BIOMETRICS 55, 452–459
June 1999

Analysis of Infant Growth Curves Using Multivariate Adaptive Splines

Heping Zhang

Department of Epidemiology and Public Health, Yale University School of Medicine,
New Haven, Connecticut 06520, U.S.A.
email: heping.zhang@yale.edu

SUMMARY. In this paper, we study the effect of cocaine use by a pregnant woman on the growth of her infant after birth. Using a data set from a retrospective study, we found that cocaine use was a marginally significant contributor to the infant growth as measured by bodyweight. From a statistical point of view, the data represent a common, though complex, structure that has received little attention in the statistical literature. To analyze these data, we adopt and further enhance an approach developed recently called MASAL (multivariate adaptive splines for the analysis of longitudinal data). In addition to the fitting of growth curves, we demonstrate particularly how to explore and estimate the underlying covariance structures for the longitudinal data that were collected from a rather irregular schedule.

KEY WORDS: Adaptive splines; Longitudinal data; Model selection; Repeated measures; Time-dependent covariates.

1. Introduction

We analyze a data set of infant growth from a retrospective study conducted by Dr. John Leventhal and his colleagues at Yale University School of Medicine, New Haven, Connecticut. Their primary aim was to study the risk factors during pregnancy that may lead to the maltreatment of the infants after birth such as physical and sexual abuse. The investigators recruited 298 children born at Yale–New Haven Hospital after reviewing the medical records for all women who had deliveries from September 1, 1989, through September 30, 1990. Detailed eligibility criteria have been reported previously elsewhere, such as in Wasserman and Leventhal (1993) and Stier et al. (1993). The main factor of the sample selection is the ascertainment of cocaine exposure. Two groups of infants were included: those whose mothers were regular cocaine users and those whose mothers were clearly not cocaine users. The group membership was classified from the infants' logs of toxicology screens and their mothers' obstetric records. In addition, efforts have been made to match the unexposed newborns with the exposed ones for date of birth, medical insurance, and the mother's parity, age, and timing of the first prenatal visit.

Our analysis includes mother's cocaine use, infant's gender, gestational age, and race (either white or black) as covariates. After birth, the infants were brought back to see their pediatricians and growth measurements were taken at each visit. We will focus on weight (in kilograms) as the response variable. Analogous analyses may be conducted for other growth endpoints such as height. To offer a brief view of the data structure, Figure 1 shows the growth curves for 30 randomly chosen children. The weights at two adjacent visits are connected by a line. The irregularity feature of these data is worth mentioning: Different children had different numbers and patterns of visits during the study period.

Multivariate adaptive splines for analysis of longitudinal data (MASAL) in Zhang (1997) is adopted in our analysis. MASAL is built upon existing works on recursive partitioning, including Breiman et al. (1984), Friedman and Silverman (1989), Friedman (1991), and Zhang (1994). This approach advances its precedents in that it handles dependent observations. Dependence creates additional methodological and computational challenges due to the adaptivity and nonlinearity of the embedded spline bases. The emphasis of this work is to show how to model the covariance structure in MASAL when longitudinal data are collected on an irregular schedule. The distinction between this work and that of Zhang (1997) lies in the present emphasis on modeling the correlation structure when the numbers and patterns of visits are irregular.

The analysis of longitudinal data is abundant in the literature, but the area is dominated by the use of parametric mixed-effects models. Further, the analyzed longitudinal data are generally collected on a relatively regular schedule. In-depth discussions of this topic have been covered by an excellent book of Diggle, Liang, and Zeger (1994) and the seminal work of Laird and Ware (1982), among others. Here, we use a method that does not require a parametric specification of the regression function. Although parametric mixed-effects models are very useful for analyzing growth curves (cf., Carter et al., 1992), the nonparametric method used here provides a fruitful complement. The term nonparametric is with regard to the mean function and, as we see later, the method can be viewed as a hybrid of nonparametric and parametric

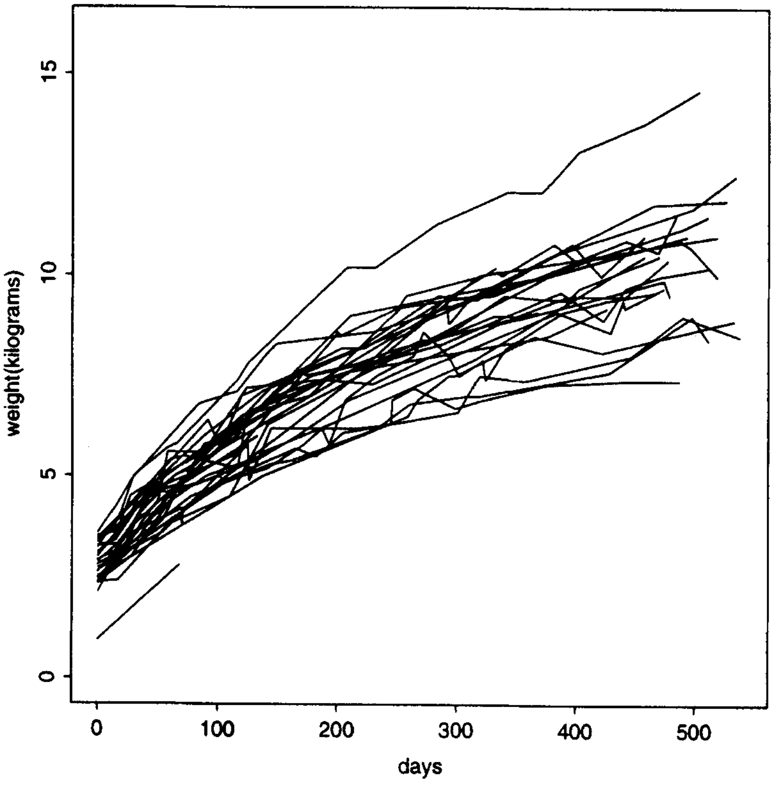

Figure 1. Connected curves for the weights of 30 randomly chosen children.

modeling. In ordinary regression with independent observations, nonparametric methods offer convenient and flexible techniques for model selection and data exploration. This is even more true for modeling growth curve data because the correlation structure within the same subject makes it less intuitive for us to specify and justify a parametric model. Thus, it is useful to utilize an automated tool that suggests, if not determines, a promising model. Not only does it have an automated feature, but the method employed here also handles messy longitudinal data that have hardly been dealt with in the literature.

We outline MASAL in Section 2 and refer to Zhang (1997) for the details. The analysis of infant growth curves appears in Section 3. We explain how to build a growth curve model using adaptive splines. In the last section, we discuss the strengths and limitations of the adopted method.

2. MASAL

In Table 1, we introduce some notation for the infant growth data displayed in Figure 1. When $x_{k,ij}$ changes with d_{ij}, then it is a time-dependent covariate. Hence, d_{ij} itself is a time-dependent covariate, and we include it as one of the p covariates.

To establish the relationship between response Y and the p covariates, we assume

$$y_{ij} = f(x_{1,ij}, \dots, x_{p,ij}) + e_{ij}, \qquad (1)$$

where f is an unknown smooth function and e_{ij} is the error term with mean zero but an unknown distribution, $j = 1, \dots, T_i, i = 1, \dots, n$.

Adaptive splines assume that the mean function f in (1) is a member of the following class of functions:

$$\left\{ f : f(\mathbf{x}) = \sum_{k=0}^{M} \beta_k B_k(\mathbf{x}), M = 0, 1, \dots \right\}, \qquad (2)$$

where B_k is a special basis function of the p covariates $\mathbf{x} = (x_1, \dots, x_p)'$ and β_k is the regression coefficient ($k = 0, 1, \dots, M$).

MASAL derives $B_k(\mathbf{x})$ from the following two functions:

$$(x_k - t)^+ \quad \text{and} \quad x_k, \qquad k = 1, \dots, p, \qquad (3)$$

where, for any number a, $a^+ = \max(a, 0)$. Specifically, $B_k(\mathbf{x})$ is either one of the functions in (3) or their product. The truncated linear function in (3) consists of two line segments connected by the knot t.

Table 1
Notation

Notation	Definition
n	Number of study subjects
T_i	Number of visits for the ith child
d_{ij}	Age of the ith child at the jth visit
$x_{k,ij}$	Value of the kth covariate for the ith child at the jth visit
y_{ij}	The response variable for the ith child at the jth visit

Note: $j = 1, \dots, T_i$, $i = 1, \dots, n$, $k = 1, \dots, p$.

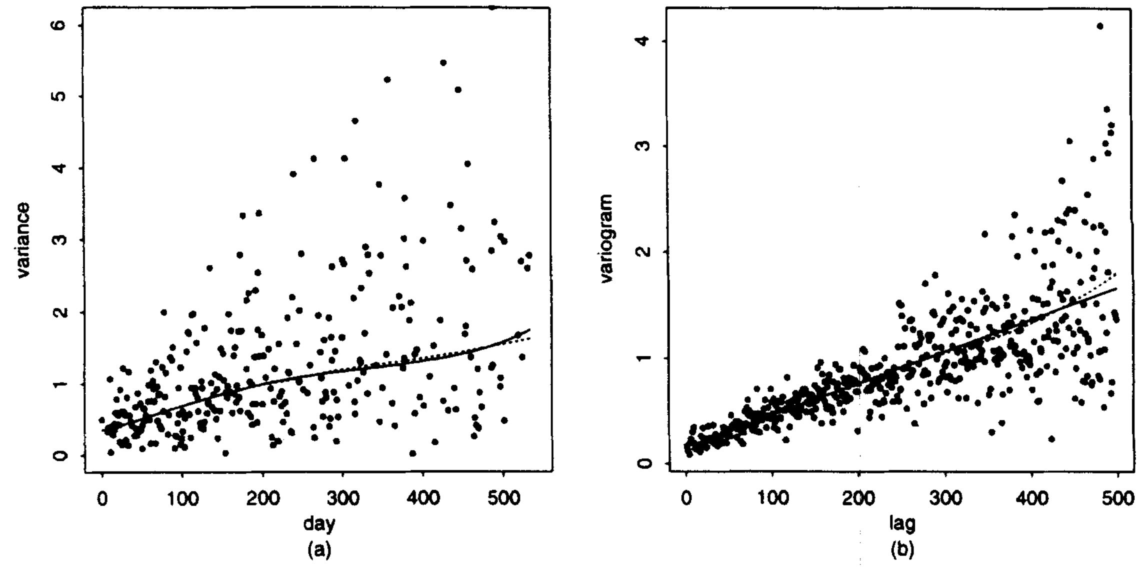

Figure 2. Variance (left) and variogram (right). The dots are the sample variance and sample variogram estimates. Panel **a** is based on those days on which there are at least five observed weights. The solid curve is the fitted variance function: $\exp(-1.078 + 0.0091d - 2.32d^2/10^5 + 2.23/10^8)$. In panel **b**, the solid line is the fitted variogram of the standardized initial residuals: $0.156 + 0.003l$. Also presented are the `loess` smooth fits (dotted curves).

For a given set of observations, MASAL selects a model from the class (2) using a forward step and then a backward step. In the forward step, terms are added to minimize the (weighted) sum of squared residuals: $WLS = \Sigma_{i=1}^{n} (\mathbf{y}_i - \hat{\mathbf{y}}_i)'\mathbf{W}_i^{-1}(\mathbf{y}_i - \hat{\mathbf{y}}_i)$, where $\mathbf{y}_i = (y_{i1}, \ldots, y_{iT_i})'$, $\hat{\mathbf{y}}_i$ contains the fitted values of $\mathbf{y}_i$, and $\mathbf{W}_i$ is the within-subject covariance matrix for $\mathbf{e}_i$, $i = 1, \ldots, n$.

After the forward step, all knots are found and each corresponding basis function will be treated as if it is a given predictor. In the backward step, we delete one least significant term from the model at a time, resulting in a reduced model. We select the final model that yields the smallest $GCV = WLS_k/[1 - (dk + 1)/\Sigma_{i=1}^{n} T_i]^2$, where WLS_k is the WLS of a reduced model with k terms and d is the penalizing parameter for model complexity, which is usually chosen between three and five. These choices of d are based on the discussion in Friedman (1991).

Since the covariance structure is usually unknown in practice, the MASAL algorithm needs to be used in an iterative manner in order to estimate the mean and the covariance structure alternately. We refer to the technical discussions of Friedman (1991), Altman (1992), and Zhang (1997). In the next section, both the forward and backward steps will be explained with specific examples.

3. Application

In this section, we analyze the growth curves of the 298 children. The first step is to specify an initial within-subject covariance $\mathbf{W}_i$ for every subject.

3.1 *Characterization of Covariance Structure*

By examining Figure 1, we see that the variance of weights appears to increase as a child grows. To explore the time trend for the variance further, we take three different approaches. First, we estimate the sample variances of weights on those days on which there are at least five observed weights (see Figure 2a). Second, on any particular day we estimate the variance using weights within a 1-week period centered on the given day. Finally, on day d between 1 and 540 days, we collect a number of cross-sectional body weights, $\mathbf{z}_d$, from all children whose last visit was after d days. If a child visited the doctor on day d, the child's weight is included in $\mathbf{z}_d$. However, if a child visited the doctor before and after, but not on day d, we include the interpolated value between the two adjacent weights in the shortest time interval containing day d. This can be viewed as cutting the growth curves displayed in Figure 1 vertically on every day and then collecting all intersected points into $\mathbf{z}_d$. We can use the sample variance of $\mathbf{z}_d$ as an estimate of daily variance of weights. By construction, we expect that these three approaches lead to different sample variance estimates. Although the variability of the estimates from these three different approaches is very different, their mean trend is remarkably similar and can be adequately described by

$$V(d) = \exp\left(\nu_0 + \nu_1 d + \nu_2 d^2 + \nu_3 d^3\right), \quad d = \text{day } 1, \ldots, \text{day } 540, \qquad (4)$$

where the ν coefficients will be estimated by the maximum likelihood approach in Section 3.2.1. In Figure 2a, the trend in (5) is very close to a `loess` fit computed by S-PLUS.

To determine $\mathbf{W}_i$, we still need to gauge the autocorrelation of weights between any two days. The time schedule is so irregular that it is difficult to examine the autocorrelation directly. As an alternative approach, we explore the variogram (Diggle et al., 1994, p. 50–51). For a stationary stochastic process $Z(t)$, the variogram is defined as

$$\gamma(l) = \frac{1}{2}\mathrm{E}\{Z(t) - Z(t-l)\}^2, \qquad l \geq 0,$$

where l is referred to as lag. If $Z(t)$ is stationary with variance τ^2, the autocorrelation ρ is a simple transformation of the variogram as follows:

$$\rho(l) = 1 - \gamma(l)/\tau^2. \tag{5}$$

To obtain the sample variogram as a function of lag, we proceed in three steps following a procedure described by Diggle et al. (1994, p. 51). First, we subtract each observed response y_{ij} on day d_{ij} from the average over a 1-week period to derive an initial residual r_{ij}, namely

$$r_{ij} = y_{ij} - \frac{\sum_{\{(k,l):|d_{kl}-d_{ij}|\leq 3\}} y_{kl}}{\#\{(k,l) : |d_{kl} - d_{ij}| \leq 3\}}, \tag{6}$$

where the denominator is the cardinal number of the set, $j = 1, \ldots, T_i$ and $i = 1, \ldots, n$. Second, we have seen that the variance is not constant over time. We standardize the residual r_{ij} by dividing it by the standard deviation of the weights on day d_{ij} according to (5). Let s_{ij} denote the standardized residual. Then the sample variogram of s_{ij} is calculated from pairs of half-squared residuals,

$$v_{ijk} = \frac{1}{2}(s_{ij} - s_{ik})^2,$$

with a lag of $l_{ijk} = d_{ij} - d_{ik}$. At each value of lag l, the sample variogram $\hat{\gamma}(l)$ is defined as the average of all v_{ijk}'s for which $d_{ij} - d_{ik} = l$. Figure 2b shows the variogram as a function of the lag between two days. A straight line can adequately describe the trend of the variogram. In light of (5) (where τ^2 is one due to the standardization), we use the linear autocorrelation

$$\rho(l) = \phi_0 + \phi_1 l, \qquad l = \text{lag } 1, \ldots, \text{lag } 539. \tag{7}$$

Like the ν coefficients in (5), ϕ_0 and ϕ_1 will also be estimated by the maximum likelihood approach in Section 3.2.1. Using (5) and (7), we can obtain the within-subject covariance matrix $\mathbf{W}_i$'s. For example, one child had visits on days 16, 52, 121, 275, 324, and 491. So the diagonal of $\mathbf{W}_i$ is $V(16), V(52), V(121)$, $V(275)$, $V(324)$, and $V(491)$. The covariance between the first two visits is

$$\sqrt{V(16)V(52)}\rho(52 - 16) = \sqrt{V(16)V(52)}\rho(36)$$

and the other covariances for this child can be derived analogously.

3.2 *Iterative Estimation for Model Parameters*

3.2.1. *The initial step.* To use MASAL, a covariance structure must be supplied. We assume that the initial residuals r_{ij} in (6) follow a normal distribution with mean zero and that their covariance matrix is a block diagonal consisting of $\mathbf{W}_i$, $i = 1, \ldots, n$. Each $\mathbf{W}_i$ is a covariance matrix of $\mathbf{r}_i = (r_{i1}, \ldots, r_{iT_i})$ determined through (5) and (7). Then we estimate the covariance parameters, ν and ϕ, embedded in $\mathbf{W}_i$ by maximizing the log likelihood

$$-\frac{1}{2}\sum_{i=1}^{n}\{T_i \log(2\pi) + \log(|\mathbf{W}_i|) + \mathbf{r}_i'\mathbf{W}_i^{-1}\mathbf{r}_i\}. \tag{8}$$

The maximum likelihood estimates for the variance and autocorrelation are, respectively,

$$V(d) = \exp(-0.53 + 0.0064d - 1.9d^2/10^5 + 2.1d^3/10^8),$$
$$\rho(l) = 0.929 - 0.0013l,$$

where d is the day of a visit and l is the lag between two visits. These are then used to form the covariance structure in the MASAL modeling.

Since we are particularly interested in the effect of cocaine use by the pregnant woman, we forced the corresponding variable c into the model. Thus, MASAL started with a model $\beta_0 + \beta_1 c$. Then $\beta_2 d + \beta_3(d - 127)^+$ was added to the model by the forward algorithm, giving a MASAL model

$$\beta_0 + \beta_1 c + \beta_2 d + \beta_3(d - 127)^+.$$

Next, $\beta_4(g_a - 28.6)^+$ is added to the model above, followed by $\beta_5 d(g_a - 28.6)^+$, and so on. Here g_a stands for the gestational age. The entered terms are either functions in (3) or their products. We stopped the forward step after collecting a total of 21 terms. Based on our experience, a maximum of 20 to 30 terms works well for most applications. Next, we used the generalized cross validation (GCV) procedure to eliminate one least significant term at a time from the 21 terms. After removing 11 terms, we obtained the first MASAL model that has the smallest GCV among the 20 nested, reduced models as

$$\begin{aligned} &0.828 + 0.028d - 0.0094(d - 120)^+ - 0.0054(d - 200)^+ \\ &\quad + (g_a - 28)^+\{0.204 + 0.0005d - 0.0007(d - 60)^+ \\ &\qquad - 0.0009(d - 490)^+\} \\ &\quad + s\{-0.0026d + 0.0022(d - 120)^+\}. \end{aligned} \tag{9}$$

To reach the final MASAL model, we need to adopt an EM-like iterative procedure by alternately estimating the covariance and the mean. If the MASAL model were a linear model, our iterative algorithm would be a special case of the EM algorithm (Laird and Ware, 1982). (See Zhang [1997] for details.)

3.2.2. *Fitted model and its interpretation.* The iterations converged at the second iteration. There is evidence suggesting that the second model is usually similar to the subsequent MASAL models. In some cases, it may take a few more iterations. Attempting four iterations seems to be a safe strategy because we have two extra iterations to evaluate the second model. The difficult part of using MASAL is to specify a reasonable covariance structure and then estimate it, which is the same situation with the use of parametric models. After this decision is made, it is trivial to run through two more iterations.

Based on the second iteration, we selected the following MASAL model:

$$\begin{aligned} \hat{f}(\mathbf{x}) &= 0.744 + 0.029d - 0.0092(d - 120)^+ - 0.0059(d - 200)^+ \\ &\quad + (g_a - 28)^+\{0.204 + 0.0005d - 0.0007(d - 60)^+ \\ &\qquad - 0.0009(d - 490)^+\} \\ &\quad + s\{-0.0026d + 0.0022(d - 120)^+\}, \end{aligned} \tag{10}$$

where d and g_a are the age of visit and gestational age, respectively, and s is the indicator for gender, with one for girls and zero for boys.

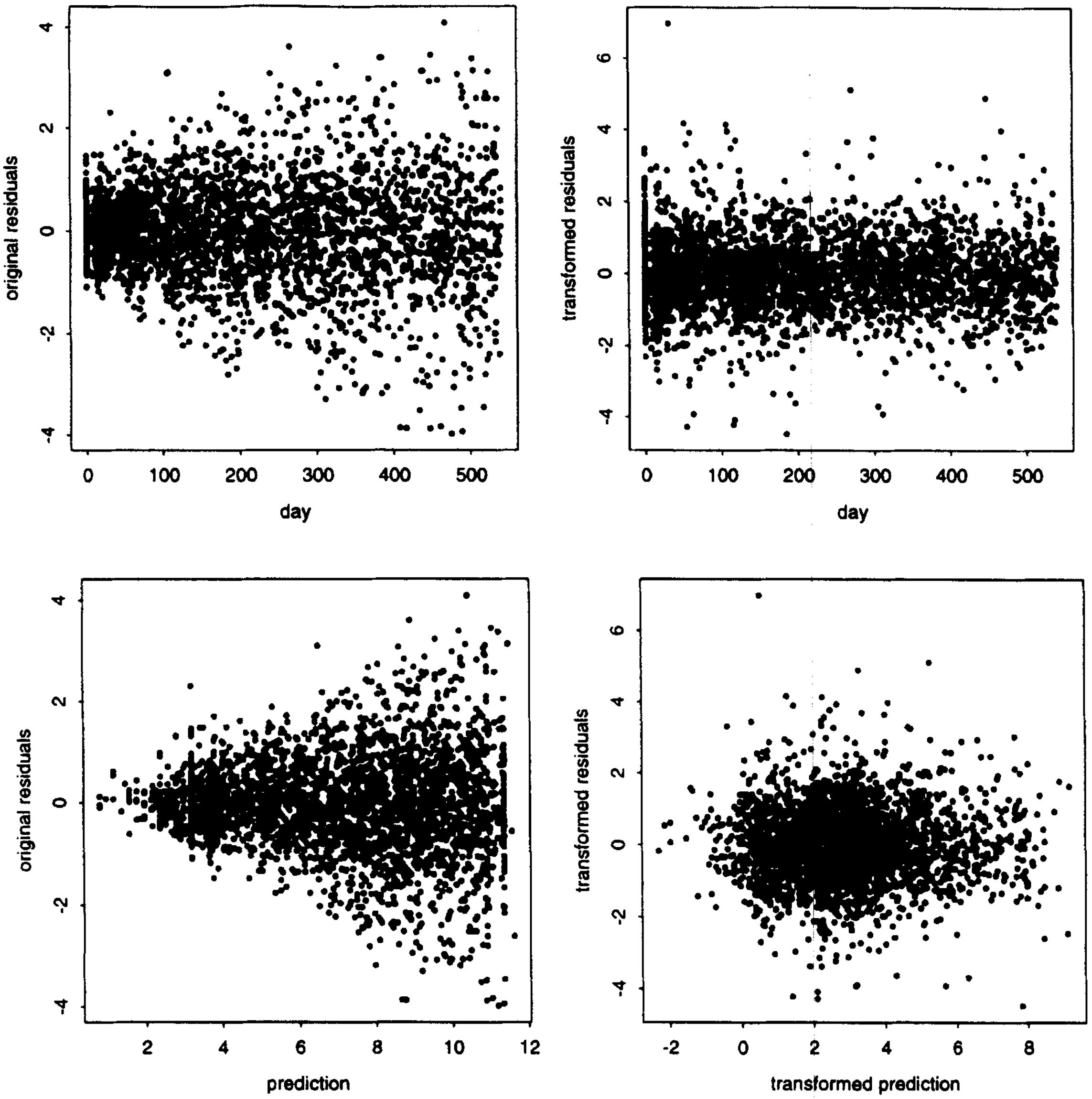

Figure 3. Residual plots against days (top) and predictions (bottom). The left panels are in the original scale and the right ones in a standardized scale.

How well does the MASAL model fit the data? We address this question graphically. In Figure 3, we plot the residuals against age and the predicted values, respectively. In the original scale, as we expected, the variability is greater at an older age or for a larger predicted value (see the two left panels). After transforming the residuals through the covariance matrix, the variability along with time is quite even. In addition, no apparent structure emerges when the transformed residuals are plotted against the transformed prediction. Thus, these residual plots support the selected MASAL model and the covariance structure (5) and (7). To further evaluate the MASAL model, we plot the fitted curves together with the observations at gestational ages of 36 and 40 weeks and for boys and girls, respectively, in Figure 4. We chose 36 and 40 weeks because a 40-week delivery is a full-term pregnancy and 36 weeks is 1 week short of a term delivery. It is clear that the fitted curves reside well in the midst of the observations, although there remain unexplained variations. Therefore, it is evident from Figures 3 and 4 that the selected MASAL model is adequate and useful.

From model (10), the velocity of growth reduces as a child grows. Beyond this common-sense knowledge, model (10) defines several interesting phases in which the velocity varies. Note that the knots for age are 60, 120, 200, and 490 days, which are about 2, 4, 8, and 16 months. In other words, not only does the velocity decrease, but also the length it lasts doubles in time. Girls grow more slowly soon after birth but start to catch up after 4 months. Gestational age affects birthweight, as immense evidence has shown. It also influences the growth dynamics. In particular, a more mature newborn tends to grow up faster at first but later experiences a slower growth as opposed to a less mature newborn. None of these

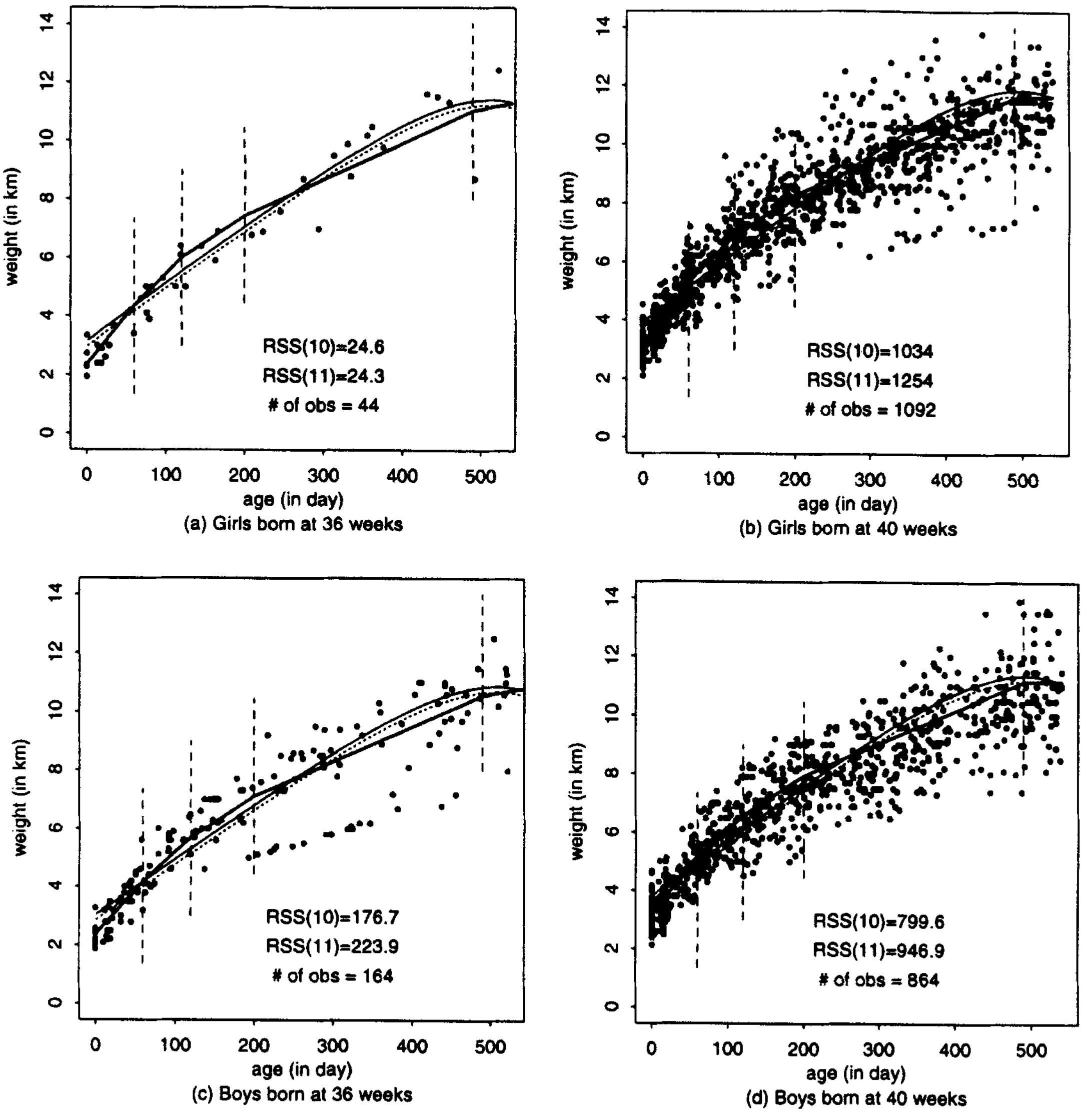

Figure 4. Observations and predictions for boys and girls born at 36 and 40 weeks. The thicker curves are from the MASAL model (10) and the vertical lines indicate the knot locations. Model (11) is drawn in thinner curves separately for the cocaine-used group (solid) and the not-used group (dashed). Along with the number of observations, the unweighted residual sum of squares (RSS) is given, respectively, for models (10) and (11) inside each panel.

implications are unexpected, while the importance of model (10) resides in its explicit mathematical characterization of the growth pattern without imposing any prior knowledge of it.

3.2.3. *Hypothesis testing.* As we mentioned earlier, we are particularly interested in the effect of cocaine use by a pregnant woman on her child's growth. This variable, denoted by c previously, did not stand out in the MASAL model. This is clearly an indication of the limited impact of this factor. We should also realize that our model building and variable selection procedures are not the same as the traditional ones. Could cocaine use contribute significantly to infant growth at a widely adopted p-value of 0.05? To answer this question, we start with model (10) and hold all of its 10 terms as fixed. Then we examine the contribution of c by including c as a main effect or an interaction term with one of the existing terms in addition to all terms already in model (10). Table 2 presents the significance of these individual terms, where the p-values are based on a two-sided t-test. Given the number of tests that were undertaken, two terms involving the interactions between cocaine use and gestational age may be worth pursuing. Overall, our data do not support the hypothesis that cocaine use by a pregnant woman influences her infant's growth significantly.

3.3 *Analysis Using Mixed Models*

As a comparison, we analyzed this data set with PROC MIXED in SAS. The initial model includes eight main-effect terms (black, the number of previous pregnancies, mother's age at delivery, mother's cocaine use, gestational age, child's gender,

Table 2
Impact of cocaine use on infant growth

Added term	Coefficient	t-Statistic	p-Value
c	0.162	2.42	0.016
cd	0.0004	1.37	0.17
$c(d-120)^+$	0.0003	0.96	0.34
$c(g_a-28)^+$	0.0166	2.69	0.007
$cd(g_a-28)^+$	0.00003	1.37	0.17
$c(d-60)^+(g_a-28)^+$	0.00005	1.68	0.093
$c(d-490+)^+(g_a-28)^+$	0.0001	0.35	0.73
csd	0.0002	1.10	0.27
$cs(d-120)^+$	0.0004	0.71	0.48
$c(d-200)^+$	0.0002	0.52	0.60

child's age at a visit, and an exponential transformation of the child's age) and four interaction terms (gender and gestational age by age and the transformation of age). The covariance is set to have an AR(1) structure. The least significant terms are removed from the model one at a time until all terms have p-values below 0.05 with exceptions clarified below. The selected model is

$$\begin{aligned} &-4.0+0.18c+0.194g_a-0.115s+0.02d-0.0008sd \\ &\quad +0.006\exp(d/100)-0.0005g_a\exp(d/100), \qquad (11) \end{aligned}$$

where the p-values for c (mother's cocaine use), s (child's gender), and $\exp(d/100)$ are, respectively, 0.03, 0.2, and 0.2. I rescaled d by 1/100 in the exponential term as a convenient way to avoid numerical difficulty; however, improvement may be possible if the scale parameter is estimated from the data. In that case, we have to use nonlinear mixed models. The remaining terms have p-values below 0.001. The terms s and $\exp(d/100)$ are included because of their corresponding significant interaction terms.

Broadly, model (11) gives rise to conclusions similar to those stated above. In Figure 4, we plot models (10) and (11) together. The general trends are consistent, but there are discrepancies. As Figure 4 shows, the unweighted residual sum of squares due to model (11) is greater than that due to model (10). In any case, the general consistency does not imply that a thorough scrutiny of the mean and covariance structures is not necessary because otherwise we cannot reveal the consistency.

A few additional remarks are warranted. First, for this particular data set, we do not have too many predictors. It may not be so time consuming to try various interactions and transformations of the predictors to improve the fit. This model-mining step is similar to what is implemented inside MASAL. The difference is who is responsible for this tedious task. When the number of predictors is large, this task could be overwhelming. Second, there are studies for which researchers may have less experience with regard to the growth pattern than people do with the growth of weight. The transformation like $\exp(d/100)$ in model (11) is not always clear. However, when the underlying growth mechanism is understood, a parameter model might have a more concise expression than a MASAL model. The MASAL model has an explicit form and offers interesting details for the growth pattern, as mentioned above. Third, the use of MASAL will not substitute for that of parametric models. Instead, it will enhance the use of parametric models by guiding data analysts to choose appropriate interactions and transformations. Finally, regardless of whether we reach consistent results from both approaches, using two complementary approaches will either assure our conclusions, reveal the weakness of each other, or provide different insights.

4. Discussion

Using the data on the growth of 298 children, we demonstrate how to use MASAL to analyze a general class of growth curves. This class of data arises from a myriad of practical problems. In the literature, some of them are referred to as repeated measures and some as longitudinal data. Some of the traditional approaches assume that the trend is linear or quadratic polynomials, and some restrict the trend to be of specified forms. In practice, the involvement of baseline and/or time-dependent covariates makes it difficult to characterize the growth trend. On the other hand, the spline model used in this work requires no knowledge of the growth trend and hence offers a flexible and powerful tool for exploring the data. In addition, the spline model adopted here has an explicit and closed formula, which allows us to extract some interesting information.

In contrast to Zhang (1997), we have emphasized the exploration and estimation of the covariance structure, which is an important and difficult step in analyzing correlated data. We have seen the usefulness of coupling some traditional ideas of modeling autocorrelation patterns with a contemporary, adaptive spline approach in the analysis of growth curves and, in general, longitudinal data.

Although the performance of MASAL warrants further study, the results of Truong (1991) and Altman (1992), among others, on related topics seem to suggest that there may not be much to lose by taking the second model. Many numerical examples have shown that the second model is a reasonable choice. It is useful, however, to compare the second model with the third one to make sure that the degree of change is relatively minor.

It has not yet been proven that the iterative estimation process converges in a rigorous sense, but many numerical examples including this application suggest that the convergence is very likely when the degrees of freedom used in estimating the covariance matrix are relatively small as compared to the sample size. Moreover, if the spline model is replaced with some parametric models and if we assume that the measurement error has a Gaussian distribution, our iterative procedure is a form of the EM algorithm (see Laird and Ware, 1982). On the basis of the Gaussian assumption, we may force the iterative algorithm to converge by requiring the maximum log-likelihood function to increase at each iteration. Therefore, the convergence of the iterative procedure is not only practically implementable, but it is also theoretically promising.

Acknowledgements

This research is supported in part by NIH grant R29 HD30712. The author wishes to thank Mr Norman Silliker for his critique on the presentation of this work and Professors

Theodore Holford and Daniel Zelterman for their insightful and generous comments. The author is also grateful to the editor, James A. Calvin, an associate editor, and two referees for their constructive and specific suggestions.

RÉSUMÉ

Dans ce papier, nous étudions l'effet de l'usage de cocaïne par une femme enceinte sur la croissance de son nourrisson après la naissance. Utilisant un jeu de données d'une étude rétrospective nous avons trouvé que la consommation de cocaïne était d'une signification marginale sur la croissance d l'enfant. La variable mesurée étant le poids de l'enfant. D'un point de vue statistique les données présentent une structure habituelle de complexité mais peu étudiée dans la littérature statistique. Pour analyser ces données nous utilisons une approche récemment développée par Zhang (1997) noté MASAL (Multivariate Adaptive Splines for the Analysis of Longitudinal Data). En plus de l'ajustement des courbes de croissance nous montrons en particulier comment explorer et estimer les structures sous-jacentes de covariance pour des données longitudinales collectées à intervalle de temps irrégulier.

REFERENCES

Altman, N. S. (1992). An iterated Cochrane–Orcutt procedure for nonparametric regression. *Journal of Statistical Computation and Simulation* **40,** 93–108.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and Regression Trees.* Belmont, California: Wadsworth.

Carter, R. L., Resnick, M. B., Ariet, M., Shieh, G., and Vonesh, E. F. (1992). A random coefficient growth curve analysis of mental development in low-birth-weight infants. *Statistics in Medicine* **11,** 243–256.

Diggle, P. J., Liang, K. Y., and Zeger, S. L. (1994). *Analysis of Longitudinal Data.* New York: Oxford University Press.

Friedman, J. H. (1991). Multivariate adaptive regression splines. *The Annals of Statistics* **19,** 1–141.

Friedman, J. H. and Silverman, B. W. (1989). Flexible parsimonious smoothing and additive modeling (with discussion). *Technometrics* **31,** 3–39.

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

Stier, D. M., Leventhal, J. M., Berg, A. T., Johnson, L., and Mezger, J. (1993). Are children born to young mothers at increased risk of maltreatment? *Pediatrics* **91,** 642–648.

Truong, Y. K. (1991). Nonparametric curve estimation with time series errors. *Journal of Statistical Planning and Inference* **28,** 167–183.

Wasserman, D. R. and Leventhal, J. M. (1993). Maltreatment of children born to cocaine-dependent mothers. *American Journal of Diseases of Children* **147,** 1324–1328.

Zhang, H. P. (1994). Maximal correlation and adaptive splines. *Technometrics* **36,** 196–201.

Zhang, H. P. (1997). Multivariate adaptive splines for the analysis of longitudinal data. *Journal of Computational and Graphical Statistics* **6,** 74–91.

Received February 1997. *Revised June* 1998.
Accepted July 1998.