J. Statist. Comput. Simul., 1995, Vol. 52, pp. 133–150
Reprints available directly from the publisher
Photocopying permitted by license only

FLEXIBLE METHODS FOR ANALYSING LONGITUDINAL DATA USING PIECEWISE CUBIC POLYNOMIALS

YONGXIAO WANG

National Research Center on Asian American Mental Health, UCLA Department of Psychology, Los Angeles, CA

JEREMY M. G. TAYLOR

Department of Biostatistics, UCLA School of Public Health, Los Angeles, CA

(Received January 29, 1993; in final form November 16, 1994)

We investigate a method for the analysis of repeated observations, that could arise in a clinical trial, in which there are many treatment groups, the number of observations per subject is variable, and the observations are unequally spaced. Changes in the mean of the outcome variable are described by curves defined in the follow-up period. We develop a practical and computationally feasible approach in which piecewise cubic polynomials with a large and fixed number of knots are used to parametrize the curves. Penalized likelihood estimates are used to reduce the variability and obtain smooth curves for different treatment groups. A leave-out-one-subject weighted cross-validation scheme is developed to choose the smoothing parameter λ which controls the smoothness of the curves. Some simplifying approximations to the cross-validation criterion are discussed. A simulation study is performed to evaluate the method. The result shows that using the λ chosen by cross-validation, the maximum penalized likelihood fit gives a smooth and acceptable estimate of the curves. The method is applied to AIDS clinical trial data.

KEY WORDS: Repeated measures, Penalized likelihood, Leave-out-one-subject, Cross-Validation, Smoothing.

1. INTRODUCTION AND MOTIVATION

In many biomedical studies, serial measurements are collected over a period of time for subjects allocated to one of several treatment groups. The aim of statistical methods applied to such longitudinal data is often to describe the changes in the mean of the response variable over time, to examine the differences among groups, and to describe the within-subject correlation structure (Diggle 1988, Laird and Ware 1982). Most statistical approaches assume that all the observations are measured at a few fixed time points. However frequently the observations are measured at irregular times, in which case, one would like to describe the changes by curves defined over time.

One approach is to assume these curves as linear or perhaps a low order polynomial, alternatively change point regression models or piecewise linear splines might be used. In this article we will describe a method which improves on these simple techniques, in particular the method we suggest is a hybrid of regression splines and penalized likelihood techniques.

The motivating data set for this paper is a double-blind randomized treatment-placebo clinical trial which is part of the ACTGO16 trial conducted during 1987–1988 (Bass *et al.*, 1992). Sixty one HIV positive patients, with 34 receiving the drug AZT and 27 receiving placebo, were followed from 6 to 18 months. The values of a serum marker Neopterin were measured before treatment and over the follow up period. Neopterin is known to immediately increase as a result of HIV infection, to progressively increase after infection and furthermore high values are associated with a greater risk of developing AIDS. The outcome variable is log-ratio Neopterin, which is defined as the difference between the log-Neopterin values of each subject and the mean of all the pretreatment log-Neopterin values of the same subject. The reason we use log-ratio Neopterin, rather than the log Neopterin values, is because there is an immediate and abrupt change in Neopterin when AZT treatment begins. This change is not well modelled by a smooth curve, however, by analyzing log-ratio Neopterin we eliminate this problem. The scatter-plot of the log-ratio data is in Figure 1. In this paper, we will use flexible parametric curves to describe the development of the outcome variable.

Various methods for smoothing a curve in the independent observation case can be found in the statistical literature, for example kernel smoothing (e.g. Silverman 1984, and Speckman 1988) and splines (e.g. Craven and Wahba 1979, and Silverman 1985). An extensive discussion of smoothing techniques for uncorrelated observations can be found in many articles and books (e.g. Hastie and Tibshirani 1990).

Many spline methods are the solution of a penalized least squares criterion. The magnitude of the penalty determines the smoothness of the fitted curves and is controlled by a smoothing parameter. A popular way for automatic selection of the smoothing parameter is a generalized cross-validation (GCV) method (Craven and Wahba, 1979).

One approach for estimating an approximation to an unknown curve is to use parametric regression splines with a small number of knots, and to regard the number and location of the knots as smoothing parameters to be chosen from the data (Friedman and Silverman 1989, Agarwal and Studden 1980). A related general approach is to use regression splines with a large number of knots at fixed positions, with penalty function techniques used to prevent the estimates from rapidly fluctuating. Other authors (Gray, 1992, Hastie and Tibshirani 1990) have used this method with between 8 and 20 knots. Hastie and Tibshirani (1990) refer to this method as generalized ridge regression and pseudo additive models, and they find that the results are typically indistinguishable from other spline solutions. Parker and Rice (1985) also have advocated this hybrid approach of a least squares spline with a penalty term, they found this modification to be very satisfactory and suggest that it is permissible to use evenly spaced knots because it considerably simplifies the programming. In this article because of its ease of implementation we follow this hybrid approach of specifying a fixed number of knots and using a penalty function to control the variability.

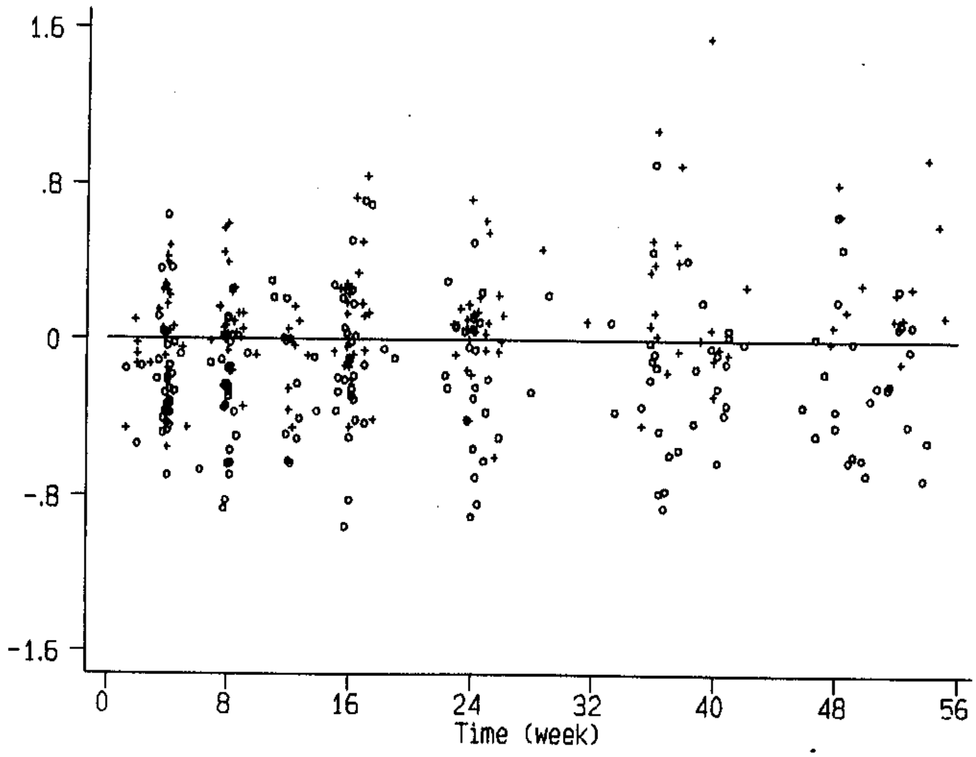


Figure 1 Log-Neopterin ratio-to-baseline versus time from randomization. "+": placebo; "o": AZT treated.

Application of flexible curve fitting techniques to longitudinal repeated measures data is less common. Various authors (Bass *et al.* 1992, DeGruttola and Tu 1992) have used piecewise linear splines with a low number of knots, with appropriate covariance structures, to model the mean function in longitudinal studies. Muller (1988) discussed kernel based and other smoothing techniques but without incorporating the serial correlation structure. Hart and Wehrly (1986) and Rice and Silverman (1991) assumed a common set of design points for all subjects. Using kernel smoothing, Hart and Wehrly (1986) suggested choosing the bandwidth, which is similar to the smoothing parameter in spline smoothing, by minimizing an estimated mean average squared error curve. Rice and Silverman (1991) chose the smoothing parameter in their spline smoothing model based on a leave-out-one-subject cross-validation method. Zeger and Diggle (1994) applied this leave-out-one-subject cross-validation method to their bandwidth selection for kernel smoothing. Instead of simplifying the cross-validation score itself, they suggested a statistic which is the estimate of the expectation of the score. Their method was applied to model the CD4 counts in HIV seroconverters in an AIDS study. All three of these papers consider only one group.

In a recent paper, Wypij, Pugh and Ware (1993) model a longitudinal data set using *B*-splines with K knots (K was chosen as 10 in their study). Because of the very large sample size in their data set, this spline approach produces smooth curves without the need for a penalty function.

In this article, under a general parametric structure for the within-subject correlations, we use piecewise cubic polynomials with smoothing by means of a penalized likelihood to estimate the curves. Thus our approach can be viewed as an improvement over using piecewise linear splines as the bias will be reduced because of the larger number of knots and in addition the curves will be smoother.

In section 2 we introduce the model with statistical assumptions, define the penalized likelihood function with a smoothing parameter, and present the asymptotic theory for the maximum likelihood estimates (MLE) and the maximum penalized likelihood estimates (MPLE). In section 3 we develop a leave-out-one-subject weighted cross-validation method and various approximations to it to choose the smoothing parameter. Section 4 gives the results of a simulation study and the analysis of the AIDS data, and Section 5 contains a brief discussion.

2. A MULTIVARIATE LINEAR MODEL

We assume M subjects from G groups, with n_i observations for the ith subject. We are thinking of applications in which n_i is in the range 6 to 10 but could be higher or lower. The total number of observations is $n = \Sigma n_i$.

Suppose that the ith subject is from group g. The model we assume for the outcome observed at time t_{ij} is

$$y_{it_{ij}} = \mu_g(t_{ij}) + e_{it_{ij}} \qquad j = 1, \ldots, n_i, \qquad i = 1, \ldots, M, \tag{2.1}$$

where $\mu_g(t)$ are the curves which describe the development of the outcome variable over time for group g. The random terms, e_{it}, are a zero mean Gaussian process, independent between subjects, but not independent within subjects. The range of t_{ij} is 0 to T.

2.1 Piecewise Cubic Polynomials

We use piecewise cubic polynomials for $\mu_g(t)$ (Hastie and Tibshirani 1990, page 22–27), which can be described as follows. Choose $K - 1$ time points between 0 and T, $0 = T_0 < T_1 < \ldots < T_K = T$ (we will refer to $T_0, T_1, \ldots, T_K$ as knots). Let $A_k = [T_k, T_{k+1}]$ be k^{th} time interval. Let $\mu_g(t)$ be a cubic polynomial in A_k continuous up to second derivatives at $T_1, \ldots, T_{K-1}$. Then $\mu_g(t)$ can be written as

$$\begin{aligned} \mu_g(t) &= \alpha_{0g} + \alpha_{1g}t + \alpha_{2g}t^2 + \alpha_{3g}t^3 + \sum_{j=4}^{K+2} \alpha_{jg}(t - T_{j-3})_+^3 \\ &= f^T(t)\alpha_g, \end{aligned} \tag{2.2}$$

where α_g is a $(K + 3)$-vector of unknown regression parameters. Furthermore, let

$\alpha^T = (\alpha_1^T, \ldots, \alpha_G^T)$, and $h_g^T(t) = (0_{1\times(g-1)(K+3)}\ f^T(t)\ 0_{1\times(G-g)(K+3)})$, then we can write the curves in a linear regression form: $\mu_g(t) = f^T(t)\alpha_g = h_g^T(t)\alpha$.

Notice that we use these piecewise cubic polynomial forms of $\mu_g(t)$ to approximate the 'true' $\mu_g(t)$ which are arbitrary unknown functions. Piecewise cubic polynomials, also called regression splines, form a large class of smooth functions, although they cerainly do not include all functions $\mu_g(t)$. It is hard to imagine a real application in which piecewise cubic polynomials with a large number of knots do not provide a sufficiently good approximation to $\mu_g(t)$. Unless the number of knots $K + 1$ is small, we use a penalized likelihood method to smooth the curves $\mu_g(t)$ since the maximum likelihood estimate will be too variable. Thus, similar to the approach of Gray (1992) and Parker and Rice (1985), our strategy is to use a large number of knots to reduce bias and a penalty function to reduce variability. In the rest of the article, unless otherwise stated, we will assume that the "true" curves in our study are piecewise cubic polynomials, that is, they can be written in the form (2.2) with K fixed. The bias associated with approximating the "true" curves by piecewise cubic polynomials with typically be small, unless there are very few knots, and particularly in comparison to the uncertainty in the estimates.

The piecewise cubic polynomial we use is an approximation to the non-parametric smoothing spline used in Craven and Wahba (1979) and Silverman (1985). The difference between a piecewise cubic polynomial and a smoothing spline is in the number of knots. The non-parametric smoothing spline, which is a solution to a specific optimization problem, uses each design point as a knot, whereas our hybrid approach uses a relatively large (between 8 and 20) but fixed number of knots. If we consider a simple cubic polynomial in which there are no knots as one extreme case, and the non-parametric smoothing spline as the other extreme, then our piecewise polynomial is intermediate of these two extremes.

This approach, although lacking the aesthetic mathematical appeal of non-parametric smoothing splines is certainly a practical and flexible method for modelling longitudinal data. Furthermore if an appropriate penalty term is subtracted from the log-likelihood in the estimation procedure then the resulting estimated curves are smooth.

An appealing feature of using a piecewise cubic polynomial is that the model is simple to fit, and its statistical properties are relatively easy to derive. Piecewise smoothing transfers the problem of estimating curves to that of estimating a finite number of unknown regression parameters. We can use the theory developed for standard linear models and ridge regression since the model can be written in a linear form.

2.2 *Log-likelihood and Penalized Log-likelihood Function*

The following notation are needed, $Y_i = (y_{it_{i1}}, \ldots, y_{it_{in_i}})^T$, $\alpha^T = (\alpha_1^T, \ldots, \alpha_G^T)$, $X_i = (h(t_{i1}), \ldots, h(t_{in_i}))^T$, then $EY_i = (\mu_g(t_{i1}), \ldots, \mu_g(t_{in_i}))^T = X_i\alpha$. Also denote $\Sigma_i = \mathrm{COV}(Y_i, Y_i)$.

Our model is $Y_i \sim N(X_i\alpha, \Sigma_i)$, where $\{Y_1, Y_2, \ldots, Y_M\}$ are independent.

The covariance structure Σ_i is assumed to have a parametric form $\Sigma_i(\phi)$, which is left general in our development, but in practice could be determined by a random-effects structure (Laird and Ware, 1982), or other more general structures derived from stochastic processes (Diggle, 1988).

The log-likelihood for the above model is

$$L(\alpha, \phi) = \text{const} - \frac{1}{2}\sum_{i=1}^{M} \log|\Sigma_i| - \frac{1}{2}\sum_{i=1}^{M} R_i^T \Sigma_i^{-1} R_i. \tag{2.3}$$

where $R_i = Y_i - X_i\alpha$.

Each curve in our model has $K + 3$ parameters. For large K this will lead to overparametization and hence instability of the estimates. One way to deal with this problem is to add a penalty term to the usual likelihood function, thus forcing additional smoothness onto the estimated curves.

The penalized log-likelihood L_p is defined by subtracting a penalty term from $L(\alpha, \phi)$. The commonly used penalty term (Wahba 1990) can be written as:

$$\begin{aligned}\text{Penalty term} &= \lambda n \sum_{g=1}^{G} \int_0^T [\mu_g''(t)]^2 \, dt \\ &= \lambda n \sum_{g=1}^{G} \int_0^T \alpha_g^T f''(t)[f''(t)]^T \alpha_g \, dt \\ &= \lambda n \sum_{g=1}^{G} \alpha_g^T \omega \alpha_g = \lambda n \alpha^T \Omega \alpha\end{aligned}$$

where $\omega = \int_0^T f''(t)[f''(t)]^T \, dt$ is a known matrix, and $\Omega = \text{diag}(\omega, \omega, \ldots, \omega)_{G\,\text{blocks}}$. Notice that if the curve for every group is a straight line then the Penalty term = 0.

The penalized log-likelihood function $L_p(\alpha, \phi; \lambda)$ is defined as:

$$L_p(\alpha, \phi; \lambda) = L(\alpha, \phi) - \lambda n \alpha^T \Omega \alpha. \tag{2.4}$$

Here λ is the smoothing parameter which controls the degree of smoothness of the estimate of $\mu_g(t)$ and can be chosen by the cross-validation scheme discussed in section 3.

One interpretation of L_p is that it is the log of the product of the likelihood and a prior for α, and hence the estimate which maximizes L_p can be interpreted as a posterior mode.

2.3 *Estimates and Their Statistical Properties*

In the rest of the article, we use $(\hat{\alpha}, \hat{\phi})$, $(\alpha_\lambda^*, \phi_\lambda^*)$ to indicate the MLE and MPLE for (α, ϕ) respectively, and denote

$$B = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{M} X_i^T \Sigma_i^{-1} X_i.$$

We can write the MLE of α as

$$\hat{\alpha} = \left(\sum_{i=1}^{M} X_i^T \hat{\Sigma}_i^{-1} X_i\right)^{-1} \left(\sum_{i=1}^{M} X_i^T \hat{\Sigma}_i^{-1} Y_i\right),$$

and the MPLE of α as

$$\alpha_\lambda^* = \left(\sum_{i=1}^{M} X_i^T \Sigma_i^{*-1} X_i + 2\lambda n \Omega\right)^{-1} \left(\sum_{i=1}^{M} X_i^T \Sigma_i^{*-1} Y_i\right)$$

where $\hat{\Sigma}_i$ and Σ_i^* are the abbreviations for $\Sigma_i(\hat{\phi})$ and $\Sigma_i(\phi_\lambda^*)$. Notice that the MPLE α_λ^* has a 'ridge' like form.

The asymptotic properties for MLE and MPLE Under certain regularity conditions including $\lambda_n \to 0$, and when M (=total number of subjects) $\to \infty$ (which implies, $n \to \infty$), we have the following

Property (i) $\sqrt{n}(\hat{\alpha} - \alpha) \xrightarrow{L} N(0, B^{-1})$

Property (ii) $\sqrt{n}(B^{-1}(B + 2\lambda_n\Omega)\alpha_{\lambda_n}^* - \alpha) \xrightarrow{L} N(0, B^{-1})$

Property (iii) $\sqrt{n}(B^{-1}(B + 2\lambda_n\Omega)\alpha_{\lambda_n}^* - \hat{\alpha}) \to 0$ *Pr.*

Property (iv) $\sqrt{n}(\phi_{\lambda_n}^* - \hat{\phi}) \to 0$ *Pr.*

Property (v) the asymptotic distributions of $\alpha_{\lambda_n}^*$ and $\phi_{\lambda_n}^*$ are independent

The proof and the regularity conditions can be found in the first author's dissertation (Wang, 1991). The necessary regularity conditions are analogous to those required to prove the consistency and asymptotic normality of maximum likelihood estimates.

Property (i) is the well known result for an MLE, properties (ii)–(v) are specific results for the MPLE.

Under the condition $\lambda_n \to 0$, Property (iv) indicates that the difference between the MLE and the MPLE of ϕ is $o_p(n^{-1/2})$. The difference between the MLE and the MPLE of α, on the other hand, is not $o_p(n^{-1/2})$ when $n^{1/2}\lambda_n$ does not go to zero as $n \to \infty$. If λ_n approaches zero faster than $n^{-1/2}$, then the MPLE and the MLE of α are asymptotically the same. This is as expected because for practical purposes when λ_n is very small, the difference between the Penalized log-likelihood and the Log-likelihood is ignorable.

3. CHOOSING A SMOOTHING PARAMETER λ BY CROSS-VALIDATION

3.1 *Review of CV Methods for Uncorrelated Observations*

In the case $y_i = \mu(t_i) + \varepsilon_i$ $(i = 1, 2, \ldots, n)$ where ε_i are uncorrelated random variables with equal variance, there is a large statistical literature on various methods to find a smooth curve as the estimate of $\mu(t)$. These methods include nonparametric splines, kernel smoothing, or piecewise polynomial splines. Despite the differences between these approaches, choosing a smoothing parameter is always a key issue in the topic. Craven and Wahba's (1979) GCV method, which was originally used for their nonparametric spline and later also for kernel smoothing, is a commonly used criterion for choosing smoothing parameters.

The penalized least square estimate of $\mu(t)$ is the minimizer of

$$\min_{\mu} \frac{1}{n} \sum_{i=1}^{n} (\mu(t_i) - y_i)^2 + \lambda \int_0^T (\mu''(u))^2 \, du.$$

The definition of the *CV* score is

$$CV(\lambda) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{\mu}_{(i)}(t_i))^2 \tag{3.1}$$

where $\hat{\mu}_{(i)}(t)$ is the penalized least square estimate of $\mu(t)$ when the i^{th} observation is omitted.

Let $(\hat{\mu}(t_1), \hat{\mu}(t_2), \ldots, \hat{\mu}(t_n))^T = A(\lambda)\,(y_1, y_2, \ldots, y_n)^T$ be the penalized least square estimate of $(\mu(t_1), \mu(t_2), \ldots, \mu(t_n))^T$, and a_{ii} be the i^{th} diagonal elements of $A(\lambda)$. Craven and Wahba (1979) showed that

$$CV(\lambda) = \frac{1}{n} \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}(t_i))^2}{(1 - a_{ii}(\lambda))^2}. \tag{3.2}$$

Thus, the *CV* score can be calculated without fitting n separate models each with one observation omitted.

GCV was obtained from (3.2) by replacing each a_{ii} by its mean, which equals $(1/n)$Trace$(A(\lambda))$: that is

$$GCV(\lambda) = \frac{\frac{1}{n} \sum_{i=1}^{n} [y_i - \hat{\mu}(t_i)]^2}{\left[1 - \frac{1}{n}\, trace(A(\lambda))\right]^2} \tag{3.3}$$

3.2 Leave-Out-One-Subject Cross-Validation and Generalized Cross-Validation for Correlated Observations

In the situation we are considering the individual observation are correlated but Y_i are independent random vectors. For this case Rice and Silverman (1991) suggest the idea of leave-out-one-subject cross-validation to choose the value of the smoothing parameter for their spline smoothing model. They assume the observations are equally spaced in time and balanced across the subjects. They calculate the *CV* score using the assumption that the data are balanced. Zeger and Diggle (1994) apply leave-out-one-subject cross-validation to choose a bandwidth in their kernel smoothing technique in the unbalanced data case. Because of the different data structure and smoothing technique, neither of the above methods can be directly applied to our unbalanced data and piecewise cubic polynomial model. In what follows, using the idea of leave-out-one-subject, we will develop an equality analogous to (3.2). We define a family of weighted *CV* criteria rather than the unweighted *CV* discussed by Rice and Silverman, and Zeger and Diggle, and derive an easily programmed approximation to CV for a specific weight.

Our model is, $Y_i = X_i\alpha + R_i$, where $\{R_1, R_2, \ldots, R_M\}$ are independent with mean zero and variance-covariance matrix $\Sigma_i(\phi)$.

There are three types of estimates of ϕ, in our procedure, they are the MLE, the MPLE and the MPLE when one subject is omitted. In developing the cross-validation criterion, we will assume that the difference between the three estimates of ϕ can be ignored when we want to smooth the curves $\mu_g(t)$. We use the same notation $\hat\Sigma_i$ to denote the estimate of Σ_i in all these cases. Ignoring-the difference can be justified because of the orthogonality between the estimates of ϕ and α and because of the *Property* (iv).

The three estimates of α are denoted by: $\hat\alpha$ for the MLE, α^*_λ for the MPLE, and $\alpha^*_{\lambda(i)}$ for the MPLE when the i^{th} subject is left out.

A weighted cross-validation criterion is defined as

$$WCV(\lambda) = \frac{1}{n}\sum_{i=1}^{M} (Y_i - X_i\alpha^*_{\lambda(i)})^T W_i^{-1}(\hat\phi)(Y_i - X_i\alpha^*_{\lambda(i)}). \tag{3.4}$$

where $W_i(\phi)$ is some positive definite matrix depending on unknown parameters ϕ. We also denote $\hat W_i = W_i(\hat\phi)$. In practice, one could use $W_i = \Sigma_i$ or $W_i = I$. The choice $W_i = \Sigma_i$ is attractive because then the covariance structure of the observations is explicitly used in estimating λ.

It is impractical to calculate $WCV(\lambda)$ by fitting M models, one for each time a subject is left out. Instead we construct a different expression for WCV which only requires the model to be fit once.

The following notation is needed for convenience.:

$$\tilde Y_i = \hat\Sigma_i^{-1/2} Y_i, \qquad \tilde X_i = \hat\Sigma_i^{-1/2} X_i, \qquad \hat C_\lambda = \sum_{k=1}^{M} \tilde X_k^T \tilde X_k + 2\lambda n\Omega.$$

LEMMA 1 $\tilde{X}_i(\hat{C}_\lambda - \tilde{X}_i^T\tilde{X}_i)^{-1} = (I_n - \tilde{X}_i\hat{C}_\lambda^{-1}\tilde{X}_i^T)^{-1}\tilde{X}_i\hat{C}_\lambda^{-1}$.
The proof requires simple matrix algebra.

LEMMA 2 $\tilde{Y}_i - \tilde{X}_i\alpha^*_{\lambda(i)} = (I_{n_i} - \tilde{X}_i\hat{C}_\lambda^{-1}\tilde{X}_i^T)^{-1}(\tilde{Y}_i - \tilde{X}_i\alpha^*_\lambda)$.

Proof. Since we are ignoring the difference between the estimates of Σ_i in the cases of the MLE, the MPLE and the MPLE when one subject is left out, then from section 2.3, $\alpha^*_\lambda = \hat{C}_\lambda^{-1}\,\Sigma_{k=1}^M\,\tilde{X}_k^T\tilde{Y}_k$ and $\alpha^*_{\lambda(i)} = (\hat{C}_\lambda - \tilde{X}_k^T\tilde{X}_i)^{-1}\,(\Sigma_{k=1}^M\,\tilde{X}_k^T\tilde{Y}_k - \tilde{X}_i^T\tilde{Y}_i)$;
Thus,

$$
\begin{aligned}
&\tilde{Y}_i - \tilde{X}_i\alpha^*_{\lambda(i)}\\
&\quad = \tilde{Y}_i - \tilde{X}_i(\hat{C}_\lambda - \tilde{X}_i^T\tilde{X}_i)^{-1}\left(\sum_{k=1}^M \tilde{X}_k^T\tilde{Y}_k - \tilde{X}_i^T\tilde{Y}_i\right)\\
&\quad = \tilde{Y}_i - (I_{n_i} - \tilde{X}_i\hat{C}_\lambda^{-1}\tilde{X}_i^T)^{-1}\tilde{X}_i\hat{C}_\lambda^{-1}\left(\sum_{k=1}^M \tilde{X}_k^T\tilde{Y}_k - \tilde{X}_i^T\tilde{Y}_i\right) \qquad \text{(from Lemma 1)}\\
&\quad = (I_{n_i} + (I_{n_i} - \tilde{X}_i\tilde{C}_\lambda^{-1}\tilde{X}_i^T)^{-1}\tilde{X}_i\hat{C}_\lambda^{-1}\tilde{X}_i^T)\tilde{Y}_i - (I_{n_i} - \tilde{X}_i\hat{C}_\lambda^{-1}\tilde{X}_i^T)^{-1}\tilde{X}_i\hat{C}_\lambda^{-1}\sum_{k=1}^M \tilde{X}_k^T\tilde{Y}_k\\
&\quad = (I_{n_i} - \tilde{X}_i\hat{C}_\lambda^{-1}\tilde{X}_i^T)^{-1}\tilde{Y}_i - (I_{n_i} - \tilde{X}_i\hat{C}_\lambda^{-1}\tilde{X}_i^T)^{-1}\tilde{X}_i\alpha^*_\lambda\\
&\quad = (I_{n_i} - \tilde{X}_i\hat{C}_\lambda^{-1}\tilde{X}_i^T)^{-1}(\tilde{Y}_i - \tilde{X}_i\alpha^*_\lambda).
\end{aligned}
$$

Directly from Lemma 2, we have the following theorem for calculating $WCV(\lambda)$:

THEOREM 1

$$
\begin{aligned}
WCV(\lambda) = \frac{1}{n}\sum_{i=1}^M (\tilde{Y}_i - \tilde{X}_i\alpha^*_\lambda)^T(I_{n_i} - \tilde{X}_i\hat{C}_\lambda^{-1}\tilde{X}_i^T)^{-1}&\\
\hat{\Sigma}_i^{1/2}\hat{W}_i^{-1}\hat{\Sigma}_i^{1/2}\,(I_{n_i} - \tilde{X}_i\hat{C}_\lambda^{-1}\tilde{X}_i^T)^{-1}(\tilde{Y}_i - \tilde{X}_i\alpha^*_\lambda).&
\end{aligned}
\tag{3.5}
$$

Note that when $W_i = \Sigma_i$, we can write the WCV as,

$$
WCV(\lambda) = \frac{1}{n}\sum_{i=1}^M (\tilde{Y}_i - \tilde{X}_i\alpha^*_\lambda)^T(I_{n_i} - \tilde{X}_i\hat{C}_\lambda^{-1}\tilde{X}_i^T)^{-2}(\tilde{Y}_i - \tilde{X}_i\alpha^*_\lambda). \tag{3.6}
$$

Calculationn of WCV(λ)

Equation (3.6) (or (3.5)) allows us to calculate $WCV(\lambda)$ for each λ by fitting the model once. But to find the best λ, that is, the minimizer of $WCV(\lambda)$, it is necessary to calculate the MPLE for many λ's. In what follows, we give a further approximate formula which allows us to approximate $WCV(\lambda)$ for all λ by only fitting the model without penalty.

The results in the *Property* (iii) suggest that the difference between α^*_λ and

$(B + 2\lambda\Omega)^{-1}B\hat{\alpha}$ is $o(n^{-1/2})$. Also, $B + 2\lambda\Omega$ can be estimated by $1/n\ \hat{C}_\lambda$ and B can be estimated by $1/n\ \hat{C}_0$, thus α_λ^* in (3.6) can be replaced by $\hat{C}_\lambda^{-1}\hat{C}_0\hat{\alpha}$. Now the approximate *WCV* criteria becomes

$$AWCV(\lambda) = \frac{1}{n}\sum_{i=1}^{M} (\tilde{Y}_i - \tilde{X}_i\hat{C}_\lambda^{-1}\hat{C}_0\hat{\alpha})^T(I_{n_i} - \tilde{X}_i\hat{C}_\lambda^{-1}\tilde{X}_i^T)^{-2}(\tilde{Y}_i - \tilde{X}_i\hat{C}_\lambda^{-1}\hat{C}_0\hat{\alpha}). \quad (3.7)$$

All the quantities in equation (3.7) can be obtained just from the MLE, and so the calculation of *AWCV* for all λ requires only one fit of the model. Thus choosing λ by minimizing $AWCV(\lambda)$ is computational feasible.

Using $W_i = \Sigma_i$, a Generalized cross-validation criterion, analogous to (3.3) can also be defined as

$$\begin{aligned} &AWGCV(\lambda)\\ &= \frac{1}{n}\sum_{i=1}^{M} (\tilde{Y}_i - \tilde{X}_i\alpha_\lambda^*)^T(\tilde{Y}_i - \tilde{X}_i\alpha_\lambda^*)\Big/\left(1 - \frac{1}{n}\,\text{trace}\ (\Sigma_{i=1}^M\ \tilde{X}_i\hat{C}_\lambda^{-1}\tilde{X}_i^T)\right)^2\\ &= \frac{1}{n}\sum_{i=1}^{M} (\tilde{Y}_i - \tilde{X}_i\hat{C}_\lambda^{-1}\hat{C}_0\hat{\alpha})^T(\tilde{Y}_i - \tilde{X}_i\hat{C}_\lambda^{-1}\hat{C}_0\hat{\alpha})\Big/\left(1 - \frac{1}{n}\,\text{trace}\ (\hat{C}_\lambda^{-1}\hat{C}_0)\right)^2. \quad (3.8)\end{aligned}$$

The formula for calculating *AWGCV* (3.8) is slightly simpler than that of *AWCV* (3.7), however the difference in computational time is insignificant compared to the time spent on the model fitting. We observed in our simulation that the difference between the choice of λ based on *AWCV* and *AWGCV* is very small, thus in this paper we focus on *AWCV*.

Rice and Silverman (1991) and Zeger and Diggle (1994) indicated that the expectation of their unweighted *CV* approximates the sum of two terms, the variance of the outcome variable and the MSE, where the mean square error is of the estimated curves to the "true" curves. Since the variance of the outcome is a constant, this argument leads to an interpretation of *CV*: to minimize the *CV* is, in an expectation sense, to minimize the MSE. The same interpretation can be modified to the weighted *CV* score defined in this paper.

4. SIMULATION STUDY AND AIDS APPLICATION

4.1 Data Generation Scheme

The generated data sets are designed to mimic the changes of a serological marker, Neopterin, to a drug in an AIDS clinical trial (Bass *et al.*, 1992). We design our data sets to have the following features: (1) each data set has a placebo group and a treatment group, and each group contains 30 subjects; (2) all subjects are followed from one time unit before treatment began until 5 time units after treatment began; (3) each subject has one pre-treatment measurement uniformly distributed over the one unit interval, and 4–8 post-treatment measurements with a tendency for more observations to be made at earlier times, that is, closer to the treatment initiation date.

The response $y_{it_{ij}}$ $\{j = 1, \ldots, n_i, i = 1, \ldots, M\}$ for person i at time t_{ij} is created as follows.

$$y_{it_{ij}} = \mu_g(t_{ij}) + e_i(t_{ij})$$

where the 'true' curves $\mu_g(t)$ are

$$\mu_1(t) = 15 + \log(t + 1), \quad \text{(The 'placebo group')}$$

$$\mu_2(t) = 15 + \log(t + 1) - [1 - \cos(\pi(t - 1)/4)]I(t > 1). \quad \text{(The 'treatment group')} \tag{4.1}$$

Let $R_i = (e_{it_{i1}}, \ldots, e_{it_{in_i}})^T$. The $(j, k)^{\text{th}}$ element of Σ_i, the variance-covariance matrix is $\sigma^2[I(j = k) + \gamma^2\rho^{-|t_{ij}-t_{ik}|}]$. We present the results for $\sigma^2 = 0.25$, $\gamma^2 = 1$ and $\rho = 0.9$, other choices of these parameters gave very similar results. As is frequently assumed in longitudinal studies, the 'true' model separates the random term e_{it} into two components: an independent measurement error term with mean zero and variance 0.25 and a random term with an $AR(1)$ correlated structure.

One hundred data sets are generated. The scatter-plots of a specific randomly chosen data set and "true" curves (4.1) are shown in Figure 2.

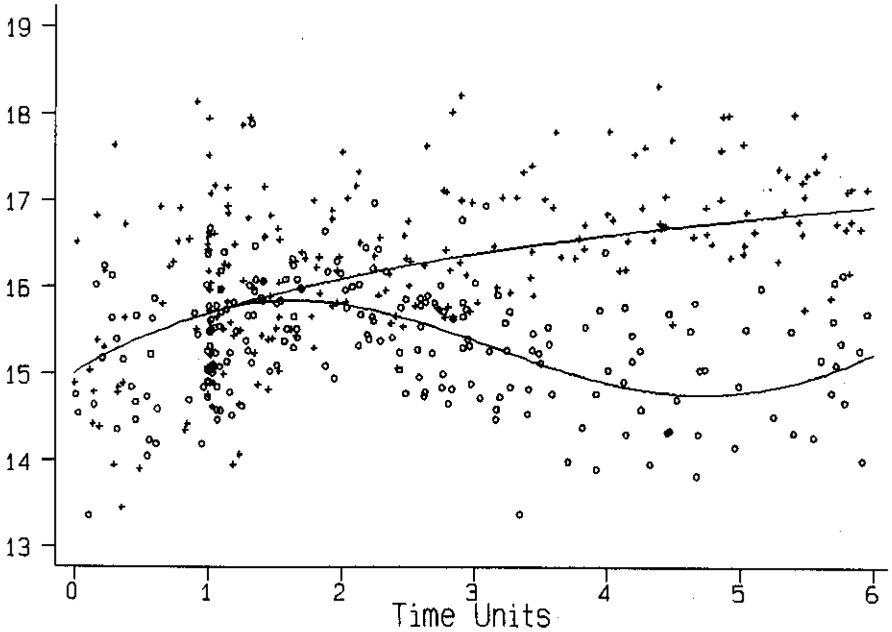


Figure 2 Scatter plot of one simulated data set and the "true" curves. The "+": placebo; the "o": treated; the upper curve: $\mu_1(t)$; the lower curve: $\mu_2(t)$.

4.2 Model Fitting Procedures

We use the Fisher-scoring algorithm (Chi and Reinsel 1989) to obtain the MLE and MPLE. This algorithm is an iterative procedure which calculates new parameter values $\alpha^{(j+1)}$, $\phi^{(j+1)}$ from current values $\alpha^{(j)}$, $\phi^{(j)}$ using

$$\begin{bmatrix}\alpha\\ \phi\end{bmatrix}^{(j+1)} = \begin{bmatrix}\alpha\\ \phi\end{bmatrix}^{(j)} - \begin{bmatrix} E\dfrac{\partial^2 L}{\partial\alpha\partial\alpha^T} & E\dfrac{\partial^2 L}{\partial\phi\partial\alpha^T}\\ E\dfrac{\partial^2 L}{\partial\alpha\partial\phi^T} & E\dfrac{\partial^2 L}{\partial\phi\partial\phi^T}\end{bmatrix}^{-1} \begin{bmatrix}\dfrac{\partial L}{\partial\alpha}\\ \dfrac{\partial L}{\partial\phi}\end{bmatrix}^{(j)}$$

Since $E(\partial^2 L/\partial\alpha\partial\phi^T) = 0$, we have in the log-likelihood ($\lambda = 0$) case,

$$\alpha^{(j+1)} = \left(\sum_{i=1}^{M} X_i^T \Sigma_i^{-1} X_i\right)^{-1} \left(\sum_{i=1}^{M} X_i^T \Sigma_i^{-1} Y_i\right)^{(j)} \tag{4.2}$$

$$\text{and } \phi^{(j+1)} = \phi^{(j)} - \Delta\phi^{(j)}, \quad \text{where } \Delta\phi^{(j)} = \left[E\frac{\partial^2 L}{\partial\phi\partial\phi^T}\right]^{-1}\left[\frac{\partial L}{\partial\phi}\right]^{(j)}$$

For the penalized log-likelihood ($\lambda > 0$) case, the only difference is that (4.2) is replaced by

$$\alpha^{(j+1)} = \left(\sum_{i=1}^{M} X_i^T \Sigma_i^{-1} X_i + 2\lambda n\Omega\right)^{-1} \left(\sum_{i=1}^{M} X_i^T \Sigma_i^{-1} Y_i\right)^{(j)}$$

SAS PROC IML was used for all computations.

The models to be fitted are assumed to have 13 knots which are chosen at spaces of half a time unit for the entire 6 time units period. For the covariance structure, we assumed that the $(j, k)^{\text{th}}$ element of Σ_i is $\sigma^2[I(j = k) + \gamma^2\rho^{-|t_{ij}-t_{ik}|}]$, where σ^2, γ^2 and ρ are the unknown parameters.

For each of the 100 data sets, we obtain the maximum likelihood estimates, then calculate λ_{CV} by minimizing *AWCV* (equation (3.7), and then using the λ_{CV} values, we obtain the MPLE.

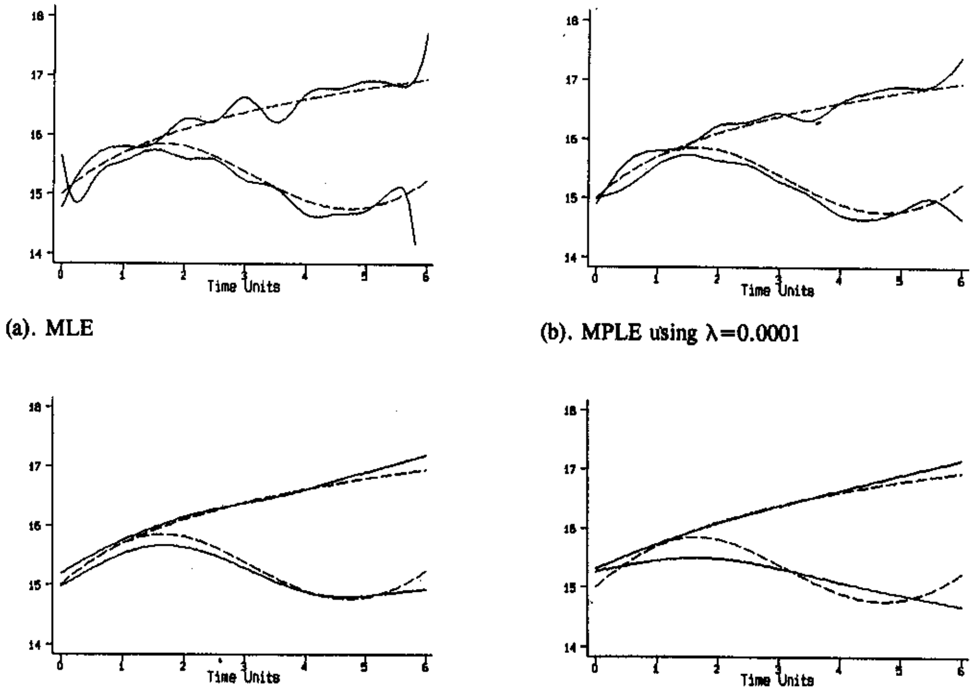
The Monte Carlo bias and variability of the estimates of $\mu_g(t)$ at specific time points are evaluated. An overall measure of the difference between the fitted curves and the true curves is calculated for each data set. This squared Integral Error is defined as SIE $= \Sigma_{g=1}^{2} \int_0^6 (\hat{\mu}_g(t) - \mu_g(t))^2\, dt$, where $\hat{\mu}_g(t)$ is the fitted curve.

In our calculation for SIE, instead of using $\mu_g(t)$ as defined in (4.1), we used a very close approximation to $\mu_g(t)$ which has a piecewise cubic polynomial form (2.2). The maximum difference of the $\mu_g(t)$ and this approximation is less than 0.01. Because of the variability of the observations, the difference between the true $\mu_g(t)$ and the approximate $\mu_g(t)$ is negligible compared to the difference between $\mu_g(t)$ and $\hat{\mu}_g(t)$.

4.3 *Results of the Simulation*

4.3.1 Illustration of the Results for One Data Set For one of the generated data sets, Figure 3(a)–(d) show the fitted curves for $\lambda = 0, 0.0001, \lambda_{CV}(=0.0094)$, and 0.1. In Figure 3(a) the estimated MLE curves show too much fluctuation around the "true" curves. In Figure 3(b) the estimated curves are under-smoothed because λ is too small. Figure 3(c) is the MPLE, when λ is chosen by Cross-validation. In Figure 3(d) the curves are the MPLE when λ is too big, notice that the curves are over-smoothed. The values of SIE corresponding to the 4 figures are (a) 1.749, (b) 0.223, (c) 0.143 and (d) 0.343 and the minimum value of SIE is 0.133 at $\lambda =$ 0.004. The closeness of the two SIE values at $\lambda = \lambda_{CV}$ and $\lambda = 0.004$ indicates that for this data set our approximate *AWCV* criteria gives rise to estimated curves which are almost optimal with respect to the SIE criteria.

4.3.2 Comparison of MPLE and MLE The MPLE always give smoother curves than the MLE since smoothing penalties are applied to the former. In what follows, we compare the MPLE and the MLE in two ways 1) pointwise and 2) overall.



(a). MLE

(b). MPLE using λ=0.0001

(c). MPLE using λ_{CV}=0.0094

(d). MPLE using λ=0.1

Figure 3 Fitted curves using different values of λ. The solid upper curves: the fitted curves of placebo; the solid lower curves: the fitted curves of treated; the dashed upper curves: the "true" curves of placebo; the dashed lower curves: the "true" curves of treated.

1). Three time points, $t = 0.25, 3, 5.75$ are chosen for the pointwise evaluations. Note that $t = 3$ is in the middle of the entire 6 time units range and is located on a knot while $t = 0.25$ and 5.75 are closer to the end points and are not located on knots.
    The box-plots of the values of the 100 fitted MLE and MPLE curves are presented in Figure 4(a), (b) and (c) for placebo and treatment group at $t = 0.25, 3$ and 5.75 respectively. It is easy to see that the 100 MLE values have larger variation than the 100 MPLE with $\lambda = \lambda_{CV}$ One point to note is that at some times, especially at $t = 5.75$ for the treatment group, the MLE is less biased than the MPLE.

2). The SIE is calculated as an index of the overall goodness of fit. The ratios of the SIE of the MPLE curves to the MLE curves are computed for all the 100 data sets. The percentiles of this ratio for the 100 data sets are: 0.03 (minimum), 0.08 (25th), 0.31 (median), 0.63 (75th) and 0.69 (maximum) respectively. These results show that although the MPLE is slightly biased it gives a substantial better overall fit to the whole curve than the MLE.

4.4 Clinical Trial Data

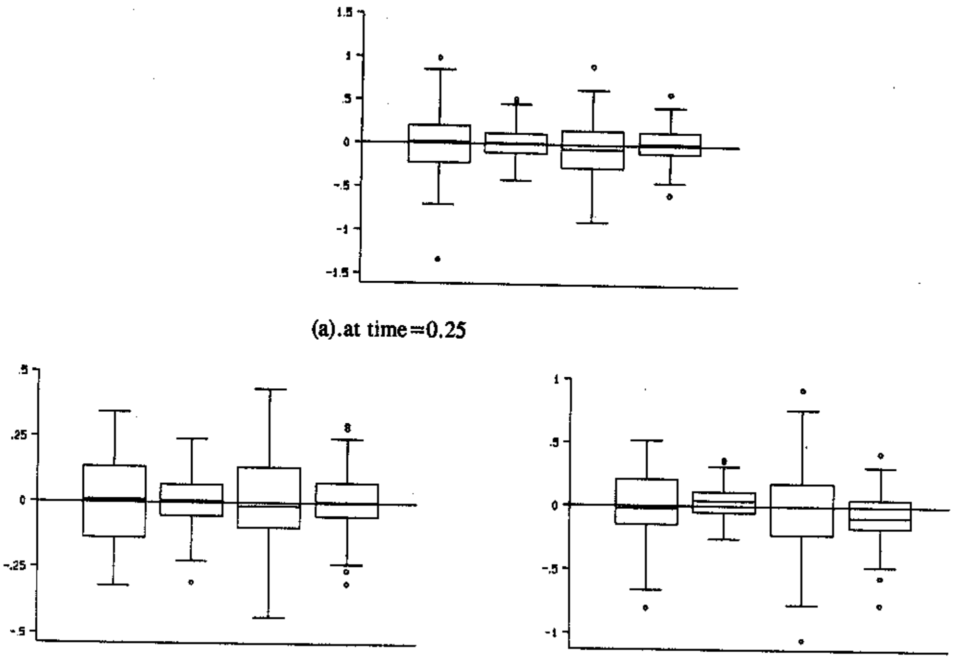
The method described above was applied to the Neopterin data given in Section 1. The covariance structure of the observations is given by a continuous time version of an $AR(1)$ process, that is, $Cov(e_{it_{ij}}, e_{it_{ik}}) = \sigma^2[I(j = k) + (\gamma_0 + \gamma_1 t_{ij})^2(\gamma_0 + \gamma_1 t_{ik})^2 \rho^{-|t_{ij}-t_{ik}|}]$, where σ^2, γ_0, γ_1 and ρ are the unknown parameters. Knots are chosen at 4 weeks intervals from $t = 0$ to week 56, that is, 15 knots total. To ensure that the curves pass through the origin, the intercept term is removed from the piecewise cubic polynomials forms we assumed for the curves. The fitted curves are shown in Figure 5. The curves indicate that there is an almost immediate reduction in Neopterin due to AZT, however there appears to be no further reduction beyond about 8 weeks.

5. DISCUSSION

One important issue in the analysis of longitudinal data is the choice of the within-subject correlation structure. Besides the $AR(1)$ covariance structure in the numerical examples above, we also considered a covariance structure derived from a random-effects model (Laird and Ware, 1982). For the simulated data sets, we observed that the choice of λ_{CV} and the estimated curves were very similar irrespective of which of these two covariance structures was assumed in fitting the model. We believe, that with respect to smoothing the mean function, that reasonable choices for the within-subject covariance structure will not change the smoothing procedure significantly.

Another point to note is that fixed or time varying covariates could easily be incorporated as linear terms in the model.

This approach has some open questions, such as how many knots to use, how



(a).at time=0.25

(b). at time=3.00

(c).at time=5.75

Figure 4 Box plots for the fitted values minus the "true" values. In each of (a), (b), and (c), the boxes from left to right: MLE for placebo; MPLE for placebo; MLE for treated; MPLE for treated.

to choose the position of the knots, whether the knots should be equally spaced, and whether the position and number of knots should depend in any way on the data. These questions may not have simple answers. Others (Gray 1992, Hastie and Tibshirani 1990) have suggested using between 8 and 20 knots, either equally spaced (Parker and Rice 1985) or with a roughly equal number of data points between each knots (Gray 1992). Our experience is that the exact number and position of the knots is not crucial provided the number is not small and a penalized likelihood approach is used.

Our approach can be extended in a number of ways. For example, a different smoothing parameter could be used for each group; an approximation similar to (3.7) would make this computationally feasible. Modifications to the model to reduce the endpoint effects common in smoothing problems are possible. One approach would be to assume that $\mu_g(t)$ is linear, rather than cubic, between T_0 and T_1 and between T_{K-1} and T_K.

In principle, statistical inference including hypothesis testing and confidence intervals for the parameters can be performed using the asymptotic distributions of MPLE in Section 2.3. One alternative for constructing confidence intervals for α is to use the Bayesian interpretation (Silverman 1985) of the penalized likelihood function. Our simulation work with confidence intervals suggests that the empirical

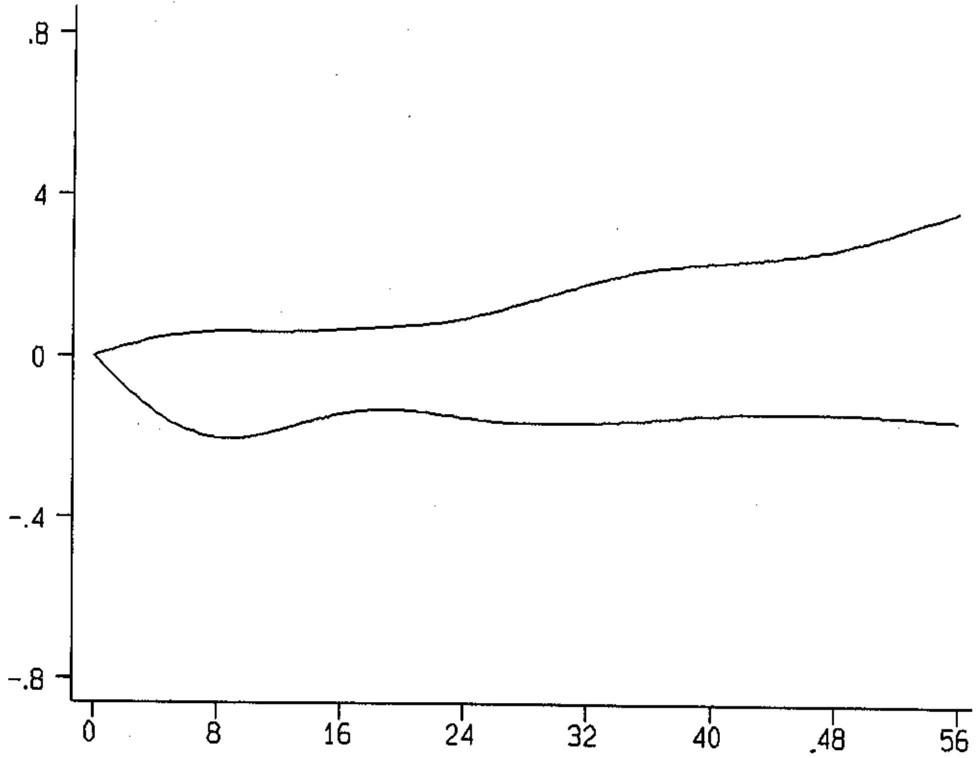


Figure 5 The MPLE of the Log-Neopterin ratio-to-baseline using $\lambda_{CV} = 0.36$. The upper curve: placebo; the lower curve: AZT treated.

coverage rates of intervals derived using the Bayesian interpretation are slightly better than these constructed from the asymptotic properties in Section 2.3, and furthermore that the coverage rates are acceptably close to the nominal rate. Further description of this is given elsewhere (Wang and Taylor 1994).

Acknowledgements

This research was partially supported by NIH Grants AI29196. The computer time was supported by IBM/UCLA Joint Project on Supercomputing.

References

Agarwal, G. and Studden, W., (1980). Asymptotic Integrated Mean Square Error Using Least Square and Bias Minimizing Splines, *The Annals of Statistics*, **8**, 1307–1325.

Bass, H. Z., Hardy, W. D., Mitsuyasu, R. T., Taylor, J. M. G., Wang, Y., Fischl, M. A., Spector, S., Richman, D., and Fahey, J. L. (1992). The effect of Zidovudine Treatment on serum neopterin and β_2-Microglobulin levels in mildly symptomatic, HIV type 1 seropositive individuals. *J. Acq. Immun. Def. Synd.* **5**, 215–221.

Chi, E. M. and Reinsel, G. C. (1989). Models for longitudinal data with random effects and AR(1) errors. *Journal of the American Statistical Association,* **84,** 452–459.

Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions. *Numerische Mathematik,* **31,** 377–403.

DeGruttola, V. and Tu, X. M. (1992). Modelling the relationship between progression of CD4-lymphoeyte count and survival time. AIDS Epidemiology: Methodological Issues: Editors Jewell, N. P., Dietz, K., and Farewell, V. T., Birkhauser, Boston. 1992. 275–296.

Diggle, P. J. (1988). An approach to the analysis of repeated measurements. *Biometrics,* **44,** 959–971.

Friedman, J. and Silverman, B. (1989). Flexible parsimonious smoothing and additive modeling, *Technometrics,* **3,** 3–21.

Gray, R. J. (1992). Flexible Methods for Analyzing Survival Data Using Splines, With Applications to Breast Cancer Prognosis. *Journal of the American Statistical Association,* **87,** 942–951.

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

Hastie, T. J. and Tibshirani, R. J. (1990). Generalized additive models. New York: Chapman & Hall.

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

Muller, H. G. (1988). Nonparametric Regression Analysis of Longitudinal Data. Berlin: Sringer-Verlag.

Parker, R. L. and Rice, J. A. (1985). Discusson of "Some aspects of the spline smoothing approach to nonparametric regression curve fitting," by Silverman, B. W. *J. R. Statist. Soc. B.,* **47,** 40–42.

Rice, J. A. and Silverman, B. W. (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *J. R. Statist. Soc. B.,* **53,** 233–243.

Silverman, B. W. (1984). Spline Smoothing: the equivalent variable kernel method. *Ann. Statist.* **12,** 898–916.

Silverman, B. W. (1985). Some aspects of the spline smoothing approach to nonparametric regression curve fitting (with discussion). *J. R. Statist. Soc. B.,* **47,** 1–52.

Speckman, P. E. (1988). Regression analysis for partially linear models. *J. R. Statist. Soc. B.* **50,** 413–436.

Wahba, G. (1990). Spline models for observational data. Philadelphia: *Society for Industrial Applied Mathematics.*

Wang, Y. (1991). Structured covariance models for longitudinal data with smoothing techniques. Unpublished Ph.D. dissertation, University of California-Los Angeles, Dept. of Biostatistics.

Wang, Y. and Taylor, J. M. G. (1994). Inference for smooth curves in longitudinal data with application to an AIDS clinical trial. *Statistics in Medicine,* to appear.

Wypij, D., Pugh, M., and Ware, J. H. (1993). Modelling pulmonary function growth with regression splines. *Statistica Sinica,* **3,** 329–350.

Zeger, S. L. and Diggle, P. J. (1994). Semiparametric models for longitudinal data with application to CD4 cell numbers in HIV seroconverters. *Biometrics,* **50,** 689–699.