Nonparametric Regression Methods for Longitudinal Data Analysis

HULIN WU

University of Rochester
Dept. of Biostatistics and Computer Biology
Rochester, New York

JIN-TING ZHANG

National University of Singapore
Dept. of Biostatistics and Applied Probability
Singapore

WILEY-
INTERSCIENCE

A JOHN WILEY & SONS, INC., PUBLICATION

This Page Intentionally Left Blank

Nonparametric Regression Methods for Longitudinal Data Analysis

Nonparametric Regression Methods for Longitudinal Data Analysis

HULIN WU

University of Rochester
Dept. of Biostatistics and Computer Biology
Rochester, New York

JIN-TING ZHANG

National University of Singapore
Dept. of Biostatistics and Applied Probability
Singapore

WILEY-
INTERSCIENCE

A JOHN WILEY & SONS, INC., PUBLICATION

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services or for technical support, please contact our Customer Care Department within the United States at (800) 762-2974, outside the United States at (317) 572-3993 or fax (317) 572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic format. For information about Wiley products, visit our web site at www.wiley.com.

Library of Congress Cataloging-in-Publication Data is available.

ISBN-13 978-0-471-48350-2
ISBN-10 0-471-48350-8

Printed in the United States of America.

10 9 8 7 6 5 4 3 2 1

To Chuan-Chuan, Isabella, and Gabriella
To Yan and Tian-Hui
To Our Parents and Teachers

This Page Intentionally Left Blank

Preface

Nonparametric regression methods for longitudinal data analysis have been a popular statistical research topic since the late 1990s. The needs of longitudinal data analysis from biomedical research and other scientific areas along with the recognition of the limitation of parametric models in practical data analysis have driven the development of more innovative nonparametric regression methods. Because of the flexibility in the form of regression models, nonparametric modeling approaches can play an important role in exploring longitudinal data, just as they have done for independent cross-sectional data analysis. Mixed-effects models are powerful tools for longitudinal data analysis. Linear mixed-effects models, nonlinear mixed-effects models and generalized linear mixed-effects models have been well developed to model longitudinal data, in particular, for modeling the correlations and within-subject/between-subject variations of longitudinal data. The purpose of this book is to survey the nonparametric regression techniques for longitudinal data analysis which are widely scattered throughout the literature, and more importantly, to systematically investigate the incorporation of mixed-effects modeling techniques into various nonparametric regression models.

The focus of this book is on modeling ideas and inference methodologies, although we also present some theoretical results for the justification of the proposed methods. The data analysis examples from biomedical research are used to illustrate the methodologies throughout the book. We regard the application of the statistical modeling technologies to practical scientific problems as important. In this book, we mainly concentrate on the major nonparametric regression and smoothing methods including local polynomial, regression spline, smoothing spline and penalized spline approaches. Linear and nonlinear mixed-effects models are incorporated in these smoothing methods to deal with continuous longitudinal data, and generalized linear and additive mixed-effects models are coupled with these nonparametric modeling techniques to handle discrete longitudinal data. Nonparametric models as well as semiparametric and time varying coefficient models are carefully investigated.

Chapter 1 provides a brief overview of the book chapters, and in particular, presents data examples from biomedical research studies which have motivated the use of nonparametric regression analysis approaches. Chapters 2 and 3 review mixed-effects models and nonparametric regression methods, the two important building blocks of the proposed modeling techniques. Chapters 4∼7 present the core contents of this book with each chapter covering one of the four major nonparametric regression methods including local polynomial, regression spline, smoothing spline and penalized spline. Chapters 8 and 9 extend the modeling techniques in Chapters 4∼7 to semiparametric and time varying coefficient models for longitudinal data analysis. The last chapter, Chapter 10, covers discrete longitudinal data modeling and analysis.

Most of the contents of this book should be comprehensible to readers with some basic statistical training. Advanced mathematics and technical skills are not necessary for understanding the key modeling ideas and for applying the analysis methods to practical data analysis. The materials in Chapters 1∼7 can be used in a lower or medium level graduate course in statistics or biostatistics. Chapters 8∼10 can be used in a higher level graduate course or as reference materials for those who intend to do research in this area.

We have tried our best to acknowledge the work of many investigators who have contributed to the development of the models and methodologies for nonparametric regression analysis of longitudinal data. However, it is beyond the scope of this project to prepare an exhaustive review of the vast literature in this active research field and we regret any oversight or omissions of particular authors or publications.

We would like to express our sincere thanks to Ms. Jeanne Holden-Wiltse for helping us with polishing and editing the manuscript. We are grateful to Ms. Susanne Steitz and Mr. Steve Quigley at John Wiley & Sons, Inc. who have made great efforts in coordinating the editing, review, and finally the publishing of this book. We would like to thank our colleagues, collaborators and friends, Zongwu Cai, Raymond Carroll, Jianqing Fan, Kai-Tai Fang, Hua Liang, James S. Marron, Yanqing Sun, Yuedong Wang, and Chunming Zhang for their fruitful collaborations and valuable inspirations. Thanks also go to Ollivier Hyrien, Hua Liang, Sally Thurston, and Naisyin Wang for their review and comments on some chapters of the book. We thank our families and loved ones who provided strong support and encouragement during the writing process of this book. We are grateful to our teachers and academic mentors, Fred W. Huffer, Jinhuai Zhang, Jianqing Fan, Kai-Tai Fang and James S. Marron, for guiding us to the beauty of statistical research. J.-T. Zhang also would like to acknowledge Professors Zhidong Bai, Louis H. Y. Chen, Kwok Pui Choi and Anthony Y. C. Kuk for their support and encouragement.

Wu's research was partially supported by grants from the National Institute of Allergy and Infectious Diseases, the National Institutes of Health (NIH). Zhang's research was partially supported by the National University of Singapore Academic

Research grant R-155-000-038-112. The book was written with partial support from the Department of Biostatistics and Computational Biology, University of Rochester, where the second author was a Visiting Professor.

HULIN WU AND JIN-TING ZHANG

University of Rochester
Department of Biostatistics and Computational Biology
Rochester, NY, USA
and
National University of Singapore
Department of Statistics and Applied Probability
Singapore

This Page Intentionally Left Blank

Contents

Guide to Notation

We use lowercase letters (e.g., a, x, and α) to denote scalar quantities, either fixed or random. Occasionally, we also use uppercase letters (e.g., X, Y) to denote random variables. Lowercase bold letters (e.g., $\mathbf{x}$ and $\mathbf{y}$) will be used for vectors and uppercase bold letters (e.g., $\mathbf{A}$ and $\mathbf{Y}$) will be used for matrices. Any vector is assumed to be a column vector. The transposes of a vector $\mathbf{x}$ and a matrix $\mathbf{X}$ are denoted as $\mathbf{x}^T$ and $\mathbf{X}^T$ respectively. Thus, a row vector is denoted as $\mathbf{x}^T$.

We use diag($\mathbf{a}$) to denote a diagonal matrix whose diagonal entries are the entries of $\mathbf{a}$, and use diag$(\mathbf{A}_1, \cdots, \mathbf{A}_n)$ to denote a block diagonal matrix. We use $\mathbf{A} \otimes \mathbf{B}$ to denote the Kronecker product, $(a_{ij}\mathbf{B})$, of two matrices $\mathbf{A}$ and $\mathbf{B}$.

The symbol "$\equiv$" means "equal by definition". The L_2-norm of a vector $\mathbf{x}$ is denoted as $||\mathbf{x}|| \equiv \sqrt{\mathbf{x}^T\mathbf{x}}$. For a function of a scalar x, $f^{(r)}(x) \equiv d^r f(x)/dx^r$ denotes the r-th derivative of $f(x)$. The estimator of $f^{(r)}(x)$ is denoted as $\hat{f}^{(r)}(x)$.

For a longitudinal data set, n denotes the number of subjects, n_i denotes the number of measurements for the i-th subject, and t_{ij} denotes the design time point for the j-th measurement of the i-th subject. The response value, the fixed-effects and random-effects covariate vectors at time t_{ij} are often denoted as y_{ij}, $\mathbf{x}_{ij}$ and $\mathbf{z}_{ij}$, respectively. We use $\mathbf{y}_i = [y_{i1}, \cdots, y_{in_i}]^T$, $\mathbf{X}_i = [\mathbf{x}_{i1}, \cdots, \mathbf{x}_{in_i}]^T$ and $\mathbf{Z}_i = [\mathbf{z}_{i1}, \cdots, \mathbf{z}_{in_i}]^T$ to denote the response vector, the fixed-effects and random-effects design matrices for the i-th subject, and use $\mathbf{y} = [\mathbf{y}_1^T, \cdots, \mathbf{y}_n^T]^T$, $\mathbf{X} = [\mathbf{X}_1^T, \cdots, \mathbf{X}_n^T]^T$ and $\mathbf{Z} =$ diag$(\mathbf{Z}_1, \cdots, \mathbf{Z}_n)$ to denote the response vector, the fixed-effects and random-effects design matrices for the whole data set. We often use $\boldsymbol{\alpha}, \boldsymbol{\beta}$ or $\boldsymbol{\alpha}(t), \boldsymbol{\beta}(t)$ to denote the fixed-effects or fixed-effects functions, and use $\mathbf{a}_i, \mathbf{b}_i$ or $v_i(t), \mathbf{v}_i(t)$ to denote the random-effects or random-effects functions. For the whole longitudinal data set, $\mathbf{b}$ often means $[\mathbf{b}_1^T, \cdots, \mathbf{b}_n^T]^T$.

Acronyms

AIC	Akaike Information Criterion
ASE	Average Squared Error
BIC	Bayesian Information Criterion
CSS	Cubic Smoothing Spline
CV	Cross-Validation
df	Degree of Freedom
GCV	Generalized Cross-Validation
GEE	Generalized Estimating Equation
GLME	Generalized Linear Mixed-Effects
GNPM	Generalized Nonparametric Population Mean
GNPME	Generalized Nonparametric Mixed-Effects
GSPM	Generalized Semiparametric Population Mean
GSAME	Generalized Semiparametric Additive Mixed-Effects
LME	Linear Mixed-Effects
Loglik	Log-likelihood
LPK	Local Polynomial Kernel
LPK-GEE	Local Polynomial Kernel GEE

LPME	Local Polynomial Mixed-Effects
MSE	Mean Squared Error
NLME	Nonlinear Mixed-Effects
NPM	Nonparametric Population Mean
NPME	Nonparametric Mixed-Effects
PCV	"Leave-One-Point-Out" Cross-Validation
SCV	"Leave-One-Subject-Out" Cross-Validation
SPM	Semiparametric Population Mean
SPME	Semiparametric Mixed-Effects
TVC	Time-Varying Coefficient

1

Introduction

Longitudinal data such as repeated measurements taken on each of a number of subjects over time arise frequently from many biomedical and clinical studies as well as from other scientific areas. Updated surveys on longitudinal data analysis can be found in Demidenko (2004) and Diggle et al. (2002), among others. Parametric mixed-effects models are a powerful tool for modeling the relationship between a response variable and covariates in longitudinal studies. Linear mixed-effects (LME) models and nonlinear mixed-effects (NLME) models are the two most popular examples. Several books have been published to summarize the achievements in these areas (Jones 1993, Davidian and Giltinan 1995, Vonesh and Chinchilli 1996, Pinheiro and Bates 2000, Verbeke and Molenberghs 2000, Diggle et al. 2002, and Demidenko 2004, among others). However, for many applications, parametric models may be too restrictive or limited, and sometimes unavailable at least for preliminary data analyses. To overcome this difficulty, nonparametric regression techniques have been developed for longitudinal data analysis in recent years. This book intends to survey the existing methods and introduce newly developed techniques that combine mixed-effects modeling ideas and nonparametric regression techniques for longitudinal data analysis.

1.1 MOTIVATING LONGITUDINAL DATA EXAMPLES

In longitudinal studies, data from individuals are collected repeatedly over time whereas cross-sectional studies only obtain one data point from each individual subject (i.e., a single time point per subject). Therefore, the key difference between longitudinal and cross-sectional data is that longitudinal data are usually correlated within a subject and independent between subjects, while cross-sectional data are often independent.

A challenge for longitudinal data analysis is how to account for within-subject correlations. LME and NLME models are powerful tools for handling such a problem when proper parametric models are available to relate a longitudinal response variable to its covariates. Many real-life data examples have been presented in the literature employing LME and NLME modeling techniques (Jones 1993, Davidian and Giltinan 1995, Vonesh and Chinchilli 1996, Pinheiro and Bates 2000, Verbeke and Molenberghs 2000, Diggle et al. 2002, and Demidenko 2004, among others). However, for many other practical data examples, proper parametric models may not exist or are difficult to find. Such examples from AIDS clinical trials and other biomedical studies will be presented and used throughout this book for illustration purposes. In these examples, LME and NLME models are no longer applicable, and nonparametric mixed-effects (NPME) modeling techniques, which are the focuses of this book, are a natural choice at least at the initial stage of exploratory analyses. Although the longitudinal data examples in this book are from biomedical and clinical studies, the proposed methodologies in this book are also applicable to panel data or clustered data from other scientific fields. All the data sets and the corresponding analysis computer codes in this book are freely accessible at the website: *http://www.urmc.rochester.edu/smd/biostat/people/faculty/WuSite/publications.htm.*

1.1.1 Progesterone Data

The progesterone data were collected in a study of early pregnancy loss conducted by the Institute for Toxicology and Environmental Health at the Reproductive Epidemiology Section of the California Department of Health Services, Berkeley, USA. Figures 1.1 and 1.2 show levels of urinary metabolite progesterone over the course of the women's menstrual cycles (days). The observations came from patients with healthy reproductive function enrolled in an artificial insemination clinic where insemination attempts were well-timed for each menstrual cycle. The data had been aligned by the day of ovulation (Day 0), determined by serum luteinizing hormone, and truncated at each end to present curves of equal length. Measurements were recorded once per day per cycle from 8 days before the day of ovulation and until 15 days after the ovulation. A woman may have one or several cycles. The length of the observation period is 24 days. Some measurements from some subjects were missing due to various reasons. The data set consists of two groups: the conceptive progesterone curves (22 menstrual cycles) and the nonconceptive progesterone curves (69 menstrual cycles). For more details about this data set, see Yen and Jaffe (1991), Brumback and Rice (1998), and Fan and Zhang (2000), among others.

Figure 1.1 (a) presents a spaghetti plot for the 22 raw conceptive progesterone curves. Dots indicate the level of progesterone observed in each cycle, and are connected with straight line segments. The problem of missing values is not serious here because each cycle curve has at least 17 out of 24 measurements. Overall, the raw curves present a similar pattern: before the ovulation day (Day 0), the raw curves

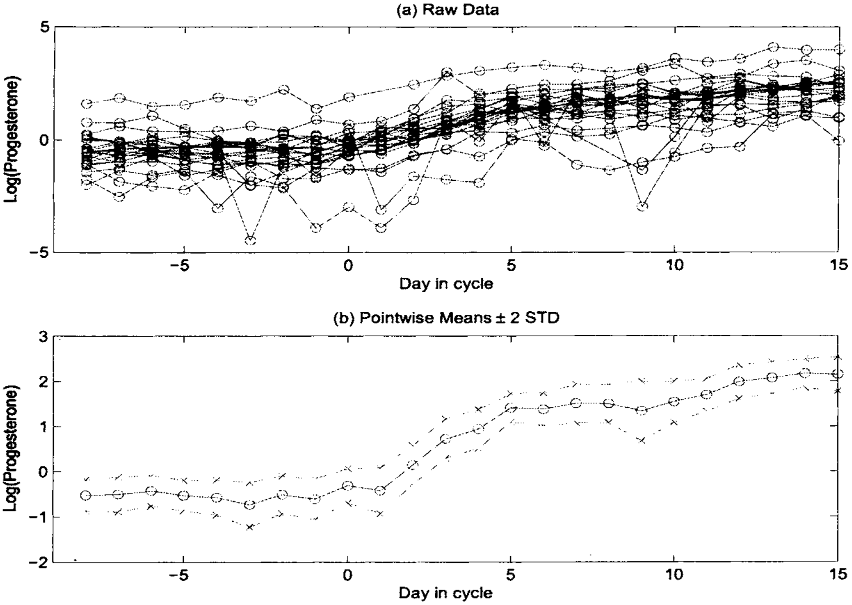

Fig. 1.1 The conceptive progesterone data.

are quite flat, but after the ovulation day, they generally move upward. However, it is easy to see that within a cycle curve, the measurements vary around some underlying curve which appears to be smooth, and for different cycles, the underlying smooth curves are different from each other. Figure 1.1 (b) presents the pointwise means (dot-dashed curve) with 95% pointwise standard deviation (SD) band (cross-dashed curves). They were obtained in a simple way: at each distinct design time point t, the mean and standard deviation were computed using the cross-sectional data at t. It can be seen that the pointwise mean curve is rather smooth, although it is not difficult to discover that there is still some noise appeared in the pointwise mean curve.

Figure 1.2 (a) presents a spaghetti plot for the 69 raw nonconceptive progesterone curves. Compared to the conceptive progesterone curves, these curves behave quite similarly before the day of ovulation, but generally show a different trend after the ovulation day. It is easy to see that, like the conceptive progesterone curves, the underlying individual cycles of the nonconceptive progesterone curves appear to be smooth, and so is their underlying mean curve. A naive estimate of the underlying mean curve is the pointwise mean curve, shown as dot-dashed curve in Figure 1.2 (b). The 95% pointwise SD band (cross-dashed curves) provides a rough estimate for the accuracy of the naive estimate.

The progesterone data have been used for illustrations of nonparametric regression methods by several authors. For example, Fan and Zhang (2000) used them to illustrate their two-step method for estimating the underlying mean function for longitudinal data or functional data, Brumback and Rice (1998) used them to illus-

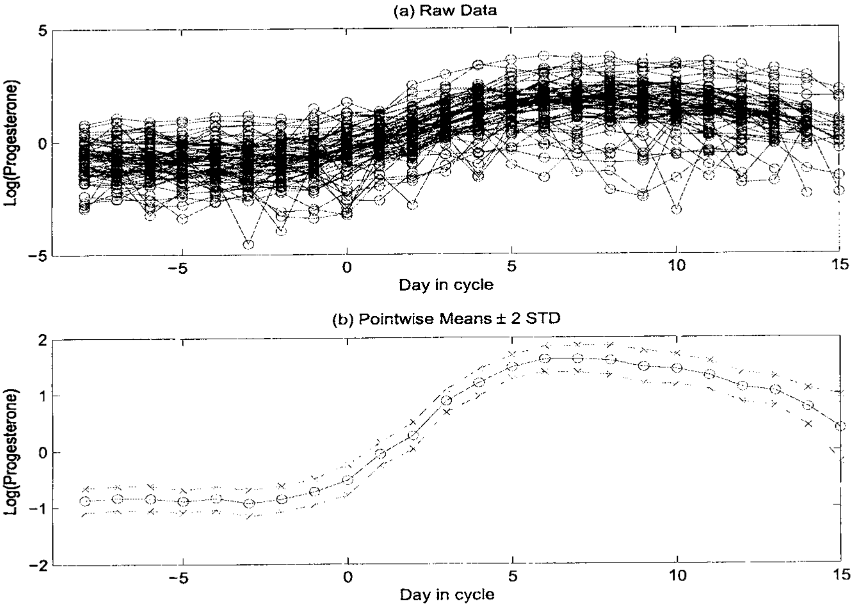

Fig. 1.2 The nonconceptive progesterone data.

trate a smoothing spline mixed-effects modeling technique for estimating both mean and individual functions, while Wu and Zhang (2002a) used them to illustrate a local polynomial mixed-effects modeling approach.

1.1.2 ACTG 388 Data

The ACTG 388 data were collected in an AIDS clinical trial study conducted by the AIDS Clinical Trials Group (ACTG). This study randomized 517 HIV-1 infected patients to three antiviral treatment arms. The data from one treatment arm will be used for illustration of the methodologies proposed in this book. This treatment arm includes 166 patients treated with highly active antiretroviral therapy (HAART) for 120 weeks during which CD4 cell counts were monitored at baseline and at weeks 4, 8, and every 8 weeks thereafter (up to 120 weeks). However, each individual patient might not exactly follow the designed schedule for measurements, and missing clinical visits for CD4 cell measurements frequently occurred. CD4 cell count is an important marker for assessing immunologic response of an antiviral regimen. Of interest are CD4 cell count trajectories over the treatment period for individual patients and for the whole treatment arm. More details about this study and scientific findings can be found in Fischl et al. (2003) and Park and Wu (2005).

The CD4 cell count data from the 166 patients during 120 weeks of treatment are plotted in Figure 1.3 (a). From this spaghetti plot, it is difficult to capture any useful information. It can be seen that the individual CD4 cell counts are quite noisy over time. We usually expect that the CD4 cell counts would increase if the antiviral treatment was effective. But from this plot, it is not easy to see any patterns among the individual patients' CD4 counts. Before a parametric model is found to fit this data set, we would have to assume that these individual curves are smooth but corrupted with noise.

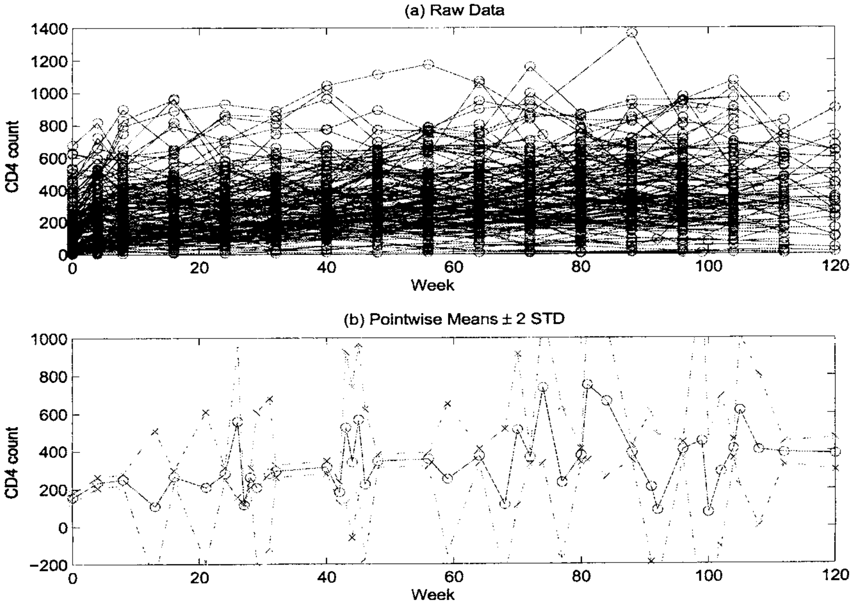

Fig. 1.3 The ACTG 388 data.

Figure 1.3 (b) presents the simple pointwise means (solid curve with dots) of the CD4 counts and their 95% pointwise SD band (cross-dashed curves). This jiggly connected pointwise mean function shows an upward trend, but it is not smooth, although the underlying mean function appears to be smooth. Moreover, the pointwise SDs are not always computable, because at some design time points (e.g., the third design time point from the right end), only a single cross-sectional data point is available. In this case, the pointwise mean is just the cross-sectional measurement itself and the pointwise SD is 0, which is not a proper measure for the accuracy of the pointwise mean. In the plot, we replaced this 0 standard deviation by the estimated standard deviation $\hat{\sigma}$ of the measurement errors, computed using all the residuals. However, this only partially solves the problem.

Without assuming parametric models for the mean and individual curves for the ACTG 388 data, nonparametric modeling techniques are then necessarily involved to handle the aforementioned problems. An example is provided by Park and Wu (2005), where they employed a kernel-based mixed-effects modeling approach.

1.1.3 MACS Data

Human immune-deficiency virus (HIV) destroys CD4 cells (T-lymphocytes, a vital component of the immune system) so that the number or percentage of CD4 cells in the blood of a patient will reduce after the subject is infected with HIV. The CD4 cell level is one of the important biomarkers to evaluate the disease progression of HIV infected subjects. To use the CD4 marker effectively in studies of new antiviral therapies or for monitoring the health status of individual subjects, it is important to build statistical models for CD4 cell count or percentage. For CD4 cell count, Lange et al. (1992) proposed Bayesian models while Zeger and Diggle (1994) employed a semiparametric model, fitted by a backfitting algorithm. For further related references, see Lange et al. (1992).

A subset of HIV monitoring data from the Multi-center AIDS Cohort Study (MACS) contains the HIV status of 283 homosexual men who were infected with HIV during the follow-up period between 1984 and 1991. Kaslow et al. (1987) presented the details for the related design, methods and medical implications of this study. The response variable is the CD4 cell percentage of a subject at a number of design time points after HIV infection. Three covariates were assessed in this study. The first one, "Smoking", takes the values of 1 or 0, according to whether a subject is a smoker or nonsmoker, respectively. The second covariate, "Age", is the age of a subject at the time of HIV infection. The third covariate, "PreCD4", is the last measured CD4 cell percentage level prior to HIV infection. All three covariates are time-independent and subject-specific. All subjects were scheduled to have clinical visits semi-annually for taking the measurements of CD4 cell percentage and other clinical status, but many subjects frequently missed their scheduled visits which resulted in unequal numbers of measurements and different measurement time points from different subjects in this longitudinal data set. We plotted the raw data from individual subjects and the simple pointwise mean of the data in Figure 1.4.

The aim of this study is to assess the effects of cigarette smoking, age at seroconversion and baseline CD4 cell percentage on the CD4 cell percentage depletion after HIV infection among the homosexual men population. From Figure 1.4, we can see that there was a trend of CD4 cell percentage depletion although the pointwise mean curve does not provide a good smooth estimate for this trend. Thus, a nonparametric modeling approach is required to characterize the CD4 cell depletion trend and to correlate this trend to the aforementioned covariates. In fact, Zeger and Diggle (1994), Wu and Chiang (2000), Fan and Zhang (2000), Rice and Wu (2001), Huang, Wu and Zhou (2002), among others have applied various nonparametric regression methods including time varying coefficient models to this data set. Similarly, we will use this data set to illustrate the proposed nonparametric regression models and smoothing methods in the succeeding chapters.