Biometrics 60, 945–953
December 2004

Smooth Random Effects Distribution in a Linear Mixed Model

Wendimagegn Ghidey,[1] Emmanuel Lesaffre,[1,*] and Paul Eilers[2]

[1]Biostatistical Centre, Catholic University of Leuven, Kapucynenvoer 35, B-3000 Leuven, Belgium
[2]Medical Statistics, University of Leiden, 2300 RC Leiden, The Netherlands
[*]*email:* emmanuel.lesaffre@med.kuleuven.ac.be

Summary. A linear mixed model with a smooth random effects density is proposed. A similar approach to *P*-spline smoothing of Eilers and Marx (1996, *Statistical Science* **11,** 89–121) is applied to yield a more flexible estimate of the random effects density. Our approach differs from theirs in that the *B*-spline basis functions are replaced by approximating Gaussian densities. Fitting the model involves maximizing a penalized marginal likelihood. The best penalty parameters minimize Akaike's Information Criterion employing Gray's (1992, *Journal of the American Statistical Association* **87,** 942–951) results. Although our method is applicable to any dimensions of the random effects structure, in this article the two-dimensional case is explored. Our methodology is conceptually simple, and it is relatively easy to fit in practice and is applied to the cholesterol data first analyzed by Zhang and Davidian (2001, *Biometrics* **57,** 795–802). A simulation study shows that our approach yields almost unbiased estimates of the regression and the smoothing parameters in small sample settings. Consistency of the estimates is shown in a particular case.

Key words: Penalized Gaussian mixture; *P*-splines; Smooth random effects distribution.

1. Introduction

In longitudinal and repeated measures studies, the pattern of change with respect to time is often modeled with the linear mixed model (Laird and Ware, 1982). This model contains fixed and random effects. The fixed effects are the population-averaged parameters. The random effects pertain to subject-specific parameters. It is classically assumed that the random effects as well as the measurement error term (within-subject variability) have a normal distribution.

Inference on the fixed effects has been found to be robust to nonnormality of the random effects (Butler and Louis, 1992; Verbeke and Lesaffre, 1997). For efficient estimation of the fixed effects and for unbiased model-based standard errors, however, selection of the correct random effects distribution could be important. Further, deviations from normality of the random effects distribution can have an important effect on inferences involving the random effects themselves. For instance, it has been shown that the empirical Bayes estimates of the random effects are distorted if normality does not hold (Verbeke and Lesaffre, 1996). Furthermore, it has been shown by Verbeke and Lesaffre (1996) that checking the normality assumption of the random effects distribution is hampered by the shrinkage of the empirical Bayes estimates.

We argue that the normality assumption may be too restrictive in practice to represent the actual between-subject distribution. Consequently, there is a need for a linear mixed model with a more flexible distributional assumption on the random effects. A popular and important technique that relaxes this assumption is given by nonparametric maximum likelihood estimation (Laird, 1978). In this case no parametric assumptions on the random effects distribution are made. But, the fact that the nonparametric maximum likelihood estimate (NPMLE) of this distribution is discrete has been criticized by some as being unrealistic. Other proposals in the literature are: smoothed nonparametric maximum likelihood estimation (Magder and Zeger, 1996), predictive recursive estimation (Tao et al., 1999), the heterogeneity linear mixed model (Verbeke and Lesaffre, 1996), the smoothing by roughening approach (Shen and Louis, 1999), and, quite recently, the semi-nonparametric (SNP) method of Zhang and Davidian (2001).

We propose an alternative and relatively simple method that originates from the *P*-spline smoothing approach of Eilers and Marx (1996). In Section 2, we introduce our penalized Gaussian linear mixed model that is based on penalized estimation of a marginal likelihood that will be discussed in Section 3. In Section 4, we prove the consistency of the regression parameters and the smoothing parameters in a particular case. In Section 5, we reanalyze the cholesterol data of Zhang and Davidian (2001). In Section 6, we compare our method to the method of Zhang and Davidian (2001) based on some simulated data. We end by some concluding remarks in Section 7.

2. The Penalized Gaussian Mixture Linear Mixed Model

Assume the classical linear mixed model (Laird and Ware, 1982),

$$\mathbf{Y}_i = X_i\boldsymbol{\beta} + Z_i\mathbf{b}_i + \boldsymbol{\epsilon}_i \quad (i = 1, \ldots, K), \tag{1}$$

where $\mathbf{Y}_i$ is an $n_i \times 1$ response vector, $\boldsymbol{\beta}$ is a $p \times 1$ vector of fixed effects, and $\mathbf{b}_i$ is a $d \times 1$ random effects vector

$\sim N(0,\, D)$, the measurement error vector $\boldsymbol{\epsilon}_i \sim N(0,\, \sigma^2 I_{n_i})$ independent of $\mathbf{b}_i$. Hence we assume independent error terms, but the approach can equally be applied to correlated errors.

The normality assumption of the random effects is relaxed to a more flexible and smooth density function. An approach similar to *P*-spline smoothing of Eilers and Marx (1996) is applied except that the *B*-splines base functions are replaced by their approximating Gaussian densities. Indeed, it has been shown that a (standardized) *B*-spline of degree q approximates a normal density as $q \to \infty$ (Unser, Aldroubi, and Eden, 1992).

Therefore, we suppose that the true distribution of the random effects is not a normal distribution anymore. However, for simplicity reasons, we assume that the true covariance matrix of the random effects is still denoted as D. Then, let the bivariate random effects in (1) be represented as $\mathbf{b}_i = R\mathbf{s}_i$, $i = 1, \ldots, K$, where R is a lower triangular matrix such that $RR^{\mathrm{T}} = D$ and $\mathbf{s}_i$ is the standardized form of $\mathbf{b}_i$. Hereby, we assume that the $\mathbf{s}_i$ extend over, say, the square $[-m,\, m] \times [-m,\, m]$ but vanish (in practical terms) outside. Observe that our method does not require to work with the standardized random effects. In fact, the choice of standardized random effects is inspired by computational arguments. Indeed, by using the $\mathbf{s}_i$ instead of the $\mathbf{b}_i$ the implementation of our method will become largely independent of the range of the random effects.

To construct the flexible random effects distribution, we take a grid of equally spaced points on the interval $[-m,\, m]$ in both dimensions (but not necessarily of the same size in each dimension). Let these grids be the means of the basis Gaussian densities, say μ_{1j}, $j = 1, \ldots, J$ for the first dimension and μ_{2l}, $l = 1, \ldots, L$ for the second dimension. Their standard deviations in the two dimensions, τ_1 and τ_2, are set to $\frac{2}{3}(\mu_{1j} - \mu_{1,j-1})$ and $\frac{2}{3}(\mu_{2l} - \mu_{2,l-1})$. This is based on the assumption that a Gaussian density which extends over $\mu \pm 3\tau$ can be approximated by a *B*-spline function of degree 3 which extends over 4 equidistant subintervals.

All together, the rectangular matrix of bivariate normal densities with means $\boldsymbol{\mu}_{jl} = (\mu_{1j},\, \mu_{2l})^{\mathrm{T}}$ and covariance matrix $D_s = \mathrm{diag}(\tau_1^2,\, \tau_2^2)$ form a two-dimensional basis for the (estimated) distribution of the standardized $\mathbf{s}_i$. On the original scale of the random effects, each of the basis Gaussian densities will have a mean vector $R\boldsymbol{\mu}_{jl}$ and a covariance matrix RD_sR^{T}.

Hence, our method assumes that the density for (a bivariate) $\mathbf{b}$ can be approximated by a mixture of these basis Gaussian densities as

$$f(\mathbf{b}) = \sum_{j=1}^{J}\sum_{l=1}^{L} c_{jl} N\big(R\boldsymbol{\mu}_{jl}, RD_sR^{\mathrm{T}}\big), \tag{2}$$

where $c_{jl} = \exp(a_{jl}) / \sum_{k=1}^{J}\sum_{m=1}^{L}\exp(a_{km})$ are (transformed) elements of a $J \times L$ matrix of coefficients with the property that $\sum_j \sum_l c_{jl} = 1$. This parameterization allows unconstrained maximization for the smoothing parameters and guarantees that all mixing coefficients c_{jl} are strictly positive. Alternative parameterizations are discussed in the last section. Observe also that when expression (2) represents the true model of the random effects, by definition its covariance matrix should be equal to D.

We call the resulting model, with the smoothed density for $\mathbf{b}$ given by (2) with the coefficients c_{jl} estimated by maximizing a penalized log likelihood (see next section), the *penalized Gaussian mixture* (*PGM*) *linear mixed model*.

It is important to note that there is a difference between our approach and that of a classical mixture model in the sense that here the means (and the variances) of the normal densities are fixed to a prespecified grid of values and that we do not estimate the number of components. Our approach is also different from classical density estimation, which would need here the latent locations of the random effects.

3. Estimation of Model Parameters

The parameters to be estimated include the fixed effects $\boldsymbol{\beta}$, $\boldsymbol{\sigma}_R = \mathrm{vec}(R)$ the stacked vector of unique elements of R, the error standard deviation (on the log scale) $\log(\sigma)$, and the vector of coefficients $\mathbf{a} = (a_{11},\, a_{12}, \ldots, a_{JL})^{\mathrm{T}}$. The total vector of parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}^{\mathrm{T}},\, \boldsymbol{\sigma}_R^{\mathrm{T}},\, \log(\sigma),\, \mathbf{a}^{\mathrm{T}})^{\mathrm{T}}$ is jointly estimated by maximizing a penalized marginal likelihood.

3.1 *The Penalized Log-Likelihood Function*

The conditional density remains in our model as $f(\mathbf{Y}_i \mid \mathbf{b}_i; \boldsymbol{\theta}) = N(X_i\boldsymbol{\beta} + Z_i\mathbf{b}_i, \sigma^2 I_{n_i})$. With the random effects density $f(\mathbf{b}_i; \boldsymbol{\theta})$ given by (2), the marginal density of $\mathbf{Y}_i$ becomes a mixture of normal densities

$$\begin{aligned} f(\mathbf{Y}_i; \boldsymbol{\theta}) &= \int f(\mathbf{Y}_i \mid \mathbf{b}; \boldsymbol{\theta}) f(\mathbf{b}; \boldsymbol{\theta})\, d\mathbf{b} \\ &= \sum_{j=1}^{J}\sum_{l=1}^{L} c_{jl} N\big(X_i\boldsymbol{\beta} + Z_iR\boldsymbol{\mu}_{jl},\, Z_iRD_sR^{\mathrm{T}}Z_i^{\mathrm{T}} + \sigma^2 I_{n_i}\big) \end{aligned} \tag{3}$$

and the marginal log likelihood becomes $\ell(\boldsymbol{\theta}; \mathbf{Y}) = \sum_{i=1}^{K} \log\{f(\mathbf{Y}_i; \boldsymbol{\theta})\}$.

When a fine grid of $J \times L$ density bases is involved, there will be overfitting resulting in a widely varying estimated bivariate surface for the random effects. Too few components lead to a relatively smooth but biasedly estimated distribution. We applied the approach by Eilers and Marx (1996) to find a compromise between smoothness and bias; i.e., take a relatively large number of basis Gaussian densities and penalize the log likelihood for overfitting with a penalty term based on finite differences of adjacent coefficients.

The penalized log-likelihood function is given by

$$\ell_p(\boldsymbol{\theta}; \mathbf{Y} \mid \boldsymbol{\lambda}) = \ell(\boldsymbol{\theta}; \mathbf{Y}) - \left[\frac{\lambda_1}{2}\sum_j\sum_l \big(\Delta_1^k a_{jl}\big)^2 + \frac{\lambda_2}{2}\sum_j\sum_l \big(\Delta_2^k a_{jl}\big)^2\right], \tag{4}$$

where Δ_i^k, $i = 1,\, 2$ is a difference operator of order k for the ith dimension, and $\boldsymbol{\lambda} = (\lambda_1,\, \lambda_2)^{\mathrm{T}}$ is a vector of penalty parameters one for each dimension.

Experience shows that the choices $k = 2$ or 3 give similar smooth density estimates. But we prefer $k = 3$ because for a large number of basis densities, the limit of the smooth distribution is a normal distribution as $\lambda \to \infty$ (Komarek, Lesaffre, and Hilton, unpublished manuscript). Observe that

the penalty term expressed in the a-coefficients will have the most effect in the areas where the latent random effects are sparse, since these will be modeled with c_{jl}'s close to zero (implying corresponding large negative a_{jl}'s). In the other areas, where there are relatively many data points, the effect of the penalty term will be less.

3.2 *Estimating* $\boldsymbol{\theta}$

For a given $\boldsymbol{\lambda}$ and a fixed basis of Gaussian densities, the penalized log-likelihood function (4) can be maximized with respect to $\boldsymbol{\theta}$ using a Newton–Raphson optimization algorithm.

Conditional maximization of the likelihood can be applied to $\boldsymbol{\theta}$, namely by swapping between $\boldsymbol{\eta} = (\boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\sigma}_R^{\mathrm{T}}, \log(\sigma))^{\mathrm{T}}$ and the vector of smoothing coefficients, i.e., $\mathbf{a}^{\mathrm{T}}$. Starting values for $\boldsymbol{\eta}$ can be obtained from the fit of a classical Gaussian linear mixed model (e.g., SAS `PROC MIXED` output). First, conditioned on $\boldsymbol{\eta}$, ℓ_p is maximized with respect to $\mathbf{a}$ and at the next step the roles of $\boldsymbol{\eta}$ and $\mathbf{a}$ are interchanged. Newton–Raphson can then be applied in each conditional maximization step. To guarantee identifiability, one of the a_{jl}'s is kept fixed (to 0) and further the constraint $\int \mathbf{b} f(\mathbf{b})\, d\mathbf{b} = 0$ needs to be imposed on the maximization algorithm. For simplicity reasons, we implemented a slightly different maximization routine. First, we kept the elements of $R(\boldsymbol{\sigma}_R)$ fixed at $\hat{R}^0$ (SAS `PROC MIXED` output) and updated the matrix at the end of the maximization. Namely, the final estimate, $\hat{R}$, was obtained by $\hat{R} = \mathrm{chol}(\mathrm{var}(\mathbf{b} \mid \hat{c}_{jl}, \hat{R}^0))$, where

$$\begin{aligned}\mathrm{var}\big(\mathbf{b} \mid \hat{c}_{jl}, \hat{R}^0\big) &= \hat{R}^0 \Bigg\{ \sum_{j=1}^{J}\sum_{l=1}^{L} \hat{c}_{jl}\mu_{jl}\mu_{jl}^{\mathrm{T}} + D_s + \left(\sum_{j=1}^{J}\sum_{l=1}^{L} \hat{c}_{jl}\mu_{jl}\right) \\ &\quad \times \left(\sum_{j=1}^{J}\sum_{l=1}^{L} \hat{c}_{jl}\mu_{jl}\right)^{\mathrm{T}} \Bigg\} \hat{R}^{0\mathrm{T}}. \end{aligned} \tag{5}$$

Although the estimate $\hat{R}$ will not be as efficient as the estimate from the maximization routine described above, our method avoids taking derivatives with respect to covariance elements. Further, our method was inspired by the results of Verbeke and Lesaffre (1997). Second, we avoided explicitly dealing with the constraint by fixing the part of $\boldsymbol{\beta}$, say $\boldsymbol{\beta}_*$, corresponding to the covariates in Z to the initial estimates $\hat{\boldsymbol{\beta}}_*^0$ (SAS `PROC MIXED` output) and updating the final estimate as $\hat{\boldsymbol{\beta}}_* = \hat{\boldsymbol{\beta}}_*^0 + \sum_{j=1}^{J}\sum_{l=1}^{L} \hat{c}_{jl}\hat{R}^0\mu_{jl}$.

3.3 *Optimal Smoothing and Standard Error of Parameter Estimates*

To find the optimal penalty coefficient $\boldsymbol{\lambda}$ we opted for the Akaike's Information Criterion (AIC). For a given $\boldsymbol{\lambda}$, $\mathrm{AIC}(\boldsymbol{\lambda}) = -2\ell_p(\boldsymbol{\theta}; \mathbf{Y} \mid \boldsymbol{\lambda}) + 2\dim(\boldsymbol{\theta} \mid \boldsymbol{\lambda})$, where $\dim(\boldsymbol{\theta} \mid \boldsymbol{\lambda})$ is the effective degrees of freedom depending on $\boldsymbol{\lambda}$. We applied a method by Gray (1992) for determining $\dim(\boldsymbol{\theta} \mid \boldsymbol{\lambda})$ as $\mathrm{trace}[(\mathbf{H}^{-1}(\boldsymbol{\theta}, \boldsymbol{\lambda})\, \mathbf{I}(\boldsymbol{\theta})]$, where $\mathbf{H}(\boldsymbol{\theta}, \lambda) = -\partial^2\ell_p(\boldsymbol{\theta}; \mathbf{Y} \mid \boldsymbol{\lambda})/\partial\boldsymbol{\theta}^2$, and $\mathbf{I}(\boldsymbol{\theta}) = -\partial^2\ell(\boldsymbol{\theta}; \mathbf{Y})/\partial\boldsymbol{\theta}^2$ are the observed Fisher information matrices based on the penalized and unpenalized likelihood, respectively. The optimal penalty vector $\boldsymbol{\lambda}$ minimizes $\mathrm{AIC}(\boldsymbol{\lambda})$.

Gray (1992) suggests to use as estimate for the asymptotic covariance matrix for the parameter estimates $\hat{\boldsymbol{\theta}}$ the matrix $\hat{\mathbf{V}} = \mathbf{H}^{-1}(\hat{\boldsymbol{\theta}}, \boldsymbol{\lambda})\mathbf{I}(\hat{\boldsymbol{\theta}})\mathbf{H}^{-1}(\hat{\boldsymbol{\theta}}, \boldsymbol{\lambda})$.

However, in small samples this estimate was often not positive definite in our case. On the other hand Verweij and Van Houwelingen (1994) suggest to use $\hat{\mathbf{V}} = \mathbf{H}^{-1}(\hat{\boldsymbol{\theta}}, \boldsymbol{\lambda})$ as estimate. They called this matrix a "pseudo-covariance matrix" arguing that penalized estimates are typically biased. Our asymptotic results (see Section 4) suggest that we will get asymptotically unbiased (or close to) estimates of our parameters, at least in a particular case. Further, simulations (results not shown) confirm that the proposal of Verweij and Van Houwelingen (1994) is close to the true covariance matrix.

4. Statistical Properties of the PGM Approach

In the Appendix we indicate that using arguments of Verbeke and Lesaffre (1997), the PGM linear mixed model delivers consistent estimates of the regression parameters. Further we show that in a particular case with the grid of Gaussian means chosen correctly, the weights c_{jl} are consistently estimated. Furthermore, in the more general case, we roughly indicate that when λ increases with K in a modest way, the penalized estimates minimize the Kullback–Leibler distance between the true random effects distribution and the assumed mixture of basis Gaussian densities.

Table 1

Estimated parameters (standard error in brackets) by the Gaussian, the Zhang and Davidian method (2001), and the PGM linear mixed models (LMM) fitted to the cholesterol data. L1, L2, and L3 are the elements of the Cholesky decomposition matrix R.

	Models		
Parameter	Gaussian LMM	Zhang and Davidian LMM	PGM LMM
β_0(intercept)	1.5969 (0.1503)	1.7131 (0.1389)	1.8326 (0.1009)
β_1(age)	0.0184 (0.0035)	0.0156 (0.0032)	0.0128 (0.0023)
β_2(sex)	−0.0630 (0.0554)	−0.0626 (0.0455)	−0.0634 (0.0477)
β_3(time)	0.2817 (0.0241)	0.2817 (0.0242)	0.2816 (0.0230)
σ	0.2084 (0.0057)	0.2081 (0.0055)	0.2077 (0.0054)
$L1$	0.3758	0.3178	0.3778
$L2$	0.0836	0.1103	0.0587
$L3$	0.1762	0.1618	0.1794
Log likelihood	−160.99	−148.60	−146.58

5. Application

5.1 *Cholesterol Example*

Zhang and Davidian (2001) illustrated their method on repeated cholesterol data from a sample of 200 subjects of the Framingham study. The cholesterol levels were measured at the beginning of the study and then every 2 years for 10 years. They fitted a semiparametric linear mixed model to the cholesterol level with age at baseline and gender as fixed effects and with a random intercept and slope,

$$Y_{ij} = \beta_0 + \beta_1 \text{age}_i + \beta_2 \text{sex}_i + \beta_3 t_{ij} + b_{0i} + b_{1i} t_{ij} + \epsilon_{ij},$$

whereby the random effects were assumed to have a smooth bivariate distribution. For numerical stability the cholesterol level for the ith subject at the jth time point, Y_{ij}, was divided by 100 and t_{ij} was taken as (time $-$ 5)/10 whereby the time is measured in years from baseline, age_i is the age at baseline for the ith subject, and $\text{sex}_i = 1$ for male and 0 otherwise. The measurement error ϵ_{ij} is assumed to have a normal distribution $N(0, \sigma^2)$. Their analysis showed that there is a bump in the distribution of the random intercept suggesting a subpopulation with a higher baseline cholesterol on average. On the other hand the estimated random slope distribution was reasonably approximated by a normal distribution.

We fitted our PGM linear mixed model to these data with the same fixed and random effects structure. We show the results for $J = 18$ and $L = 10$, inspired by the results of Zhang and Davidian, i.e., that the distribution of the random intercept is a mixture of two densities and that the distribution of the random slope seems to be close to a normal. On the standardized scale, grids of points as the means of the basis Gaussian densities were defined on the square $[-4, 4] \times [-4, 4]$ and the common covariance matrix for each Gaussian density is $D_s = \text{diag}(0.098424, 0.35117)$. Third-order difference penalties were imposed in each dimension. Optimal values for the penalty parameters were found by minimizing AIC($\boldsymbol{\lambda}$) when varying λ_1 and λ_2 on a grid of values [0.1, 1, 2, 10, 100] and [1, 5, 10, 15, 100, 1000], respectively. The best values obtained are $\lambda_1 = 1$ and $\lambda_2 = 5$.

Table 1 presents the estimated fixed effects and covariance matrix parameters by our model, the classical linear mixed model, and the semiparametric model of Zhang and Davidian (2001). The table shows that the estimates of the three models are relatively close, with the PGM model solution having a covariance matrix for the random effects closer to the covariance matrix from the linear mixed model solution. It also shows that the standard errors of our parameter estimates are not much different from those of the other models.

In Figure 1 the estimated random effects distribution surface is depicted. The plots clearly show the deviation from bivariate normality in such a way that the random intercept distribution is a mixture of two distributions while the random slope is close to normal distribution. However, while our marginal distributions are close to the fitted distributions of Zhang and Davidian, the two bivariate fitted models differ somewhat in the tails. The scatter plot for empirical Bayes estimates also show the two clusters of observations, although not as clearly separable as with the result of Zhang and Davidian.

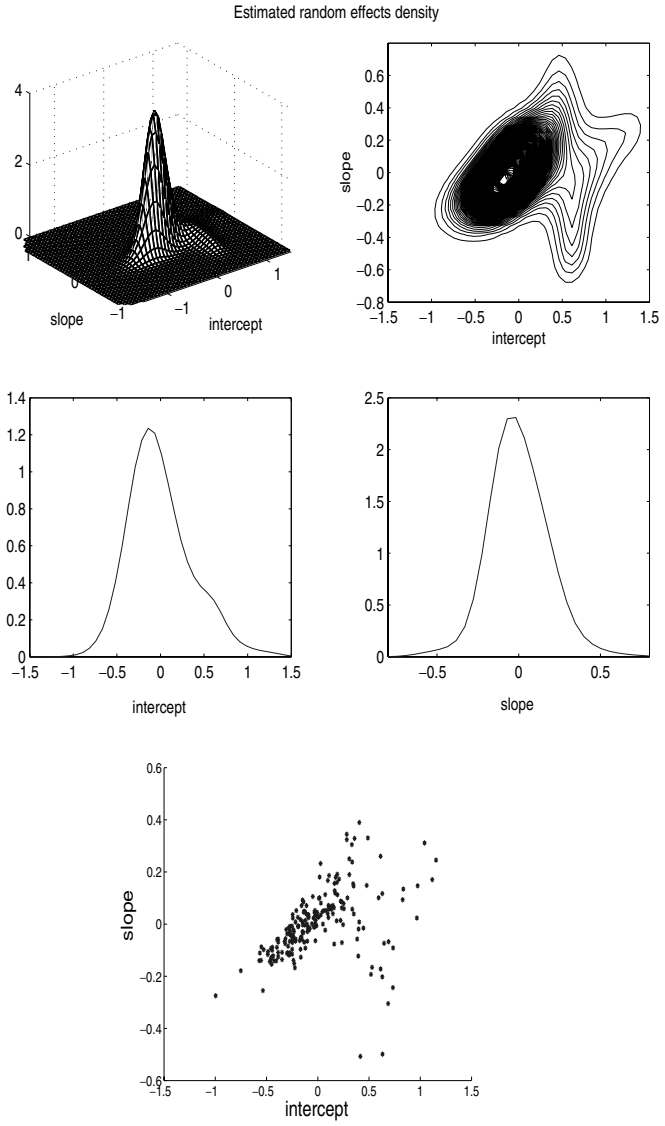

Figure 1. Estimated random effects distribution from the PGM linear mixed model fitted to the cholesterol data. In the first row the surface plot and the contour plot based on 80 contour lines are shown. The second row shows the fitted marginal distributions of the random intercept and slope, respectively. The scatter plot for the empirical Bayes estimates is given at the bottom.

6. A Simulation Study

6.1 *The Setup*

A limited simulation study has been carried out to evaluate the performance of our model for the following four random effects distributions:

- Bivariate Gaussian

$$\mathbf{b} \sim N\left\{ \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0.15 & 0.02 \\ 0.02 & 0.04 \end{bmatrix} \right\}.$$

Table 2

Mean integrated squared error by the PGM linear mixed model. Panel A shows estimates of the bivariate random effects distributions and Panel B shows estimates of the marginal distributions relative to the Zhang and Davidian SNP model for sample sizes $K = 50$ and $K = 200$.

	Sample size		
Random effects	$K = 50$	$K = 100$	$K = 200$
	Panel A		
Bivariate normal	0.0593	0.0260	0.0102
Mixture of bivariate normals	0.0908	0.0566	0.0377
Lognormal–normal	0.0317	0.0293	0.0286
$t(3)$-normal	0.0061	0.0031	0.0019

	$K = 50$		$K = 200$	
	PGM	SNP	PGM	SNP
	Panel B			
Mixture-normal				
Intercept	0.0366	0.0144	0.0068	0.0011
Slope	0.0184	0.0238	0.0051	0.0062
Lognormal-normal				
Intercept	0.0962	0.1337	0.0929	0.1402
Slope	0.0038	0.0076	0.0011	0.0016
$t(3)$-normal				
Intercept	0.0126	0.0177	0.0042	0.0093
Slope	0.0033	0.0040	0.0009	0.0009

- A mixture of two bivariate Gaussian distributions

$$\mathbf{b} \sim 0.5 \times N\left\{\begin{bmatrix}-1\\0\end{bmatrix}, \begin{bmatrix}0.15 & 0.02\\0.02 & 0.04\end{bmatrix}\right\} + 0.5 \times N\left\{\begin{bmatrix}1\\0\end{bmatrix}, \begin{bmatrix}0.15 & 0.02\\0.02 & 0.04\end{bmatrix}\right\}.$$

- A log-normal random intercept and an independent standard normal random slope: $\log(b_0) \sim N(0, 1)$ and $b_1 \sim N(0, 1)$.
- A t-distribution for the random intercept and an independent standard normal random slope: $b_0 \sim t(3)$ and $b_1 \sim N(0, 1)$.

Given the random effects, the response is generated from a normal distribution as $Y_{ij} \mid \mathbf{b}_i \sim N(\beta_0 + \beta_1 t_{ij} + b_{0i} + b_{1i} t_{ij}, \sigma^2)$, $i = 1, \ldots, K$, $j = 1, \ldots, 6$, $(\beta_0, \beta_1) = (2.35, 0.28)$, $\sigma^2 = 0.04$, and t_{ij} in $\{0, 2, 4, 6, 8, 10\}$. The choice of the simulation parameters and of the design matrix was inspired by the results from the fit of classical linear mixed model analysis to the cholesterol data. Three different sample sizes $K = 50$ (small), $K = 100$ (intermediate), and $K = 200$ (large) were considered. There were 100 simulated datasets under each combination of the random effects distributions and sample sizes.

6.2 *Model Fitting*

For the bivariate Gaussian case, we have taken $J = L = 10$. In order to attain more flexibility, $J = L = 14$ was taken for the remaining non-Gaussian cases. The means of the Gaussian density basis functions in each dimension and on the standardized scale are based on the interval $[-4, 4]$. Two penalty terms of order $k = 3$ were imposed in all cases. After an initial inspection, the optimal $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)$ corresponding to a normal random effects distribution were best searched on a grid with values [1, 10, 1000, 10,000, ...], while for the other nonnormal components the grid [1e–05, 1e–04, ..., 1e–01] was searched. Relatively large λ values are needed for approximating a normal distribution with a large number of basis Gaussian densities. Our empirical results corroborate the theoretical claim in Section 3.1.

6.2.1 *Estimated random effects distributions.* The performance of our model in estimating the true underlying distribution was evaluated by the integrated squared error (ISE) between the estimated distribution and the true distribution, given by $\int [\hat{f}(\mathbf{b}) - f(\mathbf{b})]^2 \, d\mathbf{b}$.

Table 2A presents the average ISE over all the simulations for the datasets of the different sample sizes. It shows that for the four random effects distributions, the ISE decreases as the sample size increases.

Figures 2 and 3 present the estimated marginal distributions for the nonnormally distributed random intercepts. It can be concluded that the estimated distributions show the characteristic features of the true distributions. Further, the estimated distributions approach the truth as K increases.

Additionally, we applied the SNP approach of Zhang and Davidian and compared it to our PGM results. We can conclude that the average SNP smoothed random effects distribution is closer to the true random effects distribution for a mixture of normal distributions. However, the PGM approach outperforms the SNP method in the other cases. It was also observed that the SNP approach yielded a higher variability of the smoothed distribution than the PGM approach. This is corroborated by the ISE results shown in Table 2B.

6.2.2 *Estimates for the fixed effects and covariance parameters.* Table 3 shows the average (over the different

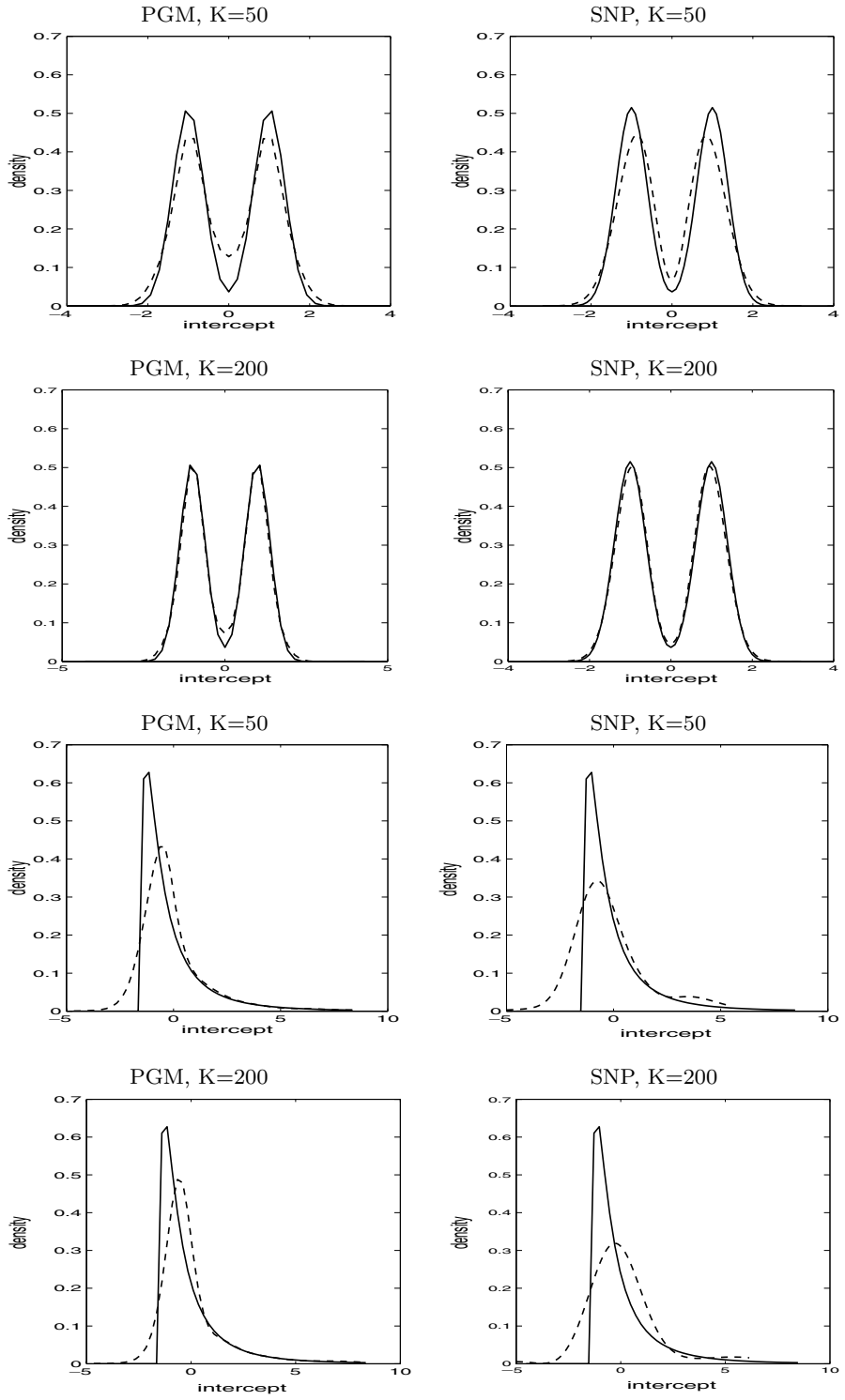

Figure 2. Penalized Gaussian mixture (PGM) and Zhang and Davidian semi-nonparametric (SNP) estimates of random intercepts simulated from mixture and log-normal distributions with sample sizes $K = 50$ and 200. Estimated distributions averaged over 100 simulations (dashed line) are superimposed to the true distributions (solid line).

simulations) mean squared errors (MSEs) in estimating $\boldsymbol{\eta}$. For all random effects distributions, the MSE decreases as the sample size increases. We also compared the PGM parameter estimates with those of the classical linear mixed model (results not shown). The MSEs were very similar implying that the fixed effects and covariance parameters are robust to the misspecification of the random effects distribution, corroborating the result of Verbeke and Lesaffre (1997).

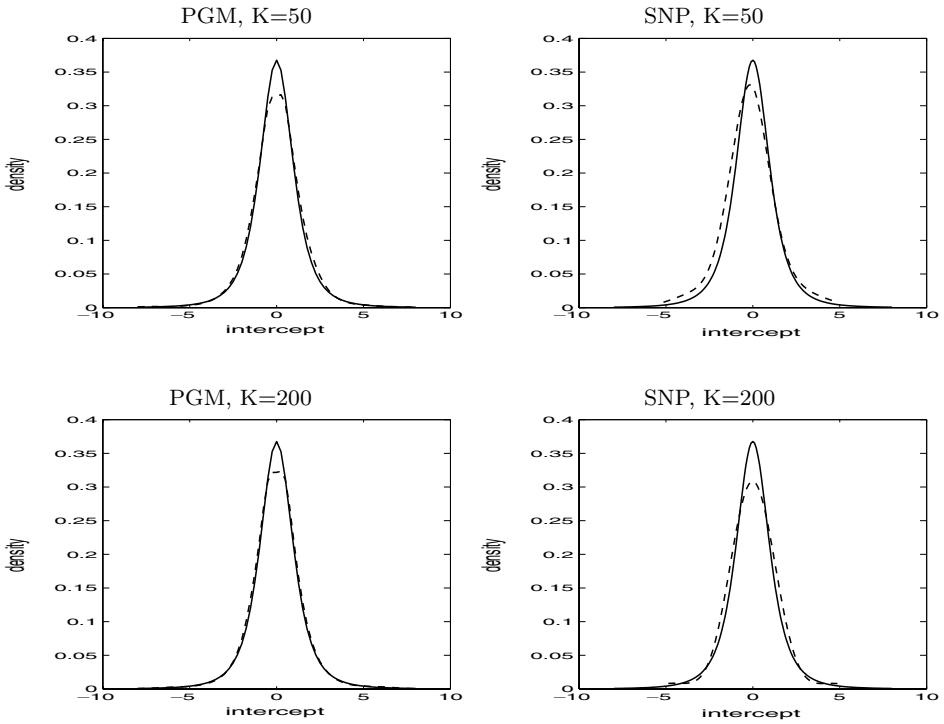

Figure 3. Penalized Gaussian mixture (PGM) and Zhang and Davidian semi-nonparametric (SNP) estimates of random intercepts simulated from Student's $t(3)$-distributions with sample sizes $K = 50$ and 200. Estimated distributions averaged over 100 simulations (dashed line) are superimposed to the true distributions (solid line).

Table 3

Mean and MSE in estimating fixed effects and covariance parameters of the PGM linear mixed model. L1, L2, and L3 are the elements of the lower triangular Cholesky decomposition matrix R.

		Mean (MSE)		
Random effects	True $\boldsymbol{\theta}$	$K = 50$	$K = 100$	$K = 200$
Bivariate normal	$\beta_0 = 2.35$	2.3510 (0.0038)	2.3550 (0.0018)	2.3531 (0.0010)
	$\beta_1 = 0.28$	0.2840 (0.0009)	0.2811 (0.0005)	0.2806 (0.0002)
	$\log(\sigma) = -1.6094$	−1.6044 (0.0030)	−1.6086 (0.0011)	−1.6094 (0.0005)
	$L1 = 0.3873$	0.3864 (0.0013)	0.3839 (0.0011)	0.3852 (0.0004)
	$L2 = 0.0516$	0.0515 (0.0009)	0.0480 (0.0005)	0.0507 (0.0002)
	$L3 = 0.1932$	0.1879 (0.0005)	0.1919 (0.0002)	0.1926 (0.0001)
Mixture of bivariate normals	$\beta_0 = 2.35$	2.3441 (0.0302)	2.3518 (0.0154)	2.3505 (0.0082)
	$\beta_1 = 0.28$	0.2790 (0.0008)	0.2792 (0.0004)	0.2790 (0.0002)
	$\log(\sigma) = -1.6094$	−1.6186 (0.0028)	−1.6153 (0.0012)	−1.6131 (0.0006)
	$L1 = 1.0724$	1.0783 (0.0041)	1.0751 (0.0020)	1.0761 (0.0008)
	$L2 = 0.0187$	0.0188 (0.0007)	0.0180 (0.0004)	0.0154 (0.0002)
	$L3 = 0.1991$	0.1965 (0.0004)	0.1964 (0.0002)	0.1994 (0.0001)
Lognormal–normal	$\beta_0 = 3.9987$	3.9983 (0.0804)	4.0150 (0.0489)	3.9900 (0.0219)
	$\beta_1 = 0.28$	0.2683 (0.0201)	0.2738 (0.0097)	0.2769 (0.0050)
	$\log(\sigma) = -1.6094$	−1.6140 (0.0021)	−1.6104 (0.0009)	−1.6094 (0.0007)
	$L1 = 2.1612$	1.7877 (0.3883)	1.8272 (0.2815)	1.7169 (0.2805)
	$L2 = 0.0000$	−0.0186 (0.0237)	−0.0041 (0.0068)	−0.0117 (0.0049)
	$L3 = 1.0000$	0.9871 (0.0105)	0.9908 (0.0059)	0.9834 (0.0029)
$t(3)$-normal	$\beta_0 = 2.35$	2.3196 (0.0443)	2.3473 (0.0290)	2.3455 (0.0123)
	$\beta_1 = 0.28$	0.2522 (0.0263)	0.2940 (0.0109)	0.2888 (0.0062)
	$\log(\sigma) = -1.6094$	−1.6144 (0.0025)	−1.6057 (0.0012)	−1.6084 (0.0007)
	$L1 = 1.7321$	1.6032 (0.1774)	1.6240 (0.0969)	1.6953 (0.0644)
	$L2 = 0.0000$	−0.0166 (0.0193)	−0.0055 (0.0111)	−0.0028 (0.0043)
	$L3 = 1.0000$	1.0049 (0.0114)	0.9806 (0.0047)	0.9940 (0.0022)

7. Concluding Remarks

It is important to stress that the PGM linear mixed model is different from a classical Gaussian mixture model. In the latter the components together with their weights need to be determined and the means and the standard deviations of the Gaussian distributions need to be estimated. It is known that maximizing the likelihood of a mixture model with varying means (and standard deviations) is not a trivial task, often resulting in a likelihood with many local modes. Also, the determination of the number of Gaussian components is not straightforward. No such numerical difficulties were encountered in the PGM method. Yet our methodology can give roughly the same information as that of a classical mixture model, without giving the relative importance of the components.

The PGM model has a number of tuning parameters, namely, the penalty parameter λ, the number of Gaussian components or knots ($J \times L$), and the order of the difference operator (k). There is no strict rule for selecting the values of J and L except that they should be large enough to fit the features in the data. Thus, in principle, J and L could be very large because overfitting is controlled by the penalty parameter, and therefore the number of knots is not crucial (see also Ruppert, 2002). With respect to k, we propose to choose $k = 3$ to exploit the property discussed in Section 3.1. The penalty parameter λ is chosen to minimize AIC. It is difficult though to give a uniform rule about which grid to take. A two-stage procedure might be used with an initial grid ranging from -6 to 6 on the log scale. In a second step a finer grid might be chosen around the best value.

The standardization of the random effects depends on the estimated random effects covariance matrix D of the classical linear mixed model. This procedure was inspired by the fact that D is consistently estimated for large sample sizes even when the random effects distribution is not normal (Verbeke and Lesaffre, 1997).

In the application and all simulations, we used a grid for the means of the Gaussian basis densities defined on $[-4, 4]$. This interval was chosen assuming that the standardized random effects distribution will have ignorable mass outside this interval. Extending the interval to $[-6, 6]$ with different random effects distributions did not change the performance of our method on the average even for the heavy tailed $t(3)$-distribution. But, if required one could start the PGM approach with the interval $[-6, 6]$ and in a second step reduce it to $[-4, 4]$ or smaller, if possible.

We have parameterized the smoothing problem in the a_{jl} coefficients as well as the penalty term. But, we have also experimented with (a) a penalty term in the c_{jl} coefficients with a parameterization in the a_{jl} coefficients, and with (b) the parameterization and the penalty terms expressed both in the c_{jl} coefficients. Our experience revealed that these alternative approaches do not offer practical advantages. Indeed, the computation times were much longer due to computational difficulties caused by estimated c_{jl} coefficients lying on the boundary (i.e., being 0). We do recognize that allowing the mixing coefficients being zero allows to discard normal components of the grid. But this advantage is counterbalanced by the statistical difficulty of having parameter estimates lying on the boundary of the parameter space rendering the calculation of the effective degrees of freedom much more complicated.

The model can be generalized to higher dimensional random effects distributions implying, however, that the number of parameters will increase geometrically, which might cause a computational burden and at times an unstable maximization algorithm. Commencing with a more stable algorithm like the EM algorithm in the first few iterations will then be helpful.

Acknowledgements

The authors thank Geert Verbeke for helpful discussions during the early developments of our model. They also thank Kris Bogaerts and Arnost Komarek for the various discussions and their helpful suggestions. Thanks go also to Irene Gijbels for sharing with us her insights into smoothing problems and finally to Daowen Zhang and Marie Davidian for providing the cholesterol data and the SAS/IML macro for fitting their model. Finally, we would like to thank the associate editor and the referees. Their comments and suggestions improved the article considerably. Financial support from the IAP research network nr. P5/24 of the Belgian State (Federal Office for Scientific, Technical and Cultural Affairs) is gratefully acknowledged.

Résumé

On propose un modèle linéaire mixte avec une densité lissées des effets aléatoires. On applique une approche similaire au lissage par P-spline de Eilers et Marx (1996) pour obtenir une estimation plus souple de la densité des effets aléatoires. La différence entre notre approche et la leur est que leur base de B-splines est remplacée par des approximations par densités Gaussiennes. L'ajustement est réalisé par maximisation d'une vraisemblance marginale pénalisée. Le meilleur paramètre de pénalisation est celui qui minimise le critère d'information de Akaïke, conformément aux résultats de Gray (Gray, 1992). Notre méthode est applicable à une structure d'effets aléatoires d'un nombre quelconque de dimensions, mais dans l'article nous n'explorons que le cas d'une structure bidimensionnelle. Notre méthodologie est conceptuellement simple et relativement facile à mettre en œuvre. Elle est appliquée à des données sur le cholestérol initialement analysées par Zhang et Davidian (2001). Une étude de simulation montre que notre méthode conduit, pour des petits échantillons, à des estimations presque sans biais des paramètres de la régression et du lissage. La convergence des estimateurs est établie dans un cas particulier.

References

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

Eilers, P. and Marx, B. (1996). Flexible smoothing with *B*-splines and penalties. *Statistical Science* **11,** 89–121.

Gray, R. (1992). Flexible methods for analysing survival data using splines, with application to breast cancer prognosis. *Journal of the American Statistical Association* **87,** 942–951.

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

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

Magder, 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.

Ruppert, D. (2002). Selecting the number of knots for penalized splines. *Journal of Computational and Graphical Statistics* **11,** 735–757.

Shen, W. and Louis, T. (1999). Empirical Bayes estimation via the smoothing by roughening approach. *Journal of Computational and Graphical Statistics* **8,** 800–823.

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

Unser, M., Aldroubi, A., and Eden, M. (1992). On the asymptotic convergence of *B*-spline wavelets to Gabor functions. *IEEE Transactions on Information Theory* **38,** 864–872.

Verbeke, G. and Lesaffre, E. (1996). A linear mixed 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.

Verweij, P. and Van Houwelingen, H. C. (1994). Penalized likelihood in Cox regression. *Statistics in Medicine* **13,** 2427–2436.

Wald, A. (1949). Note on the consistency of maximum likelihood estimates. *Annals of Mathematical Statistics* **20,** 595–601.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* **50,** 1–25.

Zhang, D. and Davidian, M. (2001). Linear mixed models with flexible distributions of random effects for longitudinal data. *Biometrics* **57,** 795–802.

Received June 2003. *Revised May* 2004.
Accepted May 2004.

Appendix

Proof of Consistency of the Parameter Estimates

We first take the particular case where a correct grid for the normal density has been chosen with the assumed standard deviation for the standardized random effects and assume that (1) is correctly specified. Hence, we assume that there exists a true vector $\mathbf{a}_0$ and also that none of the elements of $\mathbf{a}_0$ lies at infinity.

We denote the true value of $\boldsymbol{\theta}$ by $\boldsymbol{\theta}_0$. Let $\boldsymbol{\theta} = (\eta^{\mathrm{T}}, \mathbf{a}^{\mathrm{T}})^{\mathrm{T}}$, where $\boldsymbol{\eta} = (\boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\sigma}_R^{\mathrm{T}}, \sigma)^{\mathrm{T}}$. For reasons of simplicity we assume that $\lambda_1 = \lambda_2 \equiv \lambda$ such that there is only matrix Δ, which is the difference operator matrix for the penalty term (4). Assume that $\mathbf{a}_0$ satisfies $\mathbf{a}_0^{\mathrm{T}}\Delta^{\mathrm{T}}\Delta\mathbf{a}_0 = b$ where b is some fixed constant. Denote the unpenalized estimate of $\boldsymbol{\theta}$ as $\hat{\boldsymbol{\theta}}_u$ and the corresponding penalized estimate as $\hat{\boldsymbol{\theta}}_p$.

Under the conditions stated above, standard maximum likelihood theory implies that $\hat{\boldsymbol{\theta}}_u \xrightarrow{P} \boldsymbol{\theta}_0$ and hence also that $\hat{\mathbf{a}}_u \xrightarrow{P} \mathbf{a}_0$. Thus when $K \to \infty, P(\hat{\mathbf{a}}_u^{\mathrm{T}}\Delta^{\mathrm{T}}\Delta\hat{\mathbf{a}}_u \le cb) \to 1$ for a fixed $c > 1$ but arbitrary.

Because $\log L(\hat{\boldsymbol{\theta}}_u; \mathbf{Y}) \ge \log L(\hat{\boldsymbol{\theta}}_p; \mathbf{Y})$ and $\log L(\hat{\boldsymbol{\theta}}_u; \mathbf{Y}) - \frac{\lambda}{2}\hat{\mathbf{a}}_u^{\mathrm{T}}\Delta^{\mathrm{T}}\Delta\hat{\mathbf{a}}_u \le \log L(\hat{\boldsymbol{\theta}}_p; \mathbf{Y}) - \frac{\lambda}{2}\hat{\mathbf{a}}_p^{\mathrm{T}}\Delta^{\mathrm{T}}\Delta\hat{\mathbf{a}}_p$, we obtain $\hat{\mathbf{a}}_p^{\mathrm{T}}\Delta^{\mathrm{T}}\Delta\hat{\mathbf{a}}_p \le \hat{\mathbf{a}}_u^{\mathrm{T}}\Delta^{\mathrm{T}}\Delta\hat{\mathbf{a}}_u \le cb$.

Take now $\hat{\boldsymbol{\theta}}_p^*$ the penalized estimate of $\boldsymbol{\theta}$ under the constraint that $\hat{\mathbf{a}}_p^{*\mathrm{T}}\Delta^{\mathrm{T}}\Delta\hat{\mathbf{a}}_p^* \le cb$, then

$$\frac{L\left(\hat{\boldsymbol{\theta}}_p^*\right) \cdot m_{\hat{\boldsymbol{\theta}}_p^*}^K}{L(\boldsymbol{\theta}_0) \cdot m_{\boldsymbol{\theta}_0}^K} \ge 1,$$

with $m_{\boldsymbol{\theta}} = \exp(-\frac{\lambda}{2}\boldsymbol{\theta}^{\mathrm{T}}\mathbf{P}\boldsymbol{\theta})^{1/K}$ whereby $\mathbf{P}$ is a block-diagonal penalty matrix with the penalty matrix $\Delta^{\mathrm{T}}\Delta$ corresponding to the vector $\mathbf{a}$ and zero entries elsewhere. It is then clear that $m_{\boldsymbol{\theta}} < 1$ for all K and all $\boldsymbol{\theta}$.

This implies that

$$\frac{L\left(\hat{\boldsymbol{\theta}}_p^*\right)}{L(\boldsymbol{\theta}_0)} \ge \frac{m_{\boldsymbol{\theta}_0}^K}{m_{\hat{\boldsymbol{\theta}}_p^*}^K} = \frac{\exp\left(-\frac{\lambda}{2}b\right)}{\exp\left(-\frac{\lambda}{2}\hat{\boldsymbol{\theta}}_p^{*\mathrm{T}}\mathbf{P}\hat{\boldsymbol{\theta}}_p^*\right)} \ge \frac{\exp\left(-\frac{\lambda}{2}b\right)}{\exp\left(-\frac{\lambda}{2}cb\right)} > 1.$$

Using Theorem 2 of Wald (1949) shows that $\hat{\boldsymbol{\theta}}_p^* \xrightarrow{P} \boldsymbol{\theta}_0$. Finally, when $K \to \infty, P(\hat{\boldsymbol{\theta}}_p^* \ne \hat{\boldsymbol{\theta}}_p) \to 0$ which shows that $\hat{\boldsymbol{\theta}}_p \xrightarrow{P} \boldsymbol{\theta}_0$.

When some of the mixing weights are zero, i.e., some $c_{jl} = 0$, we must have that some of the a_{jl} are $-\infty$. In that case the proof can still be applied with the reduced set of coefficients (leaving out zero c_{jl}'s), but of course this will imply in general that a regular grid of Gaussian densities is not possible anymore.

In general, there exists no true vector of smoothing coefficients $\mathbf{a}_0$. This happens when the true random effects distribution is not the mixture of Gaussian distributions dictated by our choice of the grid of knots and with our choice of standard deviation for the standardized random effects. When λ is kept fixed or $\frac{\lambda}{K} \to 0$ while $K \to \infty, P(|\hat{\boldsymbol{\theta}}_p^* - \hat{\boldsymbol{\theta}}_u| > \epsilon) \to 0$ for any arbitrary $\epsilon > 0$, since the penalty part reduces in importance as $K \to \infty$. Using the above arguments, the same result holds true for $\hat{\boldsymbol{\theta}}_p$. The results of White (1982) imply that $\hat{\boldsymbol{\theta}}_u$ converges to $\boldsymbol{\theta}_0^*$, which minimizes the Kullback–Leibler distance between the true (linear mixed) model and the assumed (linear mixed) model (for which the random effects distribution is our assumed mixture of Gaussian distributions).

Even when no $\mathbf{a}_0$ exists, the proof of Verbeke and Lesaffre (1997) can easily be extended to show that $\hat{\boldsymbol{\beta}}^* \xrightarrow{P} \boldsymbol{\beta}$. Indeed, Verbeke and Lesaffre (1997) show that under misspecification of the covariance matrix the regression coefficients (fixed effects) are consistently estimated. Further, when $\frac{\lambda}{K} \to 0$ as $K \to \infty$ the penalized estimates come close to the penalized estimates.