Gaussian Variational Approximation for Overdispersed Generalized Linear Mixed Models

Aklilu H. Ghebretinsae[*] Christel Faes [†] Geert Molenberghs [‡]

Abstract

In a recent publication by Molenberghs and Demétrio (2011) a general modeling framework was proposed to model non-Gaussian data that are hierarchically structured and are overdispersed in the sense that the distributional mean-variance relationship is not fulfilled. The modeling framework extends the Generalized Linear Models with two random effects, one normally distributed random effect to accommodate the correlation in the data due to the hierarchy and one conjugate random effect to account for the overdispersion. The main difficulty with this kind of models is the computational complex estimation due to the intractable multivariate integrals, as is the case for Generalized Linear Mixed Models that involves such integrals with no analytic expression. Different estimation methods for these models were already proposed: estimation using partial marginalization, estimation in the bayesian framework, and an approximate estimation based on pseudo-likelihood. In this manuscript, we will investigate the use of Gaussian variational approximation methods as a computationally fast estimation method for the combined model. A range of over-dispersed non-gaussian mixed models are investigated.

Key Words: Gaussian variational approximation, Gamma Frailty, Weibull normal, Poisson-Normal, Hierarchical model, Random effect, Weibull model

1. Introduction

Generalized linear models are the most common class of regression models used to analyze different types of variables, including binary, counts and continuous outcomes (Nelder and Wedderburn 1972, McCullagh and Nelder 1989, Agresti 2002). The exponential family distribution provides an elegant specification of the models. The most well-known examples include linear regression, logistic regression and Poisson regression. An important extension of these models are the generalized linear mixed model by the inclusion of a normally distributed random effect, allowing to account for a multilevel structure in the data (Molenbeghs and Verbeke 2005)). A common issue with non-Gaussian data is overdispersion in the sense that the variability in the data is not well described by the distributional mean-variance relationship (Hinde and Demétrio 1998). This can happen both in the univariate or in the multi-level setting. One approach to account for overdispersion in a univariate generalized linear mixed model is by the use of a conjugate random effect, such as, for example the negative binomial (Breslow 1984, Lawless 1978) and beta-binomial model (Skellam 1948, Kleinman 1973). Molenberghs and Demétrio (2011) proposed a similar approach to account for overdisperion in a multilevel setting, by the use of two random effect, one normally distributed random effect to accommodate for the hierarchy and one strongly conjugate random effect to account for the overdispersion in the data. This introduces a new general modeling framework for the analysis of overdispersed multilevel data, and is often referred to as the *combined model* (Molenberghs et al 2007, Molenberghs et al 2010).

A difficulty in inference of these models is often encountered in both the bayesian and likelihood framework, due to the intractable multivariate integrals in the likelihood

[*] I-BioStat, Center for Statistics, Universiteit Hasselt, B-3590 Diepenbeek, Belgium

[†] I-BioStat, Center for Statistics, Universiteit Hasselt, B-3590 Diepenbeek, Belgium

[‡] I-BioStat, Center for Statistics, Universiteit Hasselt, B-3590 Diepenbeek, Belgium and I-BioStat, Katholieke Universiteit Leuven, B-3000 Leuven, Belgium

and posterior densities. This can already be problematic for the generalized linear mixed models (GLMM), because of the integrals in the marginalized likelihood with no analytic expression which need a numerical approximations. Often it is dealt with by numerical integration using adaptive and gaussian adaptive quadrature, series expansion methods including penalized quasi-likelihood and marginal quasi-likelihood, laplace approximation, etc. Different estimation techniques have been employed in modeling the combined model, commonly it is done through partial marginalization in which the conjugate random effect are first integrated out leaving the normal effects untouched and then obtain the fully marginal by numerical integration of the normal random effect using adaptive gaussian quadrature in standard software such as the SAS procedure NLMIXED (Molenberghs et al 2007, Molenberghs et al 2010), in the bayesian framework using MCMC (Ghebretinsae et al 2011) and pseudolikelihood estimation (Achmad et al 2011).

In this paper, another estimation method for the combined model is proposed, providing a fast estimation method as an alternative to the existing methods. Ormerod and Wand (2012) recently introduced variational approximation in the statistical modeling framework. Variational Inference have their roots in the statistical physics and are used to approximating intractable computations (Blei and Jordan 2006). The key idea is to introduce a set of approximating densities to the posteriors and to introduce them in such a way as to make their evaluation tractable. These approximations are then optimized so as to minimize the discrepancy between the approximation and the true posterior using some measure of the difference. The optimization is carried out by varying the functional parameters of these approximations, thus giving the approximation its name. While different variational approximations exist, we focus on Gaussian variational approximations, in which the conditional distribution of the random effects given the data are approximated by Gaussian distributions. Hall, Ormerod and Wand (2011) studied the properties of Gaussian variational approximations in the setting of generalized linear mixed models.

The general idea of variational approximation is to approximate the likelihood so that the integral problem is either fully or partially solved. It is therefore basically an approximation of the integrand. When the integral problem is fully solved, it results in optimization of the resulting approximate likelihood. When the integral problem is not completely eradicated, like e.g. in a binary GLMM as will be shown later, it is still useful in reducing the dimension of the integral to one. But approximation of the integral of the new likelihood/integrand is still required using adaptive gaussian quadrature.

The paper is organized as follows. In section 2, three datasets to illustrate the proposed methodology are introduced, namely the comet data, epilepsy data and EG data. The combined model for non-Gaussian data is introduced in Section 3. The Gaussian variational approximation estimation technique is reviewed in Section 4. Their properties are investigated via three examples, including an extended random-effects Weibull, Poisson and logistic model in Section 5, 6 and 7.

2. Motivating Examples

In this section, motivating datasets used in this manuscript are presented.

2.1 Epileptic Data

The first data considered here is obtained from a randomized, double-blind, parallel group multi-center study for the comparison of placebo with anti-epileptic drug (AED), in combination with one or two other AED's. The study is described in full detail in Faught et al (1996) and it is used in Molenberghs et al (2007). The randomization of epilepsy patients took place after a 12-week baseline period that served as a stabilization period for the use of AED's, and during which the number of seizures were counted. After that period, 45 patients were assigned to the placebo group, 44 to the active (new) treatment group. Patients were then measured on a weekly basis during 16 weeks, after which they were entered into a long term open extension study. Some patients were followed for up to 27 weeks. The outcome of interest is the number of epileptic seizures experienced during the last week, i.e, since the last time the outcome was measured. The key research question is whether or not the additional new treatment reduces the number of epileptic seizures. As a summary of the data, the subject-specific profiles for 12 randomly selected individuals in each treatment arm is presented in the upper panel of Figure 1. It is a skewed distribution with a largest observed value equal to 73 seizures in one week time.

2.2 Comet Assay Data

The second data resulted from a comet assay study and have been studied in Ghebretinsae et al (2011). The comet assay is a single cell microgel electrophoresis method detecting DNA damage in any target tissue or organ of which a single cell suspension can be prepared. Exposure to high alkali (pH $>$ 13.0) allows expression of single strand breaks and subsequent alkaline electrophoresis ensures migration of DNA fragments out of the nucleus. Visualization of this DNA migration (typical comet-like structures) is performed by a fluorescent dye. An image analysis system coupled to a microscope permits quantification of DNA damage at the single cell level. Here, the data refer to four groups of six male rats that received a daily oral dose of a compound in three dose levels (low, medium, and high) or vehicle control. On the day of necropsy, an extra group of three animals received a single dose of a positive control (200 mg/kg ethyl methanesulfonate, EMS, PC). The animals were sacrificed 3 hours after the last dose administration, their liver was removed and processed for the comet assay. For each animal, a cell suspension is prepared. From each cell suspension, three replicate samples were prepared for scoring. Fifty randomly selected, non-overlapping cells per sample were then scored for DNA damage using a semi-automated scoring system. A total of 150 liver cells were thus scored per animal. DNA damage was assessed by the software system by measuring tail migration, % tail intensity, and tail moment. The interest here is to see the toxicity of 1,2-Dimethylhydrazine dihydrochloride at the different dose levels (low, medium, and high) based on tail length. Generally, the toxicity level increase with the dose level. A summary of the data is presented in the lower left panel of Figure 1. We observe some extreme values at all dose levels.

2.3 Ethylene Glycol (EG) Data

The third dataset is from toxicology study of Ethylene glycol. Ethylene glycol (EG) also called 1,2-ethanediol is a high-volume industrial chemical with diverse applications. It is used to make antifreeze and de-icing solutions for cars, airplanes and boats, to make polyster compounds, and is used as a solvent in the paint and plastic industries. It is also used as an ingredient in photographic developing solutions, hydraulic brake fuids and in the formulation of several types of inks and many more. While EG may not be hazardous to humans in normal industrial handling, it can become dangerous when used at elevated temperatures or when ingested. Exposure to large amounts of ethylene glycol can damage the kidneys, heart, and nervous system. In addition, ingestion of antifreeze products, which consist for approximately 95 % of EG, is toxic and may result in death. The data resulted from a study in which timed-pregnant CD-1 mice were dosed by gavage with EG in distilled water as described by Price et al. (1985). Dosing occurred during the period of major organogenesis and structural development of the foetuses (gestational days 6 through 15). The doses were at 0, 750, 1500 or 3000 mg/kg/day, with 25, 24, 23 and 23 timed-pregnant mice randomly assigned to each of these dose groups, respectively. The interest here is to assess the toxicity of this chemical at the different dose levels based on a binary outcome, whether the foetus is malformed or not. Summary of the data is presented in the lower right panel of Figure 1. We observe a general trend of increasing toxicity with dose level.

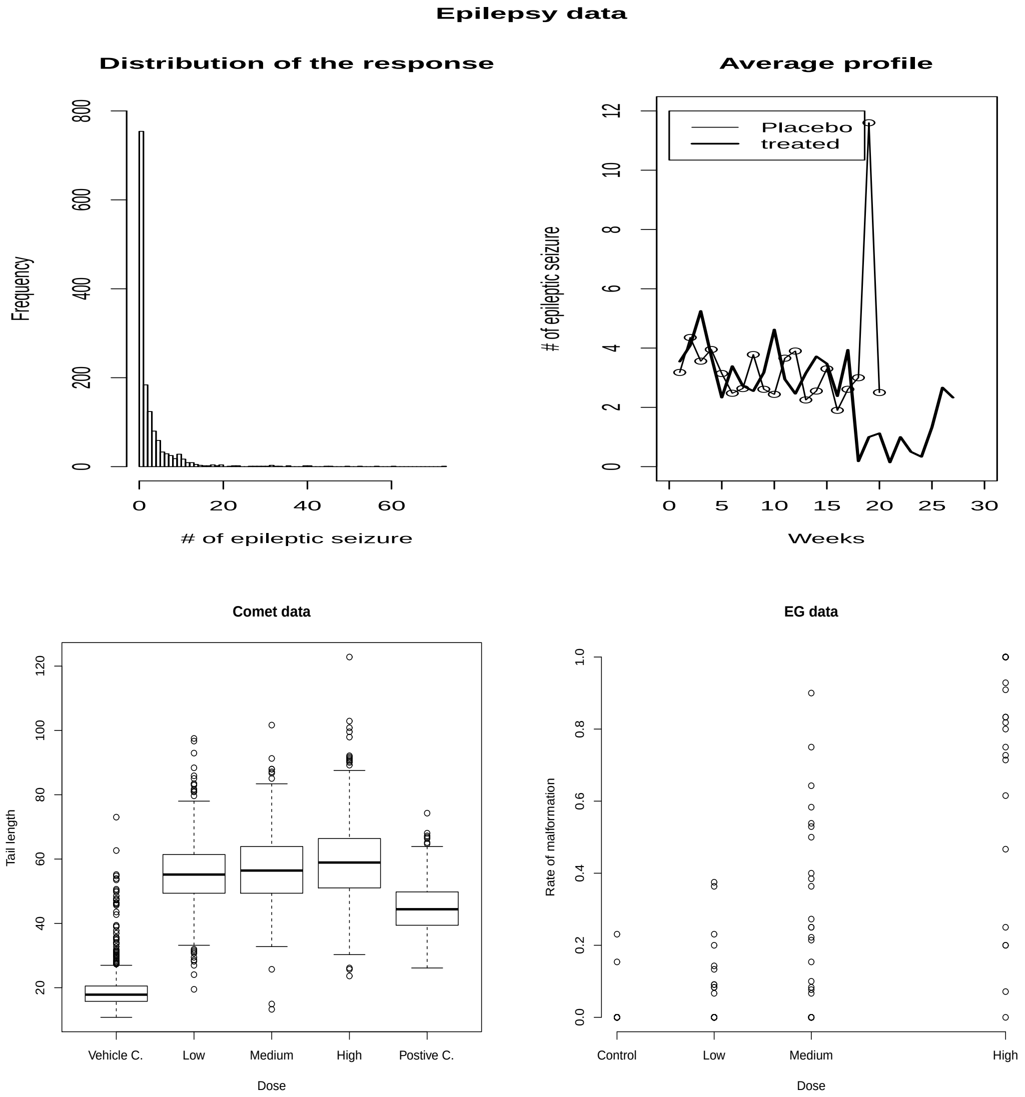

Figure 1: 1. Top: The distribution of the response and the average profile for the two treatment groups over time for Epilepsy data 2.Bottom left: box plot for comet data 3. Bottom right: scatter plot for EG data.

3. Combined Model

Molenberghs *et al* (2010) proposed an exponential family model with two random effect to accommodate simultaneously the clustering and overdispersion effects. It extends the generalized mixed model by the use of conjugate random effect for overdispersion. The general model family proposed for modeling overdispersed and correlated data is given by:

$$f_i(y_{ij}|b_i, \theta_{ij}, \xi) = \exp\left\{\phi^{-1}[y_{ij}\lambda_{ij} - \psi(\lambda_{ij})] + c(y_{ij}, \phi)\right\}, \tag{1}$$

for outcome y_{ij} on subject $i = 1, \ldots, N$ at occasion $j = 1, \ldots, n_i$. The unknown parameters λ_{ij} and ϕ are often termed natural parameter and scale parameter, respectively. The term $c(y_{ij}, \phi)$ is the normalizing constant. The function $\psi(\cdot)$ is a known function with the property that $E[y_{ij}|b_i, \theta_{ij}, \xi] = \psi'(\lambda_{ij})$ and var$(y_{ij}|b_i, \theta_{ij}, \xi) = \phi\psi''(\lambda_{ij})$. Model specification proceeds by assuming that the conditional mean of y_{ij} is given by

$$E[y_{ij}|\theta_{ij}, b_i] = \mu^c_{ij} = \theta_{ij}\kappa_{ij}, \tag{2}$$

where $\theta_{ij} \sim \mathcal{G}_{ij}(\xi_{ij}, \sigma^2_{ij})$ for some distribution $\mathcal{G}_{ij}$ with mean ξ_{ij} and variance σ^2_{ij} and $\kappa_{ij} = g(\eta_{ij}) = g(x'_{ij}\xi + z'_{ij}b_i)$ for some link function g and $b_i \sim N(0, D)$. The random variable θ_{ij} is used to account for the overdispersion in the data, while the random effect in κ_{ij} accounts for the clustered or hierarchical structure of the data. The two parameters η_{ij} and λ_{ij} refer to the linear predictor and/or the natural parameter. The basic difference is that λ_{ij} encompasses the random variables θ_{ij}, whereas η_{ij} refers to the 'GLMM part' only.

Most often, but not strictly necessary, it is assumed that the two sets of random effects, $\boldsymbol{\theta}_i$ and b_i, are independent of each other (see Molenberghs *et al* (2010) for further discussion).

Parameterization (2) such that the random effects θ_{ij} capture overdispersion, and are formulated directly at the mean scale, whereas κ_{ij} can be considered as the generalized linear mixed model component. The relationship between mean and natural parameter now is

$$\lambda_{ij} = h(\mu^c_{ij}) = h(\theta_{ij}\kappa_{ij}). \tag{3}$$

Standard GLM ideas can be applied to derive the mean and variance, combined with iterated-expectation-based calculations. For the mean, it follows that

$$E[y_{ij}] = E[\theta_{ij}]E[\kappa_{ij}] = E[h^{-1}(\lambda_{ij})]. \tag{4}$$

An important concept in regard to computational difficulty/efficiency is *conjugacy*, in the sense of Cox and Hinkley (1974, p. 370) and Lee, Nelder, and Pawitan (2006, p. 178). Conjugacy refers to the fact that the hierarchical and random-effects densities have similar algebraic forms. Conjugate distributions produce a general and closed-form solution for the corresponding marginal distribution. Molenberghs *et al* (2010) adapted conjugacy to the situation where both normal and overdispersion random effects are included. For detailed explanation on the combined model we refer to Molenberghs *et al* (2010).

In this case three types of outcomes are considered: time to event, count and binary. A Weibull model will be considered for the time to event outcome, a Poisson model for the count and a Logistic model for the binary outcomes.

4. Gaussian Variational approximation using density transformation

This section reviews the GVA approximation method using density transformation, as described by Ormerod and Wand (2012) and Hall, Ormerod and Wand (2011).

4.1 Deriving GVA lower bound likelihood

Ormerod and Wand (2012) considered the generalized mixed model of the form:

$$f(y_{ij}|b_i) \quad = \quad \exp\{y_{ij}\lambda_{ij} - \psi(\lambda_{ij}) + c(y_{ij})\}$$

with,

$$\begin{array}{c} \lambda_{ij} = \eta_{ij} = \boldsymbol{x}_{ij}{}'\boldsymbol{\xi} + \boldsymbol{z}_{ij}{}'b_i \\ \text{and} \qquad b_i \sim N(0, D) \end{array}$$

It is a reduced form of (1) in that the θ_{ij} is omitted here and the scale parameter ϕ fixed to one. Let $y_i = (y_{i1}, \ldots, y_{in_i})'$; $\boldsymbol{x}_i = (\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{in_i})'$; $\boldsymbol{\xi} = (\beta_1, \ldots, \beta_p)'$; $\boldsymbol{z}_i = (\boldsymbol{z}_{i1}, \ldots, \boldsymbol{z}_{in_i})'$ and q be the dimension of the random effect b_i.The corresponding marginal likelihood is,

$$\begin{aligned} l(\boldsymbol{\xi}, D) \quad = \quad & \sum_{i=1}^{N}\{y_i{}'\boldsymbol{x}_i'\boldsymbol{\xi} + 1_i'c(y_i)\} - \frac{N}{2}\log|D| - \frac{Nq}{2}\log(2\pi) \\ & + \sum_{i=1}^{N}\log\int \exp\{y_i{}'\boldsymbol{z}_i'b_i - \psi(\boldsymbol{x}_i'\boldsymbol{\xi} + \boldsymbol{z}_i'b_i) - \frac{1}{2}b_i{}'D^{-1}b_i\}db_i \qquad (5) \end{aligned}$$

The maximum likelihood estimates of the fixed effects $\boldsymbol{\xi}$ and covariance matrix D of the GLMM are obtained by maximizing (5). The problem in maximizing this likelihood is the presence of N integrals over the q-dimensional random effects b_i. Gaussian variational approximation method tackles this issue by introducing an extra pair of variational parameters $\boldsymbol{\mu}_i, \boldsymbol{\Lambda_i}$ for each subject i, $1 \leq i \leq N$, where $\boldsymbol{\mu}_i$ is a q-dimensional vector and $\boldsymbol{\Lambda_i}$ is $q \times q$ positive definite matrix. And new density functions $q(b_i)$ are introduced and are assumed to be Multivariate Gaussian density distribution with mean μ_i and covariance matrix Λ_i. In principle these densities can take any functional form. In this case, the marginalized likelihood can be re-written in terms of the $q(b_i)$ densities:

$$\begin{aligned} l(\boldsymbol{\xi}, D) \quad &= \quad \log\int p(y|b_i)p(b_i)db_i, \\ &= \quad \log\int p(y|b_i)p(b_i)\frac{q(b_i)}{q(b_i)}db_i, \\ &= \quad \log E_{b_i \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Lambda_i})}\left[\frac{p(y|b_i)p(b_i)}{q(b_i)}\right] \end{aligned}$$

In this expression, $E_{b_i \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Lambda_i})}(.)$ is the expected value with respect to $b_i \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Lambda_i})$.

By Jensen's inequality and concavity of the logarithm function, we then have

$$l(\boldsymbol{\xi}, D) \quad \geq \quad E_{b_i \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Lambda_i})}\left[\log\left(\frac{p(y|b_i)p(b_i)}{q(b_i)}\right)\right] = \underline{l}(\boldsymbol{\xi}, D, \boldsymbol{\mu}, \boldsymbol{\Lambda})$$

where $\underline{l}(\boldsymbol{\xi}, D, \boldsymbol{\mu}, \boldsymbol{\Lambda})$ is a lower boundary of the loglikelihood function $l(\boldsymbol{\xi}, D)$, the approximated likelihood. Alternatively the same inequality can be seen from the Kullback Leibler divergence point of view (Ormerod and Wand, 2012; Hall, Ormerod and Wand,2011). The accuracy of the approximate likelihood depends on the distance between $q(b_i)$ and $p(b_i|y)$, measured by the Kullback Leibler distance.

The idea of GVA is to approximate the posterior distribution $p(b_i|y)$ which contains integral difficulty by $q(b_i)$ in such a way that the likelihoood/integrand is integrable or easier. The integral problem is not always completely removed after applying GVA. In some cases the integral problem exists partially; however, it may still have computational advantage although numerical approximation is required. We will see this in detail in the next section.

The variational lower bound of the log-likelihood simplifies to,

$$\begin{aligned}\underline{l}(\boldsymbol{\xi}, D, \boldsymbol{\mu}, \boldsymbol{\Lambda}) &= \frac{Nq}{2} - \frac{N}{2}\log|D| + \sum_{i=1}^{N}\{y_i'(\boldsymbol{x}_i'\boldsymbol{\xi} + \boldsymbol{z}_i\mu_i) + 1_i'c(y_i) - \\ &\quad 1_i'T(\boldsymbol{x}_i'\boldsymbol{\xi} + \boldsymbol{z}_i\mu_i, \mathrm{diag}(\boldsymbol{z}_{\mathrm{i}}'\Lambda_{\mathrm{i}}\boldsymbol{z}_{\mathrm{i}})) - \frac{1}{2}(\log|\Lambda_{\mathrm{i}}| - \mu_{\mathrm{i}}'\mathrm{D}^{-1}\mu_{\mathrm{i}} - \mathrm{tr}(\mathrm{D}^{-1}\Lambda_{\mathrm{i}})\}\end{aligned}$$

where

$$T(\mu, \sigma^2) = \int \psi(\mu + \sigma x)\phi(x)$$

and ϕ is the standard normal with mean 0 and variance 1.

The variational lower bound contains the original parameter $(\boldsymbol{\xi}, D)$ and additional variational parameters $(\boldsymbol{\mu}, \boldsymbol{\Lambda})$. The ML estimates are the $(\boldsymbol{\xi}, D)$ parameters obtained by maximizing the new lower likelihood, $\underline{l}(\boldsymbol{\xi}, D, \boldsymbol{\mu}, \boldsymbol{\Lambda})$.

5. GVA for General Frailty Models

In this section, four hierarchical models are presented for time to event data:

- Weibull-Gamma frailty model;
- Weibull-Normal random intercept model;
- Weibull-Normal-Normal random intercepts model; and
- Weibull-Gamma-Normal hierarchical model.

For each of these models, a gaussian variational approximation is derived.

5.1 Weibull-Gamma frailty model

Let us consider the Weibull-Gamma model. It is well known that Weibull and Gamma distributions are conjugate. This property simplifies the computations, because the gamma frailty can be integrated out to obtain the marginal likelihood. The conditional likelihood and frailty densities are give by the following expression:

$$f(y_i|\theta_{ij}) = \prod_{j=1}^{n_i} \lambda\rho\theta_{ij}y_{ij}^{\rho-1}e^{\boldsymbol{x}_{ij}'\boldsymbol{\xi}}e^{-\lambda y_{ij}^{\rho}\theta_{ij}e^{\boldsymbol{x}_{ij}'\boldsymbol{\xi}}} \tag{6}$$

$$f(\theta_{ij}) = \frac{1}{\left(\frac{1}{\alpha}\right)^{\alpha}\Gamma(\alpha)}\theta_{ij}^{\alpha-1}e^{-\alpha\theta_{ij}} \tag{7}$$

Where (6)corresponds with a Weibull$(\lambda\theta_{ij}e^{\boldsymbol{x}_{ij}'\boldsymbol{\xi}}, \rho)$ distribution, and (7) with a Gamma$(\alpha, \frac{1}{\alpha})$ distribution. The marginal likelihood $l(\boldsymbol{\xi}, \alpha)$ is obtained by integrating the gamma random effect.

$$l(\boldsymbol{\xi}, \alpha) = \log f(y) = \log\int f(y|\boldsymbol{\theta})f(\boldsymbol{\theta})d\boldsymbol{\theta},$$

As a result, the marginal density can be expressed as:

$$\begin{aligned}l(\boldsymbol{\xi}, \alpha) &= \log p(y) = \log(\lambda\rho) + (\rho - 1)\log(y_{ij}) + \boldsymbol{x}_{ij}'\boldsymbol{\xi} \\ &\quad +(\alpha + 1)\log(\alpha) - (\alpha + 1)\log(\alpha + \lambda\rho y_{ij}^{\rho}e^{{}^{x}\boldsymbol{x}_{ij}'\boldsymbol{\xi}}).\end{aligned}$$

The required parameter of interest are obtained by maximizing this function.

5.2 Weibull-Normal Random Intercept Model

If however a normal random effect is used, instead of the conjugate gamma frailty, which we call the Weibull-Normal model.

$$\begin{aligned} f(y_{ij}|b_i) &\sim Weibull(\lambda e^{\eta_{ij}}, \rho) \\ \eta_{ij} &= \boldsymbol{x}_{ij}{}'\boldsymbol{\xi} + b_i \\ f(b_i) &\sim Normal(0, d) \end{aligned}$$

the marginal likelihood becomes:

$$\begin{aligned} L(\boldsymbol{\xi}, d) &= f(y) = \int \prod_{i=1}^{N} f(y_i|b_i) f(b_i) db_i, \\ l(\boldsymbol{\xi}, d) &= \log \int \prod_{i=1}^{N} \prod_{j=1}^{n_i} \lambda \rho y_{ij}^{\rho-1} e^{\boldsymbol{x}_{ij}{}'\boldsymbol{\xi} + b_i} e^{-\lambda y_{ij}^{\rho} e^{\boldsymbol{x}_{ij}{}'\boldsymbol{\xi} + b_i}} f(b_i) db_i. \end{aligned}$$

The N integrals of the b_i random effects do not have a tractable solution. Applying Gaussian variational approximation technique , lead to the Gaussian variational approximate likelihood $\underline{l}(\boldsymbol{\xi}, d, \boldsymbol{\mu}, \boldsymbol{\Lambda})$,

$$\begin{aligned} \underline{l}(\boldsymbol{\xi}, d, \mu, \Lambda) &= \sum_{i=1}^{N} \sum_{j=1}^{n_i} \Big[\log(\lambda) + \log(\rho) + (\rho - 1)\log(y_{ij}) + (\boldsymbol{x}_{ij}{}'\boldsymbol{\xi} + \mu_i) \\ & \quad -\lambda y_{ij}^{\rho} e^{\boldsymbol{x}_{ij}{}'\boldsymbol{\xi} + \mu_i + \frac{1}{2}\Lambda_i}\Big] - \frac{N}{2}\log(d) - \sum_{i=1}^{N} \frac{1}{2d}(\mu_i^2 + \Lambda_i) + \sum_{i=1}^{N} \frac{1}{2}\log(\Lambda_i). \end{aligned}$$

5.3 Weibull-Normal-Normal Random Intercepts Model

So far we considered just one hierarchical random effect. Extending to two or more hierarchical random effect, e.g. Weibull-Normal-Normal, is straight forward. Given the (two) random effects are independent, GVA approximation is applied to each random effect separately.

$$\begin{aligned} l(\beta, d_1, d_2, \mu, \Lambda) &= \sum_{i=1}^{N} \sum_{j=1}^{M} \sum_{k=1}^{n_{ij}} \Big[\log(\lambda) + \log(\rho) + (\rho - 1)\log(y_{ijk}) + \\ & \quad (\boldsymbol{x}'_{ijk}\boldsymbol{\xi} + \mu_i + \mu_{ij}) - \lambda y_{ijk}^{\rho} e^{\boldsymbol{x}'_{ijk}\boldsymbol{\xi} + \mu_i + \mu_{ij} + \frac{1}{2}\Lambda_i + \frac{1}{2}\Lambda_{ij}}\Big] - \\ & \quad \frac{N}{2}\log(d_1) - \sum_{i=1}^{N} \frac{1}{2d_1}(\mu_i^2 + \Lambda_i) + \sum_{i=1}^{N} \frac{1}{2}\log(\Lambda_i) - \frac{NM}{2}\log(d_2) \\ & \quad -\sum_{i=1}^{N} \sum_{j=1}^{M} \frac{1}{2d_2}(\mu_{ij}^2 + \Lambda_{ij}) + \sum_{i=1}^{N} \sum_{j=1}^{M} \frac{1}{2}\log(\Lambda_{ij}). \end{aligned}$$

5.4 Weibull-Gamma-Normal Hierarchical Model

Now let us extend to Weibull-Gamma-Normal. Omitting the gamma random effect lead to the Weibull-Normal model and excluding the normal random effect leads to the Weibull-Gamma (6) model. This model can be expressed as:

$$f(y_{ij}|\theta_{ij}, b_i) = \lambda \rho \theta_{ij} y_{ij}^{\rho-1} e^{\eta_{ij}} e^{-\lambda y_{ij}^{\rho} \theta_{ij} e^{\eta_{ij}}},$$

$$\begin{aligned}
\eta_{ij} &= \boldsymbol{x}_{ij}{}'\boldsymbol{\xi} + b_i, \\
f(\theta_{ij}) &= \frac{1}{\left(\frac{1}{\alpha}\right)^{\alpha}\Gamma(\alpha)}\theta_{ij}^{\alpha-1}e^{-\alpha\theta_{ij}}, \\
f(b_i) &= \frac{1}{(2\pi)^{1/2}|d|^{1/2}}e^{-\frac{1}{2d}b_i^2}
\end{aligned}$$

In this model, b_i is the subject-specific normal random effects to account for the clustering of observations and θ_{ij} is the gamma random effects to accommodate for overdispersion.

Here, we have one additional gamma random effect, θ_{ij}, in contrast to Weibull-Normal model. One option is to approximate the posterior density of the gamma random effect and similar to the normal random effect in such a way that the integral problem is totally eradicated. Approximation of the posterior density of Gamma random effects θ_{ij} by Gamma and lognormal distribution was attempted in which the integral problem was solved however it did not lead to good approximation. The alternative way is to first integrate the gamma random effect; Hence, the Gamma frailty with normal random effect embedded in it and then apply Gaussian variational approximation (GVA) for the normal random effect. After integrating the gamma random effect θ_{ij}, we have:

$$f(y_{ij}|b_i) = \frac{\lambda\rho y_{ij}^{\rho-1}e^{\boldsymbol{x}_{ij}{}'\boldsymbol{\xi}}\alpha^{\alpha+1}}{(\alpha+\lambda\rho y_{ij}^{\rho}e^{\boldsymbol{x}_{ij}{}'\boldsymbol{\xi}})^{\alpha+1}}.$$

Applying GVA leads to,

$$\begin{aligned}
l(\beta, d, \boldsymbol{\mu}, \boldsymbol{\Lambda}) &= \int \log\left(\frac{p(y, b_i)}{q(b_i)}\right) q(b_i)db_i \\
&= \sum_{i=1}^{N}\sum_{j=1}^{n_i}[\log(\lambda) + \log(\rho) + (\rho-1)\log(y_{ij}) + (\boldsymbol{x}_{ij}{}'\boldsymbol{\xi} + \mu_i) + \\
&\quad (\alpha+1)\log(\alpha)] - \frac{N}{2}\log(d) - \sum_{i=1}^{N}\frac{1}{2d}(\mu_i^2 + \Lambda_i) + \\
&\quad \sum_{i=1}^{N}\frac{1}{2}\log(\Lambda_i) - (\alpha+1)\sum_{i=1}^{N}\int\log(\alpha + \lambda\rho e^{\boldsymbol{x}_{ij}{}'\boldsymbol{\xi}+b_i})q(b_i)db_i
\end{aligned}$$

5.5 Illustration on Comet data

This method was applied to the comet assay data. Three models namely Weibull-Normal, Weibull-Normal-Normal and Weibull-Gamma-Normal models were considered. For the Weibull-Normal, the model is given as:

$$\eta_{ijk} = \beta_0 + \beta_1 L_{ijk} + \beta_2 M_{ijk} + \beta_3 H_{ijk} + \beta_4 PC_i + b_i + b_{ij},$$

with $b_i \sim N(0, d_1)$ and $b_{ij} \sim N(0, d_2)$. L_{ijk}, M_{ijk}, H_{ijk} and PC_{ijk} are the indicators for low dose, medium dose, high dose, and positive control groups respectively. The random intercept b_i corresponds to the rat-specific effect whereas b_{ij} corresponds to the slide-specific effect j of rat i.

Estimation was done using both GVA approximation and numerical approximation using adaptive gaussian approximation in PROC NLMIXED taken as exact/ golden standard estimate (named as exact estimate). Standard softwares like SAS PROC NLMIXED do not allow for more than one hierarchical random effect. For a specific case of single dimensional random effect at each hierarchical level, it can be modeled by using some trick.

Table 1: Exact and GVA estimation for Weibull-Normal and Weibull-Gamma-Normal Models

		Weibull Normal	
		Exact	GVA
Effect	Par.	Estimate(s.e.)	Estimate(s.e.)
Veh.	β_0	-13.7574(0.2270)	-13.7575(0.2270)
Low *vs.* veh.	β_1	-3.8319(0.2180)	-3.8316(0.2179)
Med.*vs.* veh.	β_2	-3.9243(0.2181)	-3.9241(0.2180)
High *vs.* veh.	β_3	-4.1268(0.2185)	-4.1266(0.2183)
Pos.C.*vs.* veh.	β_4	-2.9399(0.2653)	-2.9402(0.2652)
Weib. shape	ρ	4.3293(0.0477)	4.3293(0.0477)
Var.	d	0.1334(0.0384)	0.1333(0.0384)
-2l		30476	30502.6
Duration		70 sec.	7 sec.
		Weibull Gamma Normal	
		Exact	GVA
Effect	Par.	Estimate(s.e.)	Estimate(s.e.)
Veh.	β_0	-30.9295(0.7264)	-30.92917(0.7264)
Low	β_1	-42.8673(1.0005)	-42.86688(1.0004)
Med.	β_2	-43.0847(1.0045)	-43.08429(1.004)
High	β_3	-43.5321(1.0122)	-43.53161(1.0121)
Pos.C	β_4	-40.5714(0.9768)	-40.57099(0.9767)
Weib. shape	ρ	10.7070(02473)	10.7070(02473)
Var1.	$d1$	0.9764(0.1698)	0.9764(0.1698)
OD Par.	α	0.8932(0.0463)	0.8932(0.04625)
-2l		28069	28069.24
Duration		60 sec.	35 sec.

This is done by considering the subclusters as the random effect at the cluster level and specifying the same variance for the sub-clusters. However, it is applicable in specific case when we have few sub-clusters and single dimensional random effect at each hierarchical level (random intercept model).

The estimates using both GVA and exact method are given in Table 1 and 2. Implementation of Weibull-Normal-Normal model in PROC NLMIXED had convergence problem. We considered 10% of the data, to be able to evaluate the performance of Gaussian variation approximation. Interestingly, we get the result in few minutes (just using standard optimization) while using PROC NLMIXED for the small data it was taking hours. In terms of the accuracy, the parameter estimates as well as the standard error using GVA estimate were almost the same as the golden standard estimate for both Weibull-Normal and Weibull-Gamma-Normal models. It was both fast and accurate.

6. GVA for Poisson Models

6.1 Poisson-Gamma, Poisson-Normal, Poison-Gamma-Normal

A Poisson with gamma random effect, it is just a negative binomial model. If we consider Poisson-Normal model (Poisson mixed model), the marginal likelihood contains an intractable integral. Applying GVA approximation results in a new lower bound likelihood with no integral problem which is similar to Weibull-Normal.

$$\begin{aligned} Y_{ij} &\sim \text{Poisson}(\lambda_{ij}) \\ \lambda_{ij} &= \exp(\boldsymbol{x}_{ij}{}'\boldsymbol{\xi} + b_i) \end{aligned}$$

Table 2: Exact and GVA estimation for Weibull-Normal-Normal

Weibull Normal Normal			
		Exact	GVA
Effect	Par.	Estimate(s.e.)	Estimate(s.e.)
Veh.	β_0	-19.2239(0.9048)	-19.1356(0.8970)
Low *vs.* veh.	β_1	-6.9489(0.4468)	-6.9170(0.4426)
Med.*vs.* veh.	β_2	-7.1410(0.4555)	-7.1063(0.4512)
High *vs.* veh.	β_3	-7.4521(0.4670)	-7.4168(0.4627)
Pos.C.*vs.* veh.	β_4	-5.5195(0.4650)	-5.4933(0.4602)
Weib. shape	ρ	6.4192(0.2885)	6.3898(0.2861)
Var1.	$d1$	7.5943E-07(0.00026)	4.8067E-06(0.00017)
Var2.	$d2$	0.7184(0.1689)	0.6962(0.1644)
Duration		2 hr and 30 min.	10 min.

$$b_i \quad \sim \quad \text{Normal}(0, d) \tag{8}$$

The lower bound is:

$$\begin{aligned} l(\boldsymbol{\xi}, d, \mu, \boldsymbol{\Lambda}) \quad &= \quad \sum_{i=1}^{N} \sum_{j=1}^{n_i} \left[y_{ij}(\boldsymbol{x}'_{ij}\boldsymbol{\xi} + \mu_i) - e^{\boldsymbol{x}'_{ij}\boldsymbol{\xi} + \mu_i + \frac{1}{2}\Lambda_i} - \log(y_{ij!}) \right] \\ & \quad -\frac{N}{2}\log(d) - \sum_{i=1}^{N} \frac{1}{2d}(\mu_i^2 + \Lambda_i) + \sum_{i=1}^{N} \frac{1}{2}\log(\Lambda_i). \end{aligned}$$

Also here, extending to more hierarchical random effect is straight forward by applying GVA to all hierarchical random effects independently/separately.

$$\begin{aligned} l(\boldsymbol{\xi}, d_1, d_2, \boldsymbol{\mu}, \boldsymbol{\Lambda}) \quad &= \quad \sum_{i=1}^{N} \sum_{j=1}^{M} \sum_{k=1}^{n_{ij}} \left[y_{ijk}(\boldsymbol{x}'_{ijk}\boldsymbol{\xi} + \mu_i + \mu_{ij}) - e^{(\boldsymbol{x}'_{ijk}\boldsymbol{\xi} + \mu_i + \mu_{ij} + \frac{1}{2}\Lambda_i + \frac{1}{2}\Lambda_{ij})} \right. \\ & \quad - \log(y_{ijk!})] - \frac{N}{2}\log(d_1) - \sum_{i=1}^{N} \frac{1}{2d_1}(\mu_i^2 + \Lambda_i) + \sum_{i=1}^{N} \frac{1}{2}\log(\Lambda_i) \\ & \quad -\frac{NM}{2}\log(d_2) - \sum_{i=1}^{N} \sum_{j=1}^{M} \frac{1}{2d_2}(\mu_{ij}^2 + \Lambda_{ij}) + \sum_{i=1}^{N} \sum_{j=1}^{M} \frac{1}{2}\log(\Lambda_{ij}). \end{aligned}$$

For outcome y_{ijk} of cluster $i = 1, \ldots, N$ in subcluster $j = 1, \ldots, M$ measured at occasion $k = 1, \ldots, n_{ij}$. d_1 and d_2 are the variances of the first (at cluster level) and second (at subcluster) hierarchical random effects.

When an Overdispersion gamma random effect is added to the Weibull-Normal model in (8), it leads to Poisson-Gamma-Normal model of Molenberghs et al (2007), a model for repeated Poisson data with overdispersion. The gamma random effect is first integrated out and Gaussian variation approximation is applied to the normal random effect. In general, The Weibull and Poisson models have similar form.

$$\begin{aligned} Y_{ij} \quad &\sim \quad Poi(\lambda_{ij}) \\ \lambda_{ij} \quad &= \quad \theta_{ij}\exp(\boldsymbol{x}_{ij}{}'\boldsymbol{\xi} + b_i), \\ \theta_{ij} \quad &\sim \quad \text{Gamma}(\alpha, \beta), \end{aligned}$$

where Y_{ij} is the j^{th} outcome measured for subject i, $i = 1, \ldots, N$, $j = 1, \ldots, n_i$. The lower bound is given by:

$$l(\boldsymbol{\xi}, d, \boldsymbol{\mu}, \boldsymbol{\Lambda}) \quad = \quad \sum_{i=1}^{N} \sum_{j=1}^{n_i} [\log((y_{ij} + \alpha - 1)!) - \log((\alpha - 1)!) - \log(y_{ij}!) + y_{ij}\log(\beta)$$

Table 3: Exact and GVA estimation for Poisson-Normal model and Poisson-Gamma-Normal (Combined model)

Poisson-Normal			
		Exact	GVA
Effect	Parameter	Estimate(s.e.)	Estimate(s.e.)
Intercept Placebo	β_{00}	0.8179(0.1677)	0.8179(0.1675)
Slope Placebo	β_{01}	-0.0143(0.0044)	-0.0143(0.0044)
Difference in Intercept	$\beta_{10}-\beta_{00}$	-0.1703(0.2387)	-0.1704(0.2385)
Difference in Slope	$\beta_{11}-\beta_{01}$	0.0023(0.0062)	0.0023(0.0062)
Variance of RE	d	1.1568(0.1844)	1.1543(0.1839)
-2 Loglikelihood		-6810	-6808.87
Duration		11 sec.	4 sec.

Poisson Gamma Normal (Combined)			
		Exact	GVA
Effect	Parameter	Estimate(s.e.)	Estimate(s.e.)
Intercept Placebo	β_{00}	-2.9862(0.1965)	-2.9856(0.1759)
Slope Placebo	β_{01}	-0.0248(0.0077)	-0.0248(0.0077)
Difference in Intercept	$\beta_{10}-\beta_{00}$	-0.2557(0.2500)	-0.2556(0.2498)
Difference in Slope	$\beta_{11}-\beta_{01}$	0.0130(0.0107)	0.0130(0.0107)
Var.of RE	d	1.1290(0.1850)	1.1274(0,1847)
OD par.	α	2.4640(0.2113)	2.4625(0.0324)
-2 Loglikelihood		-7664	-7664.17
Duration		60 sec.	50 sec.

$$+y_{ij}(\boldsymbol{x}'_{ij}\boldsymbol{\xi}+\mu_i)-(y_{ij}+\alpha)\int\log(1+\beta e^{\boldsymbol{x}'_{ij}\boldsymbol{\xi}+b_i})q(b_i)db_i\Big]$$

$$-\frac{N}{2}\log(d)-\sum_{i=1}^{N}\frac{1}{2d}(\mu_i^2+\Lambda_i)+\sum_{i=1}^{N}\frac{1}{2}\log(\Lambda_i).$$

6.2 Illustration on Epilepsy data

Both the Poisson-Normal and Poisson-Gamma-Normal are applied to the epilepsy data. The model for Poisson-Gamma-Normal is given by:

$$\log(\lambda_{ij})=\begin{cases}(\beta_{00}+b_i)+\beta_{01}t_{ij} & \text{if} \quad \text{placebo}\\ (\beta_{10}+b_i)+\beta_{11}t_{ij} & \text{if} \quad \text{treated}\end{cases}$$

The result for Poisson-Normal and Poisson-Gamma-Normal is given in Table 3. For Poisson-Normal it was fast and accurate and for Poisson-Gamma-Normal which still require numerical approximation to the resulted GVA, it was also fast and fairly accurate. Although the approximation using both methods was fast, approximation using GVA was faster and the gain in computational time is seen in the Jimma infant data which is presented in Table 5.

7. GVA for Logistic Models

7.1 Logistic-Normal and Logistic-beta-Normal Models

For Logistic model, we have:

$$Y_{ij} \quad \sim \quad \text{bernoulli}(\pi_{ij})$$

$$\begin{aligned}\frac{\pi_{ij}}{1-\pi_{ij}} &= \exp(\boldsymbol{x}'_{ij}\boldsymbol{\xi}+b_i)\\ b_i &\sim \text{Normal}(0,d)\end{aligned} \tag{9}$$

Applying GVA results in the lower bound $\underline{l}(\boldsymbol{\xi},d,\mu,\Lambda)$. Unlike Poisson-Normal and Weibull-Normal, the GVA likelihood has still non-tractable likelihood.

$$\begin{aligned}\underline{l}(\boldsymbol{\xi},d,\mu,\Lambda) &= \sum_{i=1}^{N}\sum_{j=1}^{n_i}\left[y_{ij}(\boldsymbol{x}'_{ij}\boldsymbol{\xi}+\mu_i)-\int\log(1+e^{\boldsymbol{x}'_{ij}\boldsymbol{\xi}+b_i})q(b_i)db_i\right]\\ &\quad -\frac{N}{2}\log(d)-\sum_{i=1}^{N}\frac{1}{2d}(\mu_i^2+\Lambda_i)+\sum_{i=1}^{N}\frac{1}{2}\log(\Lambda_i).\end{aligned}$$

Considering an extended model of (9),a model combined with beta distribution/random effect:

$$\begin{aligned}Y_{ij} &\sim binary(\pi_{ij})\\ \pi_{ij} &= \theta_{ij}\frac{\exp(\boldsymbol{x}'_{ij}\boldsymbol{\xi}+b_i)}{1+\exp(\boldsymbol{x}'_{ij}\boldsymbol{\xi}+b_i)}\\ \theta_{ij} &\sim \text{Beta}(\alpha,\beta)\end{aligned}$$

The lower bound is given by:

$$\begin{aligned}\underline{l}(\boldsymbol{\xi},d,\alpha,\beta,\mu,\Lambda) &= \sum_{i=1}^{N}\sum_{j=1}^{n_i}\left[y_{ij}\log(\alpha)-\log(\alpha+\beta)-\int\log(1+e^{\boldsymbol{x}'_{ij}\boldsymbol{\xi}+b_i})q(b_i)db_i\, +\right.\\ &\quad \left. y_{ij}(\boldsymbol{x}'_{ij}\boldsymbol{\xi}+\mu_i)+(1-y_{ij})\int\log(\alpha+\beta+\beta e^{\boldsymbol{x}'_{ij}\boldsymbol{\xi}+b_i})q(b_i)db_i\right]\\ &\quad -\frac{N}{2}\log(d)-\sum_{i=1}^{N}\frac{1}{2d}(\mu_i^2+\Lambda_i)+\sum_{i=1}^{N}\frac{1}{2}log(\Lambda_i).\end{aligned}$$

For identifiability problem, we fix $\alpha/\beta=c$. The lower bound is then given by:

$$\begin{aligned}l(\boldsymbol{\xi},d,c,\boldsymbol{\mu},\boldsymbol{\Lambda}) &= \sum_{i=1}^{N}\sum_{j=1}^{n_i}\left[-\log(1+c)-\int\log(1+e^{\boldsymbol{x}'_{ij}\boldsymbol{\xi}+b_i})q(b_i)db_i\, +\right.\\ &\quad \left. y_{ij}(\boldsymbol{x}'_{ij}\boldsymbol{\xi}+\mu_i)+(1-y_{ij})\int\log(1+c+ce^{\boldsymbol{x}'_{ij}\boldsymbol{\xi}+b_i})q(b_i)db_i\right]\\ &\quad -\frac{N}{2}\log(d)-\sum_{i=1}^{N}\frac{1}{2d}(\mu_i^2+\Lambda_i)+\sum_{i=1}^{N}\frac{1}{2}\log(\Lambda_i).\end{aligned}$$

We see that the GVA for Weibull-Gamma-Normal, Poisson-Gamma-Normal and Logistic models, have similar algebraic form which still needs numerical approximation.

7.2 Illustration on EG data

We applied to the EG data. The model is given by:

$$Logit(y_{ij}=1)=\beta_0C_{ij}+\beta_1L_{ij}+\beta_2M_{ijk}+\beta_3H_{ijk}+b_i,$$

The result is presented in Table 4. The performance of the GVA interms of the accuracy of the parameter estimate as well as the standard error for the Logistic models was slightly lower as compared to the Weibull and Poisson Models.

Table 4: Exact and GVA estimation for Logistic-Normal and Logistic-Beta-Normal (Combined model)

Logistic-Normal			
		Exact	GVA
Effect	Parameter	Estimate(s.e.)	Estimate(s.e.)
Control	β_0	-6.2344(0.8860)	-6.1592(0.8543)
Low	β_1	-3.8615(0.5420)	-3.8011(0.5190)
Medium	$\beta_2 - \beta_{00}$	-1.7370(0.4332)	-1.6984(0.4121)
High	$\beta_3 - \beta_{01}$	1.5695(0.4693)	1.5274(0.4466)
Var. of RE	d	3.9988(1.0977)	3.5808(0.9360)
Duration		18 sec.	6 sec.
Logistic-Beta-Normal			
		Exact	GVA
Effect	Parameter	Estimate(s.e.)	Estimate(s.e.)
Control	β_0	-6.2344(0.8860)	-6.1592(0.8543)
Low	β_1	-3.8615(0.5420)	-3.8011(0.5190)
Medium	$\beta_2 - \beta_{00}$	-1.7370(0.4332)	-1.6984(0.4121)
High	$\beta_3 - \beta_{01}$	1.5695(0.4693)	1.5274(0.4466)
Var. of RE	d	1.3860(0.2745)	1.2756(0.2614)
OD par.	β/α	1.2957E-07(0.00011)	2.6822E-09(0.00001)
Duration		35 sec.	40 sec.

8. Discussion and Conclusion

Generalized mixed models often involve intractable integrals. Different approximating techniques exist which can be broadly categorized as approximating the integrand, the data or the integral itself. Gaussian variational approximation approximate the integrand by introducing a set of variational densities (to the posterior densities) in such a way that their evaluation is tractable. It is applicable for both bayesian and likelihood. In this paper, we focus on the likelihood frame work by approximating the posterior density of the normal random effect (by a set of normal densities). We considered three families of GLMM models; 1) The Weibull models: Weibull-Normal, Weibull-Normal-Normal and Weibull-Gamma-Normal; 2) Poisson models: Poisson-Normal, Poisson-Normal-Normal and Poisson-Gamma-Normal; 3) Logistic models.

The GVA approximation was applied to the comet data for Weibull models, epilepsy data for Poisson models and EG data for Logistic model. Estimate using adaptive numerical gaussian approximation in SAS Proc-nlmixed was taken as exact/golden standard estimate (named as exact estimate). For Weibull-Normal and Poisson-Normal, estimation using GVA was faster and very accurate (in contrast with the exact estimate). For models with higher hierarchical random effect (Weibull-Normal-Normal), normally standard software SAS Proc-nlmixed does not accommodate , it is only possible with the use of some modeling trick for cases of small number of sub-clusters. Applying to the comet data, we were having problem in convergence. Thus we were forced to deal with the reduced data yet it was taking very long time to converge. Estimating using GVA was much faster and fairly accurate. Considering Overdispersed hierarchical models (Weibull-Gamma-Normal, Poisson-Gamma-Normal and logistic models), applying GVA approximation still requires numerical approximation. It was also fast and fairly accurate for the parameters of interest.

In general, it can be a good approximation technique especially when the numerical approximation using standard software fails/very restrictive, either take long time, problem in convergency or when it doesn't allow to accommodate such futures. For instance when

Table 5: Overview of computational efficiency, duration for convergence

Models	Datasets	Exact	GVA
Weibull Normal	Comet	70 sec.	7 sec.
Weibull Gamma Normal	Comet	1 min.	35 sec.
Poisson Normal	Epilesy	11 sec.	4 sec.
Poisson Gamma Normal	Epilepsy	1 min.	50 sec.
Poisson Normal	Jimma	45min.	3 min.
Poisson Gamma Normal	Jimma	4 hr. and 40 min.	5 min.
Logistic Normal	EG	18 sec.	6 sec.
Logistic Normal	EG	35 sec.	40 sec.

we have more than two hierarchical levels where and when we have higher dimension of random effect. Table 5 gives overview of the computational efficiency in terms of the duration to convergence.

REFERENCES

Aerts, M., Geys, H., Molenberghs, G. and Ryan, L. M.(2002),*Topics in Modelling of Clustered Data*, Chapman and Hal/CRC.

Agresti, A. (2002),*Categorical Data Analysis* (2nd ed.)., New York: John Wiley & Sons.

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

Hall, P., Ormerod, J.T. and Wand, M.P. (2011), “Theory of Gaussian Variational Approximation for a Poisson Linear Mixed Model,” *Statistica Sinica*,**21**, 369-389.

Blei, D. M. and Jordan, M. I. (2006), “Variational Inference Dirichlet Process Mixtures,” *Bayesian Analysis*,**1**, 121-144.

Ghebretinsae, A.H., Faes, C., Molenberghs, G., De Boeck, M., and Geys, H. (2011), “A Bayesian, generalized frailty model for comet assays,” *Submitted for publication.*

Lee, Y. and Nelder, J.A. (1996), “Hierarchical generalized linear models,” *Journal of the Royal Statistical Society, Series B*, **58**, 619-656.

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

Kullback, S. and Leibler, R.A. (1951), “On information and sufficiency,” *The Annals of Mathematical Statistics*, **22**, 79-86.

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

McGrory, C.A. and Titterington, D.M. (2007), “Variational approximations in Bayesian model selection for finite mixture distributions,” *Computational Statistics and Data Analysis*, **51**, 5352-5367.

Molenberghs, G., Verbeke, G., and Demétrio, C. (2007), “An extended random-effects approach to modeling repeated, overdispersed count data,” *Lifetime Data Analysis*, **13**, 513–531.

Molenberghs, G., Verbeke, G., Demétrio, C.G.B., and Vieira, A. (2010), “A family of generalized linear models for repeated measures with normal and conjugate random effects,” *Statistical Science*, **00**, 000–000.

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

Wang, B. and Titterington, D.M. (2006), “Convergence properties of a general algorithm for calculating variational Bayesian estimates for a normal mixture model,” *Bayesian Analysis*,**1**, 625-650

Ormerod, J.T. and Wand, M.P. (2010), “Explaining Variation Approximations,” *The American Statistician*, **64**, 140-153.

Ormerod, J.T. and Wand, M.P. (2012), “Gaussian variational approximate inference for generalized linear mixed models,” *Journal of Computational and Graphical Statistics*, **21**, 2-17.

Ormerod, J. T. and Wand, M. P. (2009), “ Comment on paper by Rue, Martino and Chopin. J.,” *Journal of the Royal Statistical Society, Series B*, **71**, 377-378.