WORKING PAPER SERIES

WORKING PAPER NO 4, 2009

Swedish Business School at Örebro University

An efficient algorithm for the pseudo likelihood estimation of the generalized linear mixed models (GLMM) with correlated random effects

Md Moudud Alam
Swedish Business School, Örebro University, Sweden

http://www.oru.se/esi/wps

SE-701 82 Örebro
SWEDEN

ISSN 1403-0586

An efficient algorithm for the pseudo likelihood estimation of the generalized linear mixed models (GLMM) with correlated random effects

Md Moudud Alam*

December 18, 2008

Abstract

This paper presents a two-step pseudo likelihood estimation technique for generalized linear mixed models with correlated random effects. The proposed estimation technique does not require reparametarisation of the model. Multivariate Taylor's approximation has been used to approximate the intractable integrals in the likelihood function of the GLMM. Based on the analytical expression for the estimator of the covariance matrix of the random effects, a condition has been presented as to when such a covariance matrix can be estimated through the estimates of the random effects. An application of the model with a binary response variable has been presented using a real data set on credit defaults from two Swedish banks. Due to the use of two-step estimation technique, proposed algorithm outperforms the conventional pseudo likelihood algorithms in terms of computational time.

Mathematics Subject Classification: Primary 62J12; Secondary 65C60
Keywords: PQL, Laplace approximation, interdependence, cluster errors.

*Dalarna University and Örebro University; e-mail: maa@du.se

1 Introduction

The literature on estimation of the generalized linear mixed models (GLMM) is abundant. The justification of yet another paper is given from the fact that none of the estimation method currently available can provide an exact maximum likelihood estimator except for the case of Gaussian family with identity link. Moreover, most of them are computationally too heavy and all needs some restrictive assumptions about the random effects, namely they being independent. This paper presents a general (in that it works for independent or correlated random effects) algorithm to estimate the GLMM parameters using a two-step estimation procedure. The estimation procedure is derived on the second order Taylor's series approximation of the likelihood function. From the analytical view point the proposed approach can be regarded as a generalization of the Pseudo (or Penalized Quasi) Likelihood (PL or PQL) approach (Breslow & Clayton 1993, Wolfinger & O'Connell 1993) for correlated random effects. A brief discussion on the PL and other methods and algorithms for GLMM is offered in Section 2.

The PL approach has been chosen to work with because of the speed of the algorithm. It is worth noting that, anything faster than PQL is to be the fastest algorithm. However, while the available computer implementation of the PL or PQL requires independence of the random effects, between subjects, the proposed method does not require any independence assumption for the random effects terms.

The proposed two-step estimation procedure, in the first step, estimates the fixed effect parameters via a simple generalized linear model (GLM) (McCullagh & Nelder 1989) procedure while, in the second step, the random effects and the covariance parameters are estimated via a procedure similar to PQL. This two-step estimation procedure makes the mathematical derivation simple and computational implementation fast in comparison to the conventional PQL approaches.

The paper has been organized in the following way. Section two outlines on the GLMM and available estimation techniques while section three provides the mathematical explanation of the proposed estimation technique and section four gives an application of the technique to a real data set. Section five ends the paper with a concluding discussion.

2 Estimation of generalized linear mixed models

The extension of the generalized linear models with random effects terms is called the generalized linear mixed model (McCulloch & Searle 2001). The conditional independence assumption of the response variable, given the random effects, plays an important role in the formulation of the GLMM. This paper retains the conditional independence assumption. Subject to the above assumption, the joint likelihood function of the fixed effects parameters, $\boldsymbol{\beta}$, and the covariance parameter of the random effects, $\mathbf{D}$, for a GLMM is given as

$$L(\boldsymbol{\beta},\ \mathbf{D}|\mathbf{Y}) = \int L\left(\boldsymbol{\beta},\ \mathbf{D}|\mathbf{Y},\mathbf{u}\right) f\left(\mathbf{u}\right) d\mathbf{u} \tag{1}$$

where, $\mathbf{Y} = \{y_{ij}\}$, $(i = 1, 2, .., n_j$ and $j = 1, 2, .., k)$ is the vector of response variable, $\mathbf{u} = (u_1,\ u_2, ..., u_k)^T$ is the vector of random effects and the integration is taken over a k dimensional space. Under the standard assumption of the GLMM the distribution of the response variable, y_{ij}, given the random effect, u_j, belongs to a member of the exponential family of distributions. While, $f(\mathbf{u})$ is a specific pdf often it is *i.i.d* normal which converts the multivariate integral in equation (1) into a product of k univariate integrals. However, this paper drops the last assumption and considers $\mathbf{u}$ as a multivariate normal variate with mean vector, $\mathbf{0}$, and an unstructured covariance matrix, $\mathbf{D}$. Examples of such kind of correlated random effects can be given from the genetic relation's view point (see *e.g.* Searle, Casella & McCulloch (1992) pp. 383) or from the inter-industry default correlation's perspective (see *e.g.* Alam & Carling (2008)). It is worth noting that the *i.i.d* u_j's are a special form of multinormal $\mathbf{u}$, with a diagonal covariance matrix, $\mathbf{D}$, having all the diagonal elements identical. Therefore, a method derived for the unstructured $\mathbf{D}$ contains *i.i.d.* u_j's as a special case.

2.1 Estimation theory

There are several approaches to approximate integrals in (1). The PQL approach uses the Laplace approximation which is, in fact, based on the first (or sometimes second) order Taylor's approximation. The numerical integration method uses Gauss-Hermite quadrature (Broström 2007, Butler & Moffitt 1982) or Adaptive Gaussian method to evaluate the integral numerically (Littell, Milliken, Stroup & D. 1996). More computationally intensive Markov-chain Monte-Carlo (MCMC) method uses MCMC integration (McCulloch & Searle 2001) while hierarchical (h-) likelihood method bypasses the integration via h-likelihood (Lee & Nelder 1996, Lee & Nelder 2006).

Since the work in this paper is based on the arguments similar to PL or PQL, I skip this discussion by offering only a detailed theoretical discussion on PQL regarding the parameter estimates. Under the formal GLMM assumption, the conditional likelihood, $L\left(\boldsymbol{\beta},\ \mathbf{D}|\mathbf{Y},\mathbf{u}\right)$, with a canonical link is expressed as

$$L\left(\boldsymbol{\beta},\mathbf{D}|\mathbf{Y},\mathbf{u}\right) = \exp\left[\sum_{j=1}^{n_j}\sum_{i=1}^{n}\left\{\frac{y_{ij}\eta_{ij} - b\left(\eta_{ij}\right)}{a\left(\phi\right)} + c\left(y_{ij,}\phi\right)\right\}\right]$$

$$\Rightarrow \mathit{log}\left(L\left(\boldsymbol{\beta},\mathbf{D}|\mathbf{Y},\mathbf{u}\right)\right) = l = \sum_{j=1}^{n_j}\sum_{i=1}^{n}\left\{\frac{y_{ij}\eta_{ij} - b\left(\eta_{ij}\right)}{a\left(\phi\right)} + c\left(y_{ij,}\phi\right)\right\} \tag{2}$$

where, $\eta_{ij} = \mathbf{X}_{ij}\boldsymbol{\beta} + \mathbf{Z}_{ij}\mathbf{u}$ is called the linear predictor, $b\left(.\right)$ is called the cumulant function, $a\left(\phi\right)$ is called the dispersion parameter which is 1 for the Binomial and the Poisson distribution, and the conditional expectation, $E(\mathbf{Y}|\mathbf{u}) =\mu$, satisfies $g(\mu) = \eta$ for some function, $g(.)$, which is called the link function. Using matrix notations, equation (2) can be written as

$$l = \frac{\mathbf{Y}^T\left(\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\mathbf{u}\right) - \mathbf{1}^T b\left(\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\mathbf{u}\right)}{a\left(\phi\right)} + \mathbf{1}^T c\left(\mathbf{Y},\phi\right) \tag{3}$$

Then equation (1) can be re-expressed as

$$L(\boldsymbol{\beta},\ \mathbf{D}|\mathbf{Y}) = \int \exp\left[l\right] f\left(\mathbf{u}\right) d\mathbf{u} \tag{4}$$

where, following the assumption of the multivariate normal distribution of $\mathbf{u}$, $f(\mathbf{u})$ can be given as

$$f\left(\mathbf{u}\right) = \left(2\pi\right)^{-\frac{k}{2}}\left|\mathbf{D}\right|^{-\frac{1}{2}}\exp\left[-\frac{1}{2}\mathbf{u}^T\mathbf{D}^{-1}\mathbf{u}\right]$$

substituting in equation (4) we have

$$\begin{aligned} L(\boldsymbol{\beta},\ \mathbf{D}|\mathbf{Y}) &= \int \exp\left[\frac{\mathbf{Y}^T\left(\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\mathbf{u}\right) - \mathbf{1}^T b\left(\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\mathbf{u}\right)}{a\left(\phi\right)} + \mathbf{1}^T c\left(\mathbf{Y},\phi\right)\right]\left(2\pi\right)^{-\frac{k}{2}}\left|\mathbf{D}\right|^{-\frac{1}{2}}\exp\left[-\frac{1}{2}\mathbf{u}^T\mathbf{D}^{-1}\mathbf{u}\right] d\mathbf{u} \\ &= \left(2\pi\right)^{-\frac{k}{2}}\left|\mathbf{D}\right|^{-\frac{1}{2}}\exp\left[\mathbf{1}^T c\left(\mathbf{Y},\phi\right)\right]\int \exp\left[\frac{\mathbf{Y}^T\left(\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\mathbf{u}\right) - \mathbf{1}^T b\left(\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\mathbf{u}\right)}{a\left(\phi\right)} - \frac{1}{2}\mathbf{u}^T\mathbf{D}^{-1}\mathbf{u}\right] d\mathbf{u} \end{aligned}$$

Now, it is obvious from the above equation that the integral in the right hand side has the form, $I = \int_R \exp\left[-h\left(\mathbf{u}\right)\right] d\mathbf{u}$, which leaves a scope for the application of the Laplace approximation[1] to evaluate the integral. Following, the Laplace approximation based on the second order Taylor's series approximation on $h(\mathbf{u})$, the log likelihood can be expressed as

$$l_{PQL} = \frac{\mathbf{Y}^T\left(\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\widetilde{\mathbf{u}}\right) - \mathbf{1}^T b\left(\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\widetilde{\mathbf{u}}\right)}{a\left(\phi\right)} + \mathbf{1}^T c\left(\mathbf{Y},\phi\right) - \frac{1}{2}\widetilde{\mathbf{u}}^T\mathbf{D}^{-1}\widetilde{\mathbf{u}} \tag{5}$$

where, l_{PQL} is the log-likelihood function of the GLMM under PQL approach and $\widetilde{\mathbf{u}}$ is the value of $\mathbf{u}$ evaluated at the minimum of $h(\mathbf{u})$ (see also Breslow & Clayton (1993) and Wand (2002)). From equation (5) it is clear that the log-likelihood, after applying Laplace approximation, splits into two parts. The first part contains only the fixed parameters and the second part contains the covariance of the random effects. This feature of the l_{PQL} leaves a scope for the standard PQL approaches to use GLM procedure to estimate the model parameters.

2.2 Estimation in practice

Based on the PQL and the quadrature methods to evaluate the integral in (1), several similar algorithms have been developed for the estimation of the GLMM and different computer package uses different version of them. A comparative review of the different software packages for the GLMM is provided in Zhou, Perkins & Hui (1999). The implementations of those methods in the standard computer packages, *e.g.* SAS (Littell et al. 1996) or R (Broström 2007, Venables &

[1] Evans & Swartz (1995) contains a brief discussion on the Laplace approximation.

Ripley 2002), are not intended to capture a random effect having an unstructured covariance matrix as presented in model (1). A popular alternative solution in those cases is to reparametarise the unstructured $\mathbf{D}$ as $\mathbf{D} = \boldsymbol{\sigma}^2\mathbf{L}\mathbf{L}^T$ where $\mathbf{L}$ is a lower triangular matrix such as the Cholesky decomposition of $\mathbf{D}$ (Lam & Lee 2004, Lindstrom & Bates 1998). With such decomposition of $\mathbf{D}$, the transformed random effects vector, $\mathbf{v}$, defined as $\mathbf{u} = \mathbf{L}\mathbf{v}$ again become *i.i.d.* $N(0, \boldsymbol{\sigma}^2)$ thus the formal computational methods derived on diagonal $\mathbf{D}$ is applied. The elements in $\mathbf{L}$ are either estimated as the fixed parameters or are assumed to be known and their values are provided from some external sources. However, this paper presents a way to estimate $\mathbf{D}$ where no such reparametarisation is required (see section 3). The estimation technique proposed in section 3 is applicable for the GLMMs with the random effects being correlated within or between subjects while those techniques implemented in the standard computer packages, *e.g.* R and SAS, are applicable only when the random effects are independent between subjects.

Several application of PQL revealed that it becomes very slow with large data sets (see *e.g.* Carling, Rönnegård & Rosczbach (2004)) though it was still faster than any other alternative. Therefore, this paper follows the same line of PQL but with some further modification with an aim to make it more general and faster.

3 Modification to the pseudo likelihood approach

From equation (4), the marginal likelihood of $\boldsymbol{\beta}$ and $\mathbf{D}$ can be interpreted as an expectation, $E\left(\exp\left[l\right]\right)$, with respect to the multivariate normal distribution of $\mathbf{u}$. Using the multivariate version of Taylor's expansion (Raudenbush, Yang & Yosef 2000) of the function $m(\mathbf{u}) = \exp\left[l\right]$ around the marginal mean of $\mathbf{u}$, which is 0, we have

$$m(\mathbf{u}) = m\left(\mathbf{0}\right) + m^{(1)}\left(\mathbf{u}\right)_{\mathbf{u}=0}\left(\mathbf{u} - \mathbf{0}\right) + \sum_{k=2}^{\infty}\frac{1}{k!}\left[\bigotimes^{k-1}\left(\mathbf{u} - \mathbf{0}\right)^T\right] m^{(k)}\left(\mathbf{u}\right)_{\mathbf{u}=\mathbf{0}}\left(\mathbf{u} - \mathbf{0}\right)$$

$$\Rightarrow L(\boldsymbol{\beta},\ \mathbf{D}|\mathbf{Y}) = E\left(m(\mathbf{u})\right) \approx m\left(\mathbf{0}\right) + \mathbf{0} + \frac{1}{2}E\left\{\mathbf{u}^T m^{(2)}\left(\mathbf{u}\right)_{\mathbf{u}=\mathbf{0}}\,\mathbf{u}\right\} \tag{6}$$

where, $m^{(k)}\left(\mathbf{u}\right) = \frac{\partial vecm^{(k-1)}(\mathbf{u})}{\partial \mathbf{u}^T}$, $\otimes$ represents Kronecker product and the correction terms being, $\sum_{k=3}^{\infty}\frac{1}{k!}E\left\{\left[\bigotimes^{k-1}\mathbf{u}^T\right] m^{(k)}\left(\mathbf{u}\right)_{\mathbf{u}=\mathbf{0}}\,\mathbf{u}\right\}$.

Now, regarding the estimation of the fixed effects parameters, $\boldsymbol{\beta}$, already the second order term in equation (6) is relatively flat and commonly ignored in the PQL methods[2](Wand 2002). Therefore, after ignoring the second order term, the likelihood for $\boldsymbol{\beta}$ becomes a likelihood of GLM having no random effect left in it. This is not surprising since Maddala (1987) concluded

[2] Note that PQL uses the Taylor's expansion of the function $\exp\left(l - \frac{1}{2}\mathbf{u}^T D^{-1}\mathbf{u}\right)$ evaluated at the maximum point of it.

that the fixed effect approach can produce a consistent estimate of the $\boldsymbol{\beta}$ parameters even if there is an autocorrelation in the model due to the random effects, while Alam & Carling (2008) supported Maddala (1987)'s claim by empirical evidence. Therefore, the first order Taylor's approximation around $\mathbf{u} = E(\mathbf{u})$ suggests estimating the $\boldsymbol{\beta}$ parameters through a simple GLM ignoring the random effect terms.

The estimation technique suggested here for $\boldsymbol{\beta}$ might intuitively look confusing since the estimation ignores the realization of the random effects while other PQLs do consider them. Analytically, it is valid as long as the first order Taylor's approximation is reasonable. However, such approximation is not strange since in statistical literature we often see such approximation which we call the Delta method. In linear mixed model (LMM) case, the proposed estimator for fixed effects is an OLS estimator. It is well established for the balanced experiments with LMM that OLS=GLS=ML. For unbalanced experiments such equality holds if and only if there exist a matrix $\mathbf{Q}$ such that $V(\mathbf{Y})\mathbf{X} = \mathbf{XQ}$ (see Searle et al. (1992), pp. 159-161).

In order to make the scenario more clear we plot the log-likelihood for β of a logistic mixed model using simulated data with true $\mathbf{D} = \mathbf{I}$, $x_{ikt} \sim N(0,1)$, $\beta = 0.5$, $k = 3$, $t = 20$ and 40, $n_{kt} = 200$ (see section 4 for detailed model specification). Figure 1 and 2 present the plots of the conditional log-likelihoods for the single fixed effect parameter, β, evaluated, in each cases with t=20 (see Fig. 1) and 40 (see Fig. 2), at the true value of the random effects, *i.e.* $\mathbf{u} = \mathbf{u}_{true}$, (solid line) and at its marginal mean *i.e.* $\mathbf{u} = \mathbf{0}$ (dotted line). Figures 1 and 2 reveal that the both conditional log-likelihoods have their maximum at the same value for β hence they should provide the same estimate for it. The situation might be a little worse for small t (Fig. 1) but it becomes better as t increases (Fig. 2).

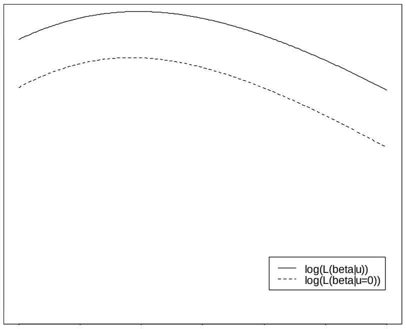

Figure 1: Conditional log-likelihood for a logistic GLMM with t=20

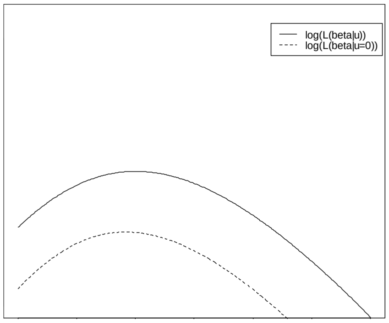

Figure 2: Conditional log-likelihood for a logistic GLMM with t=40

This paper does not, however, suggest using the same likelihood as presented in equation (6) for the estimation of the covariance parameters. This is, firstly, because there is no guarantee that higher order terms that we ignored in (6) are also flat in $\mathbf{D}$. Secondly, the simplification of the Taylor's expansion of $m(\mathbf{u})$ with higher order terms is not easy. Thus, for the estimation of $\mathbf{D}$, equation (4) will be treated in the way similar to PQL (see Section 2). From equation (4) we have

$$L(\boldsymbol{\beta},\ \mathbf{D}|\mathbf{Y}) = \int \exp[l] \frac{|\mathbf{D}^{-1}|^{1/2}}{(2\pi)^{k/2}} \exp\left[-\frac{1}{2}\mathbf{u}^T\mathbf{D}^{-1}\mathbf{u}\right] d\mathbf{u}$$

$$\Rightarrow L(\boldsymbol{\beta},\ \mathbf{D}|\mathbf{Y}) = \frac{|\mathbf{D}^{-1}|^{1/2}}{(2\pi)^{k/2}} \int \exp\left[-\left(-l+\frac{1}{2}\mathbf{u}^T\mathbf{D}^{-1}\mathbf{u}\right)\right] d\mathbf{u} \tag{7}$$

Let, $h(\mathbf{u}) = -l+\frac{1}{2}\mathbf{u}^T\mathbf{D}^{-1}\mathbf{u}$ and assume $h(\mathbf{u})$ has a single minima at $\mathbf{u} = \widetilde{\mathbf{u}}$ **.** Then, applying the multivariate version of the Laplace approximation (Evans & Swartz 1995) in equation (7) we have

$$L(\boldsymbol{\beta},\ \mathbf{D}|\mathbf{Y}) \approx \frac{|\mathbf{D}^{-1}|^{1/2}}{(2\pi)^{k/2}} (2\pi)^{k/2} \{\det[H_h(\widetilde{\mathbf{u}})]\}^{-\frac{1}{2}} \exp[-h(\widetilde{\mathbf{u}})]\ \ [\text{where,} H_h \text{ is the Hessian of } h]$$

$$\Rightarrow \ln(L(\boldsymbol{\beta},\ \mathbf{D}|\mathbf{Y})) = -\frac{1}{2}\ln(|\mathbf{D}|) - \frac{1}{2}\ln\{|H_h(\widetilde{\mathbf{u}})|\} - h(\widetilde{\mathbf{u}}) \tag{8}$$

where, $H_h(\widetilde{\mathbf{u}}) = \mathbf{Z}^T\widetilde{\mathbf{W}}\mathbf{Z}/\mathbf{a}(\phi) + \mathbf{D}^{-1}$ and $\widetilde{\mathbf{W}}$ is the diagonal weight matrix (McCullagh & Nelder 1989) evaluated at $\mathbf{u} = \widetilde{\mathbf{u}}$ (see appendix A-2). Since $\widetilde{\mathbf{u}}$ is the minima of $h(\mathbf{u})$, it can be solved from the equation $\frac{\partial h(\mathbf{u})}{\partial \mathbf{u}}|_{\mathbf{u}=\widetilde{\mathbf{u}}} = 0$ for which a Newton-Raphson algorithm leads us to solve iteratively the following equation (see appendix A.1 for detailed derivation)

$$\widetilde{\mathbf{u}}_{r+1} = \left(\mathbf{Z}^T\widetilde{\mathbf{W}_r}\mathbf{Z} + \mathbf{D}^{-1}\right)^{-1} \mathbf{Z}^T\widetilde{\mathbf{W}}_r\left(\mathbf{Y}^* - \mathbf{X}\boldsymbol{\beta}\right) \tag{9}$$

where, $\mathbf{Y}^*$ is the linearized version of the response variable which is given as: $\mathbf{Y}^* = \widetilde{\mathbf{W}}^{-1}(\mathbf{Y} - \widetilde{\boldsymbol{\mu}}) + \widetilde{\boldsymbol{\eta}}$ and the "r"'s in the subscript indicate that the matrix/vector is evaluated at the rth iteration when $\mathbf{u} = \widetilde{\mathbf{u}}_r$, $(r = 0, 1, ...)$. There is nothing new in equation (9) since Raudenbush et al. (2000) suggested exactly the same equation for estimating $\mathbf{u}$. Using further Laplace approximation it can be shown that $E(\mathbf{u}|\mathbf{Y}) = \widetilde{\mathbf{u}}$ (see Khuri (2003), pp. 548-549 for an outline of the proof). So, this $\widetilde{\mathbf{u}}$ produces the predicted values of the random effect vector, $\mathbf{u}$, given the data. It is worth noting that for known $\mathbf{D}$ matrix, equation (9) provides the predicted random effects however, for an unknown $\mathbf{D}$ we have to estimate it from (8).

The covariance matrix of the random effects, $\mathbf{D}$, can be estimated by maximizing equation (8) given a particular value of $\widetilde{\mathbf{u}}$ and $\widehat{\boldsymbol{\beta}}$. Later, we show that for a given $\widetilde{\mathbf{u}}$ and $\widehat{\boldsymbol{\beta}}$ the maximum (pseudo) likelihood estimate of $\mathbf{D}$ can be obtained analytically (see equation (12)). Since the proposed estimation procedure estimates fixed effects parameters independently of $\mathbf{u}$ and $\mathbf{D}$, let us call it a two-step pseudo likelihood (2PL) approach. It reduces the computational effort of conventional PQL since in 2PL approach we have to check only the joint convergence of $\widetilde{\mathbf{u}}$ and $\widehat{\mathbf{D}}$, while in PQL the convergence of $\widehat{\boldsymbol{\beta}}$ is also checked at the same time. In practical applications, researches generally have more simple form of $\mathbf{D}$ for instance $\mathbf{D} = \boldsymbol{\sigma}^2\mathbf{G}$ where $\mathbf{G}$ is a know matrix and the research interest is only in estimating $\boldsymbol{\sigma}^2$ (Searle et al. 1992). In those cases, the procedure becomes more simple. While, in a general case an initial estimate of $\mathbf{D}$ through the fixed effect approach as proposed in Alam & Carling (2008) would reduce the computational burden to some extent.

Now, in order to estimate $\mathbf{D}$ we have to maximize equation (8) with respect to $\mathbf{D}$. The calculation is shown below

$$\ln\left(L(\boldsymbol{\beta},\ \mathbf{D}|\mathbf{Y})\right) = -\frac{1}{2}\ln\left(|\mathbf{D}|\right) - \frac{1}{2}\ln\left\{|\mathbf{Z}^T\widetilde{\mathbf{W}}\mathbf{Z}/a\left(\phi\right) + \mathbf{D}^{-1}|\right\} + \widetilde{l} - \frac{1}{2}\widetilde{\mathbf{u}}^T\mathbf{D}^{-1}\widetilde{\mathbf{u}} \qquad (10)$$

where $\widetilde{l}$ stands for l evaluated at $\mathbf{u} = \widetilde{\mathbf{u}}$.

After taking matrix differentiation of (10) and simplifying (detailed calculation is presented in Appendix A.2) it can be shown that

$$\begin{aligned}\frac{\partial\ln\left(L(\boldsymbol{\beta},\ \mathbf{D}|\mathbf{Y})\right)}{\partial vec\left(\mathbf{D}\right)} &= -\frac{1}{2}vec\left(\mathbf{D}^{-1}\right)^T + \frac{1}{2}vec\left\{\left(\mathbf{Z}^T\widetilde{\mathbf{W}}\mathbf{Z}/a\left(\phi\right) + \mathbf{D}^{-1}\right)^{-1}\right\}^T \\ &\quad \left(\mathbf{D}^{-1}\otimes\mathbf{D}^{-1}\right) + \frac{1}{2}vec\left(\mathbf{D}^{-1}\widetilde{\mathbf{u}}\widetilde{\mathbf{u}}^T\mathbf{D}^{-1}\right)^T\end{aligned} \qquad (11)$$

Now, equation (11) can be used to find a direct solution of $\mathbf{D}$ using $\frac{\partial\ln(L(\boldsymbol{\beta},\ \mathbf{D}|\mathbf{Y}))}{\partial vec(\mathbf{D})} = 0$ for given $\widehat{\boldsymbol{\beta}}$, $\widetilde{\mathbf{u}}$ and $a\left(\phi\right)$. For binomial and Poisson distributions, we have $a\left(\phi\right) = 1$ which gives the following equation to solve for $\mathbf{D}$ (see Appendix A.3 for detailed calculation)

$$\mathbf{D} = \left(\mathbf{Z}^T\widetilde{\mathbf{W}}\mathbf{Z} + \mathbf{D}^{-1}\right)^{-1} + \widetilde{\mathbf{u}}\widetilde{\mathbf{u}}^T$$

$$\Rightarrow \widehat{\mathbf{D}} = \mathbf{H}_{h(\mathbf{u})}^{-1} + \widetilde{\mathbf{u}}\widetilde{\mathbf{u}}^T \qquad (12)$$

Summarizing the derivations presented in this section, the 2PL algorithm can be presented as

1. Step 1: Estimate $\boldsymbol{\beta}$ using a GLM procedure with $\mathbf{L}(\boldsymbol{\beta}|\mathbf{Y}, \mathbf{u} = \mathbf{0})$.

2. Step 2: Estimate $\mathbf{u}$ and $\mathbf{D}$ using the following iterative algorithm.

   (a) Initialize $\mathbf{u}$ and $\mathbf{D}$. Replace $\boldsymbol{\beta}$ with its estimate from step 1.

   (b) Update $\mathbf{u}$ by iteratively solving $\widetilde{\mathbf{u}}_{r+1} = \left(\mathbf{Z}^T\widetilde{\mathbf{W}_r}\mathbf{Z} + \mathbf{D}^{-1}\right)^{-1}\mathbf{Z}^T\widetilde{\mathbf{W}}_r\left(\mathbf{Y}^* - \mathbf{X}\boldsymbol{\beta}\right)$.

   (c) Update $\mathbf{D}$ with $\widehat{\mathbf{D}} = \mathbf{H}_{h(\mathbf{u})}^{-1} + \widetilde{\mathbf{u}}\widetilde{\mathbf{u}}^T$.

   (d) Stop if $\mathbf{D}$ converges otherwise go to (b).

Equations (9) and (12) lead us to conclude that, the covariance parameters of the random effects can be estimated consistently along with the random effects through a joint iterative procedure. All PQL based algorithms somehow relies on numerical procedure (e.g. Newton-Raphson or something similar) for jointly estimating all the model parameters, including fixed effects and the variance components, while the proposed 2PL algorithm estimates fixed effects just once and uses Newton-Raphson method only for $\mathbf{u}$; which makes the algorithm faster. At the same time, its ability to handle any type of $\mathbf{D}$ matrix makes it a generalized version of the PL/PQL type algorithms. It should be noted here again, that the procedure for the models with a free dispersion parameter, $a(\phi)$ is not discussed in this paper. With non-canonical link, the calculations presented in this section become more complex and are avoided in this paper.

The right hand side of equation (12) can be explained as $E(V(\mathbf{u}|\mathbf{Y})) + V(E(\mathbf{u}|\mathbf{Y}))$ since $\widetilde{\mathbf{u}} = E(\mathbf{u}|\mathbf{Y})$ and $E(\widetilde{\mathbf{u}}) = 0$. From such a structure the consistency arguments for $\widehat{\mathbf{D}}$ is very straightforward. Furthermore, a relation of $E(V(\mathbf{u}|\mathbf{Y}))$ to the Hessian of $h(\mathbf{u})$ reveals that $E(V(\mathbf{u}|\mathbf{Y})) \to 0$ as $n_{ij} \to \infty$. This means that given a large data set we can estimate the covariance matrix, $\mathbf{D}$, by using $\widetilde{\mathbf{u}}$ only. This feature justifies the fixed effect approach suggested in Alam and Carling (Alam & Carling 2008) for estimating $\mathbf{D}$.

4 Application with credit default data

This section presents an application of the proposed algorithm with a real data set, collected from two major Swedish banks, on credit default . This data set was first analyzed by Carling et al. (2004). In the data set there are quarterly information, between the 2^{nd} quarter of 1994 and the 2^{nd} quarter of 2000, on the borrowing companies' financial status, bank data on loan types, credit bureau data, two macro economic variables and an indicator variable stating whether a loan is default by a certain quarter. The research interest involved with the data analysis was to derive a credit risk model by incorporating industry specific default correlation. In Carling et al. (2004), only the within industry default correlation was considered while this paper aims at investigating the possibility of both within and between industry correlation.

In Carling et al. (2004) the industries were defined from some external justification while in this paper industries are defined by merging the SNI industries[3], at the first two digits level, in the way that closely resembles their (Carling et al. 2004) industry definition. Since, it was not possible to construct exactly the same industry definition presented in Carling et al. (2004) by merging SNI industries, the industry definition used in this paper differs slightly from that of Carling et al. (2004). Furthermore, with the 7 industries as presented in Carling et al. (2004), neither the fixed effects nor the random effects model with logistic link converge. This is because

[3] A detailed description about the SNI industry classification can be obtailed from Statistics Sweden's (SCB) official website, www.scb.se.

in some of the industries the relative frequencies of defaults in some quarters are very close to zero which makes the GLM estimation impossible. For that reason, the number of industries has been reduced from 7 to 6.

Following Carling et al. (2004) and considering a between industry correlation I propose the following GLMM specification for the credit default model.

$$y_{ikt}|u_k \backsim iid\ Bin\left(1, p_{ikt}\right)$$

$$\log\left(\frac{p_{ikt}}{1-p_{ikt}}\right) = \mathbf{x}_{ikt}\boldsymbol{\beta} + \mathbf{z}_{ikt}\mathbf{u}_t$$

and

$$\mathbf{u}_t \backsim iid\ \mathbf{MN}_K\left(\mathbf{0}, \mathbf{D}\right)$$

where, $i = 1, 2, ..., n_{kt},\ k = 1, 2, ..., K = 6$ (number of industries), $t = 1, 2, ..., T = 25$ (number of quarters), $\mathbf{x}_{ikt}$ is the ikt^{th} row of the observed design matrix, $\boldsymbol{\beta}$ is the vector of fixed effects parameters, $\mathbf{z}_{ikt}$ is the ikt^{th} row of the design matrix associated with the random effects and contains a 1 in its k^{th} position and 0 otherwise and $\mathbf{u}_t$ is an iid realization of the random effect $\mathbf{u}$ at quarter t. The above model is a logistic mixed model with a multinormal $\mathbf{u}$. The model presented in Carling et al. (2004) was a complementary log-log mixed model with independent u_{kt}. To insure comparability, Carling et al. (2004)'s model is reanalyzed with logistic link but with the new 6 industries. Latter we compare the results to see if the data suggests a complex correlation among the defaults. From here onwards, the logistic model with diagonal $\mathbf{D}$ will be denoted as PQLD wich resembles the model used in Carling et al. (2004) while the same model with unstructured $\mathbf{D}$ will be denoted as PQLU. The same model with cluster effects as fixed effects is also estimated and that model will be denoted as FE.

The fixed effects parameter estimates from those three models are given in the Table 1. Regarding the statistical test of significance for the fixed effects parameters, PQLD uses t-test implemented through %GLIMMIX macro (Littell et al. 1996) in SAS 9.1 while PQLU and FE use Wald Chi-squared test implemented through SAS GENMOD procedure (Olsson 2002).

Table 1 shows that most of the fixed effects parameter estimates are very close between the three models. Except for the coefficient associated with "Bank A". The big differences in estimates are found for the coefficients whose estimates are not significant (see variables and their respective coefficient estimates in rows 2-6 in Table 1). The above feature indicates that the fixed effects parameter estimates are not much sensitive to these three types of model specifications.

Table 1 Fixed effects parameter estimates

Effects	Models					
	PQLD		PQLU		FE	
	Estimate	SE	Estimate	SE	Estimate	SE
Intercept	-4.44	0.214	-4.34	0.175	-3.90	0.378
Credit Survived 1 Year	-0.18	0.137	-0.30	0.132	-0.18^{ns}	0.136
Credit Survived 2 Year	0.03^{ns}	0.137	0.05^{ns}	0.132	0.02^{ns}	0.137
Credit Survived 3 Year	0.13^{ns}	0.138	-0.16^{ns}	0.133	0.17^{ns}	0.138
Credit Survived 4 Year	0.28^{ns}	0.136	0.22^{ns}	0.132	0.28^{ns}	0.135
Credit Survived 5 Year+	0.30^{ns}	0.148	0.10^{ns}	0.141	0.31^{ns}	0.148
Short term credit	0.53	0.039	0.50	0.039	0.54	0.039
Long term credit	-0.32	0.051	-0.30	0.050	-0.31	0.051
Mixed credit	0.00	NA	0.00	NA	0.00	NA
Account. data complete	-2.60	0.090	-2.53	0.088	-2.61	0.090
,, Reported previously	0.73	0.087	0.74	0.084	0.73	0.087
,, Reported afterwards	-3.55	0.242	-3.44	0.240	-3.56	0.241
,, Missing	0.00	NA	0.00	NA	0.00	NA
Bank A	-0.09	0.040	-0.18	0.035	-0.08^{ns}	0.041
Remarks 8,11,16,25	1.09	0.055	1.08	0.054	1.09	0.055
Remark 25	1.11	0.077	1.11	0.075	1.12	0.076
Sales data missing	0.85	0.120	0.85	0.115	0.86	0.120
Sales (log$_e$)	-0.04	0.005	-0.04	0.005	-0.04	0.005
Earnings/Sales	-0.25	0.038	-0.24	0.038	-0.25	0.038
Inventory/Sales	0.54	0.106	0.53	0.102	0.53	0.106
Loan/Asset	1.02	0.041	1.04	0.040	1.01	0.041
Output gap	-0.18	0.024	-0.19	0.010	NA	NA
Yield curve	-0.23	0.060	-0.26	0.021	NA	NA

Note: SE stands for standard errors of the estimates.
^{ns} stands for not significant at 2.5% level.

The covariance matrix, **D**, estimated through PQLD is a diagonal matrix with the diagonal elements being (0.3577, 0.2592, 0.2514, 0.1097, 0.2312, 0.4203). However, the **D** matrix estimated by PQLU is an unstructured matrix. The covariance matrix, **D**, is also estimated by using the realization of the random effects obtained in PQLD. The estimates of **D** through PQLD and PQLU are given in Table 2.

Table 2 Estimated covariance matrices of the random effects

D matrix estimated through PQLU						**D** matrix estimated through PQLD					
0.28	0.16	0.18	0.16	0.16	0.26	0.31	0.16	0.18	0.09	0.15	0.21
0.16	0.25	0.25	0.17	0.23	0.15	0.16	0.24	0.22	0.09	0.18	0.13
0.18	0.25	0.26	0.18	0.22	0.16	0.18	0.22	0.23	0.09	0.17	0.13
0.16	0.17	0.18	0.14	0.15	0.13	0.09	0.09	0.09	0.05	0.06	0.06
0.16	0.23	0.22	0.15	0.22	0.16	0.15	0.18	0.17	0.06	0.19	0.13
0.26	0.15	0.16	0.13	0.16	0.36	0.21	0.13	0.13	0.06	0.13	0.31

From Table 2 we see that the covariance parameters estimates are not much different between the two models except for those parameters regarding 4^{th} industry. The close similarity in the above covariance matrix is not surprising in this case since the Hessian matrix for $\mathbf{u}$ was pointwise very closed to zero.

The computational time is often a matter of great concern especially for the large data cases. The estimation of PQLD implemented in SAS 9.1 using %GLIMMIX macro (Wolfinger & O'Connell 1993) required 44 minutes (appr.) on a Pentium 4 PC (3.19 GHz processor, 0.99 GB RAM). On the contrary, the estimation of the fixed effects part of PQLU using GENMOD procedure of SAS 9.1 required only 1.33 minutes. The arrangement of SAS GENMOD outputs for further analysis took another 3.8 seconds. The random effects part and their covariance matrix estimation, implemented in R 2.2.0 using the author's self written R codes for 2PL, took another 27.8 minutes, including the time for importing the SAS outputs, saved in a text file, to R. Though the difference in time requirement for estimating PQLD and PQLU does not vary a lot, it should be kept in mind, while comparing the time requirements, that the PQLD estimated only 7 covariance parameters, including an additional over dispersion parameter, when the PQLU estimated 21 covariance parameters.

5 Concluding discussion

Approximate likelihood methods, *e.g.* PQL and h-likelihood, are widely criticized especially for binary response models (Engel 1998). However, we should note that the Laplace approximation used in PQL approaches is an asymptotic approximation (Evans & Swartz 1995). Therefore, they may not perform well for small sample cases and they should be judged on the basis of their large sample performances which the critics of PQL have hardly considered. Moreover, for large sample cases, simulation based computation, *e.g.* MCMC methods, for GLMM is awfully time consuming. High speed computers are becoming available but still there is a need for some approximate method which can handle large data sets within a reasonable time limit.

In Alam & Carling (2008), it is claimed, on the basis of simulation, that PQL and FE methods become numerically equal when the cluster size, n_{ikt}, is between 50 and 200. Here, an analytical expression has been presented in terms of $H_{h(\mathbf{u})}$ for such approximation to hold. Such condition can always be checked using the expression, $H_h\left(\widetilde{\mathbf{u}}\right)^{-1} = \left(\mathbf{Z}^T\widetilde{\mathbf{W}}\mathbf{Z}/\mathbf{a}\left(\phi\right) + \widehat{\mathbf{D}}^{-1}\right)^{-1}.$

The estimates of the random effects, $\widetilde{\mathbf{u}}$, are explained as the conditional expectation of the random effects given the data, y. Therefore, intuitively a confusion arises while estimating the marginal covariance matrix of $\mathbf{u}$ based on $\widetilde{\mathbf{u}}$ as to why such estimation procedure should hold. Here, an analytical reason is presented since $\widehat{E}\left(V\left(\mathbf{u}|\mathbf{Y}\right)\right)$ is the inverse of the negative of a Hessian and hence $E\left(V\left(\mathbf{u}|\mathbf{Y}\right)\right) \to 0$ as $n_{ij} \to \infty$ while $V\left(E\left(\mathbf{u}|\mathbf{Y}\right)\right)$ is given by $V\left(\widetilde{\mathbf{u}}\right).$

References

Alam, M. M. & Carling, K. (2008), 'Computationally feasible estimation of the covariance structure of the generalized linear mixed models (GLMM)', *Journal of Statistical Computation and Simulation* **78**(12), 1227–1237.

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

Broström, G. (2007), 'The glmmml package', URL: http://cran.r-project.org/doc/packages/glmmML.pdf. Last accessed: 23-11-2007.

Butler, J. S. & Moffitt, R. (1982), 'A computational efficient quadrature procedure for the one-factor multinomial probit model', *Econometrica* **50**(3), 761–764.

Carling, K., Rönnegård, L. & Rosczbach, K. (2004), Is firm interdependence within industries important for portfolio credit risk?, Working Paper Series 168, Sverige Riskbank.

Engel, B. (1998), 'A simple illustration of the failure of the PQL, IRREML and APHL as approximate ML methods for mixed models for binary data', *Biometrical Journal* **40**(2), 141–154.

Evans, M. & Swartz, T. (1995), 'Methods for approximating integrals in statistics with special emphasis on bayesian integration problems', *Statistical Science* **10**(3), 254–272.

Harville, D. A. (1997), *Matrix Algebra from a Statistician's Perspective*, Springer, New York.

Khuri, A. I. (2003), *Advanced Calculus with Application in Statistics*, Wiley, Hoboken, New Jersey.

Lam, K. F. & Lee, Y. W. (2004), 'Merits of modelling multivariate survival data using random effect proportional odds model', *Biometrical Journal* **46**(1), 331–342.

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

Lee, Y. & Nelder, J. A. (2006), 'Double hierarchical generalize linear models', *Journal of the Royal Statistical Society (C)* **55**(2), 1–29.

Lindstrom, M. J. & Bates, D. M. (1998), 'Newton-raphson and em algorithms for linear mixed-effects models for repeated-measures data', *Journal of the American Statistical Association* **43**(404), 1014–1022.

Littell, R. C., Milliken, G. A., Stroup, W. W. & D., W. R. (1996), *The SAS system for mixed models*, SAS Inst. Inc., Cary, North Carolina.

Maddala, G. S. . (1987), 'Limited dependent variable models using panel data', *Journal of Human Resources* **22**(3), 307–338.

Magnus, J. R. & Neudecker, H. (1999), *Matrix Differential Calculus with Applications in Statistics and Econometrics*, Wiley, New York.

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

McCulloch, C. E. & Searle, S. R. (2001), *Generalized Linear and Mixed Models*, Wiley.

Olsson, U. (2002), *Generalized Linear Models: An Applied Approach*, Studentlitterature, Lund.

Raudenbush, S. W., Yang, M. & Yosef, M. (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), 141–157.

Searle, S. R., Casella, G. & McCulloch, C. E. (1992), *Variance Components*, Wiley, New York.

Venables, W. N. & Ripley, B. (2002), *Modern Applied Statistics with S*, Springer, New York.

Wand, M. P. (2002), 'Vector differential calculus in statistics', *The American Statistician* **56**(1), 1–8.

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

Zhou, X.-H., Perkins, A. J. & Hui, S. L. (1999), 'Comparisons of software packages for generalized linear multilevel models', *The American Statistician* **53**(3), 282–290.

A Appendix

The calculations presented in this Appendix will ferquently make use of the following properties of the matrix differentiation.

$$\left.\begin{aligned}\partial|\mathbf{D}| = |\mathbf{D}|tr\left(\mathbf{D}^{-1}\partial\mathbf{D}\right)\\ \partial\mathbf{AD} = \mathbf{A}\partial\mathbf{D}\\ \partial tr\left(\mathbf{D}\right) = tr\left(\partial\mathbf{D}\right)\\ \partial\mathbf{D}^{-1} = -\mathbf{D}^{-1}\left(\partial\mathbf{D}\right)\mathbf{D}^{-1}\\ \partial vec\left(\mathbf{D}\right) = vec\left(\partial\mathbf{D}\right)\\ tr\left(\mathbf{A}^T\mathbf{B}\right) = vec\left(\mathbf{A}\right)^T vec\left(\mathbf{B}\right)\end{aligned}\right\} \tag{A-1}$$

where, **A** and **B** are the matrices of constants, ∂ denotes differential and the derivatives are taken w.r. to **D**.

Chain Rule: Let, h be a composite function such that $h\left(\mathbf{X}\right) = g\left(F\left(\mathbf{X}\right)\right)$ when $F\left(\mathbf{X}\right) = \mathbf{b}$ then $\mathcal{D}\left(h\left(\mathbf{X}\right)\right) = \left(\mathcal{D}g\left(\mathbf{b}\right)\right)\mathcal{D}F\left(\mathbf{X}\right)$, where $\mathcal{D}$ operator stands for the matrix differentiation *i.e.* $\mathcal{D}F\left(\mathbf{X}\right) = \frac{\partial}{\partial vec(\mathbf{X})}F\left(\mathbf{X}\right)$. See Magnus & Neudecker (1999) for proof.

For further detailed about the matrix differentiation, readers are referred to advanced texts in matrix algebra *e.g.* Harville (1997) and Magnus & Neudecker (1999).

A.1 Derivation of equation (9)

We have,

$$h(\mathbf{u}) = -l + \frac{1}{2}\mathbf{u}^T\mathbf{D}^{-1}\mathbf{u}$$

$$\Rightarrow h(\mathbf{u}) = \frac{-\mathbf{Y}^T(\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\mathbf{u}) + \mathbf{1}^T b(\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\mathbf{u})}{a(\phi)} - \mathbf{1}^T c(\mathbf{Y},\phi) + \frac{1}{2}\mathbf{u}^T\mathbf{D}^{-1}\mathbf{u}$$

$$\Rightarrow \partial h(\mathbf{u}) = \frac{-\mathbf{Y}^T\mathbf{Z}\partial\mathbf{u} + \mathbf{1}^T diag\left(b^{(1)}(\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\mathbf{u})\right)\mathbf{Z}\partial\mathbf{u}}{a(\phi)} + \mathbf{u}^T\mathbf{D}^{-1}\partial\mathbf{u}$$

$$\therefore \frac{\partial h(\mathbf{u})}{\partial\mathbf{u}} = \frac{-\mathbf{Y}^T\mathbf{Z} + \mathbf{1}^T diag\left(b^{(1)}(\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\mathbf{u})\right)\mathbf{Z}}{a(\phi)} + \mathbf{u}^T\mathbf{D}^{-1}$$

Again,

$$\partial^2 h(\mathbf{u}) = \frac{1}{a(\phi)}\partial b^{(1)}(\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\mathbf{u})^T\mathbf{Z}\partial\mathbf{u} + \partial\mathbf{u}^T\mathbf{D}^{-1}\partial\mathbf{u}$$

$$\Rightarrow \partial^2 h(\mathbf{u}) = \frac{1}{a(\phi)}\partial\mathbf{u}^T\mathbf{Z}^T diag\left(b^{(2)}(\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\mathbf{u})\right)\mathbf{Z}\partial\mathbf{u} + \partial\mathbf{u}^T\mathbf{D}^{-1}\partial\mathbf{u}$$

$$\Rightarrow \frac{\partial^2 h(\mathbf{u})}{\partial\mathbf{u}\partial\mathbf{u}^T} = \mathbf{Z}^T\mathbf{W}\mathbf{Z} + \mathbf{D}^{-1}$$

where, $\mathbf{W} = \frac{1}{a(\phi)} diag\left(b^{(2)}(\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\mathbf{u})\right)$ is known as diagonal weight matrix (McCullagh & Nelder 1989).

Now, assuming $a(\phi)=1$ a Newton-Raphson algorithm for calculating the maxima w.r.t. $\mathbf{u}$ is given by

$$\widetilde{\mathbf{u}}_{r+1} = \widetilde{\mathbf{u}}_r - H^{-1}_{h(\mathbf{u})}\frac{\partial h(\mathbf{u})}{\partial\mathbf{u}^T}|\mathbf{u} = \widetilde{\mathbf{u}}_r$$

$$\Rightarrow \widetilde{\mathbf{u}}_{r+1} = \widetilde{\mathbf{u}}_r - \left(\mathbf{Z}^T\widetilde{\mathbf{W}}_r\mathbf{Z} + \mathbf{D}^{-1}\right)^{-1}\left(-\mathbf{Z}^T\mathbf{Y} + \mathbf{Z}^T diag\left(b^{(1)}(\mathbf{X}\boldsymbol{\beta}+\mathbf{Z}\widetilde{\mathbf{u}}_r)\right) + \mathbf{D}^{-1}\widetilde{\mathbf{u}}_r\right)$$

$$\Rightarrow \left(\mathbf{Z}^T\widetilde{\mathbf{W}}_r\mathbf{Z} + \mathbf{D}^{-1}\right)\widetilde{\mathbf{u}}_{r+1} = \mathbf{Z}^T\widetilde{\mathbf{W}}_r\mathbf{Z}\widetilde{\mathbf{u}}_r + \mathbf{D}^{-1}\widetilde{\mathbf{u}}_r + \mathbf{Z}^T\mathbf{Y} - \mathbf{Z}^T\widetilde{\boldsymbol{\mu}}_r - \mathbf{D}^{-1}\widetilde{\mathbf{u}}_r$$

$$\Rightarrow \widetilde{\mathbf{u}}_{r+1} = \left(\mathbf{Z}^T\widetilde{\mathbf{W}}_r\mathbf{Z} + \mathbf{D}^{-1}\right)^{-1}\mathbf{Z}^T\widetilde{\mathbf{W}}_r\left(\mathbf{Z}\widetilde{\mathbf{u}}_r + \widetilde{\mathbf{W}}_r^{-1}(\mathbf{Y}-\widetilde{\boldsymbol{\mu}}_r)\right)$$

where, $\widetilde{\boldsymbol{\mu}}_r$ is $\boldsymbol{\mu}$ evaluated at $\mathbf{u} = \widetilde{\mathbf{u}}_r$. Now, denoting $\widetilde{\mathbf{W}}_r^{-1}(\mathbf{Y}-\widetilde{\boldsymbol{\mu}}_r) + \widetilde{\eta}_r = \mathbf{Y}^*$ we have

$$\widetilde{\mathbf{u}}_{r+1} = \left(\mathbf{Z}^T\widetilde{\mathbf{W}}_r\mathbf{Z} + \mathbf{D}^{-1}\right)^{-1}\mathbf{Z}^T\widetilde{\mathbf{W}}_r(\mathbf{Y}^* - \mathbf{X}\boldsymbol{\beta})$$

Note that, to minimize $h(\mathbf{u})$ is equivalent to maximize $-h(\mathbf{u})$. In other words, $\widetilde{\mathbf{u}}$ is obtained by maximizing the joint likelihood, $L(\boldsymbol{\beta}, \mathbf{D}, \mathbf{u}|\mathbf{Y})$, w.r.t. $\mathbf{u}$.

A.2 Derivation of equation (11)

From equation (10) we have

$$\ln\left(L(\boldsymbol{\beta},\ \mathbf{D}|\mathbf{Y})\right)=-\frac{1}{2}\ln\left(|\mathbf{D}|\right)-\frac{1}{2}\ln\left(|\mathbf{Z}^T\widetilde{\mathbf{W}}\mathbf{Z}/a\left(\phi\right)+\mathbf{D}^{-1}|\right)+\widetilde{l}-\frac{1}{2}\widetilde{\mathbf{u}}^T\mathbf{D}^{-1}\widetilde{\mathbf{u}}$$

$$\Rightarrow\partial\ln\left(L(\boldsymbol{\beta},\ \mathbf{D}|\mathbf{Y})\right)=-\frac{1}{2|\mathbf{D}|}\partial|\mathbf{D}|-\frac{1}{2}\partial\ln\left(|\mathbf{Z}^T\widetilde{\mathbf{W}}\mathbf{Z}/a\left(\phi\right)+\mathbf{D}^{-1}|\right)-\frac{1}{2}\widetilde{\mathbf{u}}^T\partial\left(\mathbf{D}^{-1}\right)\widetilde{\mathbf{u}}\qquad\text{(A-2)}$$

Using these results in (A-2) and the Chain rule we have

$$\begin{aligned}\partial\ln\left(L(\boldsymbol{\beta},\ \mathbf{D}|\mathbf{Y})\right)\ &=\ -\frac{1}{2|\mathbf{D}|}|\mathbf{D}|tr\left(\mathbf{D}^{-1}\partial\mathbf{D}\right)-\frac{1}{2}\frac{1}{|\mathbf{Z}^T\widetilde{\mathbf{W}}\mathbf{Z}/a\left(\phi\right)+\mathbf{D}^{-1}|}\left(|\mathbf{Z}^T\widetilde{\mathbf{W}}\mathbf{Z}/a\left(\phi\right)+\mathbf{D}^{-1}|\right)\\&\quad vec\left(\left(\mathbf{Z}^T\widetilde{\mathbf{W}}\mathbf{Z}/a\left(\phi\right)+\mathbf{D}^{-1}\right)^{-1}\right)^T\left(-\mathbf{D}^{-1}\otimes\mathbf{D}^{-1}\right)^T\partial vec\left(\mathbf{D}\right)+\frac{1}{2}\widetilde{\mathbf{u}}^T\mathbf{D}^{-1}\left(\partial\mathbf{D}\right)\mathbf{D}^{-1}\widetilde{\mathbf{u}}\end{aligned}$$

$$\begin{aligned}\Rightarrow\quad&\partial\ln\left(L(\boldsymbol{\beta},\ \mathbf{D}|\mathbf{Y})\right)=-\frac{1}{2}tr\left(\mathbf{D}^{-1}\partial\mathbf{D}\right)+\frac{1}{2}vec\left(\left(\mathbf{Z}^T\widetilde{\mathbf{W}}\mathbf{Z}/a\left(\phi\right)+\mathbf{D}^{-1}\right)^{-1}\right)^T\left(\mathbf{D}^{-1}\otimes\mathbf{D}^{-1}\right)^T\partial vec\left(\mathbf{D}\right)\\&+\frac{1}{2}tr\left(\mathbf{D}^{-1}\widetilde{\mathbf{u}}\widetilde{\mathbf{u}}^T\mathbf{D}^{-1}\left(\partial\mathbf{D}\right)\right)\end{aligned}$$

$$\begin{aligned}\Rightarrow\quad&\partial\ln\left(L(\boldsymbol{\beta},\ \mathbf{D}|\mathbf{Y})\right)=-\frac{1}{2}vec\left(\mathbf{D}^{-1}\right)^T vec\left(\partial\mathbf{D}\right)+\frac{1}{2}vec\left(\left(\mathbf{Z}^T\widetilde{\mathbf{W}}\mathbf{Z}/a\left(\phi\right)+\mathbf{D}^{-1}\right)^{-1}\right)^T\\&\left(\mathbf{D}^{-1}\otimes\mathbf{D}^{-1}\right)^T\partial vec\left(\mathbf{D}\right)+\frac{1}{2}vec\left(\mathbf{D}^{-1}\widetilde{\mathbf{u}}\widetilde{\mathbf{u}}^T\mathbf{D}^{-1}\right)^T vec\left(\partial\mathbf{D}\right)\end{aligned}$$

$$\begin{aligned}\Rightarrow\quad&\frac{\partial\ln\left(L(\boldsymbol{\beta},\ \mathbf{D}|\mathbf{Y})\right)}{\partial vec\left(\mathbf{D}\right)^T}=-\frac{1}{2}vec\left(\mathbf{D}^{-1}\right)+\frac{\left(\mathbf{D}^{-1}\otimes\mathbf{D}^{-1}\right)vec\left(\left(\mathbf{Z}^T\widetilde{\mathbf{W}}\mathbf{Z}/a\left(\phi\right)+\mathbf{D}^{-1}\right)^{-1}\right)}{2}\\&+\frac{1}{2}vec\left(\mathbf{D}^{-1}\widetilde{\mathbf{u}}\widetilde{\mathbf{u}}^T\mathbf{D}^{-1}\right)\end{aligned}$$

A.3 Derivation of equation (12)

Assuming $a\left(\phi\right)=1$, and equating $\frac{\partial\ln(L(\boldsymbol{\beta},\ \mathbf{D}|\mathbf{Y}))}{\partial vec(\mathbf{D})^T}=0$ and using the following properties of kronicker product

$$vec\left(\mathbf{ABC}\right)=\left(\mathbf{C}^T\otimes\mathbf{A}\right)vec\left(\mathbf{B}\right)$$

and

$$\left(\mathbf{A}\otimes\mathbf{B}\right)^{-1}=\mathbf{A}^{-1}\otimes\mathbf{B}^{-1}$$

we have from A.1

$$-\frac{1}{2}vec\left(\mathbf{D}^{-1}\right)+\frac{1}{2}\left(\mathbf{D}^{-1}\otimes\mathbf{D}^{-1}\right)vec\left(\left(\mathbf{Z}^{T}\widetilde{\mathbf{W}}\mathbf{Z}+\mathbf{D}^{-1}\right)^{-1}\right)+\frac{1}{2}vec\left(\mathbf{D}^{-1}\widetilde{\mathbf{u}}\widetilde{\mathbf{u}}^{T}\mathbf{D}^{-1}\right)=\mathbf{0}$$

$$\Rightarrow -\frac{1}{2}vec\left(\mathbf{D}^{-1}\right)+\frac{1}{2}\left(\mathbf{D}^{-1}\otimes\mathbf{D}^{-1}\right)vec\left(\left(\mathbf{Z}^{T}\widetilde{\mathbf{W}}\mathbf{Z}+\mathbf{D}^{-1}\right)^{-1}\right)+\frac{1}{2}\left(\mathbf{D}^{-1}\otimes\mathbf{D}^{-1}\right)vec\left(\widetilde{\mathbf{u}}\widetilde{\mathbf{u}}^{T}\right)=\mathbf{0}$$

$$\Rightarrow -\frac{1}{2}\left(\mathbf{D}^{-1}\otimes\mathbf{D}^{-1}\right)^{-1}vec\left(\mathbf{D}^{-1}\right)+\frac{1}{2}vec\left(\left(\mathbf{Z}^{T}\widetilde{\mathbf{W}}\mathbf{Z}+\mathbf{D}^{-1}\right)^{-1}\right)-\frac{1}{2}vec\left(\widetilde{\mathbf{u}}\widetilde{\mathbf{u}}^{T}\right)=\mathbf{0}$$

$$\Rightarrow -vec\left(\mathbf{D}\right)+vec\left(\left(\mathbf{Z}^{T}\widetilde{\mathbf{W}}\mathbf{Z}+\mathbf{D}^{-1}\right)^{-1}\right)+vec\left(\widetilde{\mathbf{u}}\widetilde{\mathbf{u}}^{T}\right)=\mathbf{0}$$

$$\Rightarrow -\mathbf{D}+\left(\mathbf{Z}^{T}\widetilde{\mathbf{W}}\mathbf{Z}+\mathbf{D}^{-1}\right)^{-1}+\widetilde{\mathbf{u}}\widetilde{\mathbf{u}}^{T}=\mathbf{0}$$

Equation (12) can easily be obtined from the last equation given above.