Semiparametric Models for Longitudinal Data with Application to CD4 Cell Numbers in HIV Seroconverters

Author(s): Scott L. Zeger and Peter J. Diggle

Source: *Biometrics*, Vol. 50, No. 3 (Sep., 1994), pp. 689-699
Published by: International Biometric Society
Stable URL: http://www.jstor.org/stable/2532783
Accessed: 04-01-2016 18:43 UTC

BIOMETRICS 50, 689–699
September 1994

Semiparametric Models for Longitudinal Data with Application to CD4 Cell Numbers in HIV Seroconverters

Scott L. Zeger

Department of Biostatistics, Johns Hopkins University,
615 N. Wolfe Street, Baltimore, Maryland 21205-2179, U.S.A.

and

Peter J. Diggle

Department of Mathematics, Lancaster University,
Lancaster LA1 4YL, England

SUMMARY

The paper describes a semiparametric model for longitudinal data which is illustrated by its application to data on the time evolution of CD4 cell numbers in HIV seroconverters. The essential ingredients of the model are a parametric linear model for covariate adjustment, a nonparametric estimation of a smooth time trend, serial correlation between measurements on an individual subject, and random measurement error. A back-fitting algorithm is used in conjunction with a cross-validation prescription to fit the model. A notable feature in the application is that the onset of HIV infection is associated with a sudden drop in CD4 cells followed by a longer-term slower decay. The model is also used to estimate an individual's curve by combining his data with the population curve. Shrinkage toward the population mean trajectory is controlled in a natural way by the estimated covariance structure of the data.

1. Introduction

The average interval from infection by the human immune-deficiency virus (HIV) to AIDS is estimated to be approximately 11 years (e.g., Muñoz et al., 1989). A major difficulty in studying progression of HIV infection is that an unreasonably long follow-up is necessary to observe a large cohort over the entire incubation period from infection to disease. An alternative to monitoring the presence or absence of AIDS is to follow a surrogate marker of disease progression for shorter intervals.

HIV destroys T-lymphocytes called CD4 cells, which play a vital role in immune function. Disease progression can be assessed by measuring the number or percent of CD4 cells, which on average decrease throughout the disease incubation period. In addition, CD4 cell levels may be related to cofactors such as age or smoking. To use the CD4 marker effectively in studies of new therapies or for monitoring individual subjects, it is essential to characterize the typical time course of CD4 cell loss as well as the natural variability across subjects in the decay curves.

In this paper, we employ a semiparametric statistical methodology for: (i) estimating the average time-course of CD4 cell loss while adjusting for other covariates and allowing for time-dependence within each person's measurements; and (ii) characterizing the deviation of each person from this typical curve, thereby estimating his progression curve. The deviations are potentially useful in identifying subjects who are progressing more slowly with the hope of finding factors associated with their longer incubation periods. The individual curves can also be used in counseling about disease progression. We adopt a nonparametric model for the average time curve to allow sufficient flexibility to uncover unexpected patterns. Covariates are taken into account using a linear regression model. Finally, each subject's deviation from the typical curve is assumed to be a realization

Key words: AIDS; CD4 cell number; Cross-validation; Kernel estimation; Kernel regression; Longitudinal data analysis; Nonparametric regression; Panel studies; Repeated measures; Semiparametric modelling.

of a stationary Gaussian process. One characteristic of the CD4 example and of most longitudinal data is that observations are made at irregular times. The methodology is designed to take advantage of the extra information about the typical curve available with irregular observation times.

The motivating data are from the Multicenter AIDS Cohort Study or "MACS," which has followed nearly 5,000 gay and bisexual men from Baltimore, Pittsburgh, Chicago, and Los Angeles since 1984. The broad objective of the study is to characterize the natural history of HIV infection. The study includes 1,809 (37%) men who were infected with HIV when the study began and another 371 (7%) men who were seronegative at entry and seroconverted during the follow-up. As of October 1990, 661 men have contracted AIDS. The study team has administered interviews and physical examinations semiannually to measure psychosocial, behavioral, hematologic, serologic, and virologic variables. Participants are now in their twenty-first visit. Details of the study design and methods are reported by Kaslow et al. (1987).

Figure 1 displays 2,376 measurements of CD4 cell numbers plotted against time since seroconversion ($t = 0$) for 369 seroconverters (data from two subjects were dropped because of missing covariate information). The observations for one man are connected to make the longitudinal nature of the data clear. The purpose of this paper is to develop and apply statistical methods to characterize the typical time course of CD4 cell loss in this population and to obtain precise estimates of individual curves to be used, for example, in counseling.

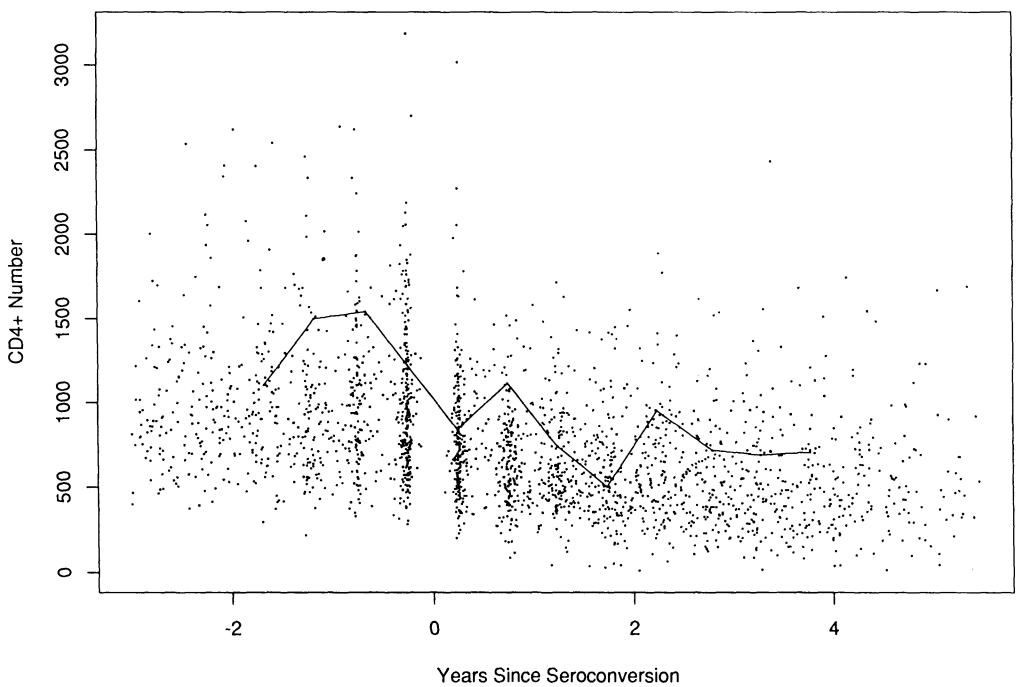

Figure 1. Multicenter AIDS Cohort Study data on CD4 numbers plotted against time from HIV seroconversion. The solid line connects the repeated observations for one man.

Several investigators have used parametric models to describe progression of HIV infection with CD4 cell numbers. Taylor et al. (1984) and Fahey et al. (1990) consider the prediction of time-to-AIDS using CD4 numbers and other markers. Lang et al. (1979) describe the patterns of CD4 cell loss. Lange, Carlin, and Gelfand (1992) take a Bayesian approach to fitting change point models to CD4 cell numbers over time.

While the methods developed in this paper are motivated by and applied to the CD4 data, they are semiparametric analogues of standard linear models for longitudinal data (e.g., Laird and Ware, 1982) and hence have application to a range of longitudinal data analysis problems. Many semiparametric and nonparametric regression methods for independent responses have appeared in the last decade. Key references are by Cleveland (1979), Silverman (1985), Hastie and Tibshirani (1986), Green (1987), and Speckman (1988). Less work has been completed for time series or longitudinal data. Müller (1988) applies nonparametric regression methods to longitudinal data but without incorporating serial correlation structure. Hart and Wehrly (1986) discuss a model similar to the one proposed here except that it does not take account of covariates and observation times are restricted

to a lattice. Raz, Turetsky, and Fein (1989) use nonparametric regression to estimate a signal in repeated evoked potential series, and Raz (1989) develops hypothesis testing tools for comparing signals estimated nonparametrically. Diggle and Hutchinson (1989), Hart (1991), Altman (1990), and Hurvich and Zeger (1990) consider nonparametric regression with serially correlated time series data. The remainder of the paper is organized as follows. Section 2 formulates our model for longitudinal data. Section 3 develops the semiparametric regression using a locally adaptive kernel estimate of the typical curve. Section 4 considers empirical Bayes estimators of the subject-specific deviations. Section 5 presents the application of our methodology to the CD4 data.

2. Model Formulation

We represent the data as an incomplete array of measurements $\{y_{ij}(t_{ij}): j = 1, \ldots, m_i; i = 1, \ldots, n\}$, in which $y_{ij}(t_{ij})$ is the jth measurement on the ith subject and t_{ij} the corresponding time (relative to seroconversion) at which this measurement was made. Where there is no ambiguity, we abbreviate by dropping the explicit dependence on time, for example to $y_{ij}(t_{ij}) = y_{ij}$. Also let $N = \Sigma_{i=1}^{n} m_i$ represent the total number of observations.

Our semiparametric model takes the form

$$Y_{ij}(t) = x_{ij}'\beta + \mu(t) + W_i(t) + Z_{ij}, \tag{1}$$

where x_{ij} is a p-element vector of covariate values and β a p-element vector of regression parameters; $\mu(t)$ is an arbitrary smooth function of time (in the CD4 example, it is of primary scientific interest); $\{W_i(t): i = 1, \ldots, n\}$ are independent replicates of a zero mean, stationary Gaussian process with covariance function $\gamma(u) = \sigma_w^2\rho(u; \theta)$; $\{Z_{ij}: j = 1, \ldots, m_i; i = 1, \ldots, n\}$ are mutually independent measurement errors, each distributed as $N(0, \sigma_z^2)$.

The two essential aspects of our formulation are the semiparametric specification for the mean value structure and the serially correlated variation about the mean, reflecting the longitudinal nature of the data. The Gaussian assumption is not required for parameter estimation and, as in classical linear modelling, is involved only at the inferential stage.

Before proceeding, we establish additional notation. First, we distinguish random variables from their realized values by using uppercase and lowercase letters, respectively. Second, for vector quantities relating to individual subjects, we write $y_i = (y_{i1}, \ldots, y_{im_i})'$, $t_i = (t_{i1}, \ldots, t_{im_i})'$, $\mu_i = (\mu(t_{i1}), \ldots, \mu(t_{im_i}))'$, and $W_i = (W_{i1}, \ldots, W_{im_i})'$. Also, let X_i be the $m_i \times p$ matrix with jth row x_{ij}' and let $\zeta_i = X_i\beta + \mu_i$ be the predicted values for person i. Let V_i be the $m_i \times m_i$ covariance matrix of Y_i, with (j, k)th element $v_{jk} = \sigma_w^2\rho(t_{ij} - t_{ik}; \theta) + \sigma_Z^2\delta_{j-k}$, where δ is the Kronecker delta. Finally, to represent quantities relating to the complete set of data, write $y = (y_1', \ldots, y_n')'$, $W = (W_1', \ldots, W_n')'$, $\mu = (\mu_1', \ldots, \mu_n')'$, $\zeta = (\zeta_1', \ldots, \zeta_n')'$, and $V = \text{diag}(V_1, \ldots, V_n)$, the $N \times N$ block-diagonal covariance matrix of Y.

3. Semiparametric Regression

3.1 *Estimation of* $\mu(t)$

We first consider the case where $\beta = 0$ in (1) so that the problem of estimating $\mu(t)$ is the nonparametric regression problem for longitudinal data. Hart and Wehrly (1986) and Rice and Silverman (1991) have considered, respectively, kernel and spline approaches to this problem. However, both have assumed a common set of design points t_j $(j = 1, \ldots, m)$ for all subjects. This is problematic in the CD4 example and in most epidemiologic cohort studies. Hart and Wehrly choose the smoothing constant (bandwidth) h to minimize an estimate of mean squared error averaged over the design points. Rice and Silverman use a cross-validatory choice of the smoothing constant.

Silverman (1984) has shown that the spline estimate is approximately equivalent to a kernel estimate with local bandwidth $h(t)$ given by

$$h(t) = d\{g(t)\}^{-.25}, \tag{2}$$

where d is a positive constant and $g(t)$ represents the density of design points in the neighborhood of t. Our approach will be to use the following locally adaptive kernel estimate, as in Müller and Stadtmüller (1987). Let $K(u)$ be a nonnegative function, symmetric about $u = 0$. For each t, choose $h(t) > 0$ and define

$$c_{ij}^*(t, h(t)) = h(t)^{-1}K\{(t_{ij} - t)/h(t)\} \tag{3}$$

and

$$c_{ij}(t) = c_{ij}^*(t, h(t)) \Big/ \sum_{i=1}^{n} \sum_{j=1}^{m_i} c_{ij}^*(t, h(t)) \tag{4}$$

so that $\Sigma_{i=1}^{n} \Sigma_{j=1}^{m_i} c_{ij}(t) = 1$. Then we use the kernel estimates

$$\hat{\mu}(t) = \sum_{i=1}^{n} \sum_{j=1}^{m_i} c_{ij}(t) y_{ij}.$$

In what follows, we use the Gaussian kernel

$$K(u) = e^{-u^2/2}.$$

In general, the results obtained by kernel estimation are much less sensitive to the precise choice of $K(u)$ than they are to the choice of bandwidth.

With regard to the local bandwidth $h(t)$ in (3), we assume that the t_{ij} can be thought of as arising from some distribution $g(t)$. Then, in the vicinity of each t_{ij} we choose $h(t)$ according to (2), replacing $g(t)$ by $\hat{g}(t, b)$, a density estimate of $g(t)$ with bandwidth b that will control the degree of local adaptation of the kernel smoother. Because of the $-.25$ power in (2), the precise choice of b is not crucial. We can treat b as a tuning constant for the method rather than as a parameter to be estimated. Given b, we then choose the more crucial smoothing parameter d by an adaptation of the cross-validatory prescription in Rice and Silverman (1991). In (4), write

$$c_i(t) = \sum_{j=1}^{m_i} c_{ij}(t). \tag{5}$$

Let $\hat{\mu}^{(k)}(t)$ be the estimate of $\mu(t)$ obtained using all the data except the measurements on the kth subject,

$$\hat{\mu}^{(k)}(t) = \sum_{i \neq k} c_{ij}(t) y_{ij} / (1 - c_i(t)).$$

We choose to minimize the quantity

$$S(d) = \sum_{i=1}^{n} \sum_{j=1}^{m_i} \{y_{ij} - \hat{\mu}^{(i)}(t_{ij})\}^2. \tag{6}$$

The rationale for minimizing (6) is the following adaptation of the result in Rice and Silverman (1991).

Proposition. Let $\mathrm{MSE}(t, d) = E[\{\hat{\mu}(t) - \mu(t)\}^2]$ be the mean squared error of $\hat{\mu}(t)$ for $\mu(t)$, and $\mathrm{MSE}^{(i)}(t, d)$ the corresponding mean squared error of $\hat{\mu}^{(i)}(t)$. Then

$$E[\{y_{ij} - \hat{\mu}^{(i)}(t_{ij})\}^2] = \mathrm{var}(y_{ij}) + \mathrm{MSE}^{(i)}(t_{ij}, d).$$

Proof. As in Rice and Silverman (1991).

This result shows that $S(d)$ defined by (6) is estimating a mean squared error averaged across the design points t_{ij}, because for large n, $\mathrm{MSE}^{(i)}(t, d)$ will be approximately equal to $\mathrm{MSE}(t, d)$.

Note that this approach to estimating $\mu(t)$ does not explicitly involve the covariance structure of the data. However, by leaving out each subject's complete vector of observations, rather than one observation at a time, the method for choosing the optimal bandwidth d does take into account the time-dependence in Y_i. In contrast, the standard method of cross-validation tends to undersmooth serially correlated data because it tracks too closely the individual subject's trajectories (Diggle and Hutchinson, 1989).

Computation of $S(d)$ is eased by the following result, again adapted from Rice and Silverman (1991):

$$y_{ij} - \hat{\mu}^{(i)}(t_{ij}) = \{y_{ij} - \hat{\mu}(t_{ij})\} + \{c_i/(1 - c_i)\} \left(\sum_{j=1}^{m_i} c_{ij} y_{ij} / c_i - \hat{\mu}(t_{ij}) \right), \tag{7}$$

where $c_{ij} = c_{ij}(t_{ij})$ and $c_i = c_i(t_{ij})$ are as defined in (4) and (5). Using (7) rather than (6) avoids the necessity for explicit computation of all the leave-one-out estimates $\hat{\mu}^{(i)}(t_{ij})$. Further computational savings can be made by bringing the design points t_{ij} into a reduced set of values.

3.2 *Estimation of* β

We now return to the case where $\beta \neq 0$. The above method for estimating $\mu(t)$ can be applied iteratively with generalized least squares regression for estimating β. This is an example of the backfitting algorithm described by Hastie and Tibshirani (1986). Given the covariance matrix V_i for the vector y_i of observations for subject i, each iteration of the algorithm has the following two steps:

1. Given the generalized least squares estimate $\hat{\beta}^{[k]}$ at the kth iteration, calculate residuals $r_{ij}^{[k]} = y_{ij} - x_{ij}'\hat{\beta}^{[k]}$ and use these residuals in place of y_{ij} to calculate the kernel estimate $\hat{\mu}^{[k]}(t)$ as described above.

2. Given $\hat{\mu}^{[k]}(t)$, let $s_{ij}^{[k]} = y_{ij} - \hat{\mu}^{[k]}(t_{ij})$ and obtain $\hat{\beta}^{[k+1]}$ by generalized least squares as follows:

$$\hat{\beta}^{[k+1]} = \left(\sum_{i=1}^{n} X_i' V_i^{-1} X_i \right)^{-1} \sum_{i=1}^{n} X_i' V_i^{-1} s_i^{[k]},$$

where $s_i^{[k]} = (s_{i1}^{[k]'}, \ldots, s_{in_i}^{[k]'})'$.

3.3 *Inferences*

We now consider the variability in $\hat{\mu}(t)$ and in the fitted values $\hat{\zeta}_{ij} = x_{ij}'\hat{\beta} + \hat{\mu}(t_{ij})$. Conditioning on the chosen bandwidth d, the estimate $\hat{\mu}(t)$ is a linear function of y which we write as $\hat{\mu}(t) = s_t' y$. Its asymptotic variance is therefore given by $\text{var}\{\hat{\mu}(t)\} = s_t' V s_t$. By partitioning the N-element vector s_t into components for each individual, i.e., $s_t = (s_{1t}', \ldots, s_{nt}')'$, we can express this as

$$\text{var}\{\hat{\mu}(t)\} = \sum_{i=1}^{n} s_{it}' V_i s_{it}. \tag{8}$$

Similarly, the fitted values are linear functions of y. Letting $\hat{\zeta}_i = X_i'\hat{\beta} + \hat{\mu}_i = G_i y$, and partitioning the $m_i \times N$ matrix G_i as $G_i = [G_{i1}, \ldots, G_{in}]$, where G_{ij} is an $m_i \times m_j$ matrix, we have that

$$\text{var}(\hat{\zeta}_i) = \sum_{k=1}^{n} G_{ik} V_k G_{ik}'. \tag{9}$$

The variance formulae (8) and (9) fail to account for estimation of parameters in V and the bandwidth d, or for the bias in $\hat{\mu}(t)$. The effect of estimating V will be small when the number of subjects is large. The effect of estimating d is most likely small when $\mu(t)$ is smooth, although less is known here. The bias in $\hat{\mu}(t)$ depends on both the value of d and the form of $\mu(t)$, becoming smaller as d becomes smaller and as $\mu(t)$ becomes smoother. For our specific application to the CD4 data, we shall use a simulation experiment to check the approximate validity of our inferences.

4. Estimation of Subject-Specific Deviations $W_i(t)$

According to our model (1), the covariance structure of the complete set of measurements y_{ij} is given by

$$\text{cov}\{Y_{ij}(t_{ij}),\ Y_{kl}(t_{kl})\} = \begin{cases} \sigma_z^2 + \sigma_w^2 & i = k,\ j = l, \\ \sigma_w^2 \rho(t_{ij} - t_{kl};\ \theta) & i = k,\ j \neq l, \\ 0 & i \neq k. \end{cases} \tag{10}$$

Then, the covariance matrix of Y_i is $V_i = \sigma_z^2 I + \sigma_w^2 R(t_i, t_i)$, where for arbitrary vectors u and v, $R(u, v)$ is a matrix with (j, k)th element $\rho(u_j - v_k;\ \theta)$.

Now, let $w_i = (w_i(u_{i1}), \ldots, w_i(u_{im_i}))'$ be the vector of realized values of $W_i(u)$ at an arbitrary set of times $u_i = (u_{i1}, \ldots, u_{im_i})'$ and W_i the corresponding random vector. Then, to estimate w_i we use the conditional expectation,

$$\hat{w}_i = E[W_i(u_i) | y_i],$$

with unknown parameters replaced by their estimated values. Note that the times u_{ij} need not be observation times. We consider the distribution of $\hat{w}_i$ when β, $\mu(t)$, and V are known rather than estimated. The joint distribution of W_i and y_i is multivariate Gaussian,

$$\begin{bmatrix} W_i \\ Y_i \end{bmatrix} \sim \text{MVN}\left(\begin{bmatrix} 0 \\ X_i\beta + \mu_i \end{bmatrix}, \begin{bmatrix} \sigma_w^2 R(u_i, u_i) & \sigma_w^2 R(u_i, t_i) \\ \sigma_w^2 R(u_i, t_i) & \sigma_z^2 I + \sigma_w^2 R(t_i, t_i) \end{bmatrix} \right).$$

Then, using standard properties of the multivariate Gaussian distribution,

$$\hat{w}_i = E[W_i | Y_i = y_i] = \sigma_w^2 R(u_i, t_i)\{\sigma_z^2 I + \sigma_w^2 R(t_i, t_i)\}^{-1}(y_i - X_i\hat{\beta} - \hat{\mu}_i), \tag{11}$$

with covariance matrix

$$\text{var}(\hat{w}_i) = \sigma_w^2 R(u_i, u_i) - \sigma_w^4 R(u_i, t_i)\{\sigma_z^2 I + \sigma_w^2 R(t_i, t_i)\}^{-1} R(t_i, u_i). \tag{12}$$

As in Section 3, we suggest that this variance formula holds approximately in large samples such as the MACS cohort when β, $\mu(t)$, and V are estimated.

The role of σ_z^2 in (11) is to smooth the subject-specific deviations toward zero. Thus if $u_i = t_i$ and $\sigma_z^2 = 0$, (11) reduces to

$$\hat{w}_i = y_i - X_i\hat{\beta} - \hat{\mu}_i,$$

the vector of residuals for the ith subject after estimating β and $\mu(t)$. If $\sigma_z^2 > 0$ these residuals are smoothed toward zero in $\hat{w}_i$, which is a reasonable reflection of the fact that the fluctuations in the residuals include a component due to measurement error. Another way of looking at this is to express the subject-specific predictor at an observation point t_{ij} as

$$\hat{p}_{ij} = \hat{\zeta}_{ij} + \hat{w}_{ij} = \alpha y_{ij} + (1 - \alpha)\hat{\zeta}_{ij},$$

with $\alpha = 1$ when $\sigma_z^2 = 0$.

5. Application to CD4 Data

We now apply the semiparametric regression methodology to the CD4 data displayed in Figure 1. There are 2,376 CD4 observations on 369 subjects ranging from 3 years before to 6 years after seroconversion. A small fraction of data (2%) for subjects taking AZT (zidovudine) was excluded because AZT is known to increase CD4 and there were too few observations to model this effect. Note that CD4 cells average roughly 1,000 for seronegatives and decrease as the virus attacks the immune system. The granularity in the time variable near zero is due to the fact that date of seroconversion was assigned to the midpoint of the 6-month interval in which it occurred. A more sophisticated approach would be to recognize the error-in-variables aspect of the time variable, for example by representing the actual time of seroconversion as $t + u$, where t is the recorded time and u is a zero-mean, uniformly distributed random variable realized independently for each subject.

The first objective of this analysis is to characterize the population average time course of CD4 decay while accounting for the following additional predictor variables: smoking (packs per day); recreational drug use (yes or no); numbers of sexual partners; and depression symptoms as measured by the CESD scale (larger values indicate increased depressive symptoms). The analysis was conducted on square-root-transformed CD4 numbers whose distribution is more nearly Gaussian.

To obtain a preliminary estimate of the autocovariance function (ACF), $\mu(t)$ was approximated by a knotted cubic spline with seven equally spaced knots. An ordinary least squares regression was used to fit the spline and the five covariates above. The ACF was estimated from the residuals. We also used this preliminary regression to check for important interactions between the covariates and the time trend $\mu(t, x)$; none were found. The preliminary ACF was used to fit the semiparametric model. The ACF was then recalculated from the semiparametric model residuals and is shown in Figure 2. The variance of square root CD4 was estimated to be 37.1. The correlation between observations 6 months apart is approximately .6; correlation decreases to roughly .35 by a lag of 4 years.

By extrapolating the ACF back to lag 0, we can estimate σ_z^2 and σ_w^2. Several extrapolation techniques were employed; they all gave similar estimates. We adopt the values $\hat{\sigma}_w^2 = 23.0$ and $\hat{\sigma}_z^2 = 14.1$. Internal quality assurance studies from the MACS estimate that the laboratory error variance for CD4 number determinations is approximately 30,000 for seronegatives, which corresponds to a variance of 7.5 on the square root scale. Hence σ_z^2 is roughly half laboratory error and half short-term temporal variation.

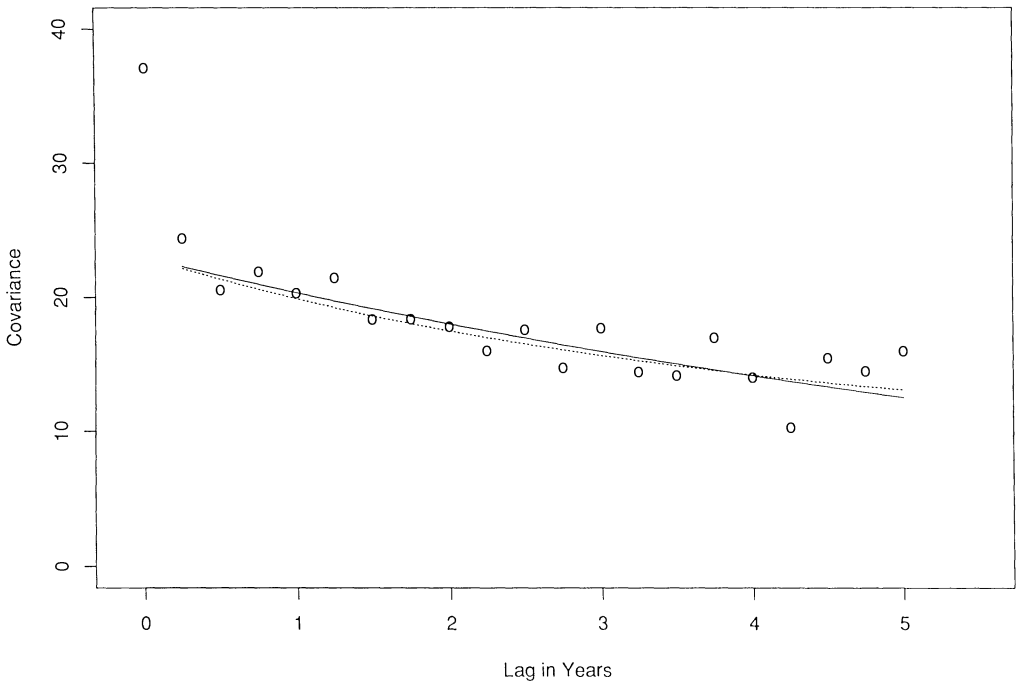

Figure 2. Empirical autocovariance function (ACF) of CD4 residuals and two fitted parametric models as described in text.
∘∘∘: empirical; ———: model A, $\alpha = 0$, $\theta = .12$; ---: model B: $\alpha = .4$, $\theta = .25$.

To be consistent with the empirical ACF in Figure 2, we model $\gamma(u)$, the autocovariance function of $W_i(t)$, as

$$\gamma(u) = \sigma_w^2\{\alpha + (1 - \alpha)e^{-\theta u}\},$$

where $0 \leq \alpha \leq 1$, $\theta > 0$, and u is the lag in years. Here the ACF will decay from σ_w^2 at lag 0 to $\alpha\sigma_w^2$ at distant lags. Unfortunately, the decay in the ACF over the observed range of 5 years is sufficiently slow to make precise estimation of α impossible. For example, Figure 2 includes fitted ACFs for two parameter settings: (A) $\alpha = 0$, $\theta = .12$, and (B) $\alpha = .4$, $\theta = .25$; both are consistent with the data. Fortunately, the semiparametric regression results depend on the ACF only for observed lags and hence do not depend on the choice of α. We will use $\alpha = .4$, $\theta = .25$ in the remainder of the analysis.

Figure 3 shows the original data with $\hat{\mu}(t)$ superimposed along with a pointwise 95% confidence interval calculated from equation (8). CD4 cell numbers are roughly constant near 1,000 cells prior to seroconversion. Within the first 6 months after seroconversion, average CD4 drops to 700. Subsequently, the rate of loss slows. It takes nearly 3 years for seroconversion to reach 500, the currently recommended level of initializing zidovudine (AZT) prophylaxis (Volberding et al., 1990).

An interesting question is whether CD4 cell numbers begin to decrease prior to seroconversion. It is possible that the virus can eliminate CD4 cells prior to the antibody reaction that indicates seroconversion. In our data, the cluster of observations 3 months prior to seroconversion ($t = -.25$) have a mean CD4 level that is the same as during the remainder of the preconversion interval as indicated by $\hat{\mu}(t)$.

The linear regression coefficients (standard errors in parentheses) for the covariates age at seroconversion (years), packs of cigarettes, recreational drug use (0: no, 1: yes), number of sexual partners, and depression score are: .037 (.18), .27 (.15), .37 (.31), .10 (.038), and −.058 (.015), respectively. Age plays little role. Smoking, recreational drug use, and increased numbers of sexual partners are associated with higher CD4 cell numbers. This may reflect immune response stimulation or simply selection bias whereby healthier men choose to continue these practices. Increased depressive symptoms are significantly associated with decreased CD4 levels. Again, a causal direction cannot be inferred from this analysis.

Figure 4 displays the observed CD4 traces for two men as well as their empirical Bayes estimates $\hat{p}_i$. Each empirical Bayes estimate compromises the individual's observations, which are contaminated by laboratory error and random daily variation, with the population-average experience. For the CD4 data, the measurement error σ_z^2 is about 40% of the total variation so that substantial

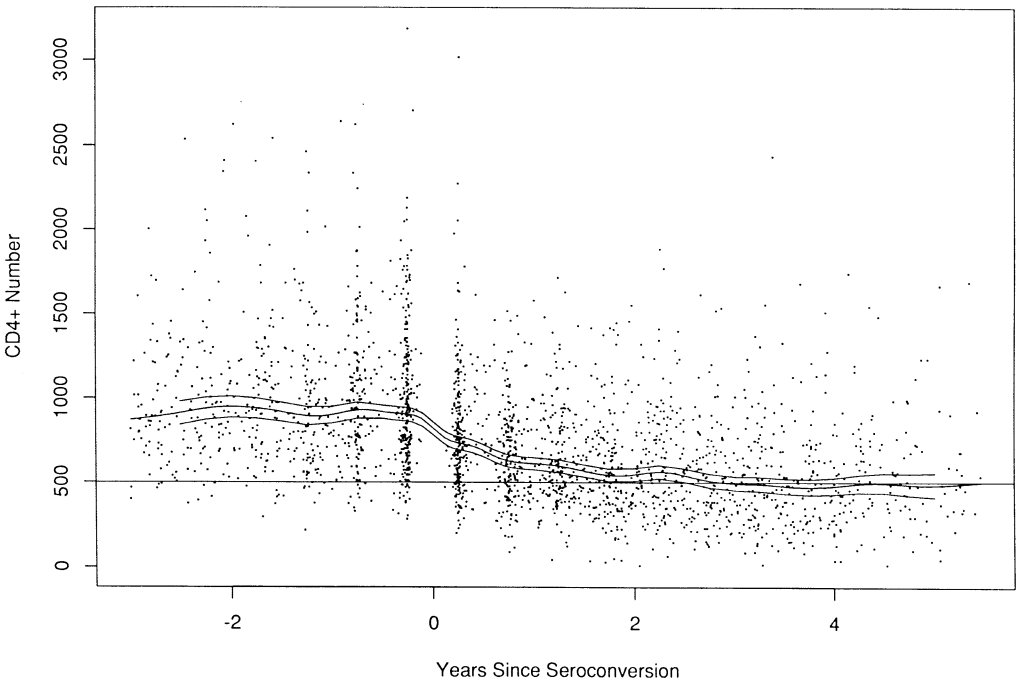

Figure 3. CD4 numbers plotted against time since HIV seroconversion, with estimated population mean curve $\hat{\mu}(t)$ plus and minus two pointwise standard errors.

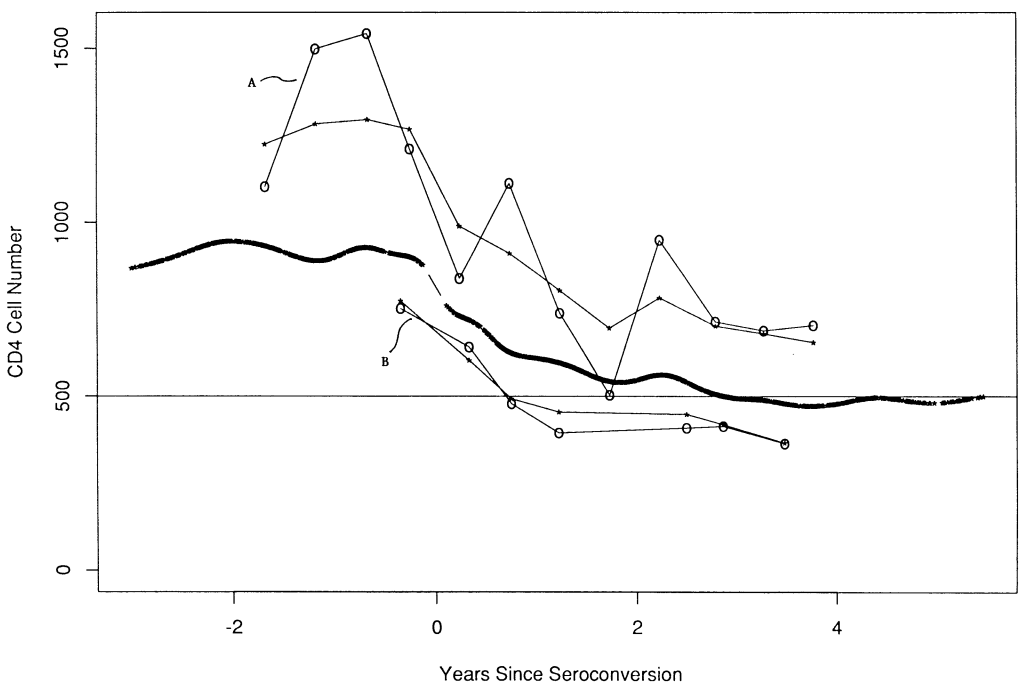

Figure 4. Raw data (∘—∘) and empirical Bayes estimates (∗—∗) for two men as well as the estimated population curve $\hat{\mu}(t)$.

shrinkage takes place. For example, the observations for participant A reach 500 cells at around 2 years after seroconversion. The smoothed curve suggests that this was a local perturbation perhaps due to laboratory error and that CD4 cells remain closer to 700 cells as much as 2 years later. AZT therapy is often initiated at 500 cells and thought to have a limited duration of effect. We believe it is preferable to initiate therapy and make other clinical decisions using $\hat{p}_i$ rather than the observed CD4 values.

The confidence intervals in Figure 3 are conditional on the estimated smoothing parameter d. The question remains whether the intervals have proper coverage given that d must be estimated, and that the estimate of $\mu(t)$ is biased. We have conducted a resampling experiment to address this issue. One hundred of the 369 individuals were selected at random and the semiparametric model was refit. The estimate of $\mu(t)$ and a 95% confidence interval were obtained for thirteen equally spaced values of t from 2 years before to 4 years after seroconversion. We determined whether the "true" $\hat{\mu}(t)$ obtained for all 369 people (shown in Figure 3) was contained within the 95% confidence interval for the sample of 100 persons at each t. Table 1 shows the actual coverage rates based on 120 repetitions

Table 1
Actual coverage rates of nominal 95% confidence intervals for $\mu(t)$

t	Actual coverage rate	t	Average coverage rate
−2.0	.93	1.0	.96
−1.5	.94	1.5	.95
−1.0	.94	2.0	.91
−.5	.92	2.5	.94
0	.94	3.0	.89
.5	.96	3.5	.94
		4.0	.87

of this experiment. The actual rates are close to the nominal 95% rate, indicating that the inferences conditioned on h are reasonable. Note that there was nontrivial variation in d over the 120 trials. The median value of d was .30; the quartiles were .14 to .41.

6. Discussion

We have extended the backfitting algorithm of Hastie and Tibshirani (1990) to semiparametric regression for longitudinal data. Correlation is explicitly accounted for by using generalized least squares to estimate β and is implicitly acknowledged by the leave-out-one-person scheme for bandwidth selection when estimating $\mu(t)$. A more complicated model in which there are multiple nonparametric curves can be accommodated by a simple extension of our approach.

The asymptotic variances for $\hat{\mu}$ presented in Section 3 appear reasonable for the CD4 example as evidenced by the brief Monte Carlo results. A detailed theoretical analysis of the properties of $\hat{\beta}$ and $\hat{\mu}(t)$ is beyond the scope of the present paper, but is given in Moyeed and Diggle (unpublished technical report, Department of Mathematics, Lancaster University, 1992). Their results add further support to the validity of our conclusions for the CD4 data, as well as showing how and when our approximations would break down in smaller samples. Bootstrapping is a viable alternative in such cases. Raz (1989) has discussed permutation tests for semiparametric regression with longitudinal data.

Moyeed and Diggle (unpublished technical report cited previously) also give a noniterative form of the backfitting algorithm for β and μ, and propose a modified estimator for β that gives somewhat better results when the explanatory variables are smoothly varying over time. The modified estimator gives worse results for time-independent covariates such as age at seroconversion in the CD4 application.

When estimating μ, we used the cross-validation criterion $S(d)$ in equation (6), which requires recomputation of $\hat{\mu}^{(i)}$ for each $i = 1, \ldots, K$. This is computationally intensive. Given a common set of observation times t_j, it is possible to estimate the covariance matrix V independently of the smoothing parameter d for $\mu(t)$. In this case, the average mean squared error of $\hat{\mu}$ can be estimated directly as shown by Hart and Wehrly (1986).

As is common in longitudinal studies, the precise specification of the model fitted to the data is somewhat empirical. In particular, the assumed autocorrelation structure can be viewed in part as a convenient and parsimonious surrogate for the cumulative effects of unobserved variables that may influence each subject's sequence of CD4 counts, rather than as a model with a direct biological interpretation. From this point of view, the near-nonidentifiability of the autocorrelation parameters is not a serious deficiency of the methodology. Of course, if we could identify additional explanatory variables we would be able to make sharper inferences about both the population mean response curve and the subject-specific trajectories.

Finally, this paper has described the average change in CD4 cell numbers with duration of

infection. We have proposed for patient counseling an empirical Bayes estimate of CD4 level that compromises a patient's own data with the population average. The time of seroconversion may not be known for some men. In such cases, it would be more appropriate to shrink toward the average CD4 at that calendar date rather than time since seroconversion.

Acknowledgements

The data for the application were collected by the Multicenter AIDS Cohort Study with centers (principal investigators) at The Johns Hopkins School of Public Health (Alfred Saah, Alvaro Muñoz); Northwestern University Medical School (John Phair); University of California, Los Angeles (Roger Detels); and University of Pittsburgh (Charles Rinaldo). The study is funded by the National Institute of Allergy and Infectious Diseases and by the National Cancer Institute. The first author gratefully acknowledges support from NIH Grant AI25529.

Résumé

L'article décrit un modèle semi-paramétrique pour les données longitudinales qui est illustré par une application à des données relatives à l'évolution du nombre des lymphocytes CD4 chez des sujets VIH-séropositifs. Le modèle a pour ingrédients essentiels: un modèle paramétrique linéaire pour l'ajustement sur les covariables, une partie non paramétrique pour une estimation lissée de la tendance au cours du temps, une corrélation entre les mesures successives pour un même individu et une erreur de mesure aléatoire. Un algorithme avec rétro-calcul est utilisé en même temps qu'une validation croisée pour ajuster le modèle. Une caractéristique remarquable de l'application est que l'infection par VIH débute par une chute soudaine des CD4 suivie d'une décroissance plus lente et plus longue. Le modèle est aussi utilisé pour estimer la courbe d'un individu en combinant ses observations avec la courbe moyenne de la population. Le rétrecissement vers la courbe moyenne de la population est controlé d'une façon naturelle en estimant la structure de la covariance des données.

References

Altman, N. (1990). Kernel smoothing of data with correlated errors. *Journal of the American Statistical Association* **85,** 749–759.

Cleveland, W. S. (1979). Robust locally-weighted regression and smoothing scatterplots. *Journal of the American Statistical Association* **74,** 829–836.

Diggle, P. J. and Hutchinson, M. F. (1989). On spline smoothing with autocorrelated errors. *Australian Journal of Statistics* **31,** 166–182.

Fahey, J. L., Taylor, J. M., Detels, R., et al. (1990). The prognostic value of cellular and serologic markers in infection with human immunodeficiency virus type I. *New England Journal of Medicine* **322,** 166–172.

Green, P. J. (1987). Penalized likelihood for general semiparametric regression models. *International Statistical Review* **55,** 245–259.

Hart, J. D. (1991). Kernel regression estimation with time series errors. *Journal of the Royal Statistical Society, Series B* **53,** 173–187.

Hart, J. D. and Wehrly, T. E. (1986). Kernel regression estimation using repeated measurements data. *Journal of the American Statistical Association* **81,** 1080–1088.

Hastie, T. and Tibshirani, R. (1986). Generalized additive models. *Statistical Science* **1,** 297–318.

Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*. London: Chapman and Hall.

Hurvich, C. and Zeger, S. L. (1990). Frequency domain selection criterion for regression with autocorrelated errors. *Journal of the American Statistical Association* **85,** 705–713.

Kaslow, R. A., Ostrow, D. G., Detels, R., et al. (1987). The Multicenter AIDS Cohort Study: Rationale, organization and selected characteristics of the participants. *American Journal of Epidemiology* **126,** 310–318.

Laird, N. M. and Ware, J. H. (1982). Random effects models for longitudinal data. *Biometrics* **38,** 963–974.

Lang, W., Perkins, H., Anderson, R., et al. (1989). Patterns of T lymphocyte changes with human immunodeficiency virus infection: From seroconversion to development of AIDS. *Journal of Acquired Immunodeficiency Syndrome* **2,** 63–69.

Lange, N., Carlin, B. P., and Gelfand, A. E. (1992). Hierarchical Bayes models for the progression of HIV infection using longitudinal CD4 T-cell numbers. *Journal of the American Statistical Association* **87,** 615–632.

Müller, H. G. (1988). *Nonparametric Regression Analysis of Longitudinal Data*. Lecture Notes in Statistics, No. 41. Berlin: Springer-Verlag.

Müller, H. G. and Stadtmüller, **U.** (1987). Estimation of heteroscedasticity in regression analysis. *Annals of Statistics* **15,** 610–625.

Muñoz, A., Wang, M.-C., Bass, S., et al. (1989). AIDS-free time after HIV-1 seroconversion in homosexual men. *American Journal of Epidemiology* **130,** 530–539.

Raz, G. (1989). Analysis of repeated measurements using nonparametric smoothing and randomization tests. *Biometrics* **45,** 851–872.

Raz, G., Turetsky, B., and Fein, G. (1989). Selecting the smoothing parameter for estimation of slowly changing evoked potential signals. *Biometrics* **45,** 745–762.

Rice, J. A. and Silverman, B. W. (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *Journal of the Royal Statistical Society, Series B* **53,** 233–243.

Silverman, B. W. (1984). Spline smoothing: The equivalent variable kernel method. *Annals of Statistics* **12,** 898–916.

Silverman, B. W. (1985). Some aspects of the spline smoothing approach to nonparametric regression curve fitting (with Discussion). *Journal of the Royal Statistical Society, Series B* **47,** 1–52.

Speckman, P. (1988). Kernel smoothing in partially linear models. *Journal of the Royal Statistical Society, Series B* **50,** 413–436.

Taylor, J., Fahey, J., Detels, R., and Giorgi, J. (1989). CD4 percentage, CD4 number, and CD4:CD8 ratios in HIV infection: Which to choose and how to use. *Journal of the Acquired Immunodeficiency Syndromes* **2,** 114–124.

Volberding, P. A., Lagakos, S. W., Koch, M. A., et al. (1990). Zidovudine in asymptomatic human immune deficiency virus infection: A controlled trial in persons with fewer than 500 CD4-positive cells per cubic millimeter. *New England Journal of Medicine* **332,** 941–949.

Received December 1991; *revised December* 1992 *and March* 1993; *accepted March* 1993.