.

For the HIV dynamic system, we estimate DE parameters from the observations of Patient 40. But it is still unclear how to estimate DE parameters from the data of total 42 patients with the generalized profiling method. Huang et al. (2005) overcome this problem by applying the Bayesian method. However, it is hard to choose the prior distributions for DE parameters, and the computation is intensive. Cao and Campbell (2006) worked on estimating DE parameters with Bayesian smoothing. The DE's are not solved numerically, and instead, smoothing splines are used to approximate the DE solutions. The pseudo likelihood is generated with the estimated smoothing splines as the mean of the observations. The DE's define the prior distribution of the smoothing coefficients using the same

penalty as is used in the L-spline smoothing. The full conditional posterior distributions for the smoothing coefficients and DE parameters are both written in closed forms, which naturally combine the information of data and DE's, and the computation is very fast. Moreover, the smoothing parameters can be estimated from the full conditional posterior distribution.

We explore several DE's to model the thermal decomposition of α-pinene. We also discuss whether some parameter values are significantly different from 0, which is equivalent to test whether any reactions happen between components. But formal statistical tests are required to be proposed.

The value of the smoothing parameter has a large effect on the DE parameter estimates from noisy data with the generalized profiling method. A small smoothing parameter leads to the parameter estimates with large biases and small variances. On the other hand, a large smoothing parameter result in the parameter estimates with small biases and large variances. Generalized cross-validation can select a good value for the smoothing parameter, which is near to the optimal value that minimizing MSE of parameter estimates. Instead of fixing the value of the smoothing parameter, another solution is to start with a small smoothing parameter value and obtain the DE parameter estimates. The obtained DE parameter estimates are then updated by increasing the smoothing parameter, in order to find a global optimal DE parameter estimates with small sampling variance.

We can now estimate ordinary DE's well from noisy data with the generalized profiling method. It is interesting to apply this method to estimate partial DE's and stochastic DE's.

The generalized profiling method has been shown to estimate statistical models well with local, global and complexity parameters. But there are several important theoretical problems that are still unsolved. First, we have understood that the criteria in the first and second level should be different. It makes sense to use the likelihood or regularized likelihood function as the criterion for the first level optimization. What should be the criterion for the second level optimization? For our three applications, it works well to use the regularized likelihood function as the first level criterion and the likelihood function as the second level criterion. But how can this be formalized theoretically? Finally, our experience shows that the optimization surface can be smoother when we estimate global parameters by viewing local parameters as functions of global parameters. But why? What will happen to the optimization surface when we estimate the local and global parameters jointly?

More applications of the generalized profiling method are required to explore. For instance, it is interesting to apply this method to estimate the classic proportional hazard model.

Appendix A

Derivative Calculations in Chapter 2

A.1 Derivative Calculations for Estimating Variances of Global and Local Parameters

The formulas (2.9) and (2.10) for $d^2F/d\boldsymbol{\theta}^2$ and $d^2F/d\boldsymbol{\theta}d\mathbf{y}$ involve the terms $\partial\hat{\mathbf{c}}/\partial\mathbf{y}$, $\partial^2\hat{\mathbf{c}}/\partial\boldsymbol{\theta}^2$ and $\partial^2\hat{\mathbf{c}}/\partial\boldsymbol{\theta}\partial\mathbf{y}$. In the following, we derive the formulas for these three terms.

We introduce the following convention, which is caller *Einstein Summation Notation.* If a Latin index is repeated in a term, then it is understood as a summation with respect to that index. For instance, instead of the expression $\sum_i a_i x_i$, we merely write $a_i x_i$. Einstein Summation Notion is also used in Appendix B.

- $\frac{\partial\hat{\mathbf{c}}}{\partial\mathbf{y}}$

Since the optimal local parameter vector $\hat{\mathbf{c}}$ satisfying $\partial H(\mathbf{c}|\boldsymbol{\theta},\mathbf{y})/\partial \mathbf{c} = 0$, and $\hat{\mathbf{c}}$ is a function of $\boldsymbol{\theta}$ and $\mathbf{y}$, we can take the $\mathbf{y}$-derivative on $\partial H(\mathbf{c}|\boldsymbol{\theta},\mathbf{y})/\partial \mathbf{c}|_{\hat{\mathbf{c}}} = 0$ as follows:

$$\frac{d}{d\mathbf{y}}\left(\frac{\partial H(\mathbf{c}|\boldsymbol{\theta},\mathbf{y})}{\partial \mathbf{c}}\bigg|_{\hat{\mathbf{c}}}\right) = \frac{\partial^2 H(\mathbf{c}|\boldsymbol{\theta},\mathbf{y})}{\partial \mathbf{c}\partial \mathbf{y}}\bigg|_{\hat{\mathbf{c}}} + \frac{\partial^2 H(\mathbf{c}|\boldsymbol{\theta},\mathbf{y})}{\partial \mathbf{c}^2}\bigg|_{\hat{\mathbf{c}}}\frac{\partial \hat{\mathbf{c}}}{\partial \mathbf{y}} = 0\,, \qquad \text{(A.1)}$$

which holds since $\partial H(\mathbf{c}|\boldsymbol{\theta},\mathbf{y})/\partial \mathbf{c}|_{\hat{\mathbf{c}}}$ is a function of $\mathbf{y}$ that is identically 0. Assuming that $\left|\partial^2 H(\mathbf{c}|\boldsymbol{\theta},\mathbf{y})/\partial \mathbf{c}^2|_{\hat{\mathbf{c}}}\right| \neq 0$, from the Implicit Function Theorem we obtain

$$\frac{\partial \hat{\mathbf{c}}}{\partial \mathbf{y}} = -\left[\frac{\partial^2 H(\mathbf{c}|\boldsymbol{\theta},\mathbf{y})}{\partial \mathbf{c}^2}\bigg|_{\hat{\mathbf{c}}}\right]^{-1}\left[\frac{\partial^2 H(\mathbf{c}|\boldsymbol{\theta},\mathbf{y})}{\partial \mathbf{c}\partial \mathbf{y}}\bigg|_{\hat{\mathbf{c}}}\right]. \qquad \text{(A.2)}$$

- $\frac{\partial \mathbf{c}^2}{\partial \boldsymbol{\theta}\partial \mathbf{y}}$

We take the y_k- derivative on both sides of Equation (2.6):

$$\begin{aligned}\frac{d^2}{d\boldsymbol{\theta}dy_k}\left(\frac{\partial H(\mathbf{c}|\boldsymbol{\theta},\mathbf{y})}{\partial \mathbf{c}}\bigg|_{\hat{\mathbf{c}}}\right) &= \frac{\partial^3 H(\mathbf{c}|\boldsymbol{\theta},\mathbf{y})}{\partial \mathbf{c}\partial\boldsymbol{\theta}\partial y_k}\bigg|_{\hat{\mathbf{c}}} + \frac{\partial^3 H(\mathbf{c}|\boldsymbol{\theta},\mathbf{y})}{\partial \mathbf{c}\partial\boldsymbol{\theta}\partial c_i}\bigg|_{\hat{\mathbf{c}}}\frac{\partial \hat{c}_i}{\partial y_k} \\ &+ \frac{\partial^3 H(\mathbf{c}|\boldsymbol{\theta},\mathbf{y})}{\partial \mathbf{c}^2\partial y_k}\bigg|_{\hat{\mathbf{c}}}\frac{\partial \hat{\mathbf{c}}}{\partial \boldsymbol{\theta}} + \frac{\partial^3 H(\mathbf{c}|\boldsymbol{\theta},\mathbf{y})}{\partial \mathbf{c}^2\partial c_i}\bigg|_{\hat{\mathbf{c}}}\frac{\partial \hat{c}_i}{\partial y_k}\frac{\partial \hat{\mathbf{c}}}{\partial \boldsymbol{\theta}} + \frac{\partial^2 H(\mathbf{c}|\boldsymbol{\theta},\mathbf{y})}{\partial \mathbf{c}^2}\bigg|_{\hat{\mathbf{c}}}\frac{\partial^2 \hat{\mathbf{c}}}{\partial \boldsymbol{\theta}\partial y_k} \\ &= 0 \qquad \text{(A.3)}\end{aligned}$$

Solving for $\frac{\partial^2 \hat{\mathbf{c}}}{\partial \boldsymbol{\theta}\partial y_k}$, we obtain the second derivative of $\hat{\mathbf{c}}$ with respect to $\boldsymbol{\theta}$ and

y_k:

$$\frac{\partial^2 \hat{\mathbf{c}}}{\partial \boldsymbol{\theta} \partial y_k} = -\left[\frac{\partial^2 H(\mathbf{c}|\boldsymbol{\theta}, \mathbf{y})}{\partial \mathbf{c}^2}\bigg|_{\hat{\mathbf{c}}}\right]^{-1}\left[\frac{\partial^3 H(\mathbf{c}|\boldsymbol{\theta}, \mathbf{y})}{\partial \mathbf{c} \partial \boldsymbol{\theta} \partial y_k}\bigg|_{\hat{\mathbf{c}}} + \frac{\partial^3 H(\mathbf{c}|\boldsymbol{\theta}, \mathbf{y})}{\partial \mathbf{c} \partial \boldsymbol{\theta} \partial c_i}\bigg|_{\hat{\mathbf{c}}} \frac{\partial \hat{c}_i}{\partial y_k} + \frac{\partial^3 H(\mathbf{c}|\boldsymbol{\theta}, \mathbf{y})}{\partial \mathbf{c}^2 \partial y_k}\bigg|_{\hat{\mathbf{c}}} \frac{\partial \hat{\mathbf{c}}}{\partial \boldsymbol{\theta}} + \frac{\partial^3 H(\mathbf{c}|\boldsymbol{\theta}, \mathbf{y})}{\partial \mathbf{c}^2 \partial c_i}\bigg|_{\hat{\mathbf{c}}} \frac{\partial \hat{c}_i}{\partial y_k} \frac{\partial \hat{\mathbf{c}}}{\partial \boldsymbol{\theta}}\right] \tag{A.4}$$

- $\frac{\partial^2 \hat{\mathbf{c}}}{\partial \boldsymbol{\theta}^2}$

  Similar to (A.4), the second partial derivative of $\mathbf{c}$ with respect to $\boldsymbol{\theta}$ and θ_j is:

$$\frac{\partial^2 \hat{\mathbf{c}}}{\partial \boldsymbol{\theta} \partial \theta_j} = -\left[\frac{\partial^2 H(\mathbf{c}|\boldsymbol{\theta}, \mathbf{y})}{\partial \mathbf{c}^2}\bigg|_{\hat{\mathbf{c}}}\right]^{-1}\left[\frac{\partial^3 H(\mathbf{c}|\boldsymbol{\theta}, \mathbf{y})}{\partial \mathbf{c} \partial \boldsymbol{\theta} \partial \theta_j}\bigg|_{\hat{\mathbf{c}}} + \frac{\partial^3 H(\mathbf{c}|\boldsymbol{\theta}, \mathbf{y})}{\partial \mathbf{c} \partial \boldsymbol{\theta} \partial c_i}\bigg|_{\hat{\mathbf{c}}} \frac{\partial \hat{c}_i}{\partial \theta_j} + \frac{\partial^3 H(\mathbf{c}|\boldsymbol{\theta}, \mathbf{y})}{\partial \mathbf{c}^2 \partial \theta_j}\bigg|_{\hat{\mathbf{c}}} \frac{\partial \hat{\mathbf{c}}}{\partial \boldsymbol{\theta}} + \frac{\partial^3 H(\mathbf{c}|\boldsymbol{\theta}, \mathbf{y})}{\partial \mathbf{c}^2 \partial c_i}\bigg|_{\hat{\mathbf{c}}} \frac{\partial \hat{c}_i}{\partial \theta_j} \frac{\partial \hat{\mathbf{c}}}{\partial \boldsymbol{\theta}}\right] \tag{A.5}$$

A.2 Matrix Calculations for Adaptive Penalized Smoothing

We provide here the results required for estimates of pointwise standard errors of the complexity function $\omega(t)$ in adaptive penalized smoothing (Section 2.4). In order to simplify notation, we define the order K_c matrix $\mathbf{B}(\lambda) = \boldsymbol{\Phi}'\mathbf{W}\boldsymbol{\Phi} + \mathbf{R}$ and order n matrix $\mathbf{A}(\lambda) = \boldsymbol{\Phi}\mathbf{B}(\lambda)^{-1}\boldsymbol{\Phi}'\mathbf{W}$. Then we can express $\mathtt{SSE}(\lambda)$ and degrees of freedom measure $\mathtt{dfe}(\lambda)$ in terms of the matrix A:

$$\mathtt{SSE}(\lambda) = \mathbf{y}'[I - \mathbf{A}(\lambda)]'[I - \mathbf{A}(\lambda)]\mathbf{y}$$

$$\mathtt{dfe}(\lambda) = n - \mathtt{Tr}(\mathbf{A}(\lambda))$$

In what follows, we suppress the explicit dependence of these three matrices on λ and the parameter vector $\boldsymbol{\theta}$ in order to keep the notation readable.

- The first derivatives with respect to the $\omega(t)$ basis coefficient θ_l of these three matrices are:

$$\begin{aligned}
\frac{\partial \mathbf{R}}{\partial \theta_l} &= \int \lambda(t)\psi_l(t)[L\boldsymbol{\phi}(t)][L\boldsymbol{\phi}(t)]'dt \\
\frac{\partial \mathbf{B}^{-1}}{\partial \theta_l} &= -\mathbf{B}^{-1}\frac{\partial \mathbf{R}}{\partial \theta_l}\mathbf{B}^{-1} \\
\frac{\partial \mathbf{A}}{\partial \theta_l} &= \boldsymbol{\Phi}\frac{\partial \mathbf{B}^{-1}}{\partial \theta_l}\boldsymbol{\Phi}'\mathbf{W}
\end{aligned}$$

- the second derivatives with respect to the smoothing function basis coefficients θ_l and θ_i are:

$$\begin{aligned}
\frac{\partial^2 \mathbf{R}}{\partial \theta_l \partial \theta_i} &= \int \lambda(t)\psi_i(t)\psi_l(t)[L\boldsymbol{\phi}(t)][L\boldsymbol{\phi}(t)]'dt \\
\frac{\partial^2 \mathbf{B}^{-1}}{\partial \theta_l \partial \theta_i} &= -\frac{\partial \mathbf{B}^{-1}}{\partial \theta_i}\frac{\partial \mathbf{R}}{\partial \theta_l}\mathbf{B}^{-1} - \mathbf{B}^{-1}\frac{\partial^2 \mathbf{R}}{\partial \theta_l \partial \theta_i}\mathbf{B}^{-1} - \mathbf{B}^{-1}\frac{\partial \mathbf{R}}{\partial \theta_l}\frac{\partial \mathbf{B}^{-1}}{\partial \theta_i} \\
\frac{\partial^2 \mathbf{A}}{\partial \theta_l \partial \theta_i} &= \boldsymbol{\Phi}\frac{\partial^2 \mathbf{B}^{-1}}{\partial \theta_l \partial \theta_i}\boldsymbol{\Phi}'\mathbf{W}
\end{aligned}$$

- The first derivative of $\mathtt{GCV}(\lambda(t)|\mathbf{y})$ with respect to $\omega(t)$ basis coefficient θ_l is

$$\frac{\partial \mathtt{GCV}(\lambda)}{\partial \theta_l} = n\left[\mathtt{dfe}\frac{\partial \mathtt{SSE}}{\partial \theta_l} - 2\mathtt{SSE}\frac{\partial \mathtt{dfe}}{\partial \theta_l}\right]\mathtt{dfe}^{-3} \tag{A.6}$$

where

$$\begin{aligned}\frac{\partial \mathtt{dfe}(\lambda)}{\partial \theta_l} &= -\mathtt{Tr}(\frac{\partial \mathbf{A}}{\partial \theta_l}) \\ \frac{\partial \mathtt{SSE}(\lambda)}{\partial \theta_l} &= -\mathbf{y}'([\frac{\partial \mathbf{A}}{\partial \theta_l}]'[\mathbf{I}-\mathbf{A}]+[\mathbf{I}-\mathbf{A}]'[\frac{\partial \mathbf{A}}{\partial \theta_l}])\mathbf{y}\end{aligned}$$

- The second derivative of $\mathtt{GCV}(\lambda(t)|\mathbf{y})$ with respect to $\omega(t)$ basis coefficients θ_l and θ_j is

$$\begin{aligned}\frac{\partial^2 \mathtt{GCV}(\lambda)}{\partial \theta_l \partial \theta_j} &= \frac{n}{\mathtt{dfe}^2}\frac{\partial^2 \mathtt{SSE}}{\partial \theta_l \partial \theta_j} - \frac{2n\mathtt{SSE}}{\mathtt{dfe}^3}\frac{\partial^2 \mathtt{dfe}}{\partial \theta_l \partial \theta_j} + \frac{6n\mathtt{SSE}}{\mathtt{dfe}^4}\frac{\partial \mathtt{dfe}}{\partial \theta_l}\frac{\partial \mathtt{dfe}}{\partial \theta_j} \\ &\quad -\frac{2n}{\mathtt{dfe}^3}\left[\frac{\partial \mathtt{dfe}}{\partial \theta_l}\frac{\partial \mathtt{SSE}}{\partial \theta_j}+\frac{\partial \mathtt{dfe}}{\partial \theta_j}\frac{\partial \mathtt{SSE}}{\partial \theta_l}\right] \qquad \text{(A.7)}\end{aligned}$$

where

$$\begin{aligned}\frac{\partial^2 \mathtt{SSE}(\lambda)}{\partial \theta_l \partial \theta_j} &= \mathbf{y}'(E'+E)\mathbf{y} \\ \frac{\partial^2 \mathtt{dfe}(\lambda)}{\partial \theta_l \partial \theta_j} &= -\mathtt{Tr}(\frac{\partial^2 \mathbf{A}}{\partial \theta_l \partial \theta_j})\end{aligned}$$

and

$$E = [\frac{\partial \mathbf{A}}{\partial \theta_l}]'[\frac{\partial \mathbf{A}}{\partial \theta_j}] - [\frac{\partial^2 \mathbf{A}}{\partial \theta_l \partial \theta_j}]'[\mathbf{I}-\mathbf{A}].$$

- The second derivative of $\mathtt{GCV}(\lambda(t)|\mathbf{y})$ with respect to $\omega(t)$ basis coefficients θ_l and $\mathbf{y}$ is

$$\frac{\partial^2 \mathtt{GCV}(\lambda)}{\partial \theta_l \partial \mathbf{y}} = n\left[\mathtt{dfe}\frac{\partial^2 \mathtt{SSE}}{\partial \theta_l \partial \mathbf{y}} - 2\frac{\partial \mathtt{SSE}}{\partial \mathbf{y}}\frac{\partial \mathtt{dfe}}{\partial \theta_l}\right]\mathtt{dfe}^{-3} \qquad \text{(A.8)}$$

where

$$\begin{aligned}\frac{\partial \mathrm{SSE}(\lambda)}{\partial \mathbf{y}} &= 2[\mathbf{I}-\mathbf{A}]'[\mathbf{I}-\mathbf{A}]\mathbf{y} \\ \frac{\partial^2 \mathrm{SSE}(\lambda)}{\partial\theta_l \partial \mathbf{y}} &= -2\left\{[\frac{\partial \mathbf{A}}{\partial \theta_l}]'[\mathbf{I}-\mathbf{A}] + [\mathbf{I}-\mathbf{A}]'\frac{\partial \mathbf{A}}{\partial \theta_l}\right\}\mathbf{y}.\end{aligned}$$

- The sampling variance of $\omega(t) = \ln \lambda(t)$ is estimated by:

$$\mathtt{Var}(\omega(t)) = (\frac{d\omega}{d\mathbf{y}})'\boldsymbol{\Sigma}(\frac{d\omega}{d\mathbf{y}}) \tag{A.9}$$

where

$$\frac{d\omega}{d\mathbf{y}} = \left(\frac{d\boldsymbol{\theta}}{d\mathbf{y}}\right)'\boldsymbol{\psi}(t) \quad \text{and} \quad \frac{d\boldsymbol{\theta}}{d\mathbf{y}} = [\frac{\partial^2 \mathtt{GCV}(\lambda)}{\partial^2 \boldsymbol{\theta}}]^{-1}\frac{\partial^2 \mathtt{GCV}(\lambda)}{\partial \boldsymbol{\theta}\partial \mathbf{y}}.$$

- Since the estimated curve $\hat{\mathbf{x}}(t) = \boldsymbol{\phi}'(t)\hat{\mathbf{c}}$, we can estimate the sampling variance of $\hat{\mathbf{x}}(t)$ by

$$\mathtt{Var}[\hat{\mathbf{x}}(t)] = \boldsymbol{\phi}'(t)\mathtt{Var}(\hat{\mathbf{c}})\boldsymbol{\phi}(t). \tag{A.10}$$

Appendix B

Derivative Calculation for Estimating Generalized Semiparametric Additive Models

we develop a method to estimate the generalized semiparametric additive models, working for arbitrarily distributed response variables. The nonparametric functions are estimated by penalized smoothing. The smoothing parameter vector $\boldsymbol{\lambda}$ controls the smoothness of functional parameters. We use the generalized profiling method to estimate three distinct groups of parameters: the coefficient vector $\mathbf{c}$, the linear coefficient vector $\boldsymbol{\beta}$, and the smoothing parameter vector $\boldsymbol{\lambda}$ and their standard deviations, assuming that observations can be in any distribution. These three parameter vectors can also be multidimensional. The unconditionally estimated standard deviation of $\boldsymbol{\beta}$ includes variation coming from $\boldsymbol{\lambda}$ and solves the

underestimation problem. Three levels of optimization procedures are conducted: first, the coefficient vector $\mathbf{c}$ is estimated, given $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$, by maximizing the regularized log likelihood function $J(\mathbf{c}|\boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{y})$. Hence, the optimal coefficient vector $\hat{\mathbf{c}}$ is a function of $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$. Next, the linear coefficient vector $\boldsymbol{\beta}$, given $\boldsymbol{\lambda}$, is estimated by maximizing the log likelihood function $H(\boldsymbol{\beta}|\boldsymbol{\lambda}, \mathbf{y})$. Therefore, the optimal linear coefficient vector $\hat{\boldsymbol{\beta}}$ is a function of $\boldsymbol{\lambda}$. Finally, the smoothing parameter vector is estimated by minimizing the criterion $F(\boldsymbol{\lambda}|\mathbf{y})$, which can be defined by any model selection methods.

The Newton-Raphson algorithm is used to do all three levels of optimization. The algorithm seems to converge quickly and stably. In the following, we write out the optimization criteria along with the gradients and Hessian matrices analytically.

The functional parameters $f_i(Z_i)$ are estimated by linear combinations of K_i B-spline basis functions:

$$f_i(Z_i) = \sum_{k=1}^{K_i} c_{ik}\phi_{ik}(Z_i) = \mathbf{c}_i'\boldsymbol{\phi}_i(Z_i)\,,$$

where $\mathbf{c}_i = (c_{i1}, \cdots, c_{iK_i})$ and $\boldsymbol{\phi}_i(Z_i) = \Big(\phi_{i1}(Z_i), \cdots, \phi_{iK_i}(Z_i)\Big)'$. Let $\boldsymbol{\Phi}_i$ be an order $n \times K_i$ matrix with the j-th row $\boldsymbol{\phi}_i(Z_{ij})'$, then the generalized semiparametric additive model (3.1) can be written in the simple matrix form:

$$\eta_j = g(\mu_j) = \boldsymbol{\Phi}\mathbf{c} + \mathbf{X}\boldsymbol{\beta}\,, \tag{B.1}$$

where $\mathbf{c} = (\mathbf{c}_1', \cdots, \mathbf{c}_P')'$, $\boldsymbol{\Phi} = (\boldsymbol{\Phi}_1, \cdots, \boldsymbol{\Phi}_P)'$ and $\mathbf{X}$ is an $n \times Q$ matrix with jk-th

entry x_{kj}.

B.1 First Optimization Level to Estimate Local Parameters

The optimization criterion in the first level is written as:

$$J(\mathbf{c}|\boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{y}) = -l(\mathbf{c}, \boldsymbol{\beta}|\mathbf{y}) + \sum_{i=1}^{P} \lambda_i \int [L_i f_i(Z_i)]^2 dZ_i \,, \tag{B.2}$$

where $l(\mathbf{c}, \boldsymbol{\beta}|\mathbf{y})$ is the log likelihood function. The second term in (3.4) penalizes the roughness of functional parameters, so a positive sign is used in front of it such that the optimal coefficient vector $\mathbf{c}$ can be estimated by minimizing $J(\mathbf{c}|\boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{y})$. L_i is a linear differential operator of order m:

$$L_i x(t) = \sum_{j=0}^{m-1} \alpha_j(t) D^j x(t) + D^m x(t) \,.$$

The penalty term $\int [L_i f_i(Z_i)]^2 dZ_i$ can be written as a quadratic function of the coefficient vector $\mathbf{c}_i$ when the differential operator is linear:

$$\int [L_i f_i(Z_i)]^2 dZ_i = \mathbf{c}_i' \mathbf{R}_i \mathbf{c}_i \,,$$

where $\mathbf{R}_i = \int [L_i\boldsymbol{\phi}_i(t)][L\boldsymbol{\phi}_i(t)]'dt$ is an order K_i matrix. Then the second term in (3.4) can be represented in the matrix form:

$$\sum_{i=1}^{P} \lambda_i \int [L_i f_i(Z_i)]^2 dZ_i = \mathbf{c}'\mathbf{R}\mathbf{c}\,,$$

where $\mathbf{c} = (\mathbf{c}_1', \cdots, \mathbf{c}_P')'$ and $\mathbf{R} = \text{diag}(\lambda_1\mathbf{R}_1, \cdots, \lambda_P\mathbf{R}_P)$. In order to attain a positive estimate for the smoothing parameter vector, we express $\boldsymbol{\lambda} = (\lambda_1, \cdots, \lambda_P)' = \exp(\boldsymbol{\theta})$, where $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_P)'$. All simulations and applications in this chapter use the second derivative to define the roughness penalty term, that is, $L = D^2$, but Ramsay and Silverman (2005) shows how to obtain better estimates by penalized smoothing with penalty terms defined by differential operators. The first and second derivatives of $J(\mathbf{c}|\boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{y})$ with respect to $\mathbf{c}$ are given in (B.3) and (B.4), respectively.

For given values of $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$, the coefficient vector $\mathbf{c}$ can be estimated by minimizing the optimization criterion (3.4) in the first level, so that the estimated $\hat{\mathbf{c}}$ can be viewed as a function of $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$. However, there is no explicit form of this function except when observations are normally distributed. That is why least squares estimations are often used in many of the literature, instead of likelihood functions. Fortunately, we can write out any order derivatives of $\hat{\mathbf{c}}$ with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$ analytically using Implicit Function Theorem, which are shown below.

- $\frac{\partial J}{\partial \mathbf{c}}$

The first derivative of $J(\mathbf{c}|\boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{y})$ with respect to $\mathbf{c}$ is:

$$\frac{\partial J}{\partial \mathbf{c}} = -\frac{\partial l}{\partial \mathbf{c}} + 2\mathbf{R}\mathbf{c} \tag{B.3}$$

- $\frac{\partial^2 J}{\partial \mathbf{c}^2}$

The second derivative of $J(\mathbf{c}|\boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{y})$ with respect to $\mathbf{c}$ is:

$$\frac{\partial^2 J}{\partial \mathbf{c}^2} = -\frac{\partial^2 l}{\partial \mathbf{c}^2} + 2\mathbf{R} \tag{B.4}$$

- $\frac{\partial \mathbf{c}}{\partial \boldsymbol{\beta}}$

For any given $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$, there exist one optimal coefficient vector $\mathbf{c}$ by minimizing $J(\mathbf{c}|\boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{y})$, so $\mathbf{c}$ is a function of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$. According (2.7), the first partial derivative of $\mathbf{c}$ with respect to $\boldsymbol{\beta}$ is:

$$\frac{\partial \mathbf{c}}{\partial \boldsymbol{\beta}} = -\left(\frac{\partial^2 J}{\partial \mathbf{c}^2}\right)^{-1} \frac{\partial^2 J}{\partial \mathbf{c} \partial \boldsymbol{\beta}} \tag{B.5}$$

where

$$\frac{\partial^2 J}{\partial \mathbf{c} \partial \boldsymbol{\beta}} = -\frac{\partial^2 l}{\partial \mathbf{c} \partial \boldsymbol{\beta}}$$

- $\frac{\partial \mathbf{c}}{\partial \boldsymbol{\theta}}$

Similarly, according (2.7), the first partial derivative of $\mathbf{c}$ with respect to

$\boldsymbol{\theta}$ is:

$$\frac{\partial \mathbf{c}}{\partial \boldsymbol{\theta}} = -\left(\frac{\partial^2 J}{\partial \mathbf{c}^2}\right)^{-1} \frac{\partial^2 J}{\partial \mathbf{c} \partial \boldsymbol{\theta}} \tag{B.6}$$

where

$$\frac{\partial^2 J}{\partial \mathbf{c} \partial \boldsymbol{\theta}} = 2 \frac{\partial \mathbf{R}}{\partial \boldsymbol{\theta}} \mathbf{c}$$

- $\frac{\partial^2 \mathbf{c}}{\partial \boldsymbol{\beta}^2}$

We can take the β_j-derivative on $D_{\boldsymbol{\beta}} D_{\mathbf{c}} J$:

$$D_{\beta_j} D_{\boldsymbol{\beta}} D_{\mathbf{c}} J = \frac{\partial^3 J}{\partial \mathbf{c} \partial \boldsymbol{\beta} \partial \beta_j} + \frac{\partial^3 J}{\partial \mathbf{c} \partial \boldsymbol{\beta} \partial c_i} \frac{\partial c_i}{\partial \beta_j} + \frac{\partial^3 J}{\partial \mathbf{c}^2 \partial c_i} \frac{\partial c_i}{\partial \beta_j} \frac{\partial \mathbf{c}}{\partial \boldsymbol{\beta}} + \frac{\partial^3 J}{\partial \mathbf{c}^2 \partial \beta_j} \frac{\partial \mathbf{c}}{\partial \boldsymbol{\beta}} + \frac{\partial^2 J}{\partial \mathbf{c}^2} \frac{\partial^2 \mathbf{c}}{\partial \boldsymbol{\beta} \partial \beta_j} = 0$$

and from the Implicit Function Theorem we obtain

$$\frac{\partial^2 \mathbf{c}}{\partial \boldsymbol{\beta} \partial \beta_j} = -\left(\frac{\partial^2 J}{\partial \mathbf{c}^2}\right)^{-1} \left[\frac{\partial^3 J}{\partial \mathbf{c} \partial \boldsymbol{\beta} \partial \beta_j} + \frac{\partial^3 J}{\partial \mathbf{c} \partial \boldsymbol{\beta} \partial c_i} \frac{\partial c_i}{\partial \beta_j} + \frac{\partial^3 J}{\partial \mathbf{c}^2 \partial c_i} \frac{\partial c_i}{\partial \beta_j} \frac{\partial \mathbf{c}}{\partial \boldsymbol{\beta}} + \frac{\partial^3 J}{\partial \mathbf{c}^2 \partial \beta_j} \frac{\partial \mathbf{c}}{\partial \boldsymbol{\beta}} \right] \tag{B.7}$$

where

$$\frac{\partial^3 J}{\partial \mathbf{c} \partial \boldsymbol{\beta}^2} = -\frac{\partial^3 l}{\partial \mathbf{c} \partial \boldsymbol{\beta}^2}$$

$$\frac{\partial^3 J}{\partial \mathbf{c}^2 \partial \boldsymbol{\beta}} = -\frac{\partial^3 l}{\partial \mathbf{c}^2 \partial \boldsymbol{\beta}}$$

$$\frac{\partial^3 J}{\partial \mathbf{c}^3} = -\frac{\partial^3 l}{\partial \mathbf{c}^3}$$

- $\frac{\partial \mathbf{c}^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\theta}}$

$$D_{\boldsymbol{\beta}} D_{\mathbf{c}} J = \frac{\partial^2 J}{\partial \mathbf{c} \partial \boldsymbol{\beta}} + \frac{\partial^2 J}{\partial \mathbf{c}^2} \frac{\partial \mathbf{c}}{\partial \boldsymbol{\beta}} = 0 \tag{B.8}$$

We then take the θ_k- derivative on $D_{\boldsymbol{\beta}} D_{\mathbf{c}} J$:

$$D_{\theta_k} D_{\boldsymbol{\beta}} D_{\mathbf{c}} J = \frac{\partial^3 J}{\partial \mathbf{c} \partial \boldsymbol{\beta} \partial \theta_k} + \frac{\partial^3 J}{\partial \mathbf{c} \partial \boldsymbol{\beta} \partial c_i} \frac{\partial c_i}{\partial \theta_k} + \frac{\partial^3 J}{\partial \mathbf{c}^2 \partial \theta_k} \frac{\partial \mathbf{c}}{\partial \boldsymbol{\beta}} + \frac{\partial^3 J}{\partial \mathbf{c}^2 \partial c_i} \frac{\partial c_i}{\partial \theta_k} \frac{\partial \mathbf{c}}{\partial \boldsymbol{\beta}} + \frac{\partial^2 J}{\partial \mathbf{c}^2} \frac{\partial^2 \mathbf{c}}{\partial \boldsymbol{\beta} \partial \theta_k} = 0 \tag{B.9}$$

Solving for $\frac{\partial^2 \mathbf{c}}{\partial \boldsymbol{\beta} \partial \theta_k}$, we obtain the second derivative of $\mathbf{c}$ with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$:

$$\frac{\partial^2 \mathbf{c}}{\partial \boldsymbol{\beta} \partial \theta_k} = -\left(\frac{\partial^2 J}{\partial \mathbf{c}^2} \right)^{-1} \left[\frac{\partial^3 J}{\partial \mathbf{c} \partial \boldsymbol{\beta} \partial \theta_k} + \frac{\partial^3 J}{\partial \mathbf{c} \partial \boldsymbol{\beta} \partial c_i} \frac{\partial c_i}{\partial \theta_k} + \frac{\partial^3 J}{\partial \mathbf{c}^2 \partial \theta_k} \frac{\partial \mathbf{c}}{\partial \boldsymbol{\beta}} + \frac{\partial^3 J}{\partial \mathbf{c}^2 \partial c_i} \frac{\partial c_i}{\partial \theta_k} \frac{\partial \mathbf{c}}{\partial \boldsymbol{\beta}} \right] \tag{B.10}$$

where

$$\frac{\partial^3 J}{\partial \mathbf{c} \partial \boldsymbol{\beta} \partial \boldsymbol{\theta}} = 0$$

$$\frac{\partial^3 J}{\partial \mathbf{c}^2 \partial \boldsymbol{\beta}} = -\frac{\partial^3 l}{\partial \mathbf{c}^2 \partial \boldsymbol{\beta}}$$

$$\frac{\partial^3 J}{\partial \mathbf{c}^2 \partial \boldsymbol{\theta}} = 2\mathbf{R}$$

$$\frac{\partial^3 J}{\partial \mathbf{c}^3} = -\frac{\partial^3 l}{\partial \mathbf{c}^3}$$

- $\frac{\partial \mathbf{c}^2}{\partial \boldsymbol{\beta} \partial \mathbf{y}}$

We take the y_k- derivative on $D_{\boldsymbol{\beta}} D_{\mathbf{c}} J$:

$$D_{y_k} D_{\boldsymbol{\beta}} D_{\mathbf{c}} J = \frac{\partial^3 J}{\partial \mathbf{c} \partial \boldsymbol{\beta} \partial y_k} + \frac{\partial^3 J}{\partial \mathbf{c} \partial \boldsymbol{\beta} \partial c_i} \frac{\partial c_i}{\partial y_k} + \frac{\partial^3 J}{\partial \mathbf{c}^2 \partial y_k} \frac{\partial \mathbf{c}}{\partial \boldsymbol{\beta}} + \frac{\partial^3 J}{\partial \mathbf{c}^2 \partial c_i} \frac{\partial c_i}{\partial y_k} \frac{\partial \mathbf{c}}{\partial \boldsymbol{\beta}} + \frac{\partial^2 J}{\partial \mathbf{c}^2} \frac{\partial^2 \mathbf{c}}{\partial \boldsymbol{\beta} \partial y_k} = 0 \tag{B.11}$$

Solving for $\frac{\partial^2 \mathbf{c}}{\partial \boldsymbol{\beta} \partial y_k}$, we obtain the second derivative of $\mathbf{c}$ with respect to $\boldsymbol{\beta}$ and y_k:

$$\frac{\partial^2 \mathbf{c}}{\partial \boldsymbol{\beta} \partial y_k} = -\left(\frac{\partial^2 J}{\partial \mathbf{c}^2}\right)^{-1}\left[\frac{\partial^3 J}{\partial \mathbf{c} \partial \boldsymbol{\beta} \partial y_k} + \frac{\partial^3 J}{\partial \mathbf{c} \partial \boldsymbol{\beta} \partial c_i}\frac{\partial c_i}{\partial y_k} + \frac{\partial^3 J}{\partial \mathbf{c}^2 \partial y_k}\frac{\partial \mathbf{c}}{\partial \boldsymbol{\beta}} + \frac{\partial^3 J}{\partial \mathbf{c}^2 \partial c_i}\frac{\partial c_i}{\partial y_k}\frac{\partial \mathbf{c}}{\partial \boldsymbol{\beta}}\right] \tag{B.12}$$

- $\frac{\partial \mathbf{c}^2}{\partial \boldsymbol{\theta} \partial \mathbf{y}}$

We take the y_k- derivative on $D_{\boldsymbol{\theta}} D_{\mathbf{c}} J$:

$$D_{y_k} D_{\boldsymbol{\theta}} D_{\mathbf{c}} J = \frac{\partial^3 J}{\partial \mathbf{c} \partial \boldsymbol{\theta} \partial y_k} + \frac{\partial^3 J}{\partial \mathbf{c} \partial \boldsymbol{\theta} \partial c_i}\frac{\partial c_i}{\partial y_k} + \frac{\partial^3 J}{\partial \mathbf{c}^2 \partial y_k}\frac{\partial \mathbf{c}}{\partial \boldsymbol{\theta}} + \frac{\partial^3 J}{\partial \mathbf{c}^2 \partial c_i}\frac{\partial c_i}{\partial y_k}\frac{\partial \mathbf{c}}{\partial \boldsymbol{\theta}} + \frac{\partial^2 J}{\partial \mathbf{c}^2}\frac{\partial^2 \mathbf{c}}{\partial \boldsymbol{\theta} \partial y_k} = 0 \tag{B.13}$$

Solving for $\frac{\partial^2 \mathbf{c}}{\partial \boldsymbol{\theta} \partial y_k}$, we obtain the second derivative of $\mathbf{c}$ with respect to $\boldsymbol{\theta}$ and y_k:

$$\frac{\partial^2 \mathbf{c}}{\partial \boldsymbol{\theta} \partial y_k} = -\left(\frac{\partial^2 J}{\partial \mathbf{c}^2}\right)^{-1}\left[\frac{\partial^3 J}{\partial \mathbf{c} \partial \boldsymbol{\theta} \partial y_k} + \frac{\partial^3 J}{\partial \mathbf{c} \partial \boldsymbol{\theta} \partial c_i}\frac{\partial c_i}{\partial y_k} + \frac{\partial^3 J}{\partial \mathbf{c}^2 \partial y_k}\frac{\partial \mathbf{c}}{\partial \boldsymbol{\theta}} + \frac{\partial^3 J}{\partial \mathbf{c}^2 \partial c_i}\frac{\partial c_i}{\partial y_k}\frac{\partial \mathbf{c}}{\partial \boldsymbol{\theta}}\right] \tag{B.14}$$

- $\frac{\partial \mathbf{c}}{\partial \mathbf{y}}$

According (2.7), the first partial derivative of $\mathbf{c}$ with respect to $\mathbf{y}$ is:

$$\frac{\partial \mathbf{c}}{\partial \mathbf{y}} = -\left(\frac{\partial^2 J}{\partial \mathbf{c}^2}\right)^{-1}\frac{\partial^2 J}{\partial \mathbf{c} \partial \mathbf{y}} \tag{B.15}$$

where

$$\frac{\partial^2 J}{\partial \mathbf{c} \partial \mathbf{y}} = -\frac{\partial^2 l}{\partial \mathbf{c} \partial \mathbf{y}} \tag{B.16}$$

- $\frac{\partial^2 \mathbf{c}}{\partial \boldsymbol{\theta}^2}$

Similar to (B.14), the second partial derivative of $\mathbf{c}$ with respect to $\boldsymbol{\theta}$ is:

$$\frac{\partial^2 \mathbf{c}}{\partial \boldsymbol{\theta} \partial \theta_j} = -\left(\frac{\partial^2 J}{\partial \mathbf{c}^2}\right)^{-1} \left[\frac{\partial^3 J}{\partial \mathbf{c} \partial \boldsymbol{\theta} \partial \theta_j} + \frac{\partial^3 J}{\partial \mathbf{c} \partial \boldsymbol{\theta} \partial c_i} \frac{\partial c_i}{\partial \theta_j} + \frac{\partial^3 J}{\partial \mathbf{c}^2 \partial c_i} \frac{\partial c_i}{\partial \theta_j} \frac{\partial \mathbf{c}}{\partial \boldsymbol{\theta}} + \frac{\partial^3 J}{\partial \mathbf{c}^2 \partial \theta_j} \frac{\partial \mathbf{c}}{\partial \boldsymbol{\theta}}\right] \tag{B.17}$$

where

$$\frac{\partial^3 J}{\partial \mathbf{c} \partial \boldsymbol{\theta}^2} = 2\mathbf{R}\mathbf{c}$$

$$\frac{\partial^3 J}{\partial \mathbf{c}^2 \partial \boldsymbol{\theta}} = 2\mathbf{R}$$

$$\frac{\partial^3 J}{\partial \mathbf{c}^3} = -\frac{\partial^3 l}{\partial \mathbf{c}^3}$$

- $\frac{\partial^3 \mathbf{c}}{\partial \boldsymbol{\beta}^3}$

We can take the β_j-derivative on $D_{\boldsymbol{\beta}} D_{\mathbf{c}} J$:

$$D_{\beta_j} D_{\boldsymbol{\beta}} D_{\mathbf{c}} J = \frac{\partial^3 J}{\partial \mathbf{c} \partial \boldsymbol{\beta} \partial \beta_j} + \frac{\partial^3 J}{\partial \mathbf{c} \partial \boldsymbol{\beta} \partial c_i} \frac{\partial c_i}{\partial \beta_j} + \frac{\partial^3 J}{\partial \mathbf{c}^2 \partial c_i} \frac{\partial c_i}{\partial \beta_j} \frac{\partial \mathbf{c}}{\partial \boldsymbol{\beta}} + \frac{\partial^3 J}{\partial \mathbf{c}^2 \partial \beta_j} \frac{\partial \mathbf{c}}{\partial \boldsymbol{\beta}} + \frac{\partial^2 J}{\partial \mathbf{c}^2} \frac{\partial^2 \mathbf{c}}{\partial \boldsymbol{\beta} \partial \beta_j} = 0$$

Then taking the β_k-derivative on $D_{\beta_j} D_{\boldsymbol{\beta}} D_{\mathbf{c}} J$, we obtain:

$$\begin{aligned}
D_{\beta_k} D_{\beta_j} D_{\boldsymbol{\beta}} D_{\mathbf{c}} J &= \frac{\partial^4 J}{\partial \mathbf{c} \partial \boldsymbol{\beta} \partial \beta_j \partial c_g} \frac{\partial c_g}{\partial \beta_k} + \frac{\partial^4 J}{\partial \mathbf{c} \partial \boldsymbol{\beta} \partial \beta_j \partial \beta_k} \\
&+ \frac{\partial^4 J}{\partial \mathbf{c} \partial \boldsymbol{\beta} \partial c_i \partial c_g} \frac{\partial c_i}{\partial \beta_j} \frac{\partial c_g}{\partial \beta_k} + \frac{\partial^4 J}{\partial \mathbf{c} \partial \boldsymbol{\beta} \partial c_i \partial \beta_k} \frac{\partial c_i}{\partial \beta_j} + \frac{\partial^3 J}{\partial \mathbf{c} \partial \boldsymbol{\beta} \partial c_i} \frac{\partial^2 c_i}{\partial \beta_j \partial \beta_k} \\
&+ \frac{\partial^4 J}{\partial \mathbf{c}^2 \partial c_i \partial c_g} \frac{\partial c_g}{\partial \beta_k} \frac{\partial c_i}{\partial \beta_j} \frac{\partial \mathbf{c}}{\partial \boldsymbol{\beta}} + \frac{\partial^3 J}{\partial \mathbf{c}^2 \partial c_i} \frac{\partial^2 c_i}{\partial \beta_j \partial \beta_k} \frac{\partial \mathbf{c}}{\partial \boldsymbol{\beta}} + \frac{\partial^4 J}{\partial \mathbf{c}^2 \partial c_i \partial \beta_k} \frac{\partial c_i}{\partial \beta_j} \frac{\partial \mathbf{c}}{\partial \boldsymbol{\beta}} \\
&+ \frac{\partial^3 J}{\partial \mathbf{c}^2 \partial c_i} \frac{\partial c_i}{\partial \beta_j} \frac{\partial^2 \mathbf{c}}{\partial \boldsymbol{\beta} \partial \beta_k} + \frac{\partial^4 J}{\partial \mathbf{c}^2 \partial \beta_j \partial c_g} \frac{\partial c_g}{\partial \beta_k} \frac{\partial \mathbf{c}}{\partial \boldsymbol{\beta}} + \frac{\partial^4 J}{\partial \mathbf{c}^2 \partial \beta_j \partial \beta_k} \frac{\partial \mathbf{c}}{\partial \boldsymbol{\beta}} \\
&+ \frac{\partial^3 J}{\partial \mathbf{c}^2 \partial \beta_j} \frac{\partial^2 \mathbf{c}}{\partial \boldsymbol{\beta} \partial \beta_k} + \frac{\partial^3 J}{\partial \mathbf{c}^2 \partial c_g} \frac{\partial c_g}{\partial \beta_k} \frac{\partial^2 \mathbf{c}}{\partial \boldsymbol{\beta} \partial \beta_j} + \frac{\partial^3 J}{\partial \mathbf{c}^2 \partial \beta_k} \frac{\partial^2 \mathbf{c}}{\partial \boldsymbol{\beta} \partial \beta_j} \\
&+ \frac{\partial^2 J}{\partial \mathbf{c}^2} \frac{\partial^3 \mathbf{c}}{\partial \boldsymbol{\beta} \partial \beta_j \partial \beta_k} = 0 \qquad \text{(B.18)}
\end{aligned}$$

Solving for $\frac{\partial^3 \mathbf{c}}{\partial \boldsymbol{\beta} \partial \beta_j \partial \beta_k}$:

$$\begin{aligned}
\frac{\partial^3 \mathbf{c}}{\partial \boldsymbol{\beta} \partial \beta_j \partial \beta_k} &= -\left\{ \frac{\partial^2 J}{\partial \mathbf{c}^2} \right\}^{-1} \left\{ \frac{\partial^4 J}{\partial \mathbf{c} \partial \boldsymbol{\beta} \partial \beta_j \partial c_g} \frac{\partial c_g}{\partial \beta_k} + \frac{\partial^4 J}{\partial \mathbf{c} \partial \boldsymbol{\beta} \partial \beta_j \partial \beta_k} \right. \\
&+ \frac{\partial^4 J}{\partial \mathbf{c} \partial \boldsymbol{\beta} \partial c_i \partial c_g} \frac{\partial c_i}{\partial \beta_j} \frac{\partial c_g}{\partial \beta_k} + \frac{\partial^4 J}{\partial \mathbf{c} \partial \boldsymbol{\beta} \partial c_i \partial \beta_k} \frac{\partial c_i}{\partial \beta_j} + \frac{\partial^3 J}{\partial \mathbf{c} \partial \boldsymbol{\beta} \partial c_i} \frac{\partial^2 c_i}{\partial \beta_j \partial \beta_k} \\
&+ \frac{\partial^4 J}{\partial \mathbf{c}^2 \partial c_i \partial c_g} \frac{\partial c_g}{\partial \beta_k} \frac{\partial c_i}{\partial \beta_j} \frac{\partial \mathbf{c}}{\partial \boldsymbol{\beta}} + \frac{\partial^3 J}{\partial \mathbf{c}^2 \partial c_i} \frac{\partial^2 c_i}{\partial \beta_j \partial \beta_k} \frac{\partial \mathbf{c}}{\partial \boldsymbol{\beta}} + \frac{\partial^4 J}{\partial \mathbf{c}^2 \partial c_i \partial \beta_k} \frac{\partial c_i}{\partial \beta_j} \frac{\partial \mathbf{c}}{\partial \boldsymbol{\beta}} \\
&+ \frac{\partial^3 J}{\partial \mathbf{c}^2 \partial c_i} \frac{\partial c_i}{\partial \beta_j} \frac{\partial^2 \mathbf{c}}{\partial \boldsymbol{\beta} \partial \beta_k} + \frac{\partial^4 J}{\partial \mathbf{c}^2 \partial \beta_j \partial c_g} \frac{\partial c_g}{\partial \beta_k} \frac{\partial \mathbf{c}}{\partial \boldsymbol{\beta}} + \frac{\partial^4 J}{\partial \mathbf{c}^2 \partial \beta_j \partial \beta_k} \frac{\partial \mathbf{c}}{\partial \boldsymbol{\beta}} \\
&+ \left. \frac{\partial^3 J}{\partial \mathbf{c}^2 \partial \beta_j} \frac{\partial^2 \mathbf{c}}{\partial \boldsymbol{\beta} \partial \beta_k} + \frac{\partial^3 J}{\partial \mathbf{c}^2 \partial c_g} \frac{\partial c_g}{\partial \beta_k} \frac{\partial^2 \mathbf{c}}{\partial \boldsymbol{\beta} \partial \beta_j} + \frac{\partial^3 J}{\partial \mathbf{c}^2 \partial \beta_k} \frac{\partial^2 \mathbf{c}}{\partial \boldsymbol{\beta} \partial \beta_j} \right\} \qquad \text{(B.19)}
\end{aligned}$$

where

$$\frac{\partial^4 J}{\partial \mathbf{c}^2 \partial \boldsymbol{\beta}^2} = -\frac{\partial^4 l}{\partial \mathbf{c}^2 \partial \boldsymbol{\beta}^2}$$

$$\frac{\partial^4 J}{\partial \mathbf{c}\partial\boldsymbol{\beta}^3} = -\frac{\partial^4 l}{\partial \mathbf{c}\partial\boldsymbol{\beta}^3}$$

$$\frac{\partial^4 J}{\partial \mathbf{c}^3\partial\boldsymbol{\beta}} = -\frac{\partial^4 l}{\partial \mathbf{c}^3\partial\boldsymbol{\beta}}$$

$$\frac{\partial^4 J}{\partial \mathbf{c}^4} = -\frac{\partial^4 l}{\partial \mathbf{c}^4}$$

- $\frac{\partial^3 \mathbf{c}}{\partial\boldsymbol{\beta}^2\partial\boldsymbol{\theta}}$

Taking the θ_k-derivative on $D_{\beta_j} D_{\boldsymbol{\beta}} D_{\mathbf{c}} J$, we obtain:

$$\begin{aligned}
& D_{\theta_k} D_{\beta_j} D_{\boldsymbol{\beta}} D_{\mathbf{c}} J \\
= \; & \frac{\partial^4 J}{\partial \mathbf{c}\partial\boldsymbol{\beta}\partial\beta_j\partial c_g}\frac{\partial c_g}{\partial\theta_k} + \frac{\partial^4 J}{\partial \mathbf{c}\partial\boldsymbol{\beta}\partial\beta_j\partial\theta_k} + \frac{\partial^4 J}{\partial \mathbf{c}^2\partial c_i\partial c_g}\frac{\partial c_g}{\partial\theta_k}\frac{\partial c_i}{\partial\beta_j}\frac{\partial \mathbf{c}}{\partial\boldsymbol{\beta}} \\
+ \; & \frac{\partial^4 J}{\partial \mathbf{c}\partial\boldsymbol{\beta}\partial c_i\partial c_g}\frac{\partial c_g}{\partial\theta_k}\frac{\partial c_i}{\partial\beta_j} + \frac{\partial^4 J}{\partial \mathbf{c}\partial\boldsymbol{\beta}\partial c_i\partial\theta_k}\frac{\partial c_i}{\partial\beta_j} + \frac{\partial^3 J}{\partial \mathbf{c}\partial\boldsymbol{\beta}\partial c_i}\frac{\partial^2 c_i}{\partial\beta_j\partial\theta_k} \\
+ \; & \frac{\partial^4 J}{\partial \mathbf{c}^2\partial c_i\partial\theta_k}\frac{\partial c_i}{\partial\beta_j}\frac{\partial \mathbf{c}}{\partial\boldsymbol{\beta}} + \frac{\partial^3 J}{\partial \mathbf{c}^2\partial c_i}\frac{\partial^2 c_i}{\partial\beta_j\partial\theta_k}\frac{\partial \mathbf{c}}{\partial\boldsymbol{\beta}} + \frac{\partial^3 J}{\partial \mathbf{c}^2\partial c_i}\frac{\partial c_i}{\partial\beta_j}\frac{\partial^2 \mathbf{c}}{\partial\boldsymbol{\beta}\partial\theta_k} \\
+ \; & \frac{\partial^4 J}{\partial \mathbf{c}^2\partial\beta_j\partial c_g}\frac{\partial c_g}{\partial\theta_k}\frac{\partial \mathbf{c}}{\partial\boldsymbol{\beta}} + \frac{\partial^4 J}{\partial \mathbf{c}^2\partial\beta_j\partial\theta_k}\frac{\partial \mathbf{c}}{\partial\boldsymbol{\beta}} + \frac{\partial^3 J}{\partial \mathbf{c}^2\partial\beta_j}\frac{\partial^2 \mathbf{c}}{\partial\boldsymbol{\beta}\partial\theta_k} \\
+ \; & \left(\frac{\partial^3 J}{\partial \mathbf{c}^2\partial c_g}\frac{\partial c_g}{\partial\theta_k} + \frac{\partial^3 J}{\partial \mathbf{c}^2\partial\theta_k}\right)\frac{\partial^2 \mathbf{c}}{\partial\boldsymbol{\beta}\partial\beta_j} + \frac{\partial^2 J}{\partial \mathbf{c}^2}\frac{\partial^3 \mathbf{c}}{\partial\boldsymbol{\beta}\partial\beta_j\partial\theta_k} = 0
\end{aligned} \tag{B.20}$$

Solving for $\frac{\partial^3 \mathbf{c}}{\partial \boldsymbol{\beta} \partial \beta_j \partial \theta_k}$:

$$
\begin{aligned}
&\frac{\partial^3 \mathbf{c}}{\partial \boldsymbol{\beta} \partial \beta_j \partial \theta_k} \\
= \; & - \left\{ \frac{\partial^2 J}{\partial \mathbf{c}^2} \right\}^{-1} \left\{ \frac{\partial^4 J}{\partial \mathbf{c} \partial \boldsymbol{\beta} \partial \beta_j \partial c_g} \frac{\partial c_g}{\partial \theta_k} + \frac{\partial^4 J}{\partial \mathbf{c} \partial \boldsymbol{\beta} \partial \beta_j \partial \theta_k} + \frac{\partial^4 J}{\partial \mathbf{c}^2 \partial c_i \partial c_g} \frac{\partial c_g}{\partial \theta_k} \frac{\partial c_i}{\partial \beta_j} \frac{\partial \mathbf{c}}{\partial \boldsymbol{\beta}} \right. \\
& + \frac{\partial^4 J}{\partial \mathbf{c} \partial \boldsymbol{\beta} \partial c_i \partial c_g} \frac{\partial c_g}{\partial \theta_k} \frac{\partial c_i}{\partial \beta_j} + \frac{\partial^4 J}{\partial \mathbf{c} \partial \boldsymbol{\beta} \partial c_i \partial \theta_k} \frac{\partial c_i}{\partial \beta_j} + \frac{\partial^3 J}{\partial \mathbf{c} \partial \boldsymbol{\beta} \partial c_i} \frac{\partial^2 c_i}{\partial \beta_j \partial \theta_k} \\
& + \frac{\partial^4 J}{\partial \mathbf{c}^2 \partial c_i \partial \theta_k} \frac{\partial c_i}{\partial \beta_j} \frac{\partial \mathbf{c}}{\partial \boldsymbol{\beta}} + \frac{\partial^3 J}{\partial \mathbf{c}^2 \partial c_i} \frac{\partial^2 c_i}{\partial \beta_j \partial \theta_k} \frac{\partial \mathbf{c}}{\partial \boldsymbol{\beta}} + \frac{\partial^3 J}{\partial \mathbf{c}^2 \partial c_i} \frac{\partial c_i}{\partial \beta_j} \frac{\partial^2 \mathbf{c}}{\partial \boldsymbol{\beta} \partial \theta_k} \\
& + \frac{\partial^4 J}{\partial \mathbf{c}^2 \partial \beta_j \partial c_g} \frac{\partial c_g}{\partial \theta_k} \frac{\partial \mathbf{c}}{\partial \boldsymbol{\beta}} + \frac{\partial^4 J}{\partial \mathbf{c}^2 \partial \beta_j \partial \theta_k} \frac{\partial \mathbf{c}}{\partial \boldsymbol{\beta}} + \frac{\partial^3 J}{\partial \mathbf{c}^2 \partial \beta_j} \frac{\partial^2 \mathbf{c}}{\partial \boldsymbol{\beta} \partial \theta_k} \\
& \left. + \left(\frac{\partial^3 J}{\partial \mathbf{c}^2 \partial c_g} \frac{\partial c_g}{\partial \theta_k} + \frac{\partial^3 J}{\partial \mathbf{c}^2 \partial \theta_k} \right) \frac{\partial^2 \mathbf{c}}{\partial \boldsymbol{\beta} \partial \beta_j} \right\}
\end{aligned}
\tag{B.21}
$$

- $\frac{\partial^3 \mathbf{c}}{\partial \boldsymbol{\beta}^2 \partial \mathbf{y}}$

Taking the y_k-derivative on $D_{\beta_j} D_{\boldsymbol{\beta}} D_{\mathbf{c}} J$, we obtain:

$$
\begin{aligned}
& D_{y_k} D_{\beta_j} D_{\boldsymbol{\beta}} D_{\mathbf{c}} J \\
= \; & \frac{\partial^4 J}{\partial \mathbf{c} \partial \boldsymbol{\beta} \partial \beta_j \partial c_g} \frac{\partial c_g}{\partial y_k} + \frac{\partial^4 J}{\partial \mathbf{c} \partial \boldsymbol{\beta} \partial \beta_j \partial y_k} + \frac{\partial^4 J}{\partial \mathbf{c}^2 \partial c_i \partial c_g} \frac{\partial c_g}{\partial y_k} \frac{\partial c_i}{\partial \beta_j} \frac{\partial \mathbf{c}}{\partial \boldsymbol{\beta}} \\
& + \frac{\partial^4 J}{\partial \mathbf{c} \partial \boldsymbol{\beta} \partial c_i \partial c_g} \frac{\partial c_g}{\partial y_k} \frac{\partial c_i}{\partial \beta_j} + \frac{\partial^4 J}{\partial \mathbf{c} \partial \boldsymbol{\beta} \partial c_i \partial y_k} \frac{\partial c_i}{\partial \beta_j} + \frac{\partial^3 J}{\partial \mathbf{c} \partial \boldsymbol{\beta} \partial c_i} \frac{\partial^2 c_i}{\partial \beta_j \partial y_k} \\
& + \frac{\partial^4 J}{\partial \mathbf{c}^2 \partial c_i \partial y_k} \frac{\partial c_i}{\partial \beta_j} \frac{\partial \mathbf{c}}{\partial \boldsymbol{\beta}} + \frac{\partial^3 J}{\partial \mathbf{c}^2 \partial c_i} \frac{\partial^2 c_i}{\partial \beta_j \partial y_k} \frac{\partial \mathbf{c}}{\partial \boldsymbol{\beta}} + \frac{\partial^3 J}{\partial \mathbf{c}^2 \partial c_i} \frac{\partial c_i}{\partial \beta_j} \frac{\partial^2 \mathbf{c}}{\partial \boldsymbol{\beta} \partial y_k} \\
& + \frac{\partial^4 J}{\partial \mathbf{c}^2 \partial \beta_j \partial c_g} \frac{\partial c_g}{\partial y_k} \frac{\partial \mathbf{c}}{\partial \boldsymbol{\beta}} + \frac{\partial^4 J}{\partial \mathbf{c}^2 \partial \beta_j \partial y_k} \frac{\partial \mathbf{c}}{\partial \boldsymbol{\beta}} + \frac{\partial^3 J}{\partial \mathbf{c}^2 \partial \beta_j} \frac{\partial^2 \mathbf{c}}{\partial \boldsymbol{\beta} \partial y_k} \\
& + \left(\frac{\partial^3 J}{\partial \mathbf{c}^2 \partial c_g} \frac{\partial c_g}{\partial y_k} + \frac{\partial^3 J}{\partial \mathbf{c}^2 \partial y_k} \right) \frac{\partial^2 \mathbf{c}}{\partial \boldsymbol{\beta} \partial \beta_j} + \frac{\partial^2 J}{\partial \mathbf{c}^2} \frac{\partial^3 \mathbf{c}}{\partial \boldsymbol{\beta} \partial \beta_j \partial y_k} = 0
\end{aligned}
\tag{B.22}
$$

Solving for $\frac{\partial^3 \mathbf{c}}{\partial \boldsymbol{\beta} \partial \beta_j \partial y_k}$:

$$
\begin{aligned}
& \frac{\partial^3 \mathbf{c}}{\partial \boldsymbol{\beta} \partial \beta_j \partial y_k} \\
= & - \left\{ \frac{\partial^2 J}{\partial \mathbf{c}^2} \right\}^{-1} \left\{ \frac{\partial^4 J}{\partial \mathbf{c} \partial \boldsymbol{\beta} \partial \beta_j \partial c_g} \frac{\partial c_g}{\partial y_k} + \frac{\partial^4 J}{\partial \mathbf{c} \partial \boldsymbol{\beta} \partial \beta_j \partial y_k} + \frac{\partial^4 J}{\partial \mathbf{c}^2 \partial c_i \partial c_g} \frac{\partial c_g}{\partial y_k} \frac{\partial c_i}{\partial \beta_j} \frac{\partial \mathbf{c}}{\partial \boldsymbol{\beta}} \right. \\
& + \frac{\partial^4 J}{\partial \mathbf{c} \partial \boldsymbol{\beta} \partial c_i \partial c_g} \frac{\partial c_g}{\partial y_k} \frac{\partial c_i}{\partial \beta_j} + \frac{\partial^4 J}{\partial \mathbf{c} \partial \boldsymbol{\beta} \partial c_i \partial y_k} \frac{\partial c_i}{\partial \beta_j} + \frac{\partial^3 J}{\partial \mathbf{c} \partial \boldsymbol{\beta} \partial c_i} \frac{\partial^2 c_i}{\partial \beta_j \partial y_k} \\
& + \frac{\partial^4 J}{\partial \mathbf{c}^2 \partial c_i \partial y_k} \frac{\partial c_i}{\partial \beta_j} \frac{\partial \mathbf{c}}{\partial \boldsymbol{\beta}} + \frac{\partial^3 J}{\partial \mathbf{c}^2 \partial c_i} \frac{\partial^2 c_i}{\partial \beta_j \partial y_k} \frac{\partial \mathbf{c}}{\partial \boldsymbol{\beta}} + \frac{\partial^3 J}{\partial \mathbf{c}^2 \partial c_i} \frac{\partial c_i}{\partial \beta_j} \frac{\partial^2 \mathbf{c}}{\partial \boldsymbol{\beta} \partial y_k} \\
& + \frac{\partial^4 J}{\partial \mathbf{c}^2 \partial \beta_j \partial c_g} \frac{\partial c_g}{\partial y_k} \frac{\partial \mathbf{c}}{\partial \boldsymbol{\beta}} + \frac{\partial^4 J}{\partial \mathbf{c}^2 \partial \beta_j \partial y_k} \frac{\partial \mathbf{c}}{\partial \boldsymbol{\beta}} + \frac{\partial^3 J}{\partial \mathbf{c}^2 \partial \beta_j} \frac{\partial^2 \mathbf{c}}{\partial \boldsymbol{\beta} \partial y_k} \\
& \left. + \left(\frac{\partial^3 J}{\partial \mathbf{c}^2 \partial c_g} \frac{\partial c_g}{\partial y_k} + \frac{\partial^3 J}{\partial \mathbf{c}^2 \partial y_k} \right) \frac{\partial^2 \mathbf{c}}{\partial \boldsymbol{\beta} \partial \beta_j} \right\}. \qquad \text{(B.23)}
\end{aligned}
$$

- $\frac{\partial^3 \mathbf{c}}{\partial \boldsymbol{\beta} \partial \boldsymbol{\theta}^2}$

Replacing β_j and with θ_k, and replacing θ_k with θ_j in (B.21), we obtain

$\frac{\partial^3 \mathbf{c}}{\partial \boldsymbol{\beta} \partial \theta_k \partial \theta_j}$:

$$\begin{aligned}
&\frac{\partial^3 \mathbf{c}}{\partial \boldsymbol{\beta} \partial \theta_k \partial \theta_j} = \\
&- \left\{ \frac{\partial^2 J}{\partial \mathbf{c}^2} \right\}^{-1} \left\{ \frac{\partial^4 J}{\partial \mathbf{c} \partial \boldsymbol{\beta} \partial \theta_k \partial c_g} \frac{\partial c_g}{\partial \theta_j} + \frac{\partial^4 J}{\partial \mathbf{c} \partial \boldsymbol{\beta} \partial \theta_k \partial \theta_j} + \frac{\partial^3 J}{\partial \mathbf{c}^2 \partial c_i} \frac{\partial c_i}{\partial \theta_k} \frac{\partial^2 \mathbf{c}}{\partial \boldsymbol{\beta} \partial \theta_j} \right. \\
&+ \frac{\partial^4 J}{\partial \mathbf{c} \partial \boldsymbol{\beta} \partial c_i \partial c_g} \frac{\partial c_g}{\partial \theta_j} \frac{\partial c_i}{\partial \theta_k} + \frac{\partial^4 J}{\partial \mathbf{c} \partial \boldsymbol{\beta} \partial c_i \partial \theta_j} \frac{\partial c_i}{\partial \theta_k} + \frac{\partial^3 J}{\partial \mathbf{c} \partial \boldsymbol{\beta} \partial c_i} \frac{\partial^2 c_i}{\partial \theta_k \partial \theta_j} \\
&+ \frac{\partial^4 J}{\partial \mathbf{c}^2 \partial \theta_k \partial c_g} \frac{\partial c_g}{\partial \theta_j} \frac{\partial \mathbf{c}}{\partial \boldsymbol{\beta}} + \frac{\partial^4 J}{\partial \mathbf{c}^2 \partial \theta_k \partial \theta_j} \frac{\partial \mathbf{c}}{\partial \boldsymbol{\beta}} + \frac{\partial^3 J}{\partial \mathbf{c}^2 \partial \theta_k} \frac{\partial^2 \mathbf{c}}{\partial \boldsymbol{\beta} \partial \theta_j} \\
&+ \frac{\partial^4 J}{\partial \mathbf{c}^2 \partial c_i \partial c_g} \frac{\partial c_g}{\partial \theta_j} \frac{\partial c_i}{\partial \theta_k} \frac{\partial \mathbf{c}}{\partial \boldsymbol{\beta}} + \frac{\partial^4 J}{\partial \mathbf{c}^2 \partial c_i \partial \theta_j} \frac{\partial c_i}{\partial \theta_k} \frac{\partial \mathbf{c}}{\partial \boldsymbol{\beta}} + \frac{\partial^3 J}{\partial \mathbf{c}^2 \partial c_i} \frac{\partial^2 c_i}{\partial \theta_k \partial \theta_j} \frac{\partial \mathbf{c}}{\partial \boldsymbol{\beta}} \\
&+ \left. \frac{\partial^3 J}{\partial \mathbf{c}^2 \partial c_g} \frac{\partial c_g}{\partial \theta_j} \frac{\partial^2 \mathbf{c}}{\partial \boldsymbol{\beta} \partial \theta_k} + \frac{\partial^3 J}{\partial \mathbf{c}^2 \partial \theta_j} \frac{\partial^2 \mathbf{c}}{\partial \boldsymbol{\beta} \partial \theta_k} \right\} \qquad \text{(B.24)}
\end{aligned}$$

- $\frac{\partial^3 \mathbf{c}}{\partial \boldsymbol{\beta} \partial \boldsymbol{\theta} \partial \mathbf{y}}$

Taking the y_j-derivative on $D_{\theta_k} D_{\boldsymbol{\beta}} D_{\mathbf{c}} J$ given in (B.13), we obtain:

$$\begin{aligned}
&D_{y_j} D_{\theta_k} D_{\boldsymbol{\beta}} D_{\mathbf{c}} J \\
&= \frac{\partial^4 J}{\partial \mathbf{c} \partial \boldsymbol{\beta} \partial \theta_k \partial c_g} \frac{\partial c_g}{\partial y_j} + \frac{\partial^4 J}{\partial \mathbf{c} \partial \boldsymbol{\beta} \partial \theta_k \partial y_j} + \frac{\partial^3 J}{\partial \mathbf{c}^2 \partial c_i} \frac{\partial c_i}{\partial \theta_k} \frac{\partial^2 \mathbf{c}}{\partial \boldsymbol{\beta} \partial y_j} \\
&+ \frac{\partial^4 J}{\partial \mathbf{c} \partial \boldsymbol{\beta} \partial c_i \partial c_g} \frac{\partial c_g}{\partial y_j} \frac{\partial c_i}{\partial \theta_k} + \frac{\partial^4 J}{\partial \mathbf{c} \partial \boldsymbol{\beta} \partial c_i \partial y_j} \frac{\partial c_i}{\partial \theta_k} + \frac{\partial^3 J}{\partial \mathbf{c} \partial \boldsymbol{\beta} \partial c_i} \frac{\partial^2 c_i}{\partial \theta_k \partial y_j} \\
&+ \frac{\partial^4 J}{\partial \mathbf{c}^2 \partial \theta_k \partial c_g} \frac{\partial c_g}{\partial y_j} \frac{\partial \mathbf{c}}{\partial \boldsymbol{\beta}} + \frac{\partial^4 J}{\partial \mathbf{c}^2 \partial \theta_k \partial y_j} \frac{\partial \mathbf{c}}{\partial \boldsymbol{\beta}} + \frac{\partial^3 J}{\partial \mathbf{c}^2 \partial \theta_k} \frac{\partial^2 \mathbf{c}}{\partial \boldsymbol{\beta} \partial y_j} \\
&+ \frac{\partial^4 J}{\partial \mathbf{c}^2 \partial c_i \partial c_g} \frac{\partial c_g}{\partial y_j} \frac{\partial c_i}{\partial \theta_k} \frac{\partial \mathbf{c}}{\partial \boldsymbol{\beta}} + \frac{\partial^4 J}{\partial \mathbf{c}^2 \partial c_i \partial y_j} \frac{\partial c_i}{\partial \theta_k} \frac{\partial \mathbf{c}}{\partial \boldsymbol{\beta}} + \frac{\partial^3 J}{\partial \mathbf{c}^2 \partial c_i} \frac{\partial^2 c_i}{\partial \theta_k \partial y_j} \frac{\partial \mathbf{c}}{\partial \boldsymbol{\beta}} \\
&+ \frac{\partial^3 J}{\partial \mathbf{c}^2 \partial c_g} \frac{\partial c_g}{\partial y_j} \frac{\partial^2 \mathbf{c}}{\partial \boldsymbol{\beta} \partial \theta_k} + \frac{\partial^3 J}{\partial \mathbf{c}^2 \partial y_j} \frac{\partial^2 \mathbf{c}}{\partial \boldsymbol{\beta} \partial \theta_k} + \frac{\partial^2 J}{\partial \mathbf{c}^2} \frac{\partial^3 \mathbf{c}}{\partial \boldsymbol{\beta} \partial \theta_k \partial y_j} = 0 \qquad \text{(B.25)}
\end{aligned}$$

Solving for $\frac{\partial^3 \mathbf{c}}{\partial \boldsymbol{\beta} \partial \theta_k \partial y_j}$:

$$\begin{aligned}
&\frac{\partial^3 \mathbf{c}}{\partial \boldsymbol{\beta} \partial \theta_k \partial y_j} = \\
&- \left\{ \frac{\partial^2 J}{\partial \mathbf{c}^2} \right\}^{-1} \left\{ \frac{\partial^4 J}{\partial \mathbf{c} \partial \boldsymbol{\beta} \partial \theta_k \partial c_g} \frac{\partial c_g}{\partial y_j} + \frac{\partial^4 J}{\partial \mathbf{c} \partial \boldsymbol{\beta} \partial \theta_k \partial y_j} + \frac{\partial^3 J}{\partial \mathbf{c}^2 \partial c_i} \frac{\partial c_i}{\partial \theta_k} \frac{\partial^2 \mathbf{c}}{\partial \boldsymbol{\beta} \partial y_j} \right. \\
&+ \frac{\partial^4 J}{\partial \mathbf{c} \partial \boldsymbol{\beta} \partial c_i \partial c_g} \frac{\partial c_g}{\partial y_j} \frac{\partial c_i}{\partial \theta_k} + \frac{\partial^4 J}{\partial \mathbf{c} \partial \boldsymbol{\beta} \partial c_i \partial y_j} \frac{\partial c_i}{\partial \theta_k} + \frac{\partial^3 J}{\partial \mathbf{c} \partial \boldsymbol{\beta} \partial c_i} \frac{\partial^2 c_i}{\partial \theta_k \partial y_j} \\
&+ \frac{\partial^4 J}{\partial \mathbf{c}^2 \partial \theta_k \partial c_g} \frac{\partial c_g}{\partial y_j} \frac{\partial \mathbf{c}}{\partial \boldsymbol{\beta}} + \frac{\partial^4 J}{\partial \mathbf{c}^2 \partial \theta_k \partial y_j} \frac{\partial \mathbf{c}}{\partial \boldsymbol{\beta}} + \frac{\partial^3 J}{\partial \mathbf{c}^2 \partial \theta_k} \frac{\partial^2 \mathbf{c}}{\partial \boldsymbol{\beta} \partial y_j} \\
&+ \frac{\partial^4 J}{\partial \mathbf{c}^2 \partial c_i \partial c_g} \frac{\partial c_g}{\partial y_j} \frac{\partial c_i}{\partial \theta_k} \frac{\partial \mathbf{c}}{\partial \boldsymbol{\beta}} + \frac{\partial^4 J}{\partial \mathbf{c}^2 \partial c_i \partial y_j} \frac{\partial c_i}{\partial \theta_k} \frac{\partial \mathbf{c}}{\partial \boldsymbol{\beta}} + \frac{\partial^3 J}{\partial \mathbf{c}^2 \partial c_i} \frac{\partial^2 c_i}{\partial \theta_k \partial y_j} \frac{\partial \mathbf{c}}{\partial \boldsymbol{\beta}} \\
&+ \left. \frac{\partial^3 J}{\partial \mathbf{c}^2 \partial c_g} \frac{\partial c_g}{\partial y_j} \frac{\partial^2 \mathbf{c}}{\partial \boldsymbol{\beta} \partial \theta_k} + \frac{\partial^3 J}{\partial \mathbf{c}^2 \partial y_j} \frac{\partial^2 \mathbf{c}}{\partial \boldsymbol{\beta} \partial \theta_k} \right\} \qquad \text{(B.26)}
\end{aligned}$$

B.2 Second Optimization Level to Estimate Global Parameters

The second level optimization criterion is be written as:

$$H(\boldsymbol{\beta}|\boldsymbol{\lambda}, \mathbf{y}) = -l(\mathbf{y}|\mathbf{c}, \boldsymbol{\beta}) \qquad \text{(B.27)}$$

- $\frac{\partial H}{\partial \boldsymbol{\beta}}$

The first derivative of $H(\boldsymbol{\beta}|\boldsymbol{\lambda}, \mathbf{y})$ with respect to $\boldsymbol{\beta}$ is:

$$\frac{\partial H}{\partial \boldsymbol{\beta}} = -\frac{\partial l}{\partial \boldsymbol{\beta}} - \left(\frac{\partial \mathbf{c}}{\partial \boldsymbol{\beta}} \right)' \frac{\partial l}{\partial \mathbf{c}} \qquad \text{(B.28)}$$

where $\frac{\partial \mathbf{c}}{\partial \boldsymbol{\beta}}$ is given in (B.5).

- $\frac{\partial^2 H}{\partial \boldsymbol{\beta}^2}$

The second derivative of $H(\boldsymbol{\beta}|\boldsymbol{\lambda}, \mathbf{y})$ with respect to $\boldsymbol{\beta}$ is:

$$\begin{aligned} \frac{\partial^2 H}{\partial \boldsymbol{\beta}^2} &= -\frac{\partial^2 l}{\partial \boldsymbol{\beta}^2} - \frac{\partial^2 l}{\partial \boldsymbol{\beta} \partial \mathbf{c}} \frac{\partial \mathbf{c}}{\partial \boldsymbol{\beta}} - \frac{\partial^2 l}{\partial \mathbf{c}^2} \left(\frac{\partial \mathbf{c}}{\partial \boldsymbol{\beta}} \right)^2 - \frac{\partial^2 l}{\partial \mathbf{c} \partial \boldsymbol{\beta}} \frac{\partial \mathbf{c}}{\partial \boldsymbol{\beta}} - \frac{\partial l}{\partial c_j} \frac{\partial^2 c_j}{\partial \boldsymbol{\beta}^2} \\ &= -\frac{\partial^2 l}{\partial \boldsymbol{\beta}^2} - \frac{\partial^2 l}{\partial \mathbf{c}^2} \left(\frac{\partial \mathbf{c}}{\partial \boldsymbol{\beta}} \right)^2 - 2 \frac{\partial^2 l}{\partial \mathbf{c} \partial \boldsymbol{\beta}} \frac{\partial \mathbf{c}}{\partial \boldsymbol{\beta}} - \frac{\partial l}{\partial c_j} \frac{\partial^2 c_j}{\partial \boldsymbol{\beta}^2} \end{aligned} \tag{B.29}$$

where $\frac{\partial^2 \mathbf{c}}{\partial \boldsymbol{\beta}^2}$ is given in (B.7).

- $\frac{\partial \boldsymbol{\beta}}{\partial \boldsymbol{\theta}}$

According (2.7), the derivative of $\boldsymbol{\beta}$ with respect to $\boldsymbol{\theta}$ is:

$$\frac{\partial \boldsymbol{\beta}}{\partial \boldsymbol{\theta}} = -\left(\frac{\partial^2 H}{\partial \boldsymbol{\beta}^2} \right)^{-1} \frac{\partial^2 H}{\partial \boldsymbol{\beta} \partial \boldsymbol{\theta}} \tag{B.30}$$

where the second derivative of $H(\boldsymbol{\beta}|\boldsymbol{\lambda}, \mathbf{y})$ with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ is:

$$\frac{\partial^2 H}{\partial \boldsymbol{\beta} \partial \boldsymbol{\theta}} = -\frac{\partial^2 l}{\partial \boldsymbol{\beta} \partial \mathbf{c}} \frac{\partial \mathbf{c}}{\partial \boldsymbol{\theta}} - \frac{\partial^2 l}{\partial^2 \mathbf{c}} \frac{\partial \mathbf{c}}{\partial \boldsymbol{\theta}} \frac{\partial \mathbf{c}}{\partial \boldsymbol{\beta}} - \frac{\partial l}{\partial c_k} \frac{\partial^2 c_k}{\partial \boldsymbol{\beta} \partial \boldsymbol{\theta}} \tag{B.31}$$

where $\frac{\partial \mathbf{c}}{\partial \boldsymbol{\theta}}$ is given in (B.6), $\frac{\partial^2 \mathbf{c}}{\partial \boldsymbol{\beta} \partial \boldsymbol{\theta}}$ is given in (B.14) and $\frac{\partial \mathbf{c}}{\partial \boldsymbol{\beta}}$ is given in (B.5).

- $\frac{\partial \boldsymbol{\beta}}{\partial \mathbf{y}}$

The partial derivative of $\boldsymbol{\beta}$ with respect to $\mathbf{y}$ is:

$$\frac{\partial \boldsymbol{\beta}}{\partial \mathbf{y}} = -\left(\frac{\partial^2 H}{\partial \boldsymbol{\beta}^2}\right)^{-1} \frac{\partial^2 H}{\partial \boldsymbol{\beta} \partial \mathbf{y}} \tag{B.32}$$

where the second derivative of $H(\boldsymbol{\beta}|\boldsymbol{\lambda}, \mathbf{y})$ with respect to $\boldsymbol{\beta}$ and $\mathbf{y}$ is:

$$\frac{\partial^2 H}{\partial \boldsymbol{\beta} \partial \mathbf{y}} = -\frac{\partial^2 l}{\partial \boldsymbol{\beta} \partial \mathbf{y}} - \frac{\partial^2 l}{\partial \boldsymbol{\beta} \partial \mathbf{c}} \frac{\partial \mathbf{c}}{\partial \mathbf{y}} - \frac{\partial^2 l}{\partial \mathbf{c} \partial \mathbf{y}} \frac{\partial \mathbf{c}}{\partial \boldsymbol{\beta}} - \frac{\partial^2 l}{\partial \mathbf{c}^2} \frac{\partial \mathbf{c}}{\partial \mathbf{y}} \frac{\partial \mathbf{c}}{\partial \boldsymbol{\beta}} - \frac{\partial l}{\partial \mathbf{c}} \frac{\partial^2 \mathbf{c}}{\partial \boldsymbol{\beta} \partial \mathbf{y}} \tag{B.33}$$

- $\frac{\partial^2 \boldsymbol{\beta}}{\partial \boldsymbol{\theta}^2}$

Since the optimal linear coefficient vector $\boldsymbol{\beta}$ satisfying $D_{\boldsymbol{\beta}} H = 0$, we can take the θ_k-derivative on $D_{\boldsymbol{\beta}} H$, as follows:

$$D_{\theta_k} D_{\boldsymbol{\beta}} H = \frac{\partial^2 H}{\partial \boldsymbol{\beta} \partial \theta_k} + \frac{\partial^2 H}{\partial \boldsymbol{\beta}^2} \frac{\partial \boldsymbol{\beta}}{\partial \theta_k} = 0 \tag{B.34}$$

We then take the θ_j-derivative on $D_{\theta_k} D_{\boldsymbol{\beta}} H$:

$$\begin{aligned} D_{\theta_j} D_{\theta_k} D_{\boldsymbol{\beta}} H = & \frac{\partial^3 H}{\partial \boldsymbol{\beta} \partial \theta_k \partial \theta_j} + \frac{\partial^3 H}{\partial \boldsymbol{\beta}^2 \partial \theta_k} \frac{\partial \boldsymbol{\beta}}{\partial \theta_j} + \frac{\partial^3 H}{\partial \boldsymbol{\beta}^2 \partial \beta_i} \frac{\partial \beta_i}{\partial \theta_j} \frac{\partial \boldsymbol{\beta}}{\partial \theta_k} \\ & + \frac{\partial^3 H}{\partial \boldsymbol{\beta}^2 \partial \theta_j} \frac{\partial \boldsymbol{\beta}}{\partial \theta_k} + \frac{\partial^2 H}{\partial \boldsymbol{\beta}^2} \frac{\partial^2 \boldsymbol{\beta}}{\partial \theta_k \partial \theta_j} = 0 \end{aligned} \tag{B.35}$$

Solving for $\frac{\partial^2 \boldsymbol{\beta}}{\partial \theta_k \partial \theta_j}$, we get:

$$\frac{\partial^2 \boldsymbol{\beta}}{\partial \theta_k \partial \theta_j} = -\left[\frac{\partial^2 H}{\partial \boldsymbol{\beta}^2}\right]^{-1} \left[\frac{\partial^3 H}{\partial \boldsymbol{\beta} \partial \theta_k \partial \theta_j} + \frac{\partial^3 H}{\partial \boldsymbol{\beta}^2 \partial \theta_k} \frac{\partial \boldsymbol{\beta}}{\partial \theta_j} + \frac{\partial^3 H}{\partial \boldsymbol{\beta}^2 \partial \beta_i} \frac{\partial \beta_i}{\partial \theta_j} \frac{\partial \boldsymbol{\beta}}{\partial \theta_k} + \frac{\partial^3 H}{\partial \boldsymbol{\beta}^2 \partial \theta_j} \frac{\partial \boldsymbol{\beta}}{\partial \theta_k}\right] \tag{B.36}$$

where

$$\frac{\partial^3 H}{\partial\boldsymbol{\beta}\partial\beta_j\partial\boldsymbol{\theta}} =$$
$$-\frac{\partial^3 l}{\partial\boldsymbol{\beta}\partial\beta_j\partial\boldsymbol{\theta}} - \frac{\partial^3 l}{\partial\boldsymbol{\beta}\partial\beta_j\partial\mathbf{c}}\frac{\partial\mathbf{c}}{\partial\boldsymbol{\theta}} - \frac{\partial^3 l}{\partial\boldsymbol{\beta}\partial c_k\partial\mathbf{c}}\frac{\partial\mathbf{c}}{\partial\boldsymbol{\theta}}\frac{\partial c_k}{\partial\beta_j} - \frac{\partial^3 l}{\partial\mathbf{c}\partial\beta_j\partial\boldsymbol{\theta}}\frac{\partial\mathbf{c}}{\partial\boldsymbol{\beta}} - \frac{\partial^2 l}{\partial c_k\partial\beta_j}\frac{\partial^2 c_k}{\partial\boldsymbol{\beta}\partial\boldsymbol{\theta}}$$
$$-\frac{\partial^2 l}{\partial\boldsymbol{\beta}\partial c_k}\frac{\partial^2 c_k}{\partial\beta_j\partial\boldsymbol{\theta}} - \frac{\partial^3 l}{\partial\mathbf{c}^2\partial c_k}\frac{\partial c_k}{\partial\boldsymbol{\theta}}\frac{\partial\mathbf{c}}{\partial\beta_j}\frac{\partial\mathbf{c}}{\partial\boldsymbol{\beta}} - \frac{\partial^2 l}{\partial\mathbf{c}^2}\frac{\partial^2\mathbf{c}}{\partial\beta_j\partial\boldsymbol{\theta}}\frac{\partial\mathbf{c}}{\partial\boldsymbol{\beta}} - \frac{\partial^2 l}{\partial\mathbf{c}\partial c_k}\frac{\partial\mathbf{c}}{\partial\beta_j}\frac{\partial^2 c_k}{\partial\boldsymbol{\beta}\partial\boldsymbol{\theta}}$$
$$-\frac{\partial^2 l}{\partial\mathbf{c}^2}\frac{\partial\mathbf{c}}{\partial\boldsymbol{\theta}}\frac{\partial^2\mathbf{c}}{\partial\boldsymbol{\beta}\partial\beta_j} - \frac{\partial l}{\partial c_k}\frac{\partial^3 c_k}{\partial\boldsymbol{\beta}\partial\beta_j\partial\boldsymbol{\theta}} - \frac{\partial^3 l}{\partial\boldsymbol{\beta}\partial\mathbf{c}\partial c_i}\frac{\partial c_i}{\partial\beta_j}\frac{\partial\mathbf{c}}{\partial\boldsymbol{\theta}}$$

$$\frac{\partial^3 H}{\partial\boldsymbol{\beta}\partial\boldsymbol{\theta}\partial\theta_j} =$$
$$-\frac{\partial^3 l}{\partial\boldsymbol{\beta}\partial\mathbf{c}\partial c_i}\frac{\partial c_i}{\partial\theta_j}\frac{\partial\mathbf{c}}{\partial\boldsymbol{\theta}} - \frac{\partial^2 l}{\partial\boldsymbol{\beta}\partial\mathbf{c}}\frac{\partial^2\mathbf{c}}{\partial\boldsymbol{\theta}\partial\theta_j} - \frac{\partial^3 l}{\partial\mathbf{c}^2\partial c_i}\frac{\partial c_i}{\partial\theta_j}\frac{\partial\mathbf{c}}{\partial\boldsymbol{\theta}}\frac{\partial\mathbf{c}}{\partial\boldsymbol{\beta}} - \frac{\partial^2 l}{\partial\mathbf{c}^2}\frac{\partial^2\mathbf{c}}{\partial\boldsymbol{\theta}\partial\theta_j}\frac{\partial\mathbf{c}}{\partial\boldsymbol{\beta}}$$
$$-\frac{\partial^2 l}{\partial\mathbf{c}^2}\frac{\partial\mathbf{c}}{\partial\boldsymbol{\theta}}\frac{\partial^2\mathbf{c}}{\partial\boldsymbol{\beta}\partial\theta_j} - \frac{\partial^2 l}{\partial c_k\partial\mathbf{c}}\frac{\partial\mathbf{c}}{\partial\theta_j}\frac{\partial^2 c_k}{\partial\boldsymbol{\beta}\partial\boldsymbol{\theta}} - \frac{\partial l}{\partial c_k}\frac{\partial^3 c_k}{\partial\boldsymbol{\beta}\partial\boldsymbol{\theta}\partial\theta_j}$$

$$\frac{\partial^3 H}{\partial\boldsymbol{\beta}^2\partial\beta_i} =$$
$$-\frac{\partial^3 l}{\partial\boldsymbol{\beta}^2\partial\beta_i} - \frac{\partial^3 l}{\partial\boldsymbol{\beta}^2\partial c_j}\frac{\partial c_j}{\partial\beta_i} - \left[\frac{\partial^3 l}{\partial\mathbf{c}^2\partial\beta_i} + \frac{\partial^3 l}{\partial\mathbf{c}^2\partial c_j}\frac{\partial c_j}{\partial\beta_i}\right]\left(\frac{\partial\mathbf{c}}{\partial\boldsymbol{\beta}}\right)^2 - 2\frac{\partial^2 l}{\partial\mathbf{c}^2}\frac{\partial\mathbf{c}}{\partial\boldsymbol{\beta}}\frac{\partial^2\mathbf{c}}{\partial\boldsymbol{\beta}\partial\beta_i}$$
$$-2\frac{\partial^3 l}{\partial\mathbf{c}\partial\boldsymbol{\beta}\partial c_j}\frac{\partial c_j}{\partial\beta_i}\frac{\partial\mathbf{c}}{\partial\boldsymbol{\beta}} - 2\frac{\partial^3 l}{\partial\mathbf{c}\partial\boldsymbol{\beta}\partial\beta_i}\frac{\partial\mathbf{c}}{\partial\boldsymbol{\beta}} - 2\frac{\partial^2 l}{\partial\mathbf{c}\partial\boldsymbol{\beta}}\frac{\partial^2\mathbf{c}}{\partial\boldsymbol{\beta}\partial\beta_i} - \frac{\partial^2 l}{\partial c_j\partial\mathbf{c}}\frac{\partial\mathbf{c}}{\partial\beta_i}\frac{\partial^2 c_j}{\partial\boldsymbol{\beta}^2}$$
$$-\frac{\partial^2 l}{\partial c_j\partial\beta_i}\frac{\partial^2 c_j}{\partial\boldsymbol{\beta}^2} - \frac{\partial l}{\partial c_j}\frac{\partial^3 c_j}{\partial\boldsymbol{\beta}^2\partial\beta_i}$$

- $\frac{\partial^2\boldsymbol{\beta}}{\partial\boldsymbol{\theta}\partial\mathbf{y}}$

Similar with $\frac{\partial^2 \boldsymbol{\beta}}{\partial \boldsymbol{\theta}^2}$, we can obtain $\frac{\partial^2 \boldsymbol{\beta}}{\partial \boldsymbol{\theta} \partial \mathbf{y}}$ as:

$$\frac{\partial^2 \boldsymbol{\beta}}{\partial \theta_k \partial \mathbf{y}} = -\left[\frac{\partial^2 H}{\partial \boldsymbol{\beta}^2}\right]^{-1}\left[\frac{\partial^3 H}{\partial \boldsymbol{\beta} \partial \theta_k \partial \mathbf{y}} + \frac{\partial^3 H}{\partial \boldsymbol{\beta}^2 \partial \theta_k}\frac{\partial \boldsymbol{\beta}}{\partial \mathbf{y}} + \frac{\partial^3 H}{\partial \boldsymbol{\beta}^2 \partial \beta_i}\frac{\partial \beta_i}{\partial \mathbf{y}}\frac{\partial \boldsymbol{\beta}}{\partial \theta_k} + \frac{\partial^3 H}{\partial \boldsymbol{\beta} \partial \beta_i \partial \mathbf{y}}\frac{\partial \beta_i}{\partial \theta_k}\right] \tag{B.37}$$

where

$$\begin{aligned}
\frac{\partial^3 H}{\partial \beta_j \partial \boldsymbol{\theta} \partial \mathbf{y}} = & -\frac{\partial^3 l}{\partial \beta_j \partial \mathbf{c} \partial \mathbf{y}}\frac{\partial \mathbf{c}}{\partial \boldsymbol{\theta}} - \frac{\partial^3 l}{\partial \beta_j \partial \mathbf{c}^2}\frac{\partial \mathbf{c}}{\partial \mathbf{y}}\frac{\partial \mathbf{c}}{\partial \boldsymbol{\theta}} - \frac{\partial^2 l}{\partial \beta_j \partial c_k}\frac{\partial^2 c_k}{\partial \boldsymbol{\theta} \partial \mathbf{y}} \\
& - \frac{\partial^3 l}{\partial c_k \partial \mathbf{c} \partial \mathbf{y}}\frac{\partial \mathbf{c}}{\partial \boldsymbol{\theta}}\frac{\partial c_k}{\partial \beta_j} - \frac{\partial^3 l}{\partial c_k \partial \mathbf{c}^2}\frac{\partial \mathbf{c}}{\partial \mathbf{y}}\frac{\partial \mathbf{c}}{\partial \boldsymbol{\theta}}\frac{\partial c_k}{\partial \beta_j} - \frac{\partial^2 l}{\partial \mathbf{c} \partial c_k}\frac{\partial^2 c_k}{\partial \boldsymbol{\theta} \partial \mathbf{y}}\frac{\partial \mathbf{c}}{\partial \beta_j} \\
& - \frac{\partial^2 l}{\partial \mathbf{c}^2}\frac{\partial \mathbf{c}}{\partial \boldsymbol{\theta}}\frac{\partial^2 \mathbf{c}}{\partial \beta_j \partial \mathbf{y}} - \frac{\partial^2 l}{\partial c_k \partial \mathbf{y}}\frac{\partial^2 c_k}{\partial \beta_j \partial \boldsymbol{\theta}} - \frac{\partial^2 l}{\partial c_k \partial \mathbf{c}}\frac{\partial \mathbf{c}}{\partial \mathbf{y}}\frac{\partial^2 c_k}{\partial \beta_j \partial \boldsymbol{\theta}} \\
& - \frac{\partial l}{\partial c_k}\frac{\partial^3 c_k}{\partial \beta_j \partial \boldsymbol{\theta} \partial \mathbf{y}}
\end{aligned}$$

$$\begin{aligned}
\frac{\partial^3 H}{\partial \boldsymbol{\beta} \partial \beta_j \partial \mathbf{y}} = & -\frac{\partial^3 l}{\partial \boldsymbol{\beta} \partial \beta_j \partial \mathbf{y}} - \frac{\partial^3 l}{\partial \boldsymbol{\beta} \partial c_k \partial \mathbf{y}}\frac{\partial c_k}{\partial \beta_j} - \frac{\partial^2 l}{\partial \boldsymbol{\beta} \partial c_k}\frac{\partial^2 c_k}{\partial \beta_j \partial \mathbf{y}} \\
& - \frac{\partial^3 l}{\partial \boldsymbol{\beta} \partial \beta_j \partial \mathbf{c}}\frac{\partial \mathbf{c}}{\partial \mathbf{y}} - \frac{\partial^3 l}{\partial \boldsymbol{\beta} \partial c_k \partial \mathbf{c}}\frac{\partial \mathbf{c}}{\partial \mathbf{y}}\frac{\partial c_k}{\partial \beta_j} - \frac{\partial^3 l}{\partial \mathbf{c} \partial \beta_j \partial \mathbf{y}}\frac{\partial \mathbf{c}}{\partial \boldsymbol{\beta}} \\
& - \frac{\partial^3 l}{\partial \mathbf{c}^2 \partial \beta_j}\frac{\partial \mathbf{c}}{\partial \mathbf{y}}\frac{\partial \mathbf{c}}{\partial \boldsymbol{\beta}} - \frac{\partial^2 l}{\partial c_k \partial \beta_j}\frac{\partial^2 c_k}{\partial \boldsymbol{\beta} \partial \mathbf{y}} - \frac{\partial^3 l}{\partial \mathbf{c}^2 \partial \mathbf{y}}\frac{\partial \mathbf{c}}{\partial \beta_j}\frac{\partial \mathbf{c}}{\partial \boldsymbol{\beta}} \\
& - \frac{\partial^3 l}{\partial \mathbf{c}^2 \partial c_k}\frac{\partial c_k}{\partial \mathbf{y}}\frac{\partial \mathbf{c}}{\partial \beta_j}\frac{\partial \mathbf{c}}{\partial \boldsymbol{\beta}} - \frac{\partial^2 l}{\partial \mathbf{c}^2}\frac{\partial^2 \mathbf{c}}{\partial \beta_j \partial \mathbf{y}}\frac{\partial \mathbf{c}}{\partial \boldsymbol{\beta}} - \frac{\partial^2 l}{\partial \mathbf{c} \partial c_k}\frac{\partial \mathbf{c}}{\partial \beta_j}\frac{\partial^2 c_k}{\partial \boldsymbol{\beta} \partial \mathbf{y}} \\
& - \frac{\partial^2 l}{\partial \mathbf{c}^2}\frac{\partial \mathbf{c}}{\partial \mathbf{y}}\frac{\partial^2 \mathbf{c}}{\partial \boldsymbol{\beta} \partial \beta_j} - \frac{\partial^2 l}{\partial \mathbf{c} \partial \mathbf{y}}\frac{\partial^2 \mathbf{c}}{\partial \boldsymbol{\beta} \partial \beta_j} - \frac{\partial l}{\partial c_k}\frac{\partial^3 c_k}{\partial \boldsymbol{\beta} \partial \beta_j \partial \mathbf{y}}
\end{aligned}$$

B.3 Third Optimization Level to Estimate Complexity Parameters

When data are distributed in the exponential family, Xiang and Wahba (1996) proposed the generalized approximate cross-validation (GACV) score to choose the proper value of the smoothing parameter vector $\boldsymbol{\lambda}$. Gu and Xiang (2001) reported that the computation for the GACV score can be numerically unstable for large n, and proposed an alternative derivation of the GACV score, which is computationally stable for all sample sizes. This new GACV score is used as the third level optimization criterion:

$$F(\boldsymbol{\lambda}|\mathbf{y}) = -\frac{1}{n}\sum_{j=1}^{n}\{y_j\eta_j - b(\eta_j)\} + \frac{\alpha}{n}\frac{\mathtt{Tr}(\boldsymbol{\Phi}\mathbf{B}^{-1}\boldsymbol{\Phi}')}{n - \mathtt{Tr}\mathbf{A}}\sum_{j=1}^{n} y_j(y_j - \mu_j)\,, \qquad \text{(B.38)}$$

where $\mathbf{B} = \boldsymbol{\Phi}'\mathbf{W}\boldsymbol{\Phi} + \mathbf{R}$, $\mathbf{A} = \boldsymbol{\Phi}\mathbf{B}^{-1}\boldsymbol{\Phi}'\mathbf{W}$, $\mathbf{W} = \mathrm{diag}(w_i)$ with $w_i = \frac{\partial^2 b(\eta_i)}{\partial \eta_i^2}$, and $\alpha \geq 1$ is a constant. Gu and Ma (2003) suggested α in the range of $1.2 \sim 1.4$ to prevent severe undersmoothing typically suffered by cross-validation methods, with little loss of general effectiveness.

- $\frac{\partial F(\boldsymbol{\lambda}|\mathbf{y})}{\partial \boldsymbol{\theta}}$

The first derivative of $F(\boldsymbol{\lambda}|\mathbf{y})$ with respect to θ_l is:

$$\begin{aligned}\frac{\partial F(\boldsymbol{\lambda}|\mathbf{y})}{\partial \theta_l} = & - \frac{1}{n}\sum_{j=1}^{n}\left\{ y_j \frac{\partial \eta_j}{\partial \theta_l} - \frac{\partial b(\eta_j)}{\partial \eta_j}\frac{\partial \eta_j}{\partial \theta_l}\right\} \\ & + \frac{\alpha}{n}\frac{\partial}{\partial \theta_l}\left\{\frac{\mathtt{Tr}(\boldsymbol{\Phi}\mathbf{B}^{-1}\boldsymbol{\Phi}')}{n-\mathtt{Tr}(\mathbf{A})}\right\}\sum_{j=1}^{n} y_j(y_j-\mu_j) \\ & - \frac{\alpha}{n}\frac{\mathtt{Tr}(\boldsymbol{\Phi}\mathbf{B}^{-1}\boldsymbol{\Phi}')}{n-\mathtt{Tr}\mathbf{A}}\sum_{j=1}^{n}\left(y_j\frac{\partial \mu_j}{\partial \theta_l}\right) \end{aligned} \tag{B.39}$$

where

$$\frac{\partial \boldsymbol{\eta}}{\partial \theta_l} = \boldsymbol{\Phi}\left[\frac{\partial \mathbf{c}}{\partial \theta_l} + \frac{\partial \mathbf{c}}{\partial \boldsymbol{\beta}}\frac{\partial \boldsymbol{\beta}}{\partial \theta_l}\right] + \mathbf{X}\frac{\partial \boldsymbol{\beta}}{\partial \theta_l}$$

$$\frac{\partial}{\partial \theta_l}\left\{\frac{\mathtt{Tr}(\boldsymbol{\Phi}\mathbf{B}^{-1}\boldsymbol{\Phi}')}{n-\mathtt{Tr}(\mathbf{A})}\right\} = \frac{\mathtt{Tr}\left(\boldsymbol{\Phi}\frac{\partial \mathbf{B}^{-1}}{\partial \theta_l}\boldsymbol{\Phi}'\right)(n-\mathtt{Tr}\mathbf{A}) + \mathtt{Tr}(\boldsymbol{\Phi}\mathbf{B}^{-1}\boldsymbol{\Phi}')\left(\mathtt{Tr}\left(\frac{\partial \mathbf{A}}{\partial \theta_l}\right)\right)}{(n-\mathtt{Tr}\mathbf{A})^2}$$

$$\frac{\partial \mathbf{B}^{-1}}{\partial \theta_l} = -\mathbf{B}^{-1}\frac{\partial \mathbf{B}}{\partial \theta_l}\mathbf{B}^{-1}$$

$$\frac{\partial \mathbf{B}}{\partial \theta_l} = \boldsymbol{\Phi}'\frac{\partial \mathbf{W}}{\partial \theta_l}\boldsymbol{\Phi} + \frac{\partial R}{\partial \theta_l}$$

$$\frac{\partial \mathbf{W}}{\partial \theta_l} = \mathrm{diag}\left(\frac{\partial^3 b(\eta_i)}{\partial \eta_i^3}\frac{\partial \eta_i}{\partial \theta_l}\right)$$

$$\frac{\partial R}{\partial \theta_l} = \mathrm{diag}\left(0,\cdots,0,\lambda_l\mathbf{R}_l,0,\cdots,0\right)$$

$$\frac{\partial \mathbf{A}}{\partial \theta_l} = \boldsymbol{\Phi}\frac{\partial \mathbf{B}^{-1}}{\partial \theta_l}\boldsymbol{\Phi}'\mathbf{W} + \boldsymbol{\Phi}\mathbf{B}^{-1}\boldsymbol{\Phi}'\frac{\partial \mathbf{W}}{\partial \theta_l}$$

$$\frac{\partial \mu_j}{\partial \theta_l} = \frac{\partial^2 b(\eta_j)}{\partial^2 \eta_j}\frac{\partial \eta_j}{\partial \theta_l}$$

- $\frac{\partial^2 F(\boldsymbol{\lambda}|\mathbf{y})}{\partial^2 \boldsymbol{\theta}}$

The second derivative of $F(\boldsymbol{\lambda}|\mathbf{y})$ with respect to $\boldsymbol{\theta}$ is:

$$\begin{aligned}\frac{\partial^2 F(\boldsymbol{\lambda}|\mathbf{y})}{\partial\theta_l\partial\theta_k} = & - \frac{1}{n}\sum_{j=1}^{n}\left\{y_j\frac{\partial^2\eta_j}{\partial\theta_l\partial\theta_k} - \frac{\partial^2 b(\eta_j)}{\partial\eta_j^2}\frac{\partial\eta_j}{\partial\theta_l}\frac{\partial\eta_j}{\partial\theta_k} - \frac{\partial b(\eta_j)}{\partial\eta_j}\frac{\partial^2\eta_j}{\partial\theta_l\partial\theta_k}\right\} \\ & + \frac{\alpha}{n}\left\{\frac{\partial^2}{\partial\theta_l\partial\theta_k}\left[\frac{\mathtt{Tr}(\boldsymbol{\Phi}\mathbf{B}^{-1}\boldsymbol{\Phi}')}{n-\mathtt{Tr}(\mathbf{A})}\right]\right\}\sum_{j=1}^{n} y_j(y_j-\mu_j) \\ & - \frac{\alpha}{n}\left\{\frac{\partial}{\partial\theta_l}\left[\frac{\mathtt{Tr}(\boldsymbol{\Phi}\mathbf{B}^{-1}\boldsymbol{\Phi}')}{n-\mathtt{Tr}(\mathbf{A})}\right]\right\}\sum_{j=1}^{n}\left(y_j\frac{\partial\mu_j}{\partial\theta_k}\right) \\ & - \frac{\alpha}{n}\left\{\frac{\partial}{\partial\theta_k}\left[\frac{\mathtt{Tr}(\boldsymbol{\Phi}\mathbf{B}^{-1}\boldsymbol{\Phi}')}{n-\mathtt{Tr}(\mathbf{A})}\right]\right\}\sum_{j=1}^{n}\left(y_j\frac{\partial\mu_j}{\partial\theta_l}\right) \\ & - \frac{\alpha}{n}\frac{\mathtt{Tr}(\boldsymbol{\Phi}\mathbf{B}^{-1}\boldsymbol{\Phi}')}{n-\mathtt{Tr}\mathbf{A}}\sum_{j=1}^{n}\left(y_j\frac{\partial^2\mu_j}{\partial\theta_l\partial\theta_k}\right) \end{aligned} \tag{B.40}$$

where

$$\frac{\partial^2\eta_j}{\partial\theta_l\partial\theta_k} = \boldsymbol{\Phi}_{(j)}\left[\frac{\partial^2\mathbf{c}}{\partial\theta_l\partial\theta_k}+\frac{\partial^2\mathbf{c}}{\partial\theta_l\partial\beta_i}\frac{\partial\beta_i}{\partial\theta_k}+\frac{\partial^2\mathbf{c}}{\partial\beta\partial\beta_i}\frac{\partial\beta_i}{\partial\theta_k}\frac{\partial\beta}{\partial\theta_l}+\frac{\partial^2\mathbf{c}}{\partial\beta\partial\theta_k}\frac{\partial\beta}{\partial\theta_l}+\frac{\partial\mathbf{c}}{\partial\beta}\frac{\partial^2\beta}{\partial\theta_l\partial\theta_k}\right]+\mathbf{X}\frac{\partial^2\boldsymbol{\beta}}{\partial\theta_l\partial\theta_k}$$

$$\begin{aligned} & \frac{\partial^2}{\partial\theta_l\partial\theta_k}\left[\frac{\mathtt{Tr}(\boldsymbol{\Phi}\mathbf{B}^{-1}\boldsymbol{\Phi}')}{n-\mathtt{Tr}(\mathbf{A})}\right] \\ = & \frac{\mathtt{Tr}\left(\boldsymbol{\Phi}\frac{\partial^2\mathbf{B}^{-1}}{\partial\theta_l\partial\theta_k}\boldsymbol{\Phi}'\right)}{n-\mathtt{Tr}\mathbf{A}} \\ + & \frac{\mathtt{Tr}\left(\boldsymbol{\Phi}\frac{\partial\mathbf{B}^{-1}}{\partial\theta_l}\boldsymbol{\Phi}'\right)\mathtt{Tr}\left(\frac{\partial\mathbf{A}}{\partial\theta_k}\right)+\mathtt{Tr}\left(\boldsymbol{\Phi}\frac{\partial\mathbf{B}^{-1}}{\partial\theta_k}\boldsymbol{\Phi}'\right)\mathtt{Tr}\left(\frac{\partial\mathbf{A}}{\partial\theta_l}\right)+\mathtt{Tr}(\boldsymbol{\Phi}\mathbf{B}^{-1}\boldsymbol{\Phi}')\mathtt{Tr}\left(\frac{\partial^2\mathbf{A}}{\partial\theta_l\partial\theta_k}\right)}{(n-\mathtt{Tr}\mathbf{A})^2} \\ + & 2\frac{\mathtt{Tr}(\boldsymbol{\Phi}\mathbf{B}^{-1}\boldsymbol{\Phi}')\mathtt{Tr}\frac{\partial\mathbf{A}}{\partial\theta_l}\mathtt{Tr}\frac{\partial\mathbf{A}}{\partial\theta_k}}{(n-\mathtt{Tr}(\mathbf{A}))^3} \end{aligned}$$

$$\frac{\partial^2 \mathbf{B}^{-1}}{\partial\theta_l \partial\theta_j} = \mathbf{B}^{-1}\frac{\partial \mathbf{B}}{\partial\theta_j}\mathbf{B}^{-1}\frac{\partial \mathbf{B}}{\partial\theta_l}\mathbf{B}^{-1} - \mathbf{B}^{-1}\frac{\partial^2 \mathbf{B}}{\partial\theta_l \partial\theta_j}\mathbf{B}^{-1} + \mathbf{B}^{-1}\frac{\partial \mathbf{B}}{\partial\theta_l}\mathbf{B}^{-1}\frac{\partial \mathbf{B}}{\partial\theta_j}\mathbf{B}^{-1}$$

$$\frac{\partial^2 \mathbf{B}}{\partial\theta_l \partial\theta_j} = \mathbf{\Phi}'\frac{\partial^2 \mathbf{W}}{\partial\theta_l \partial\theta_j}\mathbf{\Phi} + \frac{\partial^2 R}{\partial\theta_l \partial\theta_j}$$

$$\frac{\partial^2 \mathbf{W}}{\partial\theta_l \partial\theta_j} = \text{diag}\left(\frac{\partial^4 b(\eta_i)}{\partial\eta_i^4}\frac{\partial\eta_i}{\partial\theta_j}\frac{\partial\eta_i}{\partial\theta_l} + \frac{\partial^3 b(\eta_i)}{\partial\eta_i^3}\frac{\partial\eta_i^2}{\partial\theta_l \partial\theta_j}\right)$$

$$\frac{\partial^2 R}{\partial\theta_l \partial\theta_j} = \text{diag}\Big(0, \cdots, 0, \lambda_l \mathbf{R}_l, 0, \cdots, 0\Big)$$

when $l = j$; otherwise,

$$\frac{\partial^2 R}{\partial\theta_l \partial\theta_j} = 0$$

$$\frac{\partial^2 \mathbf{A}}{\partial\theta_l \partial\theta_j} = \mathbf{\Phi}\frac{\partial^2 \mathbf{B}^{-1}}{\partial\theta_l \partial\theta_j}\mathbf{\Phi}'\mathbf{W} + \mathbf{\Phi}\frac{\partial \mathbf{B}^{-1}}{\partial\theta_l}\mathbf{\Phi}'\frac{\partial \mathbf{W}}{\partial\theta_j} + \mathbf{\Phi}\frac{\partial \mathbf{B}^{-1}}{\partial\theta_j}\mathbf{\Phi}'\frac{\partial \mathbf{W}}{\partial\theta_l} + \mathbf{\Phi}\mathbf{B}^{-1}\mathbf{\Phi}'\frac{\partial^2 \mathbf{W}}{\partial\theta_l \partial\theta_j}$$

$$\frac{\partial^2 \mu_j}{\partial\theta_l \partial\theta_k} = \frac{\partial^3 b(\eta_j)}{\partial^3 \eta_j}\frac{\partial\eta_j}{\partial\theta_k}\frac{\partial\eta_j}{\partial\theta_l} + \frac{\partial^2 b(\eta_j)}{\partial^2 \eta_j}\frac{\partial^2 \eta_j}{\partial\theta_l \partial\theta_k}$$

- $\frac{\partial \boldsymbol{\theta}}{\partial \mathbf{y}}$

The partial derivative of $\boldsymbol{\theta}$ with respect to $\mathbf{y}$ is:

$$\frac{\partial \boldsymbol{\theta}}{\partial \mathbf{y}} = -\left(\frac{\partial^2 F}{\partial \boldsymbol{\theta}^2}\right)^{-1}\frac{\partial^2 F}{\partial \boldsymbol{\theta}\partial \mathbf{y}} \tag{B.41}$$

where the second derivative of $F(\boldsymbol{\lambda}|\mathbf{y})$ with respect to $\boldsymbol{\lambda}$ and $\mathbf{y}$ is:

$$\begin{aligned}\frac{\partial^2 F(\boldsymbol{\lambda}|\mathbf{y})}{\partial\theta_l\partial y_k} = & -\frac{1}{n}\sum_{j=1}^{n}\left\{\frac{\partial y_j}{\partial y_k}\frac{\partial\eta_j}{\partial\theta_l}+y_j\frac{\partial^2\eta_j}{\partial\theta_l\partial y_k}-\frac{\partial^2 b(\eta_j)}{\partial\eta_j^2}\frac{\partial\eta_j}{\partial y_k}\frac{\partial\eta_j}{\partial\theta_l}-\frac{\partial b(\eta_j)}{\partial\eta_j}\frac{\partial^2\eta_j}{\partial\theta_l\partial y_k}\right\} \\ & +\frac{\alpha}{n}\frac{\partial^2}{\partial\theta_l\partial y_k}\left\{\frac{\mathtt{Tr}(\boldsymbol{\Phi}\mathbf{B}^{-1}\boldsymbol{\Phi}')}{n-\mathtt{Tr}(\mathbf{A})}\right\}\sum_{j=1}^{n}y_j(y_j-\mu_j) \\ & +\frac{\alpha}{n}\frac{\partial}{\partial\theta_l}\left\{\frac{\mathtt{Tr}(\boldsymbol{\Phi}\mathbf{B}^{-1}\boldsymbol{\Phi}')}{n-\mathtt{Tr}(\mathbf{A})}\right\}\left[y_k-\mu_k+\sum_{j=1}^{n}y_j\left(\frac{\partial y_j}{\partial y_k}-\frac{\partial\mu_j}{\partial y_k}\right)\right] \\ & -\frac{\alpha}{n}\frac{\partial}{\partial y_k}\left\{\frac{\mathtt{Tr}(\boldsymbol{\Phi}\mathbf{B}^{-1}\boldsymbol{\Phi}')}{n-\mathtt{TrA}}\right\}\sum_{j=1}^{n}\left(y_j\frac{\partial\mu_j}{\partial\theta_l}\right) \\ & -\frac{\alpha}{n}\frac{\mathtt{Tr}(\boldsymbol{\Phi}\mathbf{B}^{-1}\boldsymbol{\Phi}')}{n-\mathtt{TrA}}\left\{\frac{\partial\mu_k}{\partial\theta_l}+\sum_{j=1}^{n}\left(y_j\frac{\partial^2\mu_j}{\partial\theta_l\partial y_k}\right)\right\} \end{aligned} \tag{B.42}$$

where

$$\frac{\partial y_j}{\partial y_k}=1$$

only when $j=k$

$$\frac{\partial\boldsymbol{\eta}}{\partial y_k}=\boldsymbol{\Phi}\left(\frac{\partial\mathbf{c}}{\partial y_k}+\frac{\partial\mathbf{c}}{\partial\beta}\frac{\partial\beta}{\partial y_k}\right)+\mathbf{X}\frac{\partial\beta}{\partial y_k}$$

$$\frac{\partial^2\eta_j}{\partial\theta_l\partial y_k}=\boldsymbol{\Phi}_{(j)}\left(\frac{\partial^2\mathbf{c}}{\partial\theta_l\partial y_k}+\frac{\partial^2\mathbf{c}}{\partial\theta_l\partial\beta_i}\frac{\partial\beta_i}{\partial y_k}+\frac{\partial^2\mathbf{c}}{\partial\beta\partial y_k}\frac{\partial\beta}{\partial\theta_l}+\frac{\partial^2\mathbf{c}}{\partial\beta\partial\beta_i}\frac{\partial\beta_i}{\partial y_k}\frac{\partial\beta}{\partial\theta_l}+\frac{\partial\mathbf{c}}{\partial\beta}\frac{\partial^2\beta}{\partial\theta_l\partial y_k}\right)+\mathbf{X}\frac{\partial^2\beta}{\partial\theta_l\partial y_k}$$

$$
\begin{aligned}
&\frac{\partial^2}{\partial\theta_l\partial y_k}\left[\frac{\mathtt{Tr}(\mathbf{\Phi B}^{-1}\mathbf{\Phi}')}{n-\mathtt{Tr}(\mathbf{A})}\right] \\
= \;& \frac{\mathtt{Tr}\left(\mathbf{\Phi}\frac{\partial^2\mathbf{B}^{-1}}{\partial\theta_l\partial y_k}\mathbf{\Phi}'\right)}{n-\mathtt{Tr}\mathbf{A}} \\
+ \;& \frac{\mathtt{Tr}\left(\mathbf{\Phi}\frac{\partial\mathbf{B}^{-1}}{\partial\theta_l}\mathbf{\Phi}'\right)\mathtt{Tr}\left(\frac{\partial\mathbf{A}}{\partial y_k}\right)+\mathtt{Tr}\left(\mathbf{\Phi}\frac{\partial\mathbf{B}^{-1}}{\partial y_k}\mathbf{\Phi}'\right)\mathtt{Tr}\left(\frac{\partial\mathbf{A}}{\partial\theta_l}\right)+\mathtt{Tr}(\mathbf{\Phi B}^{-1}\mathbf{\Phi}')\mathtt{Tr}\left(\frac{\partial^2\mathbf{A}}{\partial\theta_l\partial y_k}\right)}{(n-\mathtt{Tr}\mathbf{A})^2} \\
+ \;& 2\frac{\mathtt{Tr}(\mathbf{\Phi B}^{-1}\mathbf{\Phi}')\mathtt{Tr}\frac{\partial\mathbf{A}}{\partial\theta_l}\mathtt{Tr}\frac{\partial\mathbf{A}}{\partial y_k}}{(n-\mathtt{Tr}(\mathbf{A}))^3}
\end{aligned}
$$

$$\frac{\partial^2\mathbf{B}^{-1}}{\partial\theta_l\partial y_k} = -\frac{\partial\mathbf{B}^{-1}}{\partial y_k}\frac{\partial\mathbf{B}}{\partial\theta_l}\mathbf{B}^{-1} - \mathbf{B}^{-1}\frac{\partial^2\mathbf{B}}{\partial\theta_l\partial y_k}\mathbf{B}^{-1} - \mathbf{B}^{-1}\frac{\partial\mathbf{B}}{\partial\theta_l}\frac{\partial\mathbf{B}^{-1}}{\partial y_k}$$

$$\frac{\partial\mathbf{B}^{-1}}{\partial y_k} = -\mathbf{B}^{-1}\frac{\partial\mathbf{B}}{\partial y_k}\mathbf{B}^{-1}$$

$$\frac{\partial\mathbf{B}}{\partial y_k} = \mathbf{\Phi}'\frac{\partial\mathbf{W}}{\partial y_k}\mathbf{\Phi}$$

$$\frac{\partial^2\mathbf{B}}{\partial\theta_l\partial y_k} = \mathbf{\Phi}'\frac{\partial^2\mathbf{W}}{\partial\theta_l\partial y_k}\mathbf{\Phi}$$

$$\frac{\partial\mathbf{W}}{\partial y_k} = \mathrm{diag}\left(\frac{\partial^3 b(\eta_i)}{\partial\eta_i^3}\frac{\partial\eta_i}{\partial y_k}\right)$$

$$\frac{\partial^2\mathbf{W}}{\partial\theta_l\partial y_k} = \mathrm{diag}\left(\frac{\partial^4 b(\eta_i)}{\partial\eta_i^4}\frac{\partial\eta_i}{\partial y_k}\frac{\partial\eta_i}{\partial\theta_l}+\frac{\partial^3 b(\eta_i)}{\partial\eta_i^3}\frac{\partial^2\eta_i}{\partial\theta_l\partial y_k}\right)$$

$$\frac{\partial\mathbf{A}}{\partial y_k} = \mathbf{\Phi}\frac{\partial\mathbf{B}^{-1}}{\partial y_k}\mathbf{\Phi}'\mathbf{W}+\mathbf{\Phi B}^{-1}\mathbf{\Phi}'\frac{\partial\mathbf{W}}{\partial y_k}$$

$$\frac{\partial^2\mathbf{A}}{\partial\theta_l\partial y_k} = \mathbf{\Phi}\frac{\partial^2\mathbf{B}^{-1}}{\partial\theta_l\partial y_k}\mathbf{\Phi}'\mathbf{W}+\mathbf{\Phi B}^{-1}\mathbf{\Phi}'\frac{\partial^2\mathbf{W}}{\partial\theta_l\partial y_k}$$

$$\frac{\partial^2\mu_j}{\partial\theta_l\partial y_k} = \frac{\partial^3 b(\eta_j)}{\partial\eta_j^3}\frac{\partial\eta_j}{\partial y_k}\frac{\partial\eta_j}{\partial\theta_l}+\frac{\partial^2 b(\eta_j)}{\partial\eta_j^2}\frac{\partial^2\eta_j}{\partial\theta_l\partial y_k}$$

$$\frac{\partial \mu_j}{\partial y_k} = \frac{\partial^2 b(\eta_j)}{\partial \eta_j^2} \frac{\partial \eta_j}{\partial \partial y_k}$$

$$\frac{\partial}{\partial y_k}\left\{\frac{\mathtt{Tr}(\mathbf{\Phi B}^{-1}\mathbf{\Phi}')}{n-\mathtt{Tr}(\mathbf{A})}\right\} = \frac{\mathtt{Tr}\left(\mathbf{\Phi}\frac{\partial \mathbf{B}^{-1}}{\partial y_k}\mathbf{\Phi}'\right)(n-\mathtt{Tr}\mathbf{A}) + \mathtt{Tr}(\mathbf{\Phi B}^{-1}\mathbf{\Phi}')\left(\mathtt{Tr}\left(\frac{\partial \mathbf{A}}{\partial y_k}\right)\right)}{(n-\mathtt{Tr}\mathbf{A})^2}$$

Bibliography

Acosta, E., H. Wu, A. Walawander, J. Eron, C. Pettinelli, S. Yu, D. Neath, E. Ferguson, A. Saah, D. Kuritzkes, and J. Gerber (2004). Comparison of two indinavir/ritonavir regimens in treatment-experienced hiv-infected individuals. *Journal of Acquired Immune Deficiency Syndromes 37*, 1358–1366.

Barndorff-Nielsen, O. (1983). On a formal for the distribution of a maximum likelihood estimator. *Biometrika 70*, 343–365.

Bates, D. M. and D. B. Watts (1988). *Nonlinear Regression Analysis and Its Applications*. New York: Wiley.

Beddington, J. (1975). Mutual interference between parasites or predators and its effects on searching efficiency. *Journal of Animal Ecology 44*, 331–340.

Benson, M. (1979). Parameter fitting in dynamic models. *Ecological Modelling 6*.

Biegler, L., J. J. Damiano, and G. E. Blau (1986). Nonlinear parameter estimation: a case study comparison. *AIChE Journal 32*, 29–45.

Bock, H. G. (1981). Numerical treatment of inverse problems in chemical reac-

tion kinetics. In K. Ebert, P. Deuflhard, and W. Jager (Eds.), *Modelling of chemical reaction systems*, pp. 102–125. New York: Springer.

Bock, H. G. (1983). Recent advances in parameter identification techniques for ode. In P. Deuflhard and E. Harrier (Eds.), *Numerical Treatment of Inverse Problems in Differential and Integral Equations*, pp. 95–121. Basel: Birkhäuser.

Bock, R. and D. Thissen (1980). *Statistical problems of fitting individual growth curves. In F.E. Johnston, A.F. Roche and C. Susanne (eds.)* Human Physical Groth and Maturation: Methodologies and Factors. New York: Plenum.

Box, G., W. G. Hunter, J. F. MacGregor, and J. Erjavec (1973). Some problems asociated with the analysis of multiresponse models. *Technometrics 15*, 33–51.

Cao, J. and D. Campbell (2006). Estimating differential equations with bayesian smoothing. Technical report, Department of Mathematics and Statistics, McGill University.

Casella, G. and R. L. Berger (1990). *Statistical Inference* (Second ed.). Pacific Grove, California: Wadsworth and Brooks/Cole.

Cox, D. and N. Reid (1987). Parameter orthogonality and approximate conditional inference. *Journal of Royal Statistical Society 49*(1), 1–39.

de Boor, C. (2001). *A Practical Guide to Splines*. New York: Springer.

de Boor, C. and B. Swartz (1973). Collocation at gaussian points. *SIAM J. Numer. Anal. 10*(4), 582–606.

DiCiccio, T., M. Martin, S. Stern, and G. Young (1996). Information bias and

adjusted profile likelihoods. *Journal of Royal Statistical Society, Series B 58.*

Diggle, P., P. Heagerty, K. Liang, and S. Zeger (2002). *Analysis of longitudinal data* (Second ed.). Oxford, U.K.: Oxford University Press.

Dominici, F., A. McDermott, and T. Hastie (2004). Improved semiparametric time series models of air pollution and mortality. *Journal of the American Statistical Association 99*(468), 938–948.

Dominici, F., A. McDermott, S. Zeger, and J. Samet (2002). On the use of generalized additive models in time-series studies of air pollution and health. *American Journal of Epidemiology 156*, 193–203.

Efron, B. (2004). The estimation of prediction error: Covariance penalties and cross-validation. *Journal of the American Statistical Association 99*, 619–642.

Esposito, W. R. and C. Floudas (2000). Deterministic global optimization in nonlinear optimal control problems. *Journal of Global Optimization 17*, 97–126.

Fan, J. (1992). Design-adaptive nonparametric regression. *Journal of the American Statistical Association 87*, 998–1004.

Fan, J. and I. Gijbels (1992). Variable bandwidth and local linear regression smoothers. *Ann. Statist. 20*, 2008–2036.

Fan, J. and R. Li (2004). New estimation and model selection procedures for semiparametric modeling in longitudinal data analysis. *Journal of the American Statistical Association 99*, 710–723.

Ferguson, H., N. Reid, and D. Cox (1991). Estimaing equations based on modified profile likelihood. In V. Godambe (Ed.), *Estimating Functions*, pp. 279–

293. Oxford: Oxford University Press.

Fuguitt, R. and J. E. Hawkins (1945). The liquid-phase thermal isomerization of α-pinene. *Journal of the American Chemical Society 67*, 242–245.

Fuguitt, R. and J. E. Hawkins (1947). Rate of the thermal isomerization of α-pinene in the liquid phase. *Journal of the American Chemical Society 69*, 319–322.

Fussmann, G. F., S. P. Ellner, K. W. Shertzer, and N. G. J. Hairston (2000). Crossing the hopf bifurcation in a live predator-prey system. *Science 290*, 1358–1360.

Gasser, T. and H. G. Müller (1979). Kernel estimation of regression functions. In *Smoothing Techniques for Curve Estimation*, 757, pp. 23–68. New York: Springer-Verlag.

Gelman, A., F. Bois, and J. Jiang (1996). Physiological pharmacokinetic analysis using population modeling and informative prior distributions. *Journal of the American Statistical Association 91*, 1400–1412.

Gelman, A., J. B. Carlin, H. Stern, and D. B. Rubin (2004). *Bayesian Data Analysis*. New York: Chapman and Hall/CRC.

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

Gu, C. and P. Ma (2003). Generalized nonparametric mixed-effect models: Computation and smoothing parameter selection. To appear in Journal of Computational and Graphical Statistics.

Gu, C. and D. Xiang (2001). Cross-validating non-gaussian data: Generalized approximate cross-validation revisited. *Journal of Computational and Graph-*

ical Statistics 10, 581–591.

Hassell, M. (1978). *The Dynamics of Arthropod Predator-Prey Systems.* New York: Princeton University Press.

Himmelblau, D., C. Jones, and K. B. Bischoff (1967). Determination of rate constants for complex kinetics models. *Industrial Engineering Chemistry Fundamentals 6*, 539.

Holling, C. S. (1959). Some characteristics of simple types of predation and parasitism. *Canadian Entomologist 91*, 385–398.

Horbelt, W., J. Timmer, and H. Voss (2002). Parameter estimation in nonlinear delayed feedback systems from noisy data. *Physics letters A 299*, 513–521.

Huang, Y., D. Liu, and H. Wu (2005). Hierachical bayesian methods for estimation of parameters in a longitudinal hiv dynamic system. Submitted to Biometrics.

Jaeger, J., M. Blagov, D. Kosman, K. Kozlov, M. Myasnikova, S. Surkova, C. Vanario-Alonso, M. Samsonova, D. Sharp, and J. Reinitz (2004). Dynamical analysis of regulatory interactions in the gap gene system of drosophila melanogaster. *Genetics 167*, 1721–1737.

Jolicoeur, P., J. Pontier, and H. Abidi (1992). Asymptotic models for the longitudinal growth of human stature. *American Jounal of Human Biology* (4), 461–468.

Liang, H., W. Hardle, and R. J. Carroll (1999). Estimation in a semiparamtric partially linear errors-in-variables model. *The Annals of Statistics 27*, 1519–1535.

Liang, K. (1987). Estimating functions and approximate conditional likelihood. *Biometrika 74*, 695–702.

Lin, D. and Z. Ying (2001). Semiparamtric and nonparametric regression analysis of longitudinal data. *Journal of the American Statistical Association 96*, 103–126.

Lin, X. and R. Carroll (2001). Semiparamtric regression for clustered data using generalized estimating equations. *Journal of the American Statistical Association 96*, 1045–1056.

Lin, X. and R. Carroll (2006). Semiparamtric estimation in general repeated measures problems. *Journal of the Royal Statistical Society, Series B 68*, 69–88.

McNair, J. (1987). A reconciliation of simple and complex models of age-dependent predation. *Theoretical Population Biology 32*, 383–392.

Müller, T. G. and J. Timmer (2004). Parameter identification techniques for partial differential equations. *International Journal of Bifurcation and Chaos 14*, 2053–2060.

Murdoch, W., C. Briggs, and R. Nisbet (2003). *Consumer-Resource Dynamics*. New York: Princeton University Press.

Murdoch, W. and A. Stewart-Oaten (1975). Aggregation by parasitoids and predators: effects on equilibrium and stability. *American Naturalist 134*, 288–310.

Nadaraya, E. (1964). On estimating regression. *Theory Prob. Appl. 9*, 141–142.

Neyman, J. and E. L. Scott (1948). Consistent estimates based on partially

consistent observations. *Econometrika 16*, 1–32.

Oaten, A. and W. Murdoch (1975). Functional response and stability in predator-prey systems. *American Naturalist 109*, 289–298.

Ramsay, J. O., G. Hooker, J. Cao, and D. Campbell (2005). Estimating differential equations. Submitted to Journal of the Royal Statistical Society, Series B.

Ramsay, J. O. and B. W. Silverman (2005). *Functional Data Analysis* (Second ed.). New York: Springer.

Ramsay, T. (2005). Bias in semiparametric additive models. Technical report, University of Ottawa.

Ramsay, T., R. Burnett, and D. Krewski (2003). The effect of concurvity in generalized additive models linking mortality to ambient particulate matter. *Epidemiology 14*(1), 18–23.

Roche, A. (1991). *Growth, Maturation and Body Composition: The Fels Longitudinal Study 1929 - 1991*. Cambridge: Cambridge Press.

Rosenbrock, H. and C. Storey (1966). *Computational Techniques for Chemical Engineers*. Oxford: Pergamon Press.

Severini, T. (2000). *Likelihood Methods in Statistics*. New York: Oxford University Press.

Severini, T. and J. Staniswalis (1994). Quasi-likelihood estimation in semiparametric models. *Journal of the American Statistical Association 89*, 501–511.

Stein, C. (1981). Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics 9*, 1135.

Stewart, W. and J. Sorensen (1981). Bayesian estimation of common parameters from multiresponse data with missing observations. *Technometrics 23*, 131–141.

Swartz, J. and H. Bremermann (1975). Discussion of parameter estimation in biological modelling: Algorithms for estimation and evaluation of the estimates. *Journal of Mathematical Biology 1*, 241–275.

Tang, Y. P. (1971). On the estimation of rate constants for complex kinetic models. *Industrial Engineering Chemistry Fundamentals 10*, 321–322.

Timmer, J., H. Rust, W. Horbelt, and H. Voss (2000). Parametric, nonparametric and parametric modelling of chaotic circuit time series. *Physics Letters A 274*, 123–134.

Tuddenham, R. and M. Snyder (1954). Physical growth of california boys and girls from birth to eighteen years. *University of California Publications in Child Development 1*, 183–364.

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

Voss, H., M. M. Bünner, and M. Abel (1998). Indentification of continuous spaciotemporal systems. *Physical Review E 57*, 2820–2823.

Watson, G. (1964). Smooth regression analysis. *Sankhyā Ser. A 26.*

Xiang, D. and G. Wahba (1996). A generalized approximate cross validation for smoothing splines with non-gaussian data. *Statistica Sinica 6*, 675 – 692.

Zeger, S. and P. Diggle (1994). Semiparametric models for longitudinal data

with application to cd4 cell numbers in hiv seroconverters. *Biometrics 50*, 689–699.