Chilean Journal of Statistics
Vol. 6, No. 1, April 2015, 15–36

Estimation
Research Paper

Generalized linear longitudinal mixed models with linear covariance structure and multiplicative random effects

René Holst[1,*] and Bent Jørgensen[2]

[1] Institute of Regional Health Research, University of Southern Denmark, Winsløwparken 19, DK-5000 Odense C, Denmark,
[1] Department of Mathematics and Computer Science, University of Southern Denmark, Campusvej 55, DK-5230 Odense M, Denmark

(Received: 04 July 2014 · Accepted in final form: 29 September 2014)

Abstract

We propose a versatile class of multiplicative generalized linear longitudinal mixed models (GLLMM) with additive dispersion components, based on explicit modelling of the covariance structure. The class incorporates a longitudinal structure into the random effects models and retains a marginal as well as a conditional interpretation. The estimation procedure is based on a computationally efficient quasi-score method for the regression parameters combined with a REML-like bias-corrected Pearson estimating function for the dispersion and correlation parameters. This avoids the multidimensional integral of the conventional GLMM likelihood and allows an extension of the robust empirical sandwich estimator for use with both association and regression parameters. The method is applied to a set of otholit data, used for age determination of fish.

Keywords: Bias correction · Components of dispersion · Generalized estimating equation · Pearson estimating function · REML · Multiplicative mixed models · Tweedie distribution

Mathematics Subject Classification: Primary 62H12 · Secondary 62H10.

1. Introduction

In the conventional approach to analysis of non-normal correlated data, a distinction is made between generalized estimating equations (GEE) for longitudinal data (Liang and Zeger, 1986) and generalized linear mixed models (GLMM) with random effects (Schall, 1991; Breslow and Clayton, 1993; Wolfinger and O'Connell, 1993). With certain exceptions (e.g. McCulloch and Searle, 2001; Diggle et al., 2002; Fitzmaurice et al., 2004; Pinheiro and Bates, 2000), the literature is similarly divided into two separate strands. For example Ziegler et al. (1998) and Hall (2001) summarized the first decade of developments for GEE, whereas more recent contributions include for example Hardin and Hilbe (2003),

*Corresponding author. Email: rholst@health.sdu.dk

ISSN: 0718-7912 (print)/ISSN: 0718-7920 (online)

http://www.soche.cl/chjs

Wang and Carey (2004), Coull et al. (2006) and Wang and Hanfelt (2007). For GLMM, we refer to Jiang et al. (2007), Molenberghs et al. (2010) and references therein, along with monographs such as Verbeke and Molenberghs (2000) and Lee et al. (2006).

The GEE and GLMM models differ conceptually and computationally, as reflected in the conventional distinction between marginal and conditional models. In practice, however, one is often faced with a combination of longitudinal and random effects, where neither type of model, on its own, is adequate. There is hence a need for a versatile class of generalized linear longitudinal mixed models (GLLMM), preserving the computational efficiency of GEE, while avoiding the problems associated with the multidimensional integral defining the likelihood in conventional GLMM approaches such as Schall (1991), Breslow and Clayton (1993) and Wolfinger and O'Connell (1993).

In the present paper, we introduce a versatile and computationally efficient class of GLLMM models with a multiplicative random effects structure and a corresponding additive decomposition of the variance into dispersion components, thereby retaining much of the simplicity of classical linear mixed models, while retaining a marginal as well as a conditional interpretation. A serial correlation structure is employed within clusters, but unlike the state space models of Jørgensen et al. (1996, 1999) and Jørgensen and Song (2006), where the latent process is non-stationary, our approach is based on a stationary latent process defined by means of a linear filter, see also Jowaheer and Sutradhar (2002), Jørgensen and Song (2007) and Hall et al. (2008).

The models are defined hierarchically via conditional Tweedie distributions, much like the multiplicative mixed effects models of Ma and Jørgensen (2007) and Ma et al. (2009), thereby facilitating moment calculations by avoiding the use of additive random terms in the linear predictor of the model. The Tweedie family provides a flexible class of models for both positive continuous data, count data and positive continuous data with a point mass at zero (Jørgensen, 1997, Ch. 4). This is useful for simulation purposes, although we shall mainly use the Tweedie models as a vehicle for obtaining explicit moment calculations.

The estimation procedure for the models is based on a quasi-score method for the regression parameters, combined with bias-corrected Pearson estimating functions, which yield REML-style estimators for the association (dispersion and correlation) parameters, following earlier work by Jørgensen and Knudsen (2004), McCullagh and Tibshirani (1990) and Hall (2001). The method may be considered a so-called GEE2 variant (Liang et al., 1992; Balemi and Lee, 2009), while allowing the use of a working correlation structure and extending the robust empirical sandwich estimator for calculating the asymptotic variance for both regression and association parameters. The estimating equations are solved by an efficient Newton scoring method.

We introduce the new GLLMM model and derive its covariance structure in Section 2. This is followed by a discussion of parameter estimation in Section 3, including bias correction for association parameter estimates, and the empirical sandwich estimator. In Section 4 we present a simulation study, where we investigate aspects of robustness, and in particular the effects of misspecification of the distribution at observation level. of the method and consider the effect of the bias correction. In Section 5 we apply the method to a set of otholit data, which carry information about the age and growth patterns of fish. This is followed by a discussion of our results in Section 6.

2. Model specification

2.1 Definition

We now introduce our new class of GLLMM models, followed by a discussion of their covariance structure in Section 2.2, which is crucial for the interpretation and estimation of

the models (Section 3). The models are based on the class of Tweedie exponential dispersion models (Jørgensen, 1997, Ch. 4). A Tweedie variable $Y \sim \mathrm{Tw}_r\left(\mu, \sigma^2\right)$ with mean μ, dispersion parameter σ^2 and power parameter $r \in (-\infty, 0] \cup [1, \infty)$ is characterized among all exponential dispersion models by the power form of its variance function, $\mathrm{var}(Y) = \sigma^2 \mu^r$ (Jørgensen, 1997, Ch. 3). Special cases include the normal ($r = 0$), Poisson ($r = 1$ and $\sigma^2 = 1$) and gamma ($r = 2$) families. The case $1 < r < 2$ corresponds to compound Poisson distributions, which are non-negative and continuous, except for a positive probability at zero. The case $r < 0$ corresponds to skew distributions with support $\mathbb{R}$.

The models are built in a three-level hierarchy, with a base level accounting for the cluster random effects. The middle layer implements the longitudinal structure by means of a latent process, constructed as a linear filter of independent Tweedie variables. The response variables constitutes the top level of the model. For ease of presentation we consider a balanced design with T equidistant observation times common to a set of I independent clusters. The models are readily adapted to ragged structures.

Base level

At base level we use independent Tweedie distributed multiplicative random effects with mean 1 and dispersion parameter σ^2,

$$Z_i \sim \mathrm{Tw}_{r_1}\left(1,\, \sigma^2\right). \tag{2.1}$$

Here we require that $r_1 \geq 2$ in order to make Z_i strictly positive, ruling out Tweedie distributions with $1 \leq r_1 < 2$, which allow zero values. The restriction $r_1 \geq 2$ is hence crucial for simulation studies based on the Tweedie distribution, but not for the estimation, as we shall see below. The case $r_1 = 0$, corresponding to normal cluster effects, is applicable for normal response variables and is briefly outlined below. In view of the multiplicative form of the model (cf. (2.3)), the logarithmic values $\log Z_i$ play a role similar to that of the additive mean 0 normal random effects in conventional mixed effects models.

Middle level

The middle layer accommodates serial correlation between observations within a cluster by means of a weakly stationary latent process based on Tweedie noise,

$$Z_{it} \mid Z_i = z_i \sim z_i \mathrm{Tw}_{r_2}\left(\alpha_+^{-1}\,,\, \alpha_+^{r_2} \omega^2 z_i^{-1}\right) \tag{2.2}$$

$$= \mathrm{Tw}_{r_2}\left(\alpha_+^{-1} z_i\,,\, \alpha_+^{r_2} \omega^2 z_i^{1-r_2}\right), \tag{2.3}$$

where $\omega_2 > 0$, the Z_{it} are assumed conditionally independent given the cluster variables Z_i, and again $r_2 \geq 2$. Here $\alpha_+ = \sum_{k=0}^{\infty} \alpha_k < \infty$ with $\alpha_0 = 1$ and $\alpha_k \in [0, 1)$ for $k > 0$. The form (2.2) confirms the multiplicative nature of the model, whereas (2.3) follows from the scale transformation property of Tweedie distributions(Jørgensen, 1997, Ch. 4). The particular powers used in (2.2) serve to make the variance (2.7) below scale linearly in the dispersion parameters.

The longitudinal structure of the model is obtained via a conditionally weakly stationary latent process Q_{it}, defined by the linear filter

$$Q_{it} = \sum_{s=0}^{\infty} \alpha_s Z_{i\,t-s}\,. \tag{2.4}$$

By way of construction the latent process Q_{it} has mean 1.

Top level

At the observation level we assume that the observed variables Y_{it} are conditionally independent given $\boldsymbol{Q}_{i\bullet}$, with conditional Tweedie distribution

$$Y_{it} \mid \boldsymbol{Q}_{i\bullet} = \boldsymbol{q}_{i\bullet} \sim \mathrm{Tw}_{r_3}\left(\mu_{it} q_{it},\, \rho^2 q_{it}^{1-r_3}\right), \tag{2.5}$$

where $\mu_{it} > 0$ and $r_3 \geq 1$. Here the $\bullet$ notation denotes the vector obtained by letting the corresponding index run, so that $\boldsymbol{Y}_{i\bullet} = (Y_{i1}, \ldots, Y_{iT})^\top$ and so on. By choosing different values of r_3 from 1 and up, we can accommodate a range of different response distributions, as explained above.

By definition of the Tweedie variance function, the conditional mean and variance of (2.5) are linear, so that $E\left(Y_{it} \mid \boldsymbol{Q}_{i\bullet}\right) = q_{it}\mu_{it}$ and $\mathrm{var}\left(Y_{it} \mid \boldsymbol{Q}_{i\bullet}\right) = q_{it}\rho^2\mu_{it}^{r_3}$, which in turn implies that This property is essential for the derivation of the covariance structure and the estimating functions below. In the Poisson case ($r_3 = 1$) it is convenient to let the dispersion parameter $\rho^2 > 0$ accommodate potential over- or under-dispersion.

The marginal means μ_{it} may depend on covariates $\mu_{it} = \mu_{it}\left(\boldsymbol{x}_{it}; \boldsymbol{\beta}\right)$, where $\boldsymbol{\beta}$ denotes a vector of regression parameters. For non-negative data with positive means, the log link is a suitable choice, providing a natural interpretation of the regression parameters. Furthermore the log link, along with the multiplicative structure, imply an additive structure for the linear predictor: linear predictor:

$$\eta_{it} = \log\left(\mu_{it} q_{it}\right) = \boldsymbol{x}_{it}^\top \boldsymbol{\beta} + \log q_{it}. \tag{2.6}$$

allowing easy comparison with conventional generalized linear mixed models. In the small-dispersion limit (σ^2 and ω^2 small), the term $\log Q_{it}$ is asymptotically normal with mean zero (Jørgensen, 1997, Ch. 4), which highlights the parallels on the log scale between our model and conventional generalized linear mixed models. There is, however, no problem in using other types of link functions.

A variant of the model, applicable to normal response variables, assumes normal zero mean random cluster effects, $Z_i \sim N(0, \sigma^2)$, replacing (2.1). The noise process (2.3) is then assumed to be Gaussian, $Z_{it} \mid Z_i = z_i \sim N(z_i, \omega^2)$, while maintaining the filter (2.4) as above. At the response level (2.5) is replaced by an additive structure with identity link function, $Y_{it} \mid \boldsymbol{Q}_{i\bullet} = \boldsymbol{q}_{i\bullet} \sim N(\mu_{it} + q_{it}, \rho^2)$. Since the conditional means are again linear, the corresponding covariance structure is easily derived.

2.2 Covariance structure

An important advantage of the model is that the marginal covariance matrix of the observation vector $\boldsymbol{Y}_{\bullet\bullet}$ is available on closed form, thereby facilitating the estimating function approach to estimation and inference (cf. Section 3). Details of the derivation may be found in Appendix A.

The covariance between two given observations within the ith cluster is:

$$\mathrm{cov}\left(Y_{it},\, Y_{it'}\right) = \sigma^2\mu_{it}\mu_{it'} + \omega^2\mu_{it}\mu_{it'}\sum_{s=0}^{\infty}\alpha_s\alpha_{s+|t-t'|} + \delta_t^{t'}\rho^2\mu_{it}^{r_3},$$

where $\delta_i^{i'}$ is the Kronecker delta, being 1 for $i = i'$ and zero otherwise. This covariance does not depend on r_1 and r_2. This lends a certain degree of robustness to the estimation method (cf. Section 3), due to the fact that the estimation relies on second-moment assumptions only.

We now derive a matrix expression for the covariance matrix var $(\boldsymbol{Y}_{\bullet\bullet})$. First we consider the latent process correlation matrix, $\boldsymbol{K}(\boldsymbol{\alpha})$, with tt'th entry $\sum_{s=0}^{\infty}\alpha_s\alpha_{s+|t-t'|}$. Next let $\mathbf{1}_T$ denote the T-vector of 1s. In matrix notation, the variance-covariance matrix of the response vector for the ith cluster may then be expressed as

$$\begin{aligned}\operatorname{var}(\boldsymbol{Y}_{i\bullet}) &= \boldsymbol{\mu}_{i\bullet}\boldsymbol{\mu}_{i\bullet}^{\top} \odot \left\{\sigma^2\ \mathbf{1}_T\mathbf{1}_T^{\top} + \omega^2\,\boldsymbol{K}(\boldsymbol{\alpha})\right\} + \rho^2\ \operatorname{diag}(\boldsymbol{\mu}_{i\bullet}^{r_3}) \\ &= \sigma^2\,\boldsymbol{\mu}_{i\bullet}\boldsymbol{\mu}_{i\bullet}^{\top} + \omega^2 \operatorname{diag}(\boldsymbol{\mu}_{i\bullet})\,\boldsymbol{K}(\boldsymbol{\alpha})\ \operatorname{diag}(\boldsymbol{\mu}_{i\bullet}) + \rho^2\ \operatorname{diag}(\boldsymbol{\mu}_{i\bullet}^{r_3}), \qquad (2.7)\end{aligned}$$

say, where $\odot$ denotes the Hadamard (elementwise) product (Magnus and Neudecker, 1999, p. 45). When working under second-moment assumptions, the restrictions $r_1, r_2 \geq 2$ from above hence do not come into play.

Similar to conventional linear mixed models, we find that the covariance (2.7) has been decomposed into components of dispersion corresponding to the different sources of variation. The three covariance terms of (2.7) correspond to variation between clusters, covariances within cluster and observation error, and these three matrices are of dense, block-diagonal (one block for each i) and diagonal form, respectively. The multiplicative construction of the model hence translates into an additive covariance structure.

The models accommodated by our approach hence extend the range of possible serial correlation patterns achievable compared with the conventional GEE correlation structures usually considered. Particular covariance structures may be obtained by imposing restrictions on the linear filter parameter vector $\boldsymbol{\alpha}$ or the dispersion parameters. Table 1 lists some of the more common covariance structures and the corresponding parameter restrictions. Appendix B details $\boldsymbol{K}(\boldsymbol{\alpha})$ and its derivatives for MA(p)-type and AR(1)-type processes.

Table 1. Some standard covariance structures. The MA(p)-type and AR(p)-type refer to the latent process correlation structure conditionally on the cluster random effects. The GLMM type refers to a random intercept model.

Covariance structure	Parameter restrictions
Independent	$\omega^2 = \sigma^2 = 0$.
Exchangeable	$\omega^2 = \rho^2 = 0$ and $\alpha_s = 0$ for $s > 0$.
MA(p)-type	$\alpha_s = 0$ for $s > p$.
AR(p)-type	For $p = 1$ $\alpha_s = \alpha^s$. For $p > 1$ the α_s are given by the Yule-Walker equations.
GLMM	$\omega^2 = 0$ and $\alpha_s = 0$ for $s > 0$.

3. Estimation

3.1 General issues

The set of parameters $\boldsymbol{\theta}$ to be estimated is naturally partitioned into regression and association parameters, $\boldsymbol{\theta} = \left(\boldsymbol{\beta}^{\top}, \boldsymbol{\gamma}^{\top}\right)^{\top}$, where the regression parameters $\boldsymbol{\beta}$ usually are those of interest whereas the association parameters $\boldsymbol{\gamma}$, containing dispersion and correlation parameters, are often considered nuisance parameters. For estimation of the parameters we use a set of corresponding estimating functions denoted $\boldsymbol{\psi} = \left(\boldsymbol{\psi}_{\boldsymbol{\beta}}^{\top}, \boldsymbol{\psi}_{\boldsymbol{\gamma}}^{\top}\right)^{\top}$. These are explained in more detail in Section 3.4 and 3.5. A bias correction for $\boldsymbol{\psi}_{\boldsymbol{\gamma}}$ is given in Section 3.6.

The regression parameters are estimated by means of the quasi-score function defined

by

$$\boldsymbol{\psi}_{\boldsymbol{\beta}} = \sum_{i=1}^{I} \boldsymbol{D}_i^{\top} \boldsymbol{C}_i^{-1} \left\{ \boldsymbol{Y}_i - E\left(\boldsymbol{Y}_i\right) \right\}, \tag{3.8}$$

where $\boldsymbol{D}_i = \nabla_{\boldsymbol{\beta}} E\left(\boldsymbol{Y}_i\right) = \partial E\left(\boldsymbol{Y}_i\right) / \partial \boldsymbol{\beta}^{\top}$ and $\boldsymbol{C}_i = \operatorname{var}\left(\boldsymbol{Y}_i\right)$. Although (3.8) is similar to the well known estimating function for the regression parameters from the conventional GEE framework (Liang and Zeger, 1986), it corresponds to using the model covariance matrix (2.7) rather than the working covariance matrix. The conventional GEE working covariance matrix is built around the working correlation matrix $\boldsymbol{R}(\boldsymbol{\alpha})$ so that

$$\operatorname{var}\left(\boldsymbol{Y}_i\right) = \phi \boldsymbol{A}_i^{1/2}\left(\boldsymbol{\mu}_i\right) \boldsymbol{R}\left(\boldsymbol{\alpha}\right) \boldsymbol{A}_i^{1/2}\left(\boldsymbol{\mu}_i\right), \tag{3.9}$$

where ϕ is a dispersion parameter, $\boldsymbol{A}_i\left(\boldsymbol{\mu}_i\right) = \operatorname{diag}\left\{v\left(\boldsymbol{\mu}_{i\bullet}\right)\right\}$ and $v\left(\cdot\right)$ is the variance function. In contrast, we emphasize the decomposition of the variance into components of dispersion and associate the process correlation matrix $\boldsymbol{K}(\boldsymbol{\alpha})$ with an appropriate level in the hierarchy.

We use Pearson estimating functions for the association parameters $\boldsymbol{\gamma} = \left(\gamma_1, \ldots, \gamma_N\right)^{\top}$, which may optionally include r_3. The entire vector of functions is denoted $\boldsymbol{\psi}_{\boldsymbol{\gamma}} = \left(\psi_{\gamma_1}, \ldots, \psi_{\gamma_N}\right)^{\top}$, where $N = 3 + M$ or $N = 4 + M$ for r_3 being considered known or an element of $\boldsymbol{\gamma}$ respectively and $M = \dim\left(\boldsymbol{\alpha}\right)$. The nth component is given by

$$\psi_{\gamma_n}\left(\boldsymbol{\beta}, \boldsymbol{\gamma}\right) = \sum_{i=1}^{I} \operatorname{tr}\left\{ \boldsymbol{W}_{in} \left(\boldsymbol{r}_i \boldsymbol{r}_i^{\top} - \boldsymbol{C}_i \right) \right\},$$

where $\boldsymbol{r}_i = \boldsymbol{Y}_i - E\left(\boldsymbol{Y}_i\right)$ and $\boldsymbol{W}_{in}$ are suitable weights. This form emphasizes the model covariance matrix in contrast to the more conventional expressions of the Pearson estimating function (Jørgensen and Knudsen, 2004). As discussed by Jørgensen and Knudsen (2004), the use of Pearson estimating functions for the association parameters constitutes an extended form of residual maximum likelihood estimation. We also note at this point that since estimation is based on second-moment assumptions, and since the covariance structure (2.7) does not depend on the power parameters r_1 and r_2 from equations (2.1) and (2.1), these two parameters do not enter the estimation procedure.

3.2 Sensitivity

Cox and Reid (1987) studied parameter orthogonality in the likelihood framework corresponding to block diagonality of the Fisher information matrix. Jørgensen and Knudsen (2004) studied the corresponding property of nuisance parameter insensitivity in an estimating equation context, by means of the sensitivity matrix, defined by $\boldsymbol{S}_{\boldsymbol{\theta}} = E\left\{\nabla_{\boldsymbol{\theta}} \boldsymbol{\psi}(\boldsymbol{\theta})\right\}$. The sensitivity matrix may be partitioned into blocks corresponding to $\left(\boldsymbol{\psi}_{\boldsymbol{\beta}}^{\top}, \boldsymbol{\psi}_{\boldsymbol{\gamma}}^{\top}\right)^{\top}$ and $\left(\boldsymbol{\beta}^{\top}, \boldsymbol{\gamma}^{\top}\right)$ as follows:

$$\boldsymbol{S}_{\boldsymbol{\theta}} = \begin{bmatrix} \boldsymbol{S}_{\boldsymbol{\beta}}\left(\boldsymbol{\theta}\right) & \boldsymbol{S}_{\boldsymbol{\beta}\boldsymbol{\gamma}}\left(\boldsymbol{\theta}\right) \\ \boldsymbol{S}_{\boldsymbol{\gamma}\boldsymbol{\beta}}\left(\boldsymbol{\theta}\right) & \boldsymbol{S}_{\boldsymbol{\gamma}}\left(\boldsymbol{\theta}\right) \end{bmatrix} = \begin{bmatrix} E\left\{\nabla_{\boldsymbol{\beta}} \boldsymbol{\psi}_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \boldsymbol{\gamma})\right\} & E\left\{\nabla_{\boldsymbol{\gamma}} \boldsymbol{\psi}_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \boldsymbol{\gamma})\right\} \\ E\left\{\nabla_{\boldsymbol{\beta}} \boldsymbol{\psi}_{\boldsymbol{\gamma}}(\boldsymbol{\beta}, \boldsymbol{\gamma})\right\} & E\left\{\nabla_{\boldsymbol{\gamma}} \boldsymbol{\psi}_{\boldsymbol{\gamma}}(\boldsymbol{\beta}, \boldsymbol{\gamma})\right\} \end{bmatrix}.$$

Nuisance parameter insensitivity (for short denoted $\boldsymbol{\gamma}$-insensitivity) is defined by the upper right-hand block $\boldsymbol{S}_{\boldsymbol{\beta}\boldsymbol{\gamma}}\left(\boldsymbol{\theta}\right)$ being zero. First of all this implies efficiency stable estimation of $\boldsymbol{\beta}$, meaning that the estimation of $\boldsymbol{\gamma}$ does not affect the asymptotic variance of $\widehat{\boldsymbol{\beta}}$; see

Section 3.7. Second, it simplifies the Newton scoring algorithm (Jørgensen and Knudsen, 2004) as detailed below. Third, it implies that $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}$ varies only slowly with $\boldsymbol{\gamma}$, where $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}$ is the estimate of $\boldsymbol{\beta}$ for $\boldsymbol{\gamma}$ known. While nuisance parameter insensitivity does not ensure asymptotic independence of $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\gamma}}$, it does ease the computation of the asymptotic variance of $\widehat{\boldsymbol{\beta}}$.

Following Jørgensen and Knudsen (2004) it is easily seen that $\boldsymbol{\psi}_{\boldsymbol{\beta}}$ is $\boldsymbol{\gamma}$-insensitive. In fact, from (3.8) we see that $\boldsymbol{\psi}_{\boldsymbol{\beta}}$ depends on $\boldsymbol{\gamma}$ only via $\boldsymbol{C}_i^{-1}$ and hence $\nabla_{\boldsymbol{\gamma}}\boldsymbol{\psi}_{\boldsymbol{\beta}}(\boldsymbol{\beta},\boldsymbol{\gamma})$ has zero mean, i.e. $\boldsymbol{S}_{\boldsymbol{\beta\gamma}}(\boldsymbol{\theta}) = \mathbf{0}$.

In the rest of the paper we write $\boldsymbol{S}_{\boldsymbol{\beta}}$ for $\boldsymbol{S}_{\boldsymbol{\beta}}(\boldsymbol{\theta})$ etc, whenever the meaning is unambiguous. The remaining blocks of $\boldsymbol{S}_{\boldsymbol{\theta}}$ are detailed along with the estimating functions in Sections 3.4 and 3.5

3.3 Algorithm

Calculation of the parameter estimates is achieved by means of the Newton scoring algorithm (Jørgensen et al., 1996) in which we update the previous value of $\boldsymbol{\theta}$ by

$$\boldsymbol{\theta}^* = \boldsymbol{\theta} - \boldsymbol{S}_{\boldsymbol{\theta}}^{-1}\boldsymbol{\psi}(\boldsymbol{\theta}).$$

By the regularity of $\boldsymbol{\psi}$ along with the $\boldsymbol{\gamma}$-insensitivity of $\boldsymbol{\psi}_{\boldsymbol{\beta}}$, it follows by simple matrix manipulations that we may express the inverse of $\boldsymbol{S}_{\boldsymbol{\theta}}$ in blocks as follows:

$$\boldsymbol{S}_{\boldsymbol{\theta}}^{-1} = \begin{bmatrix} \boldsymbol{S}_{\boldsymbol{\beta}}^{-1} & \mathbf{0} \\ -\boldsymbol{S}_{\boldsymbol{\gamma}}^{-1}\boldsymbol{S}_{\boldsymbol{\gamma\beta}}\boldsymbol{S}_{\boldsymbol{\beta}}^{-1} & \boldsymbol{S}_{\boldsymbol{\gamma}}^{-1} \end{bmatrix}. \tag{3.10}$$

The algorithm therefore splits into a $\boldsymbol{\beta}$ step

$$\boldsymbol{\beta}^* = \boldsymbol{\beta} - \boldsymbol{S}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\psi}_{\boldsymbol{\beta}}(\boldsymbol{\theta}) \tag{3.11}$$

and a $\boldsymbol{\gamma}$ step

$$\begin{aligned} \boldsymbol{\gamma}^* &= \boldsymbol{\gamma} + \boldsymbol{S}_{\boldsymbol{\gamma}}^{-1}\boldsymbol{S}_{\boldsymbol{\gamma\beta}}\boldsymbol{S}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\psi}_{\boldsymbol{\beta}}(\boldsymbol{\theta}) - \boldsymbol{S}_{\boldsymbol{\gamma}}^{-1}\boldsymbol{\psi}_{\boldsymbol{\gamma}}(\boldsymbol{\theta}) \\ &= \boldsymbol{\gamma} - \boldsymbol{S}_{\boldsymbol{\gamma}}^{-1}\left\{\boldsymbol{\psi}_{\boldsymbol{\gamma}}(\boldsymbol{\theta}) - \boldsymbol{S}_{\boldsymbol{\gamma\beta}}\boldsymbol{S}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\psi}_{\boldsymbol{\beta}}(\boldsymbol{\theta})\right\}. \end{aligned} \tag{3.12}$$

Following Jørgensen and Knudsen (2004) we insert $\boldsymbol{\beta}^*$ from (3.11) into equation (3.12). Since equation (3.11) can be rewritten as $-\boldsymbol{S}_{\boldsymbol{\beta}}^{-1}(\boldsymbol{\theta})\,\boldsymbol{\psi}_{\boldsymbol{\beta}}(\boldsymbol{\theta}) = \boldsymbol{\beta}^* - \boldsymbol{\beta}$, this makes $\boldsymbol{S}_{\boldsymbol{\beta}}^{-1}(\boldsymbol{\theta}^*)\,\boldsymbol{\psi}_{\boldsymbol{\beta}}(\boldsymbol{\theta}^*) = \mathbf{0}$, where $\boldsymbol{\theta}^*$ indicates $\boldsymbol{\beta}^*$ is being used. Consequently the modified $\boldsymbol{\gamma}$ step becomes

$$\boldsymbol{\gamma}^* = \boldsymbol{\gamma} - \boldsymbol{S}_{\boldsymbol{\gamma}}^{-1}\boldsymbol{\psi}_{\boldsymbol{\gamma}}(\boldsymbol{\theta}^*). \tag{3.13}$$

Analogously we use the most recent estimate of $\boldsymbol{\gamma}$ when updating $\boldsymbol{\beta}$ in (3.11). This is however of less importance, due to the slow variation of $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}$ with $\boldsymbol{\gamma}$. Jørgensen and Knudsen (2004) coined the scheme of alternating between (3.11) and (3.13) the *chaser* algorithm, with reference to the asymmetric interdependence between $\boldsymbol{\beta}^*$ and $\boldsymbol{\gamma}^*$.

3.4 Regression parameters $\boldsymbol{\beta}$

Following Ma (1999), Ma et al. (2003) and Ma and Jørgensen (2007) we use the best linear unbiased predictor for predicting the random effects. The best linear unbiased predictor of a random variable $\boldsymbol{Q}$ given the observed data $\boldsymbol{Y}$ is defined by (Henderson, 1975; Ma, 1999)

$$\widehat{\boldsymbol{Q}} = E\left(\boldsymbol{Q}\right) + \operatorname{cov}\left(\boldsymbol{Q},\, \boldsymbol{Y}\right) \operatorname{var}\left(\boldsymbol{Y}\right)^{-1} \left\{\boldsymbol{Y} - E\left(\boldsymbol{Y}\right)\right\}. \tag{3.14}$$

The model specification using Tweedie distributions allows for derivation of the joint score function $\boldsymbol{u}\left(\boldsymbol{\theta}; \boldsymbol{Y}, \boldsymbol{Q}\right)$ (Ma and Jørgensen, 2007). We define unbiased estimating functions $\boldsymbol{\psi}_{\boldsymbol{\beta}}$ by substituting the random effects by their respective best linear unbiased predictors, i.e.

$$\boldsymbol{\psi}_{\boldsymbol{\beta}}\left(\boldsymbol{\theta}; \boldsymbol{Y}\right) = \boldsymbol{u}\left(\boldsymbol{\theta}; \boldsymbol{Y}, \widehat{\boldsymbol{Q}}\right).$$

Since $\boldsymbol{u}$ is on the form $\boldsymbol{AQ} + \boldsymbol{BY}$, for suitable matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ (Ma and Jørgensen, 2007), it follows from (3.14) and the linearity of $E\left(\cdot\right)$ and $\operatorname{cov}\left(\cdot,\, \boldsymbol{Y}\right)$ that the best linear unbiased predictor of $\boldsymbol{u}\left(\boldsymbol{\theta}; \boldsymbol{Y}, \boldsymbol{Q}\right)$ is $\boldsymbol{A}\widehat{\boldsymbol{Q}} + B\boldsymbol{Y}$. And since $\boldsymbol{u}$ is linear in both the observed and the latent variables, $\boldsymbol{Y}$ and $\boldsymbol{Q}$, we find that $\boldsymbol{\psi}_{\boldsymbol{\beta}}$ is the best linear unbiased predictor of the score function $\boldsymbol{u}$ given the data. By (3.14) we therefore arrive at the conventional GEE from expression (3.8)

$$\boldsymbol{\psi}_{\boldsymbol{\beta}} = E\left(\boldsymbol{u}\right) + \operatorname{cov}\left(\boldsymbol{u},\, \boldsymbol{Y}\right) \boldsymbol{C}^{-1} \left\{\boldsymbol{Y} - E\left(\boldsymbol{Y}\right)\right\} = \sum_{i=1}^{I} \boldsymbol{D}_i^{\top} \boldsymbol{C}_i^{-1} \left\{\boldsymbol{Y}_i - E\left(\boldsymbol{Y}_i\right)\right\}. \tag{3.15}$$

Here we have used the independence between clusters, along with the following Bartlett-type identity

$$\boldsymbol{D} = \nabla_{\boldsymbol{\beta}} E\left(\boldsymbol{Y}\right) = E\left(\boldsymbol{Y} \cdot \boldsymbol{u}\right) = \operatorname{cov}\left(\boldsymbol{Y},\, \boldsymbol{u}\right).$$

From (3.8) we furthermore obtain the sensitivity $\boldsymbol{S}_{\boldsymbol{\beta}} = E\left(\nabla_{\boldsymbol{\beta}} \boldsymbol{\psi}_{\boldsymbol{\beta}}\right)$ and variability $\boldsymbol{V}_{\boldsymbol{\beta}} = \operatorname{var}\left(\boldsymbol{\psi}_{\boldsymbol{\beta}}\right)$ as

$$\boldsymbol{S}_{\boldsymbol{\beta}} = -\boldsymbol{D}^{\top} \boldsymbol{C}^{-1} \boldsymbol{D}, \quad \boldsymbol{V}_{\boldsymbol{\beta}} = \boldsymbol{D}^{\top} \boldsymbol{C}^{-1} \boldsymbol{D}. \tag{3.16}$$

The identity $\boldsymbol{V}_{\boldsymbol{\beta}} = -\boldsymbol{S}_{\boldsymbol{\beta}}$ is characteristic for quasi-score functions. We therefore conclude that $\boldsymbol{\psi}_{\boldsymbol{\beta}}$ is optimal within the class of linear estimating functions in the sense of Crowder (1987). This also follows from (3.15) as the best linear unbiased predictor is optimal among all linear predictors.

3.5 Association parameters $\boldsymbol{\gamma}$

Our approach is akin to that of Ma and Jørgensen (2007), but deviates by allowing for correlation structures within clusters. Furthermore Ma and Jørgensen used a closed form ad-hoc estimator for the association parameters Our estimation of $\boldsymbol{\gamma}$ is based on Pearson estimating functions, following the path of Hall and Severini (1998) and Jørgensen and

Knudsen (2004). For the γ_n component it is defined by

$$\begin{aligned}\psi_{\gamma_n}(\boldsymbol{\beta},\boldsymbol{\gamma}) &= \sum_{i=1}^{I}\left\{\boldsymbol{r}_i^\top \boldsymbol{W}_{in}\boldsymbol{r}_i - \operatorname{tr}\left(\boldsymbol{W}_{in}\boldsymbol{C}_i\right)\right\}\\ &= \sum_{i=1}^{I}\left\{\operatorname{tr}\left(\boldsymbol{W}_{in}\boldsymbol{r}_i\boldsymbol{r}_i^\top\right) - \operatorname{tr}\left(\boldsymbol{W}_{in}\boldsymbol{C}_i\right)\right\}\\ &= \sum_{i=1}^{I}\operatorname{tr}\left\{\boldsymbol{W}_{in}\left(\boldsymbol{r}_i\boldsymbol{r}_i^\top - \boldsymbol{C}_i\right)\right\}, \end{aligned} \tag{3.17}$$

where $\boldsymbol{r}_i = \boldsymbol{Y}_i - E(\boldsymbol{Y}_i)$ and $\boldsymbol{W}_{in}$ are suitable weights. By linearity of $E(\cdot)$ and $\operatorname{tr}(\cdot)$ these estimating functions are clearly unbiased since $E\left(\boldsymbol{r}_i\boldsymbol{r}_i^\top\right) = \boldsymbol{C}_i$.

In the conventional GEE framework, the Pearson estimating function for the association parameters relies on a working correlation matrix used for defining $\operatorname{var}(\boldsymbol{Y}_i)$ as shown in (3.9). In contrast, we emphasize the decomposition of the variance into components of dispersion and associate the process correlation matrix $\boldsymbol{K}(\boldsymbol{\alpha})$ with an appropriate level in the hierarchy.

For $\boldsymbol{W}_{in}$ we use the weights proposed by Hall and Severini (1998),

$$\boldsymbol{W}_{in} = -\frac{\partial \boldsymbol{C}_i^{-1}}{\partial \gamma_n} = \boldsymbol{C}_i^{-1}\frac{\partial \boldsymbol{C}_i}{\partial \gamma_n}\boldsymbol{C}_i^{-1}.$$

In the Gaussian case these weights lead to the score function for the association parameters, and in general the weights are hence optimal in the small-dispersion limit Jørgensen (1987), in which dispersion models are asymptotically normal.

From (3.17) we may derive the θ_m-sensitivity of ψ_{γ_n}, namely

$$\begin{aligned} E\left(\frac{\partial}{\partial\theta_m}\psi_{\gamma_n}\right) &= E\left[\frac{\partial}{\partial\theta_m}\sum_{i=1}^{I}\operatorname{tr}\left\{\boldsymbol{W}_{in}\left(\boldsymbol{r}_i\boldsymbol{r}_i^\top - \boldsymbol{C}_i\right)\right\}\right]\\ &= \sum_{i=1}^{I}\operatorname{tr}\left[\boldsymbol{W}_{in}E\left\{\frac{\partial}{\partial\theta_m}\left(\boldsymbol{r}_i\boldsymbol{r}_i^\top - \boldsymbol{C}_i\right)\right\}\right]\\ &= -\sum_{i=1}^{I}\operatorname{tr}\left(\boldsymbol{W}_{in}\frac{\partial \boldsymbol{C}_i}{\partial\theta_m}\right). \end{aligned}$$

Here we have used that the derivatives of $\boldsymbol{r}_i$ do not depend on data so $E\left\{\left(\partial\boldsymbol{r}_i/\partial\theta_m\right)\boldsymbol{r}_i^\top\right\} = E\left\{\boldsymbol{r}_i\left(\partial\boldsymbol{r}_i^\top/\partial\theta_m\right)\right\} = 0$ and $\left(\partial\boldsymbol{W}_{in}/\partial\theta_m\right)E\left(\boldsymbol{r}_i\boldsymbol{r}_i^\top - \boldsymbol{C}_i\right) = 0$.

The nmth entries of the blocks $\boldsymbol{S}_{\boldsymbol{\gamma}}$ and $\boldsymbol{S}_{\boldsymbol{\gamma\beta}}$ are

$$\left\{\boldsymbol{S}_{\boldsymbol{\gamma}}\right\}_{nm} = -\sum_{i=1}^{I}\operatorname{tr}\left(\boldsymbol{C}_i^{-1}\frac{\partial \boldsymbol{C}_i}{\partial\gamma_n}\boldsymbol{C}_i^{-1}\frac{\partial \boldsymbol{C}_i}{\partial\gamma_m}\right) \tag{3.18}$$

and similarly

$$\{\boldsymbol{S}_{\boldsymbol{\gamma\beta}}\}_{nm} = -\sum_{i=1}^{I} \operatorname{tr}\left(\boldsymbol{C}_i^{-1}\frac{\partial \boldsymbol{C}_i}{\partial \gamma_n}\boldsymbol{C}_i^{-1}\frac{\partial \boldsymbol{C}_i}{\partial \beta_m}\right) \tag{3.19}$$

respectively.

3.6 Bias correction

The estimation of nuisance parameters may be subject to bias (McCullagh and Tibshirani, 1990; Jørgensen and Knudsen, 2004), caused by not taking into account the degrees of freedom spent on estimating the regression parameters.

In the spirit of Godambe (1960), Heyde (1997) and Jørgensen and Knudsen (2004) we adjust the estimating function for bias rather than the estimate. The corrected estimating function for γ_n becomes

$$\begin{aligned}
\breve{\psi}_{\gamma_n}(\boldsymbol{\beta},\boldsymbol{\gamma}) &= \psi_{\gamma_n}(\boldsymbol{\beta},\boldsymbol{\gamma}) + b_{\gamma_n}(\boldsymbol{\beta},\boldsymbol{\gamma}) \\
&= \sum_{i=1}^{I} \operatorname{tr}\left\{\boldsymbol{W}_{in}\left(\boldsymbol{r}_i\boldsymbol{r}_i^{\top} - \boldsymbol{C}_i\right)\right\} + \operatorname{tr}\left\{\left(\sum_{i=1}^{I}\boldsymbol{D}_i^{\top}\boldsymbol{W}_{in}\boldsymbol{D}_i\right)\left(\sum_{i=1}^{I}\boldsymbol{D}_i^{\top}\boldsymbol{C}_i^{-1}\boldsymbol{D}_i\right)^{-1}\right\} \\
&= \sum_{i=1}^{I} \operatorname{tr}\left\{\boldsymbol{W}_{in}\left(\boldsymbol{r}_i\boldsymbol{r}_i^{\top} - \boldsymbol{C}_i\right)\right\} - \operatorname{tr}\left(\boldsymbol{J}_{\boldsymbol{\beta}}^{(\gamma_n)}\boldsymbol{J}_{\boldsymbol{\beta}}^{-1}\right),
\end{aligned}$$

where $\boldsymbol{J}_{\boldsymbol{\beta}}^{(\gamma_n)} = \partial \boldsymbol{J}_{\boldsymbol{\beta}}/\partial\gamma_n$. The Godambe information $\boldsymbol{J}_{\boldsymbol{\beta}}$, see Section 3.7, plays a role in the estimating equation context analogous to that of the Fisher information in the likelihood framework, with $\boldsymbol{J}_{\boldsymbol{\beta}}^{-1}$ being the asymptotic variance of $\widehat{\boldsymbol{\beta}}$. The penalty term $b_{\gamma_n}(\boldsymbol{\beta},\boldsymbol{\gamma})$ therefore represents the $\boldsymbol{\gamma}$-dependency of $\boldsymbol{J}_{\boldsymbol{\beta}}$, weighted by the precision of the estimate $\widehat{\boldsymbol{\beta}}$. In this way it corrects for the effect upon $\psi_{\gamma_n}\left(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}},\boldsymbol{\gamma}\right)$ of using $\widehat{\boldsymbol{\beta}_{\boldsymbol{\gamma}}}$.

We note that $b_{\gamma_n}(\boldsymbol{\beta},\boldsymbol{\gamma}) = \partial \log\left|\boldsymbol{J}_{\boldsymbol{\beta}}^{-1}\right|/\partial\gamma_n$, which may be a more convenient form in some applications.

Since $b_{\gamma_n}(\boldsymbol{\beta},\boldsymbol{\gamma})$ does not depend on the data, we obtain the $\boldsymbol{\gamma}$- and $\boldsymbol{\beta}$-sensitivity of $\breve{\boldsymbol{\psi}}_{\boldsymbol{\gamma}}(\boldsymbol{\beta},\boldsymbol{\gamma})$ by amending $\boldsymbol{S}_{\boldsymbol{\gamma}}$ and $\boldsymbol{S}_{\boldsymbol{\gamma\beta}}$ respectively, with the $\boldsymbol{\gamma}$- and $\boldsymbol{\beta}$-derivatives, of the penalty term, respectively. For the nmth entries we obtain

$$\frac{\partial}{\partial\gamma_m} b_{\gamma_n}(\boldsymbol{\beta},\boldsymbol{\gamma}) = \operatorname{tr}\left(\boldsymbol{J}_{\boldsymbol{\beta}}^{(\gamma_n)}\boldsymbol{J}_{\boldsymbol{\beta}}^{-1}\boldsymbol{J}_{\boldsymbol{\beta}}^{(\gamma_m)}\boldsymbol{J}_{\boldsymbol{\beta}}^{-1} - \boldsymbol{J}_{\boldsymbol{\beta}}^{(\gamma_n,\gamma_m)}\boldsymbol{J}_{\boldsymbol{\beta}}^{-1}\right) \tag{3.20}$$

and

$$\frac{\partial}{\partial\beta_m} b_{\gamma_n}(\boldsymbol{\beta},\boldsymbol{\gamma}) = \operatorname{tr}\left(\boldsymbol{J}_{\boldsymbol{\beta}}^{(\gamma_n)}\boldsymbol{J}_{\boldsymbol{\beta}}^{-1}\boldsymbol{J}_{\boldsymbol{\beta}}^{(\beta_m)}\boldsymbol{J}_{\boldsymbol{\beta}}^{-1} - \boldsymbol{J}_{\boldsymbol{\beta}}^{(\gamma_n,\beta_m)}\boldsymbol{J}_{\boldsymbol{\beta}}^{-1}\right). \tag{3.21}$$

The derivatives of $\boldsymbol{J}_{\boldsymbol{\beta}}$ are listed in Appendix C.

3.7 Godambe information $\boldsymbol{J}_{\boldsymbol{\theta}}$

For joint inference on $\boldsymbol{\theta} = \left(\boldsymbol{\beta}^\top, \boldsymbol{\gamma}^\top\right)^\top$ we use the asymptotic property, valid under mild regularity conditions

$$\widehat{\boldsymbol{\theta}} \sim N\left(\boldsymbol{\theta}, \boldsymbol{J}_{\boldsymbol{\theta}}^{-1}\right),$$

where $\boldsymbol{J}_{\boldsymbol{\theta}}^{-1} = \boldsymbol{S}_{\boldsymbol{\theta}}^{-1}\boldsymbol{V}_{\boldsymbol{\theta}}\boldsymbol{S}_{\boldsymbol{\theta}}^{-\top}$, the inverse Godambe information or the sandwich formula.

The structure of the "bread" $\boldsymbol{S}_{\boldsymbol{\theta}}^{-1}$ in the sandwich formula is (3.10), with blocks listed in (3.16), (3.18) and (3.19). The lower blocks, associated with $\boldsymbol{\gamma}$ are however amended with terms for bias correction as given by (3.20) and (3.21).

The "meat" part $\boldsymbol{V}_{\boldsymbol{\theta}}$ is the variability of $\boldsymbol{\psi}_{\boldsymbol{\theta}}$ and may be structured analogously

$$\boldsymbol{V}_{\boldsymbol{\theta}} = \begin{bmatrix} \boldsymbol{V}_{\boldsymbol{\beta}} & \boldsymbol{V}_{\boldsymbol{\beta\gamma}} \\ \boldsymbol{V}_{\boldsymbol{\gamma\beta}} & \boldsymbol{V}_{\boldsymbol{\gamma}} \end{bmatrix}, \tag{3.22}$$

where, by symmetry, $\boldsymbol{V}_{\boldsymbol{\beta\gamma}} = \boldsymbol{V}_{\boldsymbol{\gamma\beta}}^\top$.

Using (3.10) and (3.22), $\boldsymbol{J}_{\boldsymbol{\theta}}^{-1}$ may be written as

$$\begin{aligned} \boldsymbol{J}_{\boldsymbol{\theta}}^{-1} &= \begin{bmatrix} \boldsymbol{S}_{\boldsymbol{\beta}}^{-1} & \boldsymbol{0} \\ -\boldsymbol{S}_{\boldsymbol{\gamma}}^{-1}\boldsymbol{S}_{\boldsymbol{\gamma\beta}}\boldsymbol{S}_{\boldsymbol{\beta}}^{-1} & \boldsymbol{S}_{\boldsymbol{\gamma}}^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{V}_{\boldsymbol{\beta}} & \boldsymbol{V}_{\boldsymbol{\beta\gamma}} \\ \boldsymbol{V}_{\boldsymbol{\gamma\beta}} & \boldsymbol{V}_{\boldsymbol{\gamma}} \end{bmatrix} \begin{bmatrix} \boldsymbol{S}_{\boldsymbol{\beta}}^{-1} & -\boldsymbol{S}_{\boldsymbol{\gamma}}^{-1}\boldsymbol{S}_{\boldsymbol{\gamma\beta}}^\top\boldsymbol{S}_{\boldsymbol{\beta}}^{-1} \\ \boldsymbol{0} & \boldsymbol{S}_{\boldsymbol{\gamma}}^{-1} \end{bmatrix} \\ &= \begin{bmatrix} \boldsymbol{S}_{\boldsymbol{\beta}}^{-1}\boldsymbol{V}_{\boldsymbol{\beta}}\boldsymbol{S}_{\boldsymbol{\beta}}^{-1} & \boldsymbol{S}_{\boldsymbol{\beta}}^{-1}\left(-\boldsymbol{V}_{\boldsymbol{\beta}}\boldsymbol{S}_{\boldsymbol{\beta}}^{-1}\boldsymbol{S}_{\boldsymbol{\gamma\beta}}^\top + \boldsymbol{V}_{\boldsymbol{\gamma\beta}}^\top\right)\boldsymbol{S}_{\boldsymbol{\gamma}}^{-1} \\ \boldsymbol{S}_{\boldsymbol{\gamma}}^{-1}\left(-\boldsymbol{S}_{\boldsymbol{\gamma\beta}}\boldsymbol{S}_{\boldsymbol{\beta}}^{-1}\boldsymbol{V}_{\boldsymbol{\beta}} + \boldsymbol{V}_{\boldsymbol{\gamma\beta}}\right)\boldsymbol{S}_{\boldsymbol{\beta}}^{-1} & \boldsymbol{S}_{\boldsymbol{\gamma}}^{-1}\left(\boldsymbol{L} + \boldsymbol{V}_{\boldsymbol{\gamma}}\right)\boldsymbol{S}_{\boldsymbol{\gamma}}^{-1} \end{bmatrix}, \end{aligned} \tag{3.23}$$

where $\boldsymbol{L} = \boldsymbol{S}_{\boldsymbol{\gamma\beta}}\boldsymbol{S}_{\boldsymbol{\beta}}^{-1}\left(\boldsymbol{V}_{\boldsymbol{\beta}}\boldsymbol{S}_{\boldsymbol{\beta}}^{-1}\boldsymbol{S}_{\boldsymbol{\gamma\beta}}^\top - \boldsymbol{V}_{\boldsymbol{\gamma\beta}}^\top\right) - \boldsymbol{V}_{\boldsymbol{\gamma\beta}}\boldsymbol{S}_{\boldsymbol{\beta}}^{-1}\boldsymbol{S}_{\boldsymbol{\gamma\beta}}^\top$.

The upper left-hand block of $\boldsymbol{J}_{\theta}^{-1}$ shows that the the asymptotic variance of $\widehat{\boldsymbol{\beta}}$ is unaffected by the simultaneous estimation of $\boldsymbol{\gamma}$; a consequenc of the $\boldsymbol{\gamma}$-insensitivity discussed above. On the other hand, the quantity $\boldsymbol{L}$ in the lower right-hnad block represents the inflation of the asymptotic variance of $\widehat{\boldsymbol{\gamma}}$ caused by the simultaneous estimation of $\boldsymbol{\beta}$. By (3.16) $\boldsymbol{S}_{\boldsymbol{\beta}} = -\boldsymbol{V}_{\boldsymbol{\beta}}$ and therefore the upper right-hand block of (3.23) reduces to $\boldsymbol{S}_{\beta}^{-1}\left(\boldsymbol{S}_{\boldsymbol{\gamma\beta}}^\top + \boldsymbol{V}_{\boldsymbol{\gamma\beta}}^\top\right)\boldsymbol{S}_{\boldsymbol{\gamma}}^{-1}$. If $\boldsymbol{S}_{\boldsymbol{\gamma\beta}} + \boldsymbol{V}_{\boldsymbol{\gamma\beta}} = \boldsymbol{0}$ then $\boldsymbol{S}_{\boldsymbol{\theta}} = -\boldsymbol{V}_{\boldsymbol{\theta}}$ and $\boldsymbol{\psi}$ would have been a quasi score. In this sense the matrix $\boldsymbol{S}_{\boldsymbol{\gamma\beta}}^\top + \boldsymbol{V}_{\boldsymbol{\gamma\beta}}^\top$ measures how much $\boldsymbol{\psi}_{\boldsymbol{\gamma}}$ deviates from being a quasi-score function.

Except for $\boldsymbol{V}_{\boldsymbol{\beta}}$, the blocks of (3.23) rely on 3rd and 4th moments. This seems less tractable for practical use. We may therefore employ empirical variabilities of $\breve{\boldsymbol{\psi}} = \left(\boldsymbol{\psi}_{\boldsymbol{\beta}}^\top, \breve{\boldsymbol{\psi}}_{\boldsymbol{\gamma}}^{\ \top}\right)^\top$, defined by $\widehat{\boldsymbol{V}}_{\boldsymbol{\theta},\mathrm{Emp}} = \frac{1}{I}\sum_i \breve{\boldsymbol{\psi}}_i(\widehat{\boldsymbol{\theta}})\breve{\boldsymbol{\psi}}_i(\widehat{\boldsymbol{\theta}})^\top$ instead. A variant utilizes $\boldsymbol{S}_{\boldsymbol{\beta}} = -\boldsymbol{V}_{\boldsymbol{\beta}}$ and considers only the empirical variability of $\breve{\boldsymbol{\psi}}_{\boldsymbol{\gamma}}$ defined by $\widehat{\boldsymbol{V}}_{\boldsymbol{\gamma},\mathrm{Emp}} = \frac{1}{I}\sum_i \breve{\boldsymbol{\psi}}_{\boldsymbol{\gamma},i}(\widehat{\boldsymbol{\theta}})\breve{\boldsymbol{\psi}}_{\boldsymbol{\gamma},i}(\widehat{\boldsymbol{\theta}})^\top$.

4. Simulation study

Some key properties of our method were addressed by an extensive simulation study in which 4000 data sets were simulated for each of 64 different configurations specified by combinations of the following model parameters: $r_1, r_2 \in \{2.0, 3.0\}$, and $r_3 \in \{1.0, 1.5, 2.0, 3.0\}$. The data sets were simulated with $I \in \{5, 25\}$ clusters, and each cluster consisting of latent time series of length $T \in \{10, 50\}$ having AR(1) correlation structure with $\alpha = 0.5$.

Given the latent variables the mean structure was modelled as $\log(\mu_t) = \beta_0 + \beta_1 x_t$ where $x_1, \ldots, x_T$ were assigned equidistant values in the range from -2.5 to 2.5 and $\beta_0 = 3.0$, $\beta_1 = 0.7$. All simulated data sets were fitted with and without bias correction and both assuming the correct value of the response index parameter r_3 as well as considering it an unknown parameter to be estimated. All parameters but ρ^2 and r_3 were kept fixed across all configurations at the following values: $\sigma^2 = \omega^2 = 0.5$. The dispersion parameter ρ^2 varied with the r_3 index and assumed the values 1.0, 0.5, 0.1 and 0.01 corresponding to the values 1.0, 1.5, 2.0 and 3.0 of r_3 respectively.

The simulation study focused on robustness, bias correction, efficiency and the impact of sample size for both the regression and the association parameters albeit ψ_β is a quasi score estimating function, with well known optimality properties. Parts of the vast amount of simulation results will be reported in the form of tables and figures, whereas other parts will be referred to in the text.

4.1 Robustness of estimating procedure

Simulations with varying configurations of r_1 and r_2 were used for studying the asserted robustness against the lack of knowledge about the Tweedie parameters driving the latent process. Along the same lines we investigated how the model performed across an appropriate range of values for the r_3 parameter and in particular its ability to recover the true value when considering r_3 an unknown parameter to be estimated.

Figure 1 shows the median values of the parameter estimates, for $r_1 = 3.0$ and $r_2 = 2.0$ across the range of r_3 parameter values and the sizes of data sets considered. Similar plots for other combinations of r_1 and r_2, not shown here, indicated similar patterns, although data sets for $r_2 = 3.0$ have larger bias for $\widehat{\omega}^2$ and $\widehat{\rho}^2$. However, this problem is not a real concern, since values of the Twedie power parameter as high as 3 are rare for real data (Kendal, 2004), making such values of r_2 unlikely to occur in practice.
For comparison parameter estimates are shown on the same scale across varying sample sizes.

Our approach appears reasonably robust since it allows for treating the response model parameter r_3 as part of the parameters to be estimated, and this parameter can be estimated with virtually no bias across varying specifications of the latent process. Furthermore, estimates of the parameters showed the same patterns independently of the choice of variables driving the latent process.

The asymptotic variances of the estimates of the association parameters (3.23), enable us to compute estimates of coverage probabilities for 95% asymptotic confidence intervals. Table 2 gives the coverage probabilities based on bias corrected estimates.

The coverages indicate that the standard errors of the parameters are reasonably precise for all parameters except for $\widehat{\omega}^2$ and $\widehat{\rho}^2$, suggesting that our approach is reasonably accurate. The coverages are, by convention, based on symmetric intervals. These may, however, be less suitable for dispersion parameters estimates when sample sizes are small, due to the asymmetric nature of the sampling distribution of these estimates. Not surprisingly, the coverages are too high for the smaller sample sizes, indicating overestimation of their stan-

Table 2. Summary of coverages for 95 % asymptotic confidence intervals across varying configurations, based on bias corrected estimates.

T	I	β_0	β_1	σ^2	ω^2	ρ^2	α	r_3
10	5	96.35%	98.35%	94.00%	83.60%	80.35%	99.80%	96.90%
10	25	96.10%	96.50%	94.55%	90.45%	88.25%	97.75%	96.70%
50	5	95.50%	95.35%	92.55%	89.20%	89.50%	93.00%	95.25%
50	25	94.00%	94.30%	90.80%	87.55%	89.30%	90.10%	91.25%

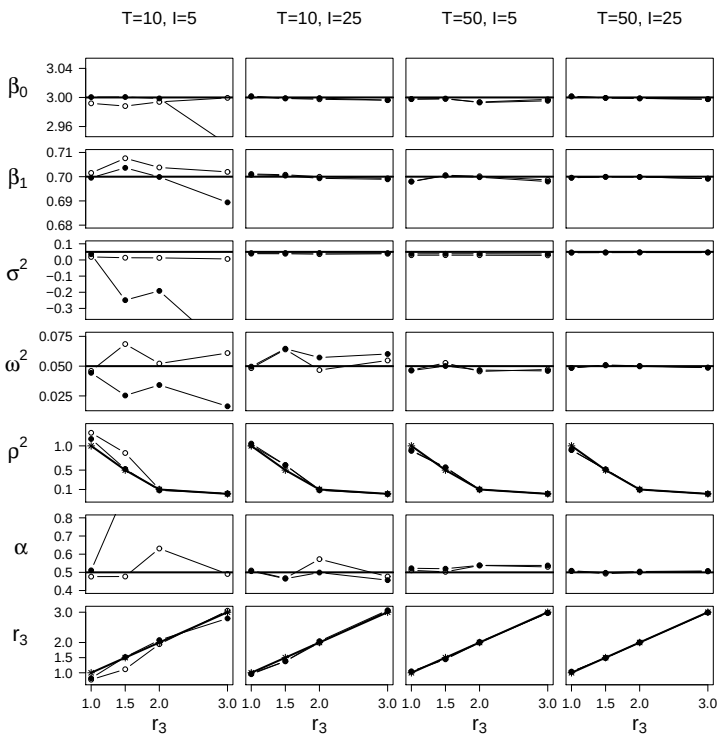

Figure 1. Median values of estimated model parameters for bias-corrected (•) and un-corrected (○) estimation compared with the true values (∗).

dard errors. From results not shown here, we observed that the estimation of r_3 improved the coverage for all parameters except for α which was unchanged.

4.2 Bias correction

The magnitude of the parameter bias correction was assessed by a duplicate analysis of the simulated data sets, except that the second estimation was done without bias correction. All parameters show the same pattern of almost negligible bias for all but the smallest sample size. For the smallest sample size, given by $I = 5$ and $T = 10$, the bias correction appears to reduce the bias for the regression parameters β_0 and β_1, the response dispersion ρ^2 and the response index r_3. The correction seemed however to have an adverse effect for the latent process parameters σ^2, ω^2 and α. With the rather short series and few clusters, this may well be interpreted as lack of information rather than a deficiency of the model.

For the larger data sets a closer inspection of the estimations, as in Figure 2, showed an almost consistent pattern of the bias correction pulling the estimates closer to their true values. The bias correction improved the coverages for σ^2, whereas it decreased for ω^2 and ρ^2 and was unchanged for the remaining parameters.

5. Data Analysis

Knowledge about the growth of fish is important for the assessment of fish biomass. For this purpose, many fisheries management programmes sample otoliths on a regular basis. An otolith is a structure located in the inner ear of fish and is built by deposit of calcium carbonate, protein and a variety of trace elements. It carries information about age and growth patterns, by means of alternating opaque and translucent bands. When viewed in

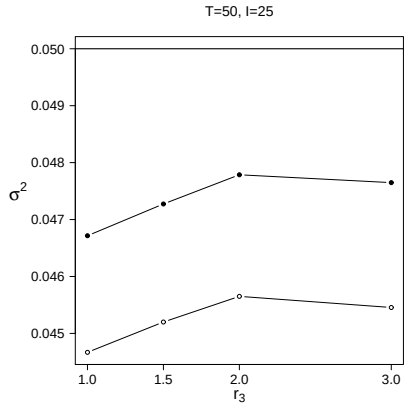

Figure 2. Median values of σ^2 for bias-corrected (•) and un-corrected (○) estimation compared with the true values (horizontal line).

transmitted light, a translucent band represents a low level of deposition of proteins in the calcium carbonate crystal structure corresponding to a period of slow growth (Mosegaard and Titus, 1987). Sub-seasonal bands, representing daily cycles, can sometimes be identified within the annual bands (Pannella, 1971).

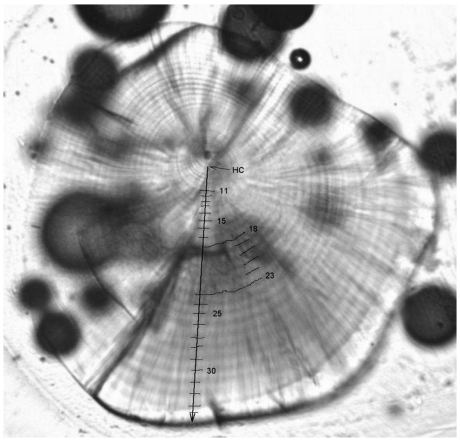

Figure 3. A photo of an otolith with radius and band marks. The trace marks for band 18–23 are identified outside the radius and only the span of these bands can be measured. HC: Hatch Check.

The data collected and analysed by Clausen et al. (2007) contains measurements of the width of daily growth bands of otoliths collected from juvenile herrings (*Clupea harengus*) (Figure 3). The age in days was determined, by counting bands, for each of the sampled specimen and along with the time of sampling they were categorized as being offspring from one of three spawner types: autumn, winter and spring. These are distinct stock components but mix on the nursery and feeding grounds. For stock assessment purposes, it is of interest to be able to discriminate between them. Clausen et al. (2007) used otolith characteristics for this purpose.

We let each fish be a cluster, whereas the sequence of bands within each fish gives the longitudinal structure. The covariance structure that simultaneously encompass the cluster effects along with a conditional serial correlation within the latent process, are suitably handled by our framework, whereas this is not the case for conventional GLMM and GEE approaches. We analysed the data from Clausen et al. (2007), in order to illustrate the application of our model to such data.

For compatibility across the collection of otoliths, the band widths are measured along similar radii on all otoliths. If the band marks are not all clearly identifiable along this

transect, two or more adjacent bands are aggregated and the total width of these bands is taken (Figure 3). The count of bands between two measurement marks is then based on intermediate band marks identified elsewhere on the otoliths. It is common practice to use the average of such aggregated bands, in place of correct measurements; a feature also found in the present data. To avoid successive values obtained from the same aggregation of bands, it sufficed to sub-sample every 8th value for our analysis.

The sampled fish have different ages and therefore display differences in the lengths of their band width series. To avoid bias caused by data selection, we truncated the sequences of band measurements to the shortest sequence within each spawning category. Furthermore the first 10 bands were left out of the analysis, as their measurements were considered too imprecise.

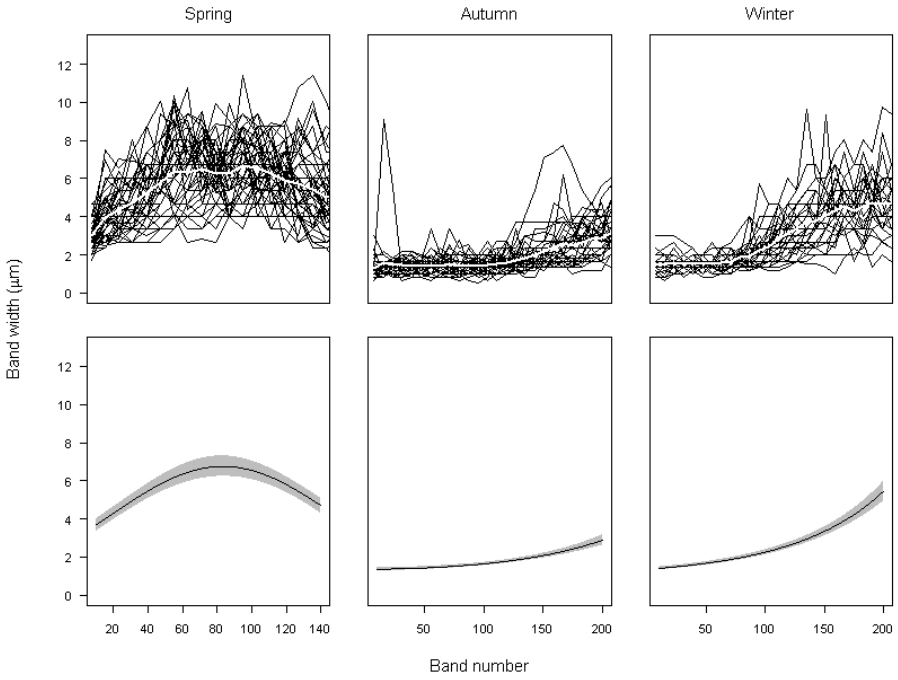

Figure 4. Width measurements at every 8th band. Thick white line indicates average measurements over otoliths. Top row: observed, bottom row: estimated.

Two variant models were estimated: one fixing the response Tweedie parameter $r_3 = 2$, corresponding to a gamma distribution and the other one estimating this parameter. Judging from exploratory plots (Figure 4) of the observed width measurements the fixed part could appropriately be modeled as 2nd order polynomials of the bands and with potential different coefficients for the three seasons: `width` $\sim$ `(band+band`^{2}`)*season` using the log link.

The initial models contained nine regression parameters. Autumn was chosen as base level for the season factor, to enable a direct comparison between autumn and winter, as these appeared most alike among the three seasons. The models were reduced to final models, with six regression parameters, through a succession of Wald tests and re-estimations (Rotnitzky and Jewell, 1990). Seasonal variation is likely to induce serial correlation between the bands. Based on inspection of the auto-correlation and the partial auto-correlation function for individual otoliths, an AR(1) model was deemed appropriate. Estimates for the parameters of the final models and standard errors are listed in Table 3. Similarly to other types of generalized linear regression models these are interpreted as linear effects on the scale of the link function, cf. equation (2.6). The dispersion

parameters σ^2, ω^2 and ρ^2 divide the total variance by means of equation (2.7) into components accounting for variance between otoliths, variance between bands within otoliths and observational error respectively. Finally the parameter α describes the latent process correlation between bands via the assumed AR(1) structure.

Table 3. Parameter estimates, standard errors (SE) and p-values for both fixed and estimated r_3 models. `aut`: autumn; `win`: winter; `spr`: spring. * The p-value for r_3 applies to the hypothesis $H_0 : r_3 = 2$.

	Fixed $r_3 = 2$			Estimated r_3		
Parameter	Est	SE	p-value	Est	SE	p-value
$\beta_{\texttt{aut+win}}$	0.3118	0.0292	< 0.0001	0.3113	0.0292	< 0.0001
$\beta_{\texttt{spr}}$	1.1304	0.0553	< 0.0001	1.1288	0.0556	< 0.0001
$\beta_{\texttt{spr:band}}$	0.0187	0.0016	< 0.0001	0.0187	0.0016	< 0.0001
$\beta_{\texttt{win:band}}$	0.0032	0.0003	< 0.0001	0.0032	0.0003	< 0.0001
$\beta_{\texttt{(aut+win):band}^2}$	1.9×10^{-5}	1.4×10^{-6}	< 0.0001	1.9×10^{-5}	1.4×10^{-6}	< 0.0001
$\beta_{\texttt{spr:band}^2}$	-0.0001	9.9×10^{-6}	0.0001	-0.0001	1.0×10^{-5}	< 0.0001
σ^2	0.0040	0.1639	0.9807	0.0042	0.1627	0.9796
ω^2	0.0277	2.6770	0.9918	0.0275	3.4696	0.9937
ρ^2	0.0153	4.1817	0.9971	0.0115	7.0517	0.9987
α	0.8321	0.0823	< 0.0001	0.8325	6.7673	0.9021
r_3				2.2700	1.7627	0.8783*

The two regressions are estimated to be almost exactly the same, whether r_3 is estimated or assumed known. This reflects the γ-insensitivity of ψ_β. From the fit we conclude that autumn and winter differ only by the 1st order term whereas autumn and spring differ by all three terms. The fitted curves from the fixed-r_3 model are plotted in Figure 4.

The association parameter estimates were very similar for the two models, except for having quite different standard errors. The rather large standard errors for the dispersion parameter estimates confirm the impression from the simulations in Section 4, that the use of empirical variabilities in the sandwich estimator (3.23) may lead to standard errors too big to be of any practical use. This seems to be the cost of avoiding the use of higher-order moments. The standard errors for the correlation parameter α in the fixed-r_3 case appear to be more realistic but raise dramatically when r_3 is estimated. In that case the standard error for r_3 seems moderate. The α parameter applies to the scale of the sub-sampling frequency. A back calculation to the day-to-day serial correlation, and assuming the AR(1) model, leads to a correlation value of about 0.97.

6. Discussion

As already discussed in Section 1, our main goal was to create a method that could unify generalized linear mixed models and generalized estimating equations in a computationally efficient way. We have achieved this by means of a versatile hierarchical modelling approach using Tweedie distributions, combined with an estimating equation approach based on second-moment assumptions. As discussed in connection with equation (2.6), the random effects part of the proposed models resembles conventional generalized linear mixed models, and the quasi-score function (3.8) is similar to a generalized estimating equation. Overall, our approach has a number of advantages compared with existing methods, such as:

(1) The models have a simple closed form expression for the covariance structure, in contrast to conventional generalized linear mixed models, where moments are not available in closed form.

(2) The models have both a conditional and marginal interpretation and therefore obviate the conventional need to distinguish between generalized estimating equations and generalized linear mixed models.
(3) The explicit hierarchical modelling approach based on Tweedie distributions allows easy simulation, in contrast to generalized estimating equations, which are not based on explicit stochastic modelling.
(4) A comprehensive range of distributions for the response level are available, controlled by the parameter r_3, and this parameter need not be specified and instead can be estimated. The estimation method does not require specification of r_1 and r_2, making the method very versatile in terms of accommodating different distributional shapes both at the observation and latent variable levels.

Our model hinges on the scale transformation properties, that characterizes the class of Tweedie distributions. This seemingly exclude binary and categorical data to be handled by the model, since their distributions are not of Tweedie type. This limitation can however be remedied by using conditionally independent Poisson response variables, leading to beta-binomial-like or Dirichlet-multinomial-like models. The topic is of considerable importance for both practioners as well as for theoretical statisticians and it is currently under investigation. Another useful extension of the method would be to allow regression modelling for the association parameters, along the lines of Davidian and Carroll (1987). It would also be straightforward to extend the model to multiple levels of random effects, adding further levels by repeated use of conditional Tweedie distributions.

Acknowledgments

The authors thank Rob Fryer for valuable comments on our approach and Lotte Worsøe Clausen and Helge Paulsen for providing us with the otolith data and guidance in understanding them. We are also grateful for comments from the associate editor and two anonymous referees.

Appendix A

Covariance structure

From the model specification (2.1)–(2.4) we derive the marginal covariance between two observations within the ith cluster. This is done in three steps by means of the law total of variance. From $E(Z_i) = 1$ and $\mathrm{var}(Z_i) = \sigma^2$ we first get

$$\begin{aligned}\mathrm{cov}(Z_{it},\, Z_{it'}) &= E\{\mathrm{cov}(Z_{it},\, Z_{it'} \mid Z_\bullet)\} + \mathrm{cov}(E(Z_{it} \mid Z_\bullet),\, E(Z_{it'} \mid Z_\bullet)) \\ &= \delta_t^{t'} E(\omega^2 Z_i) + \alpha_+^{-2}\mathrm{var}(Z_i) \\ &= \delta_t^{t'}\omega^2 + \alpha_+^{-2}\sigma^2,\end{aligned}$$

from which we obtain

$$\begin{aligned}\mathrm{cov}(Q_{it},\, Q_{it'}) &= \sum_{s=0}^{\infty}\sum_{s'=0}^{\infty}\alpha_s\alpha_{s'}\,\mathrm{cov}(Z_{i\,t-s},\, Z_{i\,t'-s'}) \\ &= \omega^2\sum_{s=0}^{\infty}\alpha_s\alpha_{s+|t-t'|} + \sigma^2,\end{aligned}$$

and finally arrive at

$$\begin{aligned}\mathrm{cov}(Y_{it},\, Y_{i't'}) &= E\{\mathrm{cov}(Y_{it},\, Y_{it'} \mid Z)\} + \mathrm{cov}\{E(Y_{it} \mid Z), E(Y_{i't'} \mid Z)\} \\ &= \delta_t^{t'}\,\mathrm{var}(Y_{it}) + \mu_{it}\mu_{ij'}\,\mathrm{cov}(Q_{it},\, Q_{it'}) \\ &= \delta_t^{t'}\,\mu_{it}^{r_3}\,\rho^2 + \mu_{it}\mu_{it'}\left(\omega_j^2\sum_{s=0}^{\infty}\alpha_s\alpha_{s+|t-t'|} + \sigma^2\right)\end{aligned}$$

Appendix B

Process correlation matrix $\boldsymbol{K}(\boldsymbol{\alpha})$ for ma(p) and ar(1) processes

The latent process linear filter $Q_{it} = \sum_{s=0}^{\infty}\alpha_s Z_{i\,t-s}$, induces the process correlation matrix $\boldsymbol{K}(\boldsymbol{\alpha})$, with tt'th entry $\{\boldsymbol{K}(\boldsymbol{\alpha})\}_{tt'} = \sum_{s=0}^{\infty}\alpha_s\alpha_{s+|t-t'|}$.

An ma(p) process is given by $\alpha_0 = 1$ and $\alpha_s = 0$ for $s > p$ and has first and second derivative matrices with tt'th entries given by

$$\left\{\frac{\partial}{\partial\alpha_k}\boldsymbol{K}(\boldsymbol{\alpha})\right\}_{tt'} = \alpha_{k+|t-t'|}\delta_{k\le p-|t-t'|} + \alpha_{k-|t-t'|}\delta_{k\ge|t-t'|}$$

and

$$\left\{\frac{\partial^2}{\partial\alpha_k\partial\alpha_m}\boldsymbol{K}(\boldsymbol{\alpha})\right\}_{tt'} = \delta_t^{t'}\delta_k^m + \delta_{|t-t'|}^{|k-m|}$$

respectively. Here $\delta_t^{t'}$, $\delta_{k\le p-|t-t'|}$ etc. are variant forms of the Kronecker delta with obvious definitions.

An AR(1) process is given by $\alpha_s = \alpha^s$; $\alpha \in (0,1)$, from which we obtain the tt'th entry of $\boldsymbol{K}(\alpha)$

$$\{\boldsymbol{K}(\alpha)\}_{tt'} = \sum_{s=0}^{\infty} \alpha^{2s+|t-t'|} = \alpha^{|t-t'|} \sum_{s=0}^{\infty} \alpha^{2s} = \frac{\alpha^{|t-t'|}}{1-\alpha^2}.$$

Consequently the kth sub- and super-diagonal of $\frac{\partial}{\partial\alpha}\boldsymbol{K}(\alpha)$ and $\frac{\partial^2}{\partial\alpha^2}\boldsymbol{K}(\alpha)$ have elements

$$\frac{\partial}{\partial\alpha}\left(\frac{\alpha^k}{1-\alpha^2}\right) = \frac{(2-k)\,\alpha^{k+1} + k\alpha^{k-1}}{(1-\alpha^2)^2}$$

and

$$\frac{\partial^2}{\partial\alpha^2}\left(\frac{\alpha^k}{1-\alpha^2}\right) = \frac{\left(k^2-5k+6\right)\alpha^{(k+2)} + \left(-2k^2+6k+2\right)\alpha^k + (k-1)\,k\alpha^{(k-2)}}{(1-\alpha^2)^3}$$

respectively.

Appendix C

Derivatives of $\boldsymbol{J}_{\boldsymbol{\beta}}$

The derivatives of $\boldsymbol{J}_{\boldsymbol{\beta}}$ involved in calculating the $\boldsymbol{\gamma}$- and $\boldsymbol{\beta}$- sensitivities of $\breve{\boldsymbol{\psi}}_{\boldsymbol{\gamma}}(\boldsymbol{\beta},\boldsymbol{\gamma})$ are

$$\begin{aligned}
\boldsymbol{J}_{\boldsymbol{\beta}}^{(\gamma_n)} &= -\sum_{i=1}^{I} \boldsymbol{D}_i^\top \boldsymbol{W}_{in} \boldsymbol{D}_i \\
\boldsymbol{J}_{\boldsymbol{\beta}}^{(\beta_m)} &= \sum_{i=1}^{I} \boldsymbol{D}_i^{(\beta_m)\top} \boldsymbol{C}_i^{-1} \boldsymbol{D}_i + \boldsymbol{D}_i^\top \boldsymbol{C}_i^{-1} \boldsymbol{D}_i^{(\beta_m)} - \boldsymbol{D}_i^\top \boldsymbol{C}_i^{-1} \boldsymbol{C}_i^{(\beta_m)} \boldsymbol{C}_i^{-1} \boldsymbol{D}_i \\
\boldsymbol{J}_{\boldsymbol{\beta}}^{(\gamma_n,\gamma_m)} &= \sum_{i=1}^{I} \boldsymbol{D}_i^\top \left(\boldsymbol{W}_{im} \boldsymbol{C}_i W_{in} + \boldsymbol{W}_{in} \boldsymbol{C}_i \boldsymbol{W}_{im} - \boldsymbol{C}_i^{-1} \boldsymbol{C}_i^{(\gamma_n,\gamma_m)} \boldsymbol{C}_i^{-1}\right) \boldsymbol{D}_i \\
\boldsymbol{J}_{\boldsymbol{\beta}}^{(\gamma_n,\beta_m)} &= -\sum_{i=1}^{I} \boldsymbol{D}_i^{(\beta_m)\top} \boldsymbol{W}_{in} \boldsymbol{D}_i + \boldsymbol{D}_i^\top \boldsymbol{W}_{in} \boldsymbol{D}_i^{(\beta_m)} + \boldsymbol{D}_i^\top \boldsymbol{W}_{in}^{(\beta_m)} \boldsymbol{D}_i,
\end{aligned}$$

where $\boldsymbol{W}_{in} = -\frac{\partial}{\partial\gamma_n}\boldsymbol{C}_i^{-1} = \boldsymbol{C}_i^{-1}\left(\frac{\partial}{\partial\gamma_n}\boldsymbol{C}_i\right)\boldsymbol{C}_i^{-1}$ and $\boldsymbol{D}_i^{(\beta_m)} = \frac{\partial}{\partial\beta_m}\boldsymbol{D}_i$ etc.

References

Balemi, A., Lee, A., 2009. Comparison of gee1 and gee2 estimation applied to clustered logistic regression. Journal of Statistical Computation and Simulation, 79(4), 361–378.

Breslow, N. E., Clayton, D. G., 1993. Approximate inference in generalized linear mixed models. Journal of the American Statistical Association, 88(421), 9–25.

Clausen, L. A. W., Bekkevold, D., Hatfield, E. M. C., Mosegaard, H., 2007. Application and validation of otolith microstructure as a stock identification method in mixed atlantic herring (clupea harengus) stocks in the north sea and western baltic. ICES J. Mar. Sci., 64(2), 377–385.

Coull, B. A., Houseman, E. A., Betensky, R. A., 2006. A computationally tractable multivariate random effects model for clustered binary data. Biometrika, 93, 587–599.

Cox, D. R., Reid, N., 1987. Parameter orthogonality and approximate conditional inference. Journal of the Royal Statistical Society, Series B, 49, 1–39.

Crowder, M., 1987. On linear and quadratic estimating functions. Biometrika, 74(3), 591–597.

Davidian, M., Carroll, R. J., 1987. Variance function estimation. Journal of the American Statistical Association, 82(400), 1079–1091.

Diggle, P. J., Heagerty, P., Liang, K.-Y., Zeger, S. L., 2002. Analysis of Longitudinal Data. Oxford University Press, Oxford, 2nd edition.

Fitzmaurice, G. M., Laird, N. M., Ware, J. H., 2004. Applied longitudinal analysis. Wiley-Interscience, Hoboken, N.J.

Godambe, V. P., 1960. An optimum property of regular maximum likelihood estimation. Annals of Mathematical Statistic, 31(4), 1208–1211.

Hall, D. B., 2001. On the application of extended quasi-likelihood to the clustered data case. The Canadian Journal of Statistics, 29(1), 77–97.

Hall, D. B., Severini, T. B., 1998. Extended generalized estimating equations for clustered data. Journal of the American Statistical Association, 93(444), 1365–1375.

Hall, P., Mller, H.-G., Yao, F., 2008. Modelling sparse generalized longitudinal observations with latent gaussian processes. Journal of the Royal Statistical Society, Series B, 70(4), 703–723.

Hardin, J. W., Hilbe, J. M., 2003. Generalized Estimating Equations. Chapman & Hall/CRC, Boca Raton, FL.

Henderson, C. R., 1975. Best linear unbiased estimation and prediction under a selection model. Biometrics, 31(2), 423–447.

Heyde, C. C., 1997. Quasi-Likelihood and Its Application: A General Approach to Optimal Parameter Estimation. Springer Series in Statistics. Springer.

Jiang, J., Luan, Y., Wang, Y.-G., 2007. Iterative estimating equations: Linear convergence and asymptotic properties. Annals of Statistics, 35(5), 2233–2260.

Jørgensen, B., 1987. Exponential dispersion models. Journal of the Royal Statistical Society. Series B (Methodological), 49(2), 127–162.

Jørgensen, B., 1997. The Theory of Dispersion Models. Chapman & Hall, London.

Jørgensen, B., Knudsen, S. J., 2004. Parameter orthogonality and bias adjustment for estimating functions. Scandinavian Journal of Statistics, 31(1), 93–114.

Jørgensen, B., Lundbye-Christensen, S., Song, P. X.-K., Sun, L., 1996. A longitudinal study for emergency room visits and air pollution for prince george, brittish columbia. Statistics in Medicine, 15, 823–836.

Jørgensen, B., Lundbye-Christensen, S., Song, P. X.-K., Sun, L., 1999. A state space model for multivariate longitudinal count data. Biometrika, 86(1), 169–181.

Jørgensen, B., Song, P.-K., 2007. Stationary state space models for longitudinal data. Canadian Journal of Statistics, 34, 1–23.

Jørgensen, B., Song, P. X.-K., 2006. Diagnosis of stationarity in state space models for longitudinal data. Far East J. Theor. Stat., 19, 43–59.

Jowaheer, V., Sutradhar, B. C., 2002. Analysing longitudinal count data with overdispersion. Biometrika, 89(2), 389–399.

Kendal, W. S., 2004. Taylor's ecological power law as a consequence of scale invariant exponential dispersion models. Ecological Complexity, 1(3), 193–209.

Lee, Y., Nelder, J. A., Pawitan, Y., 2006. Generalized. Linear Models with Random Effects: Unified Analysis via. H-likelihood. Chapman and Hall, London.

Liang, K.-Y., Zeger, S. L., 1986. Longitudinal data analysis using generalized linear models. Biometrika, 73(1), 13–22.

Liang, K.-Y., Zeger, S. L., Qaqish, B., 1992. Multivariate regression analyses for categorical data (with discussion). Journal of the Royal Statistical Society, Series B, 54(1), 3–40.

Ma, R., 1999. An orthodox BLUP approach to generalized linear mixed models. PhD thesis, The University of British Columbia.

Ma, R., Jørgensen, B., 2007. Nested generalized linear mixed models: an orthodox best linear unbiased predictor approach. Journal of the Royal Statistical Society: Series B, 69(4), 625–641.

Ma, R., Jørgensen, B., Willms, J., 2009. Clustered binary data with random cluster sizes: A dual poisson modelling approach. Statist. Mod., 9(2), 137–150.

Ma, R., Krewski, D., Burnett, R. T., 2003. Random effects cox models: a poisson modelling approach. Biometrika, 90(1), 157–169.

Magnus, J. R., Neudecker, H., 1999. Matrix Differential Calculus with Applications in Statistics and Econometrics, 2nd Edition. John Wiley & Sons, Chichester.

McCullagh, P., Tibshirani, R., 1990. A simple method for the adjustment of profile likelihoods. Journal of the Royal Statistical Society B, 52(2), 325–344.

McCulloch, C. E., Searle, S. R., 2001. Generalized, linear, and mixed models. Wiley-Interscience, New York.

Molenberghs, G., Verbeke, G., Demétrio, C. G. B., Vieira, A. M. C., 2010. A family of generalized linear models for repeated measures with normal and conjugate random effects. Statistical science, 25(3), 325–347.

Mosegaard, H., Titus, R., 1987. Daily growth rates of otoliths in yolk sac fry of two salmonid species at five different temperatures. In Kallander, S., Farnholm, B., editors, The V Congress of European Ichthyology, Stockholm, pages 221–227. Swedish Museum of Natural History, Stockholm.

Pannella, G., 1971. Fish otoliths: Daily growth layers and periodical patterns. Science, 173(4002), 1124–1127.

Pinheiro, J. C., Bates, D. M., 2000. Mixed-effects models in S and S-PLUS. Springer, New York.

Rotnitzky, A., Jewell, N. P., 1990. Hypothesis testing of regression parameters in semiparametric generalized linear models for cluster correlated data. Biometrika, 77(3), 485–497.

Schall, R., 1991. Estimation in generalized linear models with random effects. Biometrika, 78(4), 719–727.

Verbeke, G., Molenberghs, G., 2000. Linear mixed models for longitudinal data. New York: Springer.

Wang, M., Hanfelt, J. J., 2007. Orthogonal locally ancillary estimating functions for matched pair studies and errors in covariates. Journal of the Royal Statistical Society, Series B, 69, 411–428.

Wang, Y.-G., Carey, V., 2004. Unbiased estimating equations from working correlation models for irregularly timed repeated measures. Journal of the American Statistical Association, 99(467), 845–853.

Wolfinger, R., O'Connell, M., 1993. Generalized linear mixed models a pseudo-likelihood approach. Journal of Statistical Computation and Simulation, 48(3-4), 233–243.

Ziegler, A., Kastner, C., Blettner, M., 1998. The generalized estimation equations: An annotated bibliography. Biometrical Journal, 40, 115–139.