Restricted maximum likelihood estimation of variance parameters in generalized linear mixed models

Thomas Kneib

Department of Statistics
Ludwig Maximilians-University Munich

1 Introduction

Originally, restricted maximum likelihood estimation (REML) of variance parameters was introduced by Patterson & Thompson (1971) in the context of designed experiments. Harville (1974) extended their results to the more general linear mixed model. Since this time, REML-estimation has become widely used for mixed models with normally distributed response. However, REML can also be used in generalized linear mixed models (GLMM) (compare for example Breslow & Clayton (1993)). The usually presented formulae for the determination of REML-estimates via Fisher-Scoring suffer from the fact, that they involve the evaluation and manipulation of several $n \times n$-matrices, where n is the number of observations. This makes the direct usage of these formulae intractable for datasets with more than about 3000 observations due to the large amount of memory that is needed to store these matrices and due to the enormous runtime of the computations.

In this paper we describe, how some simple modifications of the usual formulae yield computationally much more appealing possibilities for the determination of REML-estimates. The main advantage of the formulae presented here, is that they reduce the dimension of the involved matrices to a value that no longer depends on the number of observations but on the number of regression parameters in the model. Moreover we can make use of matrices, that have to be computed in the determination of the regression parameters. Therefore these matrices are already available and do not have to be computed again.

In the next section we briefly describe the generalized linear mixed model assumptions. Then the estimation of the regression coefficients is reviewed, since this forms a building block for the more efficient formulae in section 4 which deals with the REML-estimation of the variance parameters. In a last section the simultaneous estimation of both regression and variance parameters is summarized in an algorithm.

2 Model

Suppose we have given observations (y_i, x_i), $i = 1, \dots, n$ on a response variable y and a set of covariates $x = (x_1, \dots, x_p)'$. Generalized linear models assume that the distribution of y_i belongs to an exponential family and that the expectation $\mu_i = \mathbb{E}(y_i)$ is linked to the linear predictor by

$$\mu_i = h(\eta_i),\ \eta_i = x_i'\beta \tag{1}$$

where h is a known response function and β is a p-dimensional vector of regression coefficients.

In a generalized linear *mixed* model we assume, that we have observations on a further set of covariates $z = (z_1, \ldots, z_q)'$ and replace the linear predictor in (1) through

$$\eta_i = x_i'\beta + z_i'b.$$

In contrast to β the q-dimensional vector b is not assumed to be fixed but to be the realization of a random vector. Therefore b is said to consist of *random effects* while β is said to consist of *fixed effects*. Note, that the covariates in x_i and z_i must not necessarily be different. It is also possible to treat covariates as fixed effects and as random effects as well.

For the further description it is more convenient to rewrite the model in matrix notation which leads to the linear predictor

$$\eta = X\beta + Zb \tag{2}$$

with the usual stacked vector η and the stacked matrices X and Z. To complete our model, we have to specify the distribution b belongs to. Usually it is assumed that b is multivariate normal with expectation zero and some positive definite covariance matrix Λ:

$$b \sim \mathrm{N}(0, \Lambda). \tag{3}$$

In the following we will always assume, that the covariance matrix Λ is uniquely determined through an r-dimensional vector of variance parameters ν, i.e. $\Lambda = \Lambda(\nu)$. Therefore, to estimate Λ, we have to estimate ν.

In addition to the estimation of the general variance parameters ν we will also discuss a special case in greater detail. In this special case we have some further assumptions about Λ and Z which lead to some simplifications in the computations described in section 4. These additional assumptions are motivated by the fact, that models with this structure arise from the reparameterisation of semiparametric regression models that are based on penalization approaches (compare e.g. Fahrmeir, Kneib & Lang (2003)). We will use this special structure of Λ and Z only in section 4 while section 3 remains completely unaffected if more general forms of Λ and Z are used.

Here we assume that the vector b may be decomposed into r independent blocks $b_1, \ldots, b_r$ with lengths q_j, $j = 1, \ldots, r$ and distributions

$$b_j \sim \mathrm{N}(0, \tau_j^2 I).$$

Therefore in this special case Λ is a diagonal matrix given by

$$\Lambda = \mathrm{diag}(\tau_1^2, \ldots, \tau_1^2, \ldots, \tau_r^2, \ldots, \tau_r^2). \tag{4}$$

Now the vector of variance parameters ν consists of the variances τ_j^2: $\nu = (\tau_1^2, \ldots, \tau_r^2)'$. If in addition the design matrix Z may be decomposed in blocks corresponding to the b_j

$$Z = \begin{pmatrix} Z_1 & 0 & \ldots & \ldots & 0 \\ 0 & Z_2 & 0 & \ldots & 0 \\ \vdots & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ 0 & \ldots & \ldots & 0 & Z_r \end{pmatrix} \tag{5}$$

we may replace the linear predictor (2) through

$$\eta = X\beta + Z_1 b_1 + \ldots + Z_r b_r.$$

Note, that if this special model is obtained from the reparameterisation of a semiparametric regression model, every term $Z_j b_j$ corresponds to one nonparametric term or subject specific random effect in the original model.

3 Estimation of regression coefficients

To estimate the regression coefficients $\delta = (\beta', b')'$, the density of the joint distribution of y and b must be maximized with respect to δ. Therefore we obtain the log-likelihood

$$\begin{aligned} l(\delta) &= \log(p(y|b)p(b)) \\ &= \sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{\phi}\omega_i - \frac{1}{2}b'\Lambda^{-1}b \end{aligned} \tag{6}$$

where θ_i is the so-called natural parameter of the exponential family the response belongs to, $b(\cdot)$ is a known function depending on that exponential family and ϕ and ω_i are the dispersion parameter and some known weights, respectively (compare Fahrmeir & Tutz (2001) pp. 18).

In analogy to the estimation in generalized linear models the maximization of (6) is carried out through a (slightly modified) Fisher-Scoring algorithm. Therefore we need the score function and the expected Fisher-information matrix. Since (6) has the form of a penalized log-likelihood whose penalty depends only on the random effects b, it is useful to to look on the derivatives separately (Fahrmeir & Tutz (2001), pp. 298/99). For the score function we obtain $s(\delta) = \partial l(\delta)/\partial\delta = (s_\beta(\delta)', s_b(\delta)')'$ with

$$s_\beta(\delta) = \frac{\partial l(\delta)}{\partial\beta} = X'D\Sigma^{-1}(y-\mu) \tag{7}$$

and

$$s_b(\delta) = \frac{\partial l(\delta)}{\partial b} = Z'D\Sigma^{-1}(y-\mu) - \Lambda^{-1}b \tag{8}$$

where

$$D = \text{diag}(D_i) = \text{diag}(\partial h(\eta_i)/\partial\eta) \tag{9}$$

and

$$\Sigma = \text{Var}(y|\delta) = \text{diag}(\sigma_i^2) = \text{diag}(\phi v(\mu_i)/\omega_i). \tag{10}$$

Here $v(\mu)$ is the variance function that is determined by the exponential family through $v(\mu) = b''(\theta)$.

For the expected Fisher information we get

$$F(\delta) = \begin{pmatrix} F_{\beta\beta}(\delta) & F_{\beta b}(\delta) \\ F_{b\beta}(\delta) & F_{bb}(\delta) \end{pmatrix}$$

with

$$\begin{aligned} F_{\beta\beta}(\delta) &= X'D\Sigma^{-1}DX \\ F_{\beta b}(\delta) = F_{b\beta}(\delta)' &= X'D\Sigma^{-1}DZ \\ F_{bb}(\delta) &= Z'D\Sigma^{-1}DZ + \Lambda^{-1}. \end{aligned}$$

Now we can estimate δ through iterating

$$\hat{\delta}^{(k+1)} = \hat{\delta}^{(k)} + F(\hat{\delta}^{(k)})^{-1}s(\hat{\delta}^{(k)}) \tag{11}$$

with some starting value $\delta^{(0)}$.

Defining the working observations

$$\tilde{y} = X\hat{\beta}^{(k)} + Z\hat{b}^{(k)} + D^{-1}(y-\mu) \tag{12}$$

leads to the equivalent estimation process of iteratively solving the linear system of equations

$$\begin{pmatrix} X'WX & X'WZ \\ Z'WX & Z'WZ + \Lambda^{-1} \end{pmatrix} \begin{pmatrix} \hat{\beta}^{(k+1)} \\ \hat{b}^{(k+1)} \end{pmatrix} = \begin{pmatrix} X'W\tilde{y} \\ Z'W\tilde{y} \end{pmatrix}$$

with the working weights

$$W = \mathrm{diag}(w_i) = D\Sigma^{-1}D. \tag{13}$$

which is called iteratively weighted least squares since it has the same form as for weighted least squares estimation but the weights change in every iteration.

4 Estimation of variance parameters

Now we turn to the estimation of the variance parameters ν. To motivate the estimation via (approximate) restricted maximum likelihood we first discuss the estimation of variance parameters for Gaussian response. Then we generalize the results to non-normal data.

In linear mixed models, that is, for Gaussian response, a well known technique for the estimation of variance parameters is maximum likelihood. To obtain the maximum likelihood estimates we consider the marginal distribution of y, which is given by

$$y \sim \mathrm{N}(X\beta, V) \tag{14}$$

with

$$V = \sigma^2 I + Z\Lambda Z'.$$

Note, that this distribution and its density explicitly depend on Λ and therefore allow the determination of derivatives with respect to ν which are needed in the estimation. This is not the case for the conditional distribution of $y|b$ which is given by

$$y|b \sim \mathrm{N}(X\beta + Zb, \sigma^2 I).$$

For Gaussian response the maximum likelihood estimator for β may be derived analytically from the marginal distribution (14) yielding

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

Plugging this expression into the marginal likelihood (the density of the marginal distribution of y) yields the profile likelihood for ν and σ^2 which may be maximized numerically using the EM-algorithm or Fisher-Scoring (compare Harville (1977) for details). However, maximum likelihood estimation does not take into account the loss of degrees of freedom due to the estimation of β and therefore the obtained estimators are usually biased towards zero. A way to overcome this bias is the usage of *restricted* maximum likelihood (Patterson & Thompson 1971). Here the estimation is based on the likelihood of some *error contrasts* $u = A'y$ rather than on the likelihood of y.

An error contrast is a linear combination $a'y$ with expectation zero, causing the distribution of $a'y$ not to depend on β. An example for a set of such error contrasts are the residuals $\hat{\varepsilon}_i = y_i - x_i'\hat{\beta}$ where the vector a consists of the ith row of the residual matrix $R = I - X(X'X)^{-1}X'$. Then the vector of all residuals $\hat{\varepsilon}$ has the distribution

$$\hat{\varepsilon} \sim \mathrm{N}(0, R\sigma^2)$$

which has the desired property $\mathbb{E}(\hat{\varepsilon}) = 0$. However, since R is inherently singular, the distribution is partially improper and therefore in general the usage of the residuals is not advisable.

Since only $n - p$ linear independent error contrasts may exist, we use error contrasts obtained from the decomposition

$$AA' = I - X(X'X)^{-1}X' \text{ with } A'A = I \tag{15}$$

where A is an $(n \times n - p)$ matrix with full column rank. It is easy to show, that the so-defined error contrasts $u = A'y$ fulfill $\mathbb{E}(u) = 0$.

Now the marginal density of u is given by (Harville 1974)

$$p(u) = \left(\frac{1}{2\pi}\right)^{\frac{n-p}{2}} |X'X|^{\frac{1}{2}} |V|^{-\frac{1}{2}} |X'V^{-1}X|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(y - X\hat{\beta})'V^{-1}(y - X\hat{\beta})\right].$$

The restricted maximum likelihood estimators of ν and σ^2 are therefore obtained through maximizing

$$l^*(\nu, \sigma^2) = -\frac{1}{2}\log(|V|) - \frac{1}{2}\log(|X'V^{-1}X|) - \frac{1}{2}(y - X\hat{\beta})'V^{-1}(y - X\hat{\beta}).$$

Note, that this *restricted log-likelihood* does not depend on the special choice of A as long as $n - p$ linear independent error contrasts are used (Verbeke & Molenberghs (2000) section 5.3).

Now it would be desirable to generalize the restricted maximum likelihood technique to non-normal data. However, this can not be done directly, since the definition of error contrasts is not possible for more general responses. This is due to the non-linear dependency of y on β in generalized linear mixed models. Therefore we consider an alternative approach that leads to the same restricted log-likelihood for the estimation of ν and σ^2 in the normal case and that may be extended to more general responses.

For the moment, consider not only b to be a random variable but also β. While the prior distribution of b is a proper distribution, given in (3), we assume that the distribution of β is noninformative in the sense that

$$p(\beta_j) \propto const, \text{ for } j = 1, \ldots, p.$$

Now it seems reasonable not only to integrate b out of the distribution of y but also β, resulting in the marginal distribution of y (as regards to b *and* β), which now has to be maximized with respect to the variance parameters. It can be shown (Harville 1974) that this approach leads exactly to the same likelihood as above but it has the advantage to be extendable to generalized linear mixed models. Therefore REML-estimation is also sometimes referred to as *marginal likelihood* estimation.

First, we approximate the logarithm of the conditional density of y $\log(p(y|\delta))$ through the Pearson χ^2-statistic, yielding

$$\begin{aligned} \log(p(y|\delta)) &\approx \sum_{i=1}^{n} \frac{(y_i - \mu_i)^2}{\omega_i v(\mu_i)/\phi} \\ &= (y - \mu)'\Sigma^{-1}(y - \mu). \end{aligned}$$

This is in fact equivalent to the Laplace approximation of $\log(p(y|\delta))$ through a quadratic function (compare Tierney & Kadane (1986)). Using the definition of the working observations $\tilde{y}$ in (12) gives

$$(y - \mu) = D(\tilde{y} - X\beta - Zb).$$

Therefore we have

$$\begin{aligned}\log(p(y|\delta)) &\approx (\tilde{y} - X\beta - Zb)' D\Sigma^{-1} D(\tilde{y} - X\beta - Zb) \\ &= (\tilde{y} - X\beta - Zb)' W(\tilde{y} - X\beta - Zb).\end{aligned}$$

Ignoring the dependance of W on the variance parameters (Breslow & Clayton 1993) gives rise to the fact, that the likelihood $l(\delta)$ in (6) may be approximated through the log-likelihood of a linear mixed model for the working observations $\tilde{y}$. To be more specific, we assume, that

$$\tilde{y}|\delta \overset{a}{\sim} \mathrm{N}(X\beta + Zb, W^{-1}).$$

Now we can determine the marginal distribution of $\tilde{y}$ (as regards to b and β) yielding an approximate restricted log-likelihood for the generalized linear mixed model:

$$l^*(\nu,\phi) = -\frac{1}{2}\log(|V|) - \frac{1}{2}\log(|X'V^{-1}X|) - \frac{1}{2}(\tilde{y} - X\hat{\beta})'V^{-1}(\tilde{y} - X\hat{\beta}) \tag{16}$$

with $V = W^{-1} + Z\Lambda Z'$ being an approximation to the marginal covariance matrix of $\tilde{y}$.

To obtain REML-estimators for the variance parameters we have to maximize (16) with respect to ν and ϕ. One way to to do this, is the Newton-Raphson algorithm, which is based on the first and second derivative of $l^*(\nu,\phi)$ with respect to the variance parameters. A modification of the Newton-Raphson algorithm is given by Fisher-Scoring. In this case the second derivative is replaced by its expectation. Since this yields much easier estimation equations, we will focus on Fisher-Scoring here.

Note, that in the following we will always give the derivatives with respect to the dispersion parameter although they are not needed in several models, e.g. for Poisson or Binomial data where the dispersion parameter is fixed. In this cases these derivatives have to be eliminated from the score-function and the expected Fisher-information.

The score-function $s^*(\nu,\phi)$ is an $r+1$-dimensional vector containing the derivatives of (16) with respect to ν and ϕ:

$$\begin{aligned}s^*(\nu,\phi) &= (s_1^*, \ldots, s_{r+1}^*)' \\ &= \left(\frac{\partial l^*(\nu,\phi)}{\partial \nu_1}, \ldots, \frac{\partial l^*(\nu,\phi)}{\partial \nu_r}, \frac{\partial l^*(\nu,\phi)}{\partial \phi}\right)'\end{aligned}$$

The first r elements of the score-function are given by (Harville 1977)

$$s_j^* = -\frac{1}{2}\operatorname{tr}\left(P\frac{\partial V}{\partial \nu_j}\right) + \frac{1}{2}(\tilde{y} - X\hat{\beta})'V^{-1}\frac{\partial V}{\partial \nu_j}V^{-1}(\tilde{y} - X\hat{\beta}) \tag{17}$$

with

$$P = V^{-1} - V^{-1}X(X'V^{-1}X)^{-1}X'V^{-1}. \tag{18}$$

Now the crucial point is, that the direct use of formulae (17) and (18) is not tractable for datasets with more than about $n = 3000$ observations, since they involve the computation and manipulation of several $n \times n$ matrices including P and V. In particular, the determination of V^{-1}, which requires order n^3 computations, makes the direct usage of (17) and (18) impractical. Also storing $n \times n$ matrices requires n^2 storing locations resulting in an enormous amount of memory that is rewuired to compute the score function.

The inversion of V may be avoided by changing from the marginal to the conditional view of the GLMM (Lin & Zhang 1999) yielding the expressions

$$s_j^* = -\frac{1}{2}\,\mathrm{tr}\left(P\frac{\partial V}{\partial \nu_j}\right) + \frac{1}{2}(\tilde{y} - X\hat{\beta} - Z\hat{b})'W\frac{\partial V}{\partial \nu_j}W(\tilde{y} - X\hat{\beta} - Z\hat{b}) \tag{19}$$

and

$$P = W - W(X\ Z)H^{-1}(X\ Z)'W \tag{20}$$

with

$$H = \begin{pmatrix} X'WX & X'WZ \\ Z'WX & Z'WZ + \Lambda^{-1} \end{pmatrix}.$$

Now we have to evaluate the derivatives $\frac{\partial V}{\partial \nu_j}$. We obtain

$$\frac{\partial V}{\partial \nu_j} = Z\frac{\partial \Lambda}{\partial \nu_j}Z' \tag{21}$$

which reduces to

$$\frac{\partial V}{\partial \tau_j^2} = Z\frac{\partial \Lambda}{\partial \tau_j^2}Z' = Z_j Z_j'$$

if we use the additional assumptions about Λ and Z given in (4) and (5).

Plugging these derivatives into formula (19) yields

$$s_j^* = -\frac{1}{2}\,\mathrm{tr}\left(PZ\frac{\partial \Lambda}{\partial \nu_j}Z'\right) + \frac{1}{2}(\tilde{y} - X\hat{\beta} - Z\hat{b})'WZ\frac{\partial \Lambda}{\partial \nu_j}Z'W(\tilde{y} - X\hat{\beta} - Z\hat{b}). \tag{22}$$

Though avoiding the inversion of V, the computation of (22) still involves the determination and multiplication of $n \times n$-matrices (especially P), and so problems described above remain eventually unchanged. To get around this, we replace P through the expression in (20) and use an elementary property of the trace. Therefore we obtain for the first part of (22):

$$\begin{aligned} -\frac{1}{2}\,\mathrm{tr}\left(PZ\frac{\partial \Lambda}{\partial \nu_j}Z'\right) &= -\frac{1}{2}\,\mathrm{tr}\left(WZ\frac{\partial \Lambda}{\partial \nu_j}Z'\right) + \frac{1}{2}\,\mathrm{tr}\left(W(X\ Z)H^{-1}(X\ Z)'WZ\frac{\partial \Lambda}{\partial \nu_j}Z'\right) \\ &= -\frac{1}{2}\,\mathrm{tr}\left(Z'WZ\frac{\partial \Lambda}{\partial \nu_j}\right) + \frac{1}{2}\,\mathrm{tr}\left(Z'W(X\ Z)H^{-1}(X\ Z)'WZ\frac{\partial \Lambda}{\partial \nu_j}\right) \end{aligned} \tag{23}$$

Note, that now the matrices inside the traces are no longer from dimension $n \times n$ but are reduced to dimension $q \times q$. Moreover most of the matrix products do not have to be evaluated explicitly, since they can be derived at low computational cost from the matrix H, which is involved in the determination of $\hat{\beta}$ and $\hat{b}$ (compare equation (11)). The SSCP-matrix

$$\begin{pmatrix} X'WX & X'WZ \\ Z'WX & Z'WZ \end{pmatrix} \tag{24}$$

can be computed from H by simply substracting Λ^{-1} from the lower right corner of H. Then most of the matrices needed to evaluate the traces turn out to be submatrices of the SSCP-matrix (24). Only H^{-1} and $\frac{\partial \Lambda}{\partial \nu_j}$ have to be computed in addition. So the main computational burden is shifted from the computation and manipulation of $n \times n$ matrices to the inversion of the $(p + q \times p + q)$-matrix H. Also the storage requirements are reduced from order n^2 to $(p + q)^2$.

Assuming the special model described in (4) and (5) yields even simpler formulae. In this case (23) reduces to

$$-\frac{1}{2}\operatorname{tr}\left(Z_j'WZ_j\right)+\frac{1}{2}\operatorname{tr}\left(Z_j'W(X\ Z)H^{-1}(X\ Z)'WZ_j\right)$$

and the second part of (22) is given through

$$\frac{1}{2}(\tilde{y}-X\hat{\beta}-Z\hat{b})'WZ_jZ_j'W(\tilde{y}-X\hat{\beta}-Z\hat{b})$$

To evaluate this expression it is important, not to determine Z_jZ_j' explicitly, since this matrix product has dimension $n\times n$. Instead it is much easier (and computationally favorable) to compute the squared norm of the vector $(\tilde{y}-X\hat{\beta}-Z\hat{b})WZ_j$ which is of dimension q_j.

If an additional dispersion parameter is present, the last entry in the score vector is given by

$$s^*_{r+1}=-\frac{1}{2}\operatorname{tr}\left(P\frac{\partial V}{\partial\phi}\right)+\frac{1}{2}(\tilde{y}-X\hat{\beta})'V^{-1}\frac{\partial V}{\partial\phi}V^{-1}(\tilde{y}-X\hat{\beta}). \tag{25}$$

For the derivative of V with respect to the dispersion parameter ϕ we get

$$\begin{aligned}\frac{\partial V}{\partial\phi} &= \frac{\partial W^{-1}}{\partial\phi}\\ &= D^{-1}\frac{\partial\Sigma}{\partial\phi}D^{-1}\\ &= D^{-1}\operatorname{diag}\left(\frac{\partial\phi v(\mu_i)/\omega_i}{\partial\phi}\right)D^{-1}\\ &= D^{-1}\operatorname{diag}\left(v(\mu_i)/\omega_i\right)D^{-1}\\ &= \frac{1}{\phi}W^{-1}.\end{aligned} \tag{26}$$

Plugging this derivative into (25) and changing from the marginal to the conditional view of the GLMM as described above yields

$$s^*_{r+1}=-\frac{1}{2\phi}\operatorname{tr}\left(PW^{-1}\right)+\frac{1}{2\phi}(\tilde{y}-X\hat{\beta}-Z\hat{b})'WW^{-1}W(\tilde{y}-X\hat{\beta}-Z\hat{b}) \tag{27}$$

The first part of (27) may again be simplified by inserting the expression for P in (20) giving

$$\begin{aligned}-\frac{1}{2\phi}\operatorname{tr}\left(PW^{-1}\right) &= -\frac{1}{2\phi}\operatorname{tr}\left(WW^{-1}\right)+\frac{1}{2\phi}\operatorname{tr}\left(W(X\ Z)H^{-1}(X\ Z)'WW^{-1}\right)\\ &= -\frac{n}{2\phi}+\frac{1}{2\phi}\operatorname{tr}\left((X\ Z)'W(X\ Z)H^{-1}\right).\end{aligned}$$

Note again, that besides of H^{-1} no new matrices have to evaluated to compute these expressions since $(X\ Z)'W(X\ Z)$ equals the SSCP-matrix (24).

The second part of (27) reduces to

$$\frac{1}{\phi}(\tilde{y}-X\hat{\beta}-Z\hat{b})'W(\tilde{y}-X\hat{\beta}-Z\hat{b})$$

which differs only by the multiplication with the inverse scale parameter from the weighted sum of squared residuals.

The expected Fisher-information $F^*(\nu, \phi) = (F^*_{jl}),\ j, l = 1, \ldots, r+1$ is an $(r+1 \times r+1)$-dimensional matrix whose upper left $r \times r$-corner consists of the negative expectations of the second derivatives $-\mathbb{E}\left(\frac{\partial^2 l^*(\nu,\phi)}{\partial\nu_j \partial\nu_l}\right)$, $j, l = 1, \ldots, r$. These are given by (Lin & Zhang 1999)

$$F^*_{jl} = -\mathbb{E}\left(\frac{\partial^2 l^*(\nu, \phi)}{\partial\nu_j \partial\nu_l}\right) = \frac{1}{2}\operatorname{tr}\left(P\frac{\partial V}{\partial \nu_j}P\frac{\partial V}{\partial \nu_l}\right). \tag{28}$$

Again using (21) and plugging in the expression for P in (20) gives rise to simpler formulae:

$$\begin{aligned} F^*_{jl} &= \frac{1}{2}\operatorname{tr}\left(WZ\frac{\partial\Lambda}{\partial\nu_j}Z'WZ\frac{\partial\Lambda}{\partial\nu_l}Z'\right) - \frac{1}{2}\operatorname{tr}\left(W(X\ Z)H^{-1}(X\ Z)'WZ\frac{\partial\Lambda}{\partial\nu_j}Z'WZ\frac{\partial\Lambda}{\partial\nu_l}Z'\right) \\ &\quad -\frac{1}{2}\operatorname{tr}\left(WZ\frac{\partial\Lambda}{\partial\nu_j}Z'W(X\ Z)H^{-1}(X\ Z)'WZ\frac{\partial\Lambda}{\partial\nu_l}Z'\right) \\ &\quad +\frac{1}{2}\operatorname{tr}\left(W(X\ Z)H^{-1}(X\ Z)'WZ\frac{\partial\Lambda}{\partial\nu_j}Z'W(X\ Z)H^{-1}(X\ Z)'WZ\frac{\partial\Lambda}{\partial\nu_l}Z'\right). \end{aligned} \tag{29}$$

First we note, that the second and the third term in (29) are equal and therefore only one of them has to be evaluated. Shifting some matrices inside the traces yields expressions that involve only H^{-1}, the derivatives of Λ and submatrices of the SSCP-matrix (24)

$$\begin{aligned} F^*_{jl} &= \frac{1}{2}\operatorname{tr}\left(Z'WZ\frac{\partial\Lambda}{\partial\nu_j}Z'WZ\frac{\partial\Lambda}{\partial\nu_l}\right) - \operatorname{tr}\left(Z'W(X\ Z)H^{-1}(X\ Z)'WZ\frac{\partial\Lambda}{\partial\nu_j}Z'WZ\frac{\partial\Lambda}{\partial\nu_l}\right) \\ &\quad +\frac{1}{2}\operatorname{tr}\left(Z'W(X\ Z)H^{-1}(X\ Z)'WZ\frac{\partial\Lambda}{\partial\nu_j}Z'W(X\ Z)H^{-1}(X\ Z)'WZ\frac{\partial\Lambda}{\partial\nu_l}\right). \end{aligned}$$

Now the largest matrix involved in the computation of F^*_{jl} is H^{-1} which is of dimension $(p + q \times p + q)$. Further reduction can be achieved for the model described in (4) and (5). Here the above expression reduces to

$$\begin{aligned} F^*_{jl}(\vartheta) &= \frac{1}{2}\operatorname{tr}(Z_l'WZ_jZ_j'WZ_l) - \operatorname{tr}(Z_l'W(X\ Z)H^{-1}(X\ Z)'WZ_jZ_j'WZ_l) \\ &\quad \frac{1}{2}\operatorname{tr}(Z_l'W(X\ Z)H^{-1}(X\ Z)'WZ_jZ_j'W(X\ Z)H^{-1}(X\ Z)'WZ_l). \end{aligned}$$

Now we turn to the elements of the expected Fisher-information that involve derivatives with respect to the dispersion parameter. Again the general form of these elements is given by

$$F^*_{r+1,j} = F^*_{j,r+1} = \frac{1}{2}\operatorname{tr}\left(P\frac{\partial V}{\partial\nu_j}P\frac{\partial V}{\partial\phi}\right) \tag{30}$$

$j = 1, \ldots, r$ and

$$F^*_{r+1,r+1} = \frac{1}{2}\operatorname{tr}\left(P\frac{\partial V}{\partial\phi}P\frac{\partial V}{\partial\phi}\right). \tag{31}$$

Using the derivatives in (21) and (26) together with the definition of P in (20), we may rewrite

(30) as

$$
\begin{aligned}
F^*_{j,r+1} &= \frac{1}{2}\,\mathrm{tr}\left(PZ\frac{\partial\Lambda}{\partial\nu_j}Z'P\frac{1}{\phi}W^{-1}\right)\\
&= \frac{1}{2\phi}\,\mathrm{tr}\left(WZ\frac{\partial\Lambda}{\partial\nu_j}Z'WW^{-1}\right) - \frac{1}{2\phi}\,\mathrm{tr}\left(WZ\frac{\partial\Lambda}{\partial\nu_j}Z'W(X\ Z)H^{-1}(X\ Z)'WW^{-1}\right)\\
&\quad -\frac{1}{2\phi}\,\mathrm{tr}\left(W(X\ Z)H^{-1}(X\ Z)'WZ\frac{\partial\Lambda}{\partial\nu_j}Z'WW^{-1}\right)\\
&\quad +\frac{1}{2\phi}\,\mathrm{tr}\left(W(X\ Z)H^{-1}(X\ Z)'WZ\frac{\partial\Lambda}{\partial\nu_j}Z'W(X\ Z)H^{-1}(X\ Z)'WW^{-1}\right)\\
&= \frac{1}{2\phi}\,\mathrm{tr}\left(Z'WZ\frac{\partial\Lambda}{\partial\nu_j}\right) - \frac{1}{\phi}\,\mathrm{tr}\left((X\ Z)'WZ\frac{\partial\Lambda}{\partial\nu_j}Z'W(X\ Z)H^{-1}\right)\\
&\quad +\frac{1}{2\phi}\,\mathrm{tr}\left((X\ Z)'W(X\ Z)H^{-1}(X\ Z)'WZ\frac{\partial\Lambda}{\partial\nu_j}Z'W(X\ Z)\right)
\end{aligned}
$$

and for the special model in (4) and (5) we obtain

$$
\begin{aligned}
F^*_{j,r+1}(\vartheta) &= \frac{1}{2\phi}\,\mathrm{tr}\left(Z_j'WZ_j\right) - \frac{1}{\phi}\,\mathrm{tr}\left((X\ Z)'WZ_jZ_j'W(X\ Z)H^{-1}\right)\\
&\quad +\frac{1}{2\phi}\,\mathrm{tr}\left((X\ Z)'W(X\ Z)H^{-1}(X\ Z)'WZ_jZ_j'W(X\ Z)\right).
\end{aligned}
$$

The expression in (31) may be rewritten as

$$
\begin{aligned}
F^*_{r+1,r+1} &= \frac{1}{2}\,\mathrm{tr}\left(P\frac{\partial V}{\partial\phi}P\frac{\partial V}{\partial\phi}\right)\\
&= \frac{1}{2\phi^2}\,\mathrm{tr}\left(PW^{-1}PW^{-1}\right)\\
&= \frac{1}{2\phi^2}\,\mathrm{tr}\left(WW^{-1}WW^{-1}\right) - \frac{1}{2\phi^2}\,\mathrm{tr}\left(W(X\ Z)H^{-1}(X\ Z)'WW^{-1}WW^{-1}\right)\\
&\quad -\frac{1}{2\phi^2}\,\mathrm{tr}\left(WW^{-1}W(X\ Z)H^{-1}(X\ Z)'WW^{-1}\right)\\
&\quad +\frac{1}{2\phi^2}\,\mathrm{tr}\left(W(X\ Z)H^{-1}(X\ Z)'WW^{-1}W(X\ Z)H^{-1}(X\ Z)'WW^{-1}\right)\\
&= \frac{n}{2\phi^2} - \frac{1}{\phi^2}\,\mathrm{tr}\left((X\ Z)'W(X\ Z)H^{-1}\right)\\
&\quad +\frac{1}{2\phi^2}\,\mathrm{tr}\left((X\ Z)'W(X\ Z)H^{-1}(X\ Z)'W(X\ Z)H^{-1}\right).
\end{aligned}
$$

To estimate the variance parameters $\vartheta = (\nu', \phi)'$ we can now proceed as follows: Beginning with some starting values $\vartheta^{(0)}$ iterate

$$\vartheta^{(k+1)} = \vartheta^{(k)} + F^*(\vartheta^{(k)})^{-1}s^*(\vartheta^{(k)})$$

until convergence.

Since the estimation of the regression coefficients depends on the variance parameters and vice versa both may be estimated through cycling between the two estimation steps, using the current values of $\hat{\delta}$ and $\hat{\vartheta}$ in the respective formulae. Compare also the following section, which describes this strategy in form of an algorithm.

5 Simultaneous estimation

In this section we will summarize the presented results in an algorithm that describes the simultaneous estimation of both regression coefficients and variance parameters.

(i) Choose starting values $\hat{\delta}^{(0)}$, $\hat{\vartheta}^{(0)}$, a termination criterion ϵ and define the iteration index $k = 0$.

(ii) Compute the working observations

$$\tilde{y} = X\hat{\beta}^{(k)} + Z\hat{b}^{(k)} + D^{-1}(y - \mu),$$

the working weights

$$W = D\Sigma^{-1}D$$

and determine $\hat{\delta}^{(k+1)} = (\hat{\beta}^{(k+1)\prime}, \hat{b}^{(k+1)\prime})'$ as the solution of the system of linear equations

$$\begin{pmatrix} X'WX & X'WZ \\ Z'WX & Z'WZ + \Lambda^{-1} \end{pmatrix} \begin{pmatrix} \hat{\beta}^{(k+1)} \\ \hat{b}^{(k+1)} \end{pmatrix} = \begin{pmatrix} X'W\tilde{y} \\ Z'W\tilde{y} \end{pmatrix}$$

where all involved matrices are evaluated at the current estimates.

(iii) Compute the score vector $s^*(\hat{\vartheta}^{(k)}) = (s_j^*)$, $j = 1, \ldots, r+1$ through

$$\begin{aligned} s_j^* &= -\frac{1}{2}\,\mathrm{tr}\left(Z'WZ\frac{\partial\Lambda}{\partial\nu_j}\right) + \frac{1}{2}\,\mathrm{tr}\left(Z'W(X\ Z)H^{-1}(X\ Z)'WZ\frac{\partial\Lambda}{\partial\nu_j}\right) \\ &+\frac{1}{2}(\tilde{y} - X\hat{\beta}^{(k)} - Z\hat{b}^{(k)})'WZ\frac{\partial\Lambda}{\partial\nu_j}Z'W(\tilde{y} - X\hat{\beta}^{(k)} - Z\hat{b}^{(k)}) \end{aligned}$$

$j = 1, \ldots, r$ and

$$\begin{aligned} s_{r+1}^* &= -\frac{n}{2\hat{\phi}^{(k)}} + \frac{1}{2\hat{\phi}^{(k)}}\,\mathrm{tr}\left((X\ Z)'W(X\ Z)H^{-1}\right) \\ &+\frac{1}{2\hat{\phi}^{(k)}}(\tilde{y} - X\hat{\beta}^{(k)} - Z\hat{b}^{(k)})'W(\tilde{y} - X\hat{\beta}^{(k)} - Z\hat{b}^{(k)}) \end{aligned}$$

and the expected Fisher-information $F^*(\hat{\vartheta}^{(k)}) = (F_{jl}^*)$, $j, l = 1, \ldots, r+1$ through

$$\begin{aligned} F_{jl}^* &= \frac{1}{2}\,\mathrm{tr}\left(Z'WZ\frac{\partial\Lambda}{\partial\nu_j}Z'WZ\frac{\partial\Lambda}{\partial\nu_l}\right) - \mathrm{tr}\left(Z'W(X\ Z)H^{-1}(X\ Z)'WZ\frac{\partial\Lambda}{\partial\nu_j}Z'WZ\frac{\partial\Lambda}{\partial\nu_l}\right) \\ &+\frac{1}{2}\,\mathrm{tr}\left(Z'W(X\ Z)H^{-1}(X\ Z)'WZ\frac{\partial\Lambda}{\partial\nu_j}Z'W(X\ Z)H^{-1}(X\ Z)'WZ\frac{\partial\Lambda}{\partial\nu_l}\right). \end{aligned}$$

$j, l = 1, \ldots, r$

$$\begin{aligned} F_{j,r+1}^* &= \frac{1}{2\hat{\phi}^{(k)}}\,\mathrm{tr}\left(Z'WZ\frac{\partial\Lambda}{\partial\nu_j}\right) - \frac{1}{\hat{\phi}^{(k)}}\,\mathrm{tr}\left((X\ Z)'WZ\frac{\partial\Lambda}{\partial\nu_j}Z'W(X\ Z)H^{-1}\right) \\ &+\frac{1}{2\hat{\phi}^{(k)}}\,\mathrm{tr}\left((X\ Z)'W(X\ Z)H^{-1}(X\ Z)'WZ\frac{\partial\Lambda}{\partial\nu_j}Z'W(XZ)\right), \end{aligned}$$

$j = 1, \ldots, r$ and

$$\begin{aligned} F_{r+1,r+1}^* &= \frac{n}{2\hat{\phi}^{(k)2}} - \frac{1}{\hat{\phi}^{(k)2}}\,\mathrm{tr}\left((X\ Z)'W(X\ Z)H^{-1}\right) \\ &+\frac{1}{2\hat{\phi}^{(k)2}}\,\mathrm{tr}\left((X\ Z)'W(X\ Z)H^{-1}(X\ Z)'W(X\ Z)H^{-1}\right). \end{aligned}$$

Update $\hat{\vartheta}^{(k+1)}$ via

$$\vartheta^{(k+1)} = \vartheta^{(k)} + F^*(\vartheta^{(k)})^{-1} s^*(\vartheta^{(k)}).$$

(iv) Compute the distance measures

$$d(\hat{\delta}^{(k)}, \hat{\delta}^{(k+1)}) = \frac{||\hat{\delta}^{(k+1)} - \hat{\delta}^{(k)}||}{||\hat{\delta}^{(k)}||}$$

and

$$d(\hat{\vartheta}^{(k)}, \hat{\vartheta}^{(k+1)}) = \frac{||\hat{\vartheta}^{(k+1)} - \hat{\vartheta}^{(k)}||}{||\hat{\vartheta}^{(k)}||}.$$

If $d(\hat{\delta}^{(k)}, \hat{\delta}^{(k+1)}) > \epsilon$ or $d(\hat{\vartheta}^{(k)}, \hat{\vartheta}^{(k+1)}) > \epsilon$ replace k by $k+1$ and go back to (ii). Otherwise stop the estimation process.

If we assume the special structure for Λ and Z as described in (4) and (5), step (iii) has to be replaced by

(iii) Compute the score vector $s^*(\hat{\vartheta}^{(k)}) = (s_j^*)$, $j = 1, \ldots, r+1$ through

$$\begin{aligned} s_j^* &= -\frac{1}{2} \operatorname{tr}\left(Z_j' W Z_j\right) + \frac{1}{2} \operatorname{tr}\left(Z_j' W (X\ Z) H^{-1} (X\ Z)' W Z_j\right) \\ &\quad + \frac{1}{2} (\tilde{y} - X\hat{\beta} - Z\hat{b})' W Z_j Z_j' W (\tilde{y} - X\hat{\beta} - Z\hat{b}) \end{aligned}$$

$j = 1, \ldots, r$ and

$$\begin{aligned} s_{r+1}^* &= -\frac{n}{2\hat{\phi}^{(k)}} + \frac{1}{2\hat{\phi}^{(k)}} \operatorname{tr}\left((X\ Z)' W (X\ Z) H^{-1}\right) \\ &\quad + \frac{1}{2\hat{\phi}^{(k)}} (\tilde{y} - X\hat{\beta}^{(k)} - Z\hat{b}^{(k)})' W (\tilde{y} - X\hat{\beta}^{(k)} - Z\hat{b}^{(k)}) \end{aligned}$$

and the expected Fisher-information $F^*(\hat{\vartheta}^{(k)}) = (F_{jl}^*)$, $j, l = 1, \ldots, r+1$ through

$$\begin{aligned} F_{jl}^*(\vartheta) &= \frac{1}{2} \operatorname{tr}(Z_l' W Z_j Z_j' W Z_l) - \operatorname{tr}(Z_l' W (X\ Z) H^{-1} (X\ Z)' W Z_j Z_j' W Z_l) \\ &\quad \frac{1}{2} \operatorname{tr}(Z_l' W (X\ Z) H^{-1} (X\ Z)' W Z_j Z_j' W (X\ Z) H^{-1} (X\ Z)' W Z_l). \end{aligned}$$

$j, l = 1, \ldots, r$

$$\begin{aligned} F_{j,r+1}^*(\vartheta) &= \frac{1}{2\hat{\phi}^{(k)}} \operatorname{tr}\left(Z_j' W Z_j\right) - \frac{1}{\hat{\phi}^{(k)}} \operatorname{tr}\left((X\ Z)' W Z_j Z_j' W (X\ Z) H^{-1}\right) \\ &\quad + \frac{1}{2\hat{\phi}^{(k)}} \operatorname{tr}\left((X\ Z)' W (X\ Z) H^{-1} (X\ Z)' W Z_j Z_j' W (X\ Z)\right). \end{aligned}$$

$j = 1, \ldots, r$ and

$$\begin{aligned} F_{r+1,r+1}^* &= \frac{n}{2\hat{\phi}^{(k)2}} - \frac{1}{\hat{\phi}^{(k)2}} \operatorname{tr}\left((X\ Z)' W (X\ Z) H^{-1}\right) \\ &\quad + \frac{1}{2\hat{\phi}^{(k)2}} \operatorname{tr}\left((X\ Z)' W (X\ Z) H^{-1} (X\ Z)' W (X\ Z) H^{-1}\right). \end{aligned}$$

Update $\hat{\vartheta}^{(k+1)}$ via

$$\vartheta^{(k+1)} = \vartheta^{(k)} + F^*(\vartheta^{(k)})^{-1} s^*(\vartheta^{(k)}).$$

References

Breslow, N. E. & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models, *Journal of the American Statistial Association* **88**: 9–25.

Fahrmeir, L., Kneib, T. & Lang, S. (2003). Penalized additive regression for space-time data: A Bayesian perspective, *Statistica Sinica (under revision)* . Available from www.stat.uni-muenchen.de/∼kneib.

Fahrmeir, L. & Tutz, G. (2001). *Multivariate Statistical Modelling based on Generalized Linear Models*, Springer, New York.

Harville, D. A. (1974). Bayesian inference for variance components using only error contrasts, *Biometrika* **61**: 383–385.

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems, *Journal of the American Statistical Association* **72**: 320–338.

Lin, X. & Zhang, D. (1999). Inference in generalized additive mixed models by using smoothing splines, *Journal of the Royal Statistical Society Series B* **61**: 381–400.

Patterson, H. D. & Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal, *Biometrika* **58**: 545–554.

Tierney, L. & Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities, *Journal of the American Statistical Association* **81**: 82–86.

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