Complex Data Modelling Using Likelihood and H-likelihood Methods

Marek Molas

Complex Data Modelling Using Likelihood and H-Likelihood Methods

Marek Molas

The studies described in this thesis were financially supported by the Department of Biostatistics of Erasmus Medical Center of Erasmus University Rotterdam

Cover design:
Florida Center for Instructional Technology http://etc.usf.edu/clipart/ (Windmill)
Steve Laube http://thelob.deviantart.com/ (Dreamcatcher)
Michal Mazurkiewicz (Design)

Printed by: HAVEKA B.V. http://www.haveka.nl

ISBN: 978-90-9027110-1

Complex Data Modelling Using Likelihood and H-Likelihood Methods

Modelleren van complexe data door middel van aannemelijkheids en h-aannemelijkheids methoden

Proefschrift

ter verkrijging van de graad van doctor aan de
Erasmus Universiteit Rotterdam
op gezag van de rector magnificus

Prof.dr. H.G. Schmidt

en volgens besluit van het College voor Promoties

De openbare verdediging zal plaatsvinden op
dinsdag 20 november 2012 om 15:30 uur

door

Marek Molas

geboren te Lublin, Polen

ERASMUS UNIVERSITEIT ROTTERDAM

Promotiecommissie:

Promotor:	Prof.dr. E.M.E.H. Lesaffre
Overige leden:	Prof.dr. P.H.C. Eilers
	Prof.dr. P.J.F. Groenen
	Prof.dr. T. Stijnen

TO MY PARENTS RYSZARD AND JOLANTA

Publications and manuscripts based on the studies described in this thesis:

Chapter 2: Molas M. and Lesaffre E. (2012). Finite Mixture Models with Fixed Weights Applied to Growth Data. Submitted to Biometrical Letters.

Chapter 3: Molas M. and Lesaffre E. (2008). A comparison of three random effects approaches to analyze repeated bounded outcome scores with an application in a stroke revalidation study. *Statistics in Medicine* **27** pp. 6612-6633.

Chapter 4: Molas M. and Lesaffre E.(2011). Hierarchical generalized linear models: the R package HGLMMM. *Journal of Statistical Software* **39** pp. 1-20.

Chapter 5: Molas M. and Lesaffre E. (2010). Hurdle models for multilevel zero-inflated data via h-likelihood. *Statistics in Medicine* **29** pp. 3294-3310.

Chapter 6: Molas M., Noh M., Lee Y., and Lesaffre E. (2012). Joint hierarchical generalized linear models with multivariate Gaussian random effects. Submitted to Computational Statistics and Data Analysis.

Contents

Chapter 1 General Introduction

An important inference framework in statistical methods is based on the likelihood function. First we will review the likelihood theory for independent observations. Then we use correlated data likelihood modeling concepts to introduce the h-likelihood framework.

1.1 Estimation based on likelihood

Suppose that a cross-sectional study was set up to examine a specific measure, say blood pressure, of individuals from a particular population. The purpose is most often to characterize the distribution of that measure, or an aspect of it such as the mean, in the population. Usually we assume that the distribution belongs to a particular class of distributions where each element of that class is specified by a d-dimensional parameter $\boldsymbol{\theta}$. For the population under consideration we assume that there exists a true, but unknown value of the parameter denoted by $\boldsymbol{\theta}_T$. The sample data from N independent individuals collected with the cross-sectional study, i.e. $\mathbf{y} = \{y_1, \ldots, y_N\}$, are assumed to throw light on $\boldsymbol{\theta}_T$. In statistical inference, the goal is to establish a mechanism to conclude from the data what the true $\boldsymbol{\theta}_T$ could be. It must be said that the assumption of a particular distribution $f_{\boldsymbol{\theta}}(\mathbf{y})$, may be too simplistic in practice. But this assumption, if the model is not taken too rigid, allows us to use an important framework of statistical inference. Namely, based on the model $f_{\boldsymbol{\theta}}(\mathbf{y})$ we can define the probability to observe $\mathbf{y}$ as a function of $\boldsymbol{\theta}$ as follows:

$$\mathcal{L}(\boldsymbol{\theta}|\mathbf{y}) = \prod_{i=1}^{N} f_{\boldsymbol{\theta}}(y_i).$$

The above expression is called the likelihood function for the sample of independent observations at hand. This function describes the plausibility of each parameter value $\boldsymbol{\theta}$ in the light of the observed data and can therefore be used to provide an estimate for the unknown parameter of the population. Such an estimate can be estimated from the value that maximizes the likelihood function, i.e. using the maximum likelihood estimator (MLE) denoted as $\hat{\boldsymbol{\theta}}$. The MLE $\hat{\boldsymbol{\theta}}$ estimator has the following properties under regularity conditions:

1. *Consistency:* The MLE is consistent. This means that, when the sample size increases, the MLE value tends to the true value $\boldsymbol{\theta}_T$ in probability.

2. *Asymptotic distribution:* The asymptotic distribution of the MLE is normal with expectation $\boldsymbol{\theta}_T$ and variance-covariance matrix equal to the inverse of the information matrix (negative 2nd derivative of the likelihood with respect to the parameters of interest). This implies that the MLE is asymptotically unbiased, meaning that for large samples, in repeated experiments the value given by the MLE will be on average equal to $\boldsymbol{\theta}_T$. The approximation improves if the derivative of the log-likelihood is close to linear in a region around the MLE or equivalently, when close to the MLE value the log-likelihood approximately quadratic. In Section 1.2 we show how the approximating distribution is derived.

3. *Efficiency:* The MLE has asymptotically the highest precision among all unbiased estimators for large samples. One says that it attains the Cramer-Rao lower bound. Note that efficiency refers to the precision the estimator exhibits in detecting $\boldsymbol{\theta}_T$ for a given sample size. Formally precision is the inverse of the standard error of the estimator. The higher the precision the estimator has, the more likely is the estimate to be closer to the true value of the parameter on average.

The discussion of the above properties, as well as the regularity conditions can be found in Severini (2000), Schervish (1995) and Pawitan (2001). The versatility of the likelihood function as a basis of the estimation, and a clear principle how to use it to obtain estimates combined with the above attractive properties of the MLE, makes likelihood-based estimation attractive. Further, not only estimation and properties of estimates might be derived but also statistical hypothesis testing is easily performed within the likelihood framework. Hypothesis testing based on the likelihood function is briefly reviewed in next section.

1.2 Hypothesis testing based on likelihood

We present here some main results, and computational aspects of likelihood-based hypothesis testing. Three tests which can be derived from the likelihood are commonly used: the likelihood ratio test, the Wald test and the score test. In this section we summarize the theory which ensures the validity of these statistical tests and show their derivation. It is customary to work with the logarithm of the likelihood. The derivative of the log-likelihood is called the score vector, i.e.

$$\mathcal{S}(\boldsymbol{\theta}|\mathbf{y}) = \frac{\partial \log \mathcal{L}(\boldsymbol{\theta}|\mathbf{y})}{\partial \boldsymbol{\theta}}.$$

The MLE of $\boldsymbol{\theta}_T$ is found by solving $\mathcal{S}(\boldsymbol{\theta}|\mathbf{y}) = \mathbf{0}$. It is important to understand that the score vector is, like the likelihood value, a random variable depending on the collected sample. The score statistic has the following properties:

$$E[\mathcal{S}(\boldsymbol{\theta}_T|\mathbf{y})] = \int \left[\frac{\partial \log \mathcal{L}(\boldsymbol{\theta}|\mathbf{y})}{\partial \boldsymbol{\theta}} |_{\boldsymbol{\theta}=\boldsymbol{\theta}_T} \right] f_{\boldsymbol{\theta}_T}(\mathbf{y}) d\mathbf{y} = \mathbf{0}, \tag{1.1}$$

$$Var[\mathcal{S}(\boldsymbol{\theta}_T|\mathbf{y})] = E[\mathcal{S}(\boldsymbol{\theta}_T|\mathbf{y}) - E(\mathcal{S}(\boldsymbol{\theta}_T|\mathbf{y}))]^2 = -E\left(\frac{\partial^2 \log \mathcal{L}(\boldsymbol{\theta}|\mathbf{y})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} |_{\boldsymbol{\theta}=\boldsymbol{\theta}_T} \right). \tag{1.2}$$

Note that these properties hold under the true value $\boldsymbol{\theta}_T$. The second expression defines the variance of the score statistic and is called the (expected) Fisher information. As a function of $\boldsymbol{\theta}$, the expected information tells how ‘hard’ it is to estimate $\boldsymbol{\theta}$. Namely, parameters with greater information can be estimated more easily, requiring a smaller sample size to achieve a required precision. By the observed Fisher information matrix we mean the negative of the Hessian of the log likelihood with respect to the parameters $\boldsymbol{\theta}$, denoted by $I(\boldsymbol{\theta}) = -\frac{\partial^2 \log \mathcal{L}(\boldsymbol{\theta}|\mathbf{y})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}$. Note that the observed Fisher information is a realization of a random variable at given $\boldsymbol{\theta}$, and has a different distribution for every $\boldsymbol{\theta}$.

We first need to see how the asymptotic distribution of the MLE $\hat{\boldsymbol{\theta}}$ is obtained in order to proceed with the derivation of likelihood-based test statistics. We now suppress the dependence of $S(\boldsymbol{\theta}|\mathbf{y})$ on the sample $\mathbf{y}$ for notational convenience. Expressing $\mathcal{S}(\boldsymbol{\theta}_T)$ around $\hat{\boldsymbol{\theta}}$ by a first order Taylor expansion (justified by the consistency of the MLE) we get:

$$\mathcal{S}(\boldsymbol{\theta}_T) \approx \mathcal{S}(\hat{\boldsymbol{\theta}}) + (\boldsymbol{\theta}_T - \hat{\boldsymbol{\theta}})\frac{\partial^2 \log \mathcal{L}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}.$$

Note that this approximation is more accurate when the score function is approximately linear, and is exact when the score function is a linear function of $\boldsymbol{\theta}$. Hence, the more quadratic the log-likelihood is around $\hat{\boldsymbol{\theta}}$ the better the previous approximation, and we have:

$$(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_T) \approx I^{-1}(\hat{\boldsymbol{\theta}})\mathcal{S}(\boldsymbol{\theta}_T). \tag{1.3}$$

Hence the score statistic follows, according to the Central Limit Theorem, asymptotically a multivariate normal distribution with mean and variance as shown in (1.1) and (1.2). This implies on its turn that the MLE is asymptotically normally distributed with mean $\boldsymbol{\theta}_T$ and variance covariance matrix $I^{-1}(\hat{\boldsymbol{\theta}})$. Further, because the observed Fisher information matrix converges to the expected Fisher information matrix and the MLE converges to $\boldsymbol{\theta}_T$, the variance covariance matrix of the MLE can be replaced by $\mathcal{I}(\boldsymbol{\theta}_T)$, which is useful in constructing statistical tests. The aim of a statistical test is to get insight whether there is enough information to reject a statement about a possible true value of the parameter. We hypothesize that the value might be $\boldsymbol{\theta}_T = \boldsymbol{\theta}_0$ and seek the rejection of the statement from the statistical test. We start from the derivation of the likelihood ratio (LR) test. The LR statistic has the following form:

$$W = 2[\log \mathcal{L}(\hat{\boldsymbol{\theta}}) - \log \mathcal{L}(\boldsymbol{\theta}_0)],$$

To show the asymptotic distribution of the above statistic we use a Taylor expansion of $\mathcal{L}(\hat{\boldsymbol{\theta}})$ around $\boldsymbol{\theta}_0$. This gives:

$$\log \mathcal{L}(\hat{\boldsymbol{\theta}}) \approx \log \mathcal{L}(\boldsymbol{\theta}_0) + (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\mathcal{S}(\boldsymbol{\theta}_0) - 0.5(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)I(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$$

We can use (1.3) and plug in the above expression. In addition we use the fact that observed information at $\hat{\boldsymbol{\theta}}$ is approximately the observed information at $\boldsymbol{\theta}_0$ to get:

$$\log \mathcal{L}(\hat{\boldsymbol{\theta}}) \approx \log \mathcal{L}(\boldsymbol{\theta}_0) + \mathcal{S}(\boldsymbol{\theta}_0)I^{-1}(\boldsymbol{\theta}_0)\mathcal{S}(\boldsymbol{\theta}_0) - 0.5\mathcal{S}(\boldsymbol{\theta}_0)I^{-1}(\boldsymbol{\theta}_0)I(\boldsymbol{\theta}_0)I^{-1}(\boldsymbol{\theta}_0)\mathcal{S}(\boldsymbol{\theta}_0)$$

We use the latter formula to write down the expression for the likelihood ratio statistic:

$$W \approx \mathcal{S}(\boldsymbol{\theta}_0)I^{-1}(\boldsymbol{\theta}_0)\mathcal{S}(\boldsymbol{\theta}_0).$$

We know that the distribution of the score statistic is normal with mean zero and variance equal to the Fisher information, this implies that the distribution of the LR test is a chi-square with d degrees of freedom under the null hypothesis $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$. The second test derived from the likelihood is the Wald test:

$$W_w = (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\mathcal{I}(\hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0).$$

To prove the distribution of the Wald test under H_0, we will use the Taylor formula for $2[\log \mathcal{L}(\hat{\boldsymbol{\theta}}) - \log \mathcal{L}(\boldsymbol{\theta}_0)]$ around $\hat{\boldsymbol{\theta}}$, this gives:

$$W \approx -2[\log \mathcal{L}(\hat{\boldsymbol{\theta}}) + \mathcal{S}(\hat{\boldsymbol{\theta}})(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}) + 0.5(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}})^T \frac{\partial^2 \log \mathcal{L}(\hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}})^T - \log \mathcal{L}(\hat{\boldsymbol{\theta}})],$$

$$W \approx (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\mathcal{I}(\hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0),$$

therefore the asymptotic distribution of the Wald statistic under H_0 is the same as that of the likelihood ratio statistic. Further, from the asymptotic distribution of the score statistic we derive the score test:

$$\mathcal{W}_s = \frac{\mathcal{S}(\boldsymbol{\theta}_0)}{\sqrt{\mathcal{I}(\boldsymbol{\theta}_0)}},$$

which has asymptotically a standard normal distribution under H_0. Equivocally,

$\mathcal{W}_s^2$ has a chi-square distribution with d degrees of freedom under H_0.

The above discussion focussed on the estimation of parameters and hypothesis tests using the likelihood of independent identically distributed observations. We now turn to the correlated case.

1.3 Likelihood inference and random effects models

Correlated data arise when sampling is performed within clusters, which have their own specific characteristics. Clusters might be hospitals in a clinical trial, herds of animals, patients who are followed up over time. In Diggle *et al.* (2002) three types of models for correlated data are discussed: marginal models, random effects models and transition models. In this thesis we are concerned with random effects models which exploit the likelihood as the basis for estimation and inference.

The class of random effects models assumes that the correlation between measurements is due to unobserved latent variables (random effects), which are responsible for heterogeneity between subjects not explained by the fixed effects. In other words, having observed all the qualities of subjects under the study, we still cannot explain all characteristics pertaining to a cluster by these covariates. These unobserved features, are the basis for the correlation between measurements within a given cluster, in addition to the observed covariates.

The latent variables in a random effects model are assumed have a distribution and are therefore coined as random. In contrast, the fixed effects are parameters in the model, which are believed to have a unique value in the non-Bayesian framework. In a random effects model one assumes that the observed data are generated in two steps. First, random effects $\mathbf{v}$ are sampled from the assumed distribution $f_{\boldsymbol{\lambda}}(\mathbf{v})$. In a second step, given realized values of the random effects $\mathbf{v}^*$, the observed data are sampled from $f_{\boldsymbol{\beta},\boldsymbol{\lambda}}(\mathbf{y}|\mathbf{v}^*)$. Therefore, given the value of the latent variable, the probability to observe the data is given by $\mathcal{L}(\boldsymbol{\beta},\boldsymbol{\lambda}|\mathbf{y}_{i\cdot})|_{\mathbf{v}_i=\mathbf{v}_i^*} = \prod_{j=1}^{n_i} f_{\boldsymbol{\beta},\boldsymbol{\lambda}}(\mathbf{y}_{ij}|\mathbf{v}_i = \mathbf{v}_i^*)$. Of course, we can obtain the same observed data under different values of latent variables $\mathbf{v}_i$. Therefore, the total likelihood to observe the data $\mathbf{y}_i$ is the sum of all conditional likelihoods for every different value of latent variables, weighted by the probability to actually have this value of the latent variable. This reasoning leads to the marginal likelihood, which is a standard objective function to estimate the parameters of a random effects model:

$$\mathcal{L}_M(\boldsymbol{\beta}, \boldsymbol{\lambda}|\mathbf{y}) = \int \dots \int \prod_{i=1}^{N} \prod_{j=1}^{n_i} f_{\boldsymbol{\beta},\boldsymbol{\lambda}}(y_{ij}|\mathbf{v}_i) f_{\boldsymbol{\lambda}}(\mathbf{v}_i) d\mathbf{v}_1 \dots d\mathbf{v}_N, \tag{1.4}$$

with N being a number of clusters and n_i number of observations in the $i-th$ cluster. To compute the marginal likelihood one needs to solve the above integral, which often does not have a closed form solutions and numerical methods need to be invoked. Observe that when N increases, the number of latent variables increases, but the number of parameters in the marginal likelihood remains the same. This allows the application of standard results of likelihood theory developed for independent observations, except for some specific designs. However, the marginal likelihood does not contain any information about the individual random effects. Inference on the individual random is based on empirical Bayes (EB) estimation. In a Bayesian approach we treat all parameters as random variables. Based on this assumption the joint likelihood of all parameters assuming that they are independent (where appropriate or necessary) needs to be established. Hence, the joint likelihood is:

$$f(\mathbf{y}, \boldsymbol{\beta}, \mathbf{v}, \boldsymbol{\lambda}) = f(\mathbf{y}|\boldsymbol{\beta}, \mathbf{v}, \boldsymbol{\lambda}) f(\boldsymbol{\beta}) f(\mathbf{v}) f(\boldsymbol{\lambda}),$$

where $f(\boldsymbol{\beta})$, $f(\mathbf{v})$, and $f(\boldsymbol{\lambda})$ are priors. In an empirical Bayes approach, we will use the estimated values of parameters $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$ from a maximum likelihood procedure. Therefore the MLE value of these parameters is assumed to occur with probability one, while other values have probability zero. This allows us to write down the joint distribution of observed and latent data, given the estimated parameters, as follows:

$$f_{\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\lambda}}}(\mathbf{y}, \mathbf{v}) = f_{\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\lambda}}}(\mathbf{y}|\mathbf{v}) f_{\hat{\boldsymbol{\lambda}}}(\mathbf{v}).$$

This forms the basis for the posterior calculation of the random effects:

$$f_{\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\lambda}}}(\mathbf{v}|\mathbf{y}) = \frac{f_{\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\lambda}}}(\mathbf{y}|\mathbf{v}) f_{\hat{\boldsymbol{\lambda}}}(\mathbf{v})}{f_{\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\lambda}}}(\mathbf{y})}.$$

Because the plug-in is used the method is called *empirical* Bayes. In a Bayesian context one says that the EB estimate is the mode of the posterior distribution (Molenberghs and Verbeke, 2005). For the meaning of the posterior distribution, we refer to the above explanation of the marginal likelihood. We noted that the same observed data might be obtained under different true values of the latent variable, therefore the posterior distribution evaluates what is the probability of the value of latent variable given the data has been obtained. Clearly random effects are treated as random parameters in this approach.

The class of mixed models, where the response variable is assumed to follow a Gaussian distribution, has been generalized to allow for Poisson, Binomial or other types of distributions. In these models the random components are typically assumed to follow (multivariate) normal distribution.

1.4 H-likelihood estimation and inference

Lee and Nelder (1996) proposed to use another approach to the estimation of correlated data model avoiding the computation of the marginal likelihood. In the same paper they coined a term hierarchical generalized linear models (HGLM), which can be handled by their approach. The HGLM model is an extension of a generalized linear model whereby the random effects can follow a conjugate Bayesian distribution. The proposed method uses the extended likelihood as the basis of estimation and inference, which is termed hierarchical likelihood or h-likelihood. This function is readily available from the definition of the model in contrast to the marginal likelihood and can be written as follows:

$$\mathcal{L}_E(\boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{v}|\mathbf{y}, \mathbf{v}) = \prod_{i=1}^{N} \prod_{j=1}^{n_i} f_{\boldsymbol{\beta},\boldsymbol{\lambda}}(y_{ij}|\mathbf{v}_i) f_{\boldsymbol{\lambda}}(\mathbf{v}_i). \tag{1.5}$$

Because we observe only $\mathbf{y}$ one can imagine that there could be different forms of $\mathbf{v}$ which later lead to the realization of the observed data. Suppose random effects $\mathbf{v}$ follow assumed distribution $f(\mathbf{v})$, and the response is generated from the distribution $f(\mathbf{y}|\eta)$. Now we can choose for different η. Suppose $\eta_1 = \mathbf{X}\boldsymbol{\beta} + \mathbf{v}$ and

$\eta_2 = \mathbf{X}\boldsymbol{\beta} \times \exp(\mathbf{v})$. Therefore we can have two likelihood functions:

$$\mathcal{L}_1(\boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{v}|\mathbf{y}, \mathbf{v}) = f(\mathbf{y}|\eta_1)f(\mathbf{v}),$$

and

$$\mathcal{L}_2(\boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{v}|\mathbf{y}, \mathbf{v}) = f(\mathbf{y}|\eta_2)f(\mathbf{v}),$$

these are two examples of extended likelihood and the question is which one shall we use for the statistical estimation and inference. By definition h-likelihood is one of the extended likelihoods when parameter $\mathbf{v}$ is canonical. The random effects $\mathbf{v}$ are canonical when they don't interfere with an estimation of fixed effects. This means that the marginal likelihood gives the same inference about fixed effects as this extended likelihood. Further, the weak canonical scale is defined the scale when random effects combine additively with fixed effects in a linear predictor as in $\mathcal{L}_1(\boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{v}|\mathbf{y}, \mathbf{v})$. This is also by definition h-likelihood even in cases when the canonical scale does not exist. Detailed discussion can be found in Lee *et al.* (2006).

The likelihood $\mathcal{L}_1(\boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{v}|\mathbf{y}, \mathbf{v})$ is called a hierarchical likelihood, as the random effects enter linearly in the linear predictor, and we use $h = \log(\mathcal{L}_1)$ to denote the log h-likelihood.

The expression of the extended likelihood contains three types of parameters: fixed $\boldsymbol{\beta}$ and random $\mathbf{v}$ parameters in the mean structure and variance components $\boldsymbol{\lambda}$. However, expression (1.5) cannot be used to estimate directly all parameters. In Lee and Nelder (1996) it is shown that an adjusted profile likelihood function is needed to estimate the variance components. The adjusted profile likelihood function is obtained from the extended likelihood by adding a correction as follows:

$$p_{\beta,v}(h) = h(\boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{v})|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}},\mathbf{v}=\hat{\mathbf{v}}} - 0.5 \log \left| \frac{D\left[h, (\boldsymbol{\beta}, \mathbf{v})\right]}{2\pi} \right|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}},\mathbf{v}=\hat{\mathbf{v}}}, \tag{1.6}$$

with

$$D[h, (\boldsymbol{\beta}, \mathbf{v})] = -\begin{pmatrix} \frac{\partial^2 h(\boldsymbol{\beta},\boldsymbol{\lambda},\mathbf{v})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T} & \frac{\partial^2 h(\boldsymbol{\beta},\boldsymbol{\lambda},\mathbf{v})}{\partial\boldsymbol{\beta}\partial\mathbf{v}^T} \\ \frac{\partial^2 h(\boldsymbol{\beta},\boldsymbol{\lambda},\mathbf{v})}{\partial\mathbf{v}\partial\boldsymbol{\beta}^T} & \frac{\partial^2 h(\boldsymbol{\beta},\boldsymbol{\lambda},\mathbf{v})}{\partial\mathbf{v}\partial\mathbf{v}^T} \end{pmatrix}.$$

The adjusted profile likelihood is maximized with respect to $\boldsymbol{\lambda}$ to obtain $\hat{\boldsymbol{\lambda}}$. The adjustment term is used in order to approximate a restricted marginal likelihood.

Maximization of unadjusted profile likelihood would provide solution equivalent to the extended likelihood for variance components. This extends the REML estimation and inference to the class of generalized linear mixed models (Noh and Lee, 2007). It can be shown that in case of linear mixed models this function provides exactly the restricted maximum likelihood. Given estimated variance components, one could use the extended likelihood to find the estimates of $\boldsymbol{\beta}$ and $\mathbf{v}$. This works well, however in some instances (especially binary data) one needs another adjusted profile likelihood to find the fixed effects $\boldsymbol{\beta}$, i.e.

$$p_v(h) = h(\boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{v})|_{\mathbf{v}=\hat{\mathbf{v}}} - 0.5 \log \left| \frac{D(h, \mathbf{v})}{2\pi} \right|_{\mathbf{v}=\hat{\mathbf{v}}}, \tag{1.7}$$

where $D(h, \mathbf{v}) = -\frac{\partial^2 h(\boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{v})}{\partial \mathbf{v} \partial \mathbf{v}^T}$. After estimates of $\boldsymbol{\lambda}$ and $\boldsymbol{\beta}$ are obtained one might use expression (1.5) to find the values of $\mathbf{v}$.

The class of HGLM contains models where the response is allowed to follow exponential family, while independent random components are assumed have a conjugate Bayesian distribution. For this class of models an extension of a Iterative Weighted Least Squares (IWLS) algorithm has been developed, which allows to estimate the fixed and random effects as well as variance components. This approach allows the analysis of complex designs of experiments and the inclusion of covariates in the variance component part or non-Gaussian random effects.

1.5 Aim of the thesis

This thesis focusses on the use of likelihood and h-likelihood methods to the estimation of complex models. In the first chapter a mixture model for cross-sectional data is scrutinized. A new approach to the mixture modeling is proposed where prior probabilities are fixed, while we change the dimension of the multinomial distribution. Further, the conditional distributions, the mean and the variance vary as a function of covariates by an application of splines.

The second chapter deals with ordinal data modelling using marginal likelihood techniques. Different random effects models for Bounded Outcome Scores (BOS) are compared in the study. BOS responses can have J-shaped or U-shaped distributions, which may require transformation in order to apply standard generalized random effects models.

The third chapter discusses the general theory of h-likelihood. It describes further the use of the package HGLMMM to estimate hierarchical generalized linear models with h-likelihood. These models were developed in Lee and Nelder (1996), Lee and Nelder (2001) and Noh and Lee (2007). Numerous examples are used for the illustration of the procedures.

In the fourth chapter, we present a further discussion of the h-likelihood approach to random effects models and contrast it to the marginal likelihood. Further, an extension of existing h-likelihood methods is presented to the estimation of hurdle models. These types of models allow handling of the zero-inflated count data.

In Chapter 5 h-likelihood methods are extended to allow for joint estimation of several HGLMs, which are linked by correlated Gaussian random effects. We also incorporate the use of Newton-Raphson procedures to estimate the correlation parameters, which is blended with existing h-likelihood algorithms.

References

Diggle, P., Heagerty, P., Liang, K.-Y., and Zeger, S. (2002). *Analysis of Longitudinal Data.* Oxford University Press, New York.

Lee, Y. and Nelder, J. (1996). Hierarchical generalized linear models (with discussion). *Journal of the Royal Statistical Society. Series B (Methodological)*, **58**, 619–678.

Lee, Y. and Nelder, J. (2001). Hierarchical generalised linear models: A synthesis of generalised linear models, random-effect models and structured dispersions. *Biometrika*, **88**, 987–1006.

Lee, Y., Nelder, J., and Pawitan, Y. (2006). *Generalized Linear Models with Random Effects.* Chapman & Hall / CRC, Boca Raton.

Molenberghs, G. and Verbeke, G. (2005). *Models for Discrete Longitudinal Data.* Springer, New York.

Noh, M. and Lee, Y. (2007). REML estimation for binary data in GLMMs. *Journal of Multivariate Analysis*, **98**, 896–915.

Pawitan, Y. (2001). *In All Likelihood: Statistical Modelling and Inference Using Likelihood.* Oxford University Press, New York.

Schervish, M. (1995). *Theory of Statistics.* Springer, New York.

Severini, T. (2000). *Likelihood Methods in Statistics.* Oxford University Press, New York.

Chapter
2

Finite Mixture Models with Fixed Weights Applied to Growth Data

Based on:
Molas M. and Lesaffre E. (2012). Finite Mixture Models with Fixed Weights Applied to Growth Data. Submitted to Biometrical Letters.

Abstract

To model cross-sectional growth data the LMS method (Cole and Green, 1992) is widely applied.

Here, we propose an alternative approach based on fitting finite mixture models (McLachlan and Peel, 2000) with K components which may perform better than the LMS method in case the data show an unusual distribution. Further, we explore fixing the weights of the mixture components in contrast to the standard approach where weights are estimated. Having fixed weights improves the speed of computation and the stability of the solution. In addition, fixing the weights provides almost as good fit as when the weights are estimated.

Our methodology combines Gaussian mixture modelling and spline smoothing

(Eilers and Marx, 1996, Ramsay,1988). The estimation of the parameters is based on the joint modeling of mean and dispersion proposed in Nelder and Pregibon (1987).

We illustrate the methodology on the Fourth Dutch Growth Study, which is a cross-sectional study that contains information on the growth of 7303 boys as a function of age. This information is used to construct centile curves so called growth curves, which describe the distribution of height as a smooth function of age. Further, we analyze simulated data showing a bimodal structure at some time point.

In its full generality, the approach allows for the replacement of the Gaussian components by any parametric density. Further, different components of the mixture can have a different probabilistic (multivariate) structure allowing for censoring and truncation.

2.1 Introduction

The Fourth Dutch Growth Study is a cross-sectional study which recorded several variables for a sample of boys and girls conducted in the Netherlands in 1997. Here we consider the height of 7303 boys, and we aimed to estimate centile curves of height as a function of age. The age ranges from 0.032 to 21.7 years, with a mean of 9.29 years. The median height is 145 cm with range 48.5 - 205.8 cm. Further details of the study can be found in van Buuren and Fredriks (2001).

The standard method for estimating growth curves is the LMS method (Cole and Green, 1992). This method transforms the data to normality, and models the median, the coefficient of variation and the skewness as a smooth function of covariates, e.g. age. It performs well in most situations. However, when the data exhibits a special structure, such as being bimodal, at some ages, and unimodal at other ages, the LMS method might not be optimal. Such data can arise when the total population splits up into subgroups that have different growth spurts.

Another approach useful in the presence of mixtures is available in the R package **gamlss.mx** (Rigby and Stasinopoulos, 2005). This package allows for a mixture of distributions to be fitted, where means and standard deviations might depend on age and the contributions of each mixture component are estimated.

We explore in this paper a simplified finite mixture modelling approach (McLachlan and Peel, 2000), where weights are fixed to be equal. This speeds up the computation time, and offers more stability of the solution. The disadvantage of the simplified computational approach is a less than optimal fit. The observed loss of fit is often minimal, and can even be avoided by adding some extra mixture components. Further, by the addition of splines we can allow for the smooth change of the density along the covariates, where means and variances of the component densities can be expressed as a non-linear function of covariates. Further, one can tune the approach with respect to the number of components, and the flexibility of splines in each part of the model. Decision about the final model is based on the Akaike Information Criterion (AIC).

In the next section we first review the estimation process of finite mixture models, using the standard case of Gaussian mixtures. Generalization to exponential family densities, multivariate distributions and censoring follows. In Section 2.3 we present an application of the modelling approach to the Fourth Dutch Growth data. A limited numerical study is performed in Section 2.4. Finally, we give some concluding remarks.

2.2 Mixture models

2.2.1 Mixtures of Gaussian Distributions

Throughout this section we will use indices: $k = 1, \ldots, K$ for the number of mixture components, $i = 1, \ldots, N$ for the number of observations, $p = 1, \ldots, P_\beta$ index is used for both fixed parameters in the mean structure and $p = 1, \ldots, P_\gamma$ for the parameters in the dispersion structure. A mixture model evaluated in data point y_i is given by:

$$g(y_i) = \sum_{k=1}^{K} w_k f_k(y_i), \tag{2.1}$$

with $f_k(y_i),\ (k = 1, \ldots, K)$ the mixture components.

In a Gaussian mixture model each $f_k(y_i)$ is assumed to be a normal density with a mean μ_k and variance $\phi_k \equiv \sigma_k^2$. Further, each mixture component $f_k(y_i)$ contributes to the total density via weight w_k. Suppose that the density described in (2.1) changes over a range of known factors (which can be continuous such as age, or discrete such as treatment). Denote by $\mathbf{z}_i$ a set of factors pertaining to observation y_i, whereby, the location parameters μ_k are a function of $\mathbf{x}_i$, and the variances ϕ_k are a function of $\mathbf{r}_i$, with $\mathbf{x}_i$ and $\mathbf{r}_i$ subsets of $\mathbf{z}_i$, then we obtain the following model:

$$g(y_i|\mathbf{x}, \mathbf{r}) = \sum_{k=1}^{K} w_k f_k(y_i|\mathbf{x}_i, \mathbf{r}_i), \tag{2.2}$$

where $g(y_i|\mathbf{x}_i, \mathbf{r}_i)$ is the distribution of the response y_i. We assume that the parameters in each of the K components are allowed to be distinct. This is of interest for the Fourth Dutch Growth Study as we wish to allow for a change of the distributional shape of the mixture along with covariates. The log-likelihood of all N subjects is

$$\log L(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K, \boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_K, w_1, \ldots, w_k; \mathbf{y}) = \sum_{i=1}^{N} \log \left[\sum_{k=1}^{K} w_k \cdot f_k(y_i|\mu_k(\mathbf{x}_i), \phi_k(\mathbf{r}_i)) \right], \tag{2.3}$$

We distinguish three types of the parameters in the above likelihood: (1) location parameters $\mu_k(\mathbf{x}_i) = \mathbf{x}_i^T \boldsymbol{\beta}_k$, (2) scale parameters $\phi_k(\mathbf{r}_i) = exp(\mathbf{r}_i^T \boldsymbol{\gamma}_k)$, and (3) weights w_k. The score equations with respect to $\boldsymbol{\beta}$ are expressed as follows:

$$\frac{\partial \log L}{\partial \beta_{kp}} = \sum_{i=1}^{N} \left(\frac{w_k \frac{\partial f_k(y_i|\mathbf{x}_i, \mathbf{r}_i)}{\partial \beta_{kp}}}{\sum_{k=1}^{K} w_k \cdot f_k(y_i|\mathbf{x}_i, \mathbf{r}_i)} \right) = \sum_{i=1}^{N} c_{ki} \frac{\partial \log(f_k(y_i|\mathbf{x}_i, \mathbf{r}_i))}{\partial \beta_{kp}}, \tag{2.4}$$

with $c_{ki} = \frac{w_k \cdot f_k(y_i|\mathbf{x}_i, \mathbf{r}_i)}{\sum_{k=1}^{K} w_k \cdot f_k(y_i|\mathbf{x}_i, \mathbf{r}_i)}$, and $p = 1, \ldots, P_\beta$. Same score equations are obtained for the parameters of the P_γ dimensional $\boldsymbol{\gamma}_k$ vectors with entries γ_{kp}, $p = 1, \ldots, P_\gamma$.

Similarly we can derive the score equations for weights w_k:

$$\frac{\partial \log L}{\partial w_k} = \sum_{i=1}^{N} c_{ki} \frac{\partial \log(w_k)}{\partial w_k} - \sum_{i=1}^{N} c_{Ki} \frac{\partial \log(w_k)}{\partial w_k}, \tag{2.5}$$

which are the score equations of a multinomial model where the parameters are w_k

given the estimated posterior weights c_{ki}. When weights do not depend on covariates the solution for w_k has a closed form, i.e.

$$\hat{w}_k = \frac{\sum_{i=1}^{N} c_{ki}}{N}. \tag{2.6}$$

However, when the weights are functions of covariates, iterative procedures are necessary.

We propose here to keep weights fixed and equal $w_k = w$. Then equations (2.5) do not need to be solved but we estimate only location and scale parameters. This speeds up the convergence, as no iteration for weights is needed and by fixing weights problems as an infinite likelihood (McLachlan and Peel, 2000)[pp. 94-97] are avoided. Thus fixing to equal weights, estimation is speeded up and numerical stability of the solution is improved. We refer to this approach as the “FMIX” approach.

For $f_k(y_i|\mathbf{x}_i, \mathbf{r}_i)$ a Gaussian probability density function:

$$f_k(y_i|\mathbf{x}_i, \mathbf{r}_i) = \mathcal{N}_k(\mu_k(\mathbf{x}_i), \phi_k(\mathbf{r}_i)), \tag{2.7}$$

the score equations become

$$\frac{\partial \log L}{\partial \beta_{kp}} = \sum_{i=1}^{N} c_{ki} \left(\frac{y_i - \mu_{ki}}{\phi_{ki}} \right) x_{ip} = \sum_{i=1}^{N} \tilde{c}_{ki}(y_i - \mu_{ki}) x_{ip}, \tag{2.8}$$

$p = 1, \ldots, P_\beta$, resembling score equations of a normal distribution with unequal variances. The solution is provided by the weighted least squares procedure:

$$\hat{\boldsymbol{\beta}}_k = (\mathbf{X}^T \widetilde{\mathbf{C}}_k \mathbf{X})^{-1} \mathbf{X}^T \widetilde{\mathbf{C}}_k \mathbf{y}, \tag{2.9}$$

$k = 1, \ldots, K$, with $\widetilde{\mathbf{C}}_k$ a diagonal matrix with entries $\tilde{c}_{ki} = \frac{c_{ki}}{\phi_{ki}}$.

For the variance structure parameters $\boldsymbol{\gamma}_k$, the general framework was described in Nelder and Pregibon (1987). The score equations have the following expressions:

$$\frac{\partial \log L}{\partial \gamma_{kp}} = \sum_{i=1}^{N} c_{ki} \left[-\frac{1}{2\phi_{ki}} + \frac{d_{ki}}{2\phi_{ki}^2} \right] \frac{\partial \phi_{ki}}{\partial \gamma_{kp}} = \sum_{i=1}^{N} \left[\frac{c_{ki}}{2} \frac{d_{ki} - \phi_{ki}}{\phi_{ki}^2} \right] \frac{\partial \phi_{ki}}{\partial \gamma_{kp}}, \tag{2.10}$$

$p = 1, \ldots, P_\gamma$, with d_{ki} the deviance residual, which for a normal distribution is equal to the squared residual $(y_i - \mu_{ki})^2$. Equation (2.10) is a score equation of a gamma generalized linear model (GLM), with a response d_{ki}, a prior weight $c_{ki}/2$, and a mean ϕ_{ki} linked to the covariates. The parameters can be estimated by Iterative Weighted Least Squares (IWLS), see also Nelder and Wedderburn (1972).

The above two (I)WLS procedures can be combined into an interchangeable IWLS algorithm to find estimates of $\boldsymbol{\beta}_k$ and $\boldsymbol{\gamma}_k$ for each k^{th} component of the mixture. In total the following estimation procedure is proposed:

1 Weights c_{ki} are computed given (starting) values of $\boldsymbol{\beta}_k(t-1)$ and $\boldsymbol{\gamma}_k(t-1)$

2 Each of K components is estimated by the above procedure yielding $\boldsymbol{\beta}_k(t)$ and $\boldsymbol{\gamma}_k(t)$

3 Iterate [1] and [2] until convergence

This approach can be proven to be equivalent to the EM-algorithm. Briefly, the E-step of the algorithm corresponds to finding the weights c_{ki}, while the M-step is solving equations (2.4) and (2.5) given the posterior weights c_{ki}. More details on the EM-algorithm in this setting can be found in McLachlan and Peel (2000)[pp. 48-51].

Upon convergence proper standard errors of each component location and scale parameters can be computed numerically using the hessian matrix of the likelihood evaluated at the maximum.

2.2.2 Mixtures of Exponential Family Distributions

We can easily extend the previous approach to mixtures of other distributions from the exponential family:

$$f(y_i) = \exp\left(\frac{y_i\theta_i - b(\theta_i)}{\phi_i} + c(y_i, \phi_i)\right). \tag{2.11}$$

In general, the score equations (2.4) become:

$$\frac{\partial \log L}{\partial \beta_{kp}} = \sum_{i=1}^{N} c_{ki} \left(\frac{y_i - \mu_{ki}}{\phi_{ki} V(\mu_{ki})} \right) \frac{\partial \mu_{ki}}{\partial \eta_{ki}} x_{ip}, \tag{2.12}$$

where $p = 1, \ldots, P_\beta$. These are the score equations of a standard GLM with a modified prior weight equal $\tilde{c}_{ki} = c_{ki}/\phi_{ki}$. Therefore a standard IWLS algorithm can be used for the estimation, see e.g. Lee *et al.* (2006) [pp. 85-87]. The estimation of the dispersion parameters $\boldsymbol{\gamma}_k$ requires the use of deviance residuals, corresponding to the distribution used as a component of the mixture (e.g. Poisson, binomial, ...). To find the estimates of dispersion parameters γ_{kp} with $p = 1, \ldots, P_\gamma$ the gamma distributed GLM is fitted as in Section 2.2.1.

2.3 Applications

2.3.1 Fourth Dutch Growth Study

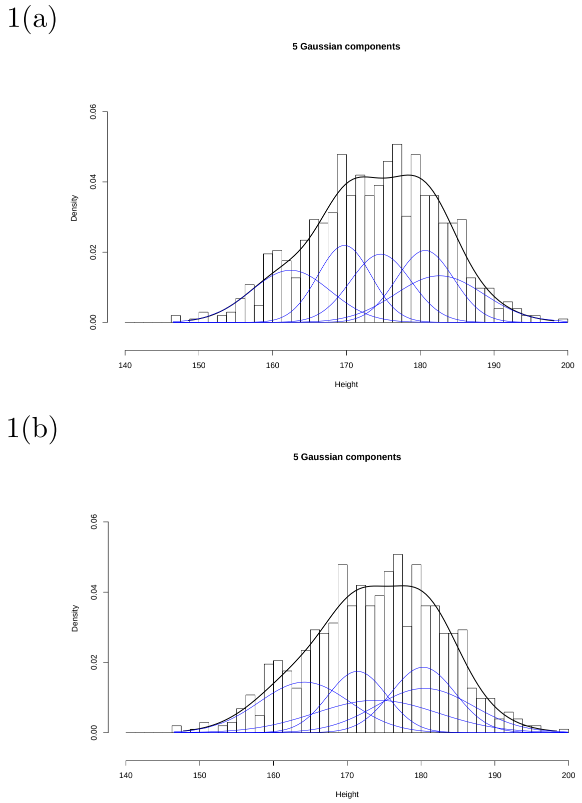
Here we present the analyses of the cross-sectional Fourth Dutch Growth Study. The objective is to model the centile curves of height as a function of age on the 7303 boys in the study.

FMIX approach

We first illustrate the FMIX approach on a part of the data i.e. the boys of age 14 to 16 and fit a flexible density distribution to this sub-sample. We used 5 Gaussian mixture components. Every component has its own mean and variance, the weights are 0.2. The assumed density for each height is then:

$$f(y_i|\mu_1 \ldots \mu_5, \sigma_1 \ldots \sigma_5) = \sum_{i=1}^{5} \frac{1}{5} \mathcal{N}(y_i|\mu_i, \sigma_i^2).$$

The result of the fit is shown in Figure 2.1(a). For comparison we present the fitted density when weights were estimated. The fitted density is presented in Figure 2.1(b). Comparing Figures 2.1(a) and 2.1(b) demonstrates the flexibility with which


1(a)

1(b)

Figure 2.1: Fourth Dutch Growth Study: estimation of the density for boys of age 14 to 16 with 1(a) fixed weights 1(b) estimated weights

mixtures are fitted.

Next, we analyzed the complete dataset to see the evolution of height with age. The relationship is clearly non-linear and therefore we used splines to capture this behaviour.

We started with restricted cubic splines (Harrell, 2001)[pp. 20-21] and cubic B-splines (Eilers and Marx, 1996), but the B-splines converge quicker and gave a better fit. In what follows we denote a B-spline of age with n degrees of freedom as $bs(age, n)$, with degrees of freedom (df) equal to the number of spline bases. The following structure was used in the final model:

$$\mu_k = bs(age, 10)\boldsymbol{\beta}_k,$$

$$\sigma_k^2 \equiv \phi_k = \exp(bs(age, 3)\gamma_k),$$

with the df determined by minimizing the AIC (smaller is better). However no claim is made that we determined the optimal df. In Table 2.1 we present the AIC of various FMIX models. We considered models with a different number of components and a different df in the mean and variance structure. The optimal model contained four mixture components with df=10 in the mean structure and df=3 in the variance structure. The result of the estimation (fitted centile curves) is presented in Figure 2.2.

Figure 2.3 presents fitted densities at different pre-selected ages of boys. To check the fit of the model worm plots (van Buuren and Fredriks, 2001) were created see Figure 2.4. A worm plot presents the expected quantiles and observed quantiles of a model for a range of covariate, augmented with confidence bands of the estimated quantiles.

The fit of the model seems appropriate, although there is still some misfit for boys younger than 35 days. Modeling at this young age is very difficult due to a fast increase in height for young boys. Further we computed the percentage of observations falling below every centile curve. This computed percentage is very close to the nominal centile i.e., below the 95%-ile lies about 95 percent of observations for the whole span of age. To fit the Gaussian mixture model with 10 bases in the mean, 3 basis functions in variance structure and 4 Gaussian components with fixed weights 2.8 minutes were necessary on a Pentium 2.33 Ghz core duo 2GB RAM.

LMS approach

A competitor to the FMIX approach, popular in the growth curves modelling is the LMS method of Cole and Green (1992). The LMS models were fitted using the R **gamlss** package, with the gamlss function with distribution 'BCCG' and an adequate number of dfs in each of the structures (mean,variance,skewness). We compared the AIC of LMS models in Table 2.1. The AIC is worse than that of FMIX models considered. The fitted centile curves with LMS approach are shown in Figure 2.2.

In the analysis of the Fourth Dutch Growth Study height of boys, the visual fit

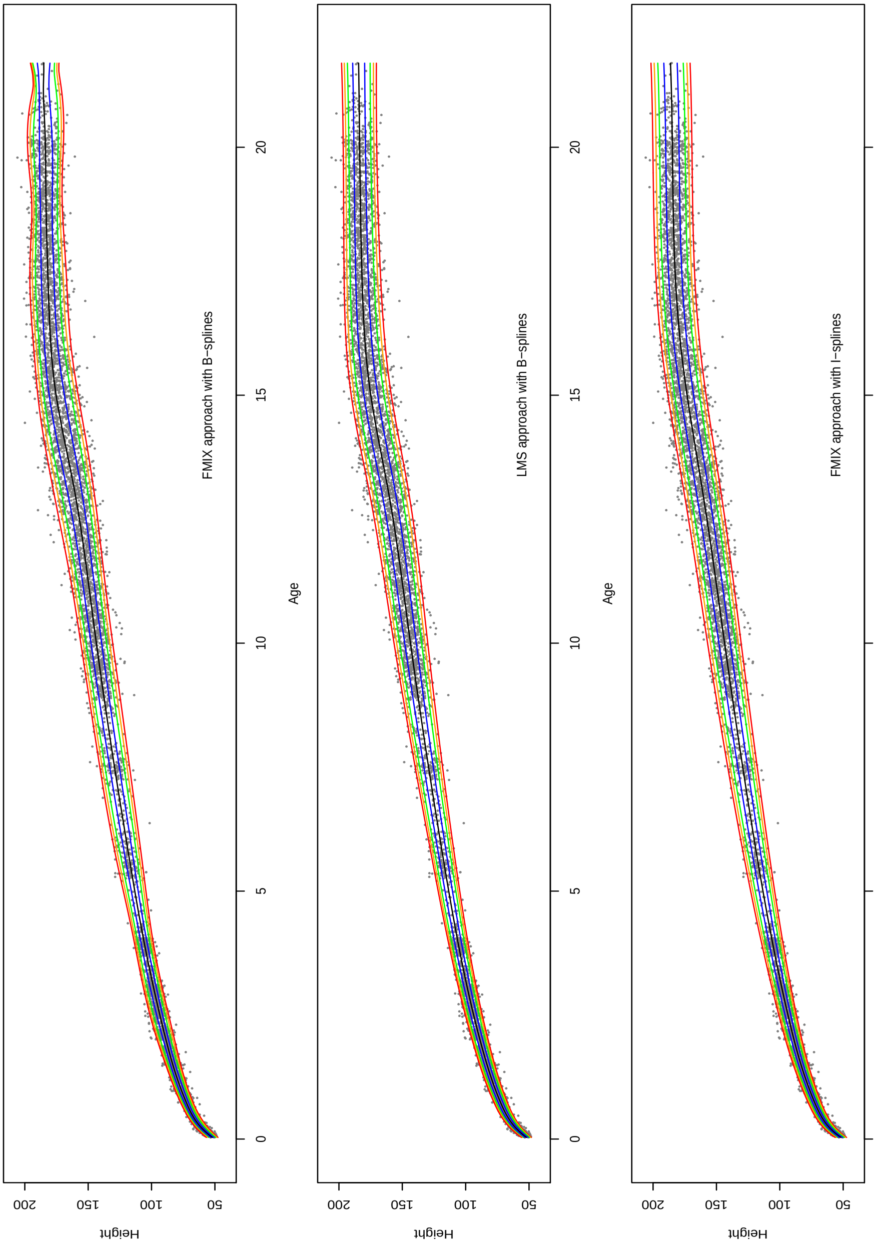


Figure 2.2: Fourth Dutch Growth Study: fitted centile curves with FMIX and LMS approaches

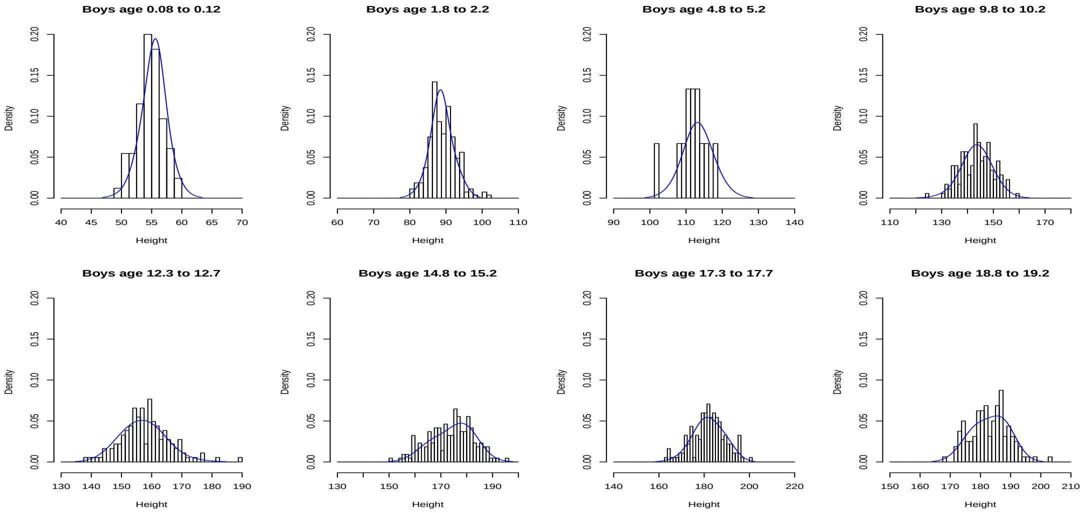


Figure 2.3: Fourth Dutch Growth Study: fitted densities at pre-selected boys ages

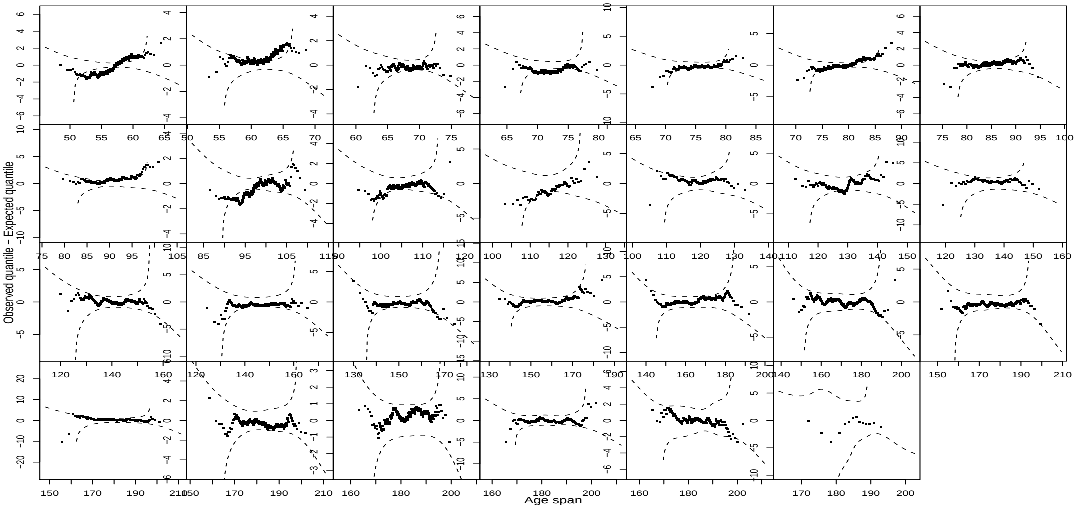


Figure 2.4: Fourth Dutch Growth Study: worm plots

Model	AIC (lower is better)
M10V3K4E	44564.9
M10V5K7	44601.9
M10V5K5	44602.3
M10V3K5	44582.3
M10V3K4	44574.1
M10V3K3	44579.1
LMS-M10V3S1	44633.7
LMS-M10V5S1	44619.3

Table 2.1: Fourth Dutch Growth Study: AIC of different models

We denote the number of dfs in the mean structure and variance structure together with a number of mixture components as follows: M10V5K7E - is a mixture model with 7 mixture components (K), 5 df of a B-spline in variance structure (V) and 10 df B-spline in the mean structure (M); E denotes that the weights of a mixture are estimated; LMS-M10V3S1 - denotes an LMS model with 10 degrees of freedom in the mean structure, 3 degrees of freedom in the variance structure and one parameter in the skewness part of the distribution

of the LMS model and the mixture approach with 4 mixture components with fixed weights did not differ much, see Figure 2.2. The LMS method required less than 5 seconds to perform the fit.

General mixture modelling

R package **gamlss.mx** allows to fit mixtures of distributions with estimated weights, which also can depend on covariates. However when trying to fit the model with splines in mean and variance structure the program failed to converge (10 and 3 dfs subsequently). We developed our own codes for fitting mixture models with estimated weights to analyze the Fourth Dutch Growth Study. We allowed for the weights to be estimated, but not to depend on covariates. The fit of the model measured by AIC was better (see Table 2.1), however visual fit of the centile curves to the data was not improved (results not shown). To fit the model 8.9 minutes were needed.

Penalized Gaussian Mixture approach

In Ghidey *et al.* (2004) the penalized Gaussian mixture model (PGMM) was introduced. This approach proposes to fix the means and variances of the components of the mixture, letting the weights of the individual distributions to be estimated. Additionally a penalty is imposed on the weights, reducing the difference between the weights estimates of the neighboring distributions. The implementation of this approach can be found in the R package **smoothSurv**. Applications in survival analysis can be found in Komarek *et al.* (2005).

This model can be applied to positive continuous data, with the exponent of the response as dependent variable. The shape of the distribution (modeled by weights) remains the same over the range of a covariates, while the mean and scale are estimated from the data and can vary with independent variables. An extension of the approach could allow the weights to depend on the covariates, thereby allowing the shape of distribution to change with a factor.

The PGMM as described in Komarek *et al.* (2005) was applied to the Fourth Dutch Growth Study assuming constant weights (results not presented). In case of growth data estimation time using this approach might be somewhat long. Further, it is unclear how to create the penalty term of the likelihood when weights change with a covariate.

2.3.2 Simulated example

In this section we present an artificially simulated dataset, however the example is motivated by the situation described in Muthen and Brown (2009). They describe a 4-class drug trial model, where the patients are assigned to either drug group or a placebo group. Further, in each drug group there are respondents to the treatment and non-respondents. Therefore, while information on the drug assignment is available in covariates, the information whether patient is a respondent or not is a latent factor and cannot be observed. Here we will consider one drug only, therefore we have respondents to the drug and non-respondents. For each group of respondents and non-respondents we simulate the different trajectory. Further we assume cross-sectional situation. Assume our response variable is a hypothet-

Model	AIC (lower the better)
LMS	20265
FMIX	17972
gamlssMX	17972

Table 2.2: Simulated data: AIC of different models

ical performance index measuring the treatment performance, while the covariate is the age of patient. We simulated dataset for 5000 individuals, with a uniform distribution of age between 0 and 40. The response originates from the following model:

$$\mu_1 = (2(40 - age)^3 + 3000)/(60 - age)^2 - 1.5,$$

for respondents and

$$\mu_2 = (2(40 - age)^3 - 3000)/(60 - age)^2,$$

for non-respondents. The dispersion parameters were set to one in both respondents and non-respondents. To this dataset we have fitted LMS, FMIX and **gamlss.mx** models. Figure 2.5 shows the fitted centile curves obtained from LMS, FMIX method with 2 components, and **gamlss.mx** with standard starting values. We plot there 2.5, 5, 10, 25, 35, 40, 45, 50, 55, 60, 65, 75, 90, 95, 97.5 percentile curves. We also computed the proportion of observations falling below the fitted centile curves. For the LMS method 43.3% of observations are below 40% centile curve, while in the FMIX approach it is 39.7%. 56% of observations are below the 60% centile curve for the LMS method, while 59.7% falls below in the FMIX model.

The FMIX method improves the fit of the LMS method by detecting the mixture of respondents and non-respondents. The fit of FMIX and **gamlss.mx** are comparable. This is due to the assumption that half the patients population are respondents and half do not respond to the drug. Therefore the correct weights are assumed in the FMIX approach. Table 2.2 presents the AIC values of the three approaches shown in this section. We conclude that FMIX and gamlssMX gave very similar results.

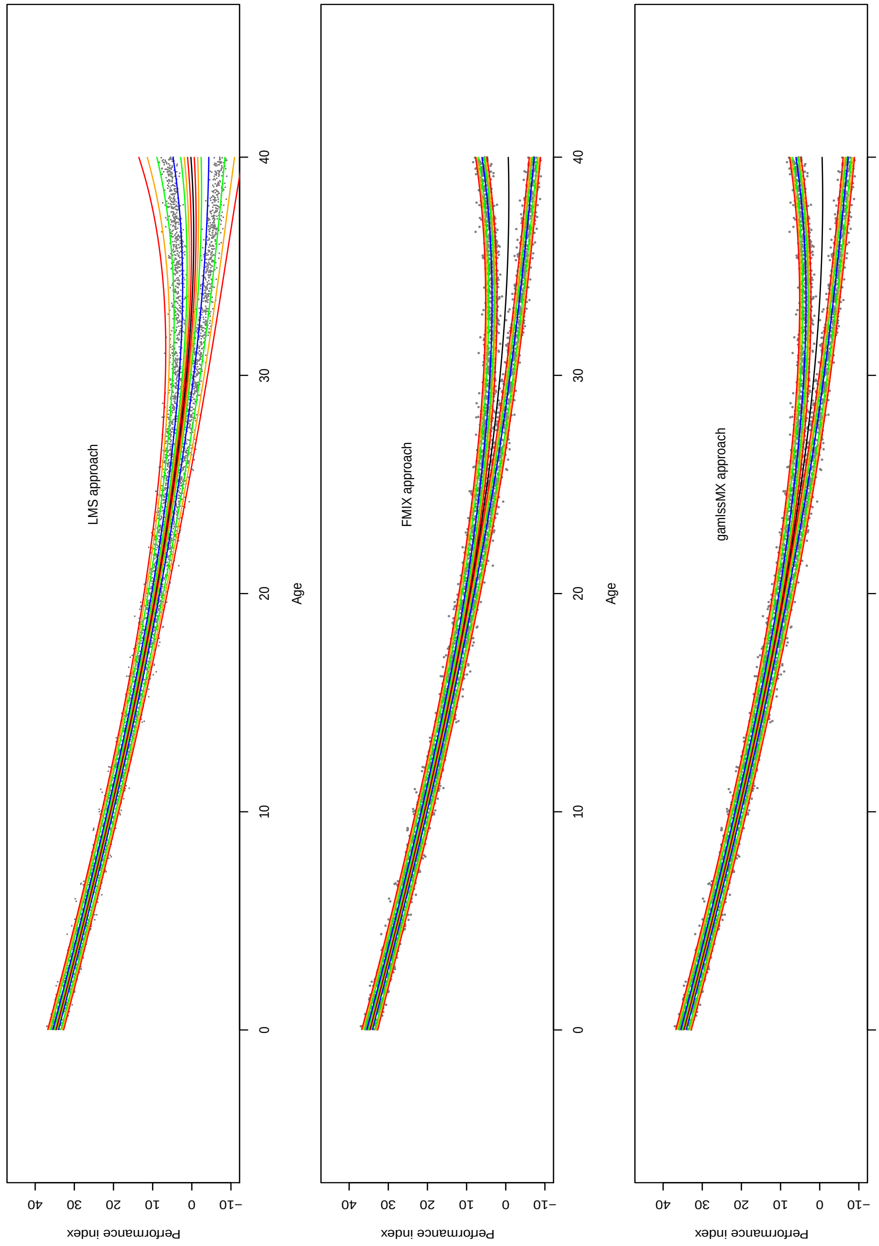


Figure 2.5: Simulated data fitted centile curves:LMS, FMIX, and gamlssMX approach

2.4 Simulation study

We performed a limited numerical comparison of the performance of the three methods: (1) FMIX (2) mixture models with estimated weights using our code and (3) **gamlss.mx** to fit mixture models with estimated weights. Approaches (2) and (3) are theoretically the same, however in practice they might perform differently.

We sampled the data from a mixture distribution. We considered three scenarios.

In Scenario 1, we sampled a data from a 75-25 mixture of normal distributions, with means -20 and 20, and standard deviations both 15. In Scenario 2 data were obtained by sampling from the 6-44-18-32 mixture of 4 Gaussian distributions with respective means of -20,10,10,20 and standard deviations all equal to 7. Finally in Scenario 3 we used 10-17.5-22.5-22.5-17.5-10 mixture of 6 Gaussian distributions with means -30,-20,-10,10,20,30 and standard deviations 3,5,7,7,5,3 respectively. In each scenario 6000 observations were sampled. No covariates were used in this simulation.

In each scenario models with varying number of components were fitted with the methods (1)-(3). In Scenario 1 we used 1-5,10,20 mixture components in fitted models, while in Scenario 2: 1-4,6,10 components. Finally in Scenario 3: 1-6,9,12,18,24 components. We computed the Kullback-Leibler (KL) divergence of the fit of the models against the true distribution. Furthermore we computed AIC of each model. These two measures were used for the comparison of the appropriateness of the model.

Under Scenario 1, the lowest KL distance was obtained for the mixture model with 2 components and estimated weights (method 2). This was the optimal model. However, there were still models close to the optimal. Model with 4 components of method 1 attained the ratio of KL of 1.11. Indicating that the Kullback-Leibler distance of this model was 1.11 times the KL distance of the optimal model. The lowest AIC was obtained by 2 components mixture model of method 2 and method 3.

Under Scenario 2, the optimal KL distance was obtained for the mixture model of 4 components with fixed weights (method 1). Further all mixture models with 6 components performed well (methods 1-3), as well as method 2 and 3 with 10 components. These had the ratio of the KL distance below 1.16. Note that models of method 2 converged at the boundary of parameter space when 3 or 4 components were used. Therefore for the models of method 2 computational problems occurred. Note that **gamlss.mx** model with two estimated components converged to maximum with log-likelihood value lower by approximately 500 points than the other models.

Under Scenario 3, the lowest KL distance was obtained for the model of 9 components of method 1. Equivalent performance was observed for the model with 6 components of method 2. Further, relatively good performed models with 9, 12, 18 components of method 2. The ratio was below 1.11. In this scenario the AIC of **gamlss.mx** models were lower than the AIC of FMIX or estimated weights models with the same number of mixture components. This was the case when 4-6,9,12,18,24 components were used for the estimation.

In summary, by using fixed weights (FMIX) we face less computational problems and we avoid obtaining infinite likelihood. It is seen that FMIX models perform as good as general mixture models when number of mixture components in the FMIX model is slightly larger than the number of components used to generate the data.

2.5 Conclusions

In this paper we proposed to use fixed weights in finite mixture models. We assume each component of a finite mixture model is parameterized by a separate set of parameters. Therefore, given prior weights (computed in the E-step of EM algorithm) every mixture component can be separately maximized. Mixture models of this type might be fitted using existing software for generalized linear models or generalized linear mixed models, which allow the inclusion of appropriate weights. The described estimation process is essentially the EM-alogrithm of Dempster *et al.* (1977).

The assumption of separate maximization of the components can be relaxed and the estimation can allow joint parameters over the mixture components. This could be of interest when one would like to keep the shape of distribution the same over the range of the covariates, and vary its mean only.

We used B-splines to model the non-linear distributions, however one could be interested in a monotone centile curves. This can be achieved by using the I-splines of Ramsay (1988) together with some reformulation of the likelihood. Monotonic centile curves are shown on the Figure 2.2. However, the reformulation of the maximization problem required the general Newton-Raphson algorithm to be used instead of interchangeable (I)WLS computational approach.

In comparison to the estimation weights approach one increases the speed of computations and stability of the estimation by fixing the weights of the mixture, whereby the total fit of the model does not deteriorate much. This was the case in the Fourth Dutch Growth Study analysis with 4 Gaussian mixture components.

Acknowledgments

We would like to thank prof. Paul Eilers for extensive consultation during preparation of this paper.

References

Cole, T. J. and Green, P. J. (1992). Smoothing reference centile curves: The LMS method and penalized likelihood. *Statistics in Medicine*, **11**, 1305–1319.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society Series B*, **39**, 1–38.

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

Ghidey, W., Lesaffre, E., and Eilers, P. (2004). Smooth random effects distribution in a linear mixed model. *Biometrics*, **60**, 945–953.

Harrell, F. E. (2001). *Regression Modelling Strategies.* Springer-Verlag, New York.

Komarek, A., Lesaffre, E., and Hilton, J. F. (2005). Accelerated failure time model for arbitrarily censored data with smoothed error distribution. *Journal of Computational and Graphical Statistics*, **14**, 726–745.

Lee, Y., Nelder, J. A., and Pawitan, Y. (2006). *Generalized Linear Models with Random Effects*. Chapman & Hall / CRC, Boca Raton.

McLachlan, G. J. and Peel, D. (2000). *Finite Mixture Models*. John Wiley and Sons, New York.

Muthen, B. and Brown, H. C. (2009). Estimating drug effects in the presence of placebo response: Casual inference using growth mixture modelling. *Statistics in Medicine*, **28**, 3363–3385.

Nelder, J. A. and Pregibon, D. (1987). An extended quasi-likelihood function. *Biometrika*, **74**, 221–232.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of Royal Statistical Society A*, **135**, 370–384.

Ramsay, J. O. (1988). Monotone regression splines in action. *Statistical Science*, **3**, 425–461.

Rigby, R. A. and Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape. *Applied Statistics*, **54**, 507–554.

van Buuren, S. and Fredriks, M. (2001). Worm plot: a simple diagnostic device for modelling growth reference curves. *Statistics in Medicine*, **20**, 1259–1277.

Chapter
3 A Comparison of Three Random Effects Approaches to Analyze Repeated Bounded Outcome Scores with an Application in a Stroke Revalidation Study

Based on:
Molas M. and Lesaffre E. (2008). A comparison of three random effects approaches to analyze repeated bounded outcome scores with an application in a stroke revalidation study. *Statistics in Medicine* **27** pp. 6612-6633.

Abstract

Discrete bounded outcome scores (BOS), i.e. discrete measurements that are restricted on a finite interval, often occur in practice. Examples are compliance measures, quality of life measures, etc. In this paper we examine three related random effects approaches to analyze longitudinal studies with a BOS as response: (1) a linear mixed effects model applied to a logistic transformed modified BOS; (2) a

model assuming that the discrete BOS is a coarsened version of a latent random variable, which after a logistic-normal transformation, satisfies a linear mixed effects model and (3) a random effects probit model. We consider also the extension whereby the variability of the BOS is allowed to depend on covariates. The methods are contrasted using a simulation study and on a longitudinal project which documents stroke rehabilitation in four European countries using measures of motor and functional recovery.

3.1 Introduction

A bounded outcome score is a measurement taking values on a finite interval. In practice a BOS is often, although not necessarily, discrete. Examples are (a) the proportion of days that patients correctly take their drug in compliance research; (b) the Barthel-index which is an Activity on Daily Living scale ranging from 0 (death or completely immobilized) to 100 (able to perform all daily activities independent) with jumps of 5; (c) visual analogue scales recorded in a discrete manner and (d) Likert scale measurements in social sciences. Here we will assume that the BOS is a numeric measurement taking values in the unit interval and that we are interested in modeling its distribution. The problem with a BOS is that its distribution can take a variety of shapes, from unimodal to J- and U-shaped leading to non-standard statistical approaches in general.

The CERISE study is used as our motivating example. This is a comparative parallel longitudinal study involving patients who experienced a stroke and revalidating in 4 European centers. Patients are followed up for 6 months, and measurements were taken at maximally 5 occasions. Specifically, we analyze the evolution of the Rivermead Motor Assessment Arm (RMAA) over time. The response is an index ranging from 0 up to 15. This score is obtained by summing up the number of positively accomplished motor tasks by the patient and is likely to have U-shaped (J-shaped) distribution. We consider patients from two centers in Nottingham (UK) and Herzogenaurach (DE), and we are interested whether recovery patterns over time differ between the centers. These data were analyzed previously by Wit *et al.* (2007) using random effect ordinal logistic model with a response categorized into 5 classes. The classification of the original response score is somewhat arbitrary, and may result in an efficiency loss. Other ways of analyzing a BOS in the univariate setting are mentioned in Lesaffre *et al.* (2007).

We contrast in this paper three random effects approaches to analyze a longitudinal study with a discrete BOS as outcome. In the first approach the BOS is modified such that it does not attain the boundary values of 0 and 1 and then a linear mixed effects model is applied on the logistically transformed modified BOS. The second approach was suggested by Lesaffre *et al.* (2007) for the univariate case. They assumed that the discrete BOS is a coarsened version of a latent continuous BOS with a logistic-normal distribution and fixed cut points. When varying the latent mean and variance a variety of distributional shapes appears. Hence in the second approach, we assume that the latent continuous BOS is logistic-normally distributed conditional on random effects (random intercept, random slope, etc) which are assumed to have a multivariate normal distribution. This model will be referred to as the random effects coarsened (CO) model. Finally, we consider the random effects ordinal probit (OP) model. This model assumes the response to have an ordinal character. However, the OP model can also be viewed as a discrete realization of a latent continuous BOS with a logistic-normal distribution but now with unknown cutpoints. For the three models we allow the measurement error variance of the (latent) BOS to depend on covariates.

The motivating longitudinal example is introduced in Section 3.2. Section 3.4 describes the three approaches and their extension whereby the residual variance is allowed to depend on covariates. Using a simulation study, we compare in Section 3.4 their performance (P(type I error), power, bias and MSE) with respect to estimating a treatment effect in a randomized controlled clinical trial. The similarity in power of the CO and OP approach, illustrated for the CO and OP models in the univariate case in Lesaffre *et al.* (2007), also appears to be case here. Therefore we show in the appendix that the power of the CO and OP model must be approximately the same for a randomized clinical trial (RCT) in the univariate case. The analysis of the longitudinal data set is shown in Section 3.5. Concluding remarks are given in Section 3.6.

3.2 Motivating example

The CERISE study is a longitudinal study which monitored, between March 2002 and September 2004, 532 consecutively admitted stroke patients from four European rehabilitation centers in Leuven (Belgium), Nottingham (United Kingdom), Zurzach (Switzerland) and Herzogenaurach (Germany). We will restrict our analysis to the comparison of two centers: Nottingham (center 2, 135 patients) and Herzogenaurach (center 4, 135 patients). Only patients satisfying specific in- and exclusion criteria were recruited in the study in order to achieve some balance between the centers. However, despite this attempt there were still quite some differences in patient characteristics at baseline between the two centers.

Patients were examined on admission, at 2, 4 and 6 months after the onset of stroke and at discharge. Thus, the time points of examination varied somewhat between patients. Also, 24 patients dropped out prematurely from each center (18%). Motor and functional recovery after stroke was assessed with a variety test batteries and the Barthel index. Here we will examine the Rivermead Motor Assessment Arm (RMAA) score, which is a measure that can assume the values $0, 1, \ldots, 15$. The two centers adhere a different revalidation scheme for the patients. More details on this study can be found in Wit *et al.* (2005).

The key question was whether the different revalidation schemes result in a different outcome on motor (measured by the RMAA score) and functional recovery taking into account the case-mix at baseline. Further, it was of interest to know which conditions at baseline were predictive for a good outcome at discharge. Finally, the likely evolution of a stroke patient given his/her baseline conditions is useful to know for the clinicians. This information requires the distribution of the BOS in time given the condition of the patient at baseline.

In Figure 3.1 the evolution of RMAA over time in the two centers is shown. Clearly at each time point the distribution has a U-shape although its appearance changes over time. Observe that the graph has been simplified by lumping together the measurements of the same visit although they happened at possibly different study time points. However, in the analysis of Section 3.5 the true time points were used.

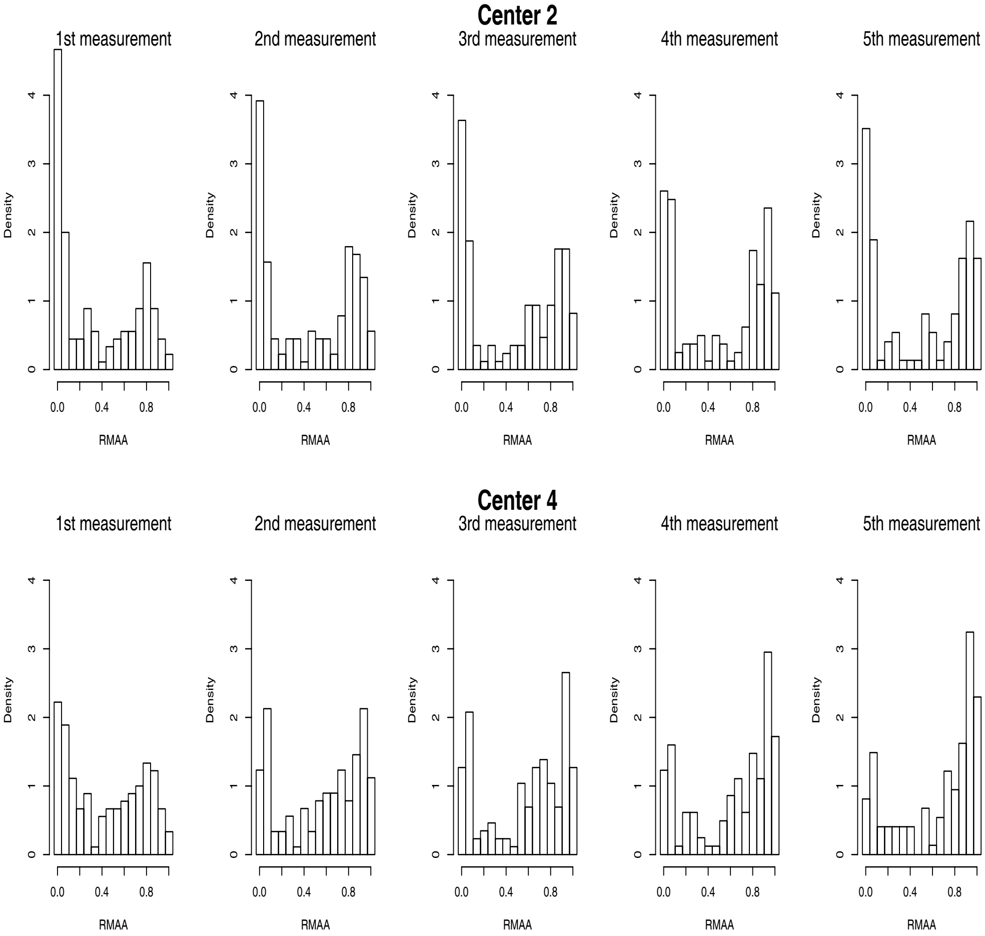


Figure 3.1: CERISE study: Evolution of Rivermead Motor Assessment Arm (RMAA) score in center Nottingham (center 2) and Herzogenaurach (center 4).

In Wit *et al.* (2005) RMAA was split up into a small number of classes and a random effects ordinal logistic regression model was used to establish the center effect taking into account the case-mix. Hereby some arbitrariness entered into the analysis when choosing the (number of) cut points to define the new outcome.

3.3 Three random effects approaches for analyzing longitudinal studies with a BOS response

Let Y_{ij} be the jth measurement of the BOS on the ith subject at the jth time point $(i = 1, \ldots, n; j = 1, \ldots, n_i)$ taking values k/m $(k = 0, \ldots, m)$ in [0,1].

3.3.1 A linear mixed effects model on a (modified) BOS

A linear mixed effects model based on Y_{ij} has the drawback that possibly some of the predicted outcomes will fall outside [0,1]. But, this problem is easily solved by modifying the Y_{ij} into $Y_{m,ij} = Y_{ij} + \varepsilon$ or $Y_{m,ij} = Y_{ij} - \varepsilon$, with ε a small positive value when $Y_{m,ij}$ is equal to 0 or 1, respectively. Therefore our first model, referred to as a logistically transformed linear mixed effects (LM) model assumes that

$$\mathrm{logit}(Y_{m,ij})|\boldsymbol{b}_i \sim \mathrm{N}(\mu_{ij}, \sigma^2), \tag{3.1}$$

with $\mathrm{logit}(p) = \log(p/(1-p))$ and

$$\mu_{ij} = \boldsymbol{\beta}^T \boldsymbol{x}_{ij} + \boldsymbol{b}_i^T \boldsymbol{w}_{ij}, \tag{3.2}$$

with $\boldsymbol{x}_{ij}$, $\boldsymbol{w}_{ij}$ a d-, p-dimensional covariate vector, respectively with $\boldsymbol{w}_{ij}$ a part of $\boldsymbol{x}_{ij}$ rendering it a 'well-formulated' model, see Morrell *et al.* (1997). Further, here $\boldsymbol{w}_{ij} = (1, t_{ij})^T$ with t_{ij} the time that the jth measurement of the ith subject was taken. Furthermore, for the random effects it is assumed that $\boldsymbol{b}_i \sim g(\boldsymbol{b}_i) \equiv \mathrm{N}_p(\boldsymbol{0}, \Sigma)$. A further extension is obtained by allowing σ to depend on covariates, see next subsection.

This approach offers a practical and computationally fast procedure. It has the drawback, though, that the original BOS is modified and that the choice of ε is subjective.

3.3.2 The coarsening model for longitudinal studies

Suppose that Y_{ij} is a coarsened version of a continuous BOS U_{ij} taking values in (0,1). More specifically, assume that $Y_{ij} = k/m \Leftrightarrow a_k \leq U_{ij} < a_{k+1}$, where the cutpoints a_k, $(k = 0, \ldots, (m+1))$ are known with $a_0 \equiv 0$ and $a_{m+1} \equiv 1$. Further, let

$$U_{ij}|\boldsymbol{b}_i \sim \mathrm{LN}(\mu_{ij}, \sigma^2), \tag{3.3}$$

where LN(μ_{ij},σ^2) is the logit-normal distribution, i.e.

$$Z_{ij} = \mathrm{logit}(U_{ij})|\boldsymbol{b}_i \sim \mathrm{N}(\mu_{ij}, \sigma^2), \tag{3.4}$$

where μ_{ij} is given by (3.2). On the logit scale the cutpoints a_k are transformed into $z_k = \mathrm{logit}(a_k)$, $(k = 0, \ldots, (m+1))$, with $z_0 = -\infty$ and $z_{m+1} = \infty$.

Further, conditional independence of U_{ij} (and Z_{ij}) given $\boldsymbol{b}_i$ is assumed such that for this model the (marginal) likelihood contribution for the ith subject given an observed vector $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{in_i})^T$ is given by

$$L(\boldsymbol{\beta}, \Sigma, \sigma^2|\boldsymbol{y}_i) = \int \left[\int_{z^l_{s(i1)}}^{z^u_{s(i1)}} f_1(z_{i1}|\boldsymbol{b}_i) dz_{i1} \times \cdots \times \int_{z^l_{s(in_i)}}^{z^u_{s(in_i)}} f_{n_i}(z_{in_i}|\boldsymbol{b}_i) dz_{in_i} \right] g(\boldsymbol{b}_i) d\boldsymbol{b}_i, \tag{3.5}$$

with $f_j(z_{ij}|\boldsymbol{b}_i)$ given by (3.4) and (3.2), $z^l_{s(ij)}$, $z^u_{s(ij)}$ the known lower, upper limit, respectively for the jth measurement of the ith subject. This model is called the random effects coarsened (CO) model. A further extension of the CO model is obtained by allowing the measurement error variance to depend on a possibly different set of covariates, i.e.

$$\log(\sigma_{ij}) = \boldsymbol{\gamma}^T \boldsymbol{x}^*_{ij}, \tag{3.6}$$

thereby replacing $L(\boldsymbol{\beta}, \Sigma, \sigma^2|\boldsymbol{y}_i)$ with $L(\boldsymbol{\beta}, \Sigma, \boldsymbol{\gamma}|\boldsymbol{y}_i)$. This model is an extension of the model suggested by Lesaffre *et al.* (2007) to a longitudinal setting.

An advantage of this approach in a univariate setting is that expressing the distribution of a BOS as a coarsened version of a continuous latent distribution model can help in planning a RCT with a BOS, see also Tsonaka *et al.* (2006) for an approach to calculate the sample size. No procedure for sample size calculation in the longitudinal setting is available now. This extension is likely to imply simulations in which the CO approach can play an important role. Further, in contrast to the previous approach, the original data are not modified. The choice of the cut points is somewhat subjective, though, but since the BOS is used as a numeric score the cut points must be in-between two possible values. In Lesaffre *et al.* (2007) a rounding mechanism was assumed, and this will be done also here. When the number of possible outcomes is high the actual choice of the cut point will be less important. Finally, some limited simulations in the univariate setting have shown that the CO approach is relatively robust with respect to the normality assumption.

3.3.3 The random effects ordinal probit model

The random effects ordinal (logit, probit) model constitutes a standard tool for analyzing repeated ordinal measures. Here we will concentrate on the random effects probit (OP) model because of its relationship with the previous two models. The OP model is defined as

$$P(Y_{ij} \leq k) = \Phi(\theta_{k+1} - \boldsymbol{\phi}^T \boldsymbol{x}_{ij} - \boldsymbol{b}_i^{*T} \boldsymbol{z}_{ij}), \tag{3.7}$$

with θ_k unknown cut points, that need to be estimated, $\boldsymbol{\phi}$ a vector of regression coefficients and $\boldsymbol{b}_i^*$ the random effects.

When the cut points on the latent scale are indeed z_k then it can be seen that the CO and OP model are related as follows (see also Lesaffre *et al.* (2007)): $\theta_k \equiv z_k/\sigma$, $\boldsymbol{\phi} \equiv \boldsymbol{\beta}/\sigma$ and $\boldsymbol{b}_i^* \equiv \boldsymbol{b}_i/\sigma$ when the measurement error variance does not depend on covariates. For σ depending on covariates, the CO model is equivalent to

$$P(Y_{ij} \leq k) = \Phi\left(\frac{z_{k+1} - \boldsymbol{\beta}^T \boldsymbol{x}_{ij} - \boldsymbol{b}_i^T \boldsymbol{z}_{ij}}{\sigma_{ij}}\right),$$

with σ_{ij} depending on covariates as in (3.6). On the other hand, the OP model is given by

$$P(Y_{ij} \leq k) = \Phi\left(\frac{\theta_{k+1} - \boldsymbol{\phi}^T \boldsymbol{x}_{ij} - \boldsymbol{b}_i^{*T} \boldsymbol{z}_{ij}}{\sigma_{ij}}\right), \tag{3.8}$$

with $\log(\sigma_{ij}) = \boldsymbol{\gamma}^{*T}\boldsymbol{x}^{**}$ whereby $\boldsymbol{x}^{*T} = (1, \boldsymbol{x}^{**T})$ and $\boldsymbol{\gamma}^T = (1, \boldsymbol{\gamma}^{*T})$. Hence, expression (3.8) becomes (3.7) for a baseline class corresponding to $\boldsymbol{x}^{**} = \mathbf{0}$. The extension of the OP model was suggested in Skrondal and S. (2004).

3.3.4 Research questions and remarks

For all above models the parameters can be estimated using a maximum likelihood procedure. For σ not depending on covariates, the analysis using the modified BOS (LM model) can be done with e.g. the SAS procedure MIXED. The CO and OP model can be fitted to the data using routines written in the SAS procedure NLMIXED. All routines can be extended relatively easy to allow for dependence of σ on covariates. The SAS routines can be obtained from the authors upon request.

A practitioner's appealing choice is probably the LM model, especially in explorative research. But the fact that the original data need to be modified is annoying and might lower its statistical performance. We question here what the impact is of modifying the original data and whether the subjective choice of ε matters.

Further, we question what the CO approach has to offer extra over the LM and OP approach. In the univariate setting (see Lesaffre *et al.* (2007)) it was observed that the statistical performance of the CO and OP approach are similar, despite the fact that in the OP approach m cut points need to be estimated.

Modeling the latent residual variance as a function of covariates can be important for the evaluation of the fixed effects, such as treatment, as exemplified in Lesaffre *et al.* (2007). The dependence of the variability on covariates might be modeled by adding extra random effects to the model. For instance, when we allow the latent measurement error to be different between two treatment groups, one could augment the random effects structure with a binary covariate indicating treatment. In the univariate setting this extension is equivalent to modeling σ as done above. However, in the longitudinal setting, introducing extra random effects will yield a different marginal covariance structure than when the measurement error variance is modeled directly and this may yield unexpected model assumptions.

The aim of this paper is not only to evaluate the impact of covariates on the outcome but also to model the distribution of the BOS in time. In this respect, a random effects ordinal model could be considered as not appropriate since it neglects the actual values of the BOS and assumes that the BOS has an ordinal nature. The estimated cut points will therefore be unrelated to the actual values of the BOS. When the latent score indeed has a normal distribution, then the estimated cut points will be on average equal to the true cut points as seen in the simulation study below. Thus, when the estimated cut points under the OP model differ 'considerably' from the assumed cut points under the CO model, it could indicate that the 'true' latent distribution of $\text{logit}(U_{ij})$ deviates from the normal distribution.

3.4 A simulation study

3.4.1 Set up

The simulation study was set up in order to compare the performance of the three random effects models. We focussed on detecting and estimating the treatment effect in a clinical trial setting. Therefore we evaluated the P(Type I error) and power in this respect. In addition, the bias and mean squared error (MSE) of the regression coefficients (and thus of treatment estimator) is compared between the three approaches. Observe, though, that the treatment effect in the LM and CO model (Δ) is defined differently from that in the OP model (Δ/σ) which makes a honest comparison with the OP approach difficult. For this reason the bias and MSE for the OP model was based on $\widehat{\Delta/\sigma} \cdot \sigma$, whenever possible.

First, for each of the random effects models the latent BOS U_{ij}, $(i = 1, \ldots, n; j = 1, \ldots, n_i)$ with $n = 200$, $n_i = 5$ was generated as follows:

$$\text{logit}(U_{ij}) = \mu_{ij} + b_{0i} + b_{1i} \cdot time_j + \epsilon_{ij}, \tag{3.9}$$

whereby $time_j \equiv t_{ij} = 0, 2, 3, 4, 6$ the time points where the BOS was measured, the measurement error $\epsilon_{ij} \sim \mathrm{N}(0, \sigma_{ij}^2)$ and the random effects $\boldsymbol{b}_i \sim \mathrm{N}_2(\mathbf{0}, \Sigma)$ with $vec(\Sigma) = (5.8, 0.15, 0.15, 0.5)$. For σ_{ij}, we have considered three cases (with the actual chosen value for the parameter):

$$\begin{aligned} log(\sigma_{ij}) &= \gamma_0\,(-0.3), && (3.10)\\ log(\sigma_{ij}) &= \gamma_0\,(-0.4) + \gamma_1\,(0.3) \cdot trt_i, && (3.11)\\ log(\sigma_{ij}) &= \gamma_0\,(0.8) + \gamma_1\,(0.4) \cdot trt_i + \gamma_2\,(-0.02) \cdot age_i. && (3.12) \end{aligned}$$

Further,

$$\mu_{ij} = \beta_0 + \beta_1 \cdot trt_i + \beta_2 \cdot time_j + \beta_3 \cdot trt_i \cdot time_j + \beta_4 \cdot age_i, \qquad (3.13)$$

with $trt_i = 0, 1$ representing two treatment groups each consisting of 100 patients. Age was randomly drawn from $\mathrm{N}(65, 9^2)$. As values for the regression coefficients, we have taken $\beta_0 = -2.5$, $\beta_1 = 0$ mimicking that we simulated from a RCT, $\beta_2 = 0.25$, $\beta_3 = 0$ when the null-hypothesis of no different evolution between the two treatment groups is true and $\beta_3 = -0.1$ in case the alternative hypothesis is true, and $\beta_4 = 0.05$. The parameter values are inspired by the results of the analysis of the CERISE data with a CO model.

In a second step the coarsened BOS Y_{ij} was created as follows: $Y_{ij} = k/m$ when $(k - 0.5)/m \leq U_{ij} < (k + 0.5)/m$, where $k = 1, \ldots, (m - 1)$. Further, $Y_{ij} = 0$ when $0 \leq U_{ij} < 0.5/m$ and $Y_{ij} = 1$ when $(m - 0.5)/m \leq U_{ij} < 1$. We use $m = 15$.

For the LM model, ε was taken 0.01 and 0.015. In addition, in the simplest setting we have investigated the performance of the LM model with ε equal to 0.001, 0.0001 and 0.00001.

Finally, we considered simulation under the correct CO model (Scenario 1) and when the latent error distribution deviates from the normal (Scenario 2). More specifically, we considered $\epsilon_{ij} \sim \frac{1}{\sqrt{3}} t_3$ to represent a symmetric distribution with heavier tails than the normal and ϵ_{ij} distributed according to a mixture of normals, i.e. $0.30\mathrm{N}(-1.21, 0.369) + 0.70\mathrm{N}(0.52, 0.369)$ to represent a skewed distribution. In both cases, the mean and variance were 0 and 1, respectively.

For all settings 1000 simulations were performed, except for the OP model with

σ depending on covariates, which requires an excessive amount of computing time. Indeed, to finalize the simulations for these models one would need about one year of computing time. Up to now we have performed only 500 simulations per setting.

3.4.2 Results

Table 3.1 shows the results for the parameters of the mean structure ($\boldsymbol{\beta}$) under Scenario 1 and with σ constant. The probability of type I error for estimating the treatment effect is roughly 0.05 for the CO and OP models under consideration, but the LM approach is somewhat conservative. The power is also the highest for the CO and OP models, about 0.06 to 0.08 higher than that of the LM model. The bias is much lower for the CO and OP models than for the LM models. However, this does not always translate in a lower MSE for the CO and OP models, but they are often comparable though in some cases the MSE of the LM model is much higher.

We repeated the scenario of Table 3.1 for LM models with a varying ε. In loose terms, we observed a worse performance of the LM terms (MSE, bias, power) when ε decreases to zero (results not shown).

Table 3.1: Scenario 1, constant σ_{ij}: P(type I Error), power, bias and MSE for mean structure

Model	Type I	Power	Bias(0)^a	Bias(A)^a	Bias(β_2)^a	MSE(0)^a	MSE(A)^a	MSE(β_2)^a
CO	0.049	0.640	-0.354	-1.227	0.916	1.894	1.952	1.072
LM1	0.042	0.570	-1.185	19.949	-53.214	1.338	1.824	3.532
LM2	0.043	0.576	-1.008	25.904	-68.761	1.157	1.886	5.323
OP	0.048	0.635	-0.404	-1.415	1.380	1.890	1.970	1.150
OPV2^b	0.048	0.630	-2.440	-1.177	3.407	1.871	2.065	1.160
OPV3^b	0.020	0.508	-2.289	-7.848	19.231	2.142	3.787	7.542
COV2	0.050	0.641	-0.323	-1.190	0.895	1.896	1.946	1.072
COV3	0.050	0.640	-0.330	-1.226	0.922	1.900	1.952	1.075
LM1V2	0.042	0.569	-1.185	19.949	-53.214	1.338	1.824	3.532
LM1V3	0.041	0.568	-1.344	54.993	-53.322	1.345	1.814	3.551
LM2V2	0.043	0.576	-1.008	25.904	-68.761	1.157	1.886	5.323
LM2V3	0.037	0.576	-1.115	68.885	-68.946	1.163	1.858	5.353

[a] values should be multiplied by 10^{-3}. [b] based on 500 simulations. C0 - coarsening model with constant variance, LM1 - linear mixed model on transformed outcomes modified by 0.01, LM2 - linear mixed model on transformed outcomes modified by 0.015, OP - ordinal probit model,OPV2 - ordinal probit model with variance dependent on treatment, OPV3 - ordinal probit model with variance dependent on treatment and age, COV2 - coarsening model with variance dependent on treatment, COV3 - coarsening model with variance dependent on treatment and age, LM1V2 - as LM1 but variance dependent on treatment, LM2V2 - as LM2 but variance dependent on treatment, LM1V3 - as LM1 but variance dependent on treatment and age, LM2V3 - as LM2 but variance dependent on treatment and age, Bias(0) - average of $\hat{\beta}_3$ - true (0) parameter value under H_0, Bias(A) - average of $\hat{\beta}_3$ - true parameter value under H_A, Bias(β_2) - average - true parameter value under H_0 for β_2, MSE(0) - mean squared error for models when $\beta_3 = 0$, MSE(A) - mean squared error for models when $\beta_3 = -0.1$, Type I - Probability of the type I error, MSE(β_2) - mean squared error for the parameter $\beta_2 = 0.25$ when data generated with $\beta_3 = 0$

For the CO and OP models, the P(type I error) for estimating the regression coefficients in expression of σ_{ij} ($\boldsymbol{\gamma}$) was roughly 0.05, while for the LM model it was between 0.10-0.15 (results not shown). Further, for all models the MSE of these estimators increased with increasing number of more covariates involved in σ_{ij}. Furthermore, it appears that for the ordinal probit model, modelling the covariance matrix of measurements error as a function of continuous covariate results highly unstable parameter estimates. A finding that is even more dominant for the simulation scenarios below.

Table 3.2 presents the results under Scenario 1 for the parameters of the mean structure when the latent score residual variance depends on treatment. Now for all models and for estimating the treatment effect P(type I error) was close to 0.05. The remaining results were similar as in the previous table, except that now the bias and MSE could not be calculated for the OP model which ignores the dependency on treatment. The reason is that this model estimates one variance parameter which does not correspond to either of the two residual variances, and hence it is not clear with which value the OP-estimated parameters must be multiplied to ensure comparability to the other estimators.

Finally, the bias of the regressor coefficients of the latent score residual variance increased when treatment was omitted in the model for σ_{ij}, while the MSE increased when age was included (results not shown).

Table 3.3 presents the results for the parameters of the mean structure under Scenario 2, when $\epsilon_{ij} \sim \frac{1}{\sqrt{3}} t_3$. Overall, similar patterns emerged as under Scenario 1, except that the OP model seems to be more vulnerable to this deviation of the assumption than the CO model. However, regarding the parameter estimators for the latent residual variance the P(type I error) was around 0.25 for all models (results not shown). In case ϵ_{ij} was generated from the mixture of two normal distributions (results not shown), the P(type I error) was around 0.04 for all models while the power was 0.60 for the CO and OP models and around 0.55 for the LM models. For the parameters of the residual variance, P(type I error) was around 0.025 for the CO models, while it was roughly 0.07-012 for the LM models.

3.4.3 Discussion of results

The P(Type I error) was closest to the nominal level of 0.05 for the CO and OP models, but was more variable for the LM model, i.e. conservative in some cases and anti-conservative in other cases. The power was always higher for the CO and OP models being roughly equivalent in performance. Further, the bias was less for the CO models and OP models in comparison to the LM models. Thus, modifying the data has had an effect. Furthermore, the MSE was sometimes lower for the LM model, but in the case when it was inflated it was much more so. Also, the choice of the ε did have a serious impact on the estimation of some regression coefficients (in bias and MSE). Finally, loosely speaking, the estimated cut points from the OP model were on average the chosen cut points by the simulation under Scenario 1.

Two conclusions emerge. Firstly, the statistical performance of the LM models is inferior to the other two models and how the BOS is modified can have a serious impact on some parts of the model. Secondly, the performance of the CO and OP models is similar, despite the fact that for the OP model 13 extra parameters have to be estimated here. This was noticed also in Lesaffre *et al.* (2007) for the univariate case. Therefore, in the Appendix we explored why the two approaches yield approximately the same power in the univariate case. For this we compared the score statistics of the treatment effect in a RCT with two groups. To this end, we employed the approach in Whitehead (1992).

Finally, it is important to mention that all simulations were done for $m = 15$. Limited simulations indicate that similar conclusions could be drawn for smaller values of m. However, the bias of the three models increase somewhat when m decreases.

Table 3.2: Scenario 1, σ_{ij} depending on treatment: P(type I Error), power, bias and MSE for mean structure

Model	Type I	Power	Bias(0)[a]	Bias(A)[a]	Bias(β_2)[a]	MSE(0)[a]	MSE(A)[a]	MSE(β_2)[a]
CO	0.051	0.668	-8.347	-6.965	5.199	2.153	1.938	1.057
LM1	0.056	0.570	-1.979	18.820	-53.621	1.442	1.678	3.551
LM2	0.052	0.570	-1.881	24.829	-69.055	1.238	1.744	5.345
OP[c]	0.054	0.667	–	–	–	–	–	–
OPV2[b]	0.057	0.607	-0.091	-0.760	0.653	2.081	2.005	1.231
OPV3[b]	0.017	0.462	-1.459	-3.048	7.066	2.408	3.057	6.325
COV2	0.051	0.606	-1.159	-1.105	1.415	2.088	1.887	0.993
COV3	0.050	0.608	-1.122	-1.165	1.394	2.090	1.881	0.994
LM1V2	0.057	0.572	-1.979	18.820	-53.621	1.442	1.678	3.551
LM1V3	0.051	0.571	-3.141	18.812	-53.455	1.393	1.675	3.520
LM2V2	0.053	0.569	-1.881	24.829	-69.055	1.238	1.744	5.345
LM2V3	0.047	0.564	-2.593	24.825	-68.903	1.187	1.742	5.317

[a] values should be multiplied by 10^{-3}. [b] based on 500 simulations. [c] bias and MSE cannot be calculated. For an explanation of the abbreviations, see Table 3.1.

Table 3.3: Scenario 2, t_3-distribution for measurement error and σ_{ij} constant: P(type I Error), power, bias and MSE for mean structure

Model	Type I	Power	Bias(0)[a]	Bias(A)[a]	Bias(β_2)[a]	MSE(0)[a]	MSE(A)[a]	MSE(β_2)[a]
CO	0.050	0.631	-0.250	1.280	3.203	1.801	1.745	0.947
LM1	0.046	0.575	-0.441	21.480	-53.427	1.249	1.728	3.485
LM2	0.044	0.585	-0.340	27.426	-68.989	1.064	1.820	5.297
OP	0.051	0.620	0.145	-9.800	32.867	2.249	2.334	2.620
OPV2[b]	0.070	0.694	-0.779	-14.614	35.867	2.224	2.544	3.078
OPV3[b]	0.029	0.543	2.386	-30.121	68.023	3.316	7.511	26.104
COV2	0.048	0.632	-0.391	0.907	3.206	1.802	1.749	0.945
COV3	0.044	0.630	-0.281	1.139	3.248	1.812	1.729	0.948
LM1V2	0.046	0.576	-0.441	21.480	-53.427	1.249	1.728	3.485
LM1V3	0.046	0.575	-0.525	21.573	-53.609	1.253	1.733	3.505
LM2V2	0.044	0.586	-0.340	27.426	-68.989	1.064	1.820	5.297
LM2V3	0.044	0.580	-0.414	27.531	-69.211	1.066	1.827	5.328

[a] values should be multiplied by 10^{-3}. [b] based on 500 simulations. For an explanation of the abbreviations, see Table 3.1.

3.5 Analysis of the CERISE longitudinal study

The longitudinal CERISE study was analyzed with the three random effects models. We preferred to use a likelihood approach because of its well-known robustness to at random dropout mechanisms and because of the relatively high percentage of patients that prematurely dropped out from the study (18%). However, we did not aim to model possible more complex dropout mechanisms.

The first objective of the analysis was to check whether the evolution in the two centers was different over time (in months) taking into account the baseline difference between centers (with center 2 corresponding to center=0). The variable of interest was therefore the interaction between center and time. The following covariates were included in the models to adjust for case-mix: gender (male=0, female=1), urinary incontinence (0=No, 1=Yes), swallowing problems (0=No, 1=Yes), presence of dysphasia which implies impairment of speech and of comprehension of speech (0=No, 1=Yes) and presence of dysarthria (0=No, 1=Yes) implying weakness or incoordination of the speech muscles.

In the analysis of the data by the linear mixed model we used the adjustment value $\epsilon = 0.01$ to modify Y_{ij}. For the analysis of the data by the CO model we assumed the following threshold values: $a_0 = 0, a_{16} = 1$ and $a_i = \frac{i-0.5}{15}$ for $i = 1 \ldots 15$. Further, in the three models we allowed the latent score residual variance to depend on center, age and time. But in order to show that modeling the variance function does have an effect on the parameter estimates of the mean structure, we first fitted the data with a constant variance, see Table 3.4. Likelihood ratio tests were used to assess the significance of these parameters. The parameter estimates of the models with the variance depending on the covariates are shown in Table 3.5.

Comparing the parameter estimates from Table 3.4 and Table 3.5 clearly show that modeling the residual variance does have an effect on the estimation of the parameters in the mean structure. It strikes us that modeling σ_{ij} is largely overlooked in the literature.

Table 3.4: Analysis of the CERISE data with the LM, CO and OP model assuming constant latent score residual variance.

	Linear Mixed Model			Coarsening Model			Ordinal Probit		
Parameter	Estimate	S.E.	P-value	Estimate	S.E.	P-value	Estimate[a]	S.E.	P-value
θ_0	-2.535	1.026	0.0141	-2.977	1.068	0.0057	-0.238 (-1.281)	1.320	0.8570
Center (Herzogenaurach)	0.821	0.276	0.0032	0.832	0.283	0.0035	0.980 (0.744)	0.351	0.0056
Time (months)	0.234	0.025	<0.0001	0.245	0.028	<0.0001	0.332 (0.252)	0.041	<0.0001
Center*Time	0.008	0.036	0.8232	0.020	0.040	0.6197	0.045 (0.034)	0.056	0.4195
Age (years)	0.038	0.014	0.0079	0.046	0.015	0.0026	0.057 (0.043)	0.019	0.0022
Gender (Female)	-0.467	0.285	0.1020	-0.581	0.295	0.0494	-0.729 (-0.553)	0.364	0.0464
Urinary Incontinence (Yes)	-1.125	0.318	0.0005	-1.211	0.329	0.0003	-1.412 (-1.071)	0.411	0.0007
Swallow (Yes)	-1.152	0.354	0.0013	-1.310	0.367	0.0004	-1.623 (-1.232)	0.458	0.0005
Dysphasia (Yes)	0.270	0.316	0.3935	0.264	0.328	0.4203	0.244 (0.185)	0.407	0.5486
Dysarthria (Yes)	-1.750	0.295	<0.0001	-1.758	0.305	<0.0001	-2.157 (-1.637)	0.384	<0.0001
Σ_{11}	3.999	0.391	<0.0001	4.112	0.433	<0.0001	6.159 (3.546)	0.785	<0.0001
Σ_{12}	0.059	0.040	0.1376	0.107	0.047	0.0221	0.283 (0.163)	0.082	0.0006
Σ_{22}	0.041	0.008	<0.0001	0.049	0.010	<0.0001	0.106 (0.061)	0.023	<0.0001
γ_0	-0.163	0.026	<0.0001	-0.276	0.032	<0.0001	-	-	-
-2logLike		*			4732.2			4609.3	
AIC		*			4760.2			4663.3	

*Likelihood and AIC are not comparable to those of the CO and OP model. [a]Within-parentheses estimates are calculated using the correspondence between the CO and OP model assuming that the CO model holds.

Further, in Table 3.4 it is shown that the parameter estimates of the LM and CO model are close to each other. For the OP model we report the estimates as obtained from PROC NLMIXED and the values calculated using the relationship between the CO and OP parameters as indicated in Section 3.3.3 (values in parentheses). The transformed OP estimates and the CO estimates are relatively close to each other, with the OP estimates most often a bit larger in absolute value. With respect to the AIC criterion the OP model seems to be preferred over the CO model. The reason for preferring the OP model lies primarily in the fact that the normal assumption for the Z_{ij} is perhaps not really satisfied. This was apparent from the estimated cut points by the OP approach (not shown), which did not correspond well to the assumed cut points of the CO model. The likelihood of the LM model (likelihood for continuous data) cannot be compared with the CO and OP likelihoods (of grouped data) and hence neither the AIC. Similar conclusions can be drawn for the estimates shown in Table 3.5.

With respect to the research questions specified in Section 3.2, we conclude that there is not enough evidence to claim that the evolution over time of the RMAA score in both centers is different, while correcting for covariates. The way the covariates influence the recovery of the patient on average is clear from the tables. Of particular interest was to see how center, age and time affect the residual variability. In this respect all models confirmed less variability in center 4, when the patient was older and towards the end of the revalidation period.

The models shown in Tables 3.4 and 3.5 were based on a selection of covariates from a larger set of possibly important covariates. Further, these models still contain some nonsignificant covariates. If the purpose was to construct a predictive model, some further pruning of the model might be envisaged. This needs further model fittings. But, the three models from Table 3.5 differ considerably in computing time. On a Pentium 4 (2.8GHz, 512 MB RAM) computer with Windows 2000 (SP 4) operating system, the LM model needed only 1 min to reach convergence, the CO model needed about 10 mins but the OP model needed about 2.5 hours. Therefore, from a practical point of view and taking into account the simulation

Table 3.5: Analysis of the CERISE data with the LM, CO and OP model, allowing the residual variance latent score to depend on center, age and time

	Linear Mixed Model			Coarsening Model			Ordinal Probit		
Parameter	Estimate	S.E.	P-value	Estimate	S.E.	P-value	Estimate^a	S.E.	P-value
θ_0	-2.060	1.038	0.0483	-2.353	1.096	0.0327	-0.228 (-0.480)	0.302	0.4512
Center (Herzogenaurach)	0.788	0.279	0.0051	0.822	0.291	0.0050	0.212(0.915)	0.095	0.0270
Time (months)	0.201	0.025	<0.0001	0.211	0.027	<0.0001	0.063(0.272)	0.018	0.0007
Center*Time	0.010	0.033	0.7563	0.016	0.035	0.6507	0.008(0.033)	0.011	0.4513
Age (years)	0.033	0.014	0.0242	0.038	0.015	0.0145	0.010(0.043)	0.005	0.0399
Gender (female)	-0.393	0.285	0.1694	-0.444	0.299	0.1381	-0.121(-0.522)	0.088	0.1701
Urinary Incontinence (yes)	-1.110	0.321	0.0006	-1.252	0.335	0.0002	-0.324(-1.398)	0.121	0.0081
Swallow (yes)	-1.199	0.357	0.0009	-1.344	0.375	0.0004	-0.371(-1.601)	0.138	0.0076
Dysphasia (yes)	0.210	0.319	0.5107	0.182	0.335	0.5873	0.030(0.129)	0.092	0.7460
Dysarthria (yes)	-1.685	0.297	<0.0001	-1.703	0.311	<0.0001	-0.457(-1.972)	0.141	0.0014
Σ_{11}	4.009	0.396	<0.0001	4.252	0.451	<0.0001	0.309(5.752)	0.156	0.0492
Σ_{12}	0.064	0.036	0.0736	0.088	0.042	0.0363	0.011(0.205)	0.007	0.0939
Σ_{22}	0.035	0.006	<0.0001	0.036	0.007	<0.0001	0.004(0.074)	0.002	0.0704
γ_0	1.220	0.193	<0.0001	1.462	0.241	<0.0001	-	-	-
γ_1 (center)	-0.383	0.053	<0.0001	-0.455	0.067	<0.0001	-0.436	0.070	<0.0001
γ_2 (age)	-0.015	0.003	<0.0001	-0.018	0.003	<0.0001	-0.015	0.003	<0.0001
γ_3 (time)	-0.081	0.016	<0.0001	-0.110	0.021	<0.0001	-0.123	0.022	<0.0001
-2logLike	*			4644.9			4531.9		
AIC	*			4678.9			4591.9		

*Likelihood and AIC are not comparable to those of the CO and OP model. ^{a}Within-parentheses estimates are calculated using the correspondence between the CO and OP model assuming that the CO model holds.

Table 3.6: Analysis of the CERISE data with the LM, CO and OP model, allowing the residual variance latent score to depend on center, age and time with three random effects

	Linear Mixed Model			Coarsening Model			Orinal Probit		
Parameter	Estimate	S.E.	P-value	Estimate	S.E.	P-value	Estimate	S.E.	P-value
θ_0	-2.888	1.039	0.0059	-3.106	1.104	0.0053	0.016(-1.513)	0.462	0.9727
Center (Herzogenaurach)	0.679	0.295	0.0219	0.765	0.309	0.0138	0.298(0.953)	0.158	0.0611
Time (months)	0.565	0.095	<0.0001	0.629	0.099	<0.0001	0.273(0.874)	0.102	0.0081
Center*Time	0.087	0.127	0.4955	0.052	0.130	0.6901	0.034(0.109)	0.061	0.5784
Age (years)	0.040	0.014	0.0066	0.043	0.015	0.0060	0.019(0.060)	0.009	0.0439
Gender (female)	-0.515	0.285	0.0724	-0.567	0.302	0.0616	-0.243(-0.776)	0.153	0.1133
UI (yes)	-1.162	0.318	0.0003	-1.340	0.336	0.0001	-0.537(-1.718)	0.228	0.0191
Swallow (yes)	-1.105	0.357	0.0021	-1.290	0.378	0.0007	-0.544(-1.739)	0.238	0.0233
Dysphasia (yes)	0.335	0.317	0.2922	0.342	0.338	0.3118	0.124(0.396)	0.149	0.4067
Dysarthria (yes)	-1.832	0.299	<0.0001	-1.893	0.316	<0.0001	-0.798(-2.551)	0.299	0.0081
Time^2	-0.052	0.013	<0.0001	-0.060	0.013	<0.0001	-0.025(-0.080)	0.010	0.0148
Center*Time^2	-0.009	0.017	0.5703	-0.004	0.017	0.8031	-0.003(-0.009)	0.008	0.7200
Σ_{11}	4.086	0.455	<0.0001	4.352	0.541	<0.0001	0.706(7.221)	0.477	0.1396
Σ_{12}	-0.171	0.160	0.2861	-0.103	0.178	0.5648	0.002(0.025)	0.033	0.9409
Σ_{22}	0.475	0.106	<0.0001	0.447	0.111	0.0001	0.096(0.985)	0.073	0.1901
Σ_{13}	0.020	0.020	0.3304	0.011	0.022	0.6310	<0.001(<0.001)	0.004	0.9988
Σ_{23}	-0.053	0.014	0.0001	-0.048	0.014	0.0006	-0.010(-0.106)	0.008	0.2007
Σ_{33}	0.006	0.002	0.0006	0.005	0.002	0.0029	0.001(0.012)	0.001	0.2098
γ_0	0.950	0.217	<0.0001	1.163	0.278	<0.0001	-	-	-
γ_1 (center)	-0.441	0.059	<0.0001	-0.508	0.076	<0.0001	-0.488	0.083	0.0000
γ_2 (age)	-0.014	0.003	<0.0001	-0.018	0.004	<0.0001	-0.012	0.004	0.0054
γ_3 (time)	-0.043	0.024	0.0736	-0.050	0.031	0.1019	-0.082	0.033	0.0130
-2logLike		*			4556.8			4444.1	
AIC		*			4600.8			4514.1	

*Likelihood and AIC are not comparable to those of the CO and OP model. ^{a}Within-parentheses estimates are calculated using the correspondence between the CO and OP model assuming that the CO model holds.

results (especially the results on bias and power), the CO model seems to offer the practitioner a good working tool for analyzing repeated BOSs.

One referee suggested further extensions of our models. In this respect we extended the models presented in the Table 3.5 by allowing the variance-covariance matrix Σ of random effects to differ between centers. These extended models were tested against the corresponding models in the Table 3.5 with a likelihood ratio test with 3 degrees of freedom. In all cases the results of the tests yield p-values higher than 0.05. Therefore, the additional flexibility in the random effects variance-covariance matrix was unnecessary. Further extensions of our model could consist of (a) allowing for the measurement error to be serially correlated, (b) extending the random effects part to a polynomial function or in general more flexible than the suggested linear evolution. Extension (a) above appears to involve quite complicated programming efforts in conjunction with the SAS Procedure NLMIXED and it is even not clear whether it is at all possible to perform the calculations. For extension (b) we were able to fit a quadratic random effects model. While this took only about 3 minutes computing time for the linear mixed model, 4 hours were needed for the CO model, and the OP model converged after 30 hours.

In Table 3.6 the parameter estimates are shown, demonstrating that the quadratic effect in time appears to be justified. However, the qualitative conclusions remain the same. Finally, we tried more complicated random effects structures such as adding a cubic term. Unfortunately, convergence could not be attained.

The better fitting of the OP model to the data suggests that the assumption of a logit-normal distribution in the CO model is not entirely appropriate. This is not too surprising since the logit-normal model represents a quite simple assumption and might be not sufficiently flexible to model complex data. Nevertheless it will be probably useful in many circumstances. Further, with some extra modeling the CO approach allows to relax the two essential model parts relatively easy, i.e. the logit link and the normal assumption. For instance a complementary log-log link function could be used or a general family of link functions. The normal assumption could e.g. be replaced by a mixture of normals as in e.g. Ghidey *et al.* (2004). Another possibility is to estimate the cut points but compatible with the observed scores. Hence, instead of assuming the rounding coarsening mechanism one could estimate $a_k, (k = 1, \ldots, m)$ with the restriction that $\hat{a}_k \leq k/m < \hat{a}_{k+1}$. Observe that the OP model also estimates the cut points but ignores the restriction.

The purpose of relaxing the model assumptions is to produce more accurate information of the distribution of the BOS in time and hence to make better predictions of the recovery of the stroke patients given his/her initial conditions. For instance, in Figure 3.2 we show the distribution of RMAA at three time points (at admission, i.e. taken as the median value of days after stroke onset that the patient was admitted to the center, month 3 and month 6) for a male 40.4 years old patient with no urinary incontinence, no swallowing problems, no dysphasia and no dysarthria at baseline treated either in Nottingham (center 2) or in Herzogenaurach (center 4), based on the fitted CO model. Observe that such a plot can not be made with the OP approach. The probability of observing a score for RMAA higher than $0.8{\times}15 = 9$ is, for center 2 equal to 0.19, 0.26 and 0.36 for the three time points and for center 4 equal to 0.29, 0.38 and 0.50, respectively.

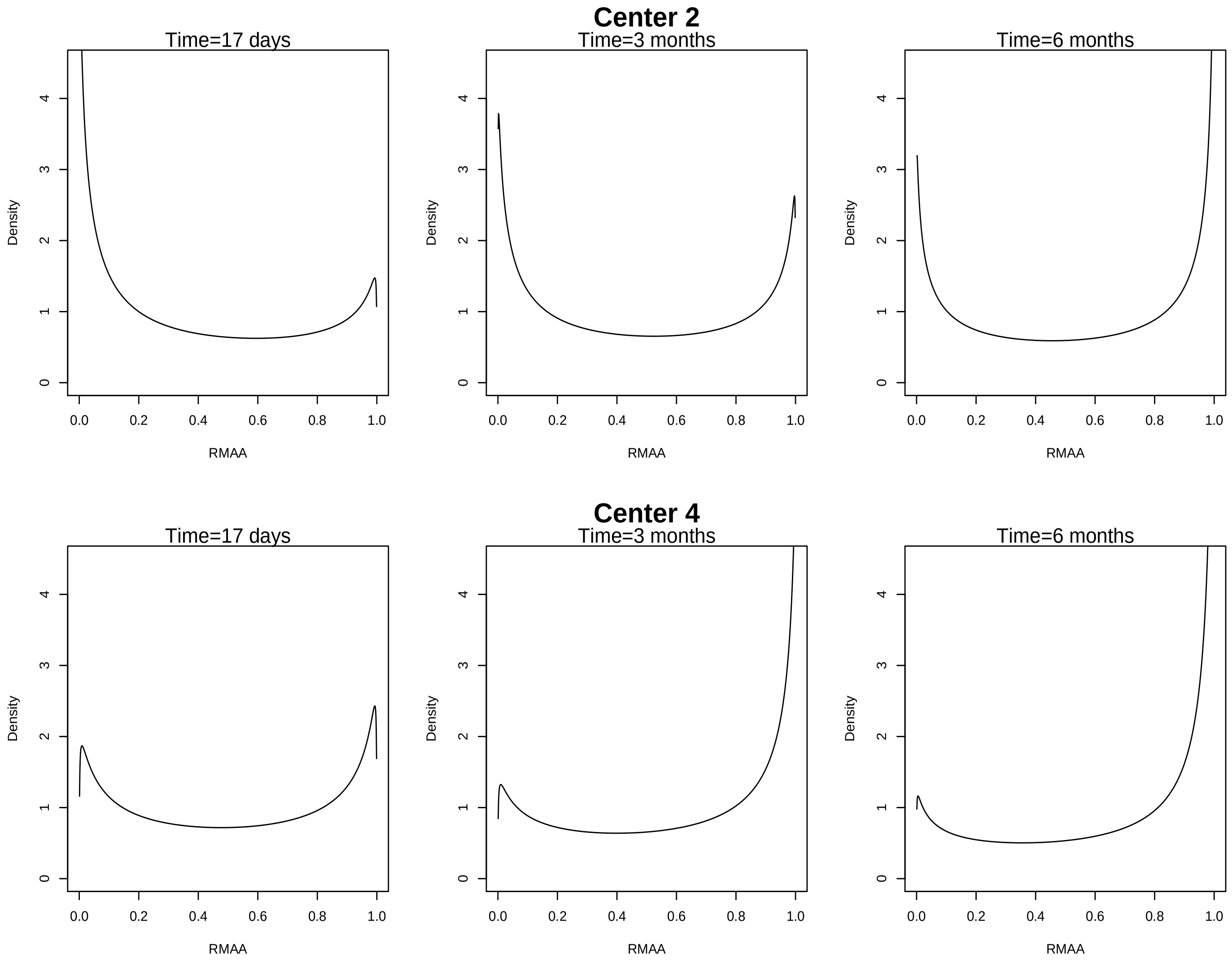


Figure 3.2: CO model: Fitted distribution of RMAA for a male patient (40.4 years old at the baseline) in center Nottingham (center 2) and Herzogenaurach (center 4) at three predefined time points.

3.6 Final remarks

The logit-normal approach taken here is one of many ways to deal with a BOS. For instance, Grootendorst (2000) regards the extreme values of 0 and 1 that the BOS can assume as an indication for a latent score censored at the boundaries leading to the Tobit model. Further, we could have modeled the BOS also using a random effects ordinal logistic model. However, for the analysis of the CERISE data the analysis corresponding to Table 3.5 needed about 6 hrs, hence not an advantage over the OP model. Furthermore, for modeling of BOSs which arise as proportions Lesaffre *et al.* (2007) suggested to use a generalized linear mixed-effects model in the univariate case. It would be of interest to see how this model can be generalized to a longitudinal setting. Finally, Arostegui *et al.* (2007) model the distribution of a univariate BOS arising from the SF-36 form (a popular health-related quality of life score) using a beta-binomial approach thereby assuming that the BOS is in fact a proportion. This could have assumed also here and again it would be interesting to see how the beta-binomial approach could be extended to a longitudinal setting.

To conclude, this paper has shown that modeling a longitudinal study with a BOS as outcome can be effectively done with the CO model. In comparison to the OP model, it has the same performance under logit-normality assumptions but is dramatically quicker especially when σ depends on covariates. The computational advantage of the CO model could be exploited in the simulation studies aiming to establish required sample size for a longitudinal study with a BOS as outcome Further, the performance of the LM model showed to be inferior in our simulations to the CO and OP model, but despite its ad hoc nature it is nevertheless useful as a first approach. Furthermore, the OP approach does not suffer from the estimation of a large number of cut points which was already observed in the univariate case. Finally, our simulations and the analysis of the CERISE data suggest that it is advisable to model also σ as a function of covariates. Ignoring the dependence of the residual variance on covariates can induce bias in estimating the parameters of the mean structure. In this sense, we experienced that the CO model is easier to work with than the OP model where the dependence of the residual variance is modeled in a less transparent manner. However, one could fit an OP model as a robustness analysis at the end of the variable selection.

APPENDIX: Score Test for CO and OP models

In this section we derive expressions for the score statistic to test the null hypothesis of no treatment effect. We work in the cross sectional setting, where we compare two groups. Two models are under consideration: Coarsening model (CO) and Ordinal Probit (OP). For both models score statistics are derived analytically and the approach to compare their value for large samples is presented. Final numerical evaluation is necessary. Analytical derivations used techniques presented in Whitehead (1992). First we derive the score statistic for the CO model, it is followed by the derivation for the OP model, and finally an approach for numerical comparisons of the asymptotic values of the two statistics is presented.

Derivation of the score test for the coarsening model

We assume that the observations come from two populations (experimental treatment and control), the residual variance is constant and equal for both populations. There are $m+1$ distinct values which can be observed, implying m valid cut-off points $z_1 \dots z_m$. For the ease of notation we assume two additional cut-off points: $z_0 = -\infty$ and $z_{m+1} = +\infty$. In the coarsening model cut-off points $z_1 \dots z_m$ are fixed and known a priori. Moreover, we use the following notation: n_{iC} denotes number of observations in the control group in the i-th class $(i = 1, \dots, (m+1))$, while n_{iE} refers to the experimental group.

The total likelihood can be expressed as follows:

$$L(\boldsymbol{\beta}, \sigma; y) = \prod_{i=1}^{m+1} [\Phi(\frac{z_i - \beta_0}{\sigma}) - \Phi(\frac{z_{i-1} - \beta_0}{\sigma})]^{n_{iC}} \times \\ \times [\Phi(\frac{z_i - \beta_0 - \beta_1}{\sigma}) - \Phi(\frac{z_{i-1} - \beta_0 - \beta_1}{\sigma})]^{n_{iE}}. \quad (3.14)$$

The following denotes the log-likelihood:

$$
\begin{aligned}
l(\boldsymbol{\beta},\sigma;y) = log(L(\boldsymbol{\beta},\sigma;y)) = l^{(C)} + l^{(E)},\\
l^{(C)} = \sum_{i=1}^{m+1} n_{iC} \times log[\Phi(z_i,\mu_C) - \Phi(z_{i-1},\mu_C)]\\
l^{(E)} = \sum_{i=1}^{m+1} n_{iE} \times log[\Phi(z_i,\mu_E) - \Phi(z_{i-1},\mu_E)],
\end{aligned}
$$

where $\Phi(z,\mu) = \Phi(\frac{z-\mu}{\sigma})$ and $\mu_C = \beta_0$, while $\mu_E = \beta_0 + \beta_1$.

We utilize the following notation. Observations in the control group come from the density $f_{\psi_1,\eta}(x)$ and observations in the experimental group come from the density $f_{\psi_2,\eta}(x)$, where η is an unknown common vector of nuisance parameters. The parameter of interest is $\theta = \frac{1}{2}(\psi_1 - \psi_2)$, and the nuisance parameter is made up of $\varphi = \frac{1}{2}(\psi_1 + \psi_2)$. This implies expressions for ψ_1 and ψ_2, which are:

$$\psi_1 = \varphi + \theta \qquad\qquad , \psi_2 = \varphi - \theta.$$

In our situation the following is used: $\theta = \beta_1$ and $\varphi = -2\beta_0 - \beta_1$, which implies $\psi_1 = -2\mu_C = -2(\beta_0)$ and $\psi_2 = -2\mu_E = -2(\beta_0 + \beta_1)$. Therefore the log-likelihood can be noted as: $l(\psi_1,\psi_2,\eta;y) = l^{(C)}(\psi_1,\eta) + l^{(E)}(\psi_2,\eta)$. The vector of nuisance parameters is $\eta = \{\sigma\}$.

To calculate the score test we are looking for the following expressions:

$$
\begin{aligned}
Z = l_\theta(0,\varphi^*,\eta^*),\\
V = -l_{\theta\theta}(0,\varphi^*,\eta^*),
\end{aligned}
$$

where φ^* and η^* are maximum likelihood estimators under the constraint that $\theta = 0$. Moreover, we show that for large sample sizes,equal allocation ratio to the control and experimental arms, and for small θ derivatives $l_{\theta\varphi}$ and $l_{\theta\eta}$ are nearly zero, which allows to use the above expression for V. To calculate l_θ we use $l_\theta = l^{(C)}_{\psi_1}\frac{\partial\psi_1}{\partial\theta} +$

$l^{(E)}_{\psi_2}\frac{\partial\psi_2}{\partial\theta} = l^{(C)}_{\psi_1} - l^{(E)}_{\psi_2}$. The above formula yields the following expression:

$$l_\theta = \sum_{i=1}^{m+1} n_{iC} \times [\frac{\phi_i(\psi_1) - \phi_{i-1}(\psi_1)}{2\sigma[\Phi_i(\psi_1) - \Phi_{i-1}(\psi_1)]}] - \sum_{i=1}^{m+1} n_{iE} \times [\frac{\phi_i(\psi_2) - \phi_{i-1}(\psi_2)}{2\sigma[\Phi_i(\psi_2) - \Phi_{i-1}(\psi_2)]}],$$

where $\Phi_i(\psi_1) = \Phi(\frac{z_i+0.5\psi_1}{\sigma})$, $\Phi_i(\psi_2) = \Phi(\frac{z_i+0.5\psi_2}{\sigma})$, $\phi_i(\psi_1) = \frac{\partial\Phi(x)}{\partial x}|_{x=\frac{z_i+0.5\psi_1}{\sigma}}$. This derivative evaluated at $(0, \varphi^*, \eta^*)$ is equal to:

$$Z = \sum_{i=1}^{m+1} (n_{iC} - n_{iE}) \times [\frac{\phi_i(\varphi^*) - \phi_{i-1}(\varphi^*)}{2\sigma^*[\Phi_i(\varphi^*) - \Phi_{i-1}(\varphi^*)]}], \tag{3.15}$$

where $\Phi_i(\varphi^*) = \Phi(\frac{z_i+0.5\varphi^*}{\sigma^*})$.
Next we derive the expression for V using the fact that $l_{\theta\theta} = l^{(C)}_{\psi_1\psi_1} + l^{(E)}_{\psi_2\psi_2}$, therefore

$$l_{\theta\theta} = \sum_{i=1}^{m+1} n_{iC}[-\frac{\phi_i(\psi_1) \times (z_i + 0.5\psi_1) - \phi_{i-1}(\psi_1) \times (z_{i-1} + 0.5\psi_1)}{4\sigma^3[\Phi_i(\psi_1) - \Phi_{i-1}(\psi_1)]} - \frac{[\phi_i(\psi_1) - \phi_{i-1}(\psi_1)]^2}{4\sigma^2[\Phi_i(\psi_1) - \Phi_{i-1}(\psi_1)]^2}]+ \\ + \sum_{i=1}^{m+1} n_{iE}[-\frac{\phi_i(\psi_2) \times (z_i + 0.5\psi_2) - \phi_{i-1}(\psi_2) \times (z_{i-1} + 0.5\psi_2)}{4\sigma^3[\Phi_i(\psi_2) - \Phi_{i-1}(\psi_2)]} - \frac{[\phi_i(\psi_2) - \phi_{i-1}(\psi_2)]^2}{4\sigma^2[\Phi_i(\psi_2) - \Phi_{i-1}(\psi_2)]^2}].$$

The above expression evaluated at $(0, \varphi^*, \eta^*)$ and mutiplied by minus one is equal to:

$$V \approx \sum_{i=1}^{m+1} (n_{iC} + n_{iE})[\frac{\phi_i(\varphi^*) \times (z_i + 0.5\varphi^*) - \phi_{i-1}(\varphi^*) \times (z_{i-1} + 0.5\varphi^*)}{4\sigma^3[\Phi_i(\varphi^*) - \Phi_{i-1}(\varphi^*)]}+ \\ + \frac{[\phi_i(\varphi^*) - \phi_{i-1}(\varphi^*)]^2}{4\sigma^2[\Phi_i(\varphi^*) - \Phi_{i-1}(\varphi^*)]^2}]. \tag{3.16}$$

As φ^* and η^* are maximum likelihood estimators, l_φ and l_η are equal zero, evaluated at $(0, \varphi^*, \eta^*)$, we have the following:

$$l_\sigma|_{(0,\varphi^*,\eta^*)} = \sum_{i=1}^{m+1} (n_{iC} + n_{iE})[\frac{\phi_i(\varphi^*) \times (z_i + 0.5\varphi^*) - \phi_{i-1}(\varphi^*) \times (z_{i-1} + 0.5\varphi^*)}{\sigma^2[\Phi_i(\varphi^*) - \Phi_{i-1}(\varphi^*)]}]. \tag{3.17}$$

The above expression allows to simplify the V formula. In the end, we obtain score statistic $\frac{Z}{\sqrt{V}}$ as follows:

$$\frac{Z}{\sqrt{V}} = \frac{\sum_{i=1}^{m+1}(n_{iC}-n_{iE}) \times [\frac{\phi_i(\varphi^*)-\phi_{i-1}(\varphi^*)}{[\Phi_i(\varphi^*)-\Phi_{i-1}(\varphi^*)]}]}{\sqrt{\sum_{i=1}^{m+1}(n_{iC}+n_{iE})[\frac{[\phi_i(\varphi^*)-\phi_{i-1}(\varphi^*)]^2}{[\Phi_i(\varphi^*)-\Phi_{i-1}(\varphi^*)]^2}]}}. \tag{3.18}$$

It remains to show that $l_{\theta\varphi}$ and $l_{\theta\eta}$ tend to zero for the validity of the approximation of the V formula. We have $l_{\theta\varphi} = l^{(C)}_{\psi_1\psi_1} - l^{(E)}_{\psi_2\psi_2}$, and therefore:

$$\begin{aligned} l_{\theta\varphi} = \sum_{i=1}^{m+1} n_{iC}[-\frac{\phi_i(\psi_1)\times(z_i+0.5\psi_1)-\phi_{i-1}(\psi_1)\times(z_{i-1}+0.5\psi_1)}{4\sigma^3[\Phi_i(\psi_1)-\Phi_{i-1}(\psi_1)]} - \frac{[\phi_i(\psi_1)-\phi_{i-1}(\psi_1)]^2}{4\sigma^2[\Phi_i(\psi_1)-\Phi_{i-1}(\psi_1)]^2}]- \\ -\sum_{i=1}^{m+1} n_{iE}[-\frac{\phi_i(\psi_2)\times(z_i+0.5\psi_2)-\phi_{i-1}(\psi_2)\times(z_{i-1}+0.5\psi_2)}{4\sigma^3[\Phi_i(\psi_2)-\Phi_{i-1}(\psi_2)]} - \frac{[\phi_i(\psi_2)-\phi_{i-1}(\psi_2)]^2}{4\sigma^2[\Phi_i(\psi_2)-\Phi_{i-1}(\psi_2)]^2}]. \end{aligned}$$

Now, when $\sum n_{iE} = \sum n_{iC} = n$ and n is large,with small θ the above expression tends to zero. Finally we present an expression for $l_{\theta\sigma} = l^{(C)}_{\psi_1\sigma} - l^{(E)}_{\psi_2\sigma}$, which is:

$$\begin{aligned} l_{\theta\sigma} = \sum_{i=1}^{m+1} n_{iC}[\frac{\phi_i(\psi_1)(z_i+0.5\psi_1)^2-\phi_{i-1}(\psi_1)(z_{i-1}+0.5\psi_1)^2}{2\sigma^4[\Phi_i(\psi_1)-\Phi_{i-1}(\psi_1)]} - \frac{\phi_i(\psi_1)-\phi_{i-1}(\psi_1)}{2\sigma^2[\Phi_i(\psi_1)-\Phi_{i-1}(\psi_1)]}+ \\ +\frac{[\phi_i(\psi_1)-\phi_{i-1}(\psi_1)][\phi_i(\psi_1)(z_i+0.5\psi_1)-\phi_{i-1}(\psi_1)(z_{i-1}+0.5\psi_1)]}{2\sigma^3[\Phi_i(\psi_1)-\Phi_{i-1}(\psi_1)]^2}]- \\ -\sum_{i=1}^{m+1} n_{iE}[\frac{\phi_i(\psi_2)(z_i+0.5\psi_2)^2-\phi_{i-1}(\psi_2)(z_{i-1}+0.5\psi_2)^2}{2\sigma^4[\Phi_i(\psi_2)-\Phi_{i-1}(\psi_2)]} - \frac{\phi_i(\psi_2)-\phi_{i-1}(\psi_2)}{2\sigma^2[\Phi_i(\psi_2)-\Phi_{i-1}(\psi_2)]}+ \\ +\frac{[\phi_i(\psi_2)-\phi_{i-1}(\psi_2)][\phi_i(\psi_2)(z_i+0.5\psi_2)-\phi_{i-1}(\psi_2)(z_{i-1}+0.5\psi_2)]}{2\sigma^3[\Phi_i(\psi_2)-\Phi_{i-1}(\psi_2)]^2}]. \end{aligned}$$

The above expression tends to zero, when $\theta \to 0$, for large n, and equal allocation of subjects to the control and experimental group. Under these conditions the approximation for the V holds.

Derivation of the score test for the ordinal probit model

We proceed in the similar fashion as in the first section. However, for the identification we fix $\beta_0 = 0$ and $\sigma = 1$. Moreover, we estimate m cut-off points $i_1 \dots i_m$. These cut-offs can be expressed as increments: $i_i = i_1 + d_i$, and $d_0 = -\infty$, $d_1 = 0$ and $d_{m+1} = +\infty$. The likelihood can be expressed as follows:

$$L(\beta_1, i_1, d_2 \dots d_m; y) = \prod_{i=1}^{m+1} [\Phi(i_1 + d_i) - \Phi(i_1 + d_{i-1})]^{n_{iC}} \times$$
$$\times [\Phi(i_1 + d_i - \beta_1) - \Phi(i_1 + d_{i-1} - \beta_1)]^{n_{iE}}.$$

Now we use the following parametrization: $\theta = \beta_1$ and $\varphi = 2i_1 - \beta_1$, which implies $\psi_1 = 2i_1$ and $\psi_2 = 2i_1 - 2\beta_1$. In similar manner as in the section one, we calculate the derivatives l_θ and $l_{\theta\theta}$ and evaluate them at point $(0, \varphi^*, \boldsymbol{\eta}^*)$, now vector $\boldsymbol{\eta} = \{d_2 \dots d_m\}$. Calculations lead to the following formulas for Z and V.

$$Z = \frac{1}{2} \sum_{i=1}^{m+1} (n_{iC} - n_{iE}) [\frac{\phi_i(\varphi^*) - \phi_{i-1}(\varphi^*)}{\Phi_i(\varphi^*) - \Phi_{i-1}(\varphi^*)}], \tag{3.19}$$

$$V \approx \frac{1}{4} \sum_{i=1}^{m+1} (n_{iC} + n_{iE}) [\frac{\phi_i(\varphi^*)(0.5\varphi^* + d_i^*) - \phi_{i-1}(\varphi^*)(0.5\varphi^* + d_{i-1}^*)}{\Phi_i(\varphi^*) - \Phi_{i-1}(\varphi^*)} +$$
$$+ \frac{[\phi_i(\varphi^*) - \phi_{i-1}(\varphi^*)]^2}{[\Phi_i(\varphi^*) - \Phi_{i-1}(\varphi^*)]^2}], \tag{3.20}$$

where $\Phi_i(\varphi^*) = \Phi(0.5\varphi^* + d_i^*)$. We show that the first term in (3.20) is equal to zero. Lets consider the ordinal probit model, which is identified by setting two first cut-off points constant, and allowing the estimation of β_0 and σ. The likelihood expression is thus the same as (3.14), just instead of having all cut-off points fixed, we fix only the first two cut-offs (these can be arbitrary two thresholds). In this model at the point $(0, \varphi^*, \eta^*)$ the derivative with respect to σ is equal to zero and has the same expression as in (3.17). Under $\theta = 0$ we have one sample problem, and the maximum likelihood estimates can be obtained from the multinomial distribution for ordinal probit model, in both identification situations. First part in (3.20) is equal to expression in (3.17), with the different parameters in the cdf and pdf functions and additional σ parameter. However in both situations $\Phi_i(\varphi^*) = \frac{\sum_{j=1}^{i}(n_{jC}+n_{jT})}{\sum_{j=1}^{m+1}(n_{jC}+n_{jE})}$, and therefore one expresion is equal to the second expression multiplied by some constant, as one of these expressions is equal to zero the other must be zero as well. This yields an expression for the ordinal probit score test as follows:

$$\frac{Z}{\sqrt{V}} = \frac{\sum_{i=1}^{m+1}(n_{iC} - n_{iE})[\frac{\phi_i(\varphi^*)-\phi_{i-1}(\varphi^*)}{\Phi_i(\varphi^*)-\Phi_{i-1}(\varphi^*)}]}{\sqrt{\sum_{i=1}^{m+1}(n_{iC} + n_{iE})[\frac{[\phi_i(\varphi^*)-\phi_{i-1}(\varphi^*)]^2}{[\Phi_i(\varphi^*)-\Phi_{i-1}(\varphi^*)]^2}]}}. \tag{3.21}$$

It can be easily shown that derivatives $l_{\theta\varphi}$ and $l_{\theta\eta}$ tend to zero when the allocation ratio to control and experimental treatment is close to one, number of subjects in each group is large and the treatment effect $\theta \to 0$, which ensures the validity of the V formula.

Comparison of OP and CO score test statistics

Both expressions (3.18) and (3.21) have the same functional form, are however functions of different estimates. All the estimates are ML estimates obtained under the constraint $\theta = 0$. Therefore the expressions in case of the ordinal probit can be obtained using ML estimators of the multinomial distribution, while in case of the coarsening model, estimates are those obtained in case normal distribution is fitted to the sample obtained under the 50-50 mixture of two normal distributions with the same variance. Therefore for the large samples we have the following results:

$$\begin{aligned}\Phi_i(\varphi^*) = \Phi(0.5\varphi^* + d_i^*) = \frac{\sum_{j=1}^{i}(n_{jC} + n_{jE})}{2n} \to \frac{n\sum_{j=1}^{i}(p_{jC} + p_{jE})}{2n} = \\ = 0.5(\Phi(\frac{z_i - \beta_0}{\sigma}) + \Phi(\frac{z_i - \beta_0 - \beta_1}{\sigma})),\end{aligned} \tag{3.22}$$

for the ordinal probit. And for the coarsening model the following holds:

$$\Phi_i(\varphi^*) = \Phi(\frac{z_i + 0.5\varphi^*}{\sigma^*}) = \Phi(\frac{z_i - \beta_0^*}{\sigma^*})$$
$$\beta_0^* \approx \frac{1}{2}(\mu_C + \mu_E) \qquad \sigma^* \approx var(\mu) + \sigma \tag{3.23}$$

Equations 3.22 and 3.23 allow us the computation of φ^*, which can be substituted into 3.21 and 3.18 respectively. Note that although the same symbol is used φ^*, it represents distinct quantities in 3.21 and 3.18, which also change with index i. Therefore, for large samples we are able to compare the values of 3.21 and 3.18. This comparison, however, must be evaluated numerically as the formulas involve the normal cumulative distribution function Φ or its inverse Φ^{-1}.

References

Arostegui, I., Nunez-Anton, V., and Quintana, J. (2007). Analysis of the Short Form-36 (SF-36): The beta binomial distribution approach. *Statistics in Medicine*, **26**, 1318–1342.

Ghidey, W., Lesaffre, E., and Eilers, P. (2004). Smooth random effects distribution in a linear mixed model. *Biometrics*, **60**, 945Ű953.

Grootendorst, P. (2000). Censoring in statistical models of health status: What happens when one can do better than ‘1’. *Quality of Life Research*, **9**, 911–914.

Lesaffre, E., Rizopoulos, D., and Tsonaka, R. (2007). The logistic-transform for bounded outcome scores. *Biostatistics*, **8**, 72–85.

Morrell, C., Pearson, J., and Brant, L. (1997). Linear transformations of linear mixed-effects models. *The American Statistician*, **51**, 338–343.

Skrondal, A. and S., R.-H. (2004). *Generalized Latent Variable Modeling: Multilevel, Longitudinal and Structural Equations Models.* Chapman & Hall/CRC, Boca Raton.

Tsonaka, S., Rizopoulos, D., and Lesaffre, E. (2006). Power and sample size calculations for discrete bounded outcome scores. *Statistics in Medicine*, **25**, 4241–4252.

Whitehead, J. (1992). *The Design and Analysis of Sequential Clinical Trials*. Ellis Horwood, Chichester.

Wit, L. D., Putman, K., Dejaeger, E., Baert, I., Berman, P., Bogaerts, K., Brinkmann, N., Connell, L., Feys, H., Jenni, W., Kaske, C., Lesaffre, E., Leys, M., Lincoln, N., Louckx, F., Schuback, B., Schupp, W., Smith, B., and Weerdt, W. D. (2005). Use of time by stroke patients: a comparison of four european rehabilitation centers. *Stroke*, **36(9)**, 1977–1983.

Wit, L. D., Putman, K., Schuback, B., Komarek, A., Angst, F., Baert, I., Berman, P., Bogaerts, K., Brinkmann, N., Connell, L., Dejaeger, E., Feys, H., Jenni, W., Kaske, C., Lesaffre, E., Leys, M., Lincoln, N., Louckx, F., Schupp, W., Smith, B., and Weerdt, W. D. (2007). Motor and functional recovery after stroke: A comparison of 4 european rehabilitation centers. *Stroke*, **38**, 2101–2107.

Chapter
4

Hierarchical Generalized Linear Models: The R Package HGLMMM

Based on:

Molas M. and Lesaffre E.(2011). Hierarchical generalized linear models: the R package HGLMMM. *Journal of Statistical Software* **39** 13 pp. 1-20

Abstract

The **R** package **HGLMMM** has been developed to fit generalized linear models with random effects using the h-likelihood approach. The response variable is allowed to follow a binomial, Poisson, Gaussian or gamma distribution. The distribution of random effects can be specified as Gaussian, gamma, inverse-gamma or beta. Complex structures as multi-membership design or multilevel designs can be handled. Further, dispersion parameters of random components and the residual dispersion (overdispersion) can be modeled as a function of covariates. Overdispersion parameter can be fixed or estimated. Fixed effects in the mean structure can be estimated using extended likelihood or a first order Laplace approximation to the

marginal likelihood. Dispersion parameters are estimated using first order adjusted profile likelihood.

4.1 Introduction

In Nelder and Wedderburn (1972) the class of generalized linear models (GLM) was developed. This class of models allows for the response to follow a distribution from the exponential family, extending modeling capabilities beyond the Gaussian response. In Henderson *et al.* (1959) the linear mixed model was suggested, which enabled to model correlation in the data. Further, it was extended to the generalized linear mixed model (see e.g., Molenberghs and Verbeke (2005)), where the response from an exponential family is combined with normal random effects. In Lee and Nelder (1996) hierarchical generalized linear models were described, which allows random effects to be not normally distributed. Further Lee and Nelder (1996) proposed the extended likelihood rather than the marginal likelihood to estimate the parameters. Later Lee and Nelder (2001) focused on the joint modelling of the mean and dispersion structure. This estimation technique relies on the Iterative Weighted Least Squares (IWLS) algorithm, where fixed effects and random effects are estimated using the extended likelihood, and dispersion parameters are obtained by maximizing the adjusted profile likelihood. A subsequent adjustment of the algorithm was proposed in Noh and Lee (2007) replacing the extended likelihood by the first order adjusted profile likelihood as a criterion to estimate fixed effects in the mean structure.

The objective of this paper is to present the **R** (**R** Development Core Team, 2009) package **HGLMMM** available from the first author upon request or from Comprehensive **R** Archive Network (CRAN). The package runs under **R** version 2.9.0 or higher. This package fits the class of generalized linear models with random effects. In the remainder of the paper we first outline the h-likelihood approach to the estimation and statistical inference. Next, we present the capabilities of the **HGLMMM** package on real-life datasets, described in Lee *et al.* (2006).

4.2 H-likelihood estimation and inference framework

Standard maximum likelihood estimation for models with random effects is based on the marginal likelihood as objective function. The parameters are estimated by a marginal likelihood procedure (MML) and their standard errors are computed from the inverse of the negative hessian matrix of the marginal likelihood. In the marginal likelihood approach random effects $\mathbf{v}$ are integrated out and only fixed effects in the mean structure $\boldsymbol{\beta}$ and dispersion parameters $\boldsymbol{\lambda}$ are retained in the maximized function. For a mixed effects model the conditional likelihood of the j^{th} $(j = 1, \ldots, n_i)$ repeated observation on the i^{th} subject $(i = 1, \ldots, N)$, i.e., y_{ij}, is given by $f_{\boldsymbol{\beta},\boldsymbol{\lambda}}(y_{ij}|\mathbf{v}_i)$. The likelihood of the i^{th} random effect is denoted as $f_{\boldsymbol{\lambda}}(\mathbf{v}_i)$. Note that $\boldsymbol{\lambda}$ contains dispersion parameters of the random components $\mathbf{v}_i$ as well as the parameters describing the residual dispersion (overdispersion) of the response y_{ij}. The marginal likelihood maximized in the MML procedure is given by

$$L_M(\boldsymbol{\beta}, \boldsymbol{\lambda}|\mathbf{y}) = \prod_{i=1}^{N} \int \prod_{j=1}^{n_i} f_{\boldsymbol{\beta},\boldsymbol{\lambda}}(y_{ij}|\mathbf{v}_i) f_{\boldsymbol{\lambda}}(\mathbf{v}_i) d\,\mathbf{v}_i. \tag{4.1}$$

Maximizing L_M or equivalently the log-likelihood $\ell_M = log(L_M)$ yields consistent estimates of the fixed effects parameters. However, the problem lies in computing the integrated likelihood. This may be a time-consuming task especially for complex models since it needs to be done for each subject and each iteration. Further, if the MLE is determined with a Newton-Raphson procedure then integrals need to be computed also for the first and second derivatives. It is also important to fine tune the likelihood calculations, see e.g., Lesaffre and Spiessens (2001).

In Lee and Nelder (1996) another approach to estimating the parameters was proposed. These authors argued to use the joint likelihood L_E for the maximization, which is directly available from the definition of the model. The joint likelihood, called also extended likelihood or h-likelihood, is then maximized jointly with respect to $\mathbf{v}$ and $\boldsymbol{\beta}$ given dispersion parameters $\boldsymbol{\lambda}$. At the maximum, standard errors are obtained in the classical way. In the notation of above, the extended likelihood is given by:

$$L_E(\boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{v}|\mathbf{y}, \mathbf{v}) = \prod_{i=1}^{N} \prod_{j=1}^{n_i} f_{\boldsymbol{\beta},\boldsymbol{\lambda}}(y_{ij}|\mathbf{v}_i) f_{\boldsymbol{\lambda}}(\mathbf{v}_i). \tag{4.2}$$

The logarithm of (4.2) is called the extended log-likelihood by Lee *et al.* (2006) and they denoted its logarithm as $h = \log\left[L_E(\boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{v}|\mathbf{y}, \mathbf{v})\right]$. We could say that this extended likelihood reflects the hierarchical character of the data.

In the h-likelihood approach the estimates of dispersion parameters are determined by maximizing the adjusted profile likelihood introduced by Cox and Reid (1987). However, for some models the approach of Lee and Nelder (1996) is not appropriate, therefore Noh and Lee (2007) proposed to replace the joint likelihood as the estimation criterion for $\boldsymbol{\beta}$ with another adjusted profile likelihood. The outline of this procedure is given in the next section.

4.2.1 Computing marginal MLEs using the h-likelihood approach

In some special cases, i.e., when the random effects are on the canonical scale (see e.g., Lee *et al.* (2006) pp. 112-114), joint maximization of the extended log-likelihood h with respect to all parameters $(\boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{v}_1, \ldots, \mathbf{v}_N)$ is equivalent to maximizing the marginal likelihood with respect to $\boldsymbol{\beta}, \boldsymbol{\lambda}$ and taking the Empirical Bayes (EB) estimates for $\mathbf{v}_1, \ldots, \mathbf{v}_N$. But, most often the two maximization procedures are not equivalent.

Noh and Lee (2007) suggest in the general case to work with a Laplace approximation to the marginal likelihood (4.1). Namely, the integral of the function $k(\mathbf{x}, \mathbf{y}) \exp\left[-ng(\mathbf{x}, \mathbf{y})\right]$ with respect to $\mathbf{x}$ can be approximated as follows:

$$\int k(\mathbf{x}, \mathbf{y}) \exp\left[-ng(\mathbf{x}, \mathbf{y})\right] d\mathbf{x} = \left| \frac{n \frac{\partial^2 g(\mathbf{x},\mathbf{y})}{\partial \mathbf{x}^\top \partial \mathbf{x}}}{2\pi} \right|_{\mathbf{x}=\hat{\mathbf{x}}}^{-\frac{1}{2}} \exp\left[-ng(\mathbf{x}, \mathbf{y})\right] k(\mathbf{x}, \mathbf{y})|_{\mathbf{x}=\hat{\mathbf{x}}}, \tag{4.3}$$

with $\hat{\mathbf{x}}$ the value of $\mathbf{x}$ that maximizes $-g(\mathbf{x}, \mathbf{y})$. This is called the Laplace approximation of the above integral (at $\hat{\mathbf{x}}$). More information on the Laplace approximation can be found in e.g., Severini (2000) (Section 2.11).

Taking in expression (4.3) $k(\mathbf{x},\mathbf{y}) = 1$; $\exp\left[ng(\mathbf{x},\mathbf{y})\right] = \exp\left[-h(\boldsymbol{\theta},\mathbf{v})\right]$ whereby $\boldsymbol{\theta} = (\boldsymbol{\beta},\boldsymbol{\lambda})$; $\mathbf{v}$ representing the stacked vector of N random effects and finally $n\frac{\partial^2 g(\mathbf{x},\mathbf{y})}{\partial\mathbf{x}^\top\partial\mathbf{x}} = -\frac{\partial^2 h(\boldsymbol{\theta},\mathbf{v})}{\partial\mathbf{v}^\top\partial\mathbf{v}}$, leads to

$$L_M(\boldsymbol{\beta},\boldsymbol{\lambda}|\mathbf{y}) = \int \exp\left[h(\boldsymbol{\beta},\boldsymbol{\lambda},\mathbf{v})\right]d\mathbf{v} \approx \left|\frac{-\frac{\partial^2 h(\boldsymbol{\beta},\boldsymbol{\lambda},\mathbf{v})}{\partial\mathbf{v}^\top\partial\mathbf{v}}}{2\pi}\right|_{\mathbf{v}=\hat{\mathbf{v}}}^{-\frac{1}{2}} \exp\left[h(\boldsymbol{\beta},\boldsymbol{\lambda},\mathbf{v})\right]|_{\mathbf{v}=\hat{\mathbf{v}}}, \quad (4.4)$$

with $\hat{\mathbf{v}}$ maximizing the extended likelihood for a given (starting) value of $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$, i.e., $\hat{\mathbf{v}}(\boldsymbol{\beta},\boldsymbol{\lambda})$. Note that the approximation improves when the number of observations per subject, n_i increases, which can be inferred from Severini (2000). Taking the logarithm of the previous expression leads to the adjusted profile (log)-likelihood

$$p_v(h) = h(\boldsymbol{\beta},\boldsymbol{\lambda},\mathbf{v})|_{\mathbf{v}=\hat{\mathbf{v}}} - 0.5\log\left|\frac{D(h,\mathbf{v})}{2\pi}\right|_{\mathbf{v}=\hat{\mathbf{v}}}, \quad (4.5)$$

with $D(h,\mathbf{v}) = -\frac{\partial^2 h(\boldsymbol{\beta},\boldsymbol{\lambda},\mathbf{v})}{\partial\mathbf{v}^\top\partial\mathbf{v}}$. The term 'adjusted profile likelihood' is chosen since $h(\boldsymbol{\beta},\boldsymbol{\lambda},\mathbf{v})|_{\mathbf{v}=\hat{\mathbf{v}}}$ is a profile (log)-likelihood of $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$ and the second term in (4.5) is a correction term to approximate the marginal log-likelihood.

The next step in the iterative procedure is to maximize the adjusted profile (log-)likelihood (4.5) with respect to $\boldsymbol{\beta}$. Note that maximizing the profile log-likelihood $h(\boldsymbol{\beta},\boldsymbol{\lambda},\mathbf{v})|_{\mathbf{v}=\hat{\mathbf{v}}}$ to find the MLE of $\boldsymbol{\beta}$ is not appropriate since this is equivalent to joint maximization of h over $\boldsymbol{\beta}$ and $\mathbf{v}$ which is most often invalid as seen above.

After obtaining $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}$ from maximization of (4.5) for a given dispersion component $\boldsymbol{\lambda}$, the estimation algorithm proceeds with estimation of $\boldsymbol{\lambda}$. Another adjusted profile likelihood is used as an objective function to find $\hat{\boldsymbol{\lambda}}$.

Let the marginal distribution of the data $\mathbf{y}$ be $f_{\boldsymbol{\beta},\boldsymbol{\lambda}}(\mathbf{y})$ (marginalized over $\mathbf{v}$), i.e., the LHS of expression (4.4) or its approximation, now seen as a probability density function (pdf) of the data. Conditional on the sufficient statistics for $\boldsymbol{\beta}$, i.e., $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}$, (see Cox and Hinkley (1974), page 21), the (marginalized) distribution of the data can be derived from:

$$f_{\boldsymbol{\lambda}}(\mathbf{y}|\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}) = \frac{f_{\boldsymbol{\beta},\boldsymbol{\lambda}}(\mathbf{y})}{f_{\boldsymbol{\beta},\boldsymbol{\lambda}}(\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}})}, \quad (4.6)$$

where $f_{\boldsymbol{\beta},\boldsymbol{\lambda}}(\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}})$ is the distribution of $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}$. One then applies the *p-formula* as derived by Barndorff-Nielsen (1980, 1983), see also Pawitan (2001), to obtain in general the distribution of the ML estimator. Namely,

$$f_{\boldsymbol{\beta},\boldsymbol{\lambda}}(\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}) = \left| -\frac{1}{2\pi} \frac{\partial^2 \log f_{\boldsymbol{\beta},\boldsymbol{\lambda}}(\mathbf{y})}{\partial \boldsymbol{\beta}^\top \partial \boldsymbol{\beta}} \right|^{\frac{1}{2}}_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}} \frac{f_{\boldsymbol{\beta},\boldsymbol{\lambda}}(\mathbf{y})}{f_{\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}},\boldsymbol{\lambda}}(\mathbf{y})}, \tag{4.7}$$

where $f_{\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}},\boldsymbol{\lambda}}(\mathbf{y})$ is the marginal profile likelihood of $\boldsymbol{\lambda}$, i.e.,

$$\log[f_{\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}},\boldsymbol{\lambda}}(\mathbf{y})] = \ell_M(\boldsymbol{\beta}_{\boldsymbol{\lambda}}, \boldsymbol{\lambda}|\mathbf{y})|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}}.$$

After substitution of expression (4.7) into (4.6) one obtains:

$$\log\left[f_{\boldsymbol{\lambda}}(\mathbf{y}|\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}})\right] = \ell_M(\boldsymbol{\beta}_{\boldsymbol{\lambda}}, \boldsymbol{\lambda}|\mathbf{y})|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}} - 0.5 \log \left| \frac{D(\ell_M, \boldsymbol{\beta})}{2\pi} \right|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}}, \tag{4.8}$$

with $D(\ell_M, \boldsymbol{\beta}) = -\frac{\partial^2 \ell_M}{\partial \boldsymbol{\beta}^\top \partial \boldsymbol{\beta}}$. In the next step, one replaces everywhere the marginal log likelihood ℓ_M by the adjusted profile log-likelihood $p_v(h)$ evaluated in $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}$. This results in:

$$\log\left[f_{\boldsymbol{\lambda}}(\mathbf{y}|\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}})\right] = h(\boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{v})|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}, \mathbf{v}=\hat{\mathbf{v}}} - 0.5 \log \left| \frac{D(h, \mathbf{v})}{2\pi} \right|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}, \mathbf{v}=\hat{\mathbf{v}}} - 0.5 \log \left| \frac{D\left[p_v(h), \boldsymbol{\beta}\right]}{2\pi} \right|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}}. \tag{4.9}$$

Finally, in Appendix 4 of Lee and Nelder (2001) it is shown that the sum of the last two terms in the above expression is equal to

$$-0.5 \log \left| \frac{D\left[h, (\boldsymbol{\beta}, \mathbf{v})\right]}{2\pi} \right|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}, \mathbf{v}=\hat{\mathbf{v}}},$$

with $D\left[h, (\boldsymbol{\beta}, \mathbf{v})\right]$ equal to $\begin{pmatrix} -\frac{\partial^2 h}{\partial \boldsymbol{\beta}^\top \partial \boldsymbol{\beta}} & -\frac{\partial^2 h}{\partial \boldsymbol{\beta}^\top \partial \mathbf{v}} \\ -\frac{\partial^2 h}{\partial \mathbf{v}^\top \partial \boldsymbol{\beta}} & -\frac{\partial^2 h}{\partial \mathbf{v}^\top \partial \boldsymbol{\beta}} \end{pmatrix}$ with dimensions equal to the sum of the dimensions of two adjustment terms in (4.9). As a result one obtains the following adjusted profile likelihood:

$$p_{\beta,v}(h) = h(\boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{v})|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}, \mathbf{v}=\hat{\mathbf{v}}} - 0.5 \log \left| \frac{D\left[h, (\boldsymbol{\beta}, \mathbf{v})\right]}{2\pi} \right|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}, \mathbf{v}=\hat{\mathbf{v}}}. \tag{4.10}$$

The latter adjusted profile likelihood is maximized with respect to $\boldsymbol{\lambda}$ to obtain $\hat{\boldsymbol{\lambda}}$. Note that this objective function is "focussed" solely on the dispersion parameters. This offers an extension of restricted maximum likelihood (REML) estimation and provides inference for the class of generalized linear mixed models, see Noh and Lee (2007). We show below that in the case of linear mixed models this function is exactly the restricted maximum likelihood.

4.2.2 The linear mixed model case

We illustrate the above estimation framework for the linear mixed model case. While in general the above calculations are approximate, in this case they are exact. Some of the expressions are based on the results shown in Harville (1977). The classical linear mixed effects model assumes:

$$\begin{aligned} f_{\boldsymbol{\beta},\boldsymbol{\lambda}_e}(\mathbf{y}_i|\mathbf{v}_i) &= \mathcal{N}(\mathbf{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\mathbf{v}_i; \boldsymbol{\Sigma}_i), \\ f_{\boldsymbol{\lambda}_v}(\mathbf{v}_i) &= \mathcal{N}(0; \boldsymbol{\Lambda}_i). \end{aligned} \tag{4.11}$$

The design matrix $\mathbf{X}_i$ contains fixed effects for the i^{th} subject, while $\boldsymbol{Z}_i$ is a design matrix for the random effects for the i^{th} subject. Matrices $\boldsymbol{\Sigma}_i$ and $\boldsymbol{\Lambda}_i$ determine the residual variance of $\mathbf{y}_i$ and the variance of random effects $\mathbf{v}_i$ respectively. Note that we have split up $\boldsymbol{\lambda}$ into $\boldsymbol{\lambda}_e$ and $\boldsymbol{\lambda}_v$, the dispersion parameters pertaining to the residual variability $\boldsymbol{\Sigma}_i$ and random effects variability $\boldsymbol{\Lambda}_i$, thereby showing the role of each of the dispersion components. Denote by $\mathbf{V}$ the total (over all subjects) marginal variance-covariance matrix $\mathbf{V} = \boldsymbol{Z}\boldsymbol{\Lambda}\boldsymbol{Z}^T + \boldsymbol{\Sigma}$, where $\mathbf{X}$ is the design matrix for fixed effects obtained by stacking $\mathbf{X}_1$ to $\mathbf{X}_N$, $\boldsymbol{Z} = diag(\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_N)$ is the design matrix of random effects, $\boldsymbol{\Lambda} = diag(\boldsymbol{\Lambda}_1, \ldots, \boldsymbol{\Lambda}_N)$ is the variance covariance matrix of the random effects and $\boldsymbol{\Sigma} = diag(\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_N)$ is the residual variance covariance matrix.

In this case the adjusted profile likelihood $p_v(h)$ becomes:

$$p_v(h) = \ell_M = -\frac{1}{2}\log|2\pi\mathbf{V}| - \frac{1}{2}(\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta})^\top\mathbf{V}^{-1}(\boldsymbol{Y} - \mathbf{X}\boldsymbol{\beta}), \tag{4.12}$$

which is the expression for the marginal likelihood of a linear mixed model. Further, the adjusted profile likelihood $p_{v,\beta}(h)$ is equal to:

$$p_{\beta,v}(h) = \log[f_{\boldsymbol{\lambda}}(\mathbf{y}|\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}})] = \ell_M|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}} - 0.5 \log \left| \frac{\mathbf{X}^\top \mathbf{V}^{-1} \mathbf{X}}{2\pi} \right|, \tag{4.13}$$

which is exactly equal to the classical REML likelihood (see e.g., Verbeke and Molenberghs (2000)) for the linear mixed model.

Note that maximization of (4.5) with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$ is actually the maximization of the marginal likelihood computed by a classical Laplace approximation. In the above h-likelihood procedure, estimation of $\boldsymbol{\beta}$ given $\boldsymbol{\lambda}$ is the same as with a classical Laplace approximation, but the estimation of $\boldsymbol{\lambda}$ is essentially different. The h-likelihood procedure gives an elegant set of IWLS equations for the estimation of the dispersion parameters, which possibly depend on covariates.

4.2.3 Application to the hierarchical generalized linear models

The above theory is applied to generalized linear models with random effects. In this class of models the assumed distribution of the response y_{ij} (conditional on random effects) belongs to the exponential family:

$$f_{\boldsymbol{\beta},\boldsymbol{\lambda}_e}(y_{ij}) = \exp\left[\frac{y_{ij}\theta_{ij} - b(\theta_{ij})}{\lambda_e} + c(y_{ij}, \lambda_e)\right]. \tag{4.14}$$

This distribution is combined with the distribution of the random component, which distribution belongs to the family of conjugate Bayesian distributions for an exponential family, i.e.,

$$f_{\lambda_v}(v_i) = \exp\left[a_1(\lambda_v)v_i - a_2(\lambda_v)b(v_i) + c_2(\lambda_v)\right], \tag{4.15}$$

which can be expressed as the distribution of a pseudo-response ψ_i as follows:

$$f_{\lambda_v}(v_i) = \exp\left[\frac{\psi_i v_i - b(v_i)}{\lambda_v} + c_2(\psi_i, \lambda_v)\right]. \tag{4.16}$$

In this class of models the response is allowed to follow a Gaussian, binomial, Poisson or gamma distribution. Their corresponding conjugate Bayesian distributions are Gaussian, beta, gamma and inverse-gamma, respectively. Note that both λ_v and λ_e are allowed to depend on covariates.

4.2.4 Implementation in the **HGLMMM** package

In this section we document the use of the function **HGLMfit**, together with accompanying functions **HGLMLRTest**, **HGLMLikeDeriv** and **BootstrapEnvelopeHGLM**. First we will describe the use of the fitting function **HGLMfit** and explain the parameters used in the invocation of the function. The following structure of the routine is used:

```
HGLMfit(DistResp = "Normal", DistRand = NULL, Link = NULL,
LapFix = FALSE, ODEst = NULL, ODEstVal = 0,
formulaMain, formulaOD, formulaRand, DataMain, DataRand,
Offset = NULL, BinomialDen = NULL, StartBeta = NULL,
StartVs = NULL, StartRGamma = NULL, INFO = TRUE,
DEBUG = FALSE, na.action, contrasts = NULL, CONV = 1e-04)
```

Below we describe the parameters of the function **HGLMfit** together with the default values in parenthesis.

`DistResp ("Normal"):` The distribution of the response as defined in (4.14) is set by the option `DistResp`. The user can set it to: `Binomial`, `Normal`, `Poisson` or `Gamma`. Note that the name of the distribution must start with a capital letter.

`DistRand (No Default):` The next option `DistRand` specifies the distribution of the random components and should be set as a vector of distribution names from the set: `Beta`, `Gamma`, `IGamma` (inverse-gamma) and `Normal`. For each random component one entry in the vector must be specified. Therefore for a model with two random effects, whereby the first random effect has a normal distribution and the second random effect has a gamma distribution, you should specify the vector c("Normal","Gamma").

Link (Canonical link): The link option is available for a gamma distribution of the response. The choice is either Log or Inverse. For the random variables of a Binomial, Normal or Poisson distribution only the default links are currently available, that is Logit, Identity and Log, respectively.

LapFix (FALSE): Having defined the structure of the model, i.e., (1) the distribution of the response, (2) random effects and (3) the link, one has to specify in the option LapFix whether the joint likelihood (4.2) will be used to estimate fixed effects in the mean structure of the response model, or the effects will be estimated by the adjusted profile likelihood (4.5), which is an approximation to the marginal likelihood. Set LapFix=TRUE for the adjusted profile likelihood and LapFix=FALSE for the joint likelihood.

ODEst (Likelihood based analysis) ODEstVal (0): Next the option ODEst determines whether the dispersion parameter λ_e in (4.14) will be estimated or held fixed. This is set to NULL by default which implies for the Poisson and Binomial distribution for the response that it is fixed. For the Normal and Gamma response, the default option is that it is estimated from the data. Specifying ODEst=TRUE implies that λ_e is estimated, while it is fixed for ODEst=FALSE. Further, the parameters ODEstVal specify either the starting values when λ_e is estimated, or the values used in the estimation when λ_e is set fixed.

formulaMain: The option formulaMain requires a two-sided formula to be specified to determine the structure of the linear predictor of the response, e.g.,
Outcome ~ Fixed.Efffect.1+Fixed.Effect.2+(1|Subject.1).
This specification sets Outcome as the response. Further, in the above example it is specified that there are two fixed effects and a random intercept with subject index Subject.1. Correlated random components are currently not allowed, therefore a structure (1+time|Subject.1) is not valid, instead
(1|Subject.1)+(time|Subject.1) needs to be entered.

`formulaOD formulaRand:` Similarly `formulaOD` specifies the covariates in the residual dispersion (overdispersion) structure by a one sided formula e.g., `~1+time`. Further, `formulaRand` requires a list of formulas determining the dependence of the dispersion parameters of the random components on covariates. For a model with two random effects the following code
`list(one=~1+mean.time,two=~1)` means that the dispersion parameter of the first random component depends on the average time, while the dispersion parameter of the second random component is constant (intercept only model).

`DataMain DataRand:` The option `DataMain` determines which data frame to use for the estimation, i.e., where the data for the mean model and residual dispersion is referred to. Likewise `DataRand` is a list of data frames. They correspond to design matrices for the dispersion parameters of random effects. Therefore for each random effect specified in `formulaMain`, a corresponding dataframe needs to be included.

`Offset (1):` In Poisson regression, an offset variable t is specified in the form $\log(\mu/t) = \eta$, where η is a linear predictor and μ a mean modeled in the Poisson regression. Therefore it is not necessary to log-transform t before entering in the model.

`BinomialDen (1):` In binomial regression one has to use option `BinomialDen` to specify the denominator for the binomial distribution. Suppose you toss a coin ten times, then you have 10 independent trials, and the denominator should then be 10. It is allowed that each observation has a different denominator.

`StartBeta (NULL) StartVs (NULL) StartRGamma (NULL):` The three following options allow to specify the starting values for fixed effects in the mean structure `StartBeta`, random effects in the mean structure `StartVs` (which is a vector of values for all random components together) and dispersion parameters `StartRGamma`. Note that starting values for the residual dispersion (overdispersion) are supplied in `ODEstVal`. Recall that the overdispersion parameter may be fixed or estimated by setting the option `ODEst`. When starting values are not supplied, for the intercept of the mean structure an appropriate sample mean of the response is used, and zeros are used for the other parameters.

`CONV (1e-04):` Setting option `CONV` determines the criterion for convergence, which is computed as the absolute difference between values of all estimated parameters in the previous iteration and in the current iteration.

The function **HGLMfit** returns an object of class `HGLM`. We refer to the help file of the package for the documentation of all elements of this list. The function **HGLMLikeDeriv** takes an object of the class `HGLM` and returns gradient values for fixed effects in the mean structure and dispersion parameters. The function **HGLMLRTest** compares two `HGLM` objects with respect to h-likelihood values, marginal likelihood $p_v(h)$ and REML likelihood $p_{\beta,v}(h)$. Likelihood ratio tests are produced. Finally, the function **BootstrapEnvelopeHGLM** creates a qq-plot of the standardized deviance residuals for the response together with bootstrap envelopes under the assumption that the given model is correct. If the qq-plot falls within the envelopes we can claim that the assumed distribution of the response is reasonable.

There are many packages/routines/programs for analysis of the hierarchical models. The algorithm based on the h-likelihood approach offers a wider choice for the random effects distributions. Further, dispersion components of random effects as well as an overdispersion parameter can be modeled as a function of covariates. This is combined with relatively modest computational requirements.

4.3 Analysis of examples

In this section we will present the analysis of the three data sets, also analyzed in Lee *et al.* (2006), using the package **HGLMMM**.

4.3.1 Salamander data

In McCullagh and Nelder (1989) a dataset on salamander mating was presented. The dependent variable represents the success of salamanders mating. There were 20 males and 20 females in each of the 3 experiments coming from the two populations called whiteside (denoted by W) and roughbutt (denoted as R). Salamanders were paired for mating six times with individuals of their own kind and from the other population. In total 360 observations were generated. Denote as μ_{ijk} the probability of successful mating. We will consider a model with crossed-random effects (also known as a multi-membership model) to analyze this dataset. The mean structure has the following expression:

$$\log\left(\frac{\mu_{ijk}}{1-\mu_{ijk}}\right) = \text{Intercept} + \text{TypeF} + \text{TypeM} + \text{TypeF} * \text{TypeM} + v_i + v_j, \quad (4.17)$$

where TypeF is the type of female (R or W), TypeM is the type of male (R or W), while v_i are the random effects corresponding to males and v_j are the random effects corresponding to females. The following analyses were also performed in Lee *et al.* (2006) (page 194-195). The program below is used to fit this model. The design matrices for the dispersion components of male and female random effects contain only an intercept and are created by the command below:

```
R> RSal <- data.frame(int = rep(1, 60))
```

The following program is invoked to perform the analysis:

```
R> modBin <- HGLMfit(DistResp = "Binomial", DistRand = c("Normal", "Normal"),
                    Link = "Logit", LapFix = TRUE, ODEst = FALSE,
                    ODEstVal = c(0), formulaMain = Mate ~ TypeF + TypeM
                    + TypeF * TypeM + (1|Female) + (1|Male),
                    formulaOD = ~ 1, formulaRand = list(one = ~ 1, two = ~ 1),
                    DataMain = salamander,DataRand = list(RSal, RSal),
                    BinomialDen = rep(1, 360), INFO = TRUE, DEBUG = FALSE)
```

A brief description of the model can be displayed by `modBin`, that uses the `print` method for objects of class `HGLM`. Printing of the object `modBin` gives the following output:

```
===== HGLM Model Information =====
Response Distribution: Binomial
Random Effect 1 : Normal / Female
Random Effect 2 : Normal / Male
Link: Logit
Overdispersion Structure is Fixed
Estimation of fixed effects: Laplace Approximation
Dataset used: salamander
Model Equation:
        Mate ~ TypeF + TypeM + TypeF * TypeM + (1 | Female) + (1 | Male)

Overdispersion Equation:
        ~1

Dispersion Equation(s):
Component 1 :~1

Component 2 :~1
```

The output gives information about the response distribution, the number of random effects, their distribution and the subject index. Further, the link function is reported. Next, information is contained whether overdispersion (residual dispersion) is fixed or estimated. The method of estimating the fixed effects in the regression equation is reported, as well as the model equation, the overdispersion equation and the dispersion equations.

The detailed results of the fit can be obtained by the command:

```
R> summary(modBin, V=TRUE)
```

If option `V=TRUE` is omitted, the results for the random effects are not printed (as in our example). Output for our object `modBin` is as follows:

```
===== Fixed Coefficients - Mean Structure =====

              Estimate Std. Error Z value Pr(>|Z|)
(Intercept)     1.0433     0.4036   2.585  0.00974 **
TypeFW         -3.0055     0.5260  -5.714 1.10e-08 ***
TypeMW         -0.7290     0.4741  -1.538  0.12413
TypeFW:TypeMW   3.7137     0.5758   6.449 1.12e-10 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

The above output gives summary of fixed effects in the mean structure.

```
===== Overdispersion Parameters Fixed =====
            Fixed Value
(Intercept)           0
```

The overdispersion parameter is held fixed at 1, which is exp(0). Note that the parameters reported for overdispersion and dispersion pertain to the logarithm of the parameters. The output shown below reports values for the parameters in the dispersion structure of random effects. To obtain the value of λ_v one again needs to exponentiate the reported estimate.

```
===== Dispersion Parameters Estimated =====

Dispersion Component: Female
            Estimate Std. Error Z value Pr(>|Z|)
(Intercept)   0.3183     0.4459   0.714    0.475

Dispersion Component: Male
            Estimate Std. Error Z value Pr(>|Z|)
(Intercept)   0.1863     0.4602   0.405    0.686
```

Finally the output shown below reports on the values of the extended likelihood (H-likelihood), $p_v(h)$ which is denoted as marginal likelihood, $p_{\beta,v}(h)$ REML likelihood, and the component $f_{\boldsymbol{\beta},\boldsymbol{\lambda}}(y_{ij}|\mathbf{v}_i)$ C-likelihood.

```
===== Likelihood Functions Value =====
H-likelihood       : -287.8858
Marginal likelihood: -209.3600
REML likelihood    : -209.5131
C-likelihood       : -136.2331
```

The H-likelihood should be used to select the random effects, the marginal likelihood is the choice for the mean structure simplification, while REML likelihood can be used for inference on the variance components. We obtain similar estimates as in Lee *et al.* (2006).

4.3.2 Cake data

The chocolate cakes preparation experiment was conducted at Iowa State College. Three recipes for preparing the batter were compared. Cakes were prepared at 6 different baking temperatures, which ranged from 175 up to 225 degrees of Centigrade. For each recipe 6 cakes were prepared baked at different temperatures, therefore 18 cakes were baked all together and they are referred to as replication. There were 15 replications, therefore in total 270 cakes were baked. Further, in each replication there are 3 different recipes. To cope with such a design we include two random effects, v_i representing the replication, and v_{ij} for the j^{th} recipe within the replication. Finally e_{ijk} stands for the error term associated with each cake. The linear predictor of the model is defined as follows:

$$\eta_{ijk} = \text{intercept} + \text{recipe}_j + \text{temp}_k + \text{recipe}_j \cdot \text{temp}_k + v_i + v_{ij}, \tag{4.18}$$

whereby $i = 1, \ldots, 15$ refers to the replicate, $j = 1, 2, 3$ is the recipe and $k =, 1 \ldots, 6$ to the temperature. The analyzes presented can also be found in Lee *et al.* (2006)

(pages 163-166). We will consider two models. In the first model we consider the breaking angle as a normally distributed response, while in the second model we assume that it has a gamma distribution. Further, in both models we assume a normal distribution for both random effects v_i and v_{ij}. The following syntax is used for a gamma response model:

```
R> cake$repbatch <- 100 * cake$Replicate + cake$Batch
R> R1Cake <- data.frame(int = rep(1, 15))
R> R2Cake <- data.frame(int = rep(1,45))
R> modCake2 <- HGLMfit(DistResp = "Gamma", DistRand = c("Normal", "Normal"),
            Link = "Log", LapFix = FALSE, ODEst = TRUE, ODEstVal = c(0),
            formulaMain = Angle ~ as.factor(Recipe) + as.factor(Temperature)
                + as.factor(Recipe) * as.factor(Temperature) +
                (1|Replicate) + (1|repbatch),
            , formulaOD = ~1, formulaRand = list(one = ~ 1, two = ~ 1),
            DataMain = cake, DataRand = list(R1Cake, R2Cake), Offset = NULL,
            , INFO = TRUE, DEBUG = FALSE)
```

The estimated gamma model yields the following (marginal) likelihood value at maximum:

```
===== Likelihood Functions Value =====
H-likelihood       : -676.3907
Marginal likelihood: -808.0586
REML likelihood    : -848.9244
C-likelihood       : -754.2644
```

The normal model yields a (marginal) likelihood value of −819.54. This value is lower than for a gamma model, and thus preferable according to the AIC criterion. Following Lee and Nelder (1998) we create the model checking plots for both models using the following code. First some graphics manipulation parameters are set to control the layout of the graphs:

```
op<-par(mfrow=c(2,2),
    oma = c(1,1,2,1),
    mar = c(3,3,4,1) +1.2
    )
```

Next we copy the standardized deviance residuals as suggested by Lee *et al.* (2006) (page 52) from the model object `modCake2`, and compute the linear predictor of the model, i.e., η together with mean and transformed mean as indicated in Nelder (1990). This is done as follows:

```
R> res <- modCake2$Details$StdDevianceResidualY
R> xmat <- model.matrix(~ as.factor(Recipe) + as.factor(Temperature) +
            as.factor(Recipe) * as.factor(Temperature), data = cake)
R> eta <- xmat%*%modCake2$Results$Beta
R> mu <- exp(eta)
R> mu <- log(mu)
```

After a few steps, standardized deviance residuals and absolute standardized deviance residuals can be plotted against scaled fitted values, together with a loess curve. In addition a qq-plot of these residuals and histogram are created.

The commands below produce scaled fitted values plot against residuals:

```
R> plot(mu, res, pch = 18, col = "black" xlab = "Scaled Fitted Values",
       ylab = "Deviance Residuals")
R> los <- loess.smooth(mu, res, span = 1/2, degree = 1,
    family = c("symmetric", "gaussian"), evaluation = 50)
R> lines(los$x, los$y, col = "black", lwd=2)
```

Next, we create a plot of scaled fitted values against absolute residuals:

```
R> plot(mu, abs(res), pch = 18, col = "black", xlab = "Scaled Fitted Values",
       ylab = "Absolute Deviance Residuals")
R> los <- loess.smooth(mu, abs(res), span = 1/2, degree = 1,
    family = c("symmetric", "gaussian"), evaluation = 50)
R> lines(los$x,los$y, col="black", lwd=2)
```

The qq-plot and histogram of residuals:

```
R> qqnorm(res, pch=18)
R> abline(h=0, v=0)
R> abline(a=0, b=1, lty=3)

R> breaks <- seq(-4, 4, by = 0.8)
R> hist(res, breaks = breaks, xlab = "Deviance Residuals",
            main = "Histogram")
```

Additional commands provide a title:

```
R> par(op)
R> mtext("Diagnostics for Cake Model Gamma",
      side = 3, line = 1.5, font = 2, cex = 2, col = 'black')
R> par(op)
R> dev.off()
```

The above code produces the diagnostic plots for the gamma model presented in Figure 4.1. Further we fitted a gamma model without interaction $\text{recipe}_j * \text{temp}_k$ and compared it with the original model using the likelihood ratio test:

```
R> HGLMLRTest(modCake3, modCake2)
```

The following output is obtained:

```
H-likelihood of model 2 is higher
Marginal likelihood comparison:
LR test p-value: 0.5034955
LR test statistics: 9.304224
LR difference df: 10
REML likelihood comparison:
LR test p-value: NA
LR test statistics: 29.88638
LR difference df: 0
```

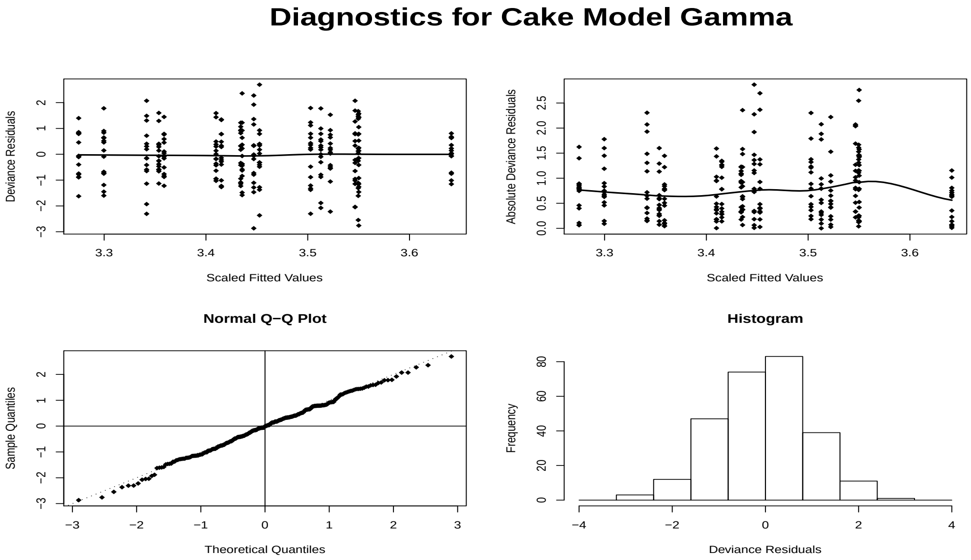


Figure 4.1: Cake data: Diagnostic plots for the gamma model.

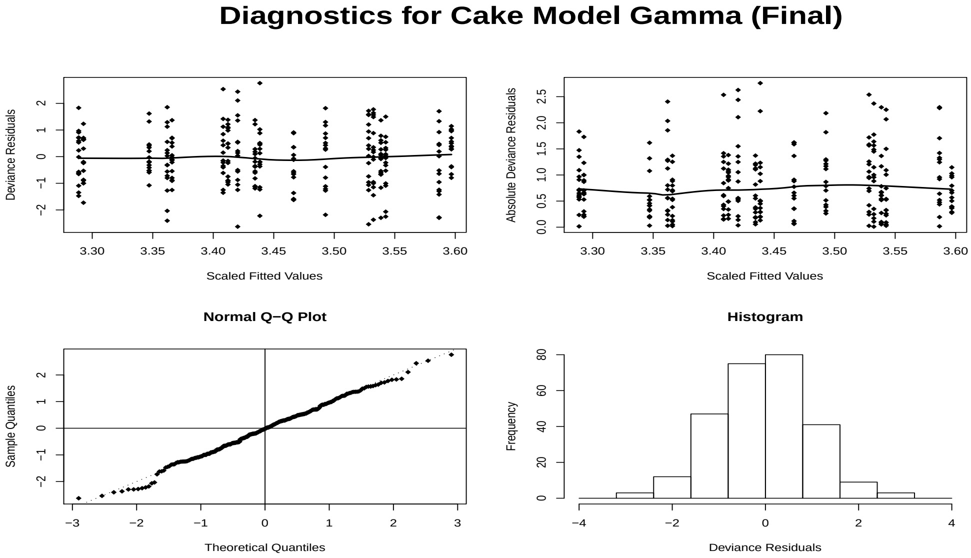


Figure 4.2: Cake data: Diagnostic plots for the final gamma model.

The chi-squared test statistics with 10 degrees of freedom indicate that there is not enough evidence for the inclusion of the interaction term into the model (p value:

0.5). Therefore the simplified model is our choice. According to the diagnostics plot on Figure 4.2 the model seems to be fitting well. Lee *et al.* (2006) analyzed this data using linear mixed model for the original, as well as the transformed response. They also indicated that the gamma model is preferable by the AIC criterion.

4.3.3 Rat data

Thirty rats were treated with one of three chemotherapy drugs. White blood cell counts (W) and red blood cell counts (R) were taken at each of four different time points. We perform the same analysis as in Lee *et al.* (2006) (pages 224-229). We will focus here on quasi-Poisson model with normal random effects. First we create the design matrix for the dispersion parameter of the random component:

```
R> Rrat <- data.frame(WBC = tapply(rat$WhiteBloodCells, rat$Subject, mean),
                      RBC = tapply(rat$RedBloodCells, rat$Subject, mean))
```

The model can be fitted by:

```
R> modRat2 <- HGLMfit(DistResp = "Poisson", DistRand = c("Normal"),
              Link = "Log", LapFix = FALSE, ODEst = TRUE,
              ODEstVal = c(0), formulaMain = Y ~ WhiteBloodCells +
              RedBloodCells + as.factor(Drug) + (1|Subject),
              formulaOD = ~ 1, formulaRand = list(one = ~ WBC + I(WBC ^ 2)),
              DataMain = rat, DataRand = list(Rrat),
              INFO = TRUE, DEBUG = FALSE)
```

The crucial option here is `ODEst=TRUE`, which requests a quasi-Poisson to be fitted instead of a likelihood based Poisson model. Note that a summary of the likelihood values is not valid now, as we are not using a likelihood based technique but extended quasi-likelihood is invoked. By a similar manipulation as in Section 4.3.2, we ask for the diagnostic plots of the model. Figure 4.3 presents diagnostic plots for the quasi-Poisson model for the residual error term $(\mathbf{y}|\mathbf{v})$, while Figure 4.4 gives diagnostics for the choice of the model for the dispersion of the random term $(\mathbf{v})$. The corresponding diagnostic plots were also presented by Lee *et al.* (2006).

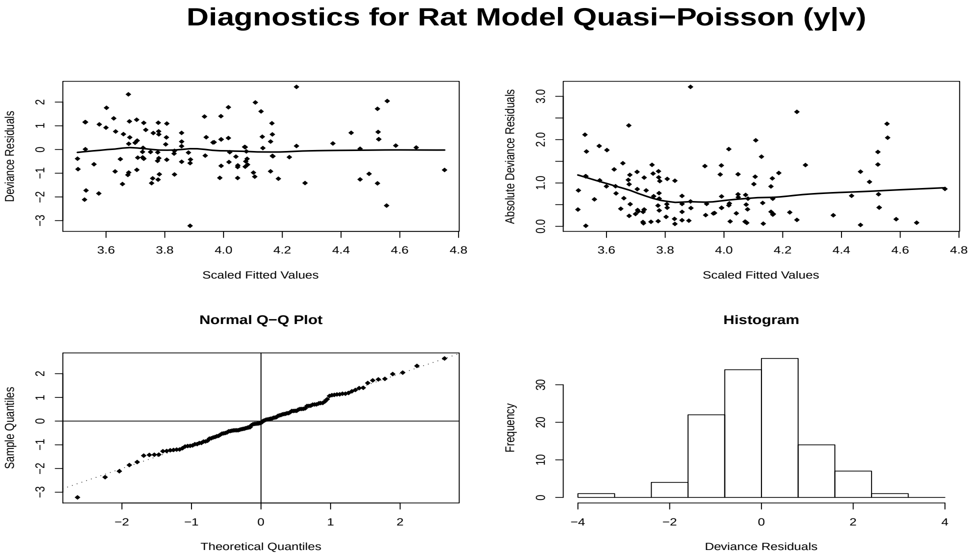


Figure 4.3: Rat data: Diagnostic plots for the quasi-Poisson model (residual component).

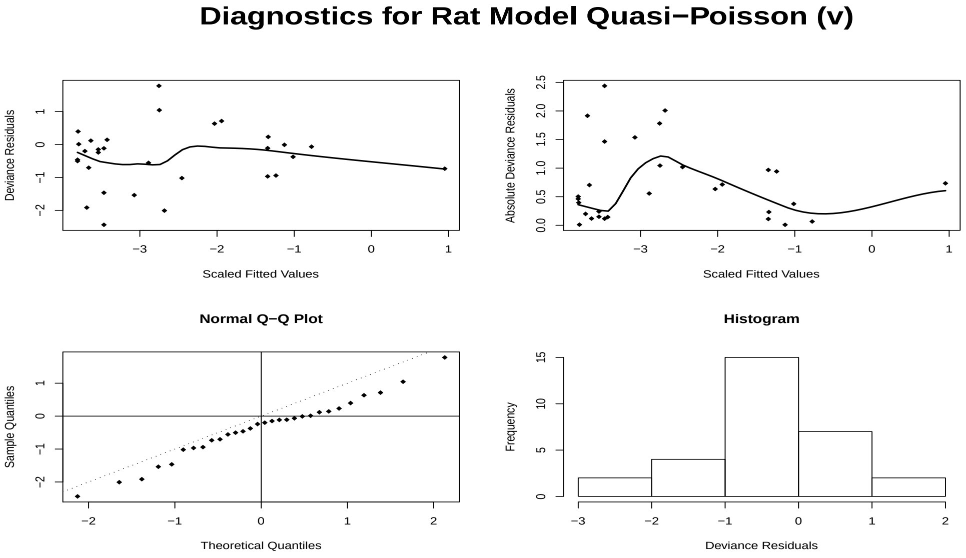


Figure 4.4: Rat data: Diagnostic plots for the quasi-Poisson model (dispersion of random component).

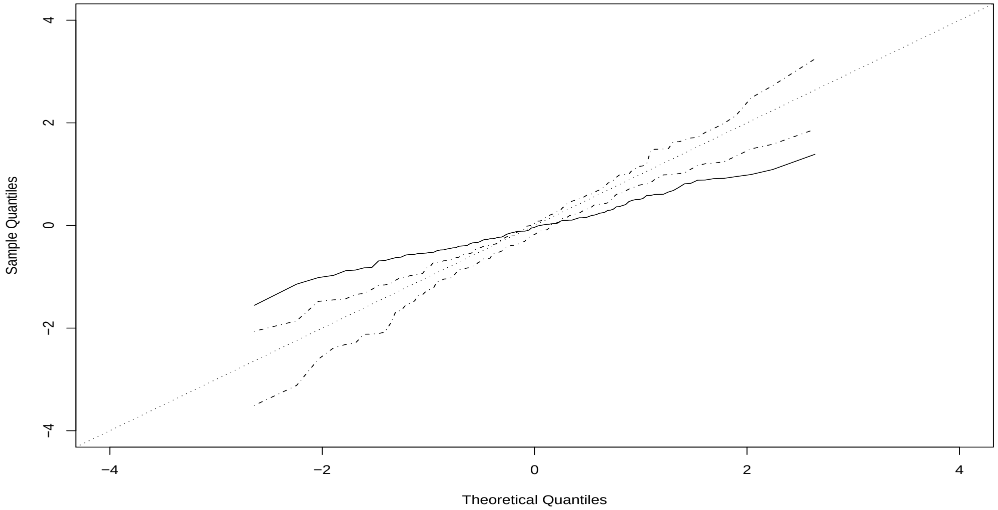


Figure 4.5: Rat data: Diagnostic plot for the likelihood based Poisson model.

Figure 4.5 presents the qq-plot for the standardized deviance residuals for the Poisson model (likelihood based) of the same structure as the above quasi-Poisson model, together with 95% bootstrap envelopes. Clearly the Poisson model does not fit well. The code to generate the plot is the following:

```
R> modRat1 <- HGLMfit(DistResp = "Poisson", DistRand = c("Normal"),
              Link = "Log", LapFix = FALSE, ODEst = FALSE,
              ODEstVal = c(0), formulaMain = Y ~ WhiteBloodCells +
              RedBloodCells + as.factor(Drug) + (1|Subject),
              formulaOD = ~ 1, formulaRand = ~ list(one=~1),
              DataMain = rat, DataRand = list(Rrat),
              INFO = TRUE, DEBUG = FALSE)

R> BootstrapEnvelopeHGLM(modRat1, 19, 9999)
```

4.4 Alternative software for hierarchical data

The methods described in this paper are applied to model hierarchy in the outcomes (resulting in unobserved heterogeneity). There are numerous statistical programs and packages to address this type of modeling. Here we would like to mention a few. Note that that is not our intention to be exhaustive, but rather to focus on most popular programs and which we are acquainted with.

First of all, to our knowledge prior to this **R** package the h-likelihood algorithms were implemented only in **GENSTAT** (Payne *et al.*, 2009) software. The capabilities of **HGLMMM** and **GENSTAT** h-likelihood methods are similar. Extra flexibility is provided by **HGLMMM** by the ability to specify several random components, allowing for a different distribution of each component. This is not available in the **GENSTAT** program. On the other hand **GENSTAT** offers a second order Laplace approximation to estimate variance components, which is not implemented in **HGLMMM**.

To analyze the salamander example the software needs to allow for a crossed random effects design. Gaussian quadrature methods are not applicable in this case as the dimension of integration is often too large. The **R** package **lme4** (Bates and Maechler, 2009a) with the function `lmer` allows for the analysis of this example with an ordinary Laplace approximation. The advantage of using **HGLMMM** is the ability to include covariates in the dispersion part of the model. This is not allowed in `lmer`. Packages **lme4** and **HGLMMM** differ in the way the dispersion parameters are estimated. The **HGLMMM** package uses the objective function which is equivalent to the REML approach in linear mixed models. The package **lme4** is however faster and more efficient. Note that **lme4** is a quite popular **R** package for the analysis of longitudinal data. The salamander data can also be analyzed by the PQL method of Breslow and Clayton (1993) as implemented e.g., in **SAS** (**SAS** Institute Inc., 2008) PROC GLIMMIX, which allows the dispersion parameters to depend on categorical covariates.

In the cake example we analyzed the breaking angle of cakes assuming a normal distribution. This analysis can be reproduced in the **R** package **nlme** (Pinheiro *et al.*, 2009) or **lme4** using a random effects model or a linear model with correlated errors. Also the **SAS** procedures MIXED, NLMIXED or GLIMMIX can be used. When the response follows a gamma distribution **SAS** procedures GLIMMIX and NLMIXED can be used, but both of them assume the random effects follow a normal distribution. On the other hand, in **HGLMMM** we easily combine non-normal responses with a non-normal random effects distribution. Such a combination can be done in **SAS** PROC NLMIXED only by using the trick of Liu and Yu (2008) and requires some additional programming. Finally, note that **SAS** PROC NLMIXED is a likelihood based procedure, while **HGLMMM** allows the utilization of the extended quasi likelihood, with a overdispersion parameter varying with covariates.

In the rat example an extended quasi likelihood analysis is presented on the number of cancer cell colonies, with the variance of the random effect distribution depending on covariates. Now it is difficult to find standard software which could repeat such an analysis. **SAS** PROC GLIMMIX comes closest, allowing for overdispersion in Poisson distribution and the variance of random effect to depend on a categorical covariate.

In summary, the package **HGLMMM** blends the distribution of the response from an exponential family with a random effects distribution from the conjugate Bayesian distributions. Further, difficult designs can be handled such as multi-membership designs or more than 2-levels in the hierarchical models. On top of that one can put covariates in mean and (over)dispersion structure. All of these features are not available in one package/software to our knowledge. The limitation of **HGLMMM** is the necessity to assume independent random components, this assumption can be relaxed in **R** packages **lme4**, **nlme** and **SAS** PROC MIXED, NLMIXED, GLIMMIX .

4.5 Future improvements

In future work we would like to adapt the codes to use **Matrix** (Bates and Maechler, 2009b) package to save memory and to avoid problems with large datasets. Different links for the binomial response such as `Probit` and `CLogLog` need to be implemented as well. In addition, future work will focus on the introduction of the correlation between random effects. In Molas and Lesaffre (2010) the routines were extended to handle hurdle models for count data.

References

Barndorff-Nielsen, O. (1980). Conditionality resolutions. *Biometrika*, **67**, 293–310.

Barndorff-Nielsen, O. (1983). On a formula for the distribution of the maximum likelihood estimator. *Biometrika*, **70**, 343–365.

Bates, D. and Maechler, M. (2009a). *lme4: Linear Mixed-Effects Models Using S4 Classes*. **R** package version 0.999375-32.

Bates, D. and Maechler, M. (2009b). *Matrix: Sparse and Dense Matrix Classes and Methods*. **R** package version 0.999375-31.

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

Cox, D. and Hinkley, D. (1974). *Theoretical Statistics*. Chapman and Hall, London.

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

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

Henderson, C., Kempthorne, O., Searle, S., and Krosigk, C. (1959). The estimation of environmental and genetic trends from records subject to culling. *Biometrics*, **15**, 192–218.

Lee, Y. and Nelder, J. (1996). Hierarchical generalized linear models (with discussion). *Journal of the Royal Statistical Society B*, **58**, 619–678.

Lee, Y. and Nelder, J. (1998). Generalized linear models for the analysis of quality-improvement experiments. *Canadian Journal of Statistics*, **26**, 95–105.

Lee, Y. and Nelder, J. (2001). Hierarchical generalised linear models: A synthesis of generalised linear models, random-effect models and structured dispersions. *Biometrika*, **88**, 987–1006.

Lee, Y., Nelder, J., and Pawitan, Y. (2006). *Generalized Linear Models with Random Effects*. Chapman & Hall / CRC, Boca Raton.

Lesaffre, E. and Spiessens, B. (2001). On the effect of the number of quadrature points in a logistic random-effects model: An example. *Applied Statistics*, **50**, 325–335.

Liu, L. and Yu, Z. (2008). A likelihood reformulation method in non-normal random effects models. *Statistics in Medicine*, **27**, 3105–3124.

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models*. Chapman and Hall, London.

Molas, M. and Lesaffre, E. (2010). Hurdle models for multilevel zero-inflated data via h-likelihood. *Statistics in Medicine*, **29**, 3294–3310.

Molenberghs, G. and Verbeke, G. (2005). *Models for Discrete Longitudinal Data*. Springer-Verlag, New York.

Nelder, J. (1990). Nearly parallel lines in residual plots. *The American Statistician*, **44**, 221–222.

Nelder, J. and Wedderburn, R. (1972). Generalized linear models. *Journal of Royal Statistical Society A*, **135**, 370–384.

Noh, M. and Lee, Y. (2007). REML estimation for binary data in GLMMs. *Journal of Multivariate Analysis*, **98**, 896–915.

Pawitan, Y. (2001). *In All Likelihood: Statistical Modelling and Inference Using Likelihood.* Clarendon Press, Oxford.

Payne, R. W., Murray, D. A., Harding, S. A., Baird, D. B., and Soutar, D. M. (2009). *GenStat for Windows (12th Edition) Introduction.* VSN International, Hemel Hempstead, UK.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., and the **R** Core team (2009). *nlme: Linear and Nonlinear Mixed Effects Models.* **R** package version 3.1-93.

R Development Core Team (2009). ***R**: A Language and Environment for Statistical Computing.* **R** Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

SAS Institute Inc. (2002-2008). ***SAS/STAT** Software, Version 9.2.* Cary, NC.

Severini, T. (2000). *Likelihood Methods in Statistics.* Oxford Univeristy Press, Oxford.

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

Chapter
5

Hurdle Models for Multilevel Zero-Inflated Data via H-likelihood

Based on:
Molas M. and Lesaffre E. (2010). Hurdle models for multilevel zero-inflated data via h-likelihood. *Statistics in Medicine* **29** pp. 3294-3310

Abstract

Count data often exhibit overdispersion. One type of overdispersion arises when there is an excess of zeros in comparison to the standard Poisson distribution. Zero-inflated Poisson and hurdle models have been proposed to perform a valid likelihood based analysis to account for the surplus of zeros. Further, data often arise in clustered, longitudinal or multiple-membership settings. The proper analysis needs to reflect the design of a study. Typically random effects are used to account for dependencies in the data. We examine the h-likelihood estimation and inference framework for hurdle models with random effects for complex designs. We extend the h-likelihood procedures to fit hurdle models, thereby extending h-likelihood to

truncated distributions. Two applications of the methodology are presented.

5.1 Introduction

While counts are often modeled with a Poisson distribution, in medical applications they usually exhibit overdispersion. One type of overdispersion arises where there is an excess of zeros compared to the Poisson distribution. To deal with this type of overdispersion the zero-inflated Poisson (ZIP) model has been suggested, see Singh (1963) for an early reference. A competitor to the ZIP model or its extensions is the hurdle model first suggested by Cragg (1971). The two above models have been extended also to correlated data structures, see e.g. Min and Agresti (2005) and references therein. In this paper we focus on further generalizations of the hurdle model and suggest another way of estimating and testing its parameters.

Zero-inflated data can be found in e.g. dental studies where caries experience (see Mwalili *et al.* (2008)) is modeled, adverse events studies or occupational accidents studies (see Min and Agresti (2005)), etc.. Zero-inflated counts can also occur in clustered data or repeated measures designs. In that case the dependence of the counts needs to be dealt with. This could be done by the inclusion of random effects in a likelihood based analysis. In Min and Agresti (2005) an extension of the ZIP and the Poisson-hurdle model to dependent data is proposed, but they focus on the latter model. Estimation and inference is based on the marginal likelihood.

In this paper we propose to use the h-likelihood approach Lee and Nelder (1996, 2001) to estimate and test the parameters of a hurdle model in a repeated measurements context. Our approach allows for multi-membership and multi-level designs. Further, other than normal distributions may be taken for the random effects but, the approach is restricted at this moment to independent random effects distributions. In the basic version the truncated part of the hurdle model is assumed to be Poisson. Overdispersion in this part can be accommodated with by the inclusion of an extra random effect. Parameter estimation is performed by an Iterative Weighted Least Squares (IWLS) algorithm. Inference is based on h-likelihood and adjusted profile likelihood functions. This allows for a restrictive maximum likelihood (REML) inference for variance components in the context of generalized linear mixed models.

In Section 5.2 we describe the two motivating data sets. The hurdle model is described in Section 5.3. The h-likelihood approach is extensively reviewed in Section 5.4. In this review we focus on the numerical aspects of the h-likelihood approach. The extension of the h-likelihood approach to the hurdle model and in a multi-level context is presented in Section 5.5. Most of the technical details of the developments are, however, deferred to the Appendix. In Section 5.6 the motivating examples are analyzed with the proposed h-likelihood approach. Concluding remarks are given in Section 5.7.

5.2 Motivating data sets

In this section we describe the two motivating data sets, both are examples of zero-inflated repeated measures data. The first example has been previously analyzed by Min and Agresti (2005). The second data set results from an intervention study to examine the effect of Tai-Chi Chuan exercises on the frequency of falls in elderly people.

5.2.1 Adverse events study

The data were obtained from a clinical trial comparing two treatments for a particular disease on the number of episodes of a certain adverse event. This example was first described by Min and Agresti (2005) on simulated data. 118 patients were randomly allocated to one of the two treatment arms. The response of the study is the frequency of episodes between consecutive visits. About 83% of the counts were zeros. Table 5.1 displays the distribution of the number of episodes of the adverse event observed between each of the six visits.

		Number of Adverse Events						
Visit	Treatment	0	1	2	3	4	5	6
1	0	51	8	0	0	0	0	0
	1	57	1	1	0	0	0	0
2	0	52	4	2	0	1	0	0
	1	47	9	3	0	0	0	0
3	0	51	6	2	0	0	0	0
	1	47	8	3	1	0	0	0
4	0	53	4	2	0	0	0	0
	1	44	7	5	2	0	0	1
5	0	52	4	3	0	0	0	0
	1	41	9	3	0	5	1	0
6	0	53	4	2	0	0	0	0
	1	42	5	5	3	2	1	1

Table 5.1: Frequency table for the Min & Agresti Min and Agresti (2005) dataset

5.2.2 Tai-Chi Chuan intervention study

Our second example is an intervention study concerned with the number of falls in elderly people. Falls are a common problem among old aged people. Between 55% to 70% of fall incidents result in physical injury. In the Department of General Practice of the Erasmus Medical Center at Rotterdam (the Netherlands) a study was set up to examine the effect of Tai-Chi Chuan exercises on reducing the frequency of falls in healthy elderly people living at home and who are at an increased risk of falling. This randomized and partially blinded clinical trial ran from February 2004 to April 2006 and enrolled in total 269 elderly people. General practitioners invited the elderly people but the GPs were not informed who attended the Tai-Chi Chuan exercises. Participants were followed up for about a year yielding a count each month. Here we lumped together the time periods into baseline and four periods of follow up, resulting in a maximum of five counts for each subject. We used as response the number of days a fall was recorded at baseline, after three, six, nine and twelve months. Further, the participants were trained in groups. The statistical analysis needs to take into account that people pertaining to the same training group are clustered. Furthermore, the repeated counts are correlated and show overdispersion. Table 5.2 presents the histograms of the frequency of falls in the two treatments over the 5 follow-up periods. It was also of interest to establish the effect of some covariates that were recorded at baseline. Here we consider age, sex, height, weight and alcohol use.

5.3 The hurdle model

There are two challenges associated with the analysis of the adverse events data described in Section 5.2. Namely the statistical model needs to account for (a) the excess of zero counts which might depend on covariates and (b) the fact that the recorded counts of the same individual are correlated. In this section we describe in detail the Poisson-hurdle model which allows for a proper likelihood based analysis when the correlated counts exhibit zero-inflation or zero-deflation. We also contrast the Poisson-hurdle model to the zero-inflated Poisson model.

The Poisson-hurdle model is composed of two submodels in such a way that the total likelihood is decomposed into two separate likelihoods which can be maximized independently. The proportion of zero counts is described by a binary model with dependence on covariates often modeled with the logit link, but also the probit and log-log link are in use. Second, the positive counts are modeled by a truncated Poisson distribution at zero. Thus the Poisson-hurdle model can be expressed as a mixture of two non-overlapping Poisson distributions: a degenerate Poisson with mean zero and a truncated Poisson defined on positive integers.

		Number of Falls															
Visit	Treatment	0	1	2	3	4	5	6	7	9	10	11	12	15	16	21	24
1	Standard	83	23	2	2	2	0	2	1	0	0	0	0	1	0	0	0
	Tai - Chi	90	24	4	2	1	2	2	0	0	0	0	1	0	0	0	0
2	Standard	77	21	8	3	3	0	0	0	0	0	0	0	0	0	0	1
	Tai - Chi	89	17	8	1	1	2	0	1	1	0	1	0	0	1	0	0
3	Standard	79	16	11	1	1	0	0	0	0	0	1	0	0	1	0	0
	Tai - Chi	87	21	5	3	0	0	1	0	0	0	0	0	0	0	1	0
4	Standard	78	20	3	0	0	1	0	0	0	1	0	0	0	0	0	0
	Tai - Chi	79	32	7	0	0	0	0	0	0	0	0	0	0	0	0	0
5	Standard	79	13	5	2	1	1	1	0	1	0	0	0	0	0	0	0
	Tai - Chi	95	12	4	0	1	0	0	0	0	0	0	0	0	0	0	0

Table 5.2: Frequency table for the Tai-Chi Chuan dataset

A competitor to the Poisson-hurdle model is the zero-inflated Poisson model. When there is overdispersion of the non-zero count part both the ZIP model as well as the Poisson-hurdle model can be extended, e.g. the zero-inflated negative binomial model (ZINB) is obtained by assuming that the mean of the Poisson distribution has itself a gamma distribution. The ZIP distribution can be expressed as a mixture of a degenerate Poisson with mean zero and a standard Poisson distribution. In this class of models the components of the mixture overlap. As a result, for mixture weights that are positive and sum to one, the ZIP model can handle only zero-inflation, see Mullahy (1986). Indeed, when the ZIP model is applied to zero-deflated data, the estimates of the ZIP model become unstable and the mixture weight for the degenerate distribution tends to zero. The problem is amplified when covariates have an impact on the zero-inflation or deflation. The hurdle model does not suffer from this handicap, though and can model both types of deviations from the Poisson distribution. These issues were exemplified by Min and Agresti (2005) in a cross-sectional setting but their findings carry over to clustered and longitudinal designs.

When no structured covariates are involved there exists a correspondence between the ZIP and Hurdle model. Parameters of one model can be expressed as a function of the parameters of the other. When structured covariates are involved e.g. continuous or splines, this correspondence seems not to hold. Further, it is not clear when the relations hold in the repeated measures designs.

The probability that a particular count occurs at the j^{th} $(j = 1, \ldots, n_i)$ occasion for the i^{th} $(i = 1, \ldots, N)$ individual under the Poisson-hurdle model is given by:

$$\begin{aligned} P(Y_{ij} = 0) &= p_{ij}, \\ P(Y_{ij} = k) &= (1 - p_{ij}) \cdot \frac{e^{-\mu_{ij}} \frac{\mu_{ij}^k}{k!}}{1 - e^{-\mu_{ij}}}. \end{aligned} \tag{5.1}$$

Further, the probabilities p_{ij} and the means of the truncated Poisson distribution, μ_{ij}, might depend on covariates. Inclusion of random effects v_{1i} and v_{2i} extends the Poisson-hurdle model to repeated measures data (as introduced in Min and Agresti

(2005)) as follows:

$$\begin{aligned} \text{logit}(1-p_{ij}) &= \mathbf{x}_{1ij}^T\boldsymbol{\beta}_1 + v_{1i}, && \text{Bernoulli Part} \\ \log(\mu_{ij}) &= \mathbf{x}_{2ij}^T\boldsymbol{\beta}_2 + v_{2i}. && \text{Truncated Poisson Part} \end{aligned} \tag{5.2}$$

Further, the random effects in (5.2) are assumed to follow a bivariate distribution. In case v_{1i} and v_{2i} are independent, the likelihood factorizes into the sum of the likelihood pertaining to the binary part and the likelihood of the truncated Poisson part of the hurdle model. Thus in that case the two likelihood parts can be maximized separately. Of course when the v_{1i} and v_{2i} are correlated, joint maximization of the two mixture components might be more appropriate.

In the Poisson-hurdle model each individual belongs to the one of the two components of the mixture, i.e. zero-part or truncated Poisson part. The repeated measures Poisson-hurdle model implies that the same individual might belong to the zero part at one timepoint, while it can change the class at another visit e.g. when a positive count is observed. Therefore the repeated measures Poisson-hurdle model is defined for each timepoint.

In Min and Agresti (2005) two ways of dealing with repeated measures hurdle models are described. First, under the assumption of bivariate Gaussian random effects, numerical integration is performed to obtain the marginal likelihood, and thereafter the marginal likelihood is maximized. Second, when an unspecified discrete distribution for the random effects is assumed, a nonparametric maximum likelihood (NPMLE) method is employed, relying on the maximization of the marginal likelihood as well. Both methods might become computationally prohibitive for high dimensional random effects distributions. Alternatively, one could refer to Bayesian methodology using MCMC sampling. But MCMC sampling requires though a lot of computational power and as a result the time until convergence might be quite long. Further, assessing convergence remains a difficult task. Here, we explore the h-likelihood approach for parameter estimation and inference in a hurdle model with a complex random effects structure.

5.4 H-likelihood

5.4.1 Introduction

Regression models typically have two kinds of parameters: (1) mean structure parameters, $\boldsymbol{\beta}$, often called regression or fixed effects parameters and (2) dispersion parameters, denoted by $\boldsymbol{\lambda}$. In a longitudinal setting random effects, denoted by $\mathbf{v}$, are often invoked to model the covariance structure between responses. The random effects describe the subject-specific deviation of the individual with respect to the population averaged evolution. Note that in a mixed effects setting there are two types of dispersion parameters: (a) $\boldsymbol{\lambda}_v$ denotes the dispersion parameter(s) pertaining to the random effects and (b) $\boldsymbol{\lambda}_e$ denotes the dispersion parameter(s) pertaining to the error distribution. A popular way to estimate parameters of a longitudinal random effects model is to integrate out the random effects from the likelihood leading to the integrated or marginal likelihood and maximize the marginal likelihood with respect to the fixed effects parameters. Estimates of the random effects are obtained in the marginal approach using an Empirical Bayes argument.

In this section we introduce the h-likelihood approach of Lee and Nelder (1996) to compute the marginal likelihood estimates of the fixed effects parameters and EB estimates of the random effects parameters, but without explicitly computing the integrated likelihood.

5.4.2 Marginal and extended likelihood

For a mixed effects model the conditional likelihood of the j^{th} $(j = 1, \dots, n_i)$ repeated observation on the i^{th} subject $(i = 1, \dots, N)$, i.e. y_{ij}, is given by $f_{\boldsymbol{\beta},\boldsymbol{\lambda}}(y_{ij}|\mathbf{v}_i)$. The likelihood of the i^{th} random effect is given by $f_{\boldsymbol{\lambda}}(\mathbf{v}_i)$. In this section we treat the dispersion parameters as one block. In Section 5.4.4 the two dispersion parameters will be treated separately. The total likelihood of the fixed and random effects is the product of all such terms over the subjects and the repeated observations within a subject, i.e.

$$L_E(\boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{v}|\mathbf{y}, \mathbf{v}) = \prod_{i=1}^{N} \prod_{j=1}^{n_i} f_{\boldsymbol{\beta},\boldsymbol{\lambda}}(y_{ij}|\mathbf{v}_i) f_{\boldsymbol{\lambda}}(\mathbf{v}_i). \tag{5.3}$$

The logarithm of (5.3) is called the extended log-likelihood by Lee *et al.* (2006) and they denoted its logarithm by $h = \log\left[L_E(\boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{v}|\mathbf{y}, \mathbf{v})\right]$. The hierarchical structure of the data, i.e. repeated measures (y_{ij}) within subjects (represented by $\mathbf{v}_i$) is clear from the expression of the extended likelihood. However maximization of (5.3) with respect to fixed effects and random effects parameters leads in general to fixed effects parameter estimates that are not consistent. This is known for a long time, see e.g. Cox (1970).

Instead the standard procedure is to maximize the marginal likelihood given by

$$L_M(\boldsymbol{\beta}, \boldsymbol{\lambda}|\mathbf{y}) = \prod_{i=1}^{N} \int \prod_{j=1}^{n_i} f_{\boldsymbol{\beta},\boldsymbol{\lambda}}(y_{ij}|\mathbf{v}_i) f_{\boldsymbol{\lambda}}(\mathbf{v}_i) d\,\mathbf{v}_i. \tag{5.4}$$

Maximizing L_M or equivalently log-likelihood $\ell_M = log(L_M)$ yields consistent estimates of the fixed effects parameters. However, the problem lies in computing the integrated likelihood. This is a time-consuming task since it needs to be done for each subject and each iteration. Further, if the MLE is determined with a Newton-Raphson procedure then integrals need to be computed also for the first and second derivatives. It is also important to fine tune the likelihood calculations, see e.g. Lesaffre and Spiessens (2001).

In Lee and Nelder (1996) the authors proposed a technique which exploits the extended likelihood for its use in estimating and testing all parameters (fixed effects and random effects) in hierarchical generalized linear models.

5.4.3 Computing marginal MLEs using the h-likelihood approach

In some special cases, i.e. when the random effects are on the canonical scale (see e.g. Lee *et al.* (2006) pp. 112-114), joint maximization of the extended log-likelihood

h with respect to all parameters $(\boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{v}_1, \ldots, \mathbf{v}_N)$ is equivalent to maximizing the marginal likelihood with respect to $\boldsymbol{\beta}, \boldsymbol{\lambda}$ and taking the EB-estimates for $\mathbf{v}_1, \ldots, \mathbf{v}_N$. But most often the two maximization procedures are not equivalent. When (5.3) is used to estimate fixed effects and random effects in the mean structure, the estimation method will be denoted as HL(0).

In the general case Noh and Lee (2007) suggest to work with a Laplace approximation to the marginal likelihood (5.4). The computation of the marginal likelihood by solving the integral by the Laplace approximation results in the adjusted profile (log)-likelihood

$$p_v(h) = h(\boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{v})|_{\mathbf{v}=\hat{\mathbf{v}}} - 0.5 \log \left| \frac{D(h, \mathbf{v})}{2\pi} \right|_{\mathbf{v}=\hat{\mathbf{v}}}, \tag{5.5}$$

with $D(h, \mathbf{v}) = -\frac{\partial^2 h(\boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{v})}{\partial \mathbf{v}^T \partial \mathbf{v}}$. The term ‘adjusted profile likelihood’ is chosen since $h(\boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{v})|_{\mathbf{v}=\hat{\mathbf{v}}}$ is a profile (log)-likelihood of $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$ and the second term of (5.5) is a correction term to approximate the marginal log-likelihood. Note that $\hat{\mathbf{v}}$ is the maximum likelihood estimator of $\mathbf{v}$ for given $(\boldsymbol{\beta}, \boldsymbol{\lambda})$ computed using h as the objective function, i.e. one could write $\hat{v} = \hat{v}_{\boldsymbol{\beta}, \boldsymbol{\lambda}}$.

The next step in the iterative procedure is to maximize the adjusted profile (log-) likelihood (5.5) with respect to $\boldsymbol{\beta}$. This approach will be denoted as HL(1). We stress that HL(0) and HL(1) denote the different ways of estimating $\boldsymbol{\beta}$, in both situations $\boldsymbol{\lambda}$ and $\mathbf{v}$ are estimated using the same objective function, i.e. the adjusted profile likelihood of order one and the extended likelihood respectively. Maximizing the profile log-likelihood $h(\boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{v})|_{\mathbf{v}=\hat{\mathbf{v}}}$ to find the MLE of $\boldsymbol{\beta}$ is not appropriate since this is equivalent to joint maximization of h over $\boldsymbol{\beta}$ and $\mathbf{v}$ which is most often invalid (see above).

After obtaining $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}}$ from maximization of (5.5) for a given dispersion component $\boldsymbol{\lambda}$, the estimation algorithm proceeds with estimation of $\boldsymbol{\lambda}$. Another adjusted profile likelihood is used as an objective function to find $\hat{\boldsymbol{\lambda}}$:

$$p_{\beta,v}(h) = h(\boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{v})|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}},\mathbf{v}=\hat{\mathbf{v}}} - 0.5 \log \left| \frac{D\left[h, (\boldsymbol{\beta}, \mathbf{v})\right]}{2\pi} \right|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}},\mathbf{v}=\hat{\mathbf{v}}} . \qquad (5.6)$$

The latter adjusted profile likelihood is maximized with respect to $\boldsymbol{\lambda}$ to obtain $\hat{\boldsymbol{\lambda}}$. Note that this objective function is "focussed" solely on the dispersion parameters. This offers an extension of REML estimation and inference to the class of generalized linear mixed models Noh and Lee (2007). It can be shown that in case of linear mixed models this function is exactly the restricted maximum likelihood.

5.4.4 A class of hierarchical generalized linear models

In this section we review the class of hierarchical generalized linear models (HGLM). The maximization procedures of the previous section were applied in this class of models by Lee and Nelder (1996).

The distribution of the response given random effects in a HGLM is assumed to belong to the exponential family, i.e.

$$f_{\boldsymbol{\beta},\lambda_e}(y_{ij}|v_i) = \exp\left[\frac{y_{ij}\theta_{ij} - b(\theta_{ij})}{\lambda_e} + c_1(y_{ij}, \lambda_e)\right], \qquad (5.7)$$

where λ_e determines the variance of the distribution in Normal or Gamma case, while it plays the role of an overdispersion parameter in case of Binomial or Poisson distribution. When we consider only likelihood inference, λ_e is fixed to 1 for a Poisson or a Binomial model. For the canonical link, $\theta_{ij} = \eta_{ij} = \mathbf{x}_{ij}^T\boldsymbol{\beta} + z_{ij}v_i$. Finally the functions $b(\theta_{ij})$ and $c_1(y_{ij}, \lambda_e)$ are determined by the chosen distribution. Note that Lee and Nelder basically restrict the random effects to the univariate case, although in Lee *et al.* (2006) a trick to construct correlated random effects is presented. We discuss this issue in Section 5.7.

In the absence of covariates a convenient choice for the random effect distribution of θ_{ij} is a Bayesian conjugate prior (see e.g. Cox and Hinkley (1974)). Lee and

Nelder assume the following conjugate prior for v_i - a part of θ_{ij}:

$$f_{\lambda_v}(v_i) = \exp\left[a_1(\lambda_v)v_i - a_2(\lambda_v)b(v_i) + c_2(\lambda_v)\right]. \tag{5.8}$$

Since the aim is to maximize h with respect of $\boldsymbol{\beta}$ but also with respect to v_i, it is convenient to rewrite distribution (5.8) such that it becomes an exponential distribution as a function of the 'parameter' v_i as follows:

$$f_{\lambda_v}(v_i) = \exp\left[\frac{\psi_i v_i - b(v_i)}{\lambda_v} + c_2(\psi_i, \lambda_v)\right], \tag{5.9}$$

with λ_v the dispersion parameter of the random component v_i. Note that in deriving (5.9) from (5.8) the property that $\lambda_v\, a_2(\lambda_v) = 1$ holds for many (standardized) random effects distributions, e.g. normal with mean zero, beta with mean 0.5, gamma and inverse gamma with mean 1.

The trick is now to consider (5.9) as the distribution of $\psi_i = \frac{a_1(\lambda_v)}{a_2(\lambda_v)}$ and not as the distribution of v_i for which it was defined. For this reason, ψ_i is called a pseudo-response. As such, ψ_i appears to have an exponential family distribution although it is not stochastic.

The extended likelihood can then be written as:

$$h(\boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{v}) = \sum_{i=1}^{N}\sum_{j=1}^{n_i}\left[\frac{y_{ij}\theta_{ij} - b(\theta_{ij})}{\lambda_e} + c_1(y_{ij}, \lambda_e)\right] + \sum_{i=1}^{N}\left[\frac{\psi_i v_i - b(v_i)}{\lambda_v} + c_2(\psi_i, \lambda_v)\right]. \tag{5.10}$$

An extended IWLS algorithm can be used to estimate hierarchical generalized linear models using the h-likelihood approach. In the next section we show the details of the algorithm for the truncated Poisson model.

5.5 Application of h-likelihood approaches to the Poisson-hurdle model

The hurdle model was introduced in Section 5.3. It was indicated that under independence of v_{1i} and v_{2i} in (5.2) the likelihood of the Poisson-hurdle model factorizes into the sum of the likelihood pertaining to a binary model and the likelihood of a truncated Poisson model. The estimation of the binary model with h-likelihood is well known (e.g. Noh and Lee (2007)). Here we describe the adaptation of h-likelihood procedures to find the estimates of the truncated Poisson model.

5.5.1 Estimation of the truncated distributions in h-likelihood

In this section we adapt the h-likelihood estimation algorithm to estimate the parameters of a truncated exponential distribution. A truncated distribution from an exponential family can be written as follows:

$$f_{\boldsymbol{\beta},\lambda_e}(y_{ij}|v_i) = \exp\left\{\frac{y_{ij}\theta_{ij} - b(\theta_{ij}) - \log[M(\theta_{ij})]}{\lambda_e} + c_1(y_{ij},\lambda_e)\right\}, \tag{5.11}$$

where $M(\theta_{ij})$ is a correction term to the exponential distribution as a result from truncation. The extended log-likelihood can be written as:

$$h = \sum_{i=1}^{N}\sum_{j=1}^{n_i}\log\left[f_{\boldsymbol{\beta},\lambda_e}(y_{ij}|v_i)\right] + \sum_{i=1}^{N}\log\left[f_{\lambda_v}(v_i)\right]. \tag{5.12}$$

We now show the modification of the algorithm for the joint maximization of (5.12) with respect to $\boldsymbol{\beta}$ and $\mathbf{v}$. In the Appendix we show a detailed description of the procedure, which can be applied to estimate $\mathbf{v}$ using h, $\boldsymbol{\beta}$ using $p_v(h)$, and $\boldsymbol{\lambda}$ using $p_{\beta,v}(h)$.

The score equations of (5.12) have the following form:

$$\frac{\partial h}{\partial \beta_p} = \sum_{i=1}^{N}\sum_{j=1}^{n_i} w_{ij}\left\{y_{ij} - \mu_{ij} - \frac{M'(\theta_{ij})}{M(\theta_{ij})}\right\}\frac{\partial \eta_{ij}}{\partial \mu_{ij}}\frac{\partial \eta_{ij}}{\partial \beta_p}, \tag{5.13}$$

$$\frac{\partial h}{\partial v_i} = \sum_{j=1}^{n_i} w_{ij}\left\{y_{ij} - \mu_{ij} - \frac{M'(\theta_{ij})}{M(\theta_{ij})}\right\}\frac{\partial \eta_{ij}}{\partial \mu_{ij}}\frac{\partial \eta_{ij}}{\partial v_i} + \left\{\frac{\psi_i - u_i}{\lambda_v}\right\}. \tag{5.14}$$

As a result, modified Henderson's equations (see e.g. Lee and Nelder (1996)) are needed, i.e.

$$\begin{pmatrix} \mathbf{X}^T\widetilde{\mathbf{W}}\mathbf{X} & \mathbf{X}^T\widetilde{\mathbf{W}}\boldsymbol{Z} \\ \boldsymbol{Z}^T\widetilde{\mathbf{W}}\mathbf{X} & \boldsymbol{Z}^T\widetilde{\mathbf{W}}\boldsymbol{Z} + \mathbf{Q} \end{pmatrix}\begin{pmatrix} \boldsymbol{\beta}^{NEW} \\ \mathbf{v}^{NEW} \end{pmatrix} = \begin{pmatrix} \mathbf{X}^T\widetilde{\mathbf{W}}\tilde{\mathbf{s}} \\ \boldsymbol{Z}^T\widetilde{\mathbf{W}}\tilde{\mathbf{s}} + \mathbf{R} \end{pmatrix}, \tag{5.15}$$

where

$$\begin{aligned} \widetilde{\mathbf{W}} &= \mathbf{W} + \text{diag}\left\{\frac{M''(\theta_{ij})}{M(\theta_{ij})} - \left[\frac{M'(\theta_{ij})}{M(\theta_{ij})}\right]^2\right\}\mathbf{V}^{-1}(\boldsymbol{\mu})\mathbf{W}, \\ \tilde{\mathbf{s}} &= \boldsymbol{\eta} + (\mathbf{W}/\widetilde{\mathbf{W}})\left[\mathbf{y} - \boldsymbol{\mu} - \frac{M'(\boldsymbol{\theta})}{M(\boldsymbol{\theta})}\right]\frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\mu}}, \end{aligned} \tag{5.16}$$

with $\boldsymbol{\beta}^{NEW}$ and $\mathbf{v}^{NEW}$ the updated vectors, $\mathbf{X}$ and $\boldsymbol{Z}$ are design matrices pertaining to the fixed effects and the random effect in the mean structure, respectively the $\mathbf{W}$ matrix is the diagonal weight matrix defined as for generalized linear models, i.e. $\mathbf{W} = \left(\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\eta}}\right)^2 V(\boldsymbol{\mu})^{-1}$. Further $V(\boldsymbol{\mu})$ is a variance function of the distribution $f_{\boldsymbol{\beta},\lambda_e}(y_{ij}|\mathbf{v}_i)$, $\mathbf{Q} = -\frac{\partial^2 log(f_{\lambda_v}(\mathbf{v}))}{\partial \mathbf{v}^T \partial \mathbf{v}}$, $\mathbf{s} = \boldsymbol{\eta} + (\mathbf{y} - \boldsymbol{\mu})\left(\frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\mu}}\right)$ and $\mathbf{R} = \mathbf{Q}\mathbf{v} + \frac{\partial \log(f_{\lambda_v}(\mathbf{v}))}{\partial \mathbf{v}}$. For details of the derivations and further adjustments necessary for the estimation of $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$ using adjusted profile likelihood, see Appendix.

5.5.2 Truncated Poisson model

Here we focus on the truncated (at zero) Poisson distribution. When Y_{ij} has a Poisson distribution truncated at zero, i.e. $Y_{ij} \sim TPoisson(\mu_{ij})$, the functions in

(5.11) are defined as:

$$\begin{aligned} \theta_{ij} &= \log(\mu_{ij}), \\ b(\theta_{ij}) &= \exp(\theta_{ij}), \\ M(\theta_{ij}) &= 1 - \exp(-\exp(\theta_{ij})), \\ \lambda_e &= 1, \end{aligned} \tag{5.17}$$

with μ_{ij} the mean of the Poisson distribution related to covariates by the log link as in (5.2).

The following extension to a three-level truncated Poisson model was inspired by the analysis of the Tai-Chi Chuan intervention data. The two level truncated Poisson model assumes a truncated Poisson distribution at each time point. However, we would like to allow for the heavy tails of the distribution (overdispersion) by the inclusion of an additional random effect.

The following model was therefore considered:

$$\begin{aligned} Y_{ijk} &\sim TPoisson(\mu_{ijk}), \\ log(\mu_{ijk}) &= \mathbf{x}_{ijk}^T\boldsymbol{\beta} + v_i + v_{ij}, \\ \mu_{ijk} &= exp(\mathbf{x}_{ijk}^T\boldsymbol{\beta})u_i u_{ij}, \end{aligned} \tag{5.18}$$

where $v_i = log(u_i)$ and $v_{ij} = log(u_{ij})$. Further, $u_i \sim$ Gamma $\left(\frac{1}{\lambda_1}, \lambda_1\right)$ and $u_{ij} \sim$ Gamma $\left(\frac{1}{\lambda_2}, \lambda_2\right)$, both with mean equal to one. The model can also accommodate a typical multi-level structure whereby there are N schools and within the $i-th$ school there are n_i classes each generating K_{ij} independent (> 0) counts. In the Tai-Chi Chuan data, $K_{ij} = 1$ for $i = 1 \ldots N$, $j = 1 \ldots n_i$. Note that the parameter μ_{ijk} is not the mean of a truncated Poisson distribution, the actual mean is given by:

$$\mu_{ijk}^{Trunc} = \mu_{ijk}\left(\frac{1}{1 - e^{-\mu_{ijk}}}\right). \tag{5.19}$$

The extended log-likelihood of the above model is then given:

$$h = \sum_{i=1}^{N}\sum_{j=1}^{n_i}\sum_{k=1}^{K_{ij}} \{y_{ijk}\theta_{ijk} - b(\theta_{ijk}) - log[M(\theta_{ijk})]\} + \sum_{i=1}^{N}\left[\frac{\psi_i v_i - b(v_i)}{\lambda_1}\right] + \sum_{i=1}^{N}\sum_{j=1}^{n_i}\left[\frac{\psi_{ij} v_{ij} - b(v_{ij})}{\lambda_2}\right]. \tag{5.20}$$

To estimate all parameters using the h-likelihood approach, similar (but more general) equations are needed than for the 2-level truncated Poisson case. We refer to the Appendix for further technical details.

5.6 Applications

5.6.1 Adverse Events Study

The adverse events study has been previously analyzed by Min and Agresti (2005) using a hurdle model. Here we repeat their analysis with a Gaussian quadrature approach which we will consider as our reference. Indeed in case of truly normally distributed random effects and with a two level structure this approach is probably most appropriate. For comparative reasons we reanalyze the dataset with the h-likelihood approach, and using also a hurdle model. The following model (as in Min and Agresti (2005)) is used:

$$\begin{aligned} logit(1-p_{ij}) &= \beta_{10} + \beta_{11}TRT_i(=2) + \beta_{12}\log(time_{ij}) + v_{1i}, && \text{Bernoulli Part} \\ log(\mu_{ij}) &= \beta_{20} + \beta_{21}TRT_i(=2) + \beta_{22}\log(time_{ij}) + v_{2i}, && \text{Truncated Poisson Part} \end{aligned} \tag{5.21}$$

where $TRT_i = 2$ denotes that treatment B is administered to the $i-th$ subject, while treatment A is a baseline group. Treatment is a time independent variable. Covariate $time_{ij}$ denotes the time between $j-th$ and $(j-1)th$ visit for the $i-th$ subject and is a time dependent covariate. In Min and Agresti (2005) the data were analyzed using correlated Gaussian random effects v_{1i} and v_{2i}, referred to in the remainder of the paper as the MA analysis. As mentioned above, a limitation of our approach at present is that we need to assume independence of v_{1i} and v_{2i}. Therefore we have analyzed the data using the approach of Min and Agresti (2005) also with independent random effects, but then we allowed the distribution of random effects to be different from normal using the method described in Liu and Yu (2008). In the Bernoulli part of the hurdle model we assumed normal random effects $v_{1i} \sim \mathcal{N}(0, \sigma_v^2)$ (dispersion parameter $\lambda_v = \sigma_v^2$) and beta random effects $v_{1i} = \text{logit}(u_{1i})$ and $u_{1i} \sim \text{Beta}\left(\frac{1}{2\lambda_v}, \frac{1}{2\lambda_v}\right)$ with mean 0.5 and variance $\frac{1}{4}\frac{\lambda_v}{1+\lambda_v}$. In the truncated Poisson part of the model we assumed normal random effects $v_{2i} \sim \mathcal{N}(0, \sigma_v^2)$ (dispersion parameter $\lambda_v = \sigma_v^2$) and gamma random effects $v_{2i} \sim \log(u_{2i})$ and $u_{1i} \sim \text{Gamma}\left(\frac{1}{\lambda_v}, \lambda_v\right)$ with mean 1 and variance λ_v. The MA analysis was executed with an adaptive Gaussian quadrature approach with twenty [AGQ(20)] quadrature points using SAS PROC NLMIXED (version 9.2) and is here considered as the reference analysis. We compared the above analyzes to the h-likelihood approach outlined in the previous section.

Originally h-likelihood procedures Lee and Nelder (1996) were based on the joint estimation of fixed and random effects given the estimates (λ) of the dispersion parameters, which were evaluated by an adjusted profile likelihood. We begin with such an analysis, which we have denoted as HL(0) in Section 5.4.3. In the same section we have seen that HL(0) can be improved by the evaluation of the fixed effects by the adjusted profile likelihood $p_v(h)$ Noh and Lee (2007). Next, given fixed effects and dispersion parameters, the extended likelihood is maximized to find random effects. This approach was called HL(1) in Section 5.4.3.

The results are presented in Table 5.3 for the Bernoulli part and in Table 5.4 for the truncated Poisson part of the model. First, we compare the HL(0) to the HL(1) method. The HL(0) method seems to be inferior especially in the case of the Bernoulli model. The estimates are attenuated towards zero and are relatively far from the estimates of the reference AGQ(20) analysis. For the truncated Poisson part of the hurdle model, HL(0) is doing better, especially when gamma random effects are used. In this case it appears to yield equivalent estimates to the HL(1) approach. Next, we compare the HL(1) method with the AGQ(20) approach with independent random effects. Point estimates obtained from both approaches are close for the fixed effects, but there is some discrepancy in the estimation of the dispersion components. Standard errors from the two methods are similar. Finally we compare the HL(1) method (with independent Gaussian random effects) to the original analysis in Min and Agresti (2005). For the Bernoulli part, the point estimates and standard errors are close. Larger differences, however, occur in the truncated Poisson part of the model, but the inferential conclusions remain the same. Especially the dispersion parameters differ. As far as the choice of the distribution of the random effects is concerned it does not have a great impact on point estimates, while slight differences are detected in standard errors. The choice of the random effects distribution matters more in the Bernoulli part of the hurdle model than in the truncated Poisson part.

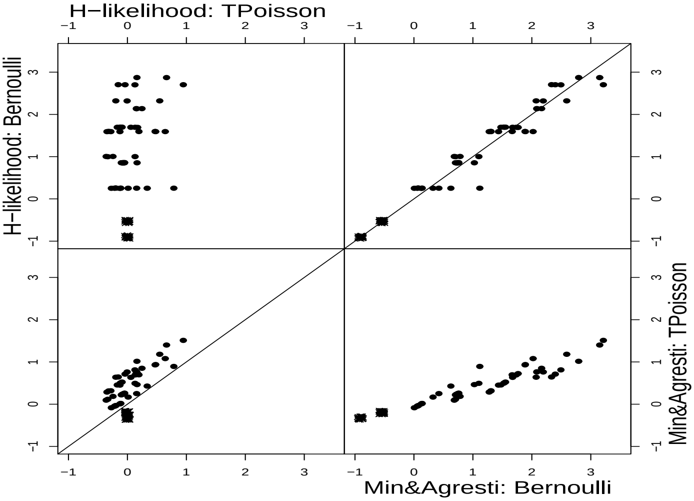


Figure 5.1: Adverse Events Study: Comparison of random effects estimates for the two parts of the hurdle model

Figure 5.1 presents the estimates of the random effects from the MA analysis and the HL(1) approach with Gaussian random effects. There seems to be a high agreement in the estimated random effects from both models for the Bernoulli part of the hurdle model. Less agreement is seen for the truncated Poisson part. Random effects depicted by squares, show the estimates of the random effects of the 64 patients who did not experience an adverse event. These patients correspond to the zero class of the Bernoulli part of the hurdle model and hence do not contribute to the estimation of the truncated Poisson part. However, the empirical Bayes rule allows for the computation of the random effects associated with the truncated Poisson model in the marginal analysis. Similarly, h-likelihood maximization can yield estimates of these random effects after they have been incorporated into h. The interpretation of such random effects in the context of the hurdle model is not immediately clear, however.

To summarize, ignoring the correlation between random effects in the two parts of the hurdle models seems to have more impact on the truncated Poisson part, than the Bernoulli part. The differences are greater in the estimates of the dispersion parameters and in the correlation between the random effects.

Finally, we computed deviance residuals for the truncated Poisson model in (5.21). The QQ-plots of these residuals showed the same pattern as residuals presented in the Section 5.6.2. We created the QQ-plots for HL(0) and HL(1) models with normal and gamma random effects, the QQ plots differ only slightly and did not indicate a preference for a particular model.

5.6.2 Tai-Chi Chuan effectiveness in the elderly

Our second example concerns the Tai-Chi Chuan intervention in preventing falls in elderly patients. We introduced the study in Section 5.2.2. The question here is to determine the factors that trigger the patient to fall or not. Further, one wished to know what factors determine the frequency of falling conditional on the fact that the patient experienced at least one fall.

The hurdle model approach offers a tool for the accommodation of the excess of zeros. The correlation between the observations on the same subject in each part of the hurdle model can be accounted for by the inclusion of a subject specific random effect. On top of this random effects structure, another random effect can be used to model the tails of the distribution of the counts. We assumed independence of random effects in the hurdle model and fitted the binary model and the truncated

Poisson model separately. We started from a complex model and removed the redundant terms subsequently by backward elimination using 5% significance level. We did not correct for multiple comparisons. Parameters describing the Tai-Chi Chuan effectiveness were always retained in the model. We implemented ordering in our modelling strategy: first, we selected the distribution for random effects using h-likelihood, then we selected the structure of the dispersion using $p_{\beta,v}(h)$ adjusted profile likelihood, and at the end regression parameters of the mean were selected based on $p_v(h)$ adjusted profile likelihood. The fitted model was the following:

$$\begin{aligned} \operatorname{logit}(1-p_{ij}) &= \mathbf{x}_{1ij}^T\boldsymbol{\beta}_1 + v_{1i}, && \text{Bernoulli Part} \\ \log\left(\frac{\mu_{ij}}{days_{ij}}\right) &= \mathbf{x}_{2ij}^T\boldsymbol{\beta}_2 + v_{2i} + v_{2ij}, && \text{Truncated Poisson Part} \end{aligned} \qquad (5.22)$$

where $days_{ij}$ denotes number of exposure days for the $i-th$ patient in the $j-th$ period. The covariates vectors $\mathbf{x}_1^t$ and $\mathbf{x}_2^t$ contain gender, time of the visit, interaction of intervention and time. In addition age is included in the Bernoulli part of the model. Table 5.5 presents the results of the estimation of the Bernoulli model. In the final model the random effects were $v_{1i} = \operatorname{logit}(u_{1i})$ and $u_{1i} \sim \text{Beta}(\frac{1}{2\lambda_{v_1}}, \frac{1}{2\lambda_{v_1}})$ with mean 0.5 and variance $\frac{1}{4}\frac{\lambda_{v_1}}{1+\lambda_{v_1}}$. The dispersion parameter λ_{v_1} is significantly different between males and females, indicating that counts of the female patients are less dispersed around the average individual evolution. There is no significant effect of the Tai-Chi Chuan intervention on the probability whether patients fall or not.

Table 5.3: Adverse Events Study: Bernoulli part of the hurdle model

Parameter	Original ($\rho = 0.848$)	AGQ(20)N	HL(0)N	HL(1)N	AGQ(20)B	HL(0)B	HL(1)B
			Estimates				
Intercept	-2.874	-2.889	-2.377	-2.961	-2.800	-2.308	-2.738
Treat(2)	0.895	0.889	0.724	0.910	0.853	0.697	0.837
log(time)	0.021	0.028	0.025	0.028	0.027	0.022	0.026
$\log(\lambda_v)$	0.998	1.005	0.723	1.039	-0.836	-1.185	-1.009
			Standard Errors				
Intercept	0.622	0.624	0.553	0.599	0.609	0.531	0.574
Treat(2)	0.417	0.418	0.363	0.420	0.408	0.327	0.378
log(time)	0.186	0.186	0.176	0.186	0.186	0.172	0.182
$\log(\lambda_v)$	0.305	0.305	0.299	0.311	0.237	0.242	0.231

Original - analysis reported in Min and Agresti (2005) (ρ - estimate of the correlation between random effects); AGQ - Adaptive Gaussian Quadrature (number of quadrature points) assuming independence between Gaussian random effects; HL(0) - H-likelihood with joint maximization with respect to random and fixed parameters in the mean structure; HL(1) - H-likelihood with $p_v(h)$ used for the estimation of the fixed effects in the mean structure; B - Beta distribution assumed for random intercept; N - Normal distribution assumed for random intercept

Table 5.4: Adverse Events Study: Truncated Poisson part of the hurdle model

Parameter	Original ($\rho = 0.848$)	AGQ(20)N	HL(0)N	HL(1)N	AGQ(20)G	HL(0)G	HL(1)G
			Estimates				
Intercept	-2.844	-2.165	-2.037	-2.188	-2.025	-2.011	-2.020
Treat(2)	0.963	0.884	0.855	0.883	0.883	0.884	0.883
log(time)	0.540	0.521	0.520	0.517	0.513	0.518	0.516
$\log(\lambda_v)$	-0.696	-1.374	-1.478	-1.171	-1.169	-1.458	-1.439
			Standard Errors				
Intercept	0.735	0.643	0.632	0.642	0.636	0.630	0.629
Treat(2)	0.352	0.326	0.319	0.335	0.325	0.316	0.315
log(time)	0.192	0.195	0.194	0.197	0.196	0.194	0.194
$\log(\lambda_v)$	0.703	0.680	0.664	0.661	0.682	0.701	0.697

Original - analysis reported in Min and Agresti (2005) (ρ - estimate of the correlation between Gaussian random effects); AGQ - Adaptive Gaussian Quadrature (number of quadrature points) assuming independence between Gaussian random effects;HL(0) - H-likelihood with joint maximization with respect to random and fixed parameters in the mean structure; HL(1) - H-likelihood with $p_v(h)$ used for the estimation of the fixed effects in the mean structure; G - Gamma distribution assumed for random intercept; N - Normal distribution assumed for random intercept

Table 5.6 reports on the point estimates and standard errors for the truncated Poisson model, the number of exposure days is used as an offset ($days_{ij}$) in (5.22). Random effects $v_{2i} \sim \mathcal{N}(0, \sigma^2_{v_{21}})$ ($\lambda_{v_{21}} = \sigma^2_{v_{21}}$), while $b_{vij} = \log(u_{2ij})$ and $u_{2ij} \sim$ Gamma $\left(\frac{1}{\lambda_{v_{22}}}, \lambda_{v_{22}}\right)$ with mean 1 and variance $\lambda_{v_{22}}$. The dispersion parameter $\lambda_{v_{22}}$ differs significantly between males and females, indicating that the frequency of falls for males is less predictable at each visit. There is a borderline significant effect of the Tai-Chi Chuan intervention for the population experiencing at least one fall.

To summarize, the Tai-Chi Chuan intervention does not affect the probability of having at least one fall, but the population might benefit from the intervention in terms of reduction of the number of falls.

In order to investigate the goodness of fit of the truncated Poisson models we computed the deviance residuals. To check the compliance with the data we used a parametric bootstrap technique. We simulated 19 datasets using the final model presented in Table 5.6. For each model a QQ-plot was created. Figure 5.2 presents the QQ-plot of the final model (solid line) together with maximum and minimum values (dotted lines) obtained from the 19 parametric bootstrap samples. It appears that there is no deviation from the assumed distribution. We refer to the Appendix for the details of the deviance residuals computation.

Table 5.5: Tai-Chi Chuan Study: Bernoulli part of hurdle model by HL(1) approach

Effect	Estimate	P-value
Intercept	-5.69	0.004
Female	0.96	0.007
Time	$-1.18 \cdot 10^{-3}$	0.132
Time*Trt(Tai-Chi)	$-0.64 \cdot 10^{-3}$	0.454
Age	0.05	0.042
γ_{10}[1]	-0.42	0.096
γ_{11}[1]	-1.22	0.001

[1] - Estimates are for the dispersion component $\log(\lambda_{v_1}) = \gamma_{10} + \gamma_{11} \times Female_i$

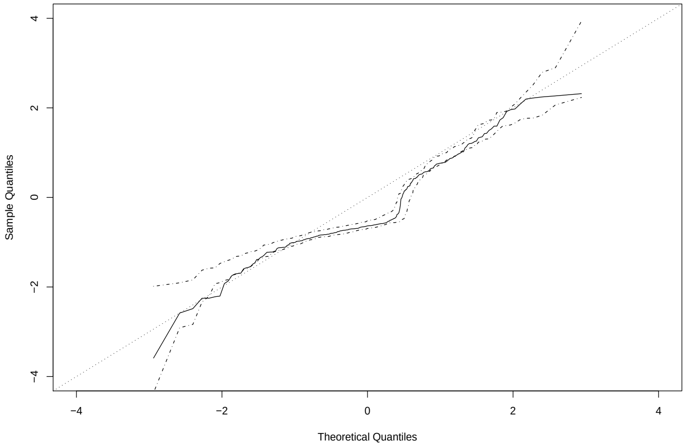


Figure 5.2: QQ-plot for truncated Poisson residuals

Table 5.6: Tai-Chi Chuan Study: Truncated Poisson part of the hurdle model by HL(1) approach

Effect	Estimate	P-value
Intercept	-3.54	<0.001
Female	-0.66	0.035
Time	$-2.48 \cdot 10^{-3}$	0.006
Time*Trt(Tai-Chi)	$-2.17 \cdot 10^{-3}$	0.048
${\gamma_{20}}^{1}$	0.11	0.670
${\gamma_{210}}^{2}$	-0.33	0.431
${\gamma_{211}}^{2}$	-1.50	0.032

[1] - Estimates are for the dispersion component $\log(\lambda_{v_{21}}) = \gamma_{21}$

[2] - Estimates are for the dispersion component $\log(\lambda_{v_{22}}) = \gamma_{220} + \gamma_{221} \times Female_i$

5.7 Discussion and conclusions

H-likelihood provides an alternative estimation and inference framework for random effects models. Lee *et al.* (2006) presented how extended likelihood and appropriate adjusted profile likelihoods can be used to estimate model parameters. The entire procedure results in an efficient estimation algorithm in terms of convergence and time. In this paper the algorithm is outlined for hierarchical generalized linear models and further extended to the estimation of the truncated distributions.

The h-likelihood framework allows random effects to have a different distribution than normal. Further, for the estimation of the dispersion parameters a REML objective function is used, now also in case of non-normal distributions of the data Noh and Lee (2007). The modeling of the dispersion parameters in terms of covariates can be performed. Furthermore, the incorporation of complex designs such as multilevel or multi-membership is possible.

We described the h-likelihood based estimation and inference for hurdle models. Using this methodology zero-inflated data of complex designs can be analyzed. An alternative tool for the analysis of such data is the zero-inflated Poisson model. While the ZIP model allows only for zero-inflation, hurdle models are appropriate for inflated, as well as for zero-deflated distributions.

The elegant Iterative Weighted Least Squares h-likelihood algorithm is obtained for models with independent random effects. Lee *et al.* (2006) describe a procedure to introduce correlation between random effects. However, the estimation of correlation does not fall within the IWLS algorithm and must be estimated by additional numerical procedures. In the correlated random effects setting the following issues need to be additionally solved: (1) the modelling of the variance covariance matrix as a function of covariates and (2) the introduction of the correlation between non-normal random effects. We will investigate the correlated random effects h-likelihood models in detail in our future work.

In this paper restricted ourselves to independent random effects. This allows for the factorization of the likelihood and separate fit of the binomial and truncated Poisson model. We focussed on the truncated Poisson distribution, however binomial distribution could have been used too. Overdispersion is allowed by the inclusion of an additional random effects as in the example of Section 5.6.2.

Currently the most accurate methods to estimate random effects models are based on the marginal likelihood, which is approximated by adaptive Gaussian quadrature algorithm. However, they might become prohibitive in case of complex designs. Breslow and Clayton (1993) proposed a penalized quasi-likelihood (PQL) which handles multilevel and crossed random effects, but their approach suffers from a biased estimation of the dispersion parameters. H-likelihood gives improved dispersion estimates compared to the PQL. Another approach is based on the evaluation of the marginal likelihood by the classical Laplace approximation. In Noh and Lee (2007) the comparison of the h-likelihood method and classical Laplace method for marginal likelihood is presented. Bias and mean squared error of estimates were compared. They showed that h-likelihood approach outperforms the estimation method based on standard Laplace approximation with respect to the estimation of dispersion and fixed effects parameters for the binary data.

Alternatively one could refer to the Bayesian methodology and evaluate the models by MCMC sampling. However, in limited analyzes for the adverse events study, we encountered problems with sampling using WINBUGS for a truncated Poisson model with gamma random effects. Namely, absorbing states occurred after 50-100 thousand iterations.

Additional work is required for the joint modeling of the two parts of the hurdle model within the h-likelihood framework, which would enable correlation between the random effects. Additionally the procedures to handle the competitor to the hurdle model, the ZIP model, within the h-likelihood need to be developed.

Appendix

Estimating equations for the truncated Poisson distribution

We present here the derivation of the estimation algorithm for the truncated Poisson distribution. We describe the situation when canonical links are used. The joint likelihood of a three level model has the following form:

$$h = \sum_{ijk} (y_{ijk}\theta_{ijk} - b(\theta_{ijk}) - log(M(\theta_{ijk}))) + \sum_{i} \left(\frac{\psi_i v_i - b(v_i)}{\lambda_1}\right) + \sum_{ij} \left(\frac{\psi_{ij} v_{ij} - b(v_{ij})}{\lambda_2}\right) \tag{5.23}$$

Now we compute the gradient and the hessian of the (5.23).

$$\begin{aligned}
\frac{\partial h}{\partial \beta_p} =& \sum_{ijk} \left(y_{ijk} - \mu_{ijk} - \frac{M'(\theta_{ijk})}{M(\theta_{ijk})}\right) \frac{\partial \theta_{ijk}}{\partial \mu_{ijk}} \frac{\partial \mu_{ijk}}{\partial \eta_{ijk}} \frac{\partial \eta_{ijk}}{\partial \beta_p} \\
=& \sum_{ijk} \left(y_{ijk} - \mu_{ijk} - \frac{M'(\theta_{ijk})}{M(\theta_{ijk})}\right) V^{-1}(\mu_{ijk}) \frac{\partial \mu_{ijk}}{\partial \eta_{ijk}} \frac{\partial \eta_{ijk}}{\partial \beta_p} \\
=& \sum_{ijk} \left(y_{ijk} - \mu_{ijk} - \frac{M'(\theta_{ijk})}{M(\theta_{ijk})}\right) w_{ijk} \frac{\partial \eta_{ijk}}{\partial \mu_{ijk}} \frac{\partial \eta_{ijk}}{\partial \beta_p} \\
\frac{\partial h}{\partial v_i} =& \sum_{ijk} \left(y_{ijk} - \mu_{ijk} - \frac{M'(\theta_{ijk})}{M(\theta_{ijk})}\right) w_{ijk} \frac{\partial \eta_{ijk}}{\partial \mu_{ijk}} \frac{\partial \eta_{ijk}}{\partial v_i} + \frac{\psi_i - u_i}{\lambda_1} \\
\frac{\partial h}{\partial v_{ij}} =& \sum_{ijk} \left(y_{ijk} - \mu_{ijk} - \frac{M'(\theta_{ijk})}{M(\theta_{ijk})}\right) w_{ijk} \frac{\partial \eta_{ijk}}{\partial \mu_{ijk}} \frac{\partial \eta_{ijk}}{\partial v_{ij}} + \frac{\psi_{ij} - u_{ij}}{\lambda_2},
\end{aligned}$$

where $\eta_{ijk} = log(\mu_{ijk}) = \mathbf{x}_{ijk}^T \boldsymbol{\beta} + z_{ijk,i}^1 v_i + z_{ijk,ij}^2 v_{ij}$ is a linear predictor and w_{ijk} is a diagonal element of the weight matrix $\mathbf{W} = \text{diag}\left[\frac{\partial \mu_{ijk}}{\partial \eta_{ijk}}\right]^2 \mathbf{V}^{-1}(\boldsymbol{\mu})$, $V(\boldsymbol{\mu}) = b''(\theta)$. Note that for canonical links models we have $\frac{\partial \mu_{ijk}}{\partial \eta_{ijk}} = w_{ijk} = V(\mu_{ijk})$. Further $\boldsymbol{Z} = [\boldsymbol{Z}^1|\boldsymbol{Z}^2]$ and $z_{ijk,i}^1$ is the $ijk-th$ row of the $i-th$ column in $\boldsymbol{Z}^1$, while $z_{ijk,ij}^2$ us the $ijk-th$ row of the $ij-th$ column in $\boldsymbol{Z}^2$. In the random part $v_i = log(u_i)$ and $v_{ij} = log(u_{ij})$. We assume $u_i \sim \text{Gamma}(\frac{1}{\lambda_1}, \lambda_1)$ and $u_{ij} \sim \text{Gamma}(\frac{1}{\lambda_2}, \lambda_2)$. Here the pseudo-responses are $\psi_i = 1$ and $\psi_{ij} = 1$. In a matrix notation the gradient can be written as follows:

$$D = \frac{\partial h}{\partial(\beta, v_i, v_{ij})} = \begin{pmatrix} \mathbf{X}^T \mathbf{W} \frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\mu}} \left(\mathbf{y} - \boldsymbol{\mu} - \frac{M'(\boldsymbol{\theta})}{M(\boldsymbol{\theta})} \right) \\ \boldsymbol{Z}^{1T} \mathbf{W} \frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\mu}} \left(\mathbf{y} - \boldsymbol{\mu} - \frac{M'(\boldsymbol{\theta})}{M(\boldsymbol{\theta})} \right) + \frac{\boldsymbol{\psi}_1 - \mathbf{u}_1}{\lambda_1} \\ \boldsymbol{Z}^{2T} \mathbf{W} \frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\mu}} \left(\mathbf{y} - \boldsymbol{\mu} - \frac{M'(\boldsymbol{\theta})}{M(\boldsymbol{\theta})} \right) + \frac{\boldsymbol{\psi}_2 - \mathbf{u}_2}{\lambda_2} \end{pmatrix}, \tag{5.24}$$

where $\boldsymbol{\psi}_1$ is a vector of ψ_i, $\boldsymbol{\psi}_2$ is a vector of ψ_{ij}, $\mathbf{u}_1$ is a vector of u_i and $\mathbf{u}_2$ is a vector of u_{ij}. In the following derivation we define $\tilde{\mathbf{W}}$ as follows:

$$\tilde{\mathbf{W}} = \mathbf{W} \left(\mathbf{I} + \text{diag} \left(\frac{M''(\theta)}{M(\theta)} - \left(\frac{M'(\theta)}{M(\theta)} \right)^2 \right) V^{-1}(\boldsymbol{\mu}) \right), \tag{5.25}$$

with diagonal elements $\tilde{w}_{ijk}$. Now we present the derivation of the hessian matrix.

$$\begin{aligned} \frac{\partial h}{\partial \beta_p \partial \beta_r} &= -\sum_{ijk} \left(1 + \left(\frac{M''(\theta_{ijk})}{M(\theta_{ijk})} - \left(\frac{M'(\theta_{ijk})}{M(\theta_{ijk})} \right)^2 \right) V^{-1}(\mu_{ijk}) \right) \frac{\partial \mu_{ijk}}{\partial \eta_{ijk}} x_{ijk,p} x_{ijk,r} = \\ & - \sum_{ijk} \tilde{w}_{ijk} x_{ijk,p} x_{ijk,r} \\ \frac{\partial h}{\partial \beta_p \partial v_i} &= -\sum_{ijk} \tilde{w}_{ijk} x_{ijk,p} z^1_{ijk,i} \\ \frac{\partial h}{\partial v_i \partial v_i} &= -\sum_{ijk} \tilde{w}_{ijk} z^1_{ijk,i} z^1_{ijk,i} - \frac{\partial u_i}{\partial v_i} (\frac{1}{\lambda_1}) \\ \frac{\partial h}{\partial v_i \partial v_{ij}} &= -\sum_{ijk} \tilde{w}_{ijk} z^1_{ijk,i} z^2_{ijk,ij}, \end{aligned}$$

where for the random components we have

$$\text{diag} \left(\frac{\partial u_1}{\partial v_1} \right) = \mathbf{W}_1(\mathbf{u}_1) = \mathbf{W}_1 = \left(\frac{\partial \mathbf{u}_1}{\partial \mathbf{v}_1} \right)^2 \mathbf{V}^{-1}(\mathbf{u}_1).$$

Below we give expression for the negative expected hessian matrix:

$$H = -E \left[\frac{\partial h}{\partial(\beta, v_1, v_2) \partial(\beta, v_1, v_2)} \right] = \begin{pmatrix} \mathbf{X}^T \tilde{\mathbf{W}} \mathbf{X} & \mathbf{X}^T \tilde{\mathbf{W}} \boldsymbol{Z}^1 & \mathbf{X}^T \tilde{\mathbf{W}} \boldsymbol{Z}^2 \\ \boldsymbol{Z}^{1T} \tilde{\mathbf{W}} \mathbf{X} & \boldsymbol{Z}^{1T} \tilde{\mathbf{W}} \boldsymbol{Z}^1 + \boldsymbol{\Lambda}_1^{-1} \mathbf{W}_1 & \boldsymbol{Z}^{1T} \tilde{\mathbf{W}} \boldsymbol{Z}^2 \\ \boldsymbol{Z}^{2T} \tilde{\mathbf{W}} \mathbf{X} & \boldsymbol{Z}^{2T} \tilde{\mathbf{W}} \boldsymbol{Z}^1 & \boldsymbol{Z}^{2T} \tilde{\mathbf{W}} \boldsymbol{Z}_2 + \boldsymbol{\Lambda}_2^{-1} \mathbf{W}_2 \end{pmatrix} \tag{5.26}$$

In the next step we derive an adjusted dependent variable used in the estimation algorithm. Estimation algorithm is a Fisher Scoring maximization step extended for the joint likelihood. It is based on previously described matrices H and D. The updating of the parameters δ is performed as follows:

$$\begin{aligned} H(\delta^{+1} - \delta^{0}) &= D \\ H\delta^{+1} &= D + H\delta^{0} \end{aligned} \tag{5.27}$$

These are in our situation as the following:

$$\begin{pmatrix} \mathbf{X}^T\tilde{\mathbf{W}}\mathbf{X} & \mathbf{X}^T\tilde{\mathbf{W}}\boldsymbol{Z}^1 & \mathbf{X}^T\tilde{\mathbf{W}}\boldsymbol{Z}^2 \\ \boldsymbol{Z}^{1T}\tilde{\mathbf{W}}\mathbf{X} & \boldsymbol{Z}^{1T}\tilde{\mathbf{W}}\boldsymbol{Z}^1 + \boldsymbol{\Lambda}_1^{-1}\mathbf{W}_1 & \boldsymbol{Z}^{1T}\tilde{\mathbf{W}}\boldsymbol{Z}^2 \\ \boldsymbol{Z}^{2T}\tilde{\mathbf{W}}\mathbf{X} & \boldsymbol{Z}^{2T}\tilde{\mathbf{W}}\boldsymbol{Z}^1 & \boldsymbol{Z}^{2T}\tilde{\mathbf{W}}\boldsymbol{Z}^2 + \boldsymbol{\Lambda}_2^{-1}\mathbf{W}_2 \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}^{+1} - \boldsymbol{\beta}^0 \\ \mathbf{v}_1^{+1} - \mathbf{v}_1^0 \\ \mathbf{v}_2^{+1} - \mathbf{v}_2^0 \end{pmatrix} = \begin{pmatrix} \mathbf{X}^T\mathbf{W}\frac{\partial\boldsymbol{\eta}}{\partial\boldsymbol{\mu}}\left(\mathbf{y} - \boldsymbol{\mu} - \frac{M'(\boldsymbol{\theta})}{M(\boldsymbol{\theta})}\right) \\ \boldsymbol{Z}^{1T}\mathbf{W}\frac{\partial\boldsymbol{\eta}}{\partial\boldsymbol{\mu}}\left(\mathbf{y} - \boldsymbol{\mu} - \frac{M'(\boldsymbol{\theta})}{M(\boldsymbol{\theta})}\right) + \frac{\boldsymbol{\psi}_1 - \mathbf{u}_1}{\lambda_1} \\ \boldsymbol{Z}^{2T}\mathbf{W}\frac{\partial\boldsymbol{\eta}}{\partial\boldsymbol{\mu}}\left(\mathbf{y} - \boldsymbol{\mu} - \frac{M'(\boldsymbol{\theta})}{M(\boldsymbol{\theta})}\right) + \frac{\boldsymbol{\psi}_2 - \mathbf{u}_2}{\lambda_2} \end{pmatrix}$$

$$\begin{pmatrix} \mathbf{X}^T\tilde{\mathbf{W}}\mathbf{X} & \mathbf{X}^T\tilde{\mathbf{W}}\boldsymbol{Z}^1 & \mathbf{X}^T\tilde{\mathbf{W}}\boldsymbol{Z}^2 \\ \boldsymbol{Z}^{1T}\tilde{\mathbf{W}}\mathbf{X} & \boldsymbol{Z}^{1T}\tilde{\mathbf{W}}\boldsymbol{Z}^1 + \boldsymbol{\Lambda}_1^{-1}\mathbf{W}_1 & \boldsymbol{Z}^{1T}\tilde{\mathbf{W}}\boldsymbol{Z}^2 \\ \boldsymbol{Z}^{2T}\tilde{\mathbf{W}}\mathbf{X} & \boldsymbol{Z}^{2T}\tilde{\mathbf{W}}\boldsymbol{Z}^1 & \boldsymbol{Z}^{2T}\tilde{\mathbf{W}}\boldsymbol{Z}^2 + \boldsymbol{\Lambda}_2^{-1}\mathbf{W}_2 \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}^{+1} \\ \mathbf{v}_1^{+1} \\ \mathbf{v}_2^{+1} \end{pmatrix} = \begin{pmatrix} \mathbf{X}^T\tilde{\mathbf{W}}\tilde{\mathbf{s}} \\ \boldsymbol{Z}^{1T}\tilde{\mathbf{W}}\tilde{\mathbf{s}} + \boldsymbol{\Lambda}_1^{-1}\mathbf{W}_1\mathbf{v}_1^0 + \frac{\boldsymbol{\psi}_1 - \mathbf{u}_1}{\lambda_1} \\ \boldsymbol{Z}^{2T}\tilde{\mathbf{W}}\tilde{\mathbf{s}} + \boldsymbol{\Lambda}_2^{-1}\mathbf{W}_2\mathbf{v}_2^0 + \frac{\boldsymbol{\psi}_2 - \mathbf{u}_2}{\lambda_2} \end{pmatrix},$$

where $\tilde{\mathbf{s}} = \boldsymbol{\eta} + \frac{\mathbf{W}}{\tilde{\mathbf{W}}}\left[\mathbf{y} - \boldsymbol{\mu} - \frac{M'(\boldsymbol{\theta})}{M(\boldsymbol{\theta})}\right]\frac{\partial\boldsymbol{\eta}}{\partial\boldsymbol{\mu}}$ is an adjusted dependent variable. This set of equations corresponds to equations (4.3) of Lee and Nelder (1996). Therefore their algorithm can be used to evaluate it.

Computation of the modification terms in the estimation of fixed parameters using adjusted profile likelihood $p_v(h)$

In this section we show how to compute adjustments to the responses $\mathbf{y}$ and pseudo-responses $\boldsymbol{\psi}$ necessary to use $p_v(h)$ as an objective function used to estimate $\boldsymbol{\beta}$. We show the computation of the modifications $\mathbf{m}$ below. In the vector notation we can define modified responses and modified pseudo-responses as follows:

$$\begin{aligned} \mathbf{y}^* &= \mathbf{y} - \mathbf{m}, \\ \boldsymbol{\psi}_1^* &= \boldsymbol{\psi}_1 + \boldsymbol{\Lambda}_1\boldsymbol{Z}^{1T}\mathbf{W}\frac{\partial\boldsymbol{\eta}}{\partial\boldsymbol{\mu}}\mathbf{m}, \\ \boldsymbol{\psi}_2^* &= \boldsymbol{\psi}_2 + \boldsymbol{\Lambda}_2\boldsymbol{Z}^{2T}\mathbf{W}\frac{\partial\boldsymbol{\eta}}{\partial\boldsymbol{\mu}}\mathbf{m}, \end{aligned} \tag{5.28}$$

where $\mathbf{W} = \left(\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\eta}}\right)^2 \mathbf{V}^{-1}(\boldsymbol{\mu})$ which in our setting is $\mathbf{W} = \text{diag}(\boldsymbol{\mu})$, $\boldsymbol{\psi}_1$ and $\boldsymbol{\psi}_2$ are vectors of pseudo-responses with components ψ_i and ψ_{ij}, in this situation they are equal to one. We set $m_i = 0.5k_i\frac{\partial \mu_i}{\partial \eta_i}$ and show the computation of k_i. Note that now the index i refers to the $i-th$ element of the $\mathbf{m}$ vector of a dimension $[(\sum_{i=1}^{N}\sum_{j=1}^{n_i} K_{ij}) \times 1]$. We use notation $N^* = \sum_{i=1}^{N}\sum_{j=1}^{n_i} K_{ij}$ for the total number of observations. We define the equivalent of hat matrix based on $\mathbf{T}_R$:

$$\mathbf{P}_R = \mathbf{T}_R(\mathbf{T}_R^{\mathbf{T}}\boldsymbol{\Sigma}_a^{-1}\mathbf{T}_R)^{-1}\mathbf{T}_R^{T}\boldsymbol{\Sigma}_a^{-1},$$

where $\mathbf{T}_R$ and $\boldsymbol{\Sigma}_a$ are defined as below:

$$\mathbf{T}_R = \begin{pmatrix} \boldsymbol{Z}^1 & \boldsymbol{Z}^2 \\ \mathbf{I}_N & 0 \\ 0 & \mathbf{I}_{\sum_{i=1}^{N} n_i} \end{pmatrix} \qquad \boldsymbol{\Sigma}_a^{-1} = \begin{pmatrix} \tilde{\mathbf{W}} & 0 & 0 \\ 0 & \mathbf{W}_1\boldsymbol{\Lambda}_1^{-1} & 0 \\ 0 & 0 & \mathbf{W}_2\boldsymbol{\Lambda}_2^{-1} \end{pmatrix}, \tag{5.29}$$

where $\mathbf{W}_s = (\frac{\partial \mathbf{u}_s}{\partial \mathbf{v}_s})^2\mathbf{V}^{-1}(\mathbf{u}_s)$ is a diagonal matrix, which specifically for the gamma distributed random effects is $\mathbf{W}_s = \mathbf{u}_s$ $(s = 1, 2)$. Matrices $\boldsymbol{\Lambda}_s$ are diagonal matrices with entries equal to $\boldsymbol{\lambda}_s$ $(s = 1, 2)$.

The computation of k_i in our model can be done as follows:

$$\begin{aligned} k_i = {} & \mathbf{P}_R[i,i](1/\tilde{\mathbf{W}}[i,i])(1/\mathbf{W}[i,i])\frac{\partial \tilde{\mathbf{W}}}{\partial \boldsymbol{\theta}}[i,i] \\ & + \sum_{j=1}^{N^*} \mathbf{P}_R[j,j](1/\tilde{\mathbf{W}}[j,j])\frac{\partial \tilde{\mathbf{W}}}{\partial \boldsymbol{\theta}}[j,j]\mathbf{A}_{11}[j,i] \\ & + \sum_{j=1}^{N^*} \mathbf{P}_R[j,j](1/\tilde{\mathbf{W}}[j,j])\frac{\partial \tilde{\mathbf{W}}}{\partial \boldsymbol{\theta}}[j,j]\mathbf{A}_{12}[j,i] \\ & + \sum_{j=1}^{n_1} \mathbf{P}_R[(N^*+j),(N^*+j)]\mathbf{A}_{21}[j,i] \\ & + \sum_{j=1}^{n_2} \mathbf{P}_R[(N^*+n_1+j),(N^*+n_1+j)]\mathbf{A}_{22}[j,i], \end{aligned} \tag{5.30}$$

where $n_1 = N$, $n_2 = \sum_{i=1}^{N} n_i$ and $N^* = \sum_{i=1}^{N}\sum_{j=1}^{n_i} K_{ij}$. The derivative $\frac{\partial \tilde{\mathbf{W}}}{\partial \boldsymbol{\theta}}$ is defined as below:

$$\frac{\partial \tilde{\mathbf{W}}}{\partial \boldsymbol{\theta}} = \mathbf{W} + \frac{M'''(\boldsymbol{\theta})}{M(\boldsymbol{\theta})} - 3\frac{M'(\boldsymbol{\theta})M''(\boldsymbol{\theta})}{M(\boldsymbol{\theta})^2} + 2\left(\frac{M'(\boldsymbol{\theta})}{M(\boldsymbol{\theta})}\right)^3$$

Next, we turn to the computation of adjustment matrices. Define the following matrix:

$$\mathbf{D} = (\mathbf{T}_R^T \boldsymbol{\Sigma}_a^{-1} \mathbf{T}_R)^{-1} \boldsymbol{Z}^T$$

and partition it into:

$$\mathbf{D} = \begin{pmatrix} \mathbf{D}^1_{(n_1 \times N^*)} \\ \mathbf{D}^2_{(n_2 \times N^*)} \end{pmatrix} \qquad\qquad \boldsymbol{Z} = \begin{pmatrix} \boldsymbol{Z}^1_{N^* \times n_1} & \boldsymbol{Z}^2_{N^* \times n_2} \end{pmatrix}$$

This allows us to define the adjustment matrices:

$$\begin{aligned} \mathbf{A}_{11} &= -\boldsymbol{Z}^1 \mathbf{D}^1 \tilde{\mathbf{W}} \mathbf{W}^{-1} \\ \mathbf{A}_{12} &= -\boldsymbol{Z}^2 \mathbf{D}^2 \tilde{\mathbf{W}} \mathbf{W}^{-1} \\ \mathbf{A}_{21} &= -\mathbf{D}^1 \tilde{\mathbf{W}} \mathbf{W}^{-1} \\ \mathbf{A}_{22} &= -\mathbf{D}^2 \tilde{\mathbf{W}} \mathbf{W}^{-1} \end{aligned} \tag{5.31}$$

This adjustment allows to use (5.5) as an objective function for the estimation of the fixed effects in the IWLS algorithm. Upon convergence hessian matrix of (5.5) can be computed and used for calculation of standard errors.

Details of the estimation of the dispersion components

In order to maximize (5.6) the following derivative must be equal to zero:

$$\frac{\partial p_{\beta,v}(h)}{\partial \lambda_v} = \frac{1}{2}\sum_{i=1}^{N} \frac{d_i - (1-q_i)\lambda_1}{\lambda_1^2} = \sum_{i=1}^{N} \frac{(1-q_i)}{2} \frac{d_i^* - \lambda_1}{\lambda_1^2}, \tag{5.32}$$

the above expression for the derivative is a score equation of a gamma distributed random variable $d_i^* = \frac{d_i}{(1-q_i)}$ with mean λ_1, variance $\frac{2\lambda_1}{(1-q_i)}$ and prior weight $\frac{(1-q_i)}{2}$.

In our model $d_i = 2(u_i - v_i - 1)$. To derive the correct prior weights q_i we begin with the calculation of p_i:

$$p_i = \tilde{p}_i + 1 + 2\left[\frac{log(\lambda_1) + digamma(1/\lambda_1)}{\lambda_1}\right], \tag{5.33}$$

where $\tilde{p}_i$ is the $N^*{+}i(i = 1 \ldots N)$ diagonal value of the hat matrix for the hierarchical models:

$$\tilde{\mathbf{P}} = \mathbf{T}(\mathbf{T}^T\mathbf{\Sigma}_a^{-1}\mathbf{T})^{-1}\mathbf{T}^{\mathbf{T}}\mathbf{\Sigma}_a^{-1}.$$

The matrix $\mathbf{T}$ is defined as below:

$$\mathbf{T} = \begin{pmatrix} \mathbf{X} & \boldsymbol{Z}^1 & \boldsymbol{Z}^2 \\ 0 & \mathbf{I}_N & 0 \\ 0 & 0 & \mathbf{I}_{\sum_{i=1}^N n_i} \end{pmatrix} \tag{5.34}$$

where $\mathbf{X}$, $\boldsymbol{Z}^1$ and $\boldsymbol{Z}^2$ are design matrices for fixed effects, first random component and second random component respectively.

Finally, to compute the prior weights q_i the values of p_i in (5.33) need to be adjusted by the following quantity:

$$-\text{trace}\left\{\tilde{\mathbf{P}}\begin{pmatrix} \frac{\partial \log(\tilde{\mathbf{W}})}{\partial \log(\boldsymbol{\lambda}_1)} & 0 & 0 \\ 0 & \frac{\partial \log(\mathbf{W}_1)}{\partial \log(\boldsymbol{\lambda}_1)} & 0 \\ 0 & 0 & \frac{\partial \log(\mathbf{W}_2)}{\partial \log(\boldsymbol{\lambda}_1)} \end{pmatrix}\right\} \tag{5.35}$$

Now we present the computation of the derivatives matrix in the adjustment term (5.35):

In the calculation we need the derivatives $\frac{\hat{\mathbf{v}}_i}{\boldsymbol{\lambda}_1}$ and $\frac{\hat{\mathbf{v}}_{ij}}{\boldsymbol{\lambda}_2}$. These have the following expressions:

$$\frac{\partial \hat{\mathbf{v}}_i}{\partial \boldsymbol{\lambda}_1} = -\left[\boldsymbol{Z}^{1T}\tilde{\mathbf{W}}\boldsymbol{Z}^1 + \mathbf{W}_1\mathbf{\Lambda}_1^{-1}\right]^{-1}\mathbf{\Lambda}_1^{-2}(\boldsymbol{\psi}_1 - \mathbf{u}_1)$$

$$\frac{\partial \hat{\mathbf{v}}_{ij}}{\partial \boldsymbol{\lambda}_2} = -\left[\boldsymbol{Z}^{2T}\tilde{\mathbf{W}}\boldsymbol{Z}^2 + \mathbf{W}_2\mathbf{\Lambda}_2^{-1}\right]^{-1}\mathbf{\Lambda}_2^{-2}(\boldsymbol{\psi}_2 - \mathbf{u}_2)$$

We define the following derivatives, which will be used in the calculations:

$$\begin{aligned}\frac{\partial \tilde{\mathbf{W}}}{\partial \boldsymbol{\theta}} =& \mathbf{W} + \frac{M'''(\boldsymbol{\theta})}{M(\boldsymbol{\theta})} - 3\frac{M'(\boldsymbol{\theta})M''(\boldsymbol{\theta})}{M(\boldsymbol{\theta})^2} + 2\left(\frac{M'(\boldsymbol{\theta})}{M(\boldsymbol{\theta})}\right)^3 \\ \frac{\partial \tilde{\mathbf{W}}}{\partial \boldsymbol{\mu}} =& \frac{\partial \tilde{\mathbf{W}}}{\partial \boldsymbol{\theta}} \text{diag}(\frac{1}{\boldsymbol{\mu}})\end{aligned}.$$

Finally we compute:

$$\begin{aligned}\frac{\partial \log(\tilde{\mathbf{W}})}{\partial \log(\lambda_1)} =& \frac{1}{\tilde{\mathbf{W}}}\frac{\partial \tilde{\mathbf{W}}}{\partial \boldsymbol{\mu}}\text{diag}(\boldsymbol{\mu})\text{diag}\left(\boldsymbol{Z}^1\frac{\partial \hat{\mathbf{v}}_i}{\partial \lambda_1}\lambda_1\right) \\ \frac{\partial \log(\mathbf{W}_1)}{\partial \log(\lambda_1)} =& \frac{\partial \mathbf{W}_1}{\partial \mathbf{u}_i}\text{diag}\left(\frac{\partial \hat{\mathbf{v}}_i}{\partial \lambda_1}\lambda_1\right) \\ \frac{\partial \log(\mathbf{W}_2)}{\partial \log(\lambda_2)} =& \frac{\partial \mathbf{W}_2}{\partial \mathbf{u}_{ij}}\text{diag}\left(\frac{\partial \hat{\mathbf{v}}_{ij}}{\partial \lambda_2}\lambda_2\right)\end{aligned}$$

This allows the proper calculation of the prior weights q_i.

Computation of the deviance residuals for the response in the truncated Poisson model

The following is the classical definition of the deviance residual Pierce and Schafer (1986):

$$2\int_{b'(\hat{\theta})}^{Y} \frac{Y - b'(\theta)}{V(\mu)} db'(\theta),$$

where $b'(\theta) = \mu + \frac{\mu exp(-\mu)}{1-exp(-\mu)}$. Further, $b'(\hat{\theta})$ is the estimate of $b'(\theta)$ and $\hat{\mu}$ is the estimate of μ. Make the following substitution:

$$b'(\theta) = \mu + \frac{\mu exp(-\mu)}{1 - exp(-\mu)}.$$

After the substitution and some derivations we need to evaluate the following integral:

$$2\int_{\hat{\mu}}^{Y'} \frac{Y - \mu - \frac{\mu exp(-\mu)}{1-exp(-\mu)}}{\mu} d\mu,$$

where $Y = Y' + \frac{Y' exp(-Y')}{1-exp(-Y')}$. The final expression for the deviance residuals is as follows:

$$2\left\{Ylog\left(\frac{Y'}{\hat{\mu}}\right) - (Y' - \hat{\mu}) - log[1 - exp(-Y')] + log[1 - exp(-\hat{\mu})]\right\}$$

This form of the deviance residuals we use in the paper.

Acknowledgments

We would like to thank Inge Logghe and Petra EM Zeeuwe of the Department of General Practice of the Erasmus Medical Centrum in Rotterdam for providing the Tai-Chi Chuan intervention data. We also acknowledge prof. Dimitris Rizopoulos for comments on the draft of the manuscript. Further, thanks goes to two anonymous referees for their valuable comments.

References

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

Cox, D. (1970). *Analysis of Binary Data.* Methuen, London.

Cox, D. and Hinkley, D. (1974). *Theoretical Statistics.* Chapman and Hall, London.

Cragg, J. (1971). Some statistical models for limited dependent variables with application to the demand for durable goods. *Econometrica*, **39**, 829–844.

Lee, Y. and Nelder, J. (1996). Hierarchical generalized linear models (with discussion). *Journal of the Royal Statistical Society. Series B (Methodological)*, **58**, 619–678.

Lee, Y. and Nelder, J. (2001). Hierarchical generalised linear models: A synthesis of generalised linear models, random-effect models and structured dispersions. *Biometrika*, **88**, 987–1006.

Lee, Y., Nelder, J., and Pawitan, Y. (2006). *Generalized Linear Models with Random Effects*. Chapman & Hall / CRC, Boca Raton.

Lesaffre, E. and Spiessens, B. (2001). On the effect of the number of quadrature points in a logistic random-effects model: an example. *Applied Statistics*, **50**, 325–335.

Liu, L. and Yu, Z. (2008). A likelihood reformulation method in non-normal random effects models. *Statistics in Medicine*, **27**, 3105–3124.

Min, Y. and Agresti, A. (2005). Random effect models for repeated measures of zero-inflated count data. *Statistical Modelling*, **5**, 1–19.

Mullahy, J. (1986). Specification and testing of some modified count data models. *Journal of Econometrics*, **33**, 341–365.

Mwalili, S., Lesaffre, E., and Declerck, D. (2008). The zero-inflated negative binomial regression model with correction for misclassification: an example in caries research. *Statistical Methods in Medical Research*, **17**, 123–139.

Noh, M. and Lee, Y. (2007). REML estimation for binary data in GLMMs. *Journal of Multivariate Analysis*, **98**, 896–915.

Pierce, D. and Schafer, D. (1986). Residuals in generalized linear models. *Journal of the American Statistical Association*, **81**, 977–86.

Singh, S. (1963). A note on inflated Poisson distribution. *Journal of the Indian Statistical Association*, **1**, 140–144.

Chapter
6

Joint Hierarchical Random Effects Models with Gaussian Random Effects

Based on:
Molas M., Noh M., Lee Y., and Lesaffre E. (2012). Joint hierarchical generalized linear models with multivariate Gaussian random effects. Submitted to Computational Statistics and Data Analysis.

Abstract

Likelihood based inference for correlated data involves the evaluation of a marginal likelihood integrating out random effects. In general this integral does not have a closed form. Moreover, its numerical evaluation might create difficulties especially when the dimension of random effects is high. H-likelihood inference has been proposed where the explicit evaluation of integral is avoided, it also allows extensions handling e.g.(1) complex design experiments, (2) REML type of inference beyond the class of a linear model and (3) overdispersion modelling. Here we extend the h-likelihood approach to multivariate generalized linear mixed models. We blend the

h-likelihood computational algorithms with a Newton-Raphson procedure for the estimation of the correlation parameters. This allows that components of the joint model are interlinked via correlated Gaussian random effects. Further, correlated random effects are allowed within each component. This approach can serve as a basis for further developments of joint double hierarchical generalized linear models with correlated random effects. The methods are illustrated with a rheumatoid arthritis study dataset, where the correlation between latent trajectories of three endpoints is evaluated.

6.1 Introduction

The h-likelihood estimation technique originated in Lee and Nelder (1996). In that paper, the authors investigated the possibility to use the extended likelihood for estimation and inference of models with random effects. The special type of an extended likelihood where a random effect appears linearly in the linear predictor is called hierarchical likelihood or h-likelihood. We refer to Lee and Nelder (2005) for an extensive discussion of h-likelihood. In Lee and Nelder (1996) also the class of the hierarchical generalized linear models (HGLM) class was defined. The HGLM is an extension of the generalized linear model (GLM) by adding random effects, which can follow any distribution from the conjugate Bayesian priors class.

In Lee and Nelder (1996) the authors proposed an augmented iterated weighted least squares algorithm (IWLS) to estimate the fixed and random effects from the hierarchical likelihood, given the variance components which are obtained from the adjusted profile likelihood. In a subsequent paper Lee and Nelder (2001a) developed a numerical procedure based on interlinked IWLS algorithms for fixed, random effects and variance components. This turned out to be a natural way for modelling dispersion or overdispersion parameters in a regression manner. Further it allowed the use of an extended quasi likelihood concept (Nelder and Pregibon, 1987). A further extension of the algorithm was proposed in Noh and Lee (2007) for a less biased estimation of the fixed effects of a mean structure using an adjusted profile likelihood. These procedures were, however, developed for independent random effects.

Correlated random effects are discussed in Lee and Nelder (2001b). Correlation between the random components is essential in the definition of joint models, where components of the model are linked via correlated random effects. A bivariate binary-normal joined model was described in Yun and Lee (2004). Another joint model in the h-likelihood framework can be found in Ha *et al.* (2003) combining a time to event endpoint and a longitudinal Gaussian outcome. There the correlation between two models is handled by a shared random effect.

In this paper we present computational details of a joint hierarchical generalized linear model (JHGLM), where the number of endpoints may vary and follow the HGLM concept, and the correlations between random effects are estimated using a Newton-Raphson algorithm. Random effects are allowed to be correlated between the models as well as within a model. We applied the method to a rheumatoid arthritis data set, where it is of interest to replace one marker measured by a physician by self-reported markers. In this study we model jointly two binary outcomes and one Gaussian outcome. To further illustrate our computational method we also analyzed a bivariate simulated dataset, with a Poisson and Gaussian outcome as the first example, and Poisson and binary outcome as the second example.

The paper is structured as follows. Section 6.2 describes the theory of a joint model and summarizes the h-likelihood method. Section 6.3 presents the estimation of a joint model in a h-likelihood framework. Section 6.4 applies the method to the rheumatoid arthritis study and two simulated datasets. In Section 6.5 concluding remarks are given. The appendix contains details of the numerical computations.

6.2 Joint model

To analyze a longitudinal study with one endpoint the class of linear mixed models (LMM) or generalized linear mixed models (GLMM) can be used. H-likelihood estimation was first applied to responses of an exponential distribution blended with independent conjugate Bayesian random effects (Lee and Nelder, 1996, 2001a). This class of models constitutes the hierarchical generalized linear models (HGLM). HGLMs allow for overdispersion parameters to be included in the model. There are two different ways to model overdispersion in HGLMs (Lee and Nelder, 2000).

For a univariate HGLM the extended likelihood can be written as, see Lee *et al.* (2006):

$$L_E(\boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{v}|\mathbf{y}, \mathbf{v}) = \prod_{i=1}^{N} \prod_{j=1}^{n_i} f_{\boldsymbol{\beta},\boldsymbol{\lambda}}(y_{ij}|\mathbf{v}_i) f_{\boldsymbol{\lambda}}(\mathbf{v}_i), \tag{6.1}$$

where y_{ij} is the observation of the i^{th} subject $(i = 1 \dots N)$ at the j^{th} time point $(j = 1 \dots n_i)$, $\boldsymbol{\beta}$ is the vector of fixed effects in the mean structure, $\mathbf{v}_i$ is the vector of latent random effects pertaining to the i^{th} subject, and $\boldsymbol{\lambda}$ contains variances of the random effects and residual variances (overdispersion parameters) of the response. Parameter $\mathbf{v}$ appears in $L_E(\boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{v}|\mathbf{y}, \mathbf{v})$ on the left side as unknowns to be estimated and on the right side as latent factors generating the data. A gentle introduction on the use of (6.1) in univariate HGLMs estimation can be found in Molas and Lesaffre (2011), where the method is also contrasted to the classical approach based on the marginal likelihood.

6.2.1 The joint HGLM

Now we write the univariate longitudinal model as a joint HGLM (JHGLM) with the extended likelihood:

$$\begin{aligned} L_E(\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\rho}, \mathbf{v}|\mathbf{y}, \mathbf{v}) &= \prod_{i=1}^{N} \prod_{j=1}^{n_i} f_{\boldsymbol{\beta},\boldsymbol{\lambda}}(y_{1ij}|\mathbf{v}_{1i}) \dots f_{\boldsymbol{\beta},\boldsymbol{\lambda}}(y_{kij}|\mathbf{v}_{ki}) \dots f_{\boldsymbol{\beta},\boldsymbol{\lambda}}(y_{Kij}|\mathbf{v}_{Ki}) \times \\ &\times \quad f_{\boldsymbol{\lambda}}(\mathbf{v}_{1i}, \dots, \mathbf{v}_{ki}, \dots, \mathbf{v}_{Ki}). \end{aligned} \tag{6.2}$$

The above extended likelihood is a simple multiplication of univariate likelihoods in (6.1), and as such there is no difference whether the models (indexed by k) are fitted jointly or separately. Let $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_k, \dots, \boldsymbol{\beta}_K)^T$ and note that $\boldsymbol{\beta}_k$ is the specific parameter vector for k^{th} univariate model, as also for $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_1, \dots, \boldsymbol{\lambda}_k, \dots, \boldsymbol{\lambda}_K)^T$.

Here we will consider the class of joint models as e.g. in Fieuws and Verbeke (2006), and impose a multivariate normal distribution on the latent variables:

$$f_{\boldsymbol{\lambda}}(\mathbf{v}_{1i}, \ldots, \mathbf{v}_{ki}, \ldots, \mathbf{v}_{Ki}) = \mathcal{MVN}_{\boldsymbol{\lambda},\boldsymbol{\rho}}(\mathbf{v}_{1i}, \ldots, \mathbf{v}_{Ki}). \tag{6.3}$$

where the vector $\boldsymbol{\lambda}$ represents the variance components and $\boldsymbol{\rho}$ represents the correlations of the joint multivariate normal distribution. To apply h-likelihood methods, the appropriate adjusted profile likelihoods of (6.2) must be used for estimation of $\boldsymbol{\lambda}$, $\boldsymbol{\rho}$ and $\boldsymbol{\beta}$. While the random effects are restricted to a normal distribution, the responses y_{kij} may have a distribution belonging to the exponential family i.e. normal, binomial, Poisson or gamma, possibly varying with k.

6.2.2 Estimation of fixed effects in the mean structure

In the general case, Noh and Lee (2007) suggest to work with a Laplace approximation to the marginal likelihood to estimate fixed effects in the mean structure. This involves the evaluation of the adjusted profile (log)-likelihood:

$$p_v(h) = h(\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\rho}, \mathbf{v})|_{\mathbf{v}=\hat{\mathbf{v}}} - 0.5 \log \left| \frac{D(h, \mathbf{v})}{2\pi} \right|_{\mathbf{v}=\hat{\mathbf{v}}}, \tag{6.4}$$

with $D(h, \mathbf{v}) = -\frac{\partial^2 h(\boldsymbol{\beta},\boldsymbol{\lambda},\boldsymbol{\rho},\mathbf{v})}{\partial \mathbf{v}^T \partial \mathbf{v}}$ and $h(\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\rho}, \mathbf{v}) = \log(L_E(\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\rho}, \mathbf{v}|\mathbf{y}, \mathbf{v}))$. The term ‘adjusted profile likelihood’ is chosen since $h(\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\rho}, \mathbf{v})|_{\mathbf{v}=\hat{\mathbf{v}}}$ is a profile (log)-likelihood of $\boldsymbol{\beta}$, variance parameters $\boldsymbol{\lambda}$ and correlation parameters $\boldsymbol{\rho}$, profiled over $\mathbf{v}$, and the second term of (6.4) is a correction term to approximate the marginal log-likelihood. This profile likelihood is used only to estimate $\boldsymbol{\beta}$ for a given $\boldsymbol{\lambda}$ and $\boldsymbol{\rho}$. Note that $\hat{\mathbf{v}}$ is the maximum likelihood estimator of $\mathbf{v}$ for given $(\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\rho})$ computed using h as the objective function, i.e. one could write $\hat{v} = \hat{v}_{\boldsymbol{\beta},\boldsymbol{\lambda},\boldsymbol{\rho}}$.

6.2.3 Estimation of correlations and variances of random components

To estimate the correlations $\boldsymbol{\rho}$ and variances $\boldsymbol{\lambda}$, a Newton-Raphson algorithm needs to be applied to maximize the appropriate adjusted profile likelihood. Let the marginal distribution of the data $\mathbf{y}$ be $f_{\boldsymbol{\beta},\boldsymbol{\lambda},\boldsymbol{\rho}}(\mathbf{y})$ (marginalized over $\mathbf{v}$), seen as a probability density function (pdf) of the data. Conditional on the sufficient statistics for $\boldsymbol{\beta}$, denoted as $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},\boldsymbol{\rho}}$, (see Cox and Hinkley (1974), page 21), the (marginalized) distribution of the data can be derived from:

$$f_{\boldsymbol{\lambda}}(\mathbf{y}|\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},\boldsymbol{\rho}}) = \frac{f_{\boldsymbol{\beta},\boldsymbol{\lambda},\boldsymbol{\rho}}(\mathbf{y})}{f_{\boldsymbol{\beta},\boldsymbol{\lambda},\boldsymbol{\rho}}(\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},\boldsymbol{\rho}})}, \tag{6.5}$$

where $f_{\boldsymbol{\beta},\boldsymbol{\lambda},\boldsymbol{\rho}}(\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},\boldsymbol{\rho}})$ is the distribution of $\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},\boldsymbol{\rho}}$. After additional steps (see Molas and Lesaffre (2010)) it can be shown that $\log(f_{\boldsymbol{\lambda}}(\mathbf{y}|\hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda},\boldsymbol{\rho}}))$ can be approximated by the following adjusted profile likelihood:

$$p_{\beta,v}(h) = h(\boldsymbol{\beta},\boldsymbol{\lambda},\boldsymbol{\rho},\mathbf{v})|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}},\mathbf{v}=\hat{\mathbf{v}}} - 0.5\log\left|\frac{D\left[h,(\boldsymbol{\beta},\mathbf{v})\right]}{2\pi}\right|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}},\mathbf{v}=\hat{\mathbf{v}}}, \tag{6.6}$$

where

$$D\left[h,(\boldsymbol{\beta},\mathbf{v})\right] = -\begin{pmatrix} \frac{\partial^2 h}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T} & \frac{\partial^2 h}{\partial\boldsymbol{\beta}\partial\mathbf{v}^T} \\ \frac{\partial^2 h}{\partial\mathbf{v}\partial\boldsymbol{\beta}^T} & \frac{\partial^2 h}{\partial\mathbf{v}\partial\mathbf{v}^T} \end{pmatrix}.$$

The latter adjusted profile likelihood is maximized with respect to $\boldsymbol{\lambda}$ and $\boldsymbol{\rho}$ to obtain $\hat{\boldsymbol{\lambda}}$ and $\hat{\boldsymbol{\rho}}$. Note that this objective function is “focussed” solely on the dispersion and correlation parameters. This offers an extension of restricted maximum likelihood (REML) estimation and provides inference for the class of generalized linear mixed models, see Noh and Lee (2007).

6.2.4 Residual dispersion and overdispersion

In (6.2) parameter $\boldsymbol{\lambda}$ contains not only the variance and correlation parameters of the random effects, but also dispersion or overdispersion parameter of each of the

K responses. In the case of a Gaussian distribution, the variance is the dispersion parameter of the response, while for binomial or Poisson data, one could include an overdispersion parameter into the model. To estimate residual dispersion parameters equation (6.6) can be used, while in case of overdispersion the h-likelihood needs to be replaced by double extended quasi likelihood (Lee and Nelder, 2001a).

6.2.5 Overview of the estimation procedure

The estimation of parameters in the h-likelihood way is the following:

1. Set $\boldsymbol{\lambda}^s$ and $\boldsymbol{\rho}^s$ at some starting values i.e. variances and correlations of the random effects and residual dispersion (overdispersion) parameters.

2. Given $\boldsymbol{\lambda}^s$ and $\boldsymbol{\rho}^s$ estimate latent variables (random effects) at maximum $\hat{\mathbf{v}}$ using (6.2) and fixed effects $\hat{\boldsymbol{\beta}}$ either using (6.2) or (6.4) where appropriate.

3. Evaluate (6.6) at new $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{v}}$ and compute $\boldsymbol{\lambda}^{s+1}$ and $\boldsymbol{\rho}^{s+1}$.

4. Iterate steps 2-3 until convergence.

6.3 The h-likelihood approach for a JHGLM

In the previous section we have defined a class of joint models, for which we will now present details on the h-likelihood estimation approach. The h-likelihood approach is a restricted maximum likelihood method (REML) of Noh and Lee (2007) extended here to joint models. Difficulty in estimation of JHGLMs with h-likelihood is two-fold: (1) extended likelihood (6.1) needs to be replaced by (6.2) and (2) additional algorithm is needed to estimate variance components and correlations of (6.3). In this section we present practical and algorithmic details which are needed to implement the scheme presented in Section 6.2.

6.3.1 Estimation of the mean structure parameters

Suppose the correlations and variance components are known, i.e. $\boldsymbol{\lambda}$ and $\boldsymbol{\rho}$ are known. Our objective is to find the estimates of fixed and random effects in the mean structure. In order to use the h-likelihood methods presented in Lee and Nelder (2001a) and Molas and Lesaffre (2011) we need to replace (6.1) by (6.2). In the algorithm, this is accomplished by an appropriate stacking of the matrices as in Lee *et al.* (2006). Suppose that we have a model with three responses, each with a random intercept and a random slope. Let $\boldsymbol{Y}_k$ denote the outcome vector of the k^{th} endpoint stacked over subjects, similarly $\mathbf{X}_k$ denotes the design matrix of the k^{th} outcome. Further, $\boldsymbol{Z}_{k1}$ is the design matrix of random intercepts and $\boldsymbol{Z}_{k2}$ is the design matrix of random slopes for the k^{th} endpoint. The design matrices of random effects are a Kronecker product of a identity matrix of the dimension equal to a number of subjects and random intercept or random time vector per subject. This implies the following matrices for the JHGLM:

$$\boldsymbol{Y} = \begin{pmatrix} \boldsymbol{Y}_1 \\ \boldsymbol{Y}_2 \\ \boldsymbol{Y}_3 \end{pmatrix}, \qquad \mathbf{X} = \begin{pmatrix} \mathbf{X}_1 & 0 & 0 \\ 0 & \mathbf{X}_2 & 0 \\ 0 & 0 & \mathbf{X}_3 \end{pmatrix}, \qquad \boldsymbol{Z} = \begin{pmatrix} \boldsymbol{Z}_{11} & \boldsymbol{Z}_{12} & 0 & 0 & 0 & 0 \\ 0 & 0 & \boldsymbol{Z}_{21} & \boldsymbol{Z}_{22} & 0 & 0 \\ 0 & 0 & 0 & 0 & \boldsymbol{Z}_{31} & \boldsymbol{Z}_{32} \end{pmatrix}.$$

Now we can define the matrices leading to the composition of the augmented IWLS (Lee and Nelder, 2001a). The same algorithm can be applied just with the above stacked matrices. Define:

$$\mathbf{T} = \begin{pmatrix} \mathbf{X} & \boldsymbol{Z} \\ \mathbf{0} & \mathbf{I} \end{pmatrix}, \qquad \boldsymbol{\Sigma}_a = \begin{pmatrix} \mathbf{W}^{-1}\boldsymbol{\Phi} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_R \end{pmatrix},$$

where $\mathbf{I}$ is the identity matrix of dimension equal to the number of columns in $\boldsymbol{Z}$, each row of $\mathbf{I}$ corresponding to a random effect. The weight matrix $\mathbf{W}$ is diagonal with elements w_{kij}, where $w_{kij} = V(\mu_{kij})^{-1}(\frac{\partial \mu_{kij}}{\partial \eta_{kij}})^2$ and $V(\mu_{kij})$ is the variance function of the k^{th} endpoint and j^{th} observation of the i^{th} subject; μ_{kij} is the mean parameter respectively and η_{kij} is the linear predictor. The matrix $\boldsymbol{\Phi}$ is a diagonal matrix of ϕ_{kij}, which represent residual error variances or overdispersion parameters. Further, $\boldsymbol{\Sigma}_R$ is the variance covariance matrix of random effects. In our case we have

6 correlated random effects, and q subjects therefore we have together $6 \times q$ estimates of random effects. Therefore if follows that:

$$\mathbf{\Sigma}_R = \begin{pmatrix} \lambda_{11} & \rho_1\sqrt{\lambda_{11}\lambda_{12}} & \rho_2\sqrt{\lambda_{11}\lambda_{21}} & \rho_3\sqrt{\lambda_{11}\lambda_{22}} & \rho_4\sqrt{\lambda_{11}\lambda_{31}} & \rho_5\sqrt{\lambda_{11}\lambda_{32}} \\ \rho_1\sqrt{\lambda_{11}\lambda_{12}} & \lambda_{12} & \rho_6\sqrt{\lambda_{12}\lambda_{21}} & \rho_7\sqrt{\lambda_{12}\lambda_{22}} & \rho_8\sqrt{\lambda_{12}\lambda_{31}} & \rho_9\sqrt{\lambda_{12}\lambda_{32}} \\ \rho_2\sqrt{\lambda_{11}\lambda_{21}} & \rho_6\sqrt{\lambda_{12}\lambda_{21}} & \lambda_{21} & \rho_{10}\sqrt{\lambda_{21}\lambda_{22}} & \rho_{11}\sqrt{\lambda_{21}\lambda_{31}} & \rho_{12}\sqrt{\lambda_{21}\lambda_{32}} \\ \rho_3\sqrt{\lambda_{11}\lambda_{22}} & \rho_7\sqrt{\lambda_{12}\lambda_{22}} & \rho_{10}\sqrt{\lambda_{21}\lambda_{22}} & \lambda_{22} & \rho_{13}\sqrt{\lambda_{22}\lambda_{31}} & \rho_{14}\sqrt{\lambda_{22}\lambda_{32}} \\ \rho_4\sqrt{\lambda_{11}\lambda_{31}} & \rho_8\sqrt{\lambda_{12}\lambda_{31}} & \rho_{11}\sqrt{\lambda_{21}\lambda_{31}} & \rho_{13}\sqrt{\lambda_{22}\lambda_{31}} & \lambda_{31} & \rho_{15}\sqrt{\lambda_{31}\lambda_{32}} \\ \rho_5\sqrt{\lambda_{11}\lambda_{32}} & \rho_9\sqrt{\lambda_{12}\lambda_{32}} & \rho_{12}\sqrt{\lambda_{21}\lambda_{32}} & \rho_{14}\sqrt{\lambda_{22}\lambda_{32}} & \rho_{15}\sqrt{\lambda_{31}\lambda_{32}} & \lambda_{32} \end{pmatrix} \bigotimes \mathbf{I}_q. \tag{6.7}$$

The augmented IWLS algorithm (Lee *et al.*, 2006) can be invoked with updating equation to obtain estimates of fixed and random effects:

$$\mathbf{T}^T\mathbf{\Sigma}_a^{-1}\mathbf{T}\boldsymbol{\xi} = \mathbf{T}^T\mathbf{\Sigma}_a^{-1}\mathbf{z}, \tag{6.8}$$

with $\boldsymbol{\xi} = (\boldsymbol{\beta}, \mathbf{v})^T$ and $\mathbf{z} = \boldsymbol{\eta} + (\mathbf{y} - \boldsymbol{\mu})(\partial\boldsymbol{\eta}/\partial\boldsymbol{\mu})$. Updating these equations until convergence will yield estimates of the fixed and random effects based on (6.2). Using the above defined matrices the procedure in Noh and Lee (2007) can be applied for the joint model to use the Laplace approximation to (6.2) as the objective function to find fixed effects. We will not describe this procedure here in detail, but refer to Noh and Lee (2007) and Molas and Lesaffre (2010) for computational and theoretical details in somewhat different settings, which can be easily extended to JHGLM by using the above defined matrices.

6.3.2 Estimation of the random effects covariance matrix

Previous section shows how to estimate the fixed and random effects in the mean structure, given the dispersion components. Here we show how to estimate the dispersion components in $\mathbf{\Sigma}_R$ using a Newton - Raphson algorithm. In order to assure that the q_R-dimensional matrix $\mathbf{\Sigma}_R$ is positive definite we will take the Cholesky decomposition of the matrix:

$$\mathbf{T}_R\mathbf{\Sigma}_R\mathbf{T}_R^T = \mathbf{D}_R \qquad \text{or} \qquad \mathbf{\Sigma}_R^{-1} = \mathbf{T}_R^T\mathbf{D}_R^{-1}\mathbf{T}_R, \tag{6.9}$$

where $\mathbf{T}_R$ is a lower triangular matrix with ones on diagonal and t_{ij} parameters in the lower triangle with $i = 2, \ldots, q_R$ and $j = 1, \ldots, (i-1)$. Further, $\mathbf{D}_R$ is a diagonal matrix with entries $d^*_{ii} = \exp(d_{ii})$. The idea of the decomposition originates from Pourahmadi (1999), and has been used by Cecere *et al.* (2006) in the context of covariance modelling. The parameters $\boldsymbol{\gamma}_R = (d_{ii}, t_{ij})$ $(i = 1, \ldots, q_R \;\; j = 1, \ldots, (i - 1))$ determine the matrix $\boldsymbol{\Sigma}_R$ and are estimated by a Newton-Raphson algorithm as follows:

1. Take a starting point vector $\boldsymbol{\gamma}^s_R$;

2. Compute the adjusted likelihood value (6.6) at $\boldsymbol{\gamma}^s_R$ and estimated fixed $\boldsymbol{\beta}^s_R$ and random effects $\mathbf{v}^s_R$ using the procedure described in Section 6.3.1;

3. Compute the score vector $\mathbf{S} = (\partial p_{\beta,v}(h)/\partial \boldsymbol{\gamma}_R)$ and hessian matrix $\mathbf{H} = (\partial^2 p_{\beta,v}(h)/\partial \boldsymbol{\gamma}_R \partial \boldsymbol{\gamma}^T_R)$ and evaluate it at $\boldsymbol{\gamma}^s_R$, $\boldsymbol{\beta}^s_R$ and $\mathbf{v}^s_R$.

4. Check if $-\mathbf{H}$ is positive definite. If not, compute the closest positive definite negative of the Hessian using the method of Higham (2002). This is done in an iterative manner based on an eigenvalues decomposition of the matrix, negative eigenvalues are set to zeros, as well as their corresponding eigenvectors. New eigenvectors and eigenvalues are used to create a new matrix. The procedure is repeated until convergence.

5. Update $\boldsymbol{\gamma}^{s+1}_R$ using:
$$\boldsymbol{\gamma}^{s+1}_R = \boldsymbol{\gamma}^s_R - \lambda_R \mathbf{H}^{-1}\mathbf{S},$$
where λ_R is equal to one;

6. Evaluate $p_{\beta,v}(h)$ at new point $\boldsymbol{\gamma}^{s+1}_R$, if the likelihood increased accept the proposal, otherwise set $\lambda_R = \lambda_R/2$ and create a new proposal, repeat until the new proposal is accepted;

7. After the new proposal $\boldsymbol{\gamma}^{s+1}_R$ is accepted, find new values of the regression parameters $\boldsymbol{\beta}^{s+1}_R$ and random parameters $\mathbf{v}^{s+1}_R$.

8. Update the value of residual (over)dispersion parameters as described in Section 6.2.4. Further details of the estimation procedures are given in the Appendix.

The computation of the score and Hessian matrices is explained in the Appendix. After updating the variance covariance matrix of the random effects, the residual dispersion components should be estimated using the approach of Chapter 7 of Lee *et al.* (2006), adapted to the JHGLM setting. This can be done by stacking matrices, and computing deviance residuals based on the distribution of each $\boldsymbol{Y}_i$.

6.3.3 Missing components

In the rheumatoid arthritis example described below, two individuals were not examined by the physician at any of the visits. Therefore, there are no data on one of the endpoints for two patients, yet they filled-in the questionnaires at all time points. We present here a trick to make it possible to avoid deletion of all the observations from these two subjects. The trick is to add observations in a way that they change the likelihood only up to the constant. Let us consider the Gaussian likelihood:

$$\log\left[f(y_{ij}|v_i)\right] = -0.5\log(2\pi\phi_{ij}) - 0.5\exp\left[\frac{(y_{ij}-\mu_{ij})^2}{\phi_{ij}}\right].$$

If we set $y_{ij}=0$ and $\mu_{ij}=0$, the new observation will not influence the estimation of the fixed structure parameters, but will allow the estimation of the random effect itself for this subject. This can be achieved by adding a row to the matrices $\boldsymbol{Y}_i$, $\mathbf{X}_i$, $\boldsymbol{Z}_{i1}$ and $\boldsymbol{Z}_{i2}$ composed only of zeros. To avoid that likelihood of the 2 subjects contributes to the estimation of residual dispersion, we set zeros to all covariates of the dispersion structure. If there is only an intercept as the design matrix in the estimation of the residual dispersion, the value one needs to be replaced by zero.

For a binary likelihood

$$\log\left[f(y_{ij}|v_i)\right] = y_{ij}\log(\mu_{ij}) + (1-y_{ij})\log(1-\mu_{ij}), \tag{6.10}$$

an observation does not contribute to the likelihood when we set $\mu_{ij} = 0.5$ again by adding a row of zeros to $\mathbf{X}_i$, $\boldsymbol{Z}_{1i}$ and $\boldsymbol{Z}_{2i}$. Note that there is no residual dispersion parameter in the Bernoulli distribution.

6.4 Illustration of the computational approach

6.4.1 Rheumatoid Arthritis Study

The example is the rheumatoid arthritis Patients rePort Onset Reactivation sTudy (RAPPORT study), which is a longitudinal study that aims to identify an increase in disease activity by self-reported questionnaires. A cohort of 159 patients is followed throughout one year. Each month these patients fill in at home a questionnaire (most of them use a web-based form), and each three months a clinical evaluation by the treating rheumatologist is performed. We consider here the self-reported measures at the time the clinical evaluation is done and examine the association of the clinical evaluation with the self-assessment questionnaires.

The self reported instruments used here are the Health Assessment Questionnaires (HAQ) and the Rheumatoid Arthritis Disease Activity Index (RADAI). Two other measurements scored on a visual analogue scale instruments (VAS) will be ignored in this paper. The HAQ measures the functional status of the patient using 20 questions from 8 categories. Each category is based on daily physical functioning activities such as dressing, rising, eating etc. The HAQ score ranges from 0 to 3, but in this study we will use a binary version of HAQ, as suggested by Aletaha *et al.* (2006). When HAQ is above 0.5 an increased activity of the disease is indicated and we classify this as ‘remission’, otherwise it is called ‘no remission’. The second self-reported endpoint (RADAI) contains 5 items, e.g. today’s disease activity in terms of swollen and tender joints or today's amount of arthritis pain. Originally RADAI is scored on a Likert scale varying between 0 and 10, but again we will use a binarized version of it with cut-off point of 2.2, as suggested in Fransen and van Riel (2009). RADAI values lower than 2.2 indicate a stable disease, while above indicate increased activity of the disease.

The examination done by the clinician was recorded in the Disease Activity Score of 28 joint counts (DAS28), which is a composite score of swollen joints count, tender joints count, a visual analog scale of the patient's assessment of general health and erythrocyte sedimentation rate at the first hour. The DAS28 score varies between 0 and 10. In this study we treat it as a Gaussian response.

The clinical and self-reported measurements taken at months $0, 3, 6, 9, 12$ were considered in this paper. Therefore, information is provided on three endpoints DAS28, HAQ at RADAI, 5 times over the span of one year. However not all measurements were recorded as planned, and the missing data that occurred here are assumed to happen according to a missing at random (MAR) mechanism. Finally gender and baseline age of the patients were included in the analysis as covariates.

Denote DAS28 as $\boldsymbol{Y}_1$, HAQ as $\boldsymbol{Y}_2$ and RADAI as $\boldsymbol{Y}_3$. As covariates we use the intercept, month of measurement, sex and age at baseline. Therefore each $\mathbf{X}_k$, where $k = 1, 2, 3$ has four columns. There are 159 patients in the study, therefore $i = 1, 2, \ldots, 159$ and $j = 1, 2, \ldots, 5$ as there are five visits. Note that not all patients gave information for each k^{th} endpoint, as well not all patients were measured at each of the 5 visits. We will consider the following multivariate model with 3 endpoints:

$$\begin{aligned} Y_{1ij} \mid v_{11i}, v_{12i} &\sim \mathcal{N}(X_{1ij}\boldsymbol{\beta}_1 + v_{11i} + v_{12i} * time_{ij}, \phi^2), \\ Y_{2ij} \mid v_{21i} &\sim \text{Bernoulli}\left(\frac{\exp(X_{2ij}\boldsymbol{\beta}_2 + v_{21i})}{1 + \exp(X_{2ij}\boldsymbol{\beta}_2 + v_{21i})}\right), \\ Y_{3ij} \mid v_{31i} &\sim \text{Bernoulli}\left(\frac{\exp(X_{3ij}\boldsymbol{\beta}_3 + v_{31i})}{1 + \exp(X_{3ij}\boldsymbol{\beta}_3 + v_{31i})}\right). \end{aligned}$$

We consider a model for DAS28 with random intercept and slope, while HAQ and RADAI have only random intercepts. Note that we were not able to fit models for

RADAI and HAQ in a univariate setting with correlated intercept and slope, but neither was this possible with SAS PROC NLMIXED and WINBUGS. The reason might be that in most clusters the binary response stayed at 0 or 1 for the whole period. Thus we assume a four dimensional latent structure:

$$\begin{pmatrix} v_{11i} \\ v_{12i} \\ v_{21i} \\ v_{31i} \end{pmatrix} \sim \mathcal{MVN} \left[\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \lambda_{11} & \rho_1\sqrt{\lambda_{11}\lambda_{12}} & \rho_2\sqrt{\lambda_{11}\lambda_{21}} & \rho_5\sqrt{\lambda_{11}\lambda_{31}} \\ \rho_1\sqrt{\lambda_{11}\lambda_{12}} & \lambda_{12} & \rho_4\sqrt{\lambda_{12}\lambda_{21}} & \rho_5\sqrt{\lambda_{12}\lambda_{31}} \\ \rho_2\sqrt{\lambda_{11}\lambda_{21}} & \rho_4\sqrt{\lambda_{12}\lambda_{21}} & \lambda_{21} & \rho_6\sqrt{\lambda_{21}\lambda_{31}} \\ \rho_3\sqrt{\lambda_{11}\lambda_{31}} & \rho_5\sqrt{\lambda_{12}\lambda_{31}} & \rho_6\sqrt{\lambda_{21}\lambda_{31}} & \lambda_{31} \end{pmatrix} \right] \tag{6.11}$$

We performed the analysis to evaluate the degree to which the DAS28 measurement correlates with HAQ and RADAI status. We are interested in correlations between the latent profiles of the measures. In case of binary measurements the latent profiles are constant linear on the linear predictor scale. In case of DAS28 latent profiles are linear with intercept and slope defined by two random effects v_{11} and v_{12}. The parameters determining the correlation of latent profiles are ρ parameters in (6.11). As two patients were not scored for DAS28 at any of the 5 time points, we used the trick described in (6.3.3) by adding two fictive observations. Further, the DAS28 random slope has a design matrix in years. The results are presented in Table 6.1. We also present the result of estimation when, three responses were fitted separately.

We are primarily interested in the correlations between the latent profiles of the endpoints. The correlation between HAQ and RADAI random intercepts was the greatest and was equal to 0.85. This indicates that two measures are highly correlated on the latent scale, implying that where the latent score is high on RADAI it is also high on HAQ over the 5 visits. The correlation between the random intercept and slope of DAS28 was equal to -0.16, being an indication that patients starting with a worse condition might have improved faster, than the patients with a better initial disease status. The correlation, however is small. DAS28 slope did not correlate strongly with the random intercept of HAQ and the random intercept of RADAI with correlations 0.12 and 0.09 respectively. The correlation between the random intercept of DAS28 and the random intercepts of HAQ and RADAI were estimated as 0.61 and 0.63. This is an indication that the latent level of DAS28 predicts moderately strong the latent level of RADAI and HAQ.

To summarize, the latent intercepts of the three outcomes play a major role, HAQ and RADAI seem to be equivalent binary measures whether disease progresses or is retained at current level, and they might as such be treated exchangeably. Further, the level of latent DAS28, predicts moderately well the status of a patient obtained from self-assessment.

Table 6.1: Results of the h-likelihood JHGLM fit: Rheumatoid Arthritis

Correlated Joint Model				
Fixed Effects				
	Estimate	S.E.	z-score	p-value
DAS28				
Intercept	1.773	0.413	4.293	<0.0001
age	0.008	0.006	1.333	0.182
sex=f	0.760	0.198	3.838	<0.0001
time	-0.010	0.008	-1.250	0.211
HAQ				
Intercept	-7.500	2.726	-2.751	0.006
age	0.073	0.042	1.730	0.084
sex=f	3.156	1.300	2.428	0.015
time	0.024	0.034	0.698	0.485
RADAI				
Intercept	-2.095	0.978	-2.142	0.032
age	0.018	0.015	1.200	0.230
sex=f	0.460	0.458	1.004	0.315
time	0.018	0.023	0.783	0.434
Correlation Matrix				
	DAS RI	DAS RS	HAQ RI	RADAI RI
DAS RI	1	-0.159	0.614	0.626
DAS RS	-0.159	1	0.117	0.095
HAQ RI	0.614	0.117	1	0.846
RADAI RI	0.626	0.095	0.846	1
Residual Variances				
DAS 28		0.466		

Independent Models for each response				
Fixed Effects				
	Estimate	S.E.	z-score	p-value
DAS28				
Intercept	1.697	0.411	4.129	<0.0001
age	0.009	0.006	1.500	0.134
sex=f	0.799	0.197	4.056	<0.0001
time	-0.011	0.008	-1.375	0.169
HAQ				
Intercept	-9.002	2.347	-3.835	<0.0001
age	0.010	0.037	2.631	0.009
sex=f	3.238	1.106	2.926	0.003
time	0.023	0.033	0.692	0.489
RADAI				
Intercept	-2.283	0.964	-2.367	0.018
age	0.022	0.015	1.479	0.139
sex=f	0.361	0.449	0.804	0.421
time	0.015	0.023	0.663	0.508
Correlation Matrix				
	DAS RI	DAS RS	HAQ RI	RADAI RI
DAS RI	1	-0.206		
DAS RS	-0.206	1		
HAQ RI			1	
RADAI RI				1
Residual Variances				
DAS 28		0.453		

6.4.2 Simulated data

We also use a simulated dataset to further exemplify our procedures. In a longitudinal study one might be interested whether the change in the albumin level correlates with the number of the loss of balance episodes in patients taking certain medications. We have assumed to follow a cohort of 100 patients over time. Each individual was measured at four visits $0, 1, 2, 3$ in the hospital, and at each visit 2 responses were measured: a count Y_1 and a Gaussian response Y_2. Y_1 might represent the loss of balance episodes generated from a Poisson distribution. The continuous Gaussian response Y_2 might represent the change in albumin level. We have used the following model to generate the data:

$$\begin{aligned}
\eta_{1ij} &= -1.5 + 0.5\text{time}_{ij} + v_{10} + v_{11}\text{time}_{ij} \\
\eta_{2ij} &= -1.5 + 0.3\text{time}_{ij} + v_{20} + v_{21}\text{time}_{ij} \\
Y_{1ij} &\sim \text{Poisson}(\exp(\eta_{1ij})) \\
Y_{2ij} &\sim \mathcal{N}(\eta_{2ij}, 0.8^2),
\end{aligned}$$

where $i = 1, 2, \ldots, N$ and $j = 1, 2, \ldots n_i$. The random effects were sampled from the following normal distribution:

$$\begin{pmatrix} v_{10} \\ v_{11} \\ v_{20} \\ v_{21} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0.8 & -0.2 & 0.36 & 0 \\ -0.2 & 0.3 & 0 & 0.2 \\ 0.36 & 0 & 4 & -1 \\ 0 & 0.2 & -1 & 1.5 \end{pmatrix} \right),$$

The results of estimation are shown in Table 6.2. The h-likelihood procedure correctly estimated the evolution of the average patient over time, the number of balance loss episodes increased over time together with the increase of the change in the albumin level. Further, among individuals evolutions of two endpoints over time were positively correlated meaning that if the albumin increase changes were above the average level over time also the number of balance loss episodes was

Table 6.2: Results of the h-likelihood and Bayesian fit to simulated data: Poisson and Gaussian models

H-likelihood				WINBUGS			
Poisson	Intercept	-1.38		Poisson	Intercept	-1.43	
	Time	0.49			Time	0.49	
Gaussian	Intercept	-1.47		Gaussian	Intercept	-1.46	
	Time	0.35			Time	0.35	
Variance Covariance Matrices							
0.16	0.019	0.4	-0.07	0.256	-0.0008	0.37	-0.06
0.019	0.21	0.018	0.15	-0.0008	0.233	0.03	0.15
0.4	0.018	5.06	-1.03	0.37	0.03	5.003	-1.004
-0.07	0.15	-1.03	1.38	-0.06	0.15	-1.004	1.368

above the average evolution over the time. The correlation between random slopes of two responses was estimated as $0.28 = \frac{0.15}{\sqrt{0.21*1.38}}$. We have managed to fit the same model using a Bayesian approach with WINBUGS with uninformative prior distributions. The SAS PROC NLMIXED is able to fit such models in theory, however we were not able to achieve convergence.

The Bayesian approach and the h-likelihood approach gave similar estimates, with the biggest difference in the values of variance covariance matrix in the Poisson model. Note that we take posterior means in the Bayesian case and we compare them with the maximum likelihood estimates.

In the next step we have categorized the Gaussian response into two categories i.e. positive or negative values. This gave the Bernoulli outcome, which we fitted jointly with the Poisson process in the h-likelihood way. We also fitted the data using WINBUGS. First we present the result of h-likelihood and Bayesian estimation for the model with only random intercepts.

In case of JHGLM with a Poisson and Bernoulli outcome the estimation process was more difficult when random intercepts and slopes were in both models. The estimation problem occurred especially when random slope was added for the binary data. In Table 6.3 we present the results of the JHGLM models estimated by h-likelihood and in a Bayesian way. The results for the random intercepts joint model with one correlation are very similar between the h-likelihood and Bayesian

Table 6.3: Results of the h-likelihood and Bayesian fit to simulated data: Poisson and binary models

H-likelihood		(RI)	WINBUGS		(RI)	H-likelihood (RI+RS)			
Poisson	Intercept	-2.31	Poisson	Intercept	-2.31	Poisson	Intercept	-1.37	
	Time	0.93		Time	0.92		Time	0.49	
Binary	Intercept	-1.82	Binary	Intercept	-1.82	Binary	Intercept	-4.69	
	Time	0.54		Time	0.55		Time	1.09	
Variance Covariance Matrices									
						0.12	0.027	2.41	-0.51
	1.28	1.02		1.31	1.02	0.027	0.2	-0.22	0.71
	1.02	6.5		1.02	6.92	2.41	-0.22	55.43	-10.85
						-0.22	0.71	-10.85	11.56

approach. In Tables 6.2 and 6.3 in the Bayesian analysis results we report posterior medians. WinBUGS was not used for the estimation of the random intercept and random slope model. When random slopes were added to the models it becomes cumbersome to estimate the binary model. This is due to low number of within patient measurements, when there are more timepoints per patient the problem is less prevalent. We have simulated data with more measurements per patient and in this case WinBUGS performed better yielding estimates of the joint models, however convergence after 35000 iterations for the variance of random slopes of the binary model can be still in question. Results were similar to those obtained by the h-likelihood method with a biggest discrepancy in the estimation of random slopes variances of binary model, less discrepancy was obtained for the estimation of the Poisson model random slopes variances. H-likelihood gave similar results to both Poisson and binary models fitted separately by standard software for longitudinal data using Laplace approximations to the marginal likelihood, with some difference in the binary model variance parameters. These differences were however much lesser than in case of the comparison of Bayesian and H-likelihood methods.

6.5 Discussion and conclusion

This paper has described the h-likelihood estimation approach for estimating multivariate repeated measures models. It extends the routines beyond the bivariate case, and allows correlated Gaussian random effects of the model components. The endpoint is allowed to be Gaussian, gamma, Poisson or binomial. Fixed effects can be estimated in our approach either by using h-likelihood or an adjusted profile likelihood.

There are numerous extensions of the presented procedure which we would like to develop further. First, an extension of covariance modelling to depend on covariates as in Pourahmadi (1999) is of interest. This could extend the univariate case of independent random effects. Next, addition of independent random effects of a conjugate Bayesian distribution would be of interest next to the correlated Gaussian random effects already present in the JHGLM. Joint double HGLM (JDHGLM) could be defined by allowing for an inclusion of correlated random effects in the variance functions. JDHGLM model with covariance modelling would provide much flexibility in the modelling assumptions.

It is important to mention the technical limitations of our R software. First, memory problems make it difficult to fit data sets of large dimensions and we will probably need sparse matrix computation methods. The second computational improvement might by speeding up the Newton-Raphson algorithm for estimating of the correlations and variances of random effects distribution.

Upon having solved the above issues, we would have a very strong computational approach to the modelling of multivariate longitudinal data. This would be an alternative to a marginal likelihood or the Bayesian approach, extending them further e.g. through the use of extended quasi likelihood functions or the availability of covariance modelling. We regard the current paper as a contribution allowing to fit JHGLMs via h-likelihood, but also as a first step allowing for all the above extensions.

Acknowledgments

We would like to thank Margot Walter and dr Jolanda Luime for providing the Rheumatoid Arthritis dataset for this paper.

Appendix

In the appendix we show how to compute the Hessian and the score matrix with respect to parameters estimated in the Newton-Raphson algorithm of Section 6.3.2.

Decomposition and derivatives of variance covariance matrix

First we give the details of the variance covariance matrix $\mathbf{\Sigma}_R$ decomposition according to Pourahmadi (1999). The matrix is decomposed as follows:

$$\mathbf{T}_R\mathbf{\Sigma}_R\mathbf{T}_R^T = \mathbf{D}_R.$$

If $\mathbf{\Sigma}_R$ is a 3 by 3 matrix we have the following matrices:

$$\mathbf{T}_R = \begin{pmatrix} 1 & 0 & 0 \\ t_{21} & 1 & 0 \\ t_{31} & t_{32} & 1 \end{pmatrix} \qquad \mathbf{D}_R = \begin{pmatrix} d_{11}^* & 0 & 0 \\ 0 & d_{22}^* & 0 \\ 0 & 0 & d_{33}^* \end{pmatrix},$$

further $d_{ii}^* = \exp(d_{ii})$. The Newton-Raphson algorithm optimizes the likelihood with respect to parameters (t_{ij}, d_{ii}). Therefore, derivatives with respect to these parameters are of interest. However first we will show derivatives with respect to (t_{ij}, d_{ii}^*). First, lets look at the derivatives $\frac{\partial \mathbf{D}_R}{\partial d_{ii}^*}$, and $\frac{\partial \mathbf{T}_R}{\partial t_{ij}}$.

$$\frac{\partial \mathbf{D}_R}{\partial d_{22}^*} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} \qquad \frac{\partial \mathbf{T}_R}{\partial t_{12}} = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

The second derivative of the above matrices with respect to any other parameters of the parameters itself, is a matrix of zeros. Now denote by $\mathbf{L} = \mathbf{T}^{-1}$, we shall drop the subscript R from notation of $\mathbf{T}$ and $\mathbf{D}$. The following derivatives need to be computed:

$$
\begin{aligned}
\frac{\partial \boldsymbol{\Sigma}_R}{\partial d^*_{ii}} &= \mathbf{L}\frac{\partial \mathbf{D}}{\partial d^*_{22}}\mathbf{L}^T \\
\frac{\partial \boldsymbol{\Sigma}_R}{\partial t_{ij}} &= -\mathbf{L}\frac{\partial \mathbf{T}}{\partial t_{ij}}\mathbf{LDL}^T - \mathbf{LDL}^T\frac{\partial \mathbf{T}}{\partial t_{ij}}\mathbf{L}^T \\
\frac{\partial^2 \boldsymbol{\Sigma}_R}{\partial d^*_{ii}\partial d^*_{jj}} &= \mathbf{L}\frac{\partial^2 \mathbf{D}}{\partial d^*_{ii}\partial d^*_{jj}}\mathbf{L}^T = \mathbf{0} \\
\frac{\partial^2 \boldsymbol{\Sigma}_R}{\partial d^*_{ii}\partial t_{kl}} &= -\mathbf{L}\frac{\partial \mathbf{T}}{\partial t_{kl}}\mathbf{L}\frac{\partial \mathbf{D}}{\partial d^*_{ii}}\mathbf{L}^T - \mathbf{L}\frac{\partial \mathbf{D}}{\partial d^*_{ii}}\mathbf{L}^T\frac{\partial \mathbf{T}}{\partial t_{kl}}\mathbf{L}^T \\
\frac{\partial^2 \boldsymbol{\Sigma}_R}{\partial t_{ij}\partial t_{kl}} &= \mathbf{L}\frac{\partial \mathbf{T}}{\partial t_{ij}}\mathbf{L}\frac{\partial \mathbf{T}}{\partial t_{kl}}\mathbf{LDL}^T + \mathbf{L}\frac{\partial \mathbf{T}}{\partial t_{kl}}\mathbf{L}\frac{\partial \mathbf{T}}{\partial t_{ij}}\mathbf{LDL}^T \\
&+ \mathbf{L}\frac{\partial \mathbf{T}}{\partial t_{kl}}\mathbf{LDL}^T\frac{\partial \mathbf{T}}{\partial t_{ij}}^T\mathbf{L}^T + \mathbf{L}\frac{\partial \mathbf{T}}{\partial t_{ij}}\mathbf{LDL}^T\frac{\partial \mathbf{T}}{\partial t_{kl}}^T\mathbf{L}^T \\
&+ \mathbf{LDL}^T\frac{\partial \mathbf{T}}{\partial t_{ij}}^T\mathbf{L}^T\frac{\partial \mathbf{T}}{\partial t_{kl}}^T\mathbf{L}^T + \mathbf{LDL}^T\frac{\partial \mathbf{T}}{\partial t_{kl}}^T\mathbf{L}^T\frac{\partial \mathbf{T}}{\partial t_{ij}}^T\mathbf{L}^T
\end{aligned}
$$

The above derivatives are used in the further computations to compute the score vector of the adjusted profile likelihood and hessian matrix with respect to the parameters of interest (d^*_{ii}, t_{ij}).

Derivatives of the adjusted profile likelihood

Lets use the motivating example of 3 endpoints with 4 multivariate random effects to demonstrate the computation of the score vector and the hessian matrix.

$$
\begin{aligned}
h &= \sum_{i=1}^{N}\sum_{j=1}^{n_{1i}}\left[\frac{y_{1ij}\theta_{1ij} - b(\theta_{1ij})}{\phi_{1ij}} + c_1(y_{1ij}, \phi_{1ij})\right] + \sum_{i=1}^{N}\sum_{j=1}^{n_{2i}}\left[\frac{y_{2ij}\theta_{2ij} - b(\theta_{2ij})}{\phi_{2ij}} + c_2(y_{2ij}, \phi_{2ij})\right] \\
&+ \sum_{i=1}^{N}\sum_{j=1}^{n_{3i}}\left[\frac{y_{3ij}\theta_{3ij} - b(\theta_{3ij})}{\phi_{3ij}} + c_3(y_{3ij}, \phi_{3ij})\right] - 2N\log(2\pi) - \frac{N}{2}\log(\det \boldsymbol{\Sigma}_R) - \frac{1}{2}(\mathbf{v}^T\boldsymbol{\Sigma}_R\mathbf{v}),
\end{aligned}
$$

note that in the above expression the multivariate normal distribution is for all random effects at once, without a summation sign over the subjects, therefore the variance covariance matrix is of form shown in (6.7) but with 4 random effects instead of 6. The adjusted profile likelihood maximized with respect to (t_{ij}, d_{ii}) is:

$$p_{\beta,v}(h) = h(\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\rho}, \mathbf{v})|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}, \mathbf{v}=\hat{\mathbf{v}}} - 0.5 \log \left| \frac{D\left[h, (\boldsymbol{\beta}, \mathbf{v})\right]}{2\pi} \right|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}, \mathbf{v}=\hat{\mathbf{v}}},$$

and

$$DD = D\left[h, (\boldsymbol{\beta}, \mathbf{v})\right] = -\begin{pmatrix} \frac{\partial^2 h}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} & \frac{\partial^2 h}{\partial \boldsymbol{\beta} \partial \mathbf{v}^T} \\ \frac{\partial^2 h}{\partial \mathbf{v} \partial \boldsymbol{\beta}^T} & \frac{\partial^2 h}{\partial \mathbf{v} \partial \mathbf{v}^T} \end{pmatrix}.$$

The derivative of the adjusted profile likelihood is:

$$\begin{aligned} \frac{\partial p_{\beta,v}(h)}{\partial d^*_{ii}} &= \frac{\partial \hat{\mathbf{v}}^T}{\partial d^*_{ii}} \boldsymbol{Z}^T \left(\frac{\boldsymbol{Y} - \boldsymbol{\mu}}{\boldsymbol{\Phi}} \right) - \frac{1}{2} \text{trace} \left\{ \boldsymbol{\Sigma}_R^{-1} \frac{\partial \boldsymbol{\Sigma}_R}{\partial d^*_{ii}} \right\} - \frac{\partial \hat{\mathbf{v}}^T}{\partial d^*_{ii}} \boldsymbol{\Sigma}_R^{-1} \hat{\mathbf{v}}^T \\ &+ \frac{1}{2} \hat{\mathbf{v}}^T \boldsymbol{\Sigma}_R^{-1} \frac{\partial \boldsymbol{\Sigma}_R}{\partial d^*_{ii}} \boldsymbol{\Sigma}_R^{-1} \mathbf{v} - \frac{1}{2} \text{trace} \left\{ DD^{-1} \frac{\partial DD}{\partial d^*_{ii}} \right\}, \end{aligned}$$

where all matrices are total matrices stacked over all three models. The hessian matrix has the following expression:

$$\begin{aligned} \frac{\partial^2 p_{\beta,v}(h)}{\partial d^*_{ii} \partial d^*_{jj}} &= -\frac{\partial \hat{\mathbf{v}}^T}{\partial d^*_{ii}} \boldsymbol{Z}^T \mathbf{W} \boldsymbol{\Phi}^{-1} \frac{\partial \hat{\mathbf{v}}}{\partial d^*_{jj}} + \frac{\partial^2 \hat{\mathbf{v}}^T}{\partial d^*_{ii} \partial d^*_{jj}} \boldsymbol{Z}^T \left(\frac{\boldsymbol{Y} - \boldsymbol{\mu}}{\boldsymbol{\Phi}} \right) \\ &- \frac{1}{2} \text{trace} \left\{ \boldsymbol{\Sigma}_R^{-1} \frac{\partial^2 \boldsymbol{\Sigma}_R}{\partial d^*_{ii} \partial d^*_{jj}} \right\} + \frac{1}{2} \text{trace} \left\{ \boldsymbol{\Sigma}_R^{-1} \frac{\partial \boldsymbol{\Sigma}_R}{\partial d^*_{ii}} \boldsymbol{\Sigma}_R^{-1} \frac{\partial \boldsymbol{\Sigma}_R}{\partial d^*_{jj}} \right\} \\ &+ \frac{\partial \hat{\mathbf{v}}^T}{\partial d^*_{ii}} \boldsymbol{\Sigma}_R^{-1} \frac{\partial \boldsymbol{\Sigma}_R}{\partial d^*_{jj}} \boldsymbol{\Sigma}_R^{-1} \hat{\mathbf{v}} - \frac{\partial \hat{\mathbf{v}}^T}{\partial d^*_{ii}} \boldsymbol{\Sigma}_R^{-1} \frac{\partial \hat{\mathbf{v}}}{\partial d^*_{jj}} - \frac{\partial^2 \hat{\mathbf{v}}^T}{\partial d^*_{jj} \partial d^*_{ii}} \boldsymbol{\Sigma}_R^{-1} \hat{\mathbf{v}} \\ &+ \frac{\partial \hat{\mathbf{v}}^T}{\partial d^*_{jj}} \boldsymbol{\Sigma}_R^{-1} \frac{\partial \boldsymbol{\Sigma}_R}{\partial d^*_{ii}} \boldsymbol{\Sigma}_R^{-1} \hat{\mathbf{v}} - \hat{\mathbf{v}}^T \boldsymbol{\Sigma}_R^{-1} \frac{\partial \boldsymbol{\Sigma}_R}{\partial d^*_{jj}} \boldsymbol{\Sigma}_R^{-1} \frac{\partial \boldsymbol{\Sigma}_R}{\partial d^*_{ii}} \boldsymbol{\Sigma}_R^{-1} \hat{\mathbf{v}} \\ &+ \frac{1}{2} \hat{\mathbf{v}}^T \boldsymbol{\Sigma}_R^{-1} \frac{\partial^2 \boldsymbol{\Sigma}_R}{\partial d^*_{ii} \partial d^*_{jj}} \boldsymbol{\Sigma}_R^{-1} \hat{\mathbf{v}} - \frac{1}{2} \text{trace} \left\{ DD^{-1} \frac{\partial^2 DD}{\partial d^*_{ii} \partial d^*_{jj}} \right\} \\ &+ \frac{1}{2} \text{trace} \left\{ DD^{-1} \frac{\partial DD}{\partial d^*_{ii}} DD^{-1} \frac{\partial DD}{\partial d^*_{jj}} \right\}. \end{aligned}$$

To compute derivatives with respect to t_{ij} the above formulas can be used as well. As we have $d^*_{ii} = \exp(d_{ii})$ the derivative with respect to d_{ii} is as follows

$$\begin{aligned}
\frac{\partial p_{\beta,v}(h)}{\partial d_{ii}} &= \frac{\partial p_{\beta,v}(h)}{\partial d^*_{ii}} \frac{\partial d^*_{ii}}{\partial d_{ii}} \\
\frac{\partial^2 p_{\beta,v}(h)}{\partial d_{ii} \partial d_{jj}} &= \frac{\partial^2 p_{\beta,v}(h)}{\partial d^*_{ii} \partial d^*_{jj}} \frac{\partial d^*_{ii}}{\partial d_{ii}} \frac{\partial d^*_{jj}}{\partial d_{jj}} + \frac{\partial p_{\beta,v}(h)}{\partial d^*_{ii}} \frac{\partial^2 d^*_{ii}}{\partial d_{ii} \partial d_{jj}}
\end{aligned}$$

In the following two sections we explain how to compute derivatives of estimates of random effects with respect to parameters of interest and the derivatives of adjustment term.

Derivatives of the estimates of the random effects

Denote by $TT_2 = \boldsymbol{Z}^T \mathbf{W} \boldsymbol{\Phi}^{-1} \boldsymbol{Z} + \boldsymbol{\Sigma}_R^{-1}$, we have:

$$\begin{aligned}
\frac{\partial \hat{\mathbf{v}}}{\partial d^*_{ii}} &= TT_2^{-1} \boldsymbol{\Sigma}_R^{-1} \frac{\partial \boldsymbol{\Sigma}_R}{\partial d^*_{ii}} \boldsymbol{\Sigma}_R^{-1} \hat{\mathbf{v}} \\
-TT_2 \frac{\partial^2 \hat{\mathbf{v}}}{\partial d^*_{ii} \partial d^*_{jj}} &= \boldsymbol{Z}^T \mathbf{W} \boldsymbol{\Phi}^{-1} \frac{\partial \mathbf{W}}{\partial \boldsymbol{\mu}} \mathrm{diag}\left(\boldsymbol{Z} \frac{\partial \hat{\mathbf{v}}}{\partial d^*_{ii}} \right) \boldsymbol{Z} \frac{\partial \hat{\mathbf{v}}}{\partial d^*_{jj}} \\
&- \boldsymbol{\Sigma}_R^{-1} \frac{\partial \boldsymbol{\Sigma}_R}{\partial d^*_{ii}} \boldsymbol{\Sigma}_R \frac{\partial \hat{\mathbf{v}}}{\partial d^*_{jj}} - \boldsymbol{\Sigma}_R^{-1} \frac{\partial \boldsymbol{\Sigma}_R}{\partial d^*_{jj}} \boldsymbol{\Sigma}_R \frac{\partial \hat{\mathbf{v}}}{\partial d^*_{ii}} \\
&+ \boldsymbol{\Sigma}_R^{-1} \frac{\partial \boldsymbol{\Sigma}_R}{\partial d^*_{ii}} \boldsymbol{\Sigma}_R^{-1} \frac{\partial \boldsymbol{\Sigma}_R}{\partial d^*_{jj}} \boldsymbol{\Sigma}_R^{-1} \hat{\mathbf{v}} + \boldsymbol{\Sigma}_R^{-1} \frac{\partial \boldsymbol{\Sigma}_R}{\partial d^*_{jj}} \boldsymbol{\Sigma}_R^{-1} \frac{\partial \boldsymbol{\Sigma}_R}{\partial d^*_{ii}} \boldsymbol{\Sigma}_R^{-1} \hat{\mathbf{v}} \\
&- \boldsymbol{\Sigma}_R^{-1} \frac{\partial^2 \boldsymbol{\Sigma}_R}{\partial d^*_{ii} \partial d^*_{jj}} \boldsymbol{\Sigma}_R^{-1} \hat{\mathbf{v}}
\end{aligned}$$

Derivatives of the adjustment term of the profile likelihood

The adjustment matrix DD has the following form:

$$DD = \begin{pmatrix} \mathbf{X}^T \mathbf{W} \boldsymbol{\Phi}^{-1} \mathbf{X} & \mathbf{X}^T \mathbf{W} \boldsymbol{\Phi}^{-1} \boldsymbol{Z} \\ \boldsymbol{Z}^T \mathbf{W} \boldsymbol{\Phi}^{-1} \mathbf{X} & \boldsymbol{Z}^T \mathbf{W} \boldsymbol{\Phi}^{-1} \boldsymbol{Z} + \boldsymbol{\Sigma}_R^{-1} \end{pmatrix}$$

The first order derivative of this matrix is as follows:

$$\begin{aligned}
UpD &= \frac{\partial \mathbf{W}}{\partial \boldsymbol{\mu}} \mathbf{W} \boldsymbol{Z} \frac{\partial \hat{\mathbf{v}}}{\partial d^*_{ii}} \\
DnD &= -\boldsymbol{\Sigma}_R^{-1} \frac{\partial \boldsymbol{\Sigma}_R}{\partial d^*_{ii}} \boldsymbol{\Sigma}_R^{-1} \\
\frac{\partial DD}{\partial d^*_{ii}} &= \begin{pmatrix} \mathbf{X}^T UpD \boldsymbol{\Phi}^{-1} \mathbf{X} & \mathbf{X}^T UpD \boldsymbol{\Phi}^{-1} \boldsymbol{Z} \\ \boldsymbol{Z}^T UpD \boldsymbol{\Phi}^{-1} \mathbf{X} & \boldsymbol{Z}^T UpD \boldsymbol{\Phi}^{-1} \boldsymbol{Z} + DnD \end{pmatrix}
\end{aligned}$$

Finally second derivative has the following form:

$$\begin{aligned}
UpD &= \frac{\partial^2 \mathbf{W}}{\partial \boldsymbol{\mu}^2} \mathbf{W} \mathbf{W} \mathrm{diag}\left(\boldsymbol{Z} \frac{\partial \hat{\mathbf{v}}}{\partial d^*_{ii}} \right) \mathrm{diag}\left(\boldsymbol{Z} \frac{\partial \hat{\mathbf{v}}}{\partial d^*_{jj}} \right) + \frac{\partial \mathbf{W}}{\partial \boldsymbol{\mu}} \frac{\partial \mathbf{W}}{\partial \boldsymbol{\mu}} \mathbf{W} \mathrm{diag}\left(\boldsymbol{Z} \frac{\partial \hat{\mathbf{v}}}{\partial d^*_{ii}} \right) \mathrm{diag}\left(\boldsymbol{Z} \frac{\partial \hat{\mathbf{v}}}{\partial d^*_{jj}} \right) \\
&+ \frac{\partial \mathbf{W}}{\partial \boldsymbol{\mu}} \mathbf{W} \mathrm{diag}\left(\boldsymbol{Z} \frac{\partial^2 \hat{\mathbf{v}}}{\partial d^*_{ii} \partial d^*_{jj}} \right) \\
DnD &= \boldsymbol{\Sigma}_R^{-1} \frac{\partial \boldsymbol{\Sigma}_R}{\partial d^*_{ii}} \boldsymbol{\Sigma}_R^{-1} \frac{\partial \boldsymbol{\Sigma}_R}{\partial d^*_{jj}} \boldsymbol{\Sigma}_R^{-1} + \boldsymbol{\Sigma}_R^{-1} \frac{\partial \boldsymbol{\Sigma}_R}{\partial d^*_{jj}} \boldsymbol{\Sigma}_R^{-1} \frac{\partial \boldsymbol{\Sigma}_R}{\partial d^*_{ii}} \boldsymbol{\Sigma}_R^{-1} \\
&- \boldsymbol{\Sigma}_R^{-1} \frac{\partial^2 \boldsymbol{\Sigma}_R}{\partial d^*_{ii} \partial d^*_{jj}} \boldsymbol{\Sigma}_R^{-1} \\
\frac{\partial DD}{\partial d^*_{ii}} &= \begin{pmatrix} \mathbf{X}^T UpD \boldsymbol{\Phi}^{-1} \mathbf{X} & \mathbf{X}^T UpD \boldsymbol{\Phi}^{-1} \boldsymbol{Z} \\ \boldsymbol{Z}^T UpD \boldsymbol{\Phi}^{-1} \mathbf{X} & \boldsymbol{Z}^T UpD \boldsymbol{\Phi}^{-1} \boldsymbol{Z} + DnD \end{pmatrix}
\end{aligned}$$

Estimation of the residual dispersion parameters

To estimate the residual dispersion parameters we use the adjusted profile likelihood (6.6). We can express the adjusted profile likelihood derivative with respect to ϕ as follows:

$$\frac{\partial p_{\boldsymbol{\beta},\mathbf{v}}(h)}{\partial \phi} = 0.5 \sum_{i=1}^{N} \sum_{j=1}^{n_i} \left\{ \frac{dres_{ij} - \phi}{\phi^2} + c_{ij}(\phi) \right\} - 0.5 \mathrm{trace} \left\{ DD^{-1} \frac{\partial DD}{\partial \phi} \right\},$$

where $dres_{ij}$ is an appropriate GLM deviance residual for the response distribution. In this derivation we ignore the derivatives $\frac{\partial \hat{\mathbf{v}}}{\partial \phi}$. Now lets define:

$$\mathbf{T} = \begin{pmatrix} \mathbf{X} & \boldsymbol{Z} \\ \mathbf{0} & \mathbf{I} \end{pmatrix} \qquad \boldsymbol{\Sigma}_T = \begin{pmatrix} \boldsymbol{\Phi} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_R \end{pmatrix} \qquad \boldsymbol{\Sigma}_0 = \begin{pmatrix} \boldsymbol{\Phi} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}$$

We have:

$$\frac{\partial p_{\boldsymbol{\beta},\mathbf{v}}(h)}{\partial \phi} = 0.5 \sum_{i=1}^{N} \sum_{j=1}^{n_i} \left\{ \frac{dres_{ij} - \phi}{\phi^2} + c_{ij}(\phi) \right\} + 0.5 \frac{\text{trace}(\mathbf{T}(\mathbf{T}^T \boldsymbol{\Sigma}_T^{-1} \mathbf{T})^{-1} \mathbf{T}^T \boldsymbol{\Sigma}_0^{-1})}{\phi},$$

Now we define q_{ij} as the $ij-th$ diagonal element of the matrix $\mathbf{T}(\mathbf{T}^T \boldsymbol{\Sigma}_T^{-1} \mathbf{T})^{-1} \mathbf{T}^T \boldsymbol{\Sigma}_0^{-1}$. This leads to:

$$\frac{\partial p_{\boldsymbol{\beta},\mathbf{v}}(h)}{\partial \phi} = 0.5 \sum_{i=1}^{N} \sum_{j=1}^{n_i} \left\{ \frac{dres_{ij} - (1 - q_{ij} - c^*_{ij})\phi}{\phi^2} \right\},$$

where $c^*_{ij} = \phi c_{ij}(\phi)$. The adjustment c^*_{ij} occurs only for the gamma distribution and is equal $c^*_{ij} = 1 + 2\frac{\log(\phi)}{\phi} + 2\frac{\text{digamma}(1/\phi)}{\phi}$. Finally we can write the derivative as follows:

$$\frac{\partial p_{\boldsymbol{\beta},\mathbf{v}}(h)}{\partial \phi} = \sum_{i=1}^{N} \sum_{j=1}^{n_i} \frac{(1 - q_{ij} - c^*_{ij})}{2} \left\{ \frac{dres^*_{ij} - \phi}{\phi^2} \right\}, \tag{6.12}$$

where $dres^*_{ij} = \frac{dres_{ij}}{(1-q_{ij}-c_{ij})}$. Equation (6.12) can be maximized by a gamma distributed GLM with $dres^*_{ij}$ as a response and $\frac{(1-q_{ij}-c^*_{ij})}{2}$ as a prior weight. The mean of the distribution is ϕ and variance $\frac{2\phi}{(1-q_{ij}-c^*_{ij})}$.

References

Aletaha, D., Machold, K., Nell, V., and Smolen, J. (2006). The perception of rheumatoid arthritis core set measures by rheumatologists. results of a survey. *Rheumatology*, **45**, 1133–1139.

Cecere, S., Jara, A., and Lesaffre, E. (2006). Analyzing the emergence times of permanent teeth: an example of modeling the covariance matrix with interval-censored data. *Statistical Modelling*, **6**, 1–15.

Cox, D. and Hinkley, D. (1974). *Theoretical Statistics*. Chapman and Hall, London.

Fieuws, S. and Verbeke, G. (2006). Pairwise fitting of mixed models for the joint modeling of multivariate longitudinal profiles. *Biometrics*, **62**, 424–431.

Fransen, J. and van Riel, P. (2009). Outcome measures in inflammatory rheumatic diseases. *Arthritis Research and Therapy*, **11:244**, 1–10.

Ha, I., Park, T., and Lee, Y. (2003). Joint modelling of repeated measures and survival time data. *Biometrical Journal*, **45**, 647–658.

Higham, N. (2002). Computing the nearest correlation matrix - a problem from finance. *IMA Journal of Numerical Analysis*, **22**, 329–343.

Lee, Y. and Nelder, J. (1996). Hierarchical generalized linear models (with discussion). *Journal of the Royal Statistical Society. Series B (Methodological)*, **58**, 619–678.

Lee, Y. and Nelder, J. (2000). Two ways of modeling overdisperion in non-normal data. *Applied Statistics*, **49**, 591–598.

Lee, Y. and Nelder, J. (2001a). Hierarchical generalised linear models: A synthesis of generalised linear models, random-effect models and structured dispersions. *Biometrika*, **88**, 987–1006.

Lee, Y. and Nelder, J. (2001b). Modelling and analysing correlated non-normal data. *Statistical Modelling*, **1**, 3–16.

Lee, Y. and Nelder, J. (2005). Likelihood for random-effect models. *Statistical and Operational Research Transactions*, **29**, 141–182.

Lee, Y., Nelder, J., and Pawitan, Y. (2006). *Generalized Linear Models with Random Effects*. Chapman & Hall / CRC, Boca Raton.

Molas, M. and Lesaffre, E. (2010). Hurdle models for multilevel zero-inflated data via h-likelihood. *Statistics in Medicine*, **29**, 3294–3310.

Molas, M. and Lesaffre, E. (2011). Hierarchical generalized linear models: The R package HGLMMM. *Journal of Statistical Software*, **39(13)**, 1–20.

Nelder, J. and Pregibon, D. (1987). An extended quasi-likelihood function. *Biometrika*, **74**, 221–232.

Noh, M. and Lee, Y. (2007). REML estimation for binary data in GLMMs. *Journal of Multivariate Analysis*, **98**, 896–915.

Pourahmadi, M. (1999). Joint mean-covariance models with applications to longitudinal data: Unconstrained parametrization. *Biometrika*, **86**, 677–690.

Yun, S. and Lee, Y. (2004). Comparison of hierarchical and marginal likelihood estimators for binary outcomes. *Computational Statistics and Data Analysis*, **45**, 639–650.

Chapter 7 Discussion, Conclusions and Further Research

We start in Chapter 2 with the use of the likelihood to model the density of the data given covariates. We developed a h-likelihood inspired algorithm to estimate the Gaussian mixture model. To increase the computational efficiency we have assumed fixed weights of the components of the mixtures. This protects from infinite likelihood problems, which can be encountered in the standard mixture modelling approach. Further it speeds up the computations without constraining the final density form if a few more mixture components are used. We considered independent data and it could be of interest to extend the computational system to a correlated datasets design.

In Chapter 3, we have explored the statistical treatment of bounded outcome scores in a longitudinal setting. As seen in that chapter, the distribution of a BOS is difficult to express by a classical parametric family. In this chapter we have extended the approach in Lesaffre *et al.* (2007), which assumes that the observed data are a coarsened version of latent continuous data, to repeated measures studies by the use of random effects. It turns out, that the likelihood of our approach may have an equivalent likelihood formulation in terms of an ordinal probit model.

However, in the proposed coarsening model less parameters need to be estimated. A simulation study showed that the statistical performance of the two approaches is similar when the data follow the coarsening model. However, the coarsening model can be computationally much more efficient. As further research, it might of interest to see how the h-likelihood approach can replace the marginal likelihood and see if there is some additional computational advantage.

In Chapter 4, we have implemented the algorithm of Lee and Nelder (2001) and Lee and Nelder (1996) in an **R** package **HGLMMM**. In this package, we allow the response to follow a Gaussian, binomial, Poisson or gamma distribution in combination with conjugate Bayesian random effects. The dispersion parameters of response as well as random effects can be easily modelled as a function of covariates and the procedures allow for complex designs. However, the implementation does not yet allow for correlated random effects, and limitations due to computer memory might occur for large data sets with many random effects. The first problem has a theoretical basis, as it seems difficult to accommodate neatly the estimation of correlations in the h-likelihood numerical procedures. The second problem is a result of choosing **R** as a programming environment and limitation of computers itself, and might be easier to circumvent.

The computational methods of hierarchical generalized linear models were extended in Chapter 5 for the modelling of zero-inflated data. We adjusted the h-likelihood computational methods to the hurdle model with random effects. This model consists of two parts, the first part represents a binary model and second part a truncated Poisson model. We have shown that the distribution of the truncated Poisson model can be expressed as an exponential family. The theory for the binomial model has already been available and we have adjusted the algorithms to work for the truncated Poisson model. All advantages of the h-likelihood computational approach apply. However, we did not allow for correlated random effects, but assumed independent random effects between two components of the joint model. Extending the h-likelihood model to correlated data became then the topic of our further research, and we have described joint models with correlated Gaussian random effects in Chapter 6. However, the introduction of correlation breaks down the neat h-likelihood algorithm and a Newton-Raphson step is required to estimate the variance covariance matrix of the correlated random effects. Given these parameter estimates, standard procedures can be adapted to estimate the rest of the unknowns. However the ability to estimate the joint hierarchical generalized linear models with correlated random effects by h-likelihood extends the REML concept of estimation to a wider class of models. In the HGLM joint models the computational issues are more severe than in the models with independent random effects, which is due to the need of an optimal code for the correlations. Further, as more responses in a longitudinal setting are analyzed over time, the dimensions of the matrices are increasing as well creating possibly a computer storage problem.

Further work on h-likelihood could be focussed on the hurdle model with correlated random effects, which could be achieved by combining the developments in Chapters 5 and 6. Another useful extension would be the modelling of the variance covariance matrix of random effects as a function of covariates. These extensions, require additional software modification, but from the theoretical point of view most of the work is in place. The hurdle model is of course just one of the possible models for analyzing correlated count data, and we are also interested to examine the h-likelihood implementation of various overdispersed/correlated count data models. Another area that has been neglected in the h-likelihood approach is modeling ordinal responses. This is definitely a topic of interest for further development. This may be achieved by first developing the computational framework in an h-likelihood context of BOS responses.

References

Lee, Y. and Nelder, J. (1996). Hierarchical generalized linear models (with discussion). *Journal of the Royal Statistical Society B*, **58**, 619–678.

Lee, Y. and Nelder, J. (2001). Hierarchical generalised linear models: A synthesis of generalised linear models, random-effect models and structured dispersions. *Biometrika*, **88**, 987–1006.

Lesaffre, E., Rizopoulos, D., and Tsonaka, R. (2007). The logistic-transform for bounded outcome scores. *Biostatistics*, **8**, 72–85.

Summary

In this thesis we scrutinize the use of likelihood based models for analysis of (longitudinal) complex data. We compare the h-likelihood estimation procedures to the standard approach, and apply h-likelihood algorithms to a wider class of problems. Models for mixture of distributions, zero-inflated count data, bounded outcome scores responses or multivariate longitudinal data are described and extended with an application of h-likelihood type solutions.

Chapter 1 describes in brief theoretical foundations to use the likelihood as a basis for estimation and inference. We present the properties of maximum likelihood estimators, and where these properties originate from. The likelihood based tests properties and derivations are discussed in detail. The other likelihood based tests i.e. Wald test and likelihood ratio test are asymptotically equivalent. We discuss the extension of the use of likelihood in repeated data models and introduce the concept of h-likelihood.

In **Chapter 2** we describe another possibility to define a mixture modelling system. We propose to fix the contribution of mixture components (weights) and do not estimate it. While we leave flexibility in a choice of number of components and the parameters of each of the components. This system might be characterized by an improved computational stability. In case of Gaussian components of the mixture system, mean of each mixture component and its variance can be modelled as a function of covariates e.g. by use of splines. The idea of joint modelling of mean and dispersion is based on the h-likelihood methods. We present the application of this theory to model the height of boys as a smooth function of age in the Fourth Dutch Growth Study. Limited simulations and hypothetical examples are presented.

Chapter 3 describes the modelling of longitudinal bounded outcome score response. Such an outcome is obtained, when certain trait is measured repeatedly for the same patient. The score is typically bounded between minimum and maximum value e.g. $0 - 100$, and often rounded to the nearest important number, which are located with equally spaced distances on the span of the score e.g. $0, 10, 20, \ldots, 100$.

Such a score might be characterized by an unusual distribution as U-shaped or J-shaped. In this chapter we compare three methods to analyze such a response using marginal likelihood. Standard method to analyze such a data could be ordinal probit model for repeated measures data. Additionally we propose to perform a logit transformation on the original score (scaled before that to $0-1$ interval) and (1) apply a mixed model on the transformed data (2) assume that the data follows mixed model on the latent scale and observed data gives information about the interval where the true latent score could lay. We compare the proposed approaches to the ordinal probit model through simulations. Application of the method is presented to the analysis of the stroke revalidation study (CERISE), where the speed of revalidation after stroke is compared between two centers.

We introduce the h-likelihood method for random effects models in **Chapter 4**, first we contrast it to the marginal likelihood approach. In the remainder of the chapter we describe our software for fitting hierarchical generalized linear models using h-likelihood algorithms. Numerous examples are presented. This chapter is focussed on practical application of a software to perform h-likelihood estimation of random effects models. The approach allows complex designs of random components e.g. crossed random effects, multilevel structures, cross-over designs. Random effects are assumed to follow a conjugate baysian family of distributions, therefore allowing distributions beyond normal. Variance components are easily modelled as a function of covariates, as well as overdispersion parameter which is estimated using extended quasi-likelihood concepts. Finally REML estimation concept is extend to exponential family distributions beyond the case of linear mixed model.

In **Chapter 5** we extend existing h-likelihood algorithm to allow the estimation of zero-inflated count data. We adapt h-likelihood approach for hurdle models. In hurdle model, the zero count is modeled separately from the positive counts. The zero count follows a binary distribution, while positive counts are assumed to follow truncated Poisson distribution. As a matter of fact the hurdle model can be looked upon as a joint model of binary response and truncated Poisson response. In this chapter we assume that both parts of the model can be estimated separately,

therefore only independent random effects are allowed. We present the application of the method to the Tai-Chi Chuan clinical trial, where it was of interest whether this type of exercise for elderly people helps prevent the number of falls they experience.

Chapter 6 extends h-likelihoods algorithms to handle joint models of multiple exponential family longitudinal endpoints. These endpoints are joined through correlated random effects. We blend the use of Newton-Raphson algorithm for the estimation of the correlation of the random effects with h-likelihood procedures. Multivariate random effects models can be handled now in h-likelihood approach. This work is a basis for further extension to joint double hierarchical generalized linear models or to modelling variance covariance matrix as a function of covariates. Further, work of Chapter 5 and Chapter 6 can be connected resulting in hurdle model with correlated random effects.

Chapter 7 gives discussion, conclusions and further research ideas.

Samenvatting

In dit proefschrift onderzoeken we het gebruik van op aannemelijkheid gebaseerde modellen voor de analyse van (longitudinale) complexe data. We vergelijken de H-aannemelijkheidsschattingsprocedure met de standaardaanpak en passen h-aannemelijkheids algoritmes toe op een bredere klasse problemen. Modellen voor mengsels van verdelingen, zero inflated count data, begrensde uitkomsten of multivariaat longitudinale data worden beschreven en uitgebreid met een toepassing van H-aannemelijkheid.

Hoofdstuk 1 beschrijft in het kort de theoretische grondslagen van het gebruik van aannemelijkheid als basis van statistische schattings- en toetsingsmethoden. We presenteren de eigenschappen van maximale-aannemelijkheidsschatters en leggen uit waar deze op gegrond zijn. We leggen in detail uit waar de toetsen gebaseerd op de aannemelijkheidsmethode vandaankomen en wat hun eigtenschappen zijn. Andere toetsen gebaseerd op de aannemelijksfuntie (Wald en aannemelijkheidsratio) zijn asymetritisch equivalent. We bespreken hoe aannemelijkheidstheorie uitgebreid kan worden tot herhaalde metingen en introduceren H-aannemelijkheid.

In **hoofdstuk 2** bespreken we een andere manier om systemen van mengsels van verdelingen te modelleren. We stellen voor om de gewichten van de verschillende componenten vast te zetten en dus niet te schatten, terwijl we wel vrijheid behouden en de keuze van het aantal componenten en de parameters van de afzonderlijke componenten. Op deze manier behalen we een verbeterde numerieke stabiliteit. Wanneer de componenten normaal verdeeld zijn kunnen we het gemiddelde en de variantie van de componenten door middel van een spline modelleren. Het idee om het gemiddelde en de spreiding gezamenlijk te modelleren is gebaseerd op H-aannemelijkheid. We presenteren een toepassing van deze theorie waar we de lengte van jongens modelleren als een gladde functie van hun leeftijd gebruikmakend van gegevens uit de Vierde Nederlandse Groei Studie. Ook worden enkele voorbeelden en simulaties gepresenteerd.

Hoofdstuk 3 beschrijft het modelleren van longitudinale begrensde uitkomsten.

Een dergelijke uitkomst kan verkregen worden wanneer een eigenschap herhaaldelijk wordt gemeten. Vaak zit de score van de uitkomst tussen een minimum en een maximum (bijvoorbeeld 0 - 100) in en wordt afgerond naar een aantal getallen die op gelijke afstand verdeeld zijn binnen dit interval (bijvoorbeeld 0, 10, 20, ..., 100). Vaak heeft zo'n score een U- of J-vormige verdeling. In dit hoofdstuk vergelijken we drie methoden om zulke uitkomsten te modeleren gebruikmakend van de marginale aannemelijkheid. De standaard methode om dergelijke data te analyseren is een ordinaal probit model voor herhaalde uitkomsten. In aanvulling hierop bekijken we ook (1) een model waarbij we de data eerst zo schalen dat ze in het interval van 0 tot 1 liggen, waarna we de logit transformatie toepassen en de resulterende scores modelleren met een gemengde effecten model. En (2) een model waarbij we een latente score veronderstellen die met een gemengd effecten model kan worden beschreven en de waargenomen data informatie geeft over de waarde van de latente score. Deze methoden worden toegepast op data van de CERISE studie waar de snelheid van het herstel na een hersenbloeding wordt vergeleken tussen twee centra.

We introduceren de H-aannemelijkheidsmethode voor random effecten modellen in **hoofdstuk 4**. We vergelijken deze methode eerst met de methode van de marginale aannemelijkheid. In de rest van het hoofdstuk beschrijven we onze software waarin het H-aannemelijkheids algoritme wordt geïmplementeerd. Verscheidene voorbeelden worden gepresenteerd. Dit hoofdstuk is gericht op de praktische applicatie de software om met H-aannemelijkheid random effect modellen te schatten. Deze aanpak staat complexe designs van de random effecten zoals gekruiste effecten, meerdere niveaus, cross-over designs toe. De random effecten worden verondersteld te komen van een geconjugeerde Bayesiaanse familie van verdelingen, waardoor ook andere verdelingen dan de normale mogelijk zijn. De variantie componenten kunnen gemodelleerd worden als een functie van covariaten en ook over-dispersie (gebaseerd over quasi aannemelijkheid) is mogelijk. Tenslotte wordt het 'REML' concept uitgebreid naar modellen buiten het lineaire gemengde effecten model.

In **hoofdstuk 5** breiden we het bestaande H-aannemelijkheids algoritme uit opdat het toegepast kan worden op zero-inflated count data. We passen het algoritme aan zodat me een hurdle model kunnen schatten. In zo'n model wordt de uitkomst nul in de modellering gescheiden van de positieve uitkomsten. De nullen worden gemodelleerd met een binaire verdeling terwijl voor de positieve uitkomsten een afgeknotte Poisson verdeling wordt gebruikt. Hier veronderstellen we dat beide delen apart gemodelleerd kunnen worden. Daarom zijn de random effecten onafhankelijk. We passen het model toe op de Tai-Chi Chuan trial waarbij gekeken wordt of met deze therapie vallen bij oudere mensen kunnen worden voorkomen.

In **hoofdstuk 6** passen we de H-aannemlijkheidsmethode toe op modellen met meerdere longitudinale uitkomsten uit een exponentiele familie. Deze worden met elkaar verbonden door gecorreleerde random effecten. We combineren het gebruik van het Newton-Raphson algoritme met het gebruik van H-aannemelijkheidsprocedures. Dit is de basis voor een verdere uitbreiding voor modellen waarin ook de correlatiestructuur als een functie van covariaten wordt gemodelleerd. Ook kunnen we het besprokene in hoofdstuk 5 en 6 combineren, zodat we hurdle modellen met gecorreleerde random effecten verkrijgen.

In **hoofdstuk 7** geven we een discussie, conclusie en ideeën voor verder onderzoek.

Acknowledgements

This thesis is a work of many who believed in my accomplishment and I am a believer. Without the people I could never get to the end of this project. First and foremost I own my gratitude to Emmanuel Lesaffre. We met first time in Diepenbeek, where he was teaching a Bayesian statistics course. I have responded to an opening for a phd candidate and he offered me a possibility to join the world of statistics. On the first paper we have worked under his guidance in Leuven, and after one year he allowed me to go with him to Rotterdam. Here, we started to read the book on h-likelihood methods together with Sten Willemsen. Next, we went to gym together, where we could talk about consulting projects while lifting heavy weights. It was the time where I lost quite a lot of weight myself. I own Emmanuel a deep gratitude for allowing me to go to Poland when I was sick, and also treating me gently during all following difficulties I had to go through. Working at home sometimes helped indeed. In that time we wrote next two papers and R package on h-likelihood. Mixture modelling paper was unfortunately rejected. Professor I would like to thank you for all your advise, humor and time spent to guide me to the end of this project.

I would like to thank my parents Ryszard and Jolanta for their constant encouragement to pursue the phd degree, and making possible for me to study master programs in Diepenbeek. Without this stage, the next could not have followed. In Diepenbeek I own my thanks to professor Geert Molenberghs, professor Marc Aerts, professor Paul Janssen, professor Tomasz Burzykowski and professor Herbert Thijs either for easing the formal procedures or direct help. Students as much as faculty staff at Limburgs Universitair Centrum and later Hasselt University greatly influenced and changed me over the years. Some friendships are still ongoing. Thanks goes also to residents of Pepperstraat building in Diepenbeek, especially Robert and Denise van Asbroeck.

In Leuven I have learned much from Steffen Fieuws, with whom we have shared an office for one year. But also from other members Kris Bogaerts, Ann Belmans,

Alejandro Jara Vallejos, Sylvia Cecere, Anhar Ullah, Michelle Ampe, Pushpike Thilakarathne, Dimitris Rizopoulos, Roula Tsonaka, professor Geert Verbeke, Luwis Diya, Arnost Komarek, David Dejardin, Joris Menten, Timothy Mutsvari, Robin van Oirbeek, Kirsten Verhaegen and Samuel Mwalili.

In Rotterdam, in the beginning I think I have started as a phd student at department of Epidemiology and Biostatistics, which later on was split into two separate departments. Throughout the years I asked for advise Wim Hop on several consulting projects, also to Wim I own interesting examples of incorrect statistics. I am also thankful for using made by me graphs (I think a bivariate likelihood function) in his teaching notes. Sten Willemsen, thanks for patiently sharing an office with me from the beginning. Thanks for guidance especially in external projects and numerous discussion which we had in the office together with Dimitris Rizopoulos and Magda Murawska. Sten many thanks for showing me the debug function in R. Maria de Ridder, Bettina Hansen and Lidia Arends thanks for support at the practicals assistance and company during lunch or (actually) epidemiology excursion to the town, invitation to your house or watching red bull air race together at Erasmus. Taye Hussien Hamza thanks for letting me experience how it is to be a one of the paranimfen and your friendship upon my arrival. Dimitris, thanks for all numerous discussions we had and R programming answers you gave me over our common years here in Rotterdam and before in Leuven. Polish members at department Magda Murawska and Karolina Sikorska made possible to talk mother language in which it is a bit easier to express things. Karolina thank you for everything, convincing me at least couple of times that I should finish the program, asking me numerous questions which made my existence worthwhile, but especially for being a great friend, jokes and time spent together in Rotterdam. With Magda we met already in Diepenbeek, I remember I visited you once in Genk these days, thanks for going to Hey concert together and visiting me couple of times in Rotterdam. I gained much from our statistical discussions. Johan de Rooi thanks for company, Ghent and Montpellier conferences and good jokes. With Baoyue Li we had many great conceptual discussions, your point of view and insight was always astounding, thanks friend, it is a pity Holland did not get the world cup in 2010, but it was close. Siti Haslinda Binti Mohd Din have done a great things for me, and her Malaysian expertise was always needed. Humor of Dymph Wijnen and her help in formal matters as well as arranging quickly ticket to Korea I try never to forget. Veronika Rockova thanks for doing a great job, especially during the repeated measures course and playing piano. Elrozy Andrinopoulou and Susan Bryan were great colleagues as well during that period. Kazem Nasserinejad and Nahid Mostafavi thanks for your support, time spent together in Rotterdam and numerous jokes. Professor Paul Eilers I shall try not to forget the Dutch croquette, thanks for your expertise and guidance. Ralph Rippe provided a nice theme for the beamer presentations, which I have been using on few conferences. Sabine Schnabel, Gianlucca Frasso, Sara Viviani, Nursel Koyuncu, Nathan Touati, Els Kinable, Puck van Osch, Nicole Erler, professor Joost van Rosmalen and professor Kees van Montfort were also in the department these days. Eline van Gent and Jacqueline Jacobs helped with their assistance in formal procedures. Nano Suwarno was always kind with IT matters. There are numerous people from Erasmus MC and many departments who contributed to this thesis as well. Participants of dutch course, members of department of Epidemiology, Medical Informatics, department of Public Health, Intensive Care or Obstetrics and Gyneacology among others.

Also I am indebted to Sweden and Korean group of h-likelihood. Professor Lars Ronnegard and his group thank you for organizing a meeting in Uppsala and Moudud Alam for being a great companion during my first Korea trip. I would like to thank professor Youngjo Lee for invitation to Seoul National University and his students Seungyoung Oh, Donghwan Lee, Karam Soh for helping me around in Korea. For trip to Pusan I thank professor Maengseok Noh and invitation to Daegu professor Il Do Ha. Professor Il Do Ha thank you for helping me with computer issues I had in Korea due to my forgetfullness. I would like also to mention professor Gilbert MacKenzie for consulting h-likelihood project at the early stage and offering me a copy of Genstat.

I would like to thank Petra EM Zeeuwe and Inge Logghe for using Tai-Chi data, Margot Walter and Jolanda Luime for Rheumatoid Arthritis data, Lisebet de Wit and Willy De Weerdt for CERISE data, and nice cooperation with Durk Berks, which resulted in a paper. I am also grateful to Magdalena Murawska and Peter Konings for agreeing to be my paranimfs and traveling quite a distance. Sten thank you for translating the summary of the thesis from english to dutch.

Time in Rotterdam was amusing due to my friends Martin Putto, Ryszard Jagodzinski, Rainey Sastrowidjojo and Maciej Szydlowski. We had good time and I learned a lot from all of you. You made the accomplishment of this project possible. I would like also to mention my family members in Poland, Agnieszka and Adam Molas, Piotr, Ilona and Malwina Rachon, Piotr and Slawomir Tofil, Daniel Tofil as well as their partners, children, parents and grandparents. Special thanks for Michal Mazurkiewicz for designing the cover of the thesis, blending and mixing images and aligning fonts. Finally I would like to thank my wife Eliza Savitri for staying late at night, so that we could chat over yahoo messenger, her teachings, patience and encouragement to finish this project. Eliza you have supported me many times not to drop the phd programme. Thank you for traveling to see me from the other side of the world, visiting Poland and meeting me in Korea. Icha terimakasih for loving me as I am, with all my problems and distractions. Saya Cinta Kamu.

Certainly many are not mentioned here, to them I owe as much as to those whose name is listed above.

Phd Portfolio

Name:	Marek Molas
Erasmus MC department:	Department of Biostatistics
Research School:	NIHES
Supervisor:	Prof. dr. Emmanuel Lesaffre
Phd period:	2007 – 2012

In-depth courses

2008	Models for longitudinal and incomplete data
2009	Multistate-models and models for competing risks
2009	Frailty models: multivariate survival analysis with applications in medicine
2009	Latex course with prof. Paul Eilers
2010	Bayesian methods and bias analysis
2010	Analysis of growth data
2010	The craft of smoothing
2010	Mixture models, cluster and discriminant analysis with R package mixAK
2011	Bayesian variable selection and model choice for structure additive regression
2011	Missing data in longitudinal studies: strategies for Bayesian modelling, sensitivity analysis and casual inference
2011	Joint modelling techniques

Long term research visits

2011	Department of Statistics, Seoul National University
2012	Department of Statistics, Seoul National University

Conferences

2007	3rd Workshop on Correlated Data Modelling - Limerick, Ireland (participation)
2007	28th International Society for Clinical Biostatistics - Alexandropoulis Greece (oral presentation)
2009	2nd International Biometric Society Channel Network - Ghent, Belgium (oral presentation)
2009	30th International Society for Clinical Biostatistics - Prague, Czech Republic (oral presentation)
2010	31st International Society for Clinical Biostatistics - Montpellier, France (oral presentation)
2012	33rd International Society for Clinical Biostatistics - Bergen, Norway (poster presentation)

Seminars

2010	Phd day presentation, Department of Biostatistics
2010	Mixed models working group presentation, Department of Biostatistics
2010	Working group HGLM meeting presentation, Uppsala, Sweden
2010	Participation in series of meetings genetics for dummies, Department of Biostatistics
2011	Phd day presentation, Department of Biostatistics
2012	Participation in spring symposium in biostatistics 2012, Rotterdam
2007– –2012	Participation in regular seminars and working group meetings organized by Center for Quantitative Methods, Erasmus MC

Teaching activities

2007– –2011	Assisting in practical sessions of NIHES Modern Methods course
2008– –2011	Assisting in practical sessions of NIHES Classical Methods course
2009	Assisting in practical sessions of NIHES Bayesian Methods course
2007– –2012	Assisting in practical sessions of NIHES Repeated Measures course
2010– –2011	Assisting in practical sessions of SPSS for undergraduate medical students

Curriculum Vitae

Marek Molas was born on 10 September 1979 in Lublin, Poland. After finishing high school in 1998 in Zamosc, he studied Econometrics at the Faculty of Economics of the Warsaw University. He completed the study with a Master Degree in 2003. Afterwards he moved to University of Hasselt, where he accomplished the masters programs in Applied Statistics and Biostatistics. In 2006 he started his doctorate under the supervision of prof. Lesaffre at the Biostatistical Centre at Katholieke Universiteit Leuven. The year after, he moved to Erasmus MC where he continued his Phd project with prof. Lesaffre at the Department of Biostatistics.