Comparison of non-stochastic approximations to the likelihood function for binomial-normal models: a research proposal

Rune Haubo Bojesen Christensen

August 11, 2010

Abstract

In this research proposal we are concerned with the accuracy, numerical stability and estimation time of three popular approximations to the log-likelihood function: the Laplace approximation, Gauss-Hermite quadrature and adaptive Gauss-Hermite quadrature. We consider generalized linear mixed models and focus on binomial-normal models with a single scalar random effects term. We outline how the accuracy of the approximations and estimators can be evaluated and compared. We describe the concepts and mathematics of the three approximations and suggest research that can provide more insight about the properties of these approximations.

Contents

1 Introduction **3**

2 Methods **5**

2.1 Accuracy of approximations and estimators 5

2.2 Asymptotic estimators for binomial observations 6

2.3 Comparison of approximations and estimators 6

2.4 Simulation of binomial-normal models 7

3 Background and research questions **8**

3.1 The Laplace approximation 8

3.2 Gauss-Hermite quadrature 9

3.3 Adaptive Gauss-Hermite quadrature 12

4 Conclusions **14**

A Estimation of regression parameters in GLMs **14**

A.1 GLM estimation 16

A.2 Binomial models 17

B Conditional mode estimation in GLMMs **18**

C Random effects estimation and ridge regression **18**

D The link between LA, PQL and the h-likelihood **19**

D.1 The PQL method 19

D.2 h-likelihood estimators 19

E ML versus REML-type approaches **19**

F Structural differences between GLMMs and NLMMs **20**

G Random effects structures **21**

1 Introduction

Mixed-effects models have proven a valuable class of models in so many areas of science and engineering where statistics are applied that they are now ubiquitous. Outside the normal linear framework evaluation of the likelihood function and optimization of it has, however, proven to be a considerable challenge and an active research area since the beginning of the 1990's with the seminal papers by Schall (1991) and Breslow and Clayton (1993). Since then a wealth of estimation methods have been proposed and compared. Among the most celebrated methods are penalized quasi likelihood (PQL) (Schall, 1991; Breslow and Clayton, 1993; Goldstein, 1986, 1989, 1991), the Laplace approximation (LA) (Liu and Pierce, 1994; Pinheiro and Bates, 1995; Pinheiro and Chao, 2006; Skaug and Fournier, 2006; Doran et al., 2007), Gauss-Hermite quadrature (GHQ) (Anderson and Aitkin, 1985; Lesaffre and Spiessens, 2001; Borjas and Sueyoshi, 1994; Hedeker and Gibbons, 1994, 1996; Lee, 2000) and adaptive Gauss-Hermite quadrature (AGQ) (Liu and Pierce, 1994; Pinheiro and Bates, 1995; Pinheiro and Chao, 2006), simulation methods and MCMC methods, possibly combined with an EM algorithm as in MCEM (Monte Carlo EM) or as in SEM (Stochastic EM) (McCulloch, 1994; Chan and Kuk, 1997; McCulloch, 1997; Booth and Hobert, 1999; Millar, 2004). Naturally there are also Bayesian attempts closely linked with MCMC methods (Zeger and Karim, 1991; Karim and Zeger, 1992). Several monographs discuss mixed models and their computation, including (McCulloch and Searle, 2001; Fahrmeir and Tutz, 2001; Diggle et al., 2002; Skrondal and Rabe-Hesketh, 2004; Demidenko, 2004; Fitzmaurice et al., 2009).

Two important classes of mixed-effects models outside the normal linear framework are the generalized linear mixed models (GLMMs) and (Gaussian) nonlinear mixed models (NLMMs). Structural differences between GLMMs and NLMMs are described in appendix F. Computational methods for these two classes have developed partially independently of each other but with a significant overlap of methodology. Their synthesis; Generalized nonlinear mixed models seem much less frequent.

Suppose we have observations from N clusters each with n_i observations such that y_{ij} is the jth observation on the ith cluster with $i = 1, \ldots, N$ and $j = 1, \ldots, n_i$. Suppose also that the random effects u_i for all i have a standard normal distribution with density, $p(u_i)$ and that the distribution of the observations conditional on the random effects has density[1] $p_{\boldsymbol{\alpha}}(y_{ij}|u_i)$, where $\boldsymbol{\alpha}$ is the parameter vector including the variance, σ_u^2 of the random effects.

The likelihood function is the joint density of the observations, $p_{\boldsymbol{\alpha}}(\boldsymbol{y})$ taken as a function of the parameters, $\boldsymbol{\alpha}$. This density is not directly available, but by standard rules of probability calculus it is given by

$$p_{\boldsymbol{\alpha}}(\boldsymbol{y}) = \int p_{\boldsymbol{\alpha}}(\boldsymbol{y}, \boldsymbol{u}) \, \mathrm{d}\boldsymbol{u} = \int p(\boldsymbol{u}) p_{\boldsymbol{\alpha}}(\boldsymbol{y}|\boldsymbol{u}) \, \mathrm{d}\boldsymbol{u} \tag{1}$$

The log-likelihood can therefore be written as

$$\ell(\boldsymbol{\alpha}; \boldsymbol{y}) = \log p_{\boldsymbol{\alpha}}(\boldsymbol{y}) = \log \int p(\boldsymbol{u}) p_{\boldsymbol{\alpha}}(\boldsymbol{y}|\boldsymbol{u}) \, \mathrm{d}\boldsymbol{u} \tag{2}$$

$$= \sum_{i=1}^{N} \log \int p(u_i) \prod_{j=1}^{n_i} p_{\boldsymbol{\alpha}}(y_{ij}|u_i) \, \mathrm{d}u_i \tag{3}$$

[1] Even though the distribution of the observations can be discrete, e.g. binomial or Poisson, we refer to the probability mass functions or probability density function of y_{ij} collectively as *density*.

where the last equality holds when observations from different clusters are assumed independent given the random effects.

The root of the computational challenge is the integral in (2), which, save for normal linear mixed models, does not have a closed form solution. Several approaches have been employed to overcome the integral. The integrand can be approximated by a function for which the integral has a closed form expression as in the Laplace approximation, the integral can be evaluated by numerical approximations, e.g. using Gauss-Hermite quadrature methods or by stochastic approximations. The accuracy of the latter methods can to some degree be improved by using more quadrature nodes or increasing the number of simulations. All these methods are based on the optimization of an approximate likelihood function. Other approximate methods such as the penalized quasi likelihood (PQL) method and relatives are defined as an algorithm and are not formulated as an approximation to an objective function.

This research proposal will be concerned with the accuracy, numerical stability and estimation time of binomial-normal models with a single scalar random effects term, that is models of the form (3). Sometimes we want results that are as accurate as possible, but we always want to know how accurate our results are for them to be scientifically meaningful. Generally there is a trade off between speed on one hand and accuracy and reliability on the other hand. Numerical stability is the ability of the method to yield the same accuracy for similar but slightly different data configurations. Predictable and understandable variations in accuracy are to some extent acceptable, but unpredictable oscillations in accuracy are not acceptable, because it would be impossible to know what the expected accuracy of the method would be on any particular data set. Focus is restricted to models of the form (3) because it is possible to evaluate the likelihood of such models with high accuracy and therefore a benchmark is available against which various approximations can be compared. The more general models of the form (2) are obviously also of practical interest, but consideration of these models is outside the scope of this proposal. Hopefully some results will generalize or some ideas for examining the accuracy of these models will emerge from the present project.

The focus will be on methods known as the Laplace approximation, Gauss-Hermite quadrature and adaptive Gauss-Hermite quadrature. This excludes stochastic methods and methods that cannot be formulated as an approximation to the likelihood function. Bayesian methods are also excluded due to the focus on the likelihood function. The three chosen methods are closely connected mathematically and computationally, so it makes sense to treat them together. They are also among the most widely implemented methods in statistical software packages and therefore of most interest to users of statistical methods.

The class of models is also restricted to binomial-normal models with logit and probit links. Some considerations will generalize to GLMMs in general, but only occasional references will be made to NLMMs.

Binomial-normal models are one of the most widely applied instances of the GLMMs and therefore of particular interest. While the logit link is often used in medical and biological applications, the probit link is almost exclusively used in applications in the social sciences and econometrics. PQL and Laplace approximations are known to be the least accurate for paired binary data and GHQ has been reported unstable for binomial data with large denominators, so the most challenging cases for the computational methods that we consider are covered by the binomial-normal models considered here. The AGQ approximation is generally believed to be numerically stable, the most accurate, but also the computationally

most intensive and therefore with the longest estimation time. Another important GLMM not considered here is the Poisson-normal model much used in biological applications.

In section 2 methodology for the evaluation and comparison of the likelihood approximations will be described. In section 3 some background on the three computational methods, LA, GHQ and AGQ will be given. We will provide insight on the behavior of these methods and outline ideas for how their properties could be studied. In section 4 we wrap up with conclusions.

2 Methods

In this section the methods that will be used to address the main research question outlined in section 1 will be described.

2.1 Accuracy of approximations and estimators

Two types of accuracies are generally of interest: the accuracy of an approximation and the accuracy of an estimator. The former is mainly a computational issue and the latter is mainly a statistical issue. The literature seems most occupied by the latter type of accuracy, but often the two are not clearly distinguished.

Let θ denote a *true* parameter, let θ_{ML} denote the ML estimator of θ, let θ_X denote some other estimator, e.g. a REML-type estimator, and let θ_X^Y denote the θ_X estimator computed with computational approximation Y, where Y is one of (PQL, LA, GHQ$_n$, AGQ$_n$) or some other computational method, and n denotes the number of quadrature nodes.

In the literature, almost exclusively, one finds comparisons of θ_X^Y to θ for some values of Y and X. This blurs the distinction between estimator and computational method and makes it difficult to tell whether any discrepancy between θ_X^Y and θ is due to bias in the estimator (X) or inaccuracy in the computational approximation (Y).

If θ is a variance parameter, e.g. σ_u^2, then θ_{ML} is *not* an unbiased estimator, especially not in situations where the information about θ is moderate and/or when θ is close to the boundary of its parameter space, i.e. when $\sigma_u = 0$. Bias in variance parameter estimators for example for the PQL estimator has been the focus of much research (e.g. Breslow and Lin (1995); Pawitan (2001); Breslow (2003)). The accuracy of a computational method can be assessed by comparing θ_X to θ_X^Y where Y is the computational method of interest. The accuracy of an estimator X can be assessed by comparing θ_X to θ. The latter requires that θ_X can be evaluated with sufficient accuracy, which can be an impediment.

A popular choice of estimator is the ML estimator, but for reasons of bias, particularly in variance parameters, some statisticians prefer a REML-type estimator. There is however more to a statistical analysis than point estimation and the literature seems devoid of discussion and assessment of the accuracy of CIs. While Wald-based CIs can be reasonably precise for some regression parameters in GLMMs and NLMMs (but not always as the Hauck-Donner effect (Hauck Jr. and Donner, 1977) attests), it seems that Wald CIs are often grossly imprecise for variance parameters (Pawitan, 2000). For further discussion of ML versus REML-type approaches see appendix E.

In this project we will focus entirely on genuine likelihood methods and particularly on the

accuracy of computational approximations Y to the ML estimator, θ_{ML}^{Y}. We will establish that $\theta_{ML}^{AGQ_n}$ is numerically equivalent to θ_{ML} for high enough n and use this as a standard against which other approximations can be compared. Properties of the ML estimator can be derived from the comparison of θ used to generate the simulations and the estimated θ_{ML}, but this is not a primary focus here.

2.2 Asymptotic estimators for binomial observations

If the response Y is binomial and the covariates x are discrete, then there is a finite and relatively small number of distinct possible values of (y_{ij}, x_{ij}). Denote these possible sets by $k = 1, \ldots, K$. Let $\ell_{(k)}^{Y}$ denote the approximation to the log-likelihood function for computational method Y to the kth set. Further, let $p_{(k)}$ denote the probability with which the kth set occurs, then the limiting log-likelihood approximation reads

$$\ell_{(lim)}^{Y} = \sum_{k=1}^{K} p_{(k)} \ell_{(k)}^{Y}$$

and the limiting θ_{MLE}^{Y} estimator is the maximizer of $\ell_{(lim)}^{Y}$.

This avoids the use of time consuming Monte Carlo simulations to obtain θ_{MLE}^{Y} and provides it with any desired accuracy. Observe that $\ell_{(lim)}^{Y}$ is the likelihood for all K data sets with $p_{(k)}$ as weights, so basically standard estimation routines for GLMMs can be used directly.

Probably the simples case is that of paired binary data. Assume that $x_{ij} \in \{0, 1\}$ is a treatment indicator variable, $y_{ij} \in \{0, 1\}$ is the binary response and $n_i = 2$, then there are four possible response patterns whose distribution depends on the model.

Confidence intervals can also be assessed with this method, for example the limiting profile likelihood curves can be drawn, potentially with the limiting Wald approximation.

The same idea applies to Poisson count data where sets with $p_{(k)} < \varepsilon$ for some small ε are ignored. It also applies to cumulative link mixed models and other multinomial type models.

Joe (2008) used this approach to compare θ_{MLE}^{LA} to θ with mentioning of θ_{MLE}^{AGQ}. There is, however, no explicit comparison of approximation θ_{MLE}^{LA} to θ_{MLE}, nor of the accuracy of the ML estimator θ_{MLE} to θ.

Monte Carlo simulations may still be needed to evaluate the robustness of the computational methods toward unbalance, outliers, starting values etc., and possibly also to evaluate average estimation times.

2.3 Comparison of approximations and estimators

There are many choices to be made during implementation of computational methods and probably different software houses have made different choices. In commercial software packages the details of the implementations are not available and the values of tuning parameters may not be publicly available. This hampers comparison of computational methods since only specific, but generally unknown implementations are being compared.

As in all other scientific research it is important that results are reproducible. This reproducibility is, however, compromised when the actual code used is unavailable with all its

implementation choices.

The properties of an estimation method depend not only on the properties of the integral approximation, but also on how the likelihood function is optimized and how convergence is judged. Some papers describe algorithms particular to their formulation (Wolfinger and Lin, 1997; Raudenbush et al., 2000; Hedeker and Gibbons, 1994). The properties of these algorithms are in general unknown. It may be that the algorithms, due to alternating steps or other approximations, lead to estimators that are not the maximizers of the approximated likelihood. We use general purpose nonlinear quasi-Newton optimizers (Nielsen, 2000; Nielsen and Mortensen, 2009; Nocedal and Wright, 2006) with accurate finite difference evaluations of the gradient when needed to optimize the approximated log-likelihood function. This ensures that the point of convergence is an optimum of the approximated log-likelihood function, although it may in general be a local optimum.

The choice of convergence criteria are important, not only because they can grossly affect the parameter estimates obtained, but also because they are important in identifying model fits that did not converge. Necessary conditions for convergence are a small gradient and positive definiteness of the Hessian matrix. These can be approximated via finite differences using Richardson's extrapolation (Richardson, 1910; Richardson and Gaunt, 1927). See Eldén et al. (2004) for an introduction to Richardson's extrapolation and Gilbert (2009) for an implementation.

An accurate evaluation of the Hessian is also important in order to obtain accurate standard errors of the model parameters. While this is only rarely mentioned, some contributions propose to use the final BFGS-updated Hessian from a quasi-Newton optimization (e.g. Joe, 2008). This Hessian depends strongly on the choice of starting values and is in general an inaccurate approximation to the true Hessian.

Since regression parameters and variance parameters are not orthogonal in GLMMs (nor in NLMMs) it seems particularly relevant to take account of the uncertainty in the variance parameters when evaluating the covariance matrix for the regression parameters.

We will strive to give accurate mathematical and computational descriptions of the methods that we implement and compare. We will make the implementation of the methods and the code for the simulation studies publicly available, so that our work is transparent and reproducible by others. This will hopefully lead to a more fair comparison of the computational methods. We will describe the convergence criteria that are used, how many models converged and how the non-convergences were handled in simulation studies. We will also seek to quantify the accuracy of the simulation results to avoid erroneous conclusions based on too few simulations.

2.4 Simulation of binomial-normal models

Simulation of the probit-normal model can be done by the following data generating mechanism: For $i = 1, \ldots, N$ and $j = 1, \ldots, n_i$ generate

$$\begin{aligned} b_i &\sim N(0, \sigma_b^2 \sigma_\varepsilon^2) \\ \varepsilon_{ij} &\sim N(0, \sigma_\varepsilon^2) \\ y_{ij}^* &= \boldsymbol{x}_{ij}^T(\boldsymbol{\beta}\sigma_\varepsilon) + b_i + \varepsilon_{ij} \\ y_{ij} &= \mathrm{I}(y_{ij}^* > 0) \ . \end{aligned}$$

Given the same seed for the pseudo random number generator, this will give the same binary observations, y_{ij} for any valid choice of σ_ε.

If ε_{ij} are generated from a logistic distribution with variance σ_ε^2, the logistic-normal model, rather than the probit-normal model, is simulated and the normalized coefficients, $\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta}\sigma_\varepsilon$ and $\tilde{\sigma}_b = \sigma_b\sigma_\varepsilon$ will be comparable in size across distributional assumptions for ε_{ij}.

3 Background and research questions

3.1 The Laplace approximation

The Laplace approximation (Tierney and Kadane, 1986; Barndorff-Nielsen and Cox, 1979, 1989) was suggested for estimation in NLMMs by Pinheiro and Bates (1995, 2000). It was also considered for GLMMs by Liu and Pierce (1993, 1994) and further developed for nested random effect structures (multilevel models) by Pinheiro and Chao (2006) for canonical links. The accuracy of the LA for binomial and Poisson GLMMs and cumulative link mixed models, in particular the proportional odds models with random effects, was studied by (Joe, 2008).

The Laplace approximation corresponds to the approximation of the log-integrand by a quadratic function for which the integral has an analytical solution:

$$\begin{aligned}
\ell(\boldsymbol{\alpha};\boldsymbol{y}) &= \log \int \exp\{\log p_{\boldsymbol{\alpha}}(\boldsymbol{y},\boldsymbol{u})\}\,\mathrm{d}\boldsymbol{u} \\
&\approx \ell_{LA}(\boldsymbol{\alpha},\boldsymbol{y}) \equiv \log \int \exp\{t(\boldsymbol{\alpha},\boldsymbol{u};\boldsymbol{y})\}\,\mathrm{d}\boldsymbol{u} \\
\ell_{LA}(\boldsymbol{\alpha},\boldsymbol{y}) &= \log p_{\boldsymbol{\alpha}}(\boldsymbol{y},\hat{\boldsymbol{u}}) + \frac{q}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{D}(\boldsymbol{\alpha},\hat{\boldsymbol{u}})|
\end{aligned}$$

where

$$\begin{aligned}
\log p_{\boldsymbol{\alpha}}(\boldsymbol{y},\boldsymbol{u}) &= \log p_{\boldsymbol{\alpha}}(\boldsymbol{y}|\boldsymbol{u}) + \log p(\boldsymbol{u}) \\
t(\boldsymbol{\alpha},\boldsymbol{u};\boldsymbol{y}) &= \log p_{\boldsymbol{\alpha}}(\boldsymbol{y},\hat{\boldsymbol{u}}) - \frac{1}{2}(\boldsymbol{u}-\hat{\boldsymbol{u}})^T\boldsymbol{D}(\boldsymbol{\alpha},\hat{\boldsymbol{u}})(\boldsymbol{u}-\hat{\boldsymbol{u}}) \\
\boldsymbol{D}(\boldsymbol{\alpha},\hat{\boldsymbol{u}}) &= -\left.\frac{\partial^2 \log p_{\boldsymbol{\alpha}}(\boldsymbol{y},\boldsymbol{u})}{\partial\boldsymbol{u}\partial\boldsymbol{u}^T}\right|_{\boldsymbol{u}=\hat{\boldsymbol{u}}} \\
\hat{\boldsymbol{u}} &= \arg\max_{\boldsymbol{u}}\{\log p_{\boldsymbol{\alpha}}(\boldsymbol{y},\boldsymbol{u})\}
\end{aligned}$$

so $t(\boldsymbol{\alpha},\boldsymbol{u};\boldsymbol{y})$ is the second order Taylor approximation of the joint log density $\log p_{\boldsymbol{\alpha}}(\boldsymbol{y},\boldsymbol{u})$ in $\boldsymbol{u}$ around the mode $\hat{\boldsymbol{u}}$ and $\boldsymbol{D}(\boldsymbol{\alpha},\hat{\boldsymbol{u}})$ is the negative Hessian evaluated at the mode.

The accuracy of the LA therefore depends on the closeness of the joint log-density to a quadratic function, or equivalently, the closeness of the joint density to a Gaussian function. Since the marginal log-density of $\boldsymbol{u}$ is exactly quadratic in $\boldsymbol{u}$, the accuracy of the LA depends on the closeness of $\log p_{\boldsymbol{\alpha}}(\boldsymbol{y}|\boldsymbol{u})$ to a quadratic function in $\boldsymbol{u}$.

If $\log p_{\boldsymbol{\alpha}}(\boldsymbol{y}|\boldsymbol{u})$ is a binomial log-density, then the closeness to a quadratic function increases with the binomial denominator and with the closeness of π_i to $1/2$. For probit and logit links $\pi_i = 0.5$ when the linear predictor (cf. appendix A.1) $\eta_i = 0$, so large absolute values of η_i leads to less accuracy. For a single scalar random effects term we may for the jth

observation on the ith cluster write $\eta_{ij} = \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \sigma u_i$, so the accuracy of the LA decreases with the absolute size of $\boldsymbol{x}_{ij}^T\boldsymbol{\beta}$ and with the size of σ. The decrease of the accuracy of the LA with the size of σ is expected from published simulation studies, while the dependency of the accuracy on the mean structure, $\boldsymbol{x}_{ij}^T\boldsymbol{\beta}$ does not seem to have been studied. It seems appropriate to study the accuracy of the LA as a function of the size of the linear predictor or, equivalently, the size of the fitted probabilities.

For GLMMs the LA can be written (cf. eq. (15) in appendix B)

$$\ell_{LA}(\boldsymbol{\alpha};\boldsymbol{y}) = \log p_{\boldsymbol{\alpha}}(\boldsymbol{y}|\boldsymbol{u}) - \frac{1}{2}\boldsymbol{u}^T\boldsymbol{u} - \frac{1}{2}\log|\boldsymbol{I}_q - \boldsymbol{V}^T\boldsymbol{\Psi}^b\boldsymbol{V} - \boldsymbol{R}| \tag{4}$$

Pinheiro and Chao (2006) considered GLMMs with canonical link functions where the $\boldsymbol{R}$ term in eq. (4) vanishes (cf. appendix A.1). Doran et al. (2007) also considers the LA for GLMMs and (appear to) ignore the $\boldsymbol{R}$ term by referring to standard GLM weights in the Fisher scoring estimation of the random effect modes effectively using the expected Hessian rather than the observed. Pinheiro and Bates (1995) similarly used the expected Hessian to estimate the random effect modes for NLMMs.

The effect of using the expected rather than the observed Hessian in the LA has not yet been studied in the literature for neither NLMMs nor GLMMs to the best knowledge of the author. For NLMMs it can be argued that the difference between the full and expected Hessian is probably small (Pinheiro and Bates, 1995; Bates and Watts, 1980), they are even identical when the random effects appear linearly in the model function.

The PQL and h-likelihood methods are both connected to the LA and in some way approximations to the LA. The connection is further described in section D.

The computational problem of the conditional mode estimation is closely related to ridge regression (e.g. Hastie et al., 2001, p.60), cf. appendix C.

Shun and McCullagh (1995); Shun (1997) and Raudenbush et al. (2000) consider higher order Laplace approximations. Although more accurate than the ordinary LA, the generalisability to complex design structures has not been investigated. For models where AGQ apply, it has not been thoroughly investigated which estimation method is the fastest. Further, Shun and McCullagh (1995); Shun (1997) showed by asymptotic arguments that the error of the ordinary LA considered above does not diminish as the sample size increase in some models where the number of random effects also increase. We are not aware of any numerical assessments of this; it may be that the LA is an adequate approximation in many practical situations despite the lack of asymptotic arguments for its validity.

3.2 Gauss-Hermite quadrature

Standard, i.e. non-adaptive Gauss-Hermite quadrature (GHQ) is a method to approximate an integral by a finite weighted sum:

$$\int f(x)\exp(-x^2)\,\mathrm{d}x \approx \sum_{h=1}^{N_{GHQ}} w_h f(x_h)\ ,$$

where the *nodes*, x_h are roots of the N_{GHQ}'th order Hermite polynomial with associated *weights*, w_h. These can be found by algorithms described by Golub and Welsch (1969); Golub (1973) or from tables in Abramowitz and Stegun (1972). The weights satisfy $\sum_{h=1}^{N} w_h \equiv \sqrt{\pi}$

for all N. Pinheiro and Bates (1995) considered GHQ for NLMMs and found that it was unreliable. The accuracy of GHQ depends on the size of the variance parameter, and in discrete GLMMs, in contrast to NLMMs, the variance parameter is (loosely) bounded above. GHQ may therefore be unreliable in NLMMs, while reliable and accurate for at least some GLMMs. Since GHQ is computationally simpler and faster than AGQ, the method is of interest.

Gauss-Hermite quadrature is exact if the integrand is a normal density. Suppose

$$\begin{aligned} f(x) &= \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \\ &= \frac{1}{\sigma\sqrt{2\pi}} \exp(-x^{*2}), \quad x^* = \frac{x-\mu}{\sqrt{2}\sigma} \end{aligned}$$

then

$$\begin{aligned} \int f(x)\,\mathrm{d}x &= \sqrt{2}\sigma \int \frac{1}{\sigma\sqrt{2\pi}} \exp(-x^{*2})\,\mathrm{d}x^* \\ &= \frac{1}{\sqrt{\pi}} \int \exp(-x^{*2})\,\mathrm{d}x^* \\ &\approx \sum_{h=1}^{N_{GHQ}} w_h \frac{1}{\sqrt{\pi}} \equiv 1 \end{aligned}$$

for any number of nodes, so Gauss-Hermite quadrature is exact in this formulation—even with a single node and for a normal density with any valid mean and standard deviation. However, if we define $f^*(x) = f(x)\exp(x^2)$, then we may write the quadrature rule as

$$\begin{aligned} \int f(x)\mathrm{d}x &= \int f^*(x)\exp(-x^2)\mathrm{d}x \\ &\approx \sum_{h=1}^{N_{GHQ}} w_h f^*(x_h) \\ &= \sum_{h=1}^{N_{GHQ}} w_h \exp(x_h^2) f(x_h) \end{aligned}$$

in which case the Gauss-Hermite quadrature is not exact for normal densities. With one quadrature node the approximation yields $\sqrt{1/2}$ for the standard normal density. With ten quadrature nodes the approximation error is 1.24e-5, but the error quickly increases with departure from $\mu = 0$ and $\sigma = 1$. For example, if $\mu = 1$ and $\sigma = .3$, then the approximation error is 0.31.

For binomial-normal models (and GLMMs in general) the integrand has the form $p(b_i)\prod_{j=1}^{n_i} p_{\boldsymbol{\alpha}}(y_{ij}|b_i)$ cf. eq. (3) with linear predictor $\eta_{ij} = x_{ij}^T\boldsymbol{\beta} + b_i = x_{ij}^T\boldsymbol{\beta} + \sigma u_i$ cf. appendix B. We may therefore integrate out b_i with a Gauss-Hermite quadrature approximation as

$$\begin{aligned} \int p(b_i)\prod_{j=1}^{n_i} p_{\boldsymbol{\alpha}}(y_{ij}|b_i)\,\mathrm{d}b_i &\approx \sum_{h=1}^{N_{GHQ}} w_h \exp(x_h^2) p(x_h) \prod_{j=1}^{n_i} p_{\boldsymbol{\alpha}}(y_{ij}|x_h) \\ &= \frac{1}{\sigma\sqrt{2\pi}} \sum_{h=1}^{N_{GHQ}} w_h \exp\{x_h^2 - x_h^2/(2\sigma^2)\} \prod_{j=1}^{n_i} p_{\boldsymbol{\alpha}}(y_{ij}|x_h)\ . \end{aligned}$$

If instead we integrate out u_i we get

$$\int p(u_i) \prod_{j=1}^{n_i} p_{\boldsymbol{\alpha}}(y_{ij}|u_i)\,\mathrm{d}u_i \approx \sum_{h=1}^{N_{GHQ}} w_h \exp(x_h^2) p(x_h) \prod_{j=1}^{n_i} p_{\boldsymbol{\alpha}}(y_{ij}|x_h)$$
$$= \frac{1}{\sqrt{2\pi}} \sum_{h=1}^{N_{GHQ}} w_h \exp(x_h^2/2) \prod_{j=1}^{n_i} p_{\boldsymbol{\alpha}}(y_{ij}|x_h)$$

Alternatively, by changing the variable of integration to $x_i^* = b_i/(\sigma\sqrt{2})$, or equivalently $x_i^* = u_i/\sqrt{2}$, we may write the Gauss-Hermite quadrature approximation as

$$\int p(b_i) \prod_{j=1}^{n_i} p_{\boldsymbol{\alpha}}(y_{ij}|b_i)\,\mathrm{d}b_i = \sigma\sqrt{2} \int \frac{1}{\sigma\sqrt{2\pi}} \exp(-x_i^{*2}) p_{\boldsymbol{\alpha}}(y_{ij}|b_i)\,\mathrm{d}x_i^*$$
$$\approx \frac{1}{\sqrt{\pi}} \sum_{h=1}^{N_{GHQ}} w_h \prod_{j=1}^{n_i} p_{\boldsymbol{\alpha}}(y_{ij}|b_h), \quad b_h = x_h \sigma\sqrt{2} \tag{5}$$

in which case it does not matter if we integrate out b_i or u_i. The point is that despite the mathematical equivalence of the integrations, the actual formulations and implementations of GHQ can make a difference. There seems to be no quantitative examination of this in the literature.

Borjas and Sueyoshi (1994) observed that underflow can occur if, in formulation (5), n_i is large and each $p_{\boldsymbol{\alpha}}(y_{ij}|\cdot)$ is sufficiently small. The lower bound on floating point number representation in double precision is around 1e-324 and the limit below which underflow may occur is around 1e-308, i.e. roughly a factor 1e16 larger than the lower bound. If $p_{\boldsymbol{\alpha}}(y_{ij}|\cdot) = 1/2$ for all j, then underflow may occur with more than $308 \log 10/\log 2 \approx 1023$ per cluster binary observations. If instead $p_{\boldsymbol{\alpha}}(y_{ij}|\cdot) = .1$, only 308 per cluster observations will lead to underflow, while if $p_{\boldsymbol{\alpha}}(y_{ij}|\cdot) = .9$, roughly 6731 per cluster observations are needed to cause underflow. More detailed examination of when underflow can occur and the consequences for likelihood approximations seem unavailable in the literature.

Lee (2000) proposed an algorithm to avoid underflow. He writes the GHQ approximation to the likelihood function in the form

$$\ell_{GHQ}(\boldsymbol{\alpha}; \boldsymbol{y}) = \sum_{i=1}^{N} \log \left\{ \frac{1}{\sqrt{\pi}} \sum_{h=1}^{N_{GHQ}} w_h \prod_{j=1}^{n_i} p_{\boldsymbol{\alpha}}(y_{ij}|x_h \sigma\sqrt{2}) \right\} \tag{6}$$

and suggests to compute this as

$$\ell_{GHQ}(\boldsymbol{\alpha}; \boldsymbol{y}) = \sum_{i=1}^{N} \sum_{j=1}^{n_i} \log \left\{ \sum_{h=1}^{N_{GHQ}} p_{\boldsymbol{\alpha}}(y_{ij}|x_h \sigma\sqrt{2}) \omega_{j-1,ih} \right\}$$

where the weights ω_{jih} are given by

$$\omega_{jih} = \frac{p_{\boldsymbol{\alpha}}(y_{ij}|x_h \sigma\sqrt{2}) \omega_{j-1,ih}}{\sum_{s=1}^{N_{GHQ}} p_{\boldsymbol{\alpha}}(y_{ij}|x_h \sigma\sqrt{2}) \omega_{j-1,is}} \tag{7}$$

and $\omega_{0ih} = w_h/\sqrt{\pi}$ for all h and i. This estimation scheme effectively interchanges the inner product and sum. This is also a well known trick to avoid numerical underflow in the

estimation of hidden Markov models (see e.g. Zucchini and MacDonald, 2009, p.46). Since the denominator of (7) is the ijth contribution to the likelihood function, the likelihood can be computed by the following algorithm

for $i = 1$ to N **do**
  for $j = 1$ to n_i **do**
    for $h = 1$ to N_{GHQ} **do**
      compute ω_{jih}
    end for
    store the denominator of (7) as C_{ij}
  end for
end for
compute $\ell(\boldsymbol{\alpha}; \boldsymbol{y}) = \sum_{ij} C_{ij}$

This estimation scheme involves more computations than the direct evaluation of the log-likelihood function (6). The computational overload has, however, not been quantified.

Hedeker and Gibbons (1994, 1996) implement gradients and Hessian of the log-likelihood function using GHQ. Lesaffre and Spiessens (2001) remarks on the potential inadequacy in the approximations of the gradient and Hessian. The reported inaccuracy in the GHQ methods may therefore be due to other computational choices than the GHQ approximation to the likelihood function.

Lesaffre and Spiessens (2001) and Rabe-Hesketh et al. (2005) find that GHQ is unreliable and gives biased estimates of variance parameters with large cluster sizes and larger variances. These authors also speculate that the reason for the failure of GHQ is that the integrand for a cluster contribution to the likelihood function is highly peaked and narrow, and so can fall, almost entirely, in between quadrature nodes.

Anderson and Aitkin (1985) is an early application of GHQ for the estimation of binomial-normal models. The paper contains a profile likelihood curve demonstrating the inappropriateness of a quadratic approximation to this.

3.3 Adaptive Gauss-Hermite quadrature

AGQ was proposed by Liu and Pierce (1994) as an improvement to GHQ. They noted that this method could prove valuable for GLMMs and remarked on the connection to the LA. Pinheiro and Bates (1995) suggested AGQ for NLMMs and remarked on the connection to GHQ, the LA and motivated AGQ as the equivalent of importance sampling for GHQ. Pinheiro and Chao (2006) extended AGQ to multilevel GLMMs with canonical link functions. Liu and Pierce (1994) and Pinheiro and Bates (1995); Pinheiro and Chao (2006) shifted and scaled the quadrature nodes by the mode of the integrand and the Hessian at the mode. Rabe-Hesketh et al. (2005) and Naylor and Smith (1988) on the other hand shifted and scaled the quadrature nodes by the mean and the variance of the integrand. Rabe-Hesketh et al. (2005) also used AGQ to approximate the integrals defining the mean and variance of the integrand. They kept the location of the quadrature nodes fixed when evaluating the finite difference approximation to the gradient and Hessian for use in a Newton scheme for the estimation of the model parameters.

Extension of quadrature methods (GHQ and AGQ) to integrals of more than one dimension is difficult since the number of quadrature nodes increases rapidly. Rabe-Hesketh et al. (2005) proposed to use spherical quadrature rules (Stroud, 1971; Naylor and Smith, 1988)

while Heiss and Winschel (2008) proposed to use a sparse grid integration rule (Smolyak, 1963; Gerstner and Griebel, 1998). Pinheiro and Chao (2006) used Cartesian quadrature for multilevel models exploiting the conditional independence structure of problem.

Following Naylor and Smith (1982) and Liu and Pierce (1994) Gauss-Hermite quadrature can be re-expressed in terms of a normal density, $\phi(t;\mu,\sigma)$ rather than $\exp(-x^2)$:

$$\begin{aligned}\int f(t)\phi(t;\mu,\sigma)\,\mathrm{d}t &= \int f(t)\frac{1}{\sigma\sqrt{2\pi}}\exp\left\{-\frac{(t-\mu)^2}{2\sigma^2}\right\}\,\mathrm{d}t \\ &= \frac{1}{\sqrt{\pi}}\int f(t)\exp(-z^2)\,\mathrm{d}z, \quad t=\mu+\sqrt{2}\sigma z \\ &\approx \sum_{h=1}^{N_{GHQ}}\frac{w_h}{\sqrt{\pi}}f(t_h), \quad t_h=\mu+\sqrt{2}\sigma x_h\end{aligned}$$

For integration of $g(t)$, let $h(t)=g(t)/\phi(t;\mu,\sigma)$, so we may write

$$\begin{aligned}\int g(t)\,\mathrm{d}t &= \int h(t)\phi(t;\mu,\sigma)\,\mathrm{d}t \\ &\approx \sum_{h=1}^{N_{GHQ}}\frac{w_h}{\sqrt{\pi}}h(t_h), \quad t_h=\mu+\sqrt{2}\sigma x_h \\ &= \sigma\sqrt{2}\sum_{h=1}^{N_{GHQ}} w_h\exp(x_h^2)g(t_h)\end{aligned}$$

subject to appropriate choice of the tuning parameters μ and σ. Liu and Pierce (1994) suggested to take μ to be the mode of $g(t)$ and $\sigma=1/D$, where

$$D=-\frac{\partial^2}{\partial t^2}\log g(t)\bigg|_{t=\mu}$$

Thus, if $g(t)$ is a Gaussian function, the quadrature rule is exact with a single node. Pinheiro and Bates (1995) suggested essentially the same approximation although using $\mathsf{E}(D)$ rather than D similar to their modification of the Laplace approximation, cf. section 3.1. There do not seem to be any quantitative assessment of the difference between using $\mathsf{E}(D)$ or D in the literature for GLMMs and NLMMs. This version of adaptive Gauss-Hermite quadrature is the Laplace approximation when a single node is used. For a reasonable number of nodes, e.g. ten, we may expect the quadrature rule to be insensitive to small differences in the choice of μ and σ. The main objective is that the integrand is reasonably covered by the quadrature nodes. The quadrature rule may therefore not be sensitive to the choice of expected versus observed Hessian, nor to whether the values of μ and σ are updated for gradient evaluations. Perhaps even more approximate estimation of μ and σ will be sufficient.

The AGQ approximation to the log-likelihood of a GLMM can be written as

$$\begin{aligned}\ell_{AGQ}(\boldsymbol{\alpha};\boldsymbol{y}) &= \sum_{i=1}^{N}\log\left\{\sigma_i\sqrt{2}\sum_{h=1}^{N_{AGQ}} w_h\exp(x_h^2)p(t_{hi})\prod_{j=1}^{n_i}p_{\boldsymbol{\alpha}}(y_{ij}|t_{hi})\right\}, \quad t_{hi}=\mu_i+\sigma_i\sqrt{2}x_h \\ &= \sum_{i=1}^{N}\log\left\{\frac{\sigma_i}{\sqrt{\pi}}\sum_{h=1}^{N_{AGQ}} w_h\exp(x_h^2)\exp(-t_{hi}^2/2)\prod_{j=1}^{n_i}p_{\boldsymbol{\alpha}}(y_{ij}|t_{hi})\right\} \qquad (8)\end{aligned}$$

Rabe-Hesketh et al. (2005, p.303) states that "... adaptive quadrature is superior [to ordinary quadrature] since it requires fewer quadrature nodes." in their discussion of GHQ and AGQ. It is not clear whether the (claimed) superiority is with respect to integration accuracy, computational speed, both or some other feature.

Approximating an integral with quadrature naturally takes less time the fewer the number of quadrature nodes. While AGQ often, but maybe not always, needs fewer nodes than GHQ to obtain the same accuracy, the shifting and scaling of the quadrature nodes used in AGQ need to be determined from the integrand and this takes time as well. Whether the more complicated process involved in AGQ takes longer or shorter time than do GHQ seems not to have been investigated from reading the literature. Such an assessment naturally depends on the particular implementation of the quadrature methods and for the AGQ how the measures of shift and scaling is obtained. It may be, however, that there are cases where GHQ is unable to provide any reasonable accuracy irrespectively of how many nodes are used, while AGQ can work adequately.

4 Conclusions

Ideally we would like to study the accuracy, the estimation time and the numerical stability of the approximations as a function of the sample size, the size of the binomial denominator, the size of the variance parameter, the size of fixed effects, the number of clusters, the number of observations per cluster, the variance in the number of within cluster observations (the degree of balance) and the presence of outliers. Further, for the LA the convergence criteria of the inner loop and choice of the observed or expected Hessian is of interest. For the quadrature rules, the no. nodes is also of interest, for GHQ the choice of formulation, and for AGQ the convergence criteria of the inner loop and the choice of Hessian (i.e. scaling of the nodes) is of interest.

The accuracy, the estimation time and numerical stability can to some extent be studied without the use of simulations by comparing the integration accuracy of single integrals. The dependency on the size of the variance parameter, the size of the binomial denominator etc. can be studied in this way. Further, the asymptotic properties of the approximations and estimators can be studied via the method outlined in section 2.2. To further study the numerical stability, robustness to imbalance and outliers, etc. simulation studies have to be employed.

A Estimation of regression parameters in GLMs

Generalized linear models (GLMs) can be fitted in various ways. In this section two popular and closely related methods are described, namely a Newton algorithm and Fisher scoring. The model and its likelihood with gradients and Hessian are described below whereas the actual algorithms are described in section A.1. In section A.2 the details of binomial models and their estimation are worked out for logit and probit links.

GLMs are models where the response follows a distribution in the exponential family including Poisson, binomial, gamma and Gaussian distributions. The expected value of the

response is linked to a linear predictor, η_i through a link function, $h(\cdot)$

$$\mathsf{E}(Y_i|x_i) = \mu_i = h^{-1}(\eta_i), \quad \eta_i = x_i^T\boldsymbol{\beta}$$

The distribution of the response is a member of the exponential family of distributions with log density of the form

$$\log p(y_i) = \frac{1}{\varphi}(y_i\theta_i - b(\theta_i)) + c(y_i, \varphi) \tag{9}$$

where the canonical parameter $\theta = \theta(\mu_i)$ is a function of the mean, $\mathsf{E}(Y_i) = b'(\theta_i) = \mu_i$, $\mathsf{Var}(Y_i) = \frac{\varphi}{w_i}b''(\theta_i) = \frac{\varphi}{w_i}\mathsf{V}(\theta_i)$, where $\mathsf{V}(\theta_i)$ is the variance function and φ is an optional dispersion parameter. The term $c(y_i, \varphi)$ is a constant with respect to θ_i and ensures that the density integrates to one. The log-likelihood for $\boldsymbol{\beta}$ can be written as

$$\ell(\boldsymbol{\beta}, y_i) = w_i \log p(y_i) \tag{10}$$

where w_i is a potential weight associated with the ith observation.

The gradient of $\ell(\boldsymbol{\beta}, y_i)$ wrt. $\boldsymbol{\beta}$, i.e. the score function, is

$$S(\boldsymbol{\beta}, y_i) = \frac{\partial}{\partial\boldsymbol{\beta}}\ell(\boldsymbol{\beta}, y_i) = w_i\left[y_i\frac{\partial\theta_i}{\partial\boldsymbol{\beta}} - b'(\theta_i)\frac{\partial\theta_i}{\partial\boldsymbol{\beta}}\right] = w_i\left[\frac{\partial\theta_i}{\partial\boldsymbol{\beta}}(y_i - \mu_i)\right] \ ,$$

where $\partial b(\theta_i)/\partial\boldsymbol{\beta} = b'(\theta_i)\partial\theta_i/\partial\boldsymbol{\beta}$ and

$$\frac{\partial\theta_i}{\partial\boldsymbol{\beta}} = \frac{\partial\theta_i}{\partial\mu_i}\frac{\partial\mu_i}{\partial\eta_i}x_i = \mathsf{V}(\theta_i)^{-1}\frac{\partial\mu_i}{\partial\eta_i}x_i \ ,$$

since

$$\mathsf{V}(\theta_i) = b''(\theta_i) = \frac{\partial b'(\theta_i)}{\partial\theta_i} = \frac{\partial\mu_i}{\partial\theta_i} \ .$$

The term $\mathsf{V}(\theta_i)$ depends on the choice of $p(y_i)$ and $\partial\mu_i/\partial\eta_i$ depends on choice of $h(\cdot)$. If the canonical link is applied, then

$$\theta_i = \theta(\mu_i) = h(\mu_i) = \eta_i \ , \tag{11}$$

so $\partial\theta_i/\partial\boldsymbol{\beta} = x_i$.

The Hessian, i.e. the second order derivative of $\ell(\boldsymbol{\beta}, y_i)$ wrt. $\boldsymbol{\beta}$ is

$$H(\boldsymbol{\beta}; y_i) = w_i\left[\frac{\partial^2\theta_i}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}(y_i - \mu_i) - \left(\frac{\partial\theta_i}{\partial\boldsymbol{\beta}}\right)^2 b''(\theta_i)\right] \tag{12}$$

where

$$\frac{\partial^2\theta_i}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T} = \frac{\partial^2\theta_i}{\partial\mu_i^2}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2 x_i x_i^T + \frac{\partial^2\mu_i}{\partial\eta_i^2}\frac{\partial\theta_i}{\partial\mu_i}x_i x_i^T$$

The first term in (12) has expectation zero since $\mathsf{E}(Y_i) = \mu_i$. When the canonical link function is applied, then, using (11):

$$\frac{\partial^2\theta_i}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T} = \frac{\partial}{\partial\boldsymbol{\beta}^T}\left(\frac{\partial\theta_i}{\partial\boldsymbol{\beta}}\right) = \frac{\partial x_i}{\partial\boldsymbol{\beta}^T} = 0$$

so the first term in (12) vanishes, and $H(\boldsymbol{\beta}; y_i) = \mathsf{E}(H(\boldsymbol{\beta}; y_i))$.

The gradient and Hessian for all data are, assuming independence, given by

$$S(\boldsymbol{\beta};\boldsymbol{y}) = \sum_i S(\boldsymbol{\beta};y_i)$$
$$H(\boldsymbol{\beta};\boldsymbol{y}) = \sum_i H(\boldsymbol{\beta};y_i)$$

A.1 GLM estimation

In matrix notation the gradient is

$$S(\boldsymbol{\beta};\boldsymbol{y}) = \boldsymbol{X}^T\boldsymbol{\Psi}^a \tag{13}$$

where in general $\boldsymbol{\Psi}^a$ is an n-vector with elements

$$\Psi_i^a = w_i(y_i-\mu_i)\mathsf{V}(\theta_i)^{-1}\frac{\partial\mu_i}{\partial\eta_i}$$

but for canonical links

$$\Psi_i^a = w_i(y_i-\mu_i)\ .$$

The Hessian can be written as

$$H(\boldsymbol{\beta},\boldsymbol{y}) = \boldsymbol{X}^T\boldsymbol{\Psi}^b\boldsymbol{X} + \boldsymbol{X}^T\boldsymbol{\Psi}^c\boldsymbol{X} = \boldsymbol{X}^T(\boldsymbol{\Psi}^b+\boldsymbol{\Psi}^c)\boldsymbol{X} \tag{14}$$

where in general $\boldsymbol{\Psi}^b$ and $\boldsymbol{\Psi}^c$ are diagonal $n\times n$ matrices with elements

$$\Psi_i^b = -w_i\mathsf{V}(\theta_i)^{-1}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2$$
$$\Psi_i^c = w_i(y_i-\mu_i)\left[\frac{\partial^2\theta_i}{\partial\mu_i^2}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2 + \frac{\partial^2\mu_i}{\partial\eta_i^2}\mathsf{V}(\theta_i)^{-1}\right]$$

but for canonical links

$$\Psi_i^b = -w_i\mathsf{V}(\theta_i)$$
$$\Psi_i^c = 0\ .$$

The Newton update reads

$$\boldsymbol{\beta}^{(i+1)} = \boldsymbol{\beta}^{(i)} - H(\boldsymbol{\beta}^{(i)};\boldsymbol{y})^{-1}S(\boldsymbol{\beta}^{(i)};\boldsymbol{y})$$

while the Fisher scoring update reads

$$\boldsymbol{\beta}^{(i+1)} = \boldsymbol{\beta}^{(i)} - \mathsf{E}[H(\boldsymbol{\beta}^{(i)};\boldsymbol{y})]^{-1}S(\boldsymbol{\beta}^{(i)};\boldsymbol{y})$$

and parenthesized superscripts denote iteration number.

A.2 Binomial models

The log-Bernoulli probability mass function reads

$$\log p(y_i) = y_i \log \frac{\pi_i}{1-\pi_i} + \log(1-\pi_i) + c(y_i)$$

where y_i is the binary response. This is of the form (9) with elements $\theta_i = \log \frac{\pi_i}{1-\pi_i}$, $\pi_i = \frac{\exp(\theta_i)}{1+\exp(\theta_i)} = [1+\exp(-\theta_i)]^{-1}$, $b(\theta_i) = -\log(1-\pi_i) = \log(1+\exp(\theta_i))$, $\frac{\partial b(\theta_i)}{\partial \theta_i} = b'(\theta_i) = \pi_i$.

For Bernoulli, i.e. binary observations $\mathsf{E}(y_i|x_i) = \pi_i = \mu_i$, while for binomial observations with denominator m_i, $\mathsf{E}(y_i|x_i) = m_i\pi_i = \mu_i$. The likelihood function for binomial observations can be obtained by using as response y_i/m_i, the ratio of success, in the Bernoulli probability mass function and m_i as weights in (10).

For Bernoulli models we have that

$$\mathsf{V}(\theta_i)^{-1} = \frac{\partial \theta_i}{\partial \pi_i} = \frac{\partial}{\partial \pi_i} \log\left(\frac{\pi_i}{1-\pi_i}\right) = [\pi_i(1-\pi_i)]^{-1}$$

and

$$\frac{\partial^2 \theta_i}{\partial \pi_i^2} = \frac{\partial}{\partial \pi_i}\left(\frac{1}{\pi_i} + \frac{1}{1-\pi_i}\right) = -\pi^{-2} + (1-\pi)^{-2}$$

Further, for the probit link

$$\frac{\partial \pi_i}{\partial \eta_i} = \frac{\partial}{\partial \eta_i}\Phi(\eta_i) = \phi(\eta_i)$$

and

$$\frac{\partial^2 \pi_i}{\partial \eta_i^2} = \frac{\partial}{\partial \eta_i}\phi(\eta_i) = -\eta_i\phi(\eta_i) \ ,$$

so

$$\frac{\partial^2 \theta_i}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} = x_i x_i^T \left\{ \phi^2(\eta_i)\left[-\pi_i^{-2} + (1-\pi_i)^{-2}\right] - \frac{\eta_i \phi(\eta_i)}{\pi_i(1-\pi_i)} \right\} \ .$$

The gradient (13) and Hessian (14) are therefore identified with elements

$$\begin{aligned}
\Psi_i^a &= w_i(y_i - \pi_i)\phi(\eta_i)[\pi_i(1-\pi_i)]^{-1} \\
\Psi_i^b &= -w_i[\pi_i(1-\pi_i)]^{-1}\phi(\eta_i)^2 \\
\Psi_i^c &= w_i(y_i - \pi_i)\left\{ \phi(\eta_i)^2\left[-\pi_i^{-2} + (1-\pi_i)^{-2}\right] - \frac{\eta_i\phi(\eta_i)}{\pi_i(1-\pi_i)} \right\}
\end{aligned}$$

For the logit link we have

$$\frac{\partial \pi_i}{\partial \eta_i} = \frac{\partial}{\partial \eta_i}[1+\exp(-\eta_i)]^{-1} = \frac{\exp(-\eta_i)}{[1+\exp(-\eta)]^2}$$

and

$$\begin{aligned}
\Psi_i^a &= w_i(y_i - \pi_i) \\
\Psi_i^b &= -w_i\pi_i(1-\pi_i) \\
\Psi_i^c &= 0
\end{aligned}$$

B Conditional mode estimation in GLMMs

In a GLMM the conditional distribution of the response given the random effects has an exponential family distribution with density, $p(\boldsymbol{y}|\boldsymbol{B}=\boldsymbol{b})$ and conditional mean satisfying

$$\mathsf{E}[\boldsymbol{y}|\boldsymbol{B}=\boldsymbol{b}] = h(\boldsymbol{\eta}), \quad \boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{b}$$

The marginal distribution of the q-dimensional random effects is multivariate normal:

$$\boldsymbol{B} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma}).$$

The linear predictor can be written as

$$\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{V}\boldsymbol{u},$$

where $\boldsymbol{V} = \boldsymbol{Z}\boldsymbol{\Lambda}$, $\boldsymbol{\Lambda}$ is the Cholesky factor of $\boldsymbol{\Sigma}$ such that $\boldsymbol{\Lambda}\boldsymbol{\Lambda}^T = \boldsymbol{\Sigma}$ and $\boldsymbol{U}$ are standard multivariate normal, $\boldsymbol{U} \sim N(\boldsymbol{0}, \boldsymbol{I}_q)$.

The joint log density is

$$\log p_{\boldsymbol{\alpha}}(\boldsymbol{y}, \boldsymbol{u}) = \log p_{\boldsymbol{\alpha}}(\boldsymbol{y}|\boldsymbol{u}) + \log p(\boldsymbol{u}),$$

where

$$\log p(\boldsymbol{u}) = -r\log(2\pi)/2 - \boldsymbol{u}^T\boldsymbol{u}/2$$

By analogy with the development in section A.1, the gradient wrt. $\boldsymbol{u}$ in the joint log density is

$$\boldsymbol{S}(\boldsymbol{u}; \boldsymbol{\alpha}, \boldsymbol{y}) = \boldsymbol{V}^T\boldsymbol{\Psi}^a - \boldsymbol{u},$$

and the Hessian is

$$\boldsymbol{H}(\boldsymbol{u}; \boldsymbol{\alpha}, \boldsymbol{y}) = \boldsymbol{V}^T\boldsymbol{\Psi}^b\boldsymbol{V} + \boldsymbol{R} - \boldsymbol{I}_q \tag{15}$$

where

$$\boldsymbol{R} = \boldsymbol{V}^T\boldsymbol{\Psi}^c\boldsymbol{V}$$

and $\boldsymbol{\Psi}^a$, $\boldsymbol{\Psi}^b$, $\boldsymbol{\Psi}^c$ are described in section A.1 and worked out for binomial models in section A.2.

C Random effects estimation and ridge regression

The computational problems is similar to a weighted version of ridge regression (Hastie et al., 2001, p.60):

$$\hat{\boldsymbol{\beta}}^{ridge} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

where the tuning parameter, λ is related to the size of the variance parameter, σ_u. The inflation of the diagonal of $\boldsymbol{Z}^T\boldsymbol{W}_{\boldsymbol{\alpha},\hat{\boldsymbol{u}}}\boldsymbol{Z}$ by $\boldsymbol{I}_r$ guaranties the positive definiteness of $\boldsymbol{D}$ and makes the computational problem well defined.

D The link between LA, PQL and the h-likelihood

D.1 The PQL method

The PQL method is motivated from the Laplace approximation by Breslow and Clayton (1993) and they also ignore the $\boldsymbol{R}$-term. The PQL estimates have not been shown to be the maximizers of a single objective function. Instead, Schall (1991) and Breslow and Clayton (1993) show that the estimates can be obtained by iteratively applying estimation methods for LMMs. Usually the REML-method is employed for estimation of the variance components. The estimators for the fixed effects parameters, $\boldsymbol{\beta}$ and the random effects $\boldsymbol{u}$ are the joint maximizers of the PQL

$$\log p_{\boldsymbol{\alpha}}(\boldsymbol{y}|\boldsymbol{u}) + \log p(\boldsymbol{u})$$

Which is the LA ignoring the last term. This could be appropriate if neither $\boldsymbol{\Psi}^b$, nor $\boldsymbol{R}$ depend much on $\boldsymbol{\beta}$. There do not seem to be any quantitative assessment of this dependency in the literature.

The variance parameters are then estimated from the REML likelihood for LMMs:

$$-\frac{1}{2}\log|\boldsymbol{V}| - \log|\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X}| - \frac{1}{2}(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})^T\boldsymbol{V}^{-1}(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})$$

where $\boldsymbol{V} = \boldsymbol{\Sigma}_\varepsilon^{-1} + \boldsymbol{Z}^T\boldsymbol{\Sigma}_u\boldsymbol{Z}$, $\boldsymbol{\Sigma}_u$ is the covariance matrix of the marginal distribution of $\boldsymbol{u}$ and $\boldsymbol{\Sigma}_\varepsilon$ is the covariance matrix of the approximated normal distribution for the conditional distribution of the response given the random effects.

D.2 h-likelihood estimators

The PQL is identical to the h-likelihood as defined by Lee and Nelder (1996, p. 621) although the h-likelihood estimator of variance parameters differ from the PQL estimator. Due to the dependence (non-orthogonality) between variance parameters and regression parameters, the h-likelihood estimator of the regression parameters is also different from the PQL estimator. In later papers, (e.g. Lee and Nelder, 2000, 2003, 2004, 2005, 2006; Lee et al., 2006, 2007) these authors describe a range of h-likelihood estimators with various adjustments or corrections. Noh and Lee (2007) adjust the PQL aka h-likelihood as defined above by $\frac{1}{2}\log|(\boldsymbol{V}_{\boldsymbol{\tau}}^T\boldsymbol{\Psi}^b\boldsymbol{V}_{\boldsymbol{\tau}} + \boldsymbol{I}_r)/(2\pi)|$ essentially obtaining the modified LA for $\boldsymbol{\beta}$ although the variance parameters are not obtained as the maximizers of the LA.

E ML versus REML-type approaches

In normal linear mixed models (LMMs) it is well known that the maximum likelihood (ML) estimator of σ_u^2 is not an unbiased estimator. This has prompted the popularity of the restricted (or residual) maximum likelihood (REML) estimator, which is a less biased estimator of σ_u^2 in some situations. Notably, the REML estimator also coincides with the estimator based on equating observed and expected mean squares in ANOVA tables that were used almost exclusively before computers became widespread. The REML estimator is not guarantied to be an unbiased estimator, nor is it always a more unbiased estimator

than the ML estimator (refs needed). The REML estimator was motivated as an estimator in the space of the residuals rather than in the observations. It has also been shown that the REML estimator of σ_u^2 can be motivated by integrating out the fixed effects regression parameters (essentially assuming a flat prior) in the likelihood function (e.g. Wolfinger, 1993; Bates and DebRoy, 2004):

$$\ell_{REML}(\boldsymbol{\tau}; \boldsymbol{y}) = \log \int \int p(\boldsymbol{u}) p_{\boldsymbol{\alpha}}(\boldsymbol{y}|\boldsymbol{u}) \, \mathrm{d}\boldsymbol{u} \, \mathrm{d}\boldsymbol{\beta} \ ,$$

where $\boldsymbol{\tau}$ is $\boldsymbol{\alpha}$ without $\boldsymbol{\beta}$. One problem by integrating out $\boldsymbol{\beta}$ is that they are no longer part of the objective function so essentially it does not contain information about these parameters anymore. The $\boldsymbol{\beta}$ parameters can, however, be retrieved as the estimators from the likelihood function (2) given the REML estimate of the variance parameters. This provides the point estimates of $\boldsymbol{\beta}$, but using the inverse Hessian matrix at the optimum as the covariance matrix leads to a covariance matrix that does not take account of the effect of σ_u. Luckily $\boldsymbol{\beta}$ and σ_u are (asymptotically?) uncorrelated in LMMs so this is not a serious deficit. The estimates of $\boldsymbol{\beta}$ can also be retrieved as the mode (which is identical to the mean in LMMs) of the conditional (aka. posterior) distribution, of $\boldsymbol{\beta}$ given the observations and estimates of $\boldsymbol{\tau}$; $p(\boldsymbol{\beta}|\boldsymbol{y})$. Observe that this density is completely specified given the observations and $\boldsymbol{\tau}$, hence there are no free parameters. Lastly, this estimator of $\boldsymbol{\beta}$ coincide with the classical ANOVA estimator.

All the things that work so amazingly well in LMMs generally fail for NLMMs and GLMMs. All attempts at proposing a REML-type estimator for NLMMs and GLMMs, of which the author is aware, are (or can be/have been) motivated, one way or another, by integrating $\boldsymbol{\beta}$ out of the likelihood function. ANOVA methods do not apply to GLMMs or NLMMs, so there is no direct link here. The REML-type estimators cannot be motivated by estimation in a residual space (not even for NLMMs where reasonable residuals can be defined in contrast to GLMMs) because this is based on the orthogonality of the residuals to $\boldsymbol{\beta}$. So while REML type estimators have been proposed for NLMMs and GLMMs, there is no general way to obtain inference for $\boldsymbol{\beta}$ and all attempts are clouded in an air of ad hoc'ary.

F Structural differences between GLMMs and NLMMs

The difference in the likelihood function between GLMMs and NLMMs is solely in the assumed density of the observations given the random effects, $p_{\boldsymbol{\alpha}}(y_{ij}|u_i)$. Suppose that this probability density function or probability mass function (density for short) is specified by a mean parameter, μ_{ij}, and that this is linked to a predictor η_{ij} by a link function, $g(\cdot)$ such that $\mu_{ij} = g(\eta_{ij})$ and $\eta_{ij} = \eta(\boldsymbol{x}_{ij}, \boldsymbol{\beta}, \sigma_u, u_i)$ is a function of fixed-effects regression parameters $\boldsymbol{\beta}$, regression variables $\boldsymbol{x}_{ij}$, the random effects, u_i and their variance, σ_u. In NLMMs $p_{\boldsymbol{\alpha}}(y_{ij}|u_i)$ is the normal density, $g(\cdot)$ is (usually) the identity function and η_{ij} is a nonlinear function of at least one of $\boldsymbol{\beta}$ and u_i. In GLMMs $p_{\boldsymbol{\alpha}}(y_{ij}|u_i)$ is any exponential-family density such as the Poisson, binomial or multinomial (in which case y_{ij} is a vector), $g(\cdot)$ is a nonlinear function of η_{ij}, usually chosen such that μ_{ij} takes on values in the appropriate range, e.g. between zero and one for binomial observations, and η is a linear function of $\boldsymbol{\beta}$ and the term $\sigma_u u_i$.

G Random effects structures

Models with more general random effects structures such as multivariate random effects, where $\boldsymbol{u}_i$ is a vector, or with multiple grouping structures can be written more generally if a matrix notation is adopted. Multivariate random effects are quite common in NLMMs where e.g. random effects for subjects appear for several fixed effects parameters describing, for example asymptotes, rates and half-time effects. In GLMMs, the most common multivariate random effect structure is the random coefficient structure where the subject-specific intercept and slope are correlated. In GLMMs it is also quite common to have multiple grouping structures, the classical situation are nested random terms, e.g. pupils in schools and schools in districts, but cross-classified, or simply "crossed" random effects are also common. This includes for instance migrating animals observed at various patches or territories, or as is common in item-response analysis, the assessment of items by respondents, or in an engineering setting where each of a number of machines are operated by each of a group of workers. In designed experiments the cross-classification can be complete, but often missing values destroy the structure. In observational studies the cross-classification is almost always incomplete. We denote the resulting structure *partially crossed random terms* following Doran et al. (2007). Even in classical nested cases, the nesting can be incomplete, for instance, if some pupils change school during the observational period. We denote this structure by *partial nesting*. From a computational perspective, however, there is no need for the distinction between partial nesting and partial crossing: the result is a lack of structure in the design matrix for the random effects. Several grouping structures does not seem as common in NLMMs.

References

Abramowitz, M. and I. A. Stegun (Eds.) (1972). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. New York: Dover Publications.

Anderson, D. A. and M. Aitkin (1985). Variance component models with binary response: Interviewer variability. *Journal of the Royal Statistical Society 47*(2), pp. 203–210.

Barndorff-Nielsen, O. and D. R. Cox (1979). Edgeworth and saddle-point approximations with statistical applications. *Journal of the Royal Statistical Society, Series B 41*(3), pp. 279–312.

Barndorff-Nielsen, O. E. and D. R. Cox (1989). *Asymptotic Techniques for Use in Statistics*. London: Chapman & Hall.

Bates, D. M. and S. DebRoy (2004). Linear mixed models and penalized least squares. *Journal of Multivariate Analysis 91*(1-17).

Bates, D. M. and D. G. Watts (1980). Relative curvature measures of nonlinearity. *Journal of the Royal Statistical Society, B 42*(1), pp. 1–25.

Booth, J. G. and J. P. Hobert (1999). Maximizing generalized linear mixed model likelihoods with an automated monte carlo em algorithm. *Journal of the Royal Statistical Society. Series B (Statistical Methodology) 61*(1), pp. 265–285.

Borjas, G. J. and G. T. Sueyoshi (1994). A two-stage estimator for probit models with structural group effects. *Journal of Econometrics 64*, pp. 165–182.

Breslow, N. (2003). Whither PQL. *UW Biostatistics Working Paper Series* (192), pp. i–xxiii.

Breslow, N. E. and D. G. Clayton (1993). Approximate Inference in Generalized Linear Mixed Models. *Journal of the American Statistical Association 88*(421), pp. 9–25.

Breslow, N. E. and X. Lin (1995). Bias correction in generalised linear mixed models with a single component of dispersion. *Biometrika 82*(1), pp. 81–91.

Chan, J. S. K. and A. Y. C. Kuk (1997). Maximum likelihood estimation for probit-linear mixed models with correlated random effects. *Biometrics 53*(1), pp. 86–97.

Demidenko, E. (2004). *Mixed Models, Theory and Applications*. Wiley.

Diggle, P. J., P. Heagerty, K.-Y. Liang, and S. L. Zeger (2002). *Analysis of Longitudinal Data* (2nd ed.). Oxford university Press.

Doran, H., D. Bates, P. Bliese, and M. Dowling (2007). Estimating the multilevel rasch model: With the lme4 package. *Journal of Statistical Software 20*(2), pp. 1–18.

Eldén, L., L. Wittmeyer-Koch, and H. B. Nielsen (2004). *Introduction to Numerical Computation — analysis and MATLAB illustrations*. Studentlitteratur.

Fahrmeir, L. and G. Tutz (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models* (Second ed.). Springer series in statistics. Springer-Verlag New York, Inc.

Fitzmaurice, G., M. Davidian, G. Verbeke, and G. Molenberghs (2009). *Longitudinal Data Analysis*. Chapman & Hall/CRC.

Gerstner, T. and M. Griebel (1998). Numerical integration using sparse grids. *Numerical Algorithms 18*(3), pp. 209–232.

Gilbert, P. (2009). *numDeriv: Accurate Numerical Derivatives*. R package version 2009.2-1.

Goldstein, H. (1986). Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika 73*(1), pp. 43–56.

Goldstein, H. (1989). Restricted unbiased iterative generalized least-squares estimation. *Biometrika 76*(3), pp. 622–3.

Goldstein, H. (1991). Nonlinear multilevel models, with an application to discrete response data. *Biometrika 78*(1), pp. 45–51.

Golub, G. H. (1973). Some modified matrix eigenvalue problems. *Siam Review 15*, pp. 318–334.

Golub, G. H. and J. H. Welsch (1969). Calculation of gaussian quadrature rules. *Mathematics of Computation 23*, pp. 221–230.

Hastie, T., R. Tibshirani, and J. Friedman (2001). *The Elements of Statistical Learning, Data Mining, Inference and Prediction*. Springer series in statistics. Springer, New York.

Hauck Jr., W. W. and A. Donner (1977). Wald's Test as Applied to Hypotheses in Logit Analysis. *Journal of the American Statistical Association 72*(360), pp. 851–853.

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

Hedeker, D. and R. D. Gibbons (1996). MIXOR: a computer program for mixed-effects ordinal regression analysis. *Computer methods and Programs in Biomedicine 49*, pp. 157–176.

Heiss, F. and V. Winschel (2008). Likelihood approximation by numerical integration on sparse grids. *Journal of Econometrics 144*, pp. 62–80.

Joe, H. (2008). Accuracy of laplace approximation for discrete response mixed models. *Comput. Stat. Data Anal. 52*(12), pp. 5066–5074.

Karim, M. R. and S. L. Zeger (1992). Generalized linear models with random effects; salamander mating revisited. *Biometrics 48*(2), pp. 631–644.

Lee, L. (2000). A numerically stable quadrature procedure for the one-factor random-component discrete choice model. *Journal of Econometrics 95*(1), pp. 117 – 129.

Lee, Y. and J. A. Nelder (1996). Hierarchical generalized linear models. *J. R. Statist. Soc. B 58*(4), pp. 619–678.

Lee, Y. and J. A. Nelder (2000). Two ways of modelling overdispersion in non-normal data. *Appl. Statist 49*, pp. 591–598.

Lee, Y. and J. A. Nelder (2003). Extended-reml estimators. *Appl Statist 30*(8), pp. 845–856.

Lee, Y. and J. A. Nelder (2004). Conditioan and marginal models: Another view. *Statistical Science 19*(2), pp. 219–238.

Lee, Y. and J. A. Nelder (2005). Likelihood for random-effect models. *SORT 29*(2), pp. 141–164.

Lee, Y. and J. A. Nelder (2006). Double hierarchical generalized linear models. *Appl. Statist. 55*, pp. 139–185.

Lee, Y., J. A. Nelder, and M. Noh (2007). H-likelihood: problems and solutions. *Stat Comput 17*, pp. 49–55.

Lee, Y., J. A. Nelder, and Y. Pawitan (2006). *Generalized Linear Models with Random Effects—Unified Analysis via H-likelihood*. Chapman & Hall/CRC.

Lesaffre, E. and B. Spiessens (2001). On the effect of the number of quadrature points in a logistic random-effects model: an example. *Applied Statistics 50*(3), pp. 325–335.

Liu, G. and D. A. Pierce (1993). Heterogeneity in mantel-haenszel-type models. *Biometrika 80*, pp. 543–556.

Liu, Q. and D. A. Pierce (1994). A note on gauss-hermite quadrature. *Biometrika 81*(3), pp. 624–629.

McCulloch, C. E. (1994). Maximum likelihood variance components estimation for binary data. *Journal of the American Statistical Association 89*(425), pp. 330–335.

McCulloch, C. E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association 92*(437), pp. 162–170.

McCulloch, C. E. and S. R. Searle (2001). *Generalized, Linear, and Mixed Models*. Wiley Series in Probability and Statistics. John Wiley & Sons, inc.

Millar, R. B. (2004). Simulated maximum likelihood applied to non-gaussian and nonlinear mixed effects and state-space models. *Australian & New Zealand Journal of Statistics 46*(4), pp. 543–554.

Naylor, J. C. and A. F. M. Smith (1982). Applications of a method for the efficient computation of posterior distributions. *Journal of the Royal Statistical Society. Series C (Applied Statistics) 31*(3), pp. 214–225.

Naylor, J. C. and A. F. M. Smith (1988). Econometric illustrations of novel numerical integration strategies for bayesian inference. *Journal of Econometrics 38*(1-2), pp. 103 – 125.

Nielsen, H. B. (2000). UCMINF - an Algorithm for Unconstrained, Nonlinear Optimization. Technical report, Informatics and Mathematical Modelling (IMM), Technical University of Denmark.

Nielsen, H. B. and S. B. Mortensen (2009). *ucminf: General-purpose unconstrained non-linear optimization*. R package version 1.0-5.

Nocedal, J. and S. J. Wright (2006). *Numerical optimization* (2nd. ed.). Springer.

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

Pawitan, Y. (2000). A reminder of the fallability of the Wald statistic: Likelihood explanation. *The American Statistician 54*(1), pp. 54–56.

Pawitan, Y. (2001). Two-staged estimation of variance components in generalized linear mixed models. *J. Statist. Comput. Simul. 69*, pp. 1–17.

Pinheiro, J. C. and D. M. Bates (1995). Approximations to the nonlinear mixed-effects model. *Jounal of Computational and Graphical Statistics 4*(1), pp. 12–35.

Pinheiro, J. C. and D. M. Bates (2000). *Mixed-Effects Models in S and S-PLUS*. Springer.

Pinheiro, J. C. and E. C. Chao (2006). Efficient laplacian and adaptive gaussian quadrature algorithms for multilevel generalized linear mixed models. *Journal of Computational and Graphical Statistics 15*(1), pp. 58–81.

Rabe-Hesketh, S., A. Skrondal, and A. Pickles (2005). Maximum likelihood estimation of limited and discrete dependent variable models with nested random effects. *Journal of Econometrics 128*, pp. 301–323.

Raudenbush, S. W., M.-L. Yang, and M. Yosef (2000). Maximum Likelihood for Generalized Linear Models with Nested Random Effects via High-Order, Multivariate Laplace Approximation. *Journal of Computational and Graphical Statistics 9*(1), pp. 141–157.

Richardson, L. F. (1910). The approximate arithmetical solution by finite differences of physical problems involving differential equations, with an application to the stressesin a masonry dam. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character 210*, pp. 307–357.

Richardson, L. F. and J. A. Gaunt (1927). The deferred approach to the limit. part i. single lattice. part ii. interpenetrating lattices. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character 226*, pp. 299–361.

Schall, R. (1991). Estimation in Generalized Linear Models with Random Effects. *Biometrika 78*(4), pp. 719–727.

Shun, Z. (1997). Another look at the salamander mating data: A modified laplace approximation approach. *Journal of the American Statistical Association 92*(437), pp. 341–349.

Shun, Z. and P. McCullagh (1995). Laplace approximation of high dimensional integrals. *Journal of the Royal Statistical Society. Series B (Methodological) 57*(4), pp. 749–760.

Skaug, H. J. and D. A. Fournier (2006). Automatic approximation of the marginal likelihood in non-gaussian hierarchical models. *Computational Statistics & Data Analysis 51*(2), pp. 699 – 709.

Skrondal, A. and S. Rabe-Hesketh (2004). *Generalized Latent Variable Modeling*. Chapman & Hall/CRC.

Smolyak, S. A. (1963). Quadrature and interpolation formulas for tensor products of certain classes of functions. *Dokl. Akad. Nauk SSSR 4*, pp. 240–243.

Stroud, A. H. (1971). *Approximate Calculation of Multiple Integrals*. Englewood Cliffs, NJ: Prentice-Hall.

Tierney, L. and J. B. Kadane (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association 81*(393), pp. 82–86.

Wolfinger, R. (1993). Laplace's approximation for nonlinear mixed models. *Biometrika 80*(4), pp. 791–795.

Wolfinger, R. D. and X. Lin (1997). Two taylor-series approximation methods for nonlinear mixed models. *Computational Statistics & Data Analysis 25*, pp. 465–490.

Zeger, S. L. and M. R. Karim (1991). Generalized linear models with random effects; a gibbs sampling approach. *Journal of the American Statistical Association 86*(413), pp. 79–86.

Zucchini, W. and I. L. MacDonald (2009). *Hidden Markov Models for Time Series, An introduction Using R*. Chapman & Hall/CRC.