Generalised Linear Models with Clustered Data: Fixed and random effects models with glmmML

Göran Broström and Henrik Holmberg
Department of Statistics
Umeå University
SE–901 87 Umeå, Sweden

glmmML Version 0.81-8, March 21, 2011

Abstract

In situations where a large data set is partitioned into many relatively small clusters, and where members within clusters have something in common, the number of parameters tend to increase with sample size, if a fixed-effects model is applied. This fact causes the standard assumptions underlying asymptotic results to be violated. The standard solution to this problem is to apply a random intercepts model, where each cluster has its own intercept. The cluster intercepts are usually taken to be a random sample from a Normal distribution with mean zero and unknown variance. In the statistical computing environment **R**, there are a few packages, most notably `lme4`, that estimates models of this kind. For binary and Poisson data, `lme4` is a de facto standard for analyzing generalized linear mixed models (*GLMM*). It also generalises from the random intercepts model to include random slopes as well as nested clustering. The package `glmmML` generalises in other directions. First, it is only implemented for the simple random intercepts model and the Binomial and Poisson distributions, but it allows for other distributions than the Normal for the random intercepts. Test of the null hypothesis of no clustering is performed by a modified likelihood ratio test and, on request, by bootstrapping. Second, it allows for estimating a fixed effects model, assuming that all cluster intercepts are distinct fixed parameters, and, as a replacement for asymptotics, a bootstrapping tecnique is implemented. The random intercepts model is fitted through maximum likelihood with adaptive Gauss-Hermite and Laplace quadrature approximations of the likelihood function. The fixed effects model is fitted through a profiling approach, which is necessary when the number of clusters is large, because the standard function `glm` in **R** will choke on a huge design matrix. In a simulation study the two approaches are compared regarding two aspects. The first aspect is test of grouping effect, and the second is performance of regression parameter estimates. The main result is that the fixed effects model has severe bias when the mixed

effects variance is positive and the number of clusters is large. It is also shown that the Laplace approximation works fairly well when cluster sizes are not too small.

1 Introduction

Data sets with many small groups where there is within-group correlation and between-group variation are commonly encountered in many fields of application, for example in medical studies with repeated observations of patients, and in demographic investigations, where it is realistic to assume that members of a family have common characteristics. In this paper we investigate the properties of mixed effects models for these situations. We concentrate on binary and count data responses, i.e., the binomial and Poisson distributions in the framework of generalised linear models.

The generalised linear model with random intercept is a well-known model with some implementations in standard software. It is available in **SAS**, **Stata**, and in **R** R Core Team (2019). There are also (at least) five **R** packages available, the `lme4` package Bates et al. (2015) includes the `lmer` function, the `MASS` package Venables and Ripley (2002) includes Ripley's `glmmPQL` function, and, adding to that, the presently discussed `glmmML` package (Broström, 2019).

So what is the motivation for the package `glmmML`? Because it is different from the others (except `lme4`) in that it, as the name implies, fits the model via a direct maximum likelihood approach. The (marginal) likelihood function is a multi-dimensional integral in the general case. Furthermore, fixed effects models can be estimated efficiently through a profiling approach. This means that even with a huge number of clusters, the estimation procedure is fast and exact.

The random effects version of `glmmML` only fits models with random intercepts. This means that the multiple integral can be expanded into several one-dimensional integrals, and the numerical integration of the log likelihood function by the Laplace or Gauss-Hermite approximations is fast and accurate. Maximisation of the log likelihood function is done using the *optim* function `vmmin` in C code version. For this purpose we use the LINPACK Fortran subroutines `dpoco`, `dposl`, and `dpodi`, combined with `blas` routines. All are found within **R**.

In the fixed effects model, testing is performed via a simple bootstrap. Under the null hypothesis of no cluster effect, the grouping factor can be randomly permuted without changing the probability distribution. This is the basic idea in estimating the p-value by simulation.

In Section 2 we define the likelihood function for the distributions we consider, i.e., the binomial and the Poisson. We then consider fixed group effects in Section 3, introducing the profile approach. In Section 4 we intro-

duce the random effects model with a symmetric mixture distribution. We show how to construct the log-likelihood function and we derive the first and second partial derivatives.

For comparisons of choice of random or fixed effects of clustering, see the forthcoming paper by Broström and Holmberg (2011).

2 The likelihood function

Assume that there are n clusters in our data, of sizes $n_i, i = 1, \ldots, n$. For each cluster we observe responses $(y_{i1}, \ldots, y_{in_i})$ and vectors of explanatory variables $(\mathbf{x}_{i1}, \ldots, \mathbf{x}_{in_i})$, where $\mathbf{x}_{ij}$ is a p-dimensional vector with the first element identically equal to unity, corresponding to the mean value of the random intercept. The random part, u_i, of the intercept is assumed to follow a distribution with density

$$h(u;\sigma) = \frac{1}{\sigma} p\left(\frac{u}{\sigma}\right), \quad -\infty < u < \infty,\ \sigma > 0, \tag{1}$$

i.e., with location zero and scale σ. It is assumed that $u_1, \ldots, u_n$ are independent.

The conditional distribution of the response, given the random intercepts $\beta_1 + u_i, i = 1, \ldots, n$, is assumed to follow a multivariate distribution according to

$$\begin{aligned} \Pr(Y_{ij} = y_{ij} \mid u_i; \mathbf{x}) = P(\boldsymbol{\beta}\mathbf{x}_{ij} + u_i, y_{ij}), \\ y_{ij} = 0, 1, \ldots;\ j = 1, \ldots, n_i,\ i = 1, \ldots, n. \end{aligned} \tag{2}$$

For instance, with the Bernoulli distribution and the *logit link*, we get

$$P(x, y) = \frac{e^{xy}}{1 + e^x}, \quad y = 0, 1;\ -\infty < x < \infty,$$

and with the *cloglog* link we have

$$P(x, y) = \big(1 - \exp(-e^x)\big)^y \exp\big(-(1 - y)e^x\big), \quad y = 0, 1;\ -\infty < x < \infty,$$

The Poisson distribution with *log link* gives rise to

$$P(x, y) = \frac{e^{xy}}{y!} e^{-e^x}, \quad y = 0, 1, 2, \ldots;\ -\infty < x < \infty \tag{3}$$

These are in fact all the possibilities that are available in `glmmML` and will be considered in this paper.

In the fixed effects model (and in the random effects model, if we condition on these effects), the likelihood function becomes

$$L\big((\boldsymbol{\beta}, \boldsymbol{\gamma}); \mathbf{y}, \mathbf{x}\big) = \prod_{i=1}^{n} \prod_{j=1}^{n_i} P(\boldsymbol{\beta}\mathbf{x}_{ij} + \gamma_i, y_{ij}). \tag{4}$$

The log likelihood function becomes

$$\ell\big((\boldsymbol{\beta},\boldsymbol{\gamma});\mathbf{y},\mathbf{x}\big)=\sum_{i=1}^{n}\sum_{j=1}^{n_i}\log P(\boldsymbol{\beta}\mathbf{x}_{ij}+\gamma_i,y_{ij}), \tag{5}$$

3 Fixed group effects

3.1 Without profiling

The partial derivatives with respect to $\beta_m,\ m=1;\ldots,p$, of the log likelihood function (5) are:

$$\begin{aligned}U_m(\boldsymbol{\beta},\boldsymbol{\gamma})&=\frac{\partial}{\partial\beta_m}\ell\big((\boldsymbol{\beta},\boldsymbol{\gamma});\mathbf{y},\mathbf{x}\big)\\&=\sum_{i=1}^{n}\sum_{j=1}^{n_i}x_{ijm}G\big(\boldsymbol{\beta}\mathbf{x}_{ij}+\gamma_i,\ y_{ij}\big),\quad m=1,\ldots,p.\end{aligned} \tag{6}$$

where

$$G(x,y)=\frac{\partial}{\partial x}\log P(x,y)=\frac{\frac{\partial}{\partial x}P(x,y)}{P(x,y)}$$

The partial derivatives with respect to $\gamma_i,\ i=1,\ldots,n$ of (5) are:

$$\begin{aligned}U_{p+i}(\boldsymbol{\beta},\boldsymbol{\gamma})&=\frac{\partial}{\partial\gamma_i}\ell\big((\boldsymbol{\beta},\boldsymbol{\gamma});\mathbf{y},\mathbf{x}\big)\\&=\sum_{j=1}^{n_i}G\big(\boldsymbol{\beta}\mathbf{x}_{ij}+\gamma_i,\ y_{ij}\big),\quad i=1,\ldots,n.\end{aligned} \tag{7}$$

The Hessian, or minus the observed information, $-I(\boldsymbol{\beta},\boldsymbol{\gamma})$ has entries

$$\begin{aligned}-I_{ms}(\boldsymbol{\beta},\boldsymbol{\gamma})&=\frac{\partial}{\partial\beta_s}U_m(\boldsymbol{\beta},\boldsymbol{\gamma})\\&=\sum_{i=1}^{n}\sum_{j=1}^{n_i}x_{ijm}x_{ijs}H\big(\boldsymbol{\beta}\mathbf{x}_{ij}+\gamma_i,\ y_{ij}\big),\quad m,s=1,\ldots,p,\end{aligned} \tag{8}$$

$$\begin{aligned}-I_{(p+i)s}(\boldsymbol{\beta},\boldsymbol{\gamma})&=\frac{\partial}{\partial\beta_s}U_{p+i}(\boldsymbol{\beta},\boldsymbol{\gamma})\\&=\sum_{j=1}^{n_i}x_{ijs}H\big(\boldsymbol{\beta}\mathbf{x}_{ij}+\gamma_i,\ y_{ij}\big),\quad s=1,\ldots,p;\ i=1,\ldots,n.\end{aligned} \tag{9}$$

$$\begin{aligned}-I_{(p+i)(p+i)}(\boldsymbol{\beta},\boldsymbol{\gamma})&=\frac{\partial}{\partial\gamma_i}U_{p+i}(\boldsymbol{\beta},\boldsymbol{\gamma})\\&=\sum_{j=1}^{n_i}H\big(\boldsymbol{\beta}\mathbf{x}_{ij}+\gamma_i,\ y_{ij}\big),\quad i=1,\ldots,n.\end{aligned} \tag{10}$$

$$-I_{(p+i)(p+k)}(\boldsymbol{\beta},\boldsymbol{\gamma}) = 0, \quad k \neq i;\ k,i = 1,\ldots,n.$$

where

$$H(x,y) = \frac{\partial}{\partial x}G(x,y)$$

When n is large, we may utilise the fact that I is partially diagonal.

3.2 With profiling

Setting (7) equal to zero defines $\boldsymbol{\gamma}$ implicitly as functions of $\boldsymbol{\beta}$, $\gamma_i = \gamma_i(\boldsymbol{\beta}),\ i = 1,\ldots,n$:

$$F\big(\boldsymbol{\beta},\gamma_i(\boldsymbol{\beta})\big) = \sum_{j=1}^{n_i} G\big(\boldsymbol{\beta}\mathbf{x}_{ij} + \gamma_i(\boldsymbol{\beta}),\ y_{ij}\big) = 0, \quad i = 1,\ldots,n. \tag{11}$$

Generally, we get no explicit form for $\gamma_i(\boldsymbol{\beta})$, but we can calculate its partial derivatives via implicit derivation. From

$$\frac{\partial}{\partial\beta_m}F\big(\boldsymbol{\beta},\gamma_i(\boldsymbol{\beta})\big) = \frac{\partial\gamma_i}{\partial\beta_m}\frac{\partial F}{\partial\gamma_i} + \frac{\partial F}{\partial\beta_m} = 0$$

we get

$$\begin{aligned}\frac{\partial\gamma_i(\boldsymbol{\beta})}{\partial\beta_m} &= -\frac{\frac{\partial F}{\partial\beta_m}}{\frac{\partial F}{\partial\gamma_i}} \\ &= -\frac{\sum_{j=i}^{n_i} x_{ijm}H\big(\boldsymbol{\beta}\mathbf{x}_{ij} + \gamma_i,\ y_{ij}\big)}{\sum_{j=1}^{n_i} H\big(\boldsymbol{\beta}\mathbf{x}_{ij} + \gamma_i,\ y_{ij}\big)}, \quad i = 1,\ldots,n;\ m = 1,\ldots,p.\end{aligned} \tag{12}$$

Replacing $\boldsymbol{\gamma}$ by $\boldsymbol{\gamma}(\boldsymbol{\beta})$ in (5) gives the profile log likelihood $\ell^{(p)}$:

$$\ell^{(p)}\big(\boldsymbol{\beta};\mathbf{y},\mathbf{x}\big) = \sum_{i=1}^{n}\sum_{j=1}^{n_i}\log P\big(\boldsymbol{\beta}\mathbf{x}_{ij} + \gamma_i(\boldsymbol{\beta}),\ y_{ij}\big), \tag{13}$$

3.2.1 Profile partial derivatives

The partial derivatives with respect to $\beta_m,\ m = 1;\ldots,p$, of the log profile likelihood function (13) becomes:

$$\begin{aligned}U_m^{(p)}(\boldsymbol{\beta}) &= \frac{\partial}{\partial\beta_m}\ell^{(p)}(\boldsymbol{\beta};\mathbf{y},\mathbf{x}) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n_i}\left(x_{ijm} + \frac{\partial\gamma_i(\boldsymbol{\beta})}{\partial\beta_m}\right)G\big(\boldsymbol{\beta}\mathbf{x}_{ij} + \gamma_i(\boldsymbol{\beta}),\ y_{ij}\big) \\ &= U_m\big(\boldsymbol{\beta},\boldsymbol{\gamma}(\boldsymbol{\beta})\big) + \sum_{i=1}^{n}\frac{\partial\gamma_i(\boldsymbol{\beta})}{\partial\beta_m}\sum_{j=1}^{n_i}G\big(\boldsymbol{\beta}\mathbf{x}_{ij} + \gamma_i(\boldsymbol{\beta}),\ y_{ij}\big) \\ &= U_m(\boldsymbol{\beta},\boldsymbol{\gamma}(\boldsymbol{\beta})),\end{aligned} \tag{14}$$

where the last equality follows from (11). Thus we get back the unprofiled partial derivatives (6).

3.2.2 The profile Hessian

From (14) we get the Hessian, or minus the information matrix

$$\begin{aligned} -I^{(p)}_{ms}(\boldsymbol{\beta}) &= \frac{\partial}{\partial \beta_s} U_m\big(\boldsymbol{\beta}, \boldsymbol{\gamma}(\boldsymbol{\beta})\big) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n_i} x_{ijm}\left(x_{ijs} + \frac{\partial \gamma_i(\boldsymbol{\beta})}{\partial \beta_s}\right) H\big(\boldsymbol{\beta}\mathbf{x}_{ij} + \gamma_i(\boldsymbol{\beta}),\ y_{ij}\big) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n_i} x_{ijm}x_{ijs}H_{ij} \\ &\quad - \sum_{i=1}^{n} \frac{\sum_{j=1}^{n_i} x_{ijm}H_{ij} \sum_{j=1}^{n_i} x_{ijs}H_{ij}}{\sum_{j=1}^{n_i} H_{ij}}, \\ &\quad m, s = 1, \ldots, p. \end{aligned} \tag{15}$$

where

$$H_{ij} = H\big(\boldsymbol{\beta}\mathbf{x}_{ij} + \gamma_i(\boldsymbol{\beta}),\ y_{ij}\big), \quad j = 1, \ldots n_i;\ i = 1, \ldots, n.$$

3.2.3 At the maximum

The following theorem by Patefield (1977) justifies the use of the profile likelihood for statistical inference.

Theorem 1 (Patefield) *The inverse Hessians from the full likelihood and from the profile likelihood for $\boldsymbol{\beta}$ are identical when*

$$(\boldsymbol{\gamma}, \boldsymbol{\beta}) = (\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\beta}}).$$

3.3 Optimisation considerations

There are a few practical things to note in the optimisation by profiling. First, a new iteration step starts by solving the n equations given by setting (7) equal to zero. The left-hand sides are simple, monotone functions of one variable and easy and fast to solve numerically, but see below. This gives $(\gamma_1, \ldots, \gamma_n)$, which are plugged into (12), (14), and (15).

Second, it is easy to see that clusters where all the responses are zero can be removed from the calculations, after noting that the corresponding $\gamma = -\infty$. Correspondingly, in the binomial case, clusters with all responses equal to one can be removed and the corresponding $\gamma = +\infty$. These extreme cases correspond to probabilities equal to zero and one, respectively.

In order to illustrate the performance boost of the profile approach over using the standard `glm` function in **R**, consider the following numerical example with 1000 clusters and five individuals in each, one covariate:

```
> dat <- data.frame(y = rbinom(5000, size = 1, prob = 0.5),
+     x = rnorm(5000), group = rep(1:1000, each = 5))
> system.time(fit1 <- glm(y ~ factor(group) + x, data = dat,
+     family = binomial))

   user  system elapsed
 86.430   0.840  87.546

> library(glmmML)
> system.time(fit2 <- glmmboot(y ~ x, cluster = group,
+     data = dat))

   user  system elapsed
  0.080   0.010   0.097
```

The huge difference in computing time is not due to lack of enough computer memory; the test was performed on a machine with 64 GB RAM. The resulting parameter estimates, standard errors and p-values are indistinguishable.

3.3.1 Profiling with the Poisson distribution

When data are Poisson, the profiling can be made explicit. The log likelihood function is, from (3) and (5),

$$\ell \propto \sum_{i=1}^{n} \sum_{j=1}^{n_i} \big\{ y_{ij}(\boldsymbol{\beta}\mathbf{x}_{ij} + \gamma_i) - \exp(\boldsymbol{\beta}\mathbf{x}_{ij} + \gamma_i) \big\}, \tag{16}$$

and equations (11) become

$$\sum_{j=1}^{n_i} \big\{ y_{ij} - \exp(\boldsymbol{\beta}\mathbf{x}_{ij} + \gamma_i) \big\} = 0, \quad i = 1, \ldots, n.$$

with solutions

$$\gamma_i = \log\left(\sum_{j=1}^{n_i} y_{ij} \right) - \log\left(\sum_{j=1}^{n_i} \exp(\boldsymbol{\beta}\mathbf{x}_{ij}) \right), \quad i = 1, \ldots, n. \tag{17}$$

Inserting (17) into (16) and simplifying results in the profile likelihood

$$\ell^{(p)} \propto \sum_{i=1}^{n} \sum_{j=1}^{n_i} y_{ij} \left\{ \boldsymbol{\beta}\mathbf{x}_{ij} - \log\left(\sum_{j=1}^{n_i} \exp(\boldsymbol{\beta}\mathbf{x}_{ij}) \right) \right\}$$

This is also recognized as a *partial likelihood* (Cox, 1975). In fact, when the responses y_{ij} are indicators, zero or one, and the clusters are interpreted as *risk sets* at times when events occur, the profile likelihood is identical to the partial likelihood in Cox regression (with Breslow's approximation for ties). This was observed by Johansen (1983).

4 Random group effects

We assume conditional independence, and, conditionally on $\mathbf{u}(=\boldsymbol{\gamma})$, an ordinary logistic regression model with *offsets* $\mathbf{u} = (u_1, \ldots, u_n)$. Since $\mathbf{u}$ is unobserved, the unconditional likelihood function is of greater interest, and we get it by "integrating out" $\mathbf{u}$:

$$L\big((\boldsymbol{\beta}, \sigma); \mathbf{y}, \mathbf{x}\big) = \int \cdots \int_{R^n} \prod_{i=1}^{n} \prod_{j=1}^{n_i} P(\boldsymbol{\beta}\mathbf{x}_{ij} + u_i, y_{ij}) \frac{1}{\sigma} p\left(\frac{u_i}{\sigma}\right) du_1 \cdots du_n.$$

Due to independence, this n-dimensional integral can be written as a product of n simple integrals:

$$L\big((\boldsymbol{\beta}, \sigma); \mathbf{y}, \mathbf{x}\big) = \prod_{i=1}^{n} \int_{-\infty}^{\infty} p(u) \prod_{j=1}^{n_i} P(\boldsymbol{\beta}\mathbf{x}_{ij} + \sigma u,\ y_{ij}) du, \tag{18}$$

where a variable substitution ($u \to \sigma u$) has taken place.

The log likelihood function thus becomes

$$\ell\big((\boldsymbol{\beta}, \sigma); \mathbf{y}, \mathbf{x}\big) = \sum_{i=1}^{n} \log \int_{-\infty}^{\infty} p(u) \prod_{j=1}^{n_i} P(\boldsymbol{\beta}\mathbf{x}_{ij} + \sigma u,\ y_{ij}) du, \tag{19}$$

and the remaining part of this section is devoted to the problem of maximising (19) with respect to $(\boldsymbol{\beta}, \sigma)$. For this we will need the score vector, and for estimation of the variance-covariance matrix of the parameter estimates we will use the Hessian of (19), or rather an approximation thereof.

For the numerical evaluation of integrals we use quadrature methods, briefly described in Subsections 4.3 and 4.4. For more detail, consult, e.g., Gray (2001). There are two ways to proceed: (i) Calculate the analytic partial first and second order derivatives of the log likelihood function (19) and make numerical approximations of them, and (ii) from the approximation of (19), calculate the analytic partial first and second order derivatives. We follow the latter route.

In the next subsection we introduce the Laplace transform of this problem, and after that the Gauss-Hermite approximation is introduced. The latter is a generalisation of the former. In both cases, a fully adaptive version is implemented.

4.1 The Laplace approximation

The integrals will be evaluated by Laplace approximation. We follow the approach of approximating only the log-likelihood function and from there calculate all the necessary derivatives.

4.1.1 The log-likelihood function

We look at one group, i.e., a fixed i for the moment. Let

$$p(u)\prod_{j=1}^{n_i} P(\boldsymbol{\beta}\mathbf{x}_{ij}+\sigma u,\ y_{ij}) = \exp\{g(u,\boldsymbol{\theta})\},$$

with $\boldsymbol{\theta}=(\boldsymbol{\beta},\sigma)$. Then the integral in (19) can be written, using the Laplace approximation, as

$$\int_{-\infty}^{\infty}\exp\{g(u,\boldsymbol{\theta})\}du \approx \sqrt{2\pi}\hat{\omega}\exp\{g(\hat{u},\boldsymbol{\theta})\},$$

where $\hat{\omega}$ and $\hat{u}$ are defined below. For that we need some partial derivatives of g.

$$\begin{aligned}
g(u,\boldsymbol{\theta}) &= \log\{p(u)\}+\sum_{j=1}^{n_i}\log P(\boldsymbol{\beta}x_{ij}+\sigma u, y_{ij}),\\
g_u(u,\boldsymbol{\theta}) &= \frac{\partial}{\partial u}\log\{p(u)\}+\sigma\sum_{j=1}^{n_i}G(\boldsymbol{\beta}x_{ij}+\sigma u, y_{ij}),\\
g_{uu}(u,\boldsymbol{\theta}) &= \frac{\partial^2}{\partial u^2}\log\{p(u)\}+\sigma^2\sum_{j=1}^{n_i}H(\boldsymbol{\beta}x_{ij}+\sigma u, y_{ij}),
\end{aligned} \tag{20}$$

where

$$\begin{aligned}
G(x,y) &= \frac{\partial}{\partial x}\log P(x,y)\\
H(x,y) &= \frac{\partial}{\partial x}G(x,y)
\end{aligned} \tag{21}$$

Now, $\hat{u}$ is defined as

$$\hat{u}=\hat{u}(\boldsymbol{\theta}) = \text{argmax}_u g(u,\boldsymbol{\theta}), \tag{22}$$

(where we emphasise that $\hat{u}$ is a function of $\boldsymbol{\theta}$), i.e., we also have

$$g_u(\hat{u},\boldsymbol{\theta}) = 0. \tag{23}$$

Then $\hat{\omega}$ is defined as

$$\begin{aligned}\hat{\omega} = \hat{\omega}(\boldsymbol{\theta}) &= \sqrt{-\frac{1}{g_{uu}(\hat{u},\boldsymbol{\theta})}} \\ &= \left(\frac{d^2}{du^2}\log\{p(u)\}\mid_{u=\hat{u}} -\sigma^2\sum_{j=1}^{n_i} H(\boldsymbol{\beta}x_{ij}+\sigma\hat{u},y_{ij})\right)^{-\frac{1}{2}},\end{aligned} \tag{24}$$

which also gives the relation

$$g_{uu}(\hat{u},\boldsymbol{\theta}) = -\frac{1}{\hat{\omega}^2(\boldsymbol{\theta})} \tag{25}$$

Thus, the contribution to the log likelihood from the ith group is approximated by

$$\begin{aligned}\ell_i(\boldsymbol{\theta}) &\approx 0.5\log(2\pi) + \log\{\hat{\omega}(\boldsymbol{\theta})\} + g(\hat{u}(\boldsymbol{\theta}),\boldsymbol{\theta}) \\ &= 0.5\log(2\pi) - 0.5\log\{g_{uu}(\hat{u}(\boldsymbol{\theta}),\boldsymbol{\theta})\} + g(\hat{u}(\boldsymbol{\theta}),\boldsymbol{\theta})\end{aligned} \tag{26}$$

4.1.2 The score vector

In the maximisation of

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n}\ell_i(\boldsymbol{\theta}),$$

we will make use of the score vector. For that purpose we will need the partial derivatives of $\hat{u}(\boldsymbol{\theta})$ and $\hat{\omega}(\boldsymbol{\theta})$. From (23) we get, by implicit differentiation,

$$\begin{aligned}\frac{\partial}{\partial\boldsymbol{\theta}}g_u(\hat{u}(\boldsymbol{\theta}),\boldsymbol{\theta}) &= g_{uu}(\hat{u},\boldsymbol{\theta})\frac{\partial\hat{u}}{\partial\boldsymbol{\theta}} + \frac{\partial g_u}{\partial\boldsymbol{\theta}} \\ &= \frac{\partial\hat{u}}{\partial\boldsymbol{\theta}}g_{uu}(\hat{u},\boldsymbol{\theta}) + g_{u\boldsymbol{\theta}}(\hat{u},\boldsymbol{\theta}) \\ &= \hat{u}_{\boldsymbol{\theta}}(\boldsymbol{\theta})g_{uu}(\hat{u},\boldsymbol{\theta}) + g_{u\boldsymbol{\theta}}(\hat{u},\boldsymbol{\theta}) = 0\end{aligned} \tag{27}$$

which gives

$$\hat{u}_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = \frac{\partial\hat{u}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} = -\frac{g_{u\boldsymbol{\theta}}(\hat{u},\boldsymbol{\theta})}{g_{uu}(\hat{u},\boldsymbol{\theta})} = \hat{\omega}^2(\boldsymbol{\theta})g_{u\boldsymbol{\theta}}(\hat{u},\boldsymbol{\theta}) = \hat{\omega}^2(\boldsymbol{\theta})g_{u\boldsymbol{\theta}}(\hat{u}(\boldsymbol{\theta}),\boldsymbol{\theta}) \tag{28}$$

In calculating the partial first order derivatives, we utilise the formula

$$\begin{aligned}\frac{\partial}{\partial\boldsymbol{\theta}}\ell(\hat{u}(\boldsymbol{\theta}),\boldsymbol{\theta}) &= \frac{\partial}{\partial\boldsymbol{\theta}}\log\hat{\omega}(\boldsymbol{\theta}) + \hat{u}_{\boldsymbol{\theta}}g_u + g_{\boldsymbol{\theta}} \\ &= \frac{\hat{\omega}_{\boldsymbol{\theta}}}{\hat{\omega}} + \hat{u}_{\boldsymbol{\theta}}g_u + g_{\boldsymbol{\theta}}.\end{aligned} \tag{29}$$

So, we will need

$$\begin{aligned}\frac{\partial}{\partial\boldsymbol{\theta}}\hat{\omega}(\hat{u}(\boldsymbol{\theta}),\boldsymbol{\theta}) = \hat{\omega}_{\boldsymbol{\theta}} &= \frac{1}{2}(-g_{uu})^{-\frac{3}{2}}\{\hat{u}_{\boldsymbol{\theta}}g_{uuu} + g_{uu\boldsymbol{\theta}}\} \\ &= \frac{1}{2}\hat{\omega}^3\{\hat{u}_{\boldsymbol{\theta}}g_{uuu} + g_{uu\boldsymbol{\theta}}\}\end{aligned} \tag{30}$$

4.1.3 The Hessian

The Hessian will be needed for variance estimation. For that purpose we will need the partial derivatives of second order of $\hat{u}(\boldsymbol{\theta})$. Therefore, we continue by calculating the partial derivatives of (27) with respect to $\boldsymbol{\theta}'$.

$$\begin{aligned}\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}g_u(\hat{u}(\boldsymbol{\theta}),\boldsymbol{\theta}) &= \hat{u}_{\boldsymbol{\theta}\boldsymbol{\theta}'}g_{uu} + \hat{u}_{\boldsymbol{\theta}}(\hat{u}_{\boldsymbol{\theta}'}g_{uuu} + g_{uu\boldsymbol{\theta}'}) \\ &\quad + \hat{u}_{\boldsymbol{\theta}'}g_{uu\boldsymbol{\theta}} + g_{u\boldsymbol{\theta}\boldsymbol{\theta}'} \\ &= 0.\end{aligned}$$

Solving for $\hat{u}_{\boldsymbol{\theta}\boldsymbol{\theta}'}$ gives

$$\hat{u}_{\boldsymbol{\theta}\boldsymbol{\theta}'} = \frac{\partial}{\partial\boldsymbol{\theta}'}\hat{u}_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = \hat{\omega}^2(\hat{u}_{\boldsymbol{\theta}}\hat{u}_{\boldsymbol{\theta}'}g_{uuu} + \hat{u}_{\theta}g_{uu\boldsymbol{\theta}'} + \hat{u}_{\theta'}g_{uu\boldsymbol{\theta}} + g_{u\boldsymbol{\theta}\boldsymbol{\theta}'}).$$

We will also need

$$\begin{aligned}\frac{\partial^2}{\partial\boldsymbol{\theta}\boldsymbol{\theta}'}\hat{\omega}(\boldsymbol{\theta}) &= \frac{3}{4}\hat{\omega}^5(\hat{u}_{\boldsymbol{\theta}'}g_{uuu} + g_{uu\boldsymbol{\theta}'})(\hat{u}_{\boldsymbol{\theta}}g_{uuu} + g_{uu\boldsymbol{\theta}}) \\ &\quad + \frac{1}{2}\hat{\omega}^3(\hat{u}_{\boldsymbol{\theta}\boldsymbol{\theta}'}g_{uuu} + \hat{u}_{\boldsymbol{\theta}}\hat{u}_{\boldsymbol{\theta}'}g_{uuuu} + \hat{u}_{\boldsymbol{\theta}}g_{uuu\boldsymbol{\theta}'} + \hat{u}_{\boldsymbol{\theta}\boldsymbol{\theta}'}g_{uuu\boldsymbol{\theta}} + g_{uu\boldsymbol{\theta}\boldsymbol{\theta}'}).\end{aligned}$$

From (29), with the same formula, we get

$$\begin{aligned}\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\ell(\hat{u}(\boldsymbol{\theta}),\boldsymbol{\theta}) &= \frac{\partial}{\partial\boldsymbol{\theta}'}\left\{\frac{\hat{\omega}_{\boldsymbol{\theta}}}{\hat{\omega}} + \hat{u}_{\boldsymbol{\theta}}g_u + g_{\boldsymbol{\theta}}\right\} \\ &= \frac{\hat{\omega}_{\boldsymbol{\theta}\boldsymbol{\theta}'}\hat{\omega} - \hat{\omega}_{\boldsymbol{\theta}}\hat{\omega}_{\boldsymbol{\theta}'}}{\hat{\omega}^2} + \hat{u}_{\boldsymbol{\theta}\boldsymbol{\theta}'}g_u \\ &\quad + \hat{u}_{\boldsymbol{\theta}}(\hat{u}_{\boldsymbol{\theta}'}g_{uu} + g_{u\boldsymbol{\theta}'}) + \hat{u}_{\boldsymbol{\theta}'}g_{u\boldsymbol{\theta}} + g_{\boldsymbol{\theta}\boldsymbol{\theta}'}.\end{aligned}$$

All the necessary derivatives can be found in A.

4.2 The Gauss-Hermite approximation

The Gauss-Hermite approximation can be viewed upon as a generalisation of the Laplace approximation; instead of using just a single point, the approximation is built around approximations in several points. More specifically, the formula is

$$\int_{-\infty}^{\infty}\exp\{g(u,\boldsymbol{\theta})\}du \approx \sqrt{2\pi}\hat{\omega}\sum_{i=1}^{n}h_i\exp\{g(\hat{u}+\sqrt{2\pi}\hat{\omega}x_i,\boldsymbol{\theta}) + x_i^2\}, \qquad (31)$$

where $x_1,\ldots,x_n$ and $h_1,\ldots,h_n$ are the *abscissas* and *weights* of the transform, and n is the number of quadrature points. When $n = 1$, this coincides with the Laplace approximation.

The constants (but functions of θ) $\hat{u}$ and $\hat{\omega}$ are the same as in the Laplace transform of Section 4.3. The calculations of the approximations of the first and second order partial derivatives (or vice versa) are straightforward (cf. Section 4.3), and we omit them here. They are implemented in the **R** package `glmmML`.

5 Conclusion

What model should be used to a particular data set, i.e., when is a mixed effects model preferable over a fixed effects model? Generally speaking, a random effects model is appropriate if the observed clusters may be regarded as a random sample from a (large, possibly infinite) pool of possible clusters. The observed clusters are of no practical interest per se, but the distribution in the pool is. Or this distribution is regarded as a nuisance that needs to be controlled for. A fixed effects model, on the other hand, is appropriate if we consider the given clusters as the full universe of clusters.

In the random effects case, we expect the number of clusters to grow as sample size grows, and the cluster sizes to remain stable. In the fixed effects approach, on the other hand, it is expected that the number of clusters is stable, while cluster size grows with sample size.

References

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

Broström, G. (2019). *glmmML: Generalized linear models with clustering*. R package version 1.1.0.

Broström, G. and Holmberg, H. (2011). Generalised linear models with clustered data: Fixed and random effects models. *Computational Statistics & Data Analysis*, 55:3123–3134.

Cox, D. (1975). Partial likelihood. *Biometrika*, 62:269–276.

Gray, R. (2001). Advanced statistical computing. Course Notes; http://pages.stat.wisc.edu/ mchung/teaching/stat471/stat_computing.pdf.

Johansen, S. (1983). An extension of Cox's regression model. *International Statistical Review*, 51:165–174.

Patefield, W. (1977). On the maximized likelihood function. *Sankhya Series B*, 39:92–96.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer, New York, fourth edition. ISBN 0-387-95457-0.

A Analytic derivatives

A.1 The score

The partial derivatives with respect to $\beta_m, m = 1; \ldots, p$, of the log likelihood function (19) are:

$$\frac{\partial}{\partial \beta_m} \ell\big((\boldsymbol{\beta}, \omega); \mathbf{y}, \mathbf{x}\big) = \sum_{i=1}^{n} \frac{\frac{\partial}{\partial \beta_m} \int_{-\infty}^{\infty} \varphi(u) \prod_{j=1}^{n_i} P(\boldsymbol{\beta}\mathbf{x}_{ij} + \sigma u,\ y_{ij}) du}{\int_{-\infty}^{\infty} \varphi(u) \prod_{j=1}^{n_i} P(\boldsymbol{\beta}\mathbf{x}_{ij} + \sigma u,\ y_{ij}) du} \tag{32}$$

The partial derivatives in the numerators are given by

$$\begin{aligned} &\frac{\partial}{\partial \beta_m} \int_{-\infty}^{\infty} \varphi(u) \prod_{j=1}^{n_i} P(\boldsymbol{\beta}\mathbf{x}_{ij} + \sigma u,\ y_{ij}) du = \\ &\qquad \int_{-\infty}^{\infty} \varphi(u) \prod_{j=1}^{n_i} P(\boldsymbol{\beta}\mathbf{x}_{ij} + \sigma u,\ y_{ij}) \sum_{j=1}^{n_i} x_{ijm} G(\boldsymbol{\beta}\mathbf{x}_{ij} + \sigma u,\ y_{ij}) du, \end{aligned} \tag{33}$$

with

$$G(x, y) = \frac{\partial}{\partial x} \log P(x, y) = \frac{\frac{\partial}{\partial x} P(x, y)}{P(x, y)} \tag{34}$$

The partial derivative with respect to $\omega = \log(\sigma)$ is

$$\frac{\partial}{\partial \omega} \ell\big((\boldsymbol{\beta}, \omega); \mathbf{y}, \mathbf{x}\big) = \sum_{i=1}^{n} \frac{\frac{\partial}{\partial \omega} \int_{-\infty}^{\infty} \varphi(u) \prod_{j=1}^{n_i} P(\boldsymbol{\beta}\mathbf{x}_{ij} + \sigma u,\ y_{ij}) du}{\int_{-\infty}^{\infty} \varphi(u) \prod_{j=1}^{n_i} P(\boldsymbol{\beta}\mathbf{x}_{ij} + \sigma u,\ y_{ij}) du} \tag{35}$$

From this, we get the partial derivatives in the numerators as

$$\begin{aligned} &\frac{\partial}{\partial \omega} \int_{-\infty}^{\infty} \varphi(u) \prod_{j=1}^{n_i} P(\boldsymbol{\beta}\mathbf{x}_{ij} + \sigma u,\ y_{ij}) du = \\ &\qquad \sigma \int_{-\infty}^{\infty} u\varphi(u) \prod_{j=1}^{n_i} P(\boldsymbol{\beta}\mathbf{x}_{ij} + \sigma u,\ y_{ij}) \sum_{j=1}^{n_i} G(\boldsymbol{\beta}\mathbf{x}_{ij} + \sigma u,\ y_{ij}) du \end{aligned} \tag{36}$$

A.2 The Hessian

Some "symbolic" notation:

$$\ell = \sum_{i=1}^{n} \log h_i \tag{37}$$

$$\begin{aligned} \frac{\partial \ell}{\partial \beta_m} &= \sum_{i=1}^{n} \frac{h_\beta(m, i)}{h_i}, \quad m = 1, \ldots, p, \\ \frac{\partial \ell}{\partial \omega} &= \sum_{i=1}^{n} \frac{h_\omega(i)}{h_i} \end{aligned}$$

Here $h_\beta(m,i),\ i=1,\ldots,n;\ m=1,\ldots,p$ are given by equation (33), and $h_\omega(i),\ i=1,\ldots,n$ are given by equation (36).

The second derivatives are needed at the solution in order to estimate standard errors.

$$\begin{aligned}
\frac{\partial^2\ell}{\partial\beta_k\partial\beta_m} &= \sum_{i=1}^{n}\left\{\frac{h_{\beta\beta}(k,m,i)}{h_i} - \frac{h_\beta(k,i)}{h_i}\frac{h_\beta(m,i)}{h_i}\right\}, \quad k,m=1,\ldots,p \\
\frac{\partial^2\ell}{\partial\beta_k\partial\omega} &= \sum_{i=1}^{n}\left\{\frac{h_{\beta\omega}(k,i)}{h_i} - \frac{h_\beta(k,i)}{h_i}\frac{h_\omega(i)}{h_i}\right\}, \quad k=1,\ldots,p \\
\frac{\partial^2\ell}{\partial\omega^2} &= \sum_{i=1}^{n}\left\{\frac{h_{\omega\omega}(i)}{h_i} - \frac{h_\omega(i)}{h_i}\frac{h_\omega(i)}{h_i}\right\}
\end{aligned} \tag{38}$$

So we need to calculate $h_{\beta\beta}$, $h_{\beta\omega}$, and $h_{\omega\omega}$,

$$\begin{aligned}
&h_{\beta\beta}(k,m,i) = \\
&\quad\frac{\partial}{\partial\beta_k}\int_{-\infty}^{\infty}\varphi(u)\prod_{j=1}^{n_i}P(\boldsymbol{\beta}\mathbf{x}_{ij}+\sigma u,\ y_{ij})\sum_{j=1}^{n_i}x_{ijm}G(\boldsymbol{\beta}\mathbf{x}_{ij}+\sigma u,\ y_{ij})du = \\
&\qquad\int_{-\infty}^{\infty}\varphi(u)\prod_{j=1}^{n_i}P(\boldsymbol{\beta}\mathbf{x}_{ij}+\sigma u,\ y_{ij})\Bigg\{\sum_{j=1}^{n_i}x_{ijk}G(\boldsymbol{\beta}\mathbf{x}_{ij}+\sigma u,\ y_{ij}) \\
&\quad\times\sum_{j=1}^{n_i}x_{ijm}G(\boldsymbol{\beta}\mathbf{x}_{ij}+\sigma u,\ y_{ij}) + \sum_{j=1}^{n_i}x_{ijk}x_{ijm}H(\boldsymbol{\beta}\mathbf{x}_{ij}+\sigma u,\ y_{ij})du \\
&\qquad\qquad\qquad\qquad\qquad\qquad k,m=1,\ldots,p;\ i=1,\ldots,n
\end{aligned}$$

where

$$H(x,y) = \frac{\partial^2}{\partial x^2}\log P(x,y) = \frac{\partial}{\partial x}G(x,y).$$

$$\begin{aligned}
&h_{\beta\omega}(k,i) = \\
&\quad\frac{\partial}{\partial\beta_k}\sigma\int_{-\infty}^{\infty}u\varphi(u)\prod_{j=1}^{n_i}P(\boldsymbol{\beta}\mathbf{x}_{ij}+\sigma u,\ y_{ij})\sum_{j=1}^{n_i}G(\boldsymbol{\beta}\mathbf{x}_{ij}+\sigma u,\ y_{ij})du = \\
&\qquad\sigma\int_{-\infty}^{\infty}u\varphi(u)\prod_{j=1}^{n_i}P(\boldsymbol{\beta}\mathbf{x}_{ij}+\sigma u,\ y_{ij})\Bigg\{\sum_{j=1}^{n_i}x_{ijk}G(\boldsymbol{\beta}\mathbf{x}_{ij}+\sigma u,\ y_{ij}) \\
&\quad\times\sum_{j=1}^{n_i}G(\boldsymbol{\beta}\mathbf{x}_{ij}+\sigma u,\ y_{ij}) + \sum_{j=1}^{n_i}x_{ijk}H(\boldsymbol{\beta}\mathbf{x}_{ij}+\sigma u,\ y_{ij})\Bigg\}du \\
&\qquad\qquad\qquad\qquad\qquad\qquad k=1,\ldots,p;\ i=1,\ldots,n
\end{aligned}$$

$$
\begin{aligned}
&h_{\omega\omega}(i) = \\
&\quad \frac{\partial}{\partial\omega}\sigma\int_{-\infty}^{\infty} u\varphi(u)\prod_{j=1}^{n_i} P(\boldsymbol{\beta}\mathbf{x}_{ij}+\sigma u,\ y_{ij})\sum_{j=1}^{n_i} G(\boldsymbol{\beta}\mathbf{x}_{ij}+\sigma u,\ y_{ij})du = \\
&\qquad \sigma\int_{-\infty}^{\infty} u\varphi(u)\prod_{j=1}^{n_i} P(\boldsymbol{\beta}\mathbf{x}_{ij}+\sigma u,\ y_{ij})\Big\{\sum_{j=1}^{n_i} G(\boldsymbol{\beta}\mathbf{x}_{ij}+\sigma u,\ y_{ij}) \\
&\quad \times\big(1+\sigma u\sum_{j=1}^{n_i} G(\boldsymbol{\beta}\mathbf{x}_{ij}+\sigma u,\ y_{ij})\big)+\sigma u\sum_{i=1}^{n_i} H(\boldsymbol{\beta}\mathbf{x}_{ij}+\sigma u,\ y_{ij})\Big\}du \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad i=1,\ldots,n
\end{aligned}
$$

B Some necessary derivatives using the Laplace or Gauss-Hermite approximation

The basic partial derivatives are:

$$
\begin{aligned}
g_\sigma &= u\sum G(\boldsymbol{\beta}\mathbf{x}_{ij}+\sigma u,\ y_{ij}) \\
g_{\beta_m} &= \sum x_{ijm}G(\boldsymbol{\beta}\mathbf{x}_{ij}+\sigma u,\ y_{ij}),\quad m=1,\ldots,p \\
g_{u\sigma} &= u\sigma\sum H(\boldsymbol{\beta}\mathbf{x}_{ij}+\sigma u,\ y_{ij})+\sum G(\boldsymbol{\beta}\mathbf{x}_{ij}+\sigma u,\ y_{ij}) \\
g_{u\beta_m} &= \sigma\sum x_{ijm}H(\boldsymbol{\beta}\mathbf{x}_{ij}+\sigma u,\ y_{ij}),\quad m=1,\ldots,p \\
g_{uu\sigma} &= 2\sigma\sum H(\boldsymbol{\beta}\mathbf{x}_{ij}+\sigma u,\ y_{ij})+u\sigma^2\sum I(\boldsymbol{\beta}\mathbf{x}_{ij}+\sigma u,\ y_{ij}) \\
g_{uu\beta_m} &= \sigma^2\sum x_{ijm}I(\boldsymbol{\beta}\mathbf{x}_{ij}+\sigma u,\ y_{ij}) \\
g_{uuu} &= \frac{d^3}{du^3}\log p(u)+\sigma^3\sum I(\boldsymbol{\beta}\mathbf{x}_{ij}+\sigma u,\ y_{ij})
\end{aligned}
\tag{39}
$$

where

$$
I(\boldsymbol{\beta}\mathbf{x}_{ij}+\sigma u,\ y_{ij}) = \frac{\partial}{\partial x}H(x,y), \tag{40}
$$

For the calculation of the Hessian we need

$$
\begin{aligned}
g_{\sigma\sigma} &= u^2\sum H(\boldsymbol{\beta}\mathbf{x}_{ij}+\sigma u,\ y_{ij}) \\
g_{\sigma\beta_m} &= u\sum x_{ijm}H(\boldsymbol{\beta}\mathbf{x}_{ij}+\sigma u,\ y_{ij}) \\
g_{\beta_m\beta_k} &= \sum x_{ijm}x_{ijk}H(\boldsymbol{\beta}\mathbf{x}_{ij}+\sigma u,\ y_{ij}),
\end{aligned}
\tag{41}
$$

and

$$
\begin{aligned}
g_{u\sigma\sigma} &= u^2\sigma\sum I(\boldsymbol{\beta}\mathbf{x}_{ij}+\sigma u,\ y_{ij})+2u\sum H(\boldsymbol{\beta}\mathbf{x}_{ij}+\sigma u,\ y_{ij}) \\
g_{u\sigma\beta_m} &= u\sigma\sum x_{ijm}I(\boldsymbol{\beta}\mathbf{x}_{ij}+\sigma u,\ y_{ij})+\sum x_{ijm}H(\boldsymbol{\beta}\mathbf{x}_{ij}+\sigma u,\ y_{ij}) \\
g_{u\beta_m\beta_k} &= \sigma\sum x_{ijm}x_{ijk}I(\boldsymbol{\beta}\mathbf{x}_{ij}+\sigma u,\ y_{ij}),
\end{aligned}
\tag{42}
$$

and

$$\begin{aligned}
g_{uu\sigma\sigma} &= u^2\sigma^2 \sum K(\boldsymbol{\beta}\mathbf{x}_{ij} + \sigma u,\ y_{ij}) \\
&\quad + 4u\sigma \sum I(\boldsymbol{\beta}\mathbf{x}_{ij} + \sigma u,\ y_{ij}) + 2 \sum H(\boldsymbol{\beta}\mathbf{x}_{ij} + \sigma u,\ y_{ij}) \\
g_{uu\sigma\beta_m} &= u\sigma^2 \sum x_{ijm} K(\boldsymbol{\beta}\mathbf{x}_{ij} + \sigma u,\ y_{ij}) + 2\sigma \sum x_{ijm} I(\boldsymbol{\beta}\mathbf{x}_{ij} + \sigma u,\ y_{ij}) \\
g_{uu\beta_m\beta_k} &= \sigma^2 \sum x_{ijm} x_{ijk} K(\boldsymbol{\beta}\mathbf{x}_{ij} + \sigma u,\ y_{ij}),
\end{aligned} \tag{43}$$

where

$$K(\boldsymbol{\beta}\mathbf{x}_{ij} + \sigma u,\ y_{ij}) = \frac{\partial}{\partial x} I(x, y), \tag{44}$$

and

$$\begin{aligned}
g_{uuuu} &= \frac{d^4}{du^4} \log p(u) + \sigma^4 \sum K(\boldsymbol{\beta}x_{ij} + \sigma u, y_{ij}) \\
g_{uuu\sigma} &= u\sigma^3 \sum K(\boldsymbol{\beta}x_{ij} + \sigma u, y_{ij}) + 3\sigma^2 \sum I(\boldsymbol{\beta}\mathbf{x}_{ij} + \sigma u,\ y_{ij}) \\
g_{uuu\beta_m} &= \sigma^3 \sum x_{ijm} K(\boldsymbol{\beta}x_{ij} + \sigma u, y_{ij})
\end{aligned} \tag{45}$$