Tutz:

Generalized semiparametrically structured mixed models

Sonderforschungsbereich 386, Paper 251 (2001)

Online unter: http://epub.ub.uni-muenchen.de/

Projektpartner

Generalized semiparametrically structured mixed models

Gerhard Tutz
Ludwig-Maximilians-Universität München
Akademiestraße 1, 80799 München
tutz@stat.uni-muenchen.de

Abstract

Generalized linear mixed models are a common tool in statistics which extends generalized linear models to situations where data are hierarchically clustered or correlated. In this article the simple but often inadequate restriction to a linear form of the predictor variables is dropped. A class of semiparametrically structured models is proposed in which the predictor decomposes into components that may be given by a parametric term, an additive form of unspecified smooth functions of covariates, varying-coefficient terms or terms which vary smoothly (or not) across the repetitions in a repeated measurement design. The class of models is explicitly designed as an extension of multivariate generalized mixed linear models such that ordinal responses may be treated within this framework. The modelling of smooth effects is based on basis functions like e.g. B-splines or radial basis functions. For the estimation of parameters a penalized marginal likelihood approach is proposed which may be based on integration techniques like Gauss-Hermite quadrature but may as well be used within the more recently developed nonparametric maximum likelihood approaches. For the maximization of the penalized marginal likelihood the EM-algorithm is adapted. Moreover, an adequate form of cross-validation is developed and shown to work satisfactorily. Various examples demonstrate the flexibility of the class of models.

Keywords: Generalized linear mixed models; Generalized semiparametrically structured mixed models; EM-algorithm; Ordinal mixed models; Cross-validation; Nonparametric maximum likelihood.

1 Introduction

In many studies samples are clustered or correlated. The clustering may be due to repeated measurements as in longitudinal studies or to subsampling from the primary sampling units in cross-sectional studies. One approach to modelling such data is the incorporation of random effects for the subjects or clusters into the linear predictor. Under the assumption of an exponential family type of distribution the incorporation of random effects extends generalized linear models (GLMs) to generalized linear mixed models (GLMMs).

The main problem in GLMMs is that the marginal distribution of the response obtained by integrating out the random effects, does not have closed form. This led to the development of several methods to obtain analytical approximation for the likelihood, like numerical integration based on Gauss-Hermite quadrature (e.g. Hinde, 1982, Anderson & Aitkin, 1985) or Monte Carlo techniques within the EM-algorithm (McCulloch, 1994, McCulloch, 1997, Booth & Hobert, 1999) or approximation methods as Taylor expansions or Laplace approximation (e.g. Breslow & Clayton, 1993, Wolfinger & O'Connell, 1993, Longford, 1994). A more recent approach is nonparametric maximum likelihood which avoids the assumption of a fixed distribution for the random effects (Aitkin, 1996, Aitkin, 1999). Most of the articles consider unidimensional, binomial or Poisson distributed, responses. Extensions to ordinal models have been considered by Harville & Mee (1984), Jansen (1990), Tutz & Hennevogl (1996) and more recently by Hartzel, Agresti & Caffo (2001).

The basis of GLMMs is the linear predictor which restricts the influence of covariates to a strictly parametric form. In regression models there is a wide body of literature where the strict parametric form is extended to more flexible forms of semi- and nonparametric regression. Overviews are given by Hastie & Tibshirani (1990), Green & Silverman (1994), Schimek (2000), for discrete data see Simonoff (1996), for multivariate responses see Fahrmeir & Tutz (2001).

Nonparametric approaches to correlated data for normally distributed responses have been considered by Rice & Silverman (1991) and for non-Gaussian

models in a generalized estimation equation framework by Wild & Yee (1996) and Berhane & Tibshirani (1998). As far as mixed methods are concerned there has been some development for nonparametric time functions within linear mixed models (Zeger & Diggle, 1994, Zhang, Lin, Raz & Sowers, 1998), but very limited work has been done to extend nonparametric regression techniques to mixed models for non-Gaussian responses. An important step in this direction is the introduction of generalized additive mixed models (GAMMs) given by Lin & Zhang (1999). They use smoothing splines to obtain smooth estimates of covariate effects. A fully Bayesian approach has been recently proposed by Fahrmeir & Lang (2001)

In this paper generalized semiparametrically structured mixed models (GSSMMs) will be considered within a multivariate framework including ordinal response models. The term "structured" means that the structure of the predictor is determined by several components, containing parametric parts as well as nonparametric parts like additive terms (which itself represents a structured form of nonparametric functions). The assumed structure determines how parametric and nonparametric parts are connected. This may be in an additve form as in partially linear models but it may also be in a multiplicative form as in varying coefficients models. In order to illustrate the type of data for which these approaches are adequate in the following potential applications are given.

Application 1: Infectious disease

Zeger & Karim (1991) considered longitudinal data on respiratory infection in Indonesian children. 275 children were examined up to six consecutive quarters. The response is respiratory infection (1: yes, 0: no), covariates were years, xerophthalmia status (1: yes, 0: no), which is an ocular manifestation of chronic vitamin A deficiency (1: yes, 0: no), gender (1: female, 0: male), height for age and presence of stunting (1: yes, 0: no).

Application 2: Injuries to the knee

In a clinical study focusing on the healing of sports related injuries of the knee $n = 123$ patients have been treated. By random design one of two therapies were chosen. In the treatment group an anti-inflammatory spray was used while in the placebo group a spray without active ingredients was used. After ten days of treatment with the spray and at three further occasions, the mobility of the knee was investigated in a standardized experiment during which the knee was actively moved by the patient. The pain occurring during the movement was assessed on a five point scale ranging from no pain to severe pain (for further information on the data see Tutz (2000). In addition to treatment the covariate age was measured.

These examples represent repeated measurements and contain at least one metric variable for which it is doubtful that the effect is linear or quadratic as would usually be assumed within a generalized linear mixed models framework. In application 2 adequate modelling should account for the ordinal response variable. Otherwise the number of parameters and smooth components would be unnecessarily high. In particular when subjective assessments represent the response as in application 2 it has to be expected that heterogenity is very high because the scaling of pain level will vary across individuals.

In Section 2 first the basic GLMMs are introduced which are appropriate for unidimensional responses like binary or Poisson data as well as multivariate extensions which are adequate for ordinal responses. Moreover, semiparametric extensions in the form of additive modelling are considered. In Section 3 the concept of penalized marginal likelihood estimation is introduced. The basic idea is to represent the smooth components as a finite sum of basis functions which are connected to knots on an equally spaced grid and to penalize the marginal likelihood by restricting the differences of coefficients that correspond to these basic functions. For common likelihood based regression problems concepts of this type have been considered in the form of P(enalized) Splines (e.g. Eilers & Marx,1996, Ruppert & Carroll,1999). The essential difference to these approaches

is that now one has a penalized marginal likelihood which can be given only in integral form. It is shown how the marginal likelihood can be maximized directly by Monte Carlo or Gauss Hermite approximation or indirectly by use of EM techniques. Moreover, a simple estimation concept for standard errors and confidence intervals is developed. In section 4 the estimation concepts are validated in a simulation study and compared to the double penalized quasi-likelihood approach of Lin & Zhang (1999). Section 5 gives some extensions to varying-coefficients modelling where interaction effects may be modelled nonparametrically.

2 Generalized structured random effects model

2.1 Linear random effects models

Before considering more complex models, in this section multivariate generalized linear models are reconsidered. They provide a general framework which allows not only unidimensional responses like binomial or Poisson distributed responses but include the multinomial case which in particular allows parsimonious modelling of ordinal response data.

Let the data be given by (y_{it}, x_{it}), $i = 1, \ldots, n$, $t = 1, \ldots, T$, with y_{it} denoting the q-dimensional response connected to observation t in cluster i and x_{it} denoting a vector of covariates which may vary across the observations within one cluster. Often the clusters correspond to individuals and the observations to repeated measurements, therefore the index t. Only for the simplicity of presentation the number of observations within one cluster T does not depend on the cluster. A generalized linear random effects model is specified by two components. First, it is assumed that the conditional density of y_{it} given the explanatory variable x_{it} *and* the random effects b_i is of the exponential family type

$$f\left(y_{it}|x_{it}\,, b_i\right) = \exp\left(y_{it}^T\theta_{it} - \kappa(\theta_{it})\right)/\phi + c(y_{it},\ \phi) \tag{1}$$

where θ_{it} denotes the natural parameter and $\kappa(.)$ the log normalization constant. The second component specifies the link between response and the covariates.

The structural assumption specifies the conditional mean by

$$\mu_{it} = \mathrm{E}(y_{it}|x_{it},\, b_i) = h\left(Z_{it}\beta + W_{it}b_i\right) \tag{2}$$

where $h : \mathbb{R}^q \mapsto \mathbb{R}^q$ is the response function and Z_{it}, W_{it} are design matrices composed from x_{it}.

The multivariate setting is chosen deliberately to incorporate ordinal models. If $Y_{it} \in \{1, \ldots, k\}$ is an ordinal response, the actual response is $y_{it}^T = (y_{it1}, \ldots, y_{itq}), q = k - 1$, with $y_{itr} = 1$ if $Y_{it} = r$ and $y_{itr} = 0$ otherwise. The most often used ordinal model is the proportional odds model (McCullagh 1980) which in its basic form is given by

$$\mathrm{P}\left(Y_i \leq r|x_i\right) = F(\gamma_{0r} + x_i^T\gamma)$$

where F is the logistic distribution function. The random effects formulation with random intercept is given by

$$\mathrm{P}\left(Y_{it} \leq r|x_{it},\, b_i\right) = F\left(\gamma_{0r} + x_{it}^T\gamma + b_i\right). \tag{3}$$

It is easily seen that the model has the form (2) by considering $\pi_{it}^T = (\pi_{it1}, \ldots, \pi_{itq})$, $\pi_{itr} = \mathrm{P}\left(Y_{it} = r|x_{it},\, b_i\right)$ as the mean response μ_{it} and specifying $Z_{it} = \left(I_{q\times q}, 1_{q\times 1}^T \otimes x_{it}\right)$, $1_{q\times 1} = (1, \ldots, 1)$, $W_{it} = 1_{q\times 1}$, $\beta^T = \left(\gamma_{01}, \ldots, \gamma_{0q}, \gamma^T\right)$. The q-dimensional response function h is also found easily (see e.g. (Fahrmeir & Tutz 2001)).

An alternative ordinal model is the sequential model or continuation ratio model which with random intercepts incorporated has the form

$$\mathrm{P}(Y_{it} = r|Y_{it} \geq r,\, x_{it},\, b_i) = F(\gamma_{0r} + x_{it}^T\gamma + b_i) \tag{4}$$

where F is an unidimensional distribution function which often is chosen as the logistic distribution function. The interpretation of the model is different from that of a cumulative model. Model (4) may be seen as a process model with start in category one and consecutive modelling of binary transitions to higher categories. The relationship to cumulative type models has been investigated by Läärä & Matthews (1985), Tutz (1991), Greenland (1994). Model (4) is again of the general form (2) but with several advantages. The linear predictor in the

cumulative model (3) has to be ordered, thus $\gamma_{0r} \leq \cdots \leq \gamma_{0q}$ has to hold. This ordering which in more complex models causes problems is not postulated in the sequential model (4). Moreover, the sequential model which basically is a model for binary transitions may be estimated by software which handles binary mixed regression models.

Univariate generalized mixed models ($q = 1$) comprise among others the binary logistic model, the Poisson model with log link, the normal model and Gamma response models. The binary logistic model with random intercept is given by

$$\mu_{it} = \mathrm{E}(y_{it}|x_{it},\, b_i) = F(\gamma_0 + x_{it}^T\gamma + b_i)$$

where $y_{it} \in \{0, 1\}$, $Z_{it} = \left(1, x_{it}^T\right)$, $\beta_0^T = \left(\gamma_0, \gamma^T\right)$ and F the logistic distribution function. A more general form where the slope may vary across clusters is given by

$$\mu_{it} = \mathrm{E}(y_{it}|x_{it},\, b_i) = F(\gamma_0 + x_{it}^T\gamma + (1,\, x_{itr})b_i)$$

where $b_i^T = \left(b_{i1},\, b_{i2}^T\right)$ represents the random effects.

The specification of the random effects model is completed by specifying the distribution of the random effect b_i. Here two approaches will be considered. First the widely used assumption of a fixed continuous distribution. Thus in addition to the additive distributional assumption (1) and the structural assumption (2) it is assumed that the cluster specific effects are independently identically distributed, with $\mathrm{E}(b_i) = 0$, $\mathrm{cov}(b_i) = Q$. A common assumption is that b_i is a multivariate normal random variable, i.e. $b_i \sim N(0, Q)$. The second approach treats the mixing distribution in a fully nonparametric way as a finite mixture yielding nonparametric maximum likelihood estimates (see Hinde & Wood, 1987, Aitkin & Aitkin, 1996, Aitkin, 1999). The approach avoids the assumption of a specific parametric form because of the possible sensitivity of conclusions to this specification (see also Heckman & Singer, 1984). Aitkin (1999) gives an extensive review of current approaches to the modelling of mixture distributions and demonstrates the usefulness of the purely nonparametric approach.

The usual assumption for random effects models is independence of observations within and between clusters given the random effects, i.e. $f(y_1, \ldots, y_n|b_1, \ldots, b_n) =$

$\prod_{i=1}^{n} f(y_i|b_i)$ with $f(y_i|b_i) = \prod_{t=1}^{T} f(y_{it}|b_i)$ where $y_i^T = (y_{i1}, \ldots, y_{iT})$.

2.2 The extension to semiparametric modelling

The linear predictor $\eta_{it}^T = (\eta_{it1}, \ldots, \eta_{itq})$ for model (2) has the form

$$\eta_{it} = \eta_{it}^S + \eta_{it}^R$$

where $\eta_{it}^S = Z_{it}\beta$ is the structured term and $\eta_i^R = W_{it}b_i$ is the random term which is connected to the random effects. The structured term for random intercepts models of the type considered in Section 2 has the components

$$\eta_{itr}^S = \gamma_{0r} + x_{it}^T\gamma,\ r = 1, \ldots, q.$$

In addition to the covariates x_{it} let now $w_{it}^T = (w_{it1}, \ldots, w_{itm})$ be an additional vector of continuous variables. Instead of using a parametric term to incorporate w_{it} a more flexible term is

$$\eta_{itr}^S = \gamma_{0r} + x_{it}^T\gamma + \sum_{j=1}^{m} \alpha_{(j)}(w_{itj})$$

where $\alpha_{(j)}$ are unspecified functions of the jth component of W. Thus the structured term decomposes into

$$\eta_{itr}^S = \eta_{itr}^L + \eta_{itr}^A \tag{5}$$

where $\eta_{itr}^L = \gamma_{0r} + x_{it}^T\gamma$ is the linear term and $\eta_{it}^A = \sum_{j=1}^{m} \alpha_{(j)}(w_{ij})$ is the non-parametric additive term. That means part of the covariates, in particular all of the categorical covariates are determined in a parametric linear form whereas the continuous variables, or part of them, are allowed to influence the outcome additively with the components being not further restricted. The model is referred to as a partial linear additive model.

For univariate models the values in (5) are unidimensional and the index may be dropped since $r = 1$. For the ordinal models the representation is restricted to the case where covariates x_{it} have weights γ that do not depend on the category. Only the intercept is category-specific, meaning that the design matrix in the linear term η_{it} has the form $Z_{it} = (I_{q\times q},\ 1_{q\times 1} \otimes x_{it}^T)$. More general models will be considered in Section 4.

3 Estimation by penalized marginal likelihood

3.1 Penalized marginal likelihood

The structural parameters in the GLMM may be estimated by maximization of the marginal likelihood which is given by

$$l(\beta,\, Q) = \sum_{i=1}^{n} \log \int f(y_i|b_i)p(b_i;\, Q)db_i.$$

For the semiparametric model this approach cannot be used directly, since the nonparametric parts of the model are functions instead of parameters. One way to obtain a parametric form is the use of regression splines. If one assumes that the knots are known and fixed, regression splines are easy to handle. But usually knots are not given. Then, however, besides the number of knots their location has to be determined, for example by forward and backward procedures, see e.g. Friedman & Silverman (1989), Friedman (1991), Stone, Hansen, Kooperberg & Truong (1997) for approaches in common regression. Alternatively, here a basis function approach is used together with discrete penalization in the spirit of P-splines (for penalized) as suggested by Eilers & Marx (1996), Ruppert & Carroll (1999), Ruppert (2000), for previous work see also Whittaker (1923), O'Sullivan (1986, 1988).

It is assumed that the smooth components $\alpha_{(j)}$ may be represented as a sum of basis functions

$$\alpha_{(j)}(w) = \sum_{s=1}^{M_j} \alpha_{js}\Phi_{js}(w) \tag{6}$$

where $\Phi_{js}(w)$, $s = 1, \ldots, m$, are basis functions connected to specific knots on an equally spaced grid. Let $w_{j(1)} < \cdots < w_{j(M_j)}$ denote the knots then Φ_{js} may be a B-spline basis (e.g. De Boor, 1978), power functions or radial basis functions of the type $\Phi_{js}(w) = \Phi\left(|w - w_{j(s)}|\right)$ where Φ is a smooth fully specified function, e.g. the Gaussian kernel function or thin plate splines with $\Phi\left(|w - w_{j(s)}|\right) = (w - w_{j(s)})^2 \log\left(|w - w_{j(s)}|\right)$.

The basic concept is to use a number of knots M_j that is high enough to ensure flexibility that is sufficient to approximate all functions of potential interest and then penalize the variation of weights $\alpha_{j1}, \ldots, \alpha_{jM_j}$. Ruppert (2000)

demonstrates nicely how a very limited number of knots, say 30, yields approximations to quite complex functions which are visually not distinguishable from the function itself.

Instead of the marginal likelihood with the parameters $\beta, Q, \{\alpha_{js}\}$ now estimation is based on the *penalized marginal likelihood* (PML)

$$l_P\left(\beta,\, Q,\, \{\alpha_{js}\}\right) = \sum_{i=1}^{n} \log \int f(y_i|b_i) p(b_i;\, Q) db_i - \frac{1}{2} \sum_{j=1}^{m} \lambda_j \sum_{j=d+1}^{M_j} \left(\Delta^d \alpha_{js}\right)^2 \quad (7)$$

where Δ is the difference operator operating on adjacent coefficients of the basis function, i.e. $\Delta\alpha_{js} = \alpha_{js} - \alpha_{j,s-1}$, $\Delta^2\alpha_{js} = \Delta(\alpha_{js} - \alpha_{j,s-1}) = \alpha_{js} - 2\alpha_{j,s-1} - \alpha_{j,s-2}$, etc. For simplicity the same order of the difference d is used for all of the components in covariate w. Although one may experiment with different orders of the penalty, satisfying results are usually obtained for $d \leq 3$. A special case, used by Ruppert & Carroll (1999) is $d = 0$ where the penalty reduces to $\Delta^0\alpha_{js} = d_{js}$. P-splines which are defined by use of a B-spline basis in (6) may fit polynomials exactly. Whatever the values of λ_j, if one uses B-splines of degree k or higher and the order of penalty is $k + 1$ a polynomial of degree k may be fitted exactly. If $\lambda_j \to \infty$ the limit of a P-splines fit is always a polynomial. If the penalty is of order k and $\lambda_j \to \infty$ the fitted curve will approach a polynomial of degree $k - 1$, given the degree of the B spline is equal to, or higher than, k (for details see Eilers & Marx, 1996). The number of knots seems not to be of significant influence. Ruppert (2000) investigated the influence of the number of knots and found that in the case of monotonic functions the number of knots has little effect. For more complex functions there is a minimal number of knots and low mean averaged squared errors are found whenever the number of knots exceeds the minimal number of knots. Highlighting on generalized cross validation for the selection of the number of knots Ruppert (2000) found a default of $\min\{n/4, 35\}$ to perform very well.

Ruppert (2000) used a slightly different version for the smooth functions which makes the connection to polynomial fitting more obvious. When using

$$\alpha_{(j)}(w) = a_{j1}w + \cdots + a_{jk}w^k + \sum_{s=1}^{M_j} \alpha_{js}(w - w_{j(s)})_+^k \quad (8)$$

with $(w)_+ = w^k I(w \leq 0)$ it is easily seen that $\alpha_{(j)}$ is a kth degree polynomial on each interval between two consecutive knots and has $(k-1)$ continuous derivatives everywhere, but the kth derivative takes a jump of size $k!\alpha_{js}$ at knot $w_{j(s)}$. When using (8) the roughness penalty is placed on the jumps α_{js}, $s = 1, \ldots, M_j$, in the kth derivative of $\alpha_{(j)}$ by use of $\lambda_j \sum_{s=1}^{M_j} \alpha_{js}^2$.

When maximizing the penalized marginal likelihood the predictor of the ith observation has the form

$$\eta_{it} = Z_{it}\beta + \Phi_i \alpha + W_{it} b_i \tag{9}$$

where the first component corresponds to the effects of x_{it} and the second component corresponds to the additive structure with Φ_i given by

$$\Phi_i = (1_{q\times 1} \otimes \text{Diag}\left(\Phi_1^T(w_{it1}), \ldots, \Phi_m^T(w_{itm})\right)$$

with $\Phi_j^T(w_{itj}) = (\Phi_{j1}(w_{itj}), \ldots, \Phi_{jM_j}(w_{itj}))$ and

$$\alpha^T = (\alpha_1^T, \ldots, \alpha_m^T), \qquad \alpha_j^T = (\alpha_{j1}, \ldots, \alpha_{jM_j}).$$

The third component is the same as in generalized linear random effects models. When using the truncated series representation (8) the matrix Φ_i may be used but with $\Phi_{js}(w)$ defined as the truncated power series, i.e. $\Phi_{js}(w) = (w - w_{j(s)})_+^k$. Then the first polynomial part in (8) is taken into the first component in (9) by using instead of $Z_{it}\beta$ the form $\tilde{Z}_{it}\tilde{\beta}$ with

$$\begin{aligned} \tilde{Z}_{it} &= \left(Z_{it},\ 1_{q\times 1} \otimes \text{Diag}(w_{it1}, \ldots, w_{it1}^k, \ldots, w_{itm}, \ldots, w_{itm}^k)\right), \\ \tilde{\beta}^T &= (\beta^T, a_{11}, \ldots, a_{1k}, \ldots, a_{m1}, \ldots, a_{mk}). \end{aligned}$$

For derivations it is useful to have the penalty term in matrix form. With D_j being the $(M_j - 1) \times M_j$ contrast matrix

$$D_j = \begin{pmatrix} -1 & 1 & & \\ & -1 & 1 & \\ & & \ddots & \ddots \\ & & & -1 \quad 1 \end{pmatrix}$$

one obtains $\sum_{s=1}^{M_j} \Delta(\alpha_{js})^2 = \sum_{s=1}^{M_j} (\alpha_{js} - \alpha_{js-1})^2 = \alpha_j^T D_j^T D_j \alpha_j$ where $\alpha_j^T = (\alpha_{j1}, \ldots, \alpha_{jM_j})$. By using $D_j^d = D_j D_j^{d-1}$ and $K_j = \left(D_j^d\right)^T D_j^d$ the total marginal likelihood (7) is given by

$$l_P\left(\beta,\, Q,\, \{\alpha_{js}\}\right) = \sum_{i=1}^{n} \log \int f(y_i|b_i) p(b_i;\, Q) db_i - \frac{1}{2} \sum_{j=1}^{m} \lambda_j \sum_{s=d+1}^{M_j} \alpha_j^T K_j \alpha_j. \quad (10)$$

For the truncated series approach (8) the matrix K_j is even simpler since it has diagonal form given by $K_j = I_{M \times M}$.

The penalized marginal likelihood used here should not be confused with penalized quasi-likelihood as used by Breslow & Clayton (1993). Here the penalization refers to the smooth component whereas in Breslow & Clayton (1993) the focus is on shrinkage estimation of the random effects and penalization refers to this shrinkage. A combination of the approach of Breslow & Clayton (1993) and penalized smooth curves yields double penalization (see Lin & Zhang (1999) in the context of smoothing splines).

3.2 Maximizing the penalized marginal likelihood

The marginal likelihood (10) depends only on the structural parameters of the model. These are given by β, α and $Q = \mathrm{cov}(b_i)$. Let q be decomposed by $Q = Q^{1/2} Q^{T/2}$ where $Q^{1/2}$ denotes the left Cholesky factor. By simple matrix algebra the linear predictor (9) may be written in the usual linear form

$$\begin{aligned} \eta_{it} &= Z_{it}\beta + \Phi_i \alpha + W_{it} Q^{\frac{1}{2}} e_i \\ &= \left[Z_{it}, \Phi_i, e_i^T \otimes W_{it}\right] \begin{bmatrix} \beta \\ \alpha \\ \theta \end{bmatrix} \end{aligned} \quad (11)$$

where $e_i \sim N(0, I)$ is the standardized random variable and $\theta = \mathrm{vec}\left(Q^{1/2}\right)$ For univariate random effects the Kronecker product simplifies to $e_i W_{it}$ and $\theta = \sqrt{\mathrm{var}(b_i)}$. By utilizing (11) all of the structural parameters are collected in $\delta^T = (\beta^T,\, \alpha^T,\, \theta^T)$.

For the maximization of the penalized likelihood l_P some approximation procedure is necessary. There are two essentially differing procedures: the direct

approach that uses a Monte Carlo or Gauss Hermite approximation of l_P (e.g. Hedeker & Gibbons, 1994, Hartzel, Liu & Agresti, 2000) or indirect approaches based on the EM algorithm (e.g. Anderson & Aitkin, 1985, Booth & Hobert, 1999). Direct approaches work well in simple cases, for example GLMMs where only the intercept is random, but for more complex models the EM algorithm is a more robust alternative.

In the following the direct approach is shortly sketched since some of the derived terms will be needed later. Let $v_g, z_g, g = 1, \ldots, G$, denote the masses and quadrature points respectively. More precisely one has $v_g = \tilde{v}_g/\sqrt{\pi}$, $z_g = \sqrt{2}\tilde{z}_g$ with $\tilde{z}_g$ denoting the gth zero of the Hermite polynomial of degree G and $\tilde{z}_g$ the corresponding weight found e.g. in Abramowitz & Stegun (1972). Then the PML is approximated by

$$l_P(\delta) = \sum_{i=1}^{n} \log \left\{ \sum_{g=1}^{G} v_g \prod_{t=1}^{T} f(y_{it}|z_g;\, \delta) \right\} - \frac{1}{2} \sum_{j=1}^{M} \lambda_j \alpha_j^T K_j \alpha_j \tag{12}$$

where the random effects e_i are replaced by the known quadrature points, or more precise $f(y_{it}|z_g, \delta)$ has predictor value $\eta_{itg} = Z_{it}\beta + \Phi_i \alpha + W_{it} Q^{1/2} z_g$. It should be noted that if the random effect is vector valued z_g is the quadrature point in multinominal Gauss-Hermite quadrature and is also a vector.

The score function derived from Gauss-Hermite approximation to the penalized marginal likelihood (7) is obtained by

$$\begin{aligned} s_P(\delta) = \partial l_P(\delta)/\partial\delta = \\ = \sum_{i=1}^{n} \sum_{g=1}^{G} c_{ig}(\delta) \partial \log f(y_i|z_g;\delta)/\partial\delta + \partial \left(\sum_{j=1}^{m} \lambda_j \alpha_j^T K_j \alpha_j \right) /\partial\delta \\ = \sum_{i=1}^{n} \sum_{g=1}^{G} c_{ig}(\delta) \sum_{t=1}^{T} Z_{itg}^T \frac{\partial h(\eta_{itg})}{\partial \eta} \Sigma_{itg}^{-1} (y_{it} - \mu_{itg}) - D\delta \end{aligned} \tag{13}$$

where

$$\begin{aligned} Z_{itg} &= \left[Z_{it}, \Phi_i, z_g^T \otimes W_{it} \right], \\ c_{ig}(\delta) &= v_g f(y_i|z_g;\alpha) / \sum_{s=1}^{M} v_s f(y_i|z_s, \alpha) \end{aligned}$$

and

$$D = \text{Diag}\left(0_{p\times p}, \lambda_1 K_1, \ldots, \lambda_m K_m, 0_{|\theta|\times|\theta|}\right)$$

where $0_{p\times p}$ is a $p \times p$ matrix of zeros (p denoting the dimension of β) and $0_{|\theta|\times|\theta|}$ is a matrix of zeros with dimension corresponding to θ. Moreover, Σ_{itg}, μ_{itg} denote evaluation at $\eta_{itg} = Z_{it}\beta + \Phi_i\alpha + z_g^T \otimes W_{it}\theta$. This is equivalent to the score function of a weighted multivariate GLM. For the derivation of the first term see Fahrmeir & Tutz (2001), Section 7.4.

The problem with (13) is that the weights $c_{ig}(\delta)$ depend on the parameter. Thus the straightforward use of the GLM framework is not possible. For example the Fisher matrix which is based on the second derivatives of l_P is difficult to obtain for complex models and therefore the usual Fisher scoring algorithm is not available. An alternative that can be used is the quasi-Newton algorithm (Hartzel, Agresti & Caffo, 2001). Hartzel, Agresti & Caffo (2001) also reduced the number of quadrature points by extending adaptive Gauss-Hermite (Liu & Pierce, 1994, Pinheiro & Bates, 1995) to multicategorical data.

The alternative indirect approach which is based on the EM algorithm avoids some of these problems. Since it is used in applications it is given more explicitely. In the E-step of the $(p+1)$th cycle one has to determine

$$\begin{aligned} \text{M}\left(\delta|\delta^{(p)}\right) &= E\left\{\log f\left(Y, E;\, \delta\right)|Y;\, \delta^{(p)}\right)\right\} \\ &= \int \log\left(f(Y, E;\, \delta)\right) f\left(E|Y,\, \delta^{(p)}\right) dE \end{aligned}$$

where

$$\log f(Y, E;\, \delta) = \sum_{i=1}^{n} \log f(y_i|e_i,\, \delta) + \sum_{i=1}^{n} \log\left(g(a_i)\right) + \frac{1}{2}\sum_{j=1}^{m} \lambda_j \alpha_j^T K_j \alpha_j$$

is the complete penalized data log likelihood with $Y = (y_1, \ldots, y_n)$ denoting the observed data and $E = (e_1, \ldots, e_n)$ denoting the unobserved data; g is the mixture distribution of the standardized random effects e_i. Since the posterior has the simple form

$$f\left(E|y, \delta^{(p)}\right) = \prod_{i=1}^{n} f\left(y_i|e_i, \delta^{(p)}\right) \prod_{i=1}^{n} g(e_i) \Big/ \prod_{i=1}^{n} \int f\left(y_i|e_i,\, \delta^{(p)}\right) g(e_i) de_i$$

$\mathrm{M}\left(\delta|\delta^{(p)}\right)$ simplifies to

$$\mathrm{M}\left(\delta|\delta^{(p)}\right) = \sum_{i=1}^{n} k_i^{-1} \int \left[\log f(y_i|e_i,\delta) + \log g(e_i)\right] f\left(y_i|e_i,\delta^{(p)}\right) g(\delta_i) de_i + \frac{1}{2}\sum_{j=1}^{m} \lambda_j \alpha_j^T K_j \alpha_j$$

where $k_i = \int f(y_i|e_i\delta^{(p)})g(e_i)$ does not depend on δ.

The integral in $M\left(\delta|\delta^{(p)}\right)$ may again be approximated in several ways: In a Monte Carlo type algorithm it is approximated by the mean over drawings from $N(0,1)$, in a Gauss-Hermite type approximation which is used in the following one has the approximation

$$\mathrm{M}\left(\delta|\delta^{(p)}\right) \approx \tilde{\mathrm{M}}\left(\delta|\delta^{(p)}\right) = \sum_{i=1}^{n}\left\{\sum_{g=1}^{G} c_{ig}\left(\log f(y_i|z_g;\delta) + \log g(z_g)\right)\right\} - \frac{1}{2}\sum_{j=1}^{m} \lambda_j \alpha_j^T K_j \alpha_j \tag{14}$$

where z_g are the G Gaussian quadrature points and the weights

$$c_{ig} = v_j f\left(y_i|z_g;\ \delta^{(p)}\right) \Big/ \sum_{s=1}^{G} f\left(y_i|z_s;\ \delta^{(p)}\right)$$

contain the masses v_g which correspond to the quadrature points z_g. The beauty of this approximation is that $\mathrm{M}(\delta|\delta^{(p)})$ again corresponds to the penalized weighted log-likelihood of a generalized linear model and therefore maximization (the M step of the EM algorithm) is simply realized within the framework of GLMs. Thus the method used by Hinde (1982), Anderson & Hinde (1988) of using the weighted log-likelihood can be extended to the penalized marginal likelihood of multivariate GSSMMs. Details of the easily implemented algorithm are given in the Appendix.

3.3 Nonparametric maximum likelihood estimates

The nonparametric maximum likelihood approach (Aitkin & Francis,1998, Aitkin, 1999) may be extended to a penalized nonparametric likelihood approach in a similar way. Let $v^T = (v_1, \ldots, v_G), \sum_i v_i = 1$, be the vector which characterizes the

proportions of the finite mixture on the mass points given by $z^T = (z_1, \ldots, z_G)$. That means v_g is the mixture proportion of mass point z_g. Instead of the predictor (11) which is based on random effects $e_i \sim N(0, I)$ now the density $f(y_{it}|z_g, \delta)$ is determined by the predictor

$$\eta_{itg} = Z_{it}\beta + \Phi_i\alpha + W_{it}z_g = Z_{itg}\delta_n$$

where z_g is one mass point from the vector $z^T = (z_1^T, \ldots, z_G^T)$ and $Z_{itg} = [Z_{it}, \Phi_i, 0, \ldots, W_{it}, \ldots]$. The corresponding vector of structural parameters $\delta_n^T = (\beta^T,\ \alpha^T,\ z^T,\ v^T)$ now includes the mass points z and the mixture proportions v. To distinguish the nonparametric parameters as well as design matrices from previous terminology sub- or superscript n is used. In the corresponding penalized likelihood the finite mixture distribution is given by

$$l_P(\delta_n) = \sum_{i=1}^{n} \log \left\{ \sum_{g=1}^{G} v_j \prod_{t=1}^{T} f(y_{it}|z_g, \delta_n) \right\}.$$

Let δ_n be split into $\delta_{n,1}^T = (\beta^T,\ \alpha^T, z^T)$ and $\delta_{n,2} = v$. Then the derivative is obtained by

$$\frac{\partial l_P(\delta_n)}{\partial \delta_{n,1}} = \sum_{i=1}^{n} \sum_{g=1}^{G} c_{ig}(\delta_n) \sum_{t=1}^{T} Z_{itg} \frac{\partial h(\eta_{itg})}{\partial \eta} \Sigma_{ig}^{-1}(y_{it} - \mu_{itg}) - D_n\delta_n$$

where $D_n = \mathrm{Diag}(0_{p\times p}, \lambda_1 K_1, \ldots, \lambda_m K_j, 0_{2m\times 2m})$ and

$$\frac{\partial l_P(\delta_n)}{\partial v_j} = \sum_{i=1}^{n} \frac{f(y_i|z_g, \delta_n) - f(y_i|z_G, \delta_n)}{\sum_s f(y_i|z_s, \delta_n)} = \left(\frac{c_{ig}(\delta_n)}{v_j} - \frac{c_{iG}(\delta_n)}{v_G} \right),$$

$j = 1, \ldots, G-1$. In the latter equation it has been used that $v_G = 1 - \sum_{g\neq G} v_g$. The total score function is given by $s_P(\delta_n) = \big(\partial l_P(\delta_n)/\partial \delta_{n,1}^T, \partial l_P(\delta_n)/\partial v_1, \ldots, \partial l_P(\delta_n)/\partial v_{G-1}\big)$. The latter part of equation $s_P(\delta_n) = 0$ yields $\hat{v}_g = \frac{1}{nT}\sum_{i=1}^{n} c_{ig}(\hat{\delta}_n)$. Thus only the first part $\partial l_P(\delta_n)/\partial \delta_{n,1} = 0$ has to be solved by finding a solution $\hat{\delta}_{n,1}$.

3.4 Inference

Inference may be based on the score function $s_P(\delta)$. However, the function $s_P(\delta)$ is not a score function in the usual sense since

$$\mathrm{E}(s_P(\delta)) = D\delta$$

which is not equal to zero. Let $s_P(\delta)$ be decomposed into

$$s_P(\delta) = s(\delta) - D\delta$$

with $s(\delta)$ denoting the first term in (13) which corresponds to a weighted multivariate GLM. Then a first order Taylor approximation yields

$$(\delta - \hat{\delta}) = \left(-\frac{\partial s_P(\delta)}{\partial \delta^T}\right)^{-1} s_P(\delta) = -\left(\frac{\partial s(\delta)}{\partial \delta^T} - D\right)^{-1} (s(\delta) - D\delta).$$

Using the approximation $\mathrm{cov}(s(\delta) - D\delta)) = \mathrm{cov}(s(\delta)) \approx \sum_{i=1}^{n} s_i(\hat{\delta}) s_i(\hat{\delta})^T = S(\hat{\delta})$ where $s_i(\delta) = \sum_{g=1}^{G} c_{ig}(\delta) \sum_{i=1}^{T} Z_{itg}^T (\partial h(\eta_{itg})/\partial \eta) \Sigma_{itg}^{-1} (y_{it} - \mu_{itg})$ and

$$\mathrm{E}\left(-\frac{\partial s(\delta)}{\partial \delta^T}\right) = \mathrm{cov}(s(\delta)) \approx \sum_{i=1}^{n} s_i(\hat{\delta}) s_i(\hat{\delta})^T = S(\hat{\delta})$$

one obtains as approximation of $\mathrm{cov}(\hat{\delta})$ the sandwich matrix

$$\widehat{\mathrm{cov}}(\hat{\delta}) = \left(S(\hat{\delta}) + D\right)^{-1} S(\hat{\delta}) \left(S(\hat{\delta}) + D\right)^{-1}. \tag{15}$$

For the special case $\lambda_1 = \cdots = \lambda_m = 0$ which corresponds to a strictly parametric setting without smoothing one obtains $D = 0$ and therefore

$$\mathrm{cov}(\hat{\delta}) = S(\hat{\delta})^{-1}$$

which is equivalent to the proposal of Gourieroux & Monfort (1989) for the computation of standard errors in mixed models.

The simple form (15) avoids the tedious task of deriving the derivative of $s_P(\delta)$ which corresponds to the second derivative of the likelihood. The complexity of these derivations has been the reason to use the EM algorithm instead of a direct method based on these derivations.

3.5 Choice of smoothing parameters

One way of selecting the smoothing parameters is cross-validation. The basic idea is to leave one observation out and measure the discrepancy between this observation and its prediction based on the remaining $n - 1$ observations. After this is done for every single observation the total discrepancy is built. Since

observations within one cluster are correlated leaving one out means to leave out one cluster at a time (see also Rice & Silverman, 1991). The observation of one cluster is given by $y_i^T = (y_{i1}, \ldots, y_{iT})$ drawn from the marginal distribution (given Z_{it}, Φ_i, W_{it} but not the random effect b_i). An appropriate measure for a wide range of distributions is the Kullback-Leibler discrepancy $\mathrm{KL}(f_{\psi^*}, f_\psi) = \mathrm{E}_{\psi^*} \log(f_{\psi^*}/f_\psi)$ where ψ^*, ψ are parameters that determine the densities f_{ψ^*}, f_ψ. If the first argument in $\mathrm{KL}(.,.)$ is replaced by the degenerated distribution δ_i which puts mass one on observation vector y_i and the second argument by the estimated density one obtains

$$\mathrm{KL}(\delta_i, \hat{f}) = -\log f(y_i; \hat{\eta}_i)$$

with $f(y_i; \hat{\eta}_i) = \int \prod_{i=1}^T f(y_{it}; \hat{\eta}_{it}) db_i$ where the predictor $\hat{\eta}_i^T = (\hat{\eta}_{i1}, \ldots, \hat{\eta}_{iT})$ is given by $\hat{\eta}_{it} = Z_{it}\hat{\beta} + \Phi_i\hat{\alpha} + W_{it}b_i$. For the evaluation of $f(y_i; \hat{\eta}_i)$ the same Gauss-Hermite or nonparametric approximation is used as in (12). One obtains the cross-validation criterion

$$\mathrm{CV}(\{\lambda_j\}) = -\sum_{i=1}^n \log f(y_i; \hat{\eta}_i)$$

which implicitely takes the correlation between $y_{i1}, \ldots, y_{it}$ into account. For categorical data with $q+1$ response categories one has $y_i^T = (y_{i1}^T, \ldots, y_{iT}^T)$ with components $y_{it}^T = (y_{i1}, \ldots, y_{i,q+1})$ and CV has the familiar form. In the less time consuming form of m-fold cross validation the data are partitioned into m blocks of approximately n/m observation and the discrepancy between observations and prediction is computed for one block with estimates based on the remaining $m-1$ blocks.

4 Applications and simulations

4.1 Application to infectious disease data

For the binary response variable in the infectious disease data set (Application 1 in Section 1) Lin & Zhang (1999) considered the semiparametric logistic model

$$\mathrm{logit}\,(\mathrm{P}(y_{it} = 1|b_i) = x_{it}^T\gamma + \alpha(\mathrm{age}_{it}) + b_i$$

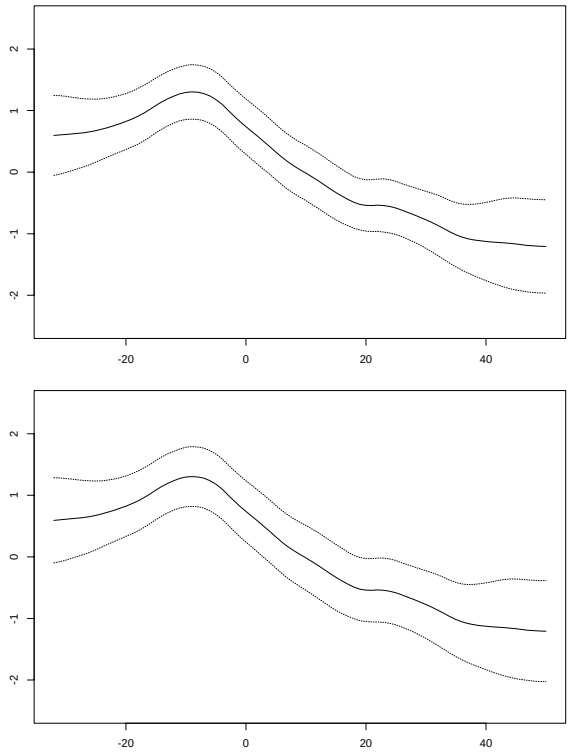

Figure 1: Estimated effect of age for infectious disease data based on $GH(8)$ (above) and $NPML(3)$ (below) with estimated 0.95 confidence bounds

where x_{it} contains an intercept, xerophthalmia, seasonal sine and cosine, gender height and stunting. For comparison the same model is fitted with α specified as a B-spline basis of degree 2 with 20 knots and differences of first order. The model was fitted by Gauss-Hermite quadrature and nonparametric maximum likelihood, abbreviated by $GH(G)$ and $NPML(G)$ where G denotes the number of quadrature or mass points used. Table 1 shows the parameter estimates for two specific choices, $GH(8)$ and $NPML(3)$; in addition the double penalized quasi-likelihood estimates ($DPQL$) from Lin & Zhang (1999) are given. It is seen that estimates are quite comparable. The same holds for Gauss-Hermite with $G \geq 6$ and $NPML$ with $G \geq 3$ (not given). Although (frequentist) standard errors from the $DPQL$ approach are slightly smaller than for the other approaches the conclusions on the relevance of covariates are substantially the same.

The estimated effect of age which was modeled smoothly is given in Fig. 1 with smoothing parameter chosen by cross validation. The curves which are based on Gauss-Hermite quadrature and nonparametric maximum likelihood are also

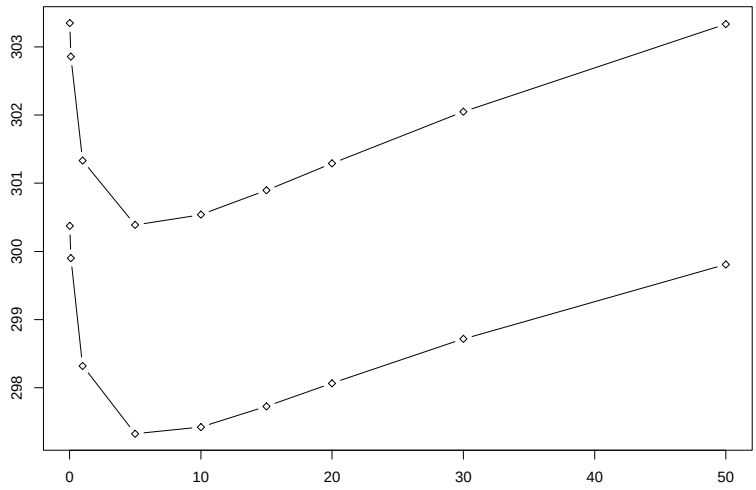

Figure 2: 11-fold cross-validation based on Kullback-Leibler loss for infectious disease data (GH(8): upper curve, NPML(3): lower curve)

quite comparable to the curves given in Lin & Zhang (1999). The same holds for the standard errors for Gauss-Hermite quadrature ($G \geq 6$) and nonparametric approaches with less or equal four mass points and $DPQL$. For the nonparametric approach standard errors are somewhat increased if more than four points are used. Usually for more than three or four mass points only three of them have weights which can be considered as non-zero. This was also the case in the analysis of the present data set.

The curves in Figure 1 are based on 11-fold cross validation. Fig. 2 shows that cross validation works well yielding distinct minima. The smoothing parameters which yields the minimal value is about the same for Gauss-Hermite and nonparametric techniques. However, since the nonparametric approach is more flexible the distance between observation and fitted value is smaller yielding smaller distances than for the Gauss-Hermite procedure.

4.2 Simulation study

In a small simulation study the underlying model was

$$\text{logit}(\text{P}(y_{it} = 1|b_i) = \gamma_0 + x_i\gamma + \alpha(w_i) + b_i$$

where $\gamma_0 = 0$, $\gamma = 1$, $\alpha(w) = \sin(w)$, $b_i \sim N(0, \sigma^2), \sigma^2 = 1$. The binary variable x_i takes value 0.5 for half of the subjects and -0.5 for the other half from $n = 80$

	$GH(8)$		$NPML(3)$		Lin/Zhang	
Intercept	-3.009	(0.232)	-2.934		-2.92	(0.23)
Vitamin A	0.502	(0.544)	0.545	(0.535)	0.52	(0.46)
Seasonal sine	-0.152	(0.214)	-0.126	(0.234)	-0.16	(0.17)
Seasonal cosine	-0.594	(0.175)	-0.575	(0.179)	-0.58	(0.17)
Gender	-0.521	(0.251)	-0.483	(0.244)	-0.50	(0.24)
Height	-0.030	(0.026)	-0.043	(0.027)	-0.03	(0.02)
Stunted	0.424	(0.465)	0.299	(0.473)	0.39	(0.43)
$\hat{\sigma}$	0.888	(0.505)				

Table 1: Estimates for infectious disease data based on Gauss-Hermite quadrature, the nonparametric maximum likelihood approach and double penalized quasi-likelihood, the latter taken from Lin & Zhang (1999)

subjects. For each subject $T = 5$ observations were drawn. The estimates are based on a B-spline basis of degree 2 with 40 knots and differences of first order. Table 2 gives the estimated fixed effects and the empirical and estimated standard errors for Gauss-Hermite based techniques ($GH(8)$) and the nonparametric maximum likelihood approach ($NPML(3)$). It is seen that both approaches worked quite well in estimating the fixed parameters. For the nonparametric approach $\hat{\gamma}_0$ and $\hat{\sigma}$ were computed from the estimated mass points and the corresponding weights. Moreover, estimated standard errors are quite comparable to empirical standard errors for fixed effects. Only in the case of σ which reflects the heterogeneity the estimated standard error is distinctly larger than the empirical standard error. Fig. 3 depicts the true curve and the mean estimated curve. The variation of estimates are seen from box plots. In addition, estimated and empirical standard errors are given by plotting mean curve $\pm 1.96\hat{\sigma}$. For both approaches the estimates work quite well and estimated standard errors reflect the underlying empirical error well.

Gauss-Hermite (GH(8))			
Parameters	Mean	Empirical SE	Estimated SE
γ_0	-0.016	0.145	0.141 (0.013)
γ_1	0.971	0.240	0.245 (0.023)
σ	1.088	0.246	0.471 (0.082)
Nonparametric maximum likelihood (NPML(3))			
γ_0	-0.015	0.172	
γ_1	0.988	0.241	0.246 (0.024)
σ	0.897	0.330	

Table 2: Means and standard errors over 100 replications

5 Extensions

In Section 2.2 the flexibility of the predictor has been increased by specifying a partial linear form of the predictor. This extension represents a basic form of semiparametrically structured smoothing of random effects models which has been used to derive the estimation concept. In the following several extension of the predictor form are given which may be treated within the general estimation framework.

5.1 Smooth components varying across observations

In many applications the variables w_{tj} in the additive part of the predictor will be cluster-level covariates i.e. $w_{tj} = w_j$, and the index t for the subject or the repeated measurement can be dropped. In order to avoid identification problems in the following w_j is considered to be a cluster-level variable and moreover only one smooth component is incorporated

The first extension reflects that the effect of a variable is not necessarily constant across the observations within a cluster. If the observations are repeated

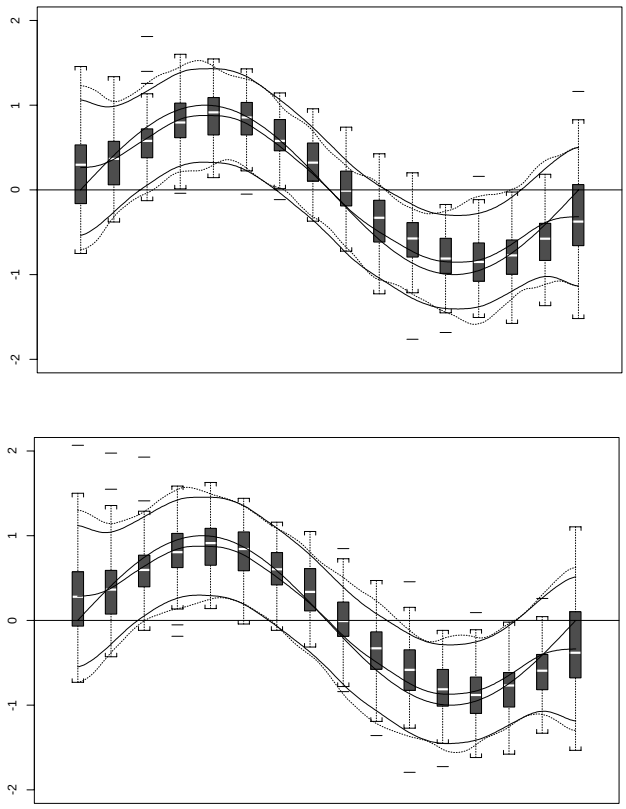

Figure 3: True and estimated nonparametric functions with empirical standard errors (drawn) and estimated standard errors (dotted) for Gauss Hermite(GH(8), upper picture), and nonparametric maximum likelihood (NP(3), lower picture)

measurements it is well conceivable that the effect decreases or increases across time. Then the corresponding unspecified function of w has the form $\alpha^{(t)}(w)$. That means w has different influence for each observation within a cluster. In a parametric setting one may specify $\alpha^{(t)}(w) = w\alpha_t$ where α_t is a coefficient which varies across time if observations are repeated measurements. Then the model may be considered as a mixed model with varying coefficients and can be fitted with local smoothing techniques (Tutz, 1999).

In the nonparametric case only in very large data sets (T small, n very large) one has enough information to use the form $\alpha^{(t)}(w)$ in full generality. If there is

enough information the function $\alpha^{(t)}$ can be represented in the form

$$\alpha^{(t)}(w) = \sum_{s=1}^{M} \alpha_s^{(t)} \Phi_s(w) \tag{16}$$

where the use of Φ_s means that the same knots are used for each t, but the weights $\alpha_s^{(t)}$ obviously have to depend on t. To reduce this general approach so that it works also for reasonable sample sizes one can decompose the weights into

$$\alpha_s^{(t)} = \alpha^{(t)} + \alpha_s \tag{17}$$

where the first term depends only on the index t which stands for the observation within a cluster and the second term depends on the knot index s. Simplification (17) yields

$$\alpha_{(t)}^{(t)}(w) = \alpha^{(t)} \sum_{s=1}^{M} \Phi_s(w) + \sum_{s=1}^{M} \alpha_s \Phi_s(w).$$

If Φ_s are chosen as B-splines $c = \sum_s \Phi_s(w)$ is a constant which can be omitted and one obtains

$$\alpha^{(t)}(w) = \alpha^{(t)} + \sum_{s=1}^{M} \alpha_s \Phi_s(w).$$

Thus $\alpha(w) = \sum_{s=1}^{M} \alpha_s \Phi_s(w)$ is a function which represents the basic effect of w not depending on the observation t and $\alpha^{(t)}$ represents the shifting of the function $\alpha(w)$ connected to observation t. Thus $\alpha^{(t)}$ is just an intercept for observation t. The latter form represents an extension which is easily incorporated into the framework of the model as given in Section 2 by replacing the predictor $\eta_{it} = Z_{it}\beta + \Phi_i\alpha + W_{it}b_i$ by

$$\eta_{it} = Z_{it}\beta + \Phi_t\alpha_V + \Phi_i\alpha + W_{it}b_i$$

where

$$\Phi_t = (I_{q\times 1} \otimes 1_t^T)$$

with

$$1_t^T = (0, \ldots, 1, \ldots, 0) \text{ of length T and 1 at position } t,$$
$$\alpha_V^T = (\alpha^{(1)}, \ldots, \alpha^{(T)}).$$

The vector α_V may be treated as a fixed effect and is incorporated into the linear term by considering the predictor $\eta_{it} = \tilde{Z}_{it}\tilde{\beta} + \Phi_i\alpha + W_{it}b_i$ with $\tilde{Z}_{it} = (Z_{it},\, \Phi_t)$, $\tilde{\beta}^T = (\beta^T,\, \alpha_V^T)$. Then the estimation procedure developed in Section 3 applies directly.

However, if T is large, for example in panel studies, even the simplification (17) may not work well because the number of parameters is too high. Then the parameters $\alpha^{(1)}, \ldots, \alpha^{(T)}$ which are connected to variable w vary strongly across $t = 1, \ldots, T$ and reflect noise instead of structure. This effect may be avoided by penalizing the variation of $\alpha^{(t)}$ across t by using the extended roughness penalty

$$\begin{aligned} \kappa\left(\{\alpha_s\}, \{\alpha^{(t)}\}\right) = \frac{1}{2}\lambda \sum_{s=d+1}^{M_j} \left(\Delta_s^d \alpha_s\right)^2 \\ + \frac{1}{2}\tilde{\lambda} \sum_{s=d+1}^{T} \left(\Delta_t^d \alpha^{(t)}\right)^2 \end{aligned} \tag{18}$$

where the subindex in Δ_t^d refers to the index which is used in the penalization, i.e. $\Delta_t \alpha^{(t)} = \alpha^{(t)} - \alpha^{(t-1)}$. The new second penalty term penalizes the variation of $\alpha^{(t)}$ across $t = 1, \ldots, T$. The incorporation of the extended penalization is a straightforward exercise which adds no new structure to the estimation and inference concepts derived in Section 3.

5.2 Varying-coefficients modelling

The extension to effects which vary across observations as given in section 5.1 may be seen in the more general framework of varying coefficients models. Hastie & Tibshirani (1993) extended the modelling of covariate effects by introducing a form of interaction of variables where the effect of one variable is modified smoothly by a so-called effect modifier. Let u_i be a set of variables which modify the effects of covariates x_{it}. Then the rth component of the predictor in its general form is given by

$$\eta_{itr} = \eta_{it}^L + \eta_{it}^A + \eta_{it}^V + \eta_{it}^R$$

where $\eta_{it}^L, \eta_{it}^A, \eta_{it}^R$ represent the linear, the additive and the random term and η_{it}^V is given by

$$\eta_{it}^V = \sum_j x_{itj} v_{(j)}(u_{ij})$$

where $v_{(j)}(u)$ is an unspecified function. Thus the effects of covariates j (or part of them) may vary with the variables u_{ij}. Since the functional form of $v_{(j)}(u)$ has not been determined a priori, this is a semiparametric model containing parametric parts (weighted variables x_{itj}) and nonparametric (unspecified functions u_{ij}). However, in contrast to partially linear models the structure is now multiplicative. Use of an expansion in basis functions for $v_{(j)}(u)$ and an additional penalty term in the marginal likelihood is straightforward. The models given in Section 5.1 may be seen as varying-coefficients models where the variation is across the repeated measurements. The corresponding varying-coefficients term is given by

$$\eta_{it}^V = \alpha^{(t)}$$

which is equivalent to a time variation of the intercept. For repeated measurements with many repetitions time variation is the essential interaction effect which may also have a semiparametric form. The term

$$\eta_{it}^V = x_{itj} \alpha_{(j)}^{(t)}$$

specifics that the effect of variable x_{itj} varies across time. In order to avoid problems of identification problems also the intercept should be specified as varying across time if $x_{itj}\alpha_{(j)}^{(t)}$ is included in the predictor.

In common regression modelling models which contain varying coefficients form have been considered by Eilers & Marx (1999), Tutz & Scholz (2000) and Kauermann & Tutz (2000). For mixed models Fahrmeir & Lang (2001) include varying coefficients within a fully Bayesian framework.

5.3 Application to knee injury data

The ordinal response in Application 2, Section 1, has five response categories which reflect the level of pain suffered when the knee is actively bent. Since the

	Cumulative model		Sequential model	
$\alpha^{(1)}$	4.529	(0.589)	4.361	(0.562)
$\alpha^{(2)}$	2.617	(0.541)	2.569	(0.520)
$\alpha^{(3)}$	1.462	(0.534)	1.431	(0.492)
γ_G	0.708	(0.337)	0.604	(0.331)
γ_T	−1.997	(0.286)	−2.047	(0.278)
σ	8.897	(0.678)	8.509	(0.660)

Table 3: Estimated parameters for knee injury data (standard errors in brackets)

pain level should decrease during therapy the response 1 stands for severe pain and 5 for no pain. In particular for the sequential model this is a more natural ordering although for symmetric distribution functions in the link the reverse ordering yields an equivalent model. The models considered are the cumulative model

$$\mathrm{P}(Y_{it} \leq r | x_{it}) = \eta_{itr}$$

and the sequential model

$$\mathrm{P}(Y_{it} = r | Y_{it} \leq r,\ x_{it}) = \eta_{itr}$$

where the nonparametric predictor has the form

$$\eta_{itr} = \gamma_{0r} + \alpha^{(t)} + \alpha(\mathrm{Age}) + \mathrm{treatment} \times \gamma_t + \mathrm{gender} \times \gamma_G$$

Table 3 gives the estimated parametric effects for both models where $\alpha^{(t)}$ has not been penalized but α(age) is estimated by penalization ($\lambda = 8$, B-splines of degree 2, 34 inner knots, first differences penalized).

It is seen that for both models the heterogeneity is rather strong reflecting that the scaling of pain levels is a subjective response that varies across individuals. If age is incorporated in quadratic form the effect is even stronger, yielding $\hat{\sigma} = 9.423$. Therapy is distinctly superior to placebo and there are ordered time effects which reduce the probability of high pain levels over time thereby reflecting continuity in improvement. Figure 4 shows the smoothly estimated age effect which obviously is nonlinear. In particular for people above 35 years of age the

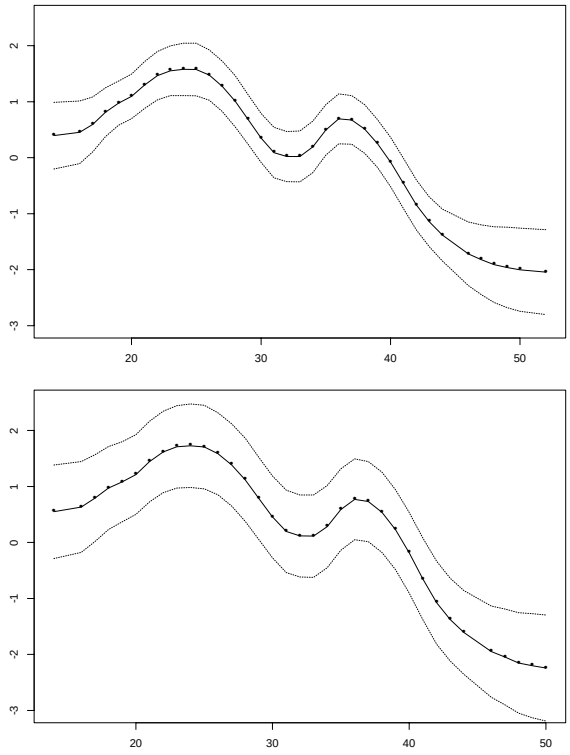

Figure 4: Estimated effect of age in knee injury data (above: sequential model, below: cumulative model)

probability for high pain levels is reduced, maybe reflecting higher tolerance for pain or different scaling of pain levels. In the study the focus is on the effect of therapy but of course the effect should be adjusted to potentially differing scaling of pain.

Since the same effect has been found in other data sets it is noteworthy that the confidence intervals for the cumulative model are slightly larger than for the sequential model.

6 Concluding remarks

A concept of generalized semiparametrically structured mixed models has been developed which allows to incorporate into the predictor additive terms of unknown functional form as well as nonparametric interactions in the form of vary-

ing coefficients. For additive models the approach represents an alternative to the double penalized approach by Lin & Zhang (1999). The latter is based on the Laplace approximation considered by Breslow & Clayton (1993). Although this approach may produce biased estimators for non-normal data (Breslow & Lin, 1995, Lin & Breslow, 1996) it encompasses nested and crossed designs whereas the present approach, because of the high dimensional integration needed in crossed design, is adequate only for nested designs. As is seen from the comparison with the approach of Lin & Zhang (1999) in this case the results are quite similar.

The penalized marginal likelihood approach considered here is wider by including ordinal response variables and more general semiparametric terms in the predictor. Moreover, it applies under normally distributed random effects as well as in the form of nonparametric penalized marginal likelihood. In applications the performance of these approaches was comparable if the number of mass points in the nonparametric approach is kept low. If the number is increased by including mass points which have low weights standard errors become larger.

The computation in the present approach makes use of GLM tools which makes implementation rather easy. The approach encompasses ordinal response models of the cumulative and sequential type. The latter is even easier to handle because it may be embedded into binary GLMs by considering transitions between adjacent categories. As a consequence it is more stable because no order restrictions in the predictor are necessary.

The approach can easily be extended to multicategorical mixed models with nominal categories which have been proposed recently by Hartzel, Agresti & Caffo (2001). For nominal models one has to estimate smooth functions for each of the categories. Thus estimation procedures are more complex but the same principles apply. For the simpler case of nominal data without mixing penalized functions approach have already been shown to work by Tutz & Scholz (2000).

Appendix

The M-step consists of maximizing $\tilde{M}(\delta|\delta^{(p)})$ from (14) which is equivalent to solving the equation $\partial\tilde{M}(\delta|\delta^{(p)})/\partial\delta = 0$. $\tilde{M}(\delta|\delta^{(p)})$ is given by

$$\tilde{M}\left(\delta|\delta^{(p)}\right) = \sum_{i=1}^{n}\sum_{g=1}^{G}\sum_{t=1}^{T} c_{ig}\log(y_{it}|z_g;\delta) + \log g(z_g) - \frac{1}{2}\sum_{j=1}^{m}\lambda_j\alpha_j^T K_j\alpha_j$$

where $y_{it}|z_g$ has predictor

$$\eta_{itg} = [Z_{it},\ \Phi_i,\ z_g^T \otimes W_{it}] \begin{bmatrix}\beta\\ \alpha\\ \theta\end{bmatrix}.$$

Thus one obtains

$$\begin{aligned}
\frac{\partial\tilde{M}\left(\delta|\delta^{(p)}\right)}{\partial\beta} &= \sum_{i=1}^{n}\sum_{g=1}^{G}\sum_{t=1}^{T} c_{ig} Z_{it}^T \frac{\partial h(\eta_{itg})}{\partial\eta}\Sigma_{itg}^{-1}(y_{it}-\mu_{itg}),\\
\frac{\partial\tilde{M}\left(\delta|\delta^{(p)}\right)}{\partial\alpha_j} &= \sum_{i=1}^{n}\sum_{g=1}^{G}\sum_{t=1}^{T} c_{ig} \Phi_i^T \frac{\partial h(\eta_{itg})}{\partial\eta}\Sigma_{itg}^{-1}(y_{it}-\mu_{itg}) - \lambda_j K_j\alpha_j,\\
\frac{\partial\tilde{M}\left(\delta|\delta^{(p)}\right)}{\partial\theta} &= \sum_{i=1}^{n}\sum_{g=1}^{G}\sum_{t=1}^{T} c_{ig} \left(z_g^T \otimes W_{it}\right)^T \frac{\partial h(\eta_{it|g})}{\partial\eta}\Sigma_{itg}^{-1}(y_{it}-\mu_{itg}),
\end{aligned}$$

where $\Sigma_{itg}(\mu_{itg})$ are the covariance (expectations) of y_{it} evaluated at $\eta_{itg} = Z_{it}\beta + \Phi_i\alpha^+ z_g^T \otimes W_{it}\theta$.

In closed form one obtains

$$\begin{aligned}
\frac{\partial\tilde{M}\left(\delta|\delta^{(p)}\right)}{\partial\delta} = &\sum_{i=1}^{n}\sum_{g=1}^{G}\sum_{t=1}^{T} c_{ig}\left(Z_{it}, \Phi_i, z_g^T \otimes W_{it}\right)^T \frac{\partial h(\eta_{it|g})}{\partial\eta}\Sigma_{itg}^{-1}(y_{it}-\mu_{itg})\\
&- \text{Diag}\left(0_{p\times p}, \lambda_1 K_1, \ldots, \lambda_m K_m, 0_{|\theta|\times|\theta|}\right)\left(\beta^T, \alpha^T, \theta^T\right)^T
\end{aligned}$$

where $0_{p\times p}$ is a $p\times p$-matrix of zeros (p denoting the dimension of β) and $0_{|\theta|\times|\theta|}$ is a matrix of zeros with dimensions corresponding to θ.

The first part corresponds to the weighted score function of a (multivariate) GLM with observations y_{it} given η_{itg} for $i = 1,\ldots,n, t = 1,\ldots,T, g = 1,\ldots,G$.

The second part is the block-diagonal matrix representing the penalty term. The corresponding Fisher matrix is given by

$$F\left(\delta|\delta^{(p)}\right) = \sum_{i=1}^{n}\sum_{g=1}^{G}\sum_{t=1}^{T} c_{ig}\left(Z_{it}, \Phi_i, z_g^T \otimes W_{it}\right)^T \frac{\partial h(\eta_{itg})}{\partial \eta} \Sigma_{itr}^{-1} \left(\frac{\partial h(\eta_{itg})}{\partial \eta}\right)^T \left(Z_{it}, \Phi_i, z_i^T \otimes W_{it}\right) - \mathrm{Diag}\left(0_{p\times p}, \lambda_1\kappa, \ldots, \lambda_m\kappa, 0_{|\theta|}\right).$$

Maximization of $\tilde{\mathrm{M}}(\delta|\delta^{(p)})$ may be obtained by pseudo-Fisher-scoring iterations

$$\tilde{\mathrm{M}}\left(\delta_{(s+1)}|\delta^{(p)}\right) = \mathrm{M}\left(\delta_{(s)}|\delta^{(p)}\right) + F^{-1}\left(\delta|\delta^{(p)}\right) \frac{\partial\, \mathrm{M}\left(\delta_{(s)}|\delta^{(p)}\right)}{\partial\delta}$$

The estimate $\delta_{(s+1)}$ which results after iterations $s = 1, 2, \ldots$ have converged represents the next value $\delta^{(p+1)}$ in the EM cycle. Thus one EM cycle contains a completed Fisher scoring algorithm.

Acknowledgment:

Support from the SFB 386 sponsored by Deutsche Forschungsgemeinschaft is gratefully acknowledged. The first version of this paper was written while the author was visiting the University of Florida, Gainesville. I am grateful for computational work done by Ludwig Heigenhauser.

References

Abramowitz, M. and Stegun, I. (1972). *Handbook of Mathematical Functions*. New York: Dover.

Aitkin, M. (1996). A general maximum likelihood analysis of overdispersion in generalized linear models. *Statistics and Computing* **6**, 251–262.

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

Aitkin, M. and Aitkin, I. (1996). A hybrid EM/Gauss-Newton algorithm for maximum likelihood in mixture distributions. *Statistics and Computing* **6**, 127–130.

Aitkin, M. and Francis, B. J. (1998). Fitting generalized linear variance component models by nonparametric maximum likelihood. *The GLIM Newsletter* **29** (in press).

Anderson, D. A. and Aitkin, M. (1985). Variance component models with binary response: Interviewer variability. *Journal of the Royal Statistical Society* **Ser. B 47**, 203–210.

Anderson, D. A. and Hinde, J. P. (1988). Random effects in generalized linear models and the EM algorithm. *Comm. Statist. A – Theory Methods* **17**, 3847–3856.

Berhane, K. and Tibshirani, R. (1998). Generalized additive models for longitudinal data. *The Canadian Journal of Statistics* **26**, 517–535.

Booth, J. G. and Hobert, J. P. (1999). Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *J. R. Statist. Soc* **B 61**, 265–285.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed model. *Journal of the American Statistical Association* **88**, 9–25.

Breslow, N. E. and Lin, X. (1995). Bias correction in generalized linear mixed models with a single component of dispersion. *Biometrika* **82**, 81–91.

De Boor, C. (1978). *A Practical Guide to Splines*. New York: Springer-Verlag.

Eilers, P. H. and Marx, B. D. (1999). Generalized linear additive smooth structures. Preprint.

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

Fahrmeir, L. and Lang, S. (2001). Bayesian inference for generalized additive mixed models based on Markov random field priors. *Applied Statistics* (to appear).

Fahrmeir, L. and Tutz, G. (2001). *Multivariate Statistical Modelling based on Generalized Linear Models* (2nd ed.). New York: Springer.

Friedman, J. (1991). Multivariate adaptive regression splines (with discussion). *Ann. Statist.* **19**, 1–14.

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

Gourieroux, C. and Monfort, A. (1989). Simulation based inference in models with heterogeneity. Document de Travail INSEE/ENSAE. 8902.

Green, D. J. and Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*. London: Chapman & Hall.

Greenland, S. (1994). Alternative models for ordinal logistic regression. *Statistics in Medicine* **13**, 1665–1677.

Hartzel, J., Agresti, A., and Caffo, B. (2001). Multinomial logit random effects models. Preprint.

Hartzel, J., Liu, I., and Agresti, A. (2000). Describing heterogenous effects in stratified ordinal contingency tables, with applications to multi-center clinical trials. *Computational Statistics & Data Analysis* (to appear).

Harville, D. A. and Mee, R. W. (1984). A mixed-model procedure for analyzing ordered categorical data. *Biometrics* **40**, 393–408.

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

Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society* **B 55**, 757–796.

Heckman, J. J. and Singer, B. (1984). A method for minimizing the impact of distributional assumptions in econometric models of duration. *Econometrica* **52**, 271–320.

Hedeker, D. and Gibbons, R. B. (1994). A random-effects ordinal regression model for multilevel analysis. *Biometrics* **50**, 933–944.

Hinde, J. (1982). Compound poisson regression models. In R. Gilchrist (Ed.), *GLIM 1982 Internat. Conf. Generalized Linear Models*, New York, pp. 109–121. Springer-Verlag.

Hinde, J. P. and Wood, A. T. A. (1987). Binomial variance component models with a non-parametric assumption concerning random effects. In R. Crouchley (Ed.), *Longitudinal Data Analysis*. Avebebury, Aldershot, Hants.

Jansen, J. (1990). On the statistical analysis of ordinal data when extravariation is present. *Applied Statistics* **39**, 74–85.

Kauermann, G. and Tutz, G. (2000). Local likelihood estimation in varying coefficient models including additive bias correction. *Journal of Nonparametric Statistics* **12**, 343–371.

Läärä, E. and Matthews, J. N. (1985). The equivalence of two models for ordinal data. *Biometrika* **72**, 206–207.

Lin, X. and Breslow, N. E. (1996). Bias correction in generalized linear mixed models with multiple components of dispersion. *Journal of the American Statistical Association* **91**, 1007–1016.

Lin, X. and Zhang, D. (1999). Inference in generalized additive mixed models by using smoothing splines. *Journal of the Royal Statistical Society* **B61**,

381–400.

Liu, Q. and Pierce, D. A. (1994). A note on Gauss-Hermite quadrature. *Biometrika* **81**, 624–629.

Longford, N. T. (1994). Random coefficient models. In G. Arminger, C. Glogg, & M. Sobel (Eds.), *Handbook of Statistical Modelling for the Behavioural Sciences*. New York: Plenum.

McCullagh, P. (1980). Regression model for ordinal data (with discussion). *Journal of the Royal Statistical Society* **B 42**, 109–127.

McCulloch, C. E. (1994). Maximum likelihood variance components estimation for binary data. *J. Am. Statist. Assoc.* **89**, 330–335.

McCulloch, C. E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association* **92**, 162–170.

O'Sullivan, F. (1986). A statistical perspective on ill-posed inverse problems (with discussion). *Statistical Science* **1**, 505–527.

O'Sullivan, F. (1988). Fast computation of fully automated log-density and log hazard estimators. *Journal of Scientific and Statistical Computation* **42**, 363–379.

Pinheiro, J. C. and Bates, D. M. (1995). Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of Computational and Graphical Statistics* **4**, 12–35.

Rice, J. A. and Silverman, B. W. (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *Journal of the Royal Statistical Society* **B 53**, 233–243.

Ruppert, D. (2000). Selecting the number of knots for penalized splines. Preprint.

Ruppert, D. and Carroll, R. J. (1999). Spatially-adaptive penalties for spline fitting. *Australian Journal of Statistics* **42**, 205–223.

Schimek, M. (2000). *Smoothing and Regression. Approaches, Computation and Application*. New York: Wiley.

Simonoff, J. S. (1996). *Smoothing Methods in Statistics*. New York: Springer-Verlag.

Stone, C., Hansen, M., Kooperberg, C., and Truong, Y. (1997). Polynomial splines and their tensor products in extended linear modeling. *The Annals of Statistics* **25**, 1371–1470.

Tutz, G. (1991). Sequential models in ordinal regression. *Computational Statistics & Data Analysis* **11**, 275–295.

Tutz, G. (1999). Varying coefficients in generalized linear random effects models: A local likelihood approach. Discussion Paper 171, SFB 386, Universität München.

Tutz, G. (2000). *Die Analyse kategorialer Daten – eine anwendungsorientierte Einführung in Logit-Modellierung und kategoriale Regression*. München: Oldenbourg Verlag.

Tutz, G. and Hennevogl, W. (1996). Random effects in ordinal regression models. *Computational Statistics and Data Analysis* **22**, 537–557.

Tutz, G. and Scholz, T. (2000). Semiparametric modelling of multicategorial data. Discussion Paper Nr. 209 SFB 386, LMU München.

Whittaker, E. T. (1923). On a new method of graduation. *Proc. Edinborough Math. Assoc.* **78**, 81–89.

Wild, C. J. and Yee, T. W. (1996). Additive extensions to generalized estimating equation methods. *Journal of the Royal Statistical Society* **B58**, 711–725.

Wolfinger, R. and O'Connell, M. (1993). Generalized linear mixed models; a pseudolikelihood approach. *Journal Statist. Comput. Simulation* **48**, 233–243.

Zeger, S. L. and Diggle, P. J. (1994). Semi-parametric models for longitudinal data with application to CD4 cell numbers in HIV seroconverters. *Biometrics* **50**, 689–699.

Zeger, S. L. and Karim, M. R. (1991). Generalized linear models with random effects; a Gibbs' sampling approach. *Journal of the American Statistical Association* **86**, 79–95.

Zhang, D., Lin, X., Raz, J., and Sowers, M. (1998). Semi-parametric stochastic mixed models for longitudinal data. *Journal of the American Statistical Association* **93**, 710–719.