A Graphical Method of Exploring the Mean Structure in Longitudinal Data Analysis

Julian J. FARAWAY

In a longitudinal study, individuals are observed over some period of time. The investigator wishes to model the responses over this time as a function of various covariates measured on these individuals. The times of measurement may be sparse and not coincident across individuals. When the covariate values are not extensively replicated, it is very difficult to propose a parametric model linking the response to the covariates because plots of the raw data are of little help. Although the response curve may only be observed at a few points, we consider the underlying curve $y(t)$. We fit a regression model $y(t) = x^T\beta(t) + \epsilon(t)$ and use the coefficient functions $\beta(t)$ to suggest a suitable parametric form. Estimates of $y(t)$ are constructed by simple interpolation, and appropriate weighting is used in the regression. We demonstrate the method on simulated data to show its ability to recover the true structure and illustrate its application to some longitudinal data from the Panel Study of Income Dynamics.

Key Words: Curve estimation; Exploratory data analysis; Functional data analysis; Nonparametric regression; Repeated measures.

1. INTRODUCTION

Longitudinal data arise when a subject is observed over a period of time, $\mathcal{T}$, or over some other continuous variable. Suppose the response of the ith individual is $y_i(t)$, where $t \in \mathcal{T}$. Now in practice one never observes the whole continuous function $y_i(t)$ but only at a finite number of points $t_{ij} \in \mathcal{T}$, where $i = 1, \ldots, n$ and $j = 1, \ldots, m_i$. In many applications m_i is quite small, and sometimes the points of observation, t_{ij}, differ from individual to individual. Even when an experiment has been designed to make the points of observation the same, data is often missing and the pattern will be broken. Now even though the m_i may be relatively small and there might seem to be little hope of reconstructing $y_i(t)$, we take the point of view that it is worthwhile considering the underlying function rather than the vector $y_i(t_{ij})$, $j = 1, \ldots, m_i$.

Suppose that we observe covariates x_i, a vector of length p, associated with each individual i. We wish to develop a model of the form

$$y_i(t) = f(t, x_i) + \epsilon_i(t).$$

Julian J. Faraway is Associate Professor, Department of Statistics, University of Michigan, Ann Arbor, MI 48109-1027 (E-mail: faraway@umich.edu).

Journal of Computational and Graphical Statistics, Volume 8, Number 1, Pages 60–68

Our main objective is to determine an appropriate form for the mean structure $f()$. With a sufficiently large amount of data (in the sense that the range $\mathcal{T}$ is more densely covered) we might attempt to formulate an at least partially nonparametric model. When there is less data, or a compact formulation of the model is required, then a fully parametric form for $f()$ will be desirable. The primary difficulty is in choosing a suitable parametric form. This is the problem this article sets out to solve.

The error structure is contained in $\epsilon_i(t)$. I assume that each $\epsilon_i(t)$ is an independent realization of a stochastic process with mean 0 and covariance function $\gamma(s, t), s, t \in \mathcal{T}$. So I allow for correlated errors within individuals, but not between individuals. The errors do not depend on the covariates. The purpose of this article, however, is not to come up with novel methods for determining the form of $\gamma()$ as this problem has been investigated by other authors—see, for example, Izenman and Williams (1989) and Grambsch, Randall, Bostick, Potter, and Louis (1995) among others.

Little work exists on exploratory methods for finding an appropriate mean structure for longitudinal models. When there are no covariates or sufficiently large numbers of individuals that have the same values of the covariates, then it is possible to plot the data to obtain a suggestion for the appropriate form for $f()$. This is described in Rice and Silverman (1991) and Diggle, Liang, and Zeger (1995). When the covariates are continuous in nature, these methods will not be effective. When the points of measurement, t_{ij}, are in common across individuals, then it may be possible to avoid specifying a functional form for the mean structure by saturating the model with enough parameters to model the response at each time point, but this also requires some replication of covariate values. The alternative is to use contextual information as much as possible, but in practice, often the parametric form must be guessed and experimented with which is less than satisfactory. Other semiparametric approaches to related problems can be found in Hoover, Rice, Wu, and Yang (1996), Brumbach and Rice (1996), and Zhang, Lin, Raz, and Sowers (1998). A fully functional approach is described in Ramsay and Silverman (1997).

The method we propose here is exploratory and graphical in nature. Parametric assumptions are avoided because the method is intended to suggest a suitable parameterization. The spirit of the method is similar to the alternating conditional expectation (ACE) method of Breiman and Friedman (1985) and the generalized additive models of Hastie and Tibshirani (1990), which although nonparametric in nature can be used to suggest suitable parametric forms for functions of covariates in regression problems. The method can also be used as a diagnostic as well as an exploratory method when the appropriate mean structure is thought to be known, but when some caution is required.

Methodology is laid out in Section 2. We check the efficacy of the method for some simulated data in Section 3, where we know the true model, and demonstrate the method on some real data in Section 4. A discussion follows in Section 5.

2. METHOD

2.1 Functional Regression Analysis

Suppose that we may observe the whole functional responses, $y_i(t)$, which are assumed to arise from the model

$$y = X\beta + \epsilon, \tag{2.1}$$

where β is a vector of functions $(\beta_1(t), \dots \beta_p(t))^T$ and X is the familiar $n \times p$ design matrix formed from the p vector valued covariates x_i, $i = 1, \dots n$. As in scalar (i.e., scalar y) regression, the first column of X will usually all be ones and categorical predictors can be handled by assigning appropriate dummy variables. Various transformations of the predictors can be incorporated in X as in scalar regression. Also y is a vector of response functions $(y_1(t), \dots y_n(t))^T$, and ϵ is a vector of error functions $(\epsilon_1(t), \dots, \epsilon_n(t))^T$.

Suppose we choose $\hat{\beta}$ to minimize $\sum_{i=1}^n ||y_i - x_i^T \beta||^2$, where $|| \cdot ||$ is the L_2 norm on $\mathcal{T}$. In parametric longitudinal analysis, distributional assumptions are made and the model can be fit using maximum likelihood methods. Of course, this fit will be more efficient than using the least squares criterion, but at the price of making some possibly unjustifiable assumptions. By considering each $t \in \mathcal{T}$ separately, and provided that X has full rank, as in the usual regression situation, it is clear that the solution is

$$\hat{\beta} = (X^T X)^{-1} X^T y.$$

The optimality of this estimator is discussed in Ramsay and Dalzell (1991). Let $\hat{y} = X\hat{\beta}$ and $\hat{\epsilon} = y - \hat{y}$. We could also use a weight matrix W to get $\hat{\beta} = (X^T W X)^{-1} X^T W y$. Because the solution is computed pointwise, the weight matrix can change as t varies. We will need this feature later. See Faraway (1997) or Ramsay and Silverman (1997) for a development of this model when the $y(t)$ are observed more densely.

2.2 Range of Representation

Not all longitudinal models of interest may be represented in the form (2.1). Clearly covariates that are explicitly a function of time cannot be represented within this model without some further generalization. For simplicity consider a situation where there is only one (univariate) covariate z. The model with an intercept will take the form

$$y(t) = \beta_0(t) + z\beta_1(t) + \epsilon(t).$$

The functional form of $\beta_0(t)$ and $\beta_1(t)$ are estimated which may then suggest a suitable parametric model. A wide range of parametric forms fall within this setup. Further generalization is possible by allowing for transformations of z as in

$$y(t) = \beta_0(t) + \sum_l g_l(z)\beta_l(t) + \epsilon(t), \tag{2.2}$$

where the $g_l()$'s must be specified. We provide no way of estimating the correct forms for the $g_l()$; in practice, some experimentation and perhaps luck would be needed to find these forms. It is conceivable that, with a sufficiently large number of individuals, we might attempt some nonparametric estimation of $g_l(z)$, but we do not investigate that in this article.

Models such as

$$y(t) = \log(zt + z^2t^2) + \epsilon(t)$$

do not fall within the factorized form (2.2) and lie outside the range of representation. We would have no chance of discovering such a model using our method, although prior substantive knowledge would likely be necessary to choose such a model anyway. In some cases, an appropriate transformation of the response might allow the model to be expressed in a representable form.

2.3 INTERPOLATION AND WEIGHTING

In practice, we do not observe $y_i(t), i = 1, \ldots, n$, only $y_{ij}, j = 1, \ldots m_i$ at times t_{ij} where

$$y_{ij} = x_i^T \beta(t_{ij}) + \epsilon_{ij},$$

where $\epsilon_{ij} = \epsilon_i(t_{ij})$.

To estimate β, we will first estimate $y_i(t)$ on a fine grid of points $t_j, j = 1, \ldots m$, where the grid of points is in common for all i. Call these estimates $y_i^l(t)$.

We will use linear interpolation to construct the $y_i^l(t)$ using a constant for the estimate before the first and after the last observation in the range $\mathcal{T}$; that is, $y_i^l(t) = y_{i1}$ for $t < t_{i1}$ and $y_i^l(t) = y_{im_i}$ for $t > t_{im_i}$. This is not the standard nonparametric regression problem. We wish to estimate $y(t)$, not $Ey(t)$. Certainly, linear interpolation is a poor method of estimating $Ey(t)$, but that is not our objective. We wish to estimate $\beta(t)$. One might contemplate some smoothing, but there are several reasons for preferring linear interpolation, such as:

1. The m_i might be very small, in which case more sophisticated methods of nonparametric regression will not be appropriate—our method of linear interpolation will work with only one point.
2. Not smoothing the estimates at all means that the bias will be minimized. At the points of observation t_{ij} there is no bias at all. When $\hat{\beta}(t)$ is constructed, the bias will be further reduced by the averaging. We can always smooth the estimates of $\beta(t)$, but if we smooth the y_{ij} too much, we may lose features that cannot later be recovered. One exception to this idea of avoiding smoothing the y_{ij} is that we should be beware of outliers. We might seek to eliminate these by a prescreening.
3. In the usual nonparametric regression problem, the errors at each time point are assumed to be independent. In that case, linear interpolation is a very poor method asymptotically. However, for that problem, the objective is to estimate the mean curve $Ey(t)$. Here, we are more interested in the observed curve $y(t)$ and the appropriate asymptotics will be different. Smoothing might filter out some measurement error, but we would rather retain the random effects and serial correlation parts of $y(t)$ so that the error structure of $\epsilon(t)$ can be analyzed.

4. Linear interpolation is simple and easily implemented without special software. At a particular point t, we have

$$y_i^l = x_i^T\beta + b_i + \epsilon_i, \quad i = 1, \dots n,$$

where $\text{var}(\epsilon) = \Omega$ (which is a diagonal matrix by the assumption of independence between individuals) and biases b_i. The presence of the biases b_i is what distinguishes this from the usual regression setup. At the points t_{ij}, the bias is zero. How should the weight matrix W be chosen in this case?

The estimator is

$$\hat{\beta} = (X^TWX)^{-1}Wy^l = Ay^l$$

so that $\text{bias}(\hat{\beta}) = Ab$ and $\text{var}(\hat{\beta}) = A\Omega A^T$, where $A = (X^TWX)^{-1}W$. Notice that the bias of $\hat{\beta}$ is a linear combination of the biases b_i. Because some cancellation is likely to occur, the bias of $\hat{\beta}$ is likely to be reduced. Now $\text{MSE}(\hat{\beta}) = A[bb^T + \Omega]A^T$, which suggests choosing $W = [bb^T + \Omega]^{-1}$ which gives $\text{MSE}(\hat{\beta}) = [X^T[bb^T + \Omega]^{-1}X]^{-1}$.

It would now be nice to have estimates of the bias and variance. However, given the lack of information about the true mean response and the error structure combined with possibly very few observations per individual, it is very difficult to find good estimates of the bias and variance. Various, more ambitious methods than the one I will recommend in the following failed because of a lack of stability in these estimates.

Note that $\text{MSE}(\hat{\beta})$ has nonzero off-diagonal elements so a separate estimate of the bias is required to approximate the optimal weight matrix. However, since it is difficult to estimate the bias we propose to ignore the off-diagonal elements—we will use weighted least squares rather than generalized least squares.

If we assume that $y(t)$ is locally quadratic, this suggests that the bias at t might be estimated by a term of the form

$$b_i(t) = y_i(t) - y_i^l(t) = \pm\gamma\{t^2 + t_ut_l - t(t_u + t_l)\}$$

for $t \in [t_l, t_u]$, where the closest bracketing timepoints are at t_l and t_u. This expression may be derived by computing the distance between a line and a parabola and by observing that $b_i(t_l) = b_i(t_u) = 0$. $\gamma > 0$ is a tuning parameter which should be chosen with respect to the speculated amount of curvature in $y(t)$. We cannot determine the sign of $b_i(t)$ because we do not know whether $y_i(t)$ has positive or negative curvature in the local interval. This also illustrates why we chose to ignore the off-diagonal elements of W, whereas the diagonal elements of bias are squared and so the sign is irrelevant. Bias is not necessarily greatest in the middle of a large gap between observations, but without any further information it is safest to downweight any $y_i^l(t)$ that is estimated using relatively distant observations. For $t < t_{i1}$, we suggest $\gamma(t_{i1} - t)^2$ and for $t > t_{im_i}$, $\gamma(t - t_{im_i})^2$.

Now suppose we assume that the error process, $\epsilon(t)$, has variance, σ^2, which is constant in time, then we will choose the weights as

$$w_i(t) = \frac{1}{\sigma^2 + b_i^2(t)} \equiv \frac{1}{1 + b_i^2(t)}.$$

The last equivalency arises by observing that the multiplication of weights by a constant term has no effect on the calculation of the regression coefficients and that we may absorb constant multipliers of the tuning parameter γ. If we know that the variance is not constant in time and takes the form $\sigma^2(t) \equiv \sigma^2 v(t)$, for known $v(t)$, then an appropriate form for the weights will be

$$w_i(t) = \frac{1}{v(t) + b_i^2(t)}.$$

The $w_i(t)$ may be plotted to guide the selection of γ. With larger datasets, γ can be made bigger to localize the effect of observations. Since the aim of the analysis is exploratory, manual adjustment of γ is adequate—there is no necessity for smoothing parameter selection methods. Other weighting systems can easily be implemented if more information is available.

2.4 CONSTRUCTING THE PARAMETRIC MODEL

Pointwise standard errors may easily be constructed using the standard regression formula: $\mathrm{se}(\hat{\beta}_k(t)) = \sqrt{(X^T W X)^{-1}_{kk}}\hat{\sigma}(t)$. We may then plot $\hat{\beta}_k(t) \pm 2\mathrm{se}(\hat{\beta}_k(t))$ against t for each $k = 1, \ldots, p$. We can now

1. Gain insight on a suitable parametric form for $\beta(t)$, using the pointwise confidence bands as guide to what the reasonable forms are.
2. Check these estimated $\hat{\beta}(t)$ against an already proposed parametric model looking for possible inadequacies.

It would be better to use simultaneous rather than pointwise confidence bands, but this would require more knowledge of the error structure than we care to presume.

3. SIMULATION

To see how well our method could recover the true $\beta(t)$, we generated data from a known model:

$$Y_{ij} = \beta_0(t_{ij}) + z_{1i}\beta_1(t_{ij}) + z_{2i}\beta_2(t_{ij}) + \epsilon_{ij}, \tag{3.1}$$

where

$$\begin{aligned} \beta_0(t) &= t/2 \\ \beta_1(t) &= \begin{cases} .5 - t & \text{if } t < .5 \\ 0 & \text{otherwise} \end{cases} \\ \beta_2(t) &= \begin{cases} t - .5 & \text{if } t > .5 \\ 0 & \text{otherwise,} \end{cases} \end{aligned}$$

and z_{1i} and z_{2i} are iid $U[-1, 1]$ and ϵ_{ij} is iid $N(0, .1^2)$, $i = 1, \ldots 20$ and $j = 1, \ldots 5$. The plot of the raw data is shown in the first panel of Figure 1.

Notice that it would be almost impossible to guess the correct form of the model by looking at the plot of the raw data. There is no suggestion of a change in trend at $t = .5$.

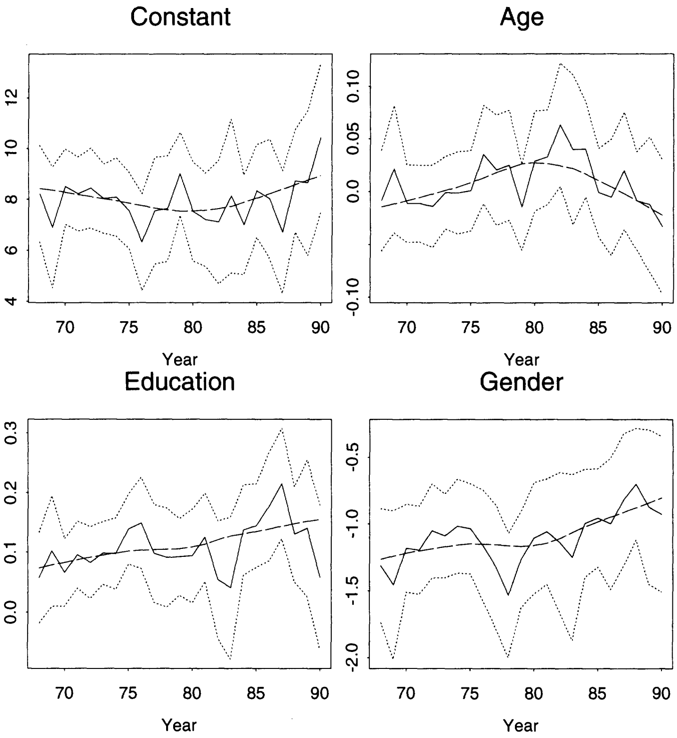

Figure 1. The first panel shows the data and the remaining three panels show the estimated $\beta(t)$ (solid), two standard error bands (dotted), and the true $\beta(t)$ (dashed).

However, our method recovers the true functional form very well. We see there is linear increase in the response over time unrelated to either covariate, $\hat{\beta}_0(t)$, that the effect of z_1 decreases linearly over time to zero at $t = .5$ and is constant after that, $\hat{\beta}_1(t)$, and the complimentary effect for z_2, $\hat{\beta}_2(t)$. We should not be too concerned that the true $\beta(t)$ ventures slightly outside the bands as these are pointwise, not simultaneous, confidence bands.

Obviously one example is not completely convincing, but we have tried many more. I invite the reader to explore their own simulated models using S-Plus software available from www.stat.lsa.umich.edu/~faraway.

4. EXAMPLE

The Panel Study of Income Dynamics (PSID), begun in 1968, is a longitudinal study of a representative sample of U.S. individuals (Hill 1992). The study is conducted at the Survey Research Center, Institute for Social Research, University of Michigan, and is still continuing. There are currently 8,700 households in the study, and many variables are measured. I chose to analyze a random subset of this data, consisting of 85 heads of household who were aged between 25–39 in 1968 and had complete data for at least 11 of the years between 1968 and 1990. The variables included were annual

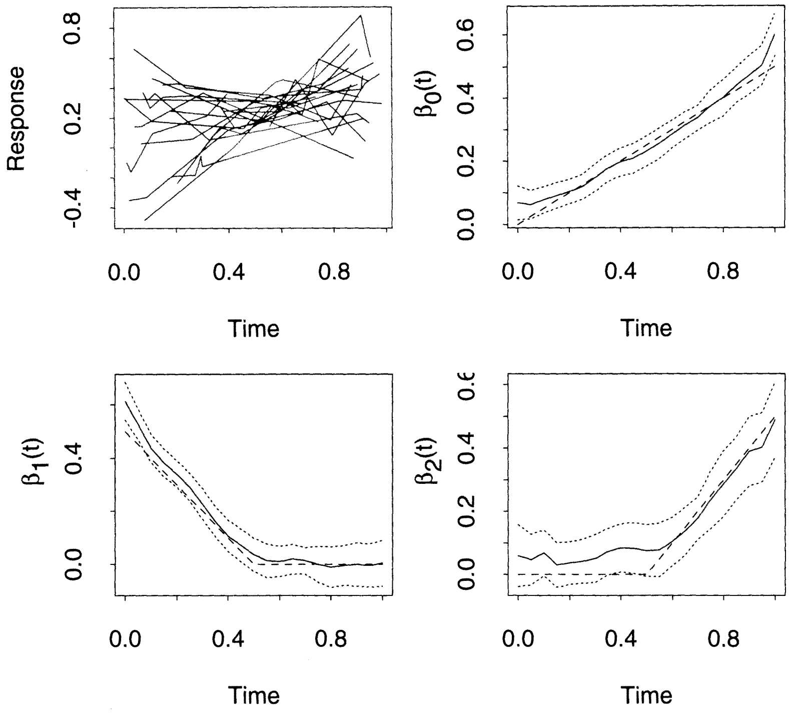

Figure 2. The coefficient functions $\hat{\beta}(t)$ (solid line) for the model (4.1) with ±2 standard errors (dotted lines) and lowess smooth of $\hat{\beta}(t)$ (dashed line). For gender, the effect of being female is observed.

income, gender, years of education and age in 1968. The analysis here is intended to demonstrate the methodology. Because we have selected a subsample and have excluded various important variables, this should not be construed as a complete attempt to draw conclusions about the subject matter. We will attempt to model the annual income over the period of interest as a function of the covariates mentioned above. We will consider a model for individual i of the form:

$$\log(\text{income})_i(t) = \beta_0(t) + \beta_g(t)\text{gender}_i + \beta_e\text{education}_i + \beta_a\text{age}_i\epsilon_i(t) \qquad (4.1)$$

Since in this example the response, income, is measured at yearly intervals, it is convenient (although not necessary) to make the grid of timepoints at which the functions will be estimated coincident with the years. For this data it it also convenient to make γ very large so that missing observations, although interpolated are effectively not used. A plot of the estimated coefficient functions is shown in Figure 2 with a lowess-smoothed fit to aid interpretation.

The constant function $\hat{\beta}_0(t)$ is not the mean income at each timepoint, so we should not be disturbed that this does not increase monotonically with time. Age may well not be important at all given that the zero line lies within the two standard error bands, but there is some indication of a quadratic effect. The effect of Education seems approximately linear with the relative effect of Education increasing over time. The coding of the gender variable meant that the effect of being female is represented. We can see that women earn substantially less although the difference is decreasing approximately linearly with time.

This analysis gives an idea of what functional forms for a parametric longitudinal model might be appropriate. Some more work needs to be done in specifying a model, but at least we now have good indication of how to do this.

5. DISCUSSION

Our method is not appropriate for all types of longitudinal data. For example, where each individual is measured at the same small number of time points, then there would be little value in using our method. However, although data may seem to have been collected at the same time points, the concordance is lost when a different origin is used—for example, patients may be measured at equal time points from the start of treatment, but if chronological age is used as the origin, the times of measurement are no longer the same for each individual. We also have not described how to handle time-varying covariates, although a simple interpolation approach for this may also be feasible. Nevertheless, the method provides a simple way to explore certain types of longitudinal data.

[Received January 1997. Revised March 1998.]

REFERENCES

Breiman, L., and Friedman, J. (1985), "Estimating Optimal Transformations for Multiple Regression and Correlation" (with discussion), *Journal of the American Statistical Association*, 80, 580–619.

Brumbach, B., and Rice, J. (1996), "Smoothing Spline Models for the Analysis of Nested and Crossed Samples of Curves," Technical Report 474, University of California, Berkeley, Dept. of Statistics.

Diggle, P., Liang, K., and Zeger, S. (1995), *Analysis of Longitudinal Data*, Oxford: Oxford University Press.

Faraway, J. (1997), "Regression Analysis for a Functional Response," *Technometrics*, 39, 254–261.

Grambsch, P., Randall, B., Bostick, R., Potter, J., and Louis, T. (1995), "Modeling the Labeling Index Distribution: An Application of Functional Data Analysis," *Journal of the American Statistical Association*, 90, 813–821.

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

Hill, M. S. (1992), *The Panel Study of Income Dynamics: A User's Guide*, Newbury Park, CA: Sage Publications.

Hoover, D., Rice, J., Wu, C., and Yang, L. (1996), "Nonparametric Smoothing Estimates of Time-Varying Coefficient Models With Longitudinal Data," Technical Report 459, University of California, Berkeley, Dept. of Statistics.

Izenman, A., and Williams, J. (1989), "A Class of Linear Spectral Models and Analyses for the Study of Longitudinal Data," *Biometrics*, 45, 831–849.

Ramsay, J., and Dalzell, C. (1991), "Some Tools for Functional Data Analysis," *Journal of the Royal Statistical Society*, Ser. B, 53, 539–572.

Ramsay, J., and Silverman, B. (1997), *Functional Data Analysis*, New York: Springer.

Rice, J., and Silverman, B. (1991), "Estimating the Mean and Covariance Structures Nonparametrically When the Data are Curves," *Journal of the Royal Statistical Society*, Ser. B, 57, 673–690.

Zhang, D., Lin, X., Raz, J., and Sowers, M. (1998), "Semiparametric Stochastic Mixed Models for Longitudinal Data," *Journal of the American Statistical Association*, 93, 710–719.