BIOMETRICS 57, 795–802
September 2001

Linear Mixed Models with Flexible Distributions of Random Effects for Longitudinal Data

Daowen Zhang* and Marie Davidian

Department of Statistics, North Carolina State University,
Box 8203, Raleigh, North Carolina 27695-8203, U.S.A.
**email:* dzhang2@stat.ncsu.edu

SUMMARY. Normality of random effects is a routine assumption for the linear mixed model, but it may be unrealistic, obscuring important features of among-individual variation. We relax this assumption by approximating the random effects density by the seminonparameteric (SNP) representation of Gallant and Nychka (1987, *Econometrics* **55,** 363–390), which includes normality as a special case and provides flexibility in capturing a broad range of nonnormal behavior, controlled by a user-chosen tuning parameter. An advantage is that the marginal likelihood may be expressed in closed form, so inference may be carried out using standard optimization techniques. We demonstrate that standard information criteria may be used to choose the tuning parameter and detect departures from normality, and we illustrate the approach via simulation and using longitudinal data from the Framingham study.

KEY WORDS: Longitudinal data; Multimodality; Random effects; Seminonparametric density; Semiparametric mixed effects model; Skewness.

1. Introduction

Longitudinal data are collected in clinical trials, epidemiology, and agriculture, and interest may focus on population effects of within- and among-individual covariates on the response as well as on individual-specific behavior. A routine framework for analysis is the linear mixed effects model, where random effects are incorporated to accommodate among-subject variation (Laird and Ware, 1982). Fundamental assumptions in the standard version of the model are that within-subject errors and random effects are normally distributed. Under these assumptions, inference on fixed model parameters and random effects can be carried out using widely available software (e.g., SAS's `proc mixed`; Littell et al., 1996).

Although within-subject conditional normality may be realistic, the normality assumption on the random effects may be too restrictive to provide an accurate representation of among-individual variation. Figure 1a shows cholesterol levels over time for 200 randomly selected individuals from the Framingham study. The data consist of participants' cholesterol levels measured at the beginning of the study and then every 2 years for 10 years, age at baseline, and gender. A standard objective for such data is to characterize change in cholesterol over time, the effect of baseline covariates, and among-subject variation in cholesterol levels. Figure 1a suggests that cholesterol increases linearly over time for most subjects but with substantial intersubject variation. To explore this informally, we fit individual profiles by simple linear regression over time. A pooled residual plot supports the assumption of within-individual normality; however, although estimated subject-specific slopes appear normal, Figure 1b suggests that variation in intercepts may not be normally distributed. This pattern may be only partially explained by the available covariates.

Although inference on fixed effects may be robust to nonnormality of random effects (Butler and Louis, 1992; Verbeke and Lesaffre, 1997), it is natural to be concerned about efficiency and validity of inference on individual effects. Moreover, estimation of the random effects distribution under less restrictive assumptions may provide considerable insight; e.g., a multimodal or skewed estimate may suggest exclusion of important covariates and reveal critical features of inherent subject heterogeneity. Thus, considerable interest has focused on relaxing the normality assumption and jointly estimating the random effects distribution and model parameters. Under minimal assumptions on the distribution, one obtains the discrete nonparametric maximum likelihood estimate (e.g., Laird, 1978; Aitken, 1999). However, for continuous responses, it is reasonable to suppose that the random effects are continuous, so recent proposals have been made under the assumption of a smooth random effects density. Magder and Zeger (1996) propose a smooth nonparametric maximum likelihood approach, which entails some computational burden and uses somewhat *ad hoc* fitting and assessment of the fit. Tao et al. (1999) estimate the density of a scalar random effect via their predictive recursive algorithm. Verbeke and Lesaffre (1996) use a mixture of normals, which they implement via an EM algorithm (Verbeke and Molenberghs, 2000, Chapter 12).

In this article, we propose an alternative method that is particularly attractive for linear mixed models. Assuming a smooth random effects density, we describe in Section 2 a

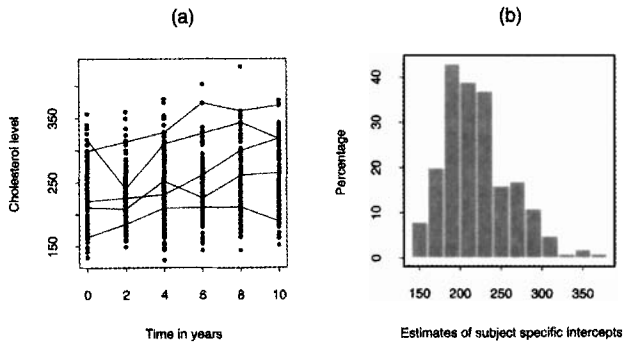

Figure 1. Framingham cholesterol data. **a.** Longitudinal cholesterol levels for 200 subjects, with trajectories highlighted for 5 random subjects. **b.** Histogram of subject-specific intercept estimates from individual least squares fits.

semiparametric linear mixed model using the seminonparametric (SNP) representation of Gallant and Nychka (1987) for the density of random effects. Davidian and Gallant (1993) use this for nonlinear mixed models; however, in that setting, inference is complicated by the need for intractable integration via intensive numerical techniques and for imposition of identifiability constraints on the SNP density. But as we show in Section 3, for the linear mixed model, the SNP allows expression of the marginal likelihood of the data in closed form, and we suggest a parameterization of the SNP representation that eases the complication of ensuring identifiability. This approach facilitates straightforward implementation with standard optimization routines. In Section 4, we illustrate the method for the Framingham data in Figure 1, and we present simulation results in Section 5.

2. Semiparametric Linear Mixed Model

Suppose Y_{ij}, $i = 1,\ldots,m$, $j = 1,\ldots,n_i$, the response for subject i at time t_{ij}, satisfies

$$Y_{ij} = x_{ij}^{\mathrm{T}}\beta + s_{ij}^{\mathrm{T}}b_i + \epsilon_{ij}, \tag{1}$$

where β $(p \times 1)$ is a vector of fixed effects associated with covariate vector x_{ij}; b_i are q-dimensional, mutually independent, subject-specific random effects associated with covariate vector s_{ij}; and $\epsilon_{ij} \sim \mathrm{N}(0, \sigma^2)$ are mutually independent errors independent of the b_i.

To facilitate our further development, represent the random effects as

$$b_i = \mu + RZ_i, \tag{2}$$

where μ is a $(q \times 1)$ vector of parameters, R is a $(q \times q)$ lower triangular matrix, and Z_i is a $(q \times 1)$ random vector. Rather than assume standard multivariate normality for Z_i, we assume that Z_i, and hence b_i, has a density belonging to a class of smooth densities discussed in detail by Gallant and Nychka (1987). The mathematical details are complex, but, practically speaking, densities in this class are sufficiently differentiable that they do not exhibit unusual behavior such as kinks, jumps, or oscillation but may be skewed, multimodal, and fat- or thin-tailed relative to the normal; moreover, this class contains the normal. As discussed in Davidian and Giltinan (1995, Chapter 7), densities in this class may be approximated by a truncated series expansion, and the resulting estimation methods and the density approximation are referred to as seminonparametric (SNP). Applying this to (2), for inference, we propose to represent the density of Z_i by the standard SNP density,

$$h_K(z) = P_K^2(z)\varphi(z) = \left\{ \sum_{|\lambda| \le K} a_\lambda z^\lambda \right\}^2 \varphi(z), \tag{3}$$

where $\lambda = (\lambda_1,\ldots,\lambda_q)$ is a q-dimensional vector of nonnegative integers, z^λ is the monomial $z^\lambda = z_1^{\lambda_1}\cdots z_q^{\lambda_q}$ of order $|\lambda| = \Sigma_{k=1}^q \lambda_k$, $\varphi(z)$ is q-dimensional standard normal density, K is the order of the polynomial $P_K(z)$, and the coefficients a_λ satisfy conditions discussed in the next paragraph. For example, when $K = 2$, $q = 2$, $P_2(z) = a_{00} + a_{10}z_1 + a_{01}z_2 + a_{20}z_1^2 + a_{11}z_1z_2 + a_{02}z_2^2$. When $K = 0$, $P_K(z) \equiv 1$ under the conditions discussed next ($a_{00} = 1$ in this case), so (3) includes the normal as a special case, and (1) reduces to the usual linear mixed model with $b_i \sim \mathrm{N}_q(\mu, RR^{\mathrm{T}})$. Note that, for the sake of identifiability, the formulation of the model using (2) thus requires that x_{ij} not contain s_{ij}, but this does not pose any practical restriction, as we illustrate in Section 4. The order K acts as a tuning parameter controlling the degree of flexibility of shape of the resulting density $h_K(z)$. As demonstrated in Sections 4 and 5, K need be no larger than one or two to approximate complicated shapes, including multimodality and skewness.

For $h_K(z)$ to be a density, the coefficients a_λ of $P_K(z)$ must be chosen so that $\int h_K(z)dz = 1$. Previously, (e.g., Davidian and Giltinan, 1995, Chapter 7), $h_K(z)$ was defined alternatively as (3) divided by the appropriate normalizing constant, with the leading constant a_{00} of the polynomial set to one to achieve identifiability. We instead ensure $\int h_K(z)dz = 1$ by imposing

$$\mathrm{E}\{P_K^2(U)\} = 1, \quad \text{where } U \sim \mathrm{N}_q(0, I); \tag{4}$$

the consequences for improved numerical stability are discussed in Section 3.2.

This expectation is easy to calculate by representing it in an equivalent way. To demonstrate, consider $q = 2$; then there are $d = (K+1)(K+2)/2$ distinct terms in $P_K(u)$, and we may write

$$P_K(u) = \sum_{0 \le i+j \le K} a_{ij}u_1^i u_2^j = \sum_{\ell=0}^{K}\sum_{j=0}^{\ell} a_{\ell-j,j}u_1^{\ell-j}u_2^j$$
$$= \sum_{i=1}^{d} a_i u_1^{i_1}u_2^{i_2},$$

where a_i and the powers i_1, i_2 are easily determined numerically. For example, with $K = 3$, a_i is the ith element of $a = (a_{00}, a_{10}, a_{01}, a_{20}, a_{11}, a_{02}, a_{30}, a_{21}, a_{12}, a_{03})^{\mathrm{T}}$ and i_1, i_2 are just the subscripts corresponding to a_i. Letting U_a be the random vector whose ith element is $U_1^{i_1}U_2^{i_2}$, where U_1 and U_2 are independent standard normal, we have $P_K(U) = a^{\mathrm{T}}U_a$, so that

$$\mathrm{E}\{P_K^2(U)\} = a^{\mathrm{T}}\mathrm{E}(U_aU_a^{\mathrm{T}})a = a^{\mathrm{T}}Aa, \tag{5}$$

where A is the matrix with (i,j) element $\mathrm{E}(U_1^{i_1+j_1})\mathrm{E}(U_1^{i_2+j_2})$

and the superscripts correspond to a_i and a_j. These expectations are straightforward by standard recursive formulas (e.g., Johnson and Kotz, 1994), and it follows that A is a sparse, positive definite matrix. When $K = 2$, an alternative derivation of (5) is straightforward, as suggested by a reviewer. Writing $P_2(u) = a_{00} + u^{\mathrm{T}}a^{(1)} + u^{\mathrm{T}}A^{(2)}u$, where $u = (u_1, \ldots, u_q)^{\mathrm{T}}$ and, e.g., with $q = 2$, $a^{(1)} = (a_{10}, a_{01})^{\mathrm{T}}$ and

$$A^{(2)} = \begin{pmatrix} a_{20} & a_{11}/2 \\ a_{11}/2 & a_{02} \end{pmatrix}.$$

Now $u^{\mathrm{T}}A^{(2)}u = \mathrm{trace}(uu^{\mathrm{T}}A^{(2)}) = \mathrm{vec}(uu^{\mathrm{T}})^{\mathrm{T}}\mathrm{vec}(A^{(2)})$, so that $P_2(u) = a_{00} + u^{\mathrm{T}}a^{(1)} + \mathrm{vec}(uu^{\mathrm{T}})^{\mathrm{T}}\mathrm{vec}(A^{(2)}) = \{a_{00}, a^{(1)T}, \mathrm{vec}(A^{(2)})^{\mathrm{T}}\}\{1, u^{\mathrm{T}}, \mathrm{vec}(uu^{\mathrm{T}})^{\mathrm{T}}\}^{\mathrm{T}}$. Thus, that $P_2^2(u)$ is of a quadratic form follows, and it may be shown that the expectation is given by (5).

The above arguments may be generalized to any q and K and automated for computational purposes. We may thus represent (4) as

$$a^{\mathrm{T}}Aa = 1. \tag{6}$$

Letting $r = \mathrm{vech}(R)$ denote the nonzero elements of R, the parameters of interest are $\theta = (\beta^{\mathrm{T}}, \mu^{\mathrm{T}}, a^{\mathrm{T}}, r^{\mathrm{T}}, \sigma)^{\mathrm{T}}$, with a subject to the constraint (6), as well as the tuning parameter K, the order of the polynomial $P_K(z)$.

3. Estimation Procedure

3.1 *Likelihood Function*

Substituting (2) in (1); defining $v_{ij} = (x_{ij}^{\mathrm{T}}, s_{ij}^{\mathrm{T}})^{\mathrm{T}}$, $\delta = (\beta^{\mathrm{T}}, \mu^{\mathrm{T}})^{\mathrm{T}}$, $Y_i = (Y_{i1}, \ldots, Y_{in_i})^{\mathrm{T}}$, and ϵ_i similarly; and letting V_i $\{n_i \times (p+q)\}$ and S_i $(n_i \times q)$ be the matrices with rows v_{ij}^{T} and s_{ij}^{T}, respectively, we may express the model as

$$Y_i = V_i\delta + S_iRZ_i + \epsilon_i. \tag{7}$$

Given K, the log likelihood of θ is

$$\ell(\theta; Y) = \sum_{i=1}^{m} \log\{f(Y_i; \theta)\},$$

where $f(Y_i; \theta)$ is the marginal density of Y_i for subject i. Keeping in mind dependence of $P_K(z)$ on θ, we have

$$f(Y_i; \theta) = \int f(Y_i \mid z; \theta)P_K^2(z)\varphi(z)dz,$$

where $f(Y_i \mid z; \theta)$ is the normal density with mean $V_i\delta + S_iRz$ and covariance $\sigma^2 I$. Clearly, $f(Y_i \mid Z_i; \theta)\varphi(Z_i)$ can be identified as the joint density of Y_i and Z_i when Z_i in (7) is assumed to be standard q-variate normal. Under these conditions, let $g(Y_i; \theta)$ denote the marginal density of Y_i and $g(Z_i \mid Y_i; \theta)$ be the conditional density of Z_i given Y_i if Z_i were normal. Then $f(Y_i \mid Z_i; \theta)\varphi(Z_i) = g(Y_i; \theta)g(Z_i \mid Y_i; \theta)$ and hence

$$\begin{aligned} f(Y_i; \theta) &= g(Y_i; \theta) \int P_K^2(z)g(z \mid Y_i; \theta)dz \\ &= g(Y_i; \theta)\mathrm{E}_{Z_i|Y_i;\theta}\left\{P_K^2(Z_i)\right\}, \end{aligned}$$

where $\mathrm{E}_{Z_i|Y_i;\theta}(\cdot)$ represents expectation with respect to the conditional distribution of Z_i given Y_i with Z_i normal under θ. It is straightforward to show that $g(Y_i; \theta)$ is the normal density with mean $V_i\delta$ and covariance $\sigma^2(I + S_i\Gamma\Gamma^{\mathrm{T}}S_i^{\mathrm{T}})$, where $\Gamma = R/\sigma$, and that $g(z \mid Y_i; \theta)$ is the normal density with mean and covariance $\mu_i = \sigma^{-1}\Sigma_i\Gamma^{\mathrm{T}}S_i^{\mathrm{T}}(Y_i - V_i\delta)$ and $\Sigma_i = (I + \Gamma^{\mathrm{T}}S_i^{\mathrm{T}}S_i\Gamma)^{-1}$, respectively. Thus, the log likelihood may be expressed alternatively as

$$\ell(\theta; Y) = \sum_{i=1}^{m} \log\{g(Y_i; \theta)\} + \sum_{i=1}^{m} \log\left[\mathrm{E}_{Z_i|Y_i;\theta}\left\{P_K^2(Z_i)\right\}\right]. \tag{8}$$

The first term in (8) is just the usual log-likelihood function for the linear mixed model (7) when Z_i is assumed normal, which is available in closed form from above. The second term involves calculation of moments of a normal random vector with mean μ_i and variance matrix Σ_i, $i = 1, \ldots, m$, so also has a closed-form expression. Thus, the log likelihood for θ under the semiparametric linear mixed model has a convenient, closed-form representation. A simple example is given in the Appendix.

Calculation of the required moments in the second term may be accomplished using the series representation of the moment generating function of a normal random variable (see the Appendix). The calculation can be intensive for q large, but in this situation, it may not be prudent to entertain the semiparametric model for any smooth representation of random effects. For moderate q, the calculations are not prohibitive and are straightforward to implement.

3.2 *Maximizing the Likelihood Function*

For given K, obtaining the maximum likelihood estimator (MLE) for θ involves maximizing $\ell(\theta; Y)$ subject to the quadratic constraint (6). One possibility is to take a Lagrange multiplier approach and maximize $\ell(\theta; Y) - \lambda(a^{\mathrm{T}}Aa - 1)$ with respect to (θ, λ) via standard optimization techniques. However, this introduces yet another parameter, which may contribute to numerical instability.

Alternatively, we have found that the following reparameterization leads to quite stable implementation in practice. Because A in (6) is positive definite, there exists B positive definite such that $A = B^2$. With $c = Ba$ $(d \times 1)$, (6) becomes $c^{\mathrm{T}}c = 1$. Note that $-c$ (and hence $-a$) yields an identical density for z; thus, c must lie on a half-unit sphere in $\Re^d$. Hence, $c = (c_1, \ldots, c_d)^{\mathrm{T}}$ can be represented using the polar coordinate transformation

$$\begin{aligned} c_1 &= \sin(\phi_1), \\ c_2 &= \cos(\phi_1)\sin(\phi_2), \\ &\vdots \\ c_{d-1} &= \cos(\phi_1)\cos(\phi_2)\cdots\cos(\phi_{d-2})\sin(\phi_{d-1}), \\ c_d &= \cos(\phi_1)\cos(\phi_2)\cdots\cos(\phi_{d-1}), \end{aligned}$$

where $-\pi/2 < \phi_t \leq \pi/2$ for $t = 1, 2, \ldots, d-1$ to guarantee that c can take all values on a half-unit sphere in $\Re^d$. Letting $\phi = (\phi_1, \ldots, \phi_{d-1})^{\mathrm{T}}$, the parameter of interest for fixed K is $\theta = (\beta^{\mathrm{T}}, \mu^{\mathrm{T}}, \phi^{\mathrm{T}}, r^{\mathrm{T}}, \sigma)^{\mathrm{T}}$, where θ now denotes this vector with dimension one less than before. With this parameterization and given K, there is no constraint, and standard optimization techniques may be used. We discuss determination of K in the next section.

In practice, it is crucial to have a good initial value for θ and especially ϕ for any optimization approach. Such a value may be obtained by maximizing a penalized log likelihood for

θ in the original parameterization such as $\ell_p(\theta;Y) = \ell(\theta;Y) - Q(a^{\mathrm{T}}Aa - 1)^2$, where Q is a positive constant such as $Q = N = \Sigma_{i=1}^{m} n_i$, the total number of observations. Once the corresponding value for θ is obtained, a may be transformed to ϕ and the other parameters, yielding a starting value under the new parameterization. Alternatively, when q and K are small, a grid search for ϕ may be carried out over $(-\pi/2, \pi/2]$ and the log likelihood maximized in the remaining parameters. As we will demonstrate, K required to provide suitable flexibility is small.

Once the MLE $\hat{\theta}$ for the particular given K is obtained, the inverse of the observed information may be used to construct estimates of uncertainty for functions of model parameters in the usual way. Population inference proceeds by identifying fixed effects of interest within the parameterization of the model, as we illustrate for the cholesterol data in Section 4. As is standard, inference on individuals may be based on a so-called empirical Bayes approach, where individual posterior modes for b_i are estimated by finding the maximizer $\hat{Z}_i$ of the posterior density $f(Z_i \mid Y_i;\theta) \propto f(Y_i \mid Z_i;\theta)P_K^2(Z_i)\varphi(Z_i)$ with $\theta = \hat{\theta}$ and calculating $\hat{b}_i = \hat{\mu} + \hat{R}\hat{Z}_i$. Alternatively, Z_i may be estimated by $\mathrm{E}(Z_i \mid Y_i)$ evaluated at $\hat{\theta}$, which has a straightforward closed-form expression (see the Appendix).

3.3 *Choosing the Tuning Parameter K*

The previous procedure is based on a given value for the tuning parameter K controlling the flexibility of representation of the random effects density. Although fixing a choice of K yields a representation flexible enough to approximate a wide variety of densities, including the normal, an objective selection method is preferable. Following other authors (e.g., Davidian and Gallant, 1993), we propose to select K by inspection of information criteria evaluated at $\hat{\theta}$ over a series of fits for different given K, including $K = 0$ (normality). These criteria all take the form of a penalized log likelihood $-\ell(\theta;Y)/N + C(N)(p_{\mathrm{net}}/N)$, where p_{net} is the number of (free) parameters in the model excluding K: the Akaike Information Criterion (AIC) with $C(N) = 1$, the Schwarz Information Criterion (BIC) with $C(N) = 0.5 \log N$, and the Hannan–Quinn criterion (HQ) with $C(N) = \log\log N$. For a given criterion, K minimizing the penalized log likelihood is preferred; AIC tends to prefer larger models, while BIC prefers smaller ones, with HQ intermediate. Davidian and Gallant (1993) advocate inspection of the estimated density for K selected for each criterion, from which a visual selection may be made, and propose HQ in the event an automatic rule is desired. We evaluate the objective performance of these criteria via simulation in Section 5.

4. Application

We illustrate the proposed methods by applying them to the Framingham cholesterol data introduced in Section 1. Figure 1a and preliminary analysis indicate that, although baseline cholesterol may depend on individual subject characteristics, the linear trend over time is somewhat similar across participants regardless of age or gender. Moreover, the residuals from the individual fits in Section 1 exhibit no particular pattern over time. Based on these observations, we consider the semiparametric linear mixed model

$$Y_{ij} = b_{0i} + \beta_1 \mathrm{age}_i + \beta_2 \mathrm{sex}_i + b_{1i} t_{ij} + \epsilon_{ij}. \qquad (9)$$

Here, Y_{ij} is cholesterol level divided by 100 at the jth time for subject i and t_{ij} is (time−5)/10, with time measured in years from baseline, where transformations of level and time were adopted for reasons of numerical stability; ϵ_{ij} are independent $\mathrm{N}(0, \sigma^2)$; age_i is age at baseline; sex_i is a gender indicator (0 = female, 1 = male); and $b_i = (b_{0i}, b_{1i})^{\mathrm{T}} = \mu + RZ_i$, where $\mu = (\mu_0, \mu_1)^{\mathrm{T}}$, $\mathrm{vech}(R) = (r_{11}, r_{21}, r_{22})^{\mathrm{T}}$, and $Z_i = (Z_{0i}, Z_{1i})^{\mathrm{T}}$

Table 1
Fits of model (9), (10) *to the cholesterol data for* $K = 0, 1, 2$; d_{11}, d_{21}, *and* d_{22} *are the distinct elements of the matrix* D

	$K = 0$		$K = 1$		$K = 2$	
Parameter	Estimate	SE	Estimate	SE	Estimate	SE
β_1 (age)	0.0184	0.0035	0.0156	0.0032	0.0161	0.0030
β_2 (sex)	−0.0630	0.0554	−0.0626	0.0455	−0.0765	0.0442
γ_0 (intercept)	1.5969	0.1503	1.7131	0.1389	1.7026	0.1314
γ_1 (time)	0.2817	0.0241	0.2817	0.0242	0.2821	0.0242
σ	0.2084	0.0057	0.2081	0.0055	0.2079	0.0055
d_{11}	0.1412	0.0153	0.1401	0.0165	0.1402	0.0179
d_{21}	0.0314	0.0100	0.0294	0.0098	0.0297	0.0105
d_{22}	0.0380	0.0116	0.0392	0.0109	0.0395	0.0112
μ_0	1.5969	0.1503	1.9110	0.1384	1.5987	0.1462
μ_1	0.2817	0.0241	0.2529	0.0316	0.0870	0.0720
r_{11}	0.3758	0.0203	0.3178	0.0198	0.3583	0.0264
r_{21}	0.0836	0.0253	0.1103	0.0232	0.1378	0.0280
r_{22}	0.1762	0.0313	0.1618	0.0250	0.1616	0.0253
ϕ_1	—	—	0.5240	0.0727	0.5844	0.2076
ϕ_2	—	—	−0.8021	0.1069	−0.2381	0.1266
ϕ_3	—	—	—	—	0.9290	0.1513
ϕ_4	—	—	—	—	−0.3755	0.1357
ϕ_5	—	—	—	—	0.5187	0.2792

Table 2
Model selection criteria for $K = 0, 1, 2$ for the cholesterol data; smaller values are preferred

Criterion	$K = 0$	$K = 1$	$K = 2$
−Log likelihood	160.9864	148.5971	146.8909
AIC	0.1619	0.1519	0.1532
BIC	0.1808	0.1756	0.1840
HQ	0.1691	0.1609	0.1648

has density (3) for the choices of K discussed below. Thus, $x_{ij} = (\text{age}_i, \text{sex}_i)^{\mathrm{T}}$, and $s_{ij} = (1, t_{ij})^{\mathrm{T}}$. Model (9) is a simple linear random coefficient model with baseline effects of age and sex; from the construction of (1), we may rewrite the model in the familiar form

$$Y_{ij} = \gamma_0 + \beta_1 \text{age}_i + \beta_2 \text{sex}_i + \gamma_1 t_{ij} + u_{0i} + u_{1i} t_{ij} + \epsilon_{ij}, \quad (10)$$

where $\gamma_0 = \mathrm{E}(b_{0i}) = \mu_0 + r_{11}\mathrm{E}(Z_{0i})$, $\gamma_1 = \mathrm{E}(b_{1i}) = \mu_1 + r_{21}\mathrm{E}(Z_{0i}) + r_{22}\mathrm{E}(Z_{1i})$, and $u_i = (u_{0i}, u_{1i})^{\mathrm{T}}$ has mean zero and covariance $D = \mathrm{var}(b_i) = R\,\mathrm{var}(Z_i)R^{\mathrm{T}}$. Of course, when $K = 0$, $\mathrm{E}(Z_i) = 0$ and $\mathrm{var}(Z_i) = I$.

Table 1 presents the results of fits for $K = 0, 1, 2$. The estimates and standard errors are similar for all three models. All information criteria given in Table 2 prefer the model with $K = 1$, supporting the contention of a departure from normality. Figure 2a depicts the estimated bivariate density of the random effects b_i for this preferred fit, which shows the presence of a second mode or bump. Figure 2b provides another perspective on this feature; following the shape of the density in Figure 1a, the subject-specific empirical Bayes estimates of the b_i clump into two distinct groups. Figure 2c and 2d show the estimated marginal densities of b_{0i} and b_{1i} and offer support for the informal observations given in Section 1: the distribution of slopes appears normal while the shape of the density for intercepts shows evidence of skewness, as in Figure 1b. As (9) includes baseline age and gender effects, the estimated density suggests that the apparent nonnormal pattern for intercepts reflects something other than heterogeneity due to these characteristics.

Taken as a whole, these results suggest the possibility of a subpopulation of individuals with higher baseline cholesterol, even after adjusting for the effects of age and gender. A potential explanation is that an important, unavailable covariate has failed to be taken into account. In any event, the ability to estimate the random effects density gives the data analyst considerable insight and raises issues for further investigation. Moreover, the form of the estimate suggests that inference on individual effects under the usual normality assumption could be misleading. Note that inspection of marginal densities of random effects only would not reveal the extent and form of the apparent departure from normality; with $q = 2$, it is possible to depict the joint density as in Figure 2a, allowing full view of the bimodality of the estimate.

5. Simulation Results

We conducted simulation studies to investigate the performance of the proposed methods. We report here on results for the model

$$Y_{ij} = t_{ij}\beta_1 + w_i\beta_2 + b_i + \epsilon_{ij}, \quad i = 1, \ldots, m = 100,\ j = 1, \ldots, 5, \quad (11)$$

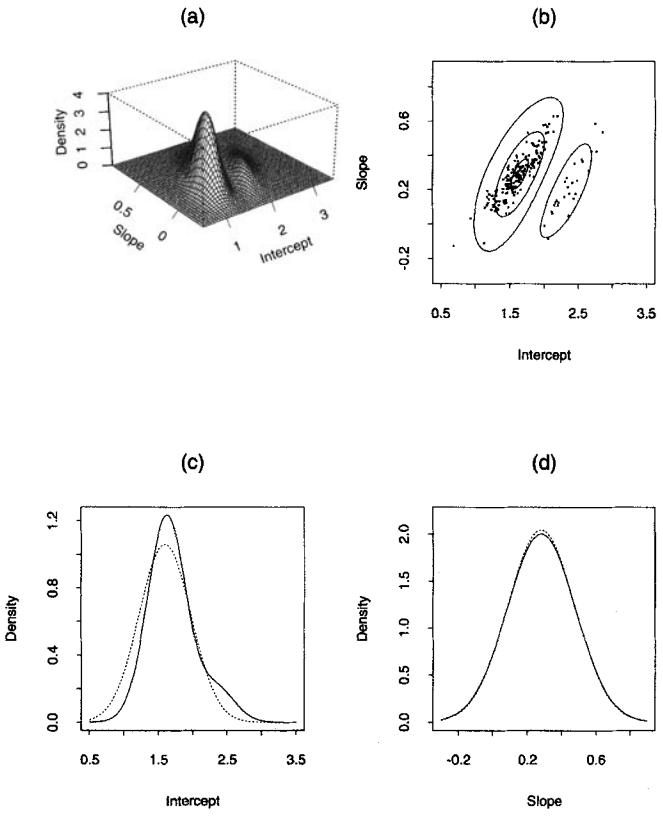

Figure 2. Fit of model (1) to the cholesterol data with $K = 1$. **a.** Estimated density of b_i. **b.** Contour plot of density in **a** with subject-specific estimated posterior modes for b_i superimposed (contours are 10, 50, and 90%; some extreme subjects lie outside the figure). **c** and **d.** Corresponding estimated marginal densities for components of b_i (solid) with normal with same moments superimposed (dashed).

where $t_{ij} = j - 3$, $w_i = 1$ if $i \leq 50$ and is zero otherwise, $\beta_1 = 2$, $\beta_2 = 1$, $\epsilon_{ij} \sim \mathrm{N}(0, 0.5^2)$, and the true distribution of b_i is the mixture of normals $(2\pi)^{-1/2}[0.7\exp\{-(x+3)^2/2\} + 0.3\exp\{-(x-2)^2/2\}]$. Under this specification, the true intercept is $\mathrm{E}(b_i) = -1.5$, and $\mathrm{var}(b_i) = 6.25$. Note that t_{ij} represents a covariate with values changing within individuals and the same for all individuals, while w_i is an individual-level covariate. The choice $m = 100$ corresponds to a situation where the amount of information available to estimate both fixed model parameters and the density may not seem great. For each of 100 Monte Carlo data sets, (11) was fit three times under the semiparametric assumptions of Section 2, with the density of b_i represented by the SNP approximation with $K = 0, 1$, and 2, and the three information criteria in Section 3.3 were calculated for each fit. Preliminary study revealed that richer models ($K > 2$) were never selected by the information criteria. To evaluate the objective use of the criteria, the fit preferred by each of AIC, BIC, and HQ was recorded.

None of AIC, BIC, or HQ selected the normal specification ($K = 0$) for any of the 100 data sets, demonstrating the ability of these selection methods to detect an obvious departure

Table 3

Simulation results, 100 data sets: MC ave. and MC SD are average and standard deviation of the estimates, respectively; Ave. SE is average of estimated standard errors; RE is Monte Carlo mean square error for the indicated fit divided by that for $K = 0$; true values of parameters are in parentheses

	$K = 0$			Preferred by BIC				Preferred by HQ			
	MC ave.	MC SD	Ave. SE	MC ave.	MC SD	Ave. SE	RE	MC ave.	MC SD	Ave. SE	RE
(a) Mixture Scenario											
β_1 (2)	2.000	0.017	0.016	2.000	0.017	0.016	1.00	2.000	0.017	0.016	1.00
β_2 (1)	1.158	0.472	0.493	1.034	0.234	0.209	0.21	1.028	0.230	0.208	0.23
E(b) (−1.5)	−1.614	0.369	0.349	−1.552	0.275	0.269	0.52	−1.549	0.273	0.269	0.52
var(b) (6.25)	6.045	0.638	0.862	6.098	0.654	0.690	1.01	6.099	0.655	0.695	1.00
σ (0.5)	0.498	0.018	0.018	0.498	0.018	0.018	1.00	0.498	0.018	0.018	1.00
(b) Normal Scenario											
β_1 (2)	2.000	0.017	0.016	2.000	0.017	0.016	1.00	2.000	0.017	0.016	1.00
β_2 (1)	0.994	0.512	0.489	0.987	0.533	0.487	1.17	0.990	0.550	0.479	1.08
E(b) (−1.5)	−1.491	0.363	0.346	−1.487	0.373	0.345	1.09	−1.489	0.380	0.343	1.05
var(b) (6.25)	5.955	0.789	0.849	5.957	0.790	0.861	1.00	5.958	0.790	0.863	1.00
σ (0.5)	0.498	0.018	0.018	0.498	0.018	0.018	1.00	0.498	0.018	0.018	1.00

from normality. Otherwise, 35% of the time, the AIC selected $K = 1$, and 65% of the time, it selected $K = 2$; these percentages were 76 and 24% for BIC and 56 and 44% for HQ, respectively, demonstrating the tendency of AIC (BIC) to prefer larger (smaller) models, with HQ intermediate. The first part of Table 3, part a, shows results for all 100 data sets using $K = 0$ (assuming normality), and the remaining two sections give summaries using for each of the 100 data sets the choice of K preferred by BIC and HQ, respectively, thus representing performance if these criteria were used as automatic selection rules. Results for AIC are similar and are excluded for brevity. Parameter estimates are, for the most part, unbiased in all cases. The efficiency of estimation under the incorrect normality assumption for the intercept and treatment effect β_2 is quite poor relative to that when the density is estimated. Because a main focus of such an analysis may well be evaluation of treatment effect, this suggests that adopting the normality assumption routinely may lead to inefficient inferences on fixed effects of primary interest. In contrast, inferences on β_1 corresponding to the within-individual time effect and on the variance components are unaffected. Because β_2 is associated with a subject-level covariate and random effects also represent subject-level heterogeneity, it is perhaps not surprising that failure to characterize the latter correctly would impact the former, while β_1 is associated with a within-individual effect that is in some sense orthogonal to intersubject differences. Similar behavior was observed by Tao et al. (1999). The variance σ^2 also represents an intrasubject effect. Because all approaches attempt to estimate var(b_i) from the apparent intersubject variation, the point estimate is similarly unaffected; however, estimation of E(b_i) is compromised when normality is incorrectly assumed, likely due to association with estimation of β_2. Note that, for the cholesterol data (see Table 1), the estimate of the time effect is virtually unchanged across fits while those for the subject-level age and sex effects and the intercept change noticeably, potentially reflecting the phenomenon observed here.

The advantage of estimating the random effects density may be appreciated from Figure 3. Figure 3a shows the Monte Carlo average of estimated densities over the 100 data sets along with the true density for $K = 0$ (the normal fits) and those preferred under BIC and HQ, as in Table 3; that for the AIC fits is virtually identical. The figure demonstrates that the additional flexibility afforded by the SNP representation is sufficient to capture quite accurately the true underlying features of the random effects, even with only 100 subjects. This observation is further supported by Figure 3b, which shows the 100 estimates from the fits preferred by HQ. Given the sample size, it is not unreasonable that allowing the additional flexibility to represent more complex densities would result in occasional overmodeling; however, note that only 3 of the 100 fits preferred by HQ includes a third, spurious mode. In practice, one would likely combine visual inspection of the estimate with the information criteria to select a feasible model, focusing on the dominant features, and take care not to overinterpret such behavior.

To gauge performance under the opposite situation, we conducted a simulation under the same scenario as in (11) but with the true random effects distribution as $b_i \sim \mathrm{N}(-1.5, 6.25)$. Eighty-four, 89, and 97% of the time the AIC, HQ, and BIC criteria, respectively, correctly selected $K = 0$. AIC selected $K = 1$ and $K = 2$ for 7 and 9% of the data sets; these figures were 5 and 6% for HQ and 3 and 0% for BIC. In those data sets where $K > 0$ was chosen, examination of the estimated densities under the preferred choice of K reveals that only three are bimodal. An important feature of the SNP density is that larger K does not necessarily imply a greater number of modes but may just be a less parsimonious representation. Moreover, K functions strictly as a tuning parameter and should not be interpreted as reflecting underlying features such as number of subpopulations.

Summaries of the Monte Carlo results are given in Table 3, part b. The inefficiency of the fits preferred by BIC and HQ is obviously due to the few data sets where $K > 0$ was preferred. For the automatic use of HQ as a selection rule,

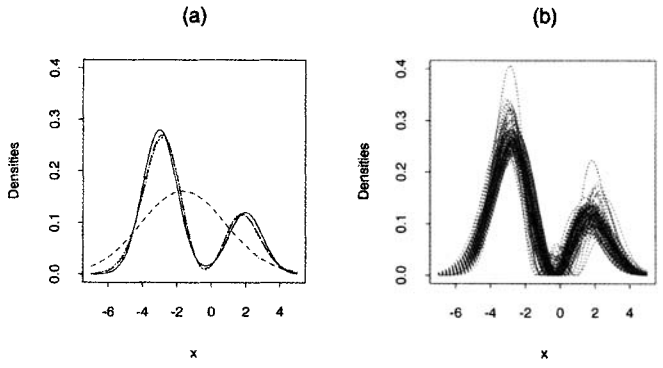

Figure 3. Simulation results based on 100 datasets. **a.** True density (solid line) and Monte Carlo average estimated densities for 100 data sets using $K = 0$ (short dashed line; normal) and using the fits preferred by BIC (dotted line) and HQ (long dashed line). **b.** Estimated densities for the fits preferred by HQ for the 100 data sets.

note that this results in only minimal loss of efficiency relative to the correct $K = 0$ model. The apparent conclusion is that the price to pay for estimating the random effects density when the normality assumption holds is mild; similar results are reported in Hu, Tsiatis, and Davidian (1998).

6. Discussion

We have proposed an approach to a semiparametric linear mixed model where the random effects are assumed to have a smooth density in which the form of the random effects density is represented by the SNP truncated series expansion. As we have demonstrated, the expression for the SNP density allows the marginal likelihood of the data to be written in a closed form. Moreover, we have also proposed a new parameterization of this density representation that imposes identifiability constraints in a straightforward manner, depending on quantities (moments of normal distributions) that may be computed efficiently and is attractive for stable computation. Standard optimization techniques may be used to estimate jointly the fixed model parameters and the density. The degree of flexibility of the representation is controlled by a scalar tuning parameter, and the representation admits the usual normal model as a special case. Use of standard information criteria to select the tuning parameter, along with visual inspection of the estimated density, has worked well in practice to provide reliable estimates of the density.

An alternative to the SNP approach is to represent the random effects density by a mixture of normals (Verbeke and Lesaffre, 1996). Although these authors use an EM algorithm for implementation, it is worth noting that the marginal likelihood may also be expressed in closed form and the number of normal mixtures plays the role of a tuning parameter. This approach requires that the obvious constraints be imposed on the mixing probabilities in order that standard optimization techniques be used.

As observed in other contexts and approaches (e.g., Hu et al., 1998; Tao et al., 1999), there is potential to gain efficiency in estimating certain parameters when the normality assumption does not hold, with only a small price to pay for the extra complication of estimating the density when it is normal. An additional major advantage of all approaches that relax the assumption on the random effects density is the insight the estimate provides.

When n_i is large, one may wish to relax the assumption on ϵ_{ij} in (1) to include a subject-specific, mean-zero Gaussian process $W_i(t)$ to model underlying biological behavior, as in Zhang et al. (1998), so that the model becomes

$$Y_{ij} = x_{ij}^{\mathrm{T}}\beta + s_{ij}^{\mathrm{T}}b_i + W_i(t_{ij}) + e_{ij},$$

where $e_{ij} \sim \mathrm{N}(0, \sigma^2)$ is measurement error. This modification is accommodated straightforwardly by the proposed approach; in this case, the likelihood function will still have a closed form similar to (8).

We have implemented the approach using SAS `proc iml` using the `nlpqn` optimizer (SAS Institute, 1989); code is available from the authors on request.

Acknowledgements

This work was supported by NIH grants R01-CA85848 and R01-AI31789. The authors are grateful to the associate editor for suggestions that improved the presentation.

Résumé

La normalité des effets aléatoires est une hypothèse de routine dans les modèles mixtes linéaires; mais elle peut s'avérer irréaliste, masquant alors des caractéristiques importantes de la variabilité inter-individuelle. Nous assouplissons cette hypothèse en approchant la distribution des effets aléatoires par la représentation semi-nonparamétrique de Gallant et Nychka (1987) dépendant d'un paramètre d'ajustement choisi par l'utilisateur, qui inclut la distribution normale comme un cas particulier et permet une plus grande flexibilité en incorporant un large éventail de distributions non-normales. Un avantage est que la vraisemblance marginale peut s'exprimer de manière explicite, si bien que l'inférence peut être conduite par des techniques classiques d'optimisation. Nous démontrons que des mesures standards de l'information peuvent servir à choisir le paramètre d'ajustement et détecter des écarts à la normalité. Nous illustrons cette approche par des simulations et à partir des données longitudinales de l'étude Framingham.

References

Aitken, M. (1999). A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics* **55,** 117–128.

Butler, S. M. and Louis, T. A. (1992). Random effects models with nonparametric priors. *Statistics in Medicine* **11,** 1981–2000.

Davidian, M. and Gallant, A. R. (1993). The nonlinear mixed effects model with a smooth random effects density. *Biometrika* **80,** 475–488.

Davidian, M. and Giltinan, D. M. (1995). *Nonlinear Models for Repeated Measurement Data.* London: Chapman and Hall.

Gallant, A. R. and Nychka, D. W. (1987). Seminonparametric maximum likelihood estimation. *Econometrica* **55,** 363–390.

Hu, P., Tsiatis, A. A., and Davidian, M. (1998). Estimating the parameters in the Cox model when covariate variables are measured with error. *Biometrics* **54,** 1407–1419.

Johnson, N. L. and Kotz, S. (1994). *Continuous Univariate Distributions,* Volume 1, 2nd edition. New York: John Wiley and Sons.

Laird, N. M. (1978). Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association* **73,** 805–811.

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

Littell, R. C., Milliken, G. A., Stroup, W. W., and Wolfinger, R. D. (1996). *SAS System for Mixed Models.* Cary, North Carolina: SAS Institute.

Madger, L. S. and Zeger, S. L. (1996). A smooth nonparametric estimate of a mixing distribution using mixtures of Gaussians. *Journal of the American Statistical Association* **91,** 1141–1151.

SAS Institute. (1989) *SAS/IML Software: Usage and Reference,* Version 6, 1st edition. Cary, North Carolina: SAS Institute.

Tao, H., Palta, M., Yandell, B. S., and Newton, M. A. (1999). An estimation method for the semiparametric mixed effects model. *Biometrics* **55,** 102–110.

Verbeke, G. and Lesaffre, E. (1996). A linear mixed-effects model with heterogeneity in the random-effects population. *Journal of the American Statistical Association* **91,** 217–221.

Verbeke, G. and Lesaffre, E. (1997). The effect of misspecifying the random effects distribution in linear mixed models for longitudinal data. *Computational Statistics and Data Analysis* **23,** 541–556.

Verbeke, G. and Molenberghs, G. (2000). *Linear Mixed Models for Longitudinal Data.* New York: Springer.

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.

Received February 2001. *Revised March* 2001. *Accepted March* 2001.

Appendix

To exhibit the required calculations discussed in Section 3.1, we consider $q = 2$. To compute the second term in (8), we need to calculate expectations of the form $\mathrm{E}(Z_1^{\alpha_1} Z_2^{\alpha_2})$, where expectation is with respect to the conditional distribution of $(Z_1, Z_2)^{\mathrm{T}}$ given Y_i, $\mathrm{N}(\mu_i, \Sigma_i)$; α_1 and α_2 are nonnegative integers; and μ_i and Σ_i are given in Section 3.1. This is feasible through use of the moment generating function $m(t) = \exp(\mu_i^{\mathrm{T}} t + t^{\mathrm{T}} \Sigma_i t/2)$, where $t = (t_1, t_2)^{\mathrm{T}}$. The function $m(t)$ may be written as $m(t) = \Sigma_{n=0}^{\infty} a_n t_1^n \, \Sigma_{n=0}^{\infty} b_n t_2^n \, \Sigma_{n=0}^{\infty} c_n t_1^n t_2^n$. Thus, $\mathrm{E}(Z_1^{\alpha_1} Z_2^{\alpha_2}) = \alpha_1! \alpha_2! \, \Sigma_{i=0}^{\min(\alpha_1, \alpha_2)} a_{\alpha_1 - i} b_{\alpha_2 - i} c_i$, where $\{a_n\}$, $\{b_n\}$, and $\{c_n\}$ are determined by μ_i and Σ_i.

In simple cases, the form of the ith expectation in the second term in (8) is easy to express. For example, when $q = 2$ and $K = 1$, this is given by $a_{00}^2 + 2a_{00}a_{10}\mu_{1i} + 2a_{00}a_{01}\mu_{2i} + a_{10}^2(\mu_{1i}^2 + \sigma_{11i}) + 2a_{10}a_{01}\sigma_{12i} + a_{01}^2(\mu_{2i}^2 + \sigma_{22i})$, where μ_{1i}, μ_{2i}, σ_{11i}, σ_{12i}, and σ_{22i} are the obvious elements of μ_i and Σ_i.

These same calculations may be used to evaluate $\mathrm{E}(Z_i \mid Y_i) = \mathrm{E}_{Z_i|Y_i;\theta}\{Z_i P_K^2(Z_i)\} / \mathrm{E}_{Z_i|Y_i;\theta}\{P_K^2(Z_i)\}$, discussed in Section 3.2.