Feasible estimation of generalized linear mixed models (GLMM) with weak dependency between groups

Md Moudud Alam*

October 6, 2010

Abstract

This paper presents a two-step pseudo likelihood estimation technique for generalized linear mixed models with the random effects being correlated between groups. The core idea is to deal with the intractable integrals in the likelihood function by multivariate Taylor's approximation. The accuracy of the estimation technique is assessed in a Monte-Carlo study. An application of it with a binary response variable is presented using a real data set on credit defaults from two Swedish banks. Thanks to the use of two-step estimation technique, the proposed algorithm outperforms conventional pseudo likelihood algorithms in terms of computational time.

Mathematics Subject Classification: Primary 62J12; Secondary 65C60

Keywords: PQL, Laplace approximation, interdependence, cluster errors, credit risk model.

*Dalarna University and Örebro University; contact: Dalarna University, School of Technology and Business Studies, 781 88 Borlange, Sweden; e-mail: maa@du.se

1 Introduction

This paper presents a computationally feasible estimation technique for the generalized linear mixed models (GLMM) with the random effects being correlated between groups. There are computational procedures to fit GLMM using (approximate) likelihood methods such as Pseudo likelihood (Wolfinger & O'Connell 1993, PL) and Penalized quasi likelihood (Breslow & Clayton 1993, PQL), h-likelihood (Lee & Nelder 2006) and Markov-chain Monte-Carlo (MCMC). However, they are computationally heavy. Especially for large data sets, the MCMC procedures are awfully time consuming. The PL, PQL and h-likelihood are faster than MCMC, but still slow for large data sets. This paper presents a general, in that it works for both the independent and the correlated random effects, algorithm to estimate the GLMM parameters using a two-step estimation procedure.

To estimate the fixed effects parameters, we maximize an approximate marginal likelihood. The marginal likelihood of a GLMM is explained as the expected conditional likelihood, thus the expectation is taken over the first order Taylor's approximation of the conditional likelihood. Therefore, the fixed effects parameters can be estimated by a generalized linear model (McCullagh & Nelder 1989, GLM) procedure. For the estimation of the variance and covariance parameters, a Laplace approximation to the marginal likelihood is applied. The estimation approach can be regarded a generalization of the PL or PQL approach for correlated random effects and we refer to it as the two-step pseudo likelihood (2PL) approach.

Since we approximate the true likelihood, the 2PL estimator may not be optimal. In a simulation study we show that the 2PL does provide precise estimates of the model parameters, especially when the variance and covariance parameters are small in absolute value and the sample size is large. In a real data example, we find that the 2PL method works much faster than the PQL method and that it produces reasonable estimates of the model parameters.

The paper has been organized in the following way. Section two derives the 2PL estimation procedure. Section three discusses the properties of 2PL estimate. Section four provides an application of the method to credit risk modelling using a real data set obtained from two major Swedish banks. Section four also discusses the computational advantage of 2PL over PQL. Section five concludes.

2 Derivation of the 2PL procedure

For a GLMM, the joint likelihood function of the fixed effects parameters, $\boldsymbol{\beta}$, conditional dispersion parameter, ϕ, and the covariance parameter of the random effects, $\mathbf{D}$, is given as

$$L(\boldsymbol{\beta},\phi,\mathbf{D}|\mathbf{Y}) = \int \prod_{t=1}^{T} L\left(\boldsymbol{\beta},\phi,\mathbf{D}|\mathbf{Y},\mathbf{u}_t\right) f\left(\mathbf{u}_t\right) d\mathbf{u}_t \tag{1}$$

where $\mathbf{Y} = \{y_{ijt}\}$ $(i = 1, 2, .., n_{jt},\ j = 1, 2, .., k,\ t = 1, 2, ..., T)$ is the vector of the response variable, $\mathbf{u}_t = (u_{1t},\ u_{2t}, ..., u_{kt})^T$ is the vector of the random effects with $\mathbf{u}_t \perp \mathbf{u}_{t'}\ \forall t \neq t'$, k is the number of groups (or cross sectional clusters), T is the occasions (or time) and n_{jt} is the number of observations at time t and cluster j. For the time being, we drop the $t-$dimension but we will come back whenever it is relevant.

Under the standard assumption of the GLMM, after omitting the $t-$dimension, the distribution of the response variable, y_{ij}, given the random effects follows exponential family of distributions. Often, $\mathbf{u}$ is assumed *i.i.d* normal which renders the multivariate integral in equation (1) a product of k univariate integrals. However, in this paper the $\mathbf{u}$ is a multivariate normal variate with mean vector, $\mathbf{0}$, and an unstructured covariance matrix, $\mathbf{D}$. Examples of the correlated random effects, as above, arises in the genetic relations (see *e.g.* Searle, Casella & McCulloch (1992) pp. 383) and in the inter-industry default correlations in credit risk modeling (see *e.g.* Alam & Carling (2008)).

Under the above GLMM assumptions, the conditional likelihood, $L(\boldsymbol{\beta},\phi,\ \mathbf{D}|\mathbf{Y},\mathbf{u})$, with a canonical link is expressed as

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

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

where $\eta_{ij} = \mathbf{X}_{ij}\boldsymbol{\beta} + \mathbf{Z}_{ij}\mathbf{u}$ is called the linear predictor, $b(.)$ is called the cumulant function, $a(\phi)$ 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 the link function, $g(.)$. Using matrix notations, equation (2) can be put as

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

So, equation (1) can be re-expressed as

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

With $\mathbf{u}$ multivariate normal $f(\mathbf{u})$ is

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

From equation (4), the marginal likelihood of $\boldsymbol{\beta}$, ϕ and $\mathbf{D}$ can be interpreted as an expectation, $E(\exp[l])$, with respect to the multivariate normal distribution of $\mathbf{u}$. The multivariate version of the Taylor's expansion of the function $m(\mathbf{u}) = \exp[l]$ around the marginal mean of $\mathbf{u}$ (being 0) gives

$$L(\boldsymbol{\beta},\phi,\ \mathbf{D}|\mathbf{Y}) = E\left(m(\mathbf{u})\right) \approx m\left(\mathbf{0}\right) + \mathbf{0} + \frac{1}{2}E\left\{\mathbf{u}^T m^{(2)}\left(\mathbf{u}\right)_{\mathbf{u}=\mathbf{0}}\mathbf{u}\right\} \tag{5}$$

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

Regarding the estimation of the fixed effects parameters, already the second order term in equation (5) is fairly flat in $\boldsymbol{\beta}$ and commonly ignored in the PQL methods (Breslow & Clayton 1993). In the same spirit, ignoring the second order term, the likelihood for $\boldsymbol{\beta}$ becomes the likelihood of a GLM that is a model without random effects. In other words, the first order Taylor's approximation of the conditional likelihood around $\mathbf{u} = E\left(\mathbf{u}\right)$ suggests estimating the $\boldsymbol{\beta}$ parameters by a simple GLM.

We do not, however, use the same likelihood as presented in equation (5) for the estimation of the covariance parameters. This is, firstly, because there is no guarantee that the higher order terms that we ignored in (5) are also flat in $\mathbf{D}$. Secondly, the simplification of the Taylor's expansion of $m\left(\mathbf{u}\right)$ 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 (Breslow & Clayton 1993). From equation (4) we have

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

$$\Rightarrow L(\boldsymbol{\beta},\phi,\ \mathbf{D}|\mathbf{Y}) = \frac{|\mathbf{D}^{-1}|^{1/2}}{(2\pi)^{k/2}}\int \exp\left[-\left(-l + \frac{1}{2}\mathbf{u}^T\mathbf{D}^{-1}\mathbf{u}\right)\right]d\mathbf{u} \tag{6}$$

Let $h\left(\mathbf{u}\right) = -l + \frac{1}{2}\mathbf{u}^T\mathbf{D}^{-1}\mathbf{u}$ and assume that $h\left(\mathbf{u}\right)$ has a single minima at $\mathbf{u} = \widetilde{\mathbf{u}}$**.** Then by applying the multivariate version of the Laplace approximation (Evans & Swartz 1995) in equation (6) we have

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

$$\Rightarrow \ln\left(L(\boldsymbol{\beta},\phi,\ \mathbf{D}|\mathbf{Y})\right) = -\frac{1}{2}\ln\left(|\mathbf{D}|\right) - \frac{1}{2}\ln\left\{|H_h\left(\widetilde{\mathbf{u}}\right)|\right\} - h\left(\widetilde{\mathbf{u}}\right) \tag{7}$$

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

$$\widetilde{\mathbf{u}}_{r+1} = \left(\mathbf{Z}^T\widetilde{\mathbf{W}_r}\mathbf{Z} + \mathbf{D}^{-1}\right)^{-1}\mathbf{Z}^T\widetilde{\mathbf{W}}_r\left(\mathbf{Y}^* - \mathbf{X}\boldsymbol{\beta}\right) \tag{8}$$

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 r^{th} iteration when $\mathbf{u}=\widetilde{\mathbf{u}}_r$. 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 given the data. However, it remains to estimate the unknown $\mathbf{D}$ matrix by maximizing (7).

From equation (7), we have

$$\ln\left(L(\boldsymbol{\beta},\phi,\ \mathbf{D}|\mathbf{Y})\right) = -\frac{1}{2}\ln\left(|\mathbf{D}|\right) - \frac{1}{2}\ln\left\{|\mathbf{Z}^T\widetilde{\mathbf{W}}\mathbf{Z}/a(\phi)+\mathbf{D}^{-1}|\right\} + \widetilde{l} - \frac{1}{2}\widetilde{\mathbf{u}}^T\mathbf{D}^{-1}\widetilde{\mathbf{u}} \tag{9}$$

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

After taking matrix differentiation of (9) and simplifying (detailed calculation is presented in Appendix A.2) it can be shown that

$$\begin{aligned}\frac{\partial \ln\left(L(\boldsymbol{\beta},\phi,\ \mathbf{D}|\mathbf{Y})\right)}{\partial vec(\mathbf{D})} &= -\frac{1}{2}vec\left(\mathbf{D}^{-1}\right)^T + \frac{1}{2}vec\left\{\left(\mathbf{Z}^T\widetilde{\mathbf{W}}\mathbf{Z}/a(\phi)+\mathbf{D}^{-1}\right)^{-1}\right\}^T \\ &\quad \left(\mathbf{D}^{-1}\otimes\mathbf{D}^{-1}\right) + \frac{1}{2}vec\left(\mathbf{D}^{-1}\widetilde{\mathbf{u}}\widetilde{\mathbf{u}}^T\mathbf{D}^{-1}\right)^T\end{aligned} \tag{10}$$

Equation (10) can be used to find a direct solution of $\mathbf{D}$ by equating $\frac{\partial \ln(L(\boldsymbol{\beta},\ \mathbf{D}|\mathbf{Y}))}{\partial vec(\mathbf{D})} = 0$ for given $\widehat{\boldsymbol{\beta}}$, $\widetilde{\mathbf{u}}$ and $a(\phi)$. For binomial and Poisson distributions, $a(\phi) = 1$ which gives the following equation to solve for $\mathbf{D}$ (see Appendix A.3 for detailed calculation)

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

$$\Rightarrow \widehat{\mathbf{D}} = \mathbf{H}^{-1}_{h(\mathbf{u})} + \widetilde{\mathbf{u}}\widetilde{\mathbf{u}}^T \tag{11}$$

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

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

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

- **i)** Initialize $\mathbf{u}$ and $\mathbf{D}$. Replace $\boldsymbol{\beta}$ with its estimate from step 1. Use corrections for $\boldsymbol{\beta}$ unless $|\mathbf{D}|$ is very small.
- **ii)** Update $\mathbf{u}$ by iteratively solving $\widetilde{\mathbf{u}}_{r+1} = \left(\mathbf{Z}^T\widetilde{\mathbf{W}_r}\mathbf{Z}+\mathbf{D}^{-1}\right)^{-1}\mathbf{Z}^T\widetilde{\mathbf{W}}_r\left(\mathbf{Y}^*-\mathbf{X}\boldsymbol{\beta}\right)$.
- **iii)** Update $\mathbf{D}$ with $\widehat{\mathbf{D}} = \mathbf{H}^{-1}_{h(\mathbf{u})} + \widetilde{\mathbf{u}}\widetilde{\mathbf{u}}^T$; for T independent realization of $\mathbf{u}_t$ it is $\widehat{\mathbf{D}} = \frac{1}{T}\sum_t\left(\mathbf{H}^{-1}_{h(\mathbf{u}_t)} + \widetilde{\mathbf{u}}_t\widetilde{\mathbf{u}}_t^T\right)$.

iv) Stop if $\mathbf{D}$ converges, otherwise go to (ii).

Equations (8) and (11) 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 the likelihood based algorithms, such as PQL and double extended quasi likelihood (Lee & Nelder 2003, DEQL), somehow relies on iterative procedure for jointly estimating all the model parameters, including fixed effects and the variance components, while the proposed 2PL algorithm estimates the 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 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 complicated and are avoided in this paper.

3 Large sample properties

Equation (5) reveals that $\widehat{\boldsymbol{\beta}}_{2PL}$, the 2PL estimator of $\boldsymbol{\beta}$, is the maximum likelihood estimator up to a first order Taylor's approximation of the conditional likelihood. The error term in (5) is given by

$$E = \sum_{q=2}^{\infty} \frac{1}{k!} E \left\{ \left[\bigotimes^{q-1} \mathbf{u}^T \right] m^{(q)}(\mathbf{u})_{\mathbf{u}=\mathbf{0}} \mathbf{u} \right\} \tag{12}$$

In the expression all terms being odd are 0. The second order error term is $\frac{1}{2} tr\left(m^{(2)}(\mathbf{u})_{\mathbf{u}=\mathbf{0}} \mathbf{D}\right)$. Thus $\widehat{\boldsymbol{\beta}}_{2PL}$ is the MLE if $E = 0$ or it converges to the MLE if $\lim_{n\to\infty} \frac{\partial}{\partial \boldsymbol{\beta}} tr\left(m^{(2)}(\mathbf{u})_{\mathbf{u}=\mathbf{0}} \mathbf{D}\right) \to 0$. In some trivial cases this occurs.

1. If $\mathbf{D} = \mathbf{0}$ which means that neither the observations nor the groups are not correlated.
2. If $m^{(k)}(\mathbf{u})$ is constant as a function of $\boldsymbol{\beta}$, which can be verified for a practical problem in hand.
3. For logit and probit GLMM under $H_0 : \boldsymbol{\beta} = \mathbf{0}$ which is not very interesting other than for hypothesis testing.

Otherwise, $\widehat{\boldsymbol{\beta}}_{2PL}$ is the MLE with an order of accuracy $O(\mathbf{D})$.

A second line of argument for the consistency of $\widehat{\boldsymbol{\beta}}_{2PL}$ can be provided from the generalized estimation equations' (GEE) view point (Liang & Zeger 1986). Given that the mean of $\mathbf{Y}$ is expressed as $g(\mathbf{X}\boldsymbol{\beta}^*)$ the consistency of $\widetilde{\boldsymbol{\beta}}^*$ estimated through a simple GLM is established in Liang & Zeger (1986). The marginal model parameter, $\boldsymbol{\beta}^*$, however, is not always identical to the conditional model parameter $\boldsymbol{\beta}$ and should be interpreted differently. A relation between $\boldsymbol{\beta}^*$

and $\boldsymbol{\beta}$ is given in Zeger, Liang & Albert (1988) e.g. for logit model, $\log\left(\frac{\mu_{ij}}{1-\mu_{ij}}\right) = a\left(\mathbf{D}\right) X_{ij}\boldsymbol{\beta}^{*}$ where, $a\left(D\right) = |\mathbf{I} + c^{2}\mathbf{D}Z_{ij}Z_{ij}^{T}|^{-\frac{K}{2}}$ and $c = \frac{16\sqrt{3}}{15\pi}$. Similar expressions are also available for other GLMMs (Zeger et al. 1988). Therefore, when $a\left(D\right)$ is large, the 2PL estimator of $\boldsymbol{\beta}$ can be interpreted as the marginal mean parameter and a correction is needed in calculating $\hat{\eta}_{ij}$ which is used for the calculations of $\widetilde{\mathbf{u}}$ and $\widehat{\mathbf{D}}$ in the second step in the 2PL. In applications it is often the case that the variance components turn out to be small and consequently the above correction is negligible (see section 4). It is shown (Zeger et al. 1988) that, for a logistic model, the above adjustment gives an estimate of $\boldsymbol{\beta}^{*}$ with a proximity of 2% around the true GEE estimates even with $|\mathbf{D}| = \mathbf{4}$, which is very large.

Given $\boldsymbol{\beta}$ and $\mathbf{D}$, $\widetilde{\mathbf{u}} = E\left(\mathbf{u}|\mathbf{Y}\right)$. Thus given a consistent estimator of $\boldsymbol{\beta}$ and $\mathbf{D}$, $\widetilde{\mathbf{u}}$ is a consistent predictor for $E\left(\mathbf{u}|\mathbf{y}\right)$. On the other hand, given a consistent estimator for $\boldsymbol{\beta}$ and $\mathbf{u}$, the 2PL estimator for $\mathbf{D}$ is equivalent to PQL estimator which in turn is known to be biased for large $\mathbf{D}$ and with relatively small number of observations per group (Lee & Nelder 2003). However, for the elements of $\mathbf{D}$ being not very large, the PQL estimator is found, through simulations and application to real data sets, to be useful (Breslow & Clayton 1993). In special cases, when the log-likelihood is a quadratic function in $\mathbf{u}$, the PQL objective function is the exact marginal likelihood. In the other cases, the consistency arguments of $\mathbf{D}$ is subject to the appropriateness of the quadratic approximation to the log-likelihood.

In order to assess the preciseness of 2PL estimator, without any correction, we conduct a small simulation study. In the simulation we use a logistic mixed model where, $E\left(y_{ijt}|u_{j}\right) = \mu_{ijt}$, $\log\left(\frac{\mu_{ijt}}{1-\mu_{ijt}}\right) = \eta_{ijt} = \beta_{0} + \beta_{1}x_{ijt}$ and $\left(u_{1t}, u_{2t}, ..., u_{kt}\right)^{T} = \mathbf{u}_{t} \sim N\left(\mathbf{0}, \mathbf{D}\right)$. In practical implementation we use $T = 10$ and 20, $k = 7$, $n_{jt} = 10, 50$ and 100, $\beta_{0} = 0$, $\beta_{1} = 0.5$, $x_{ijt} \sim N\left(0, 1\right)$ and the covariance matrix of $\mathbf{u}_{t}$ being $\mathbf{D}$ and $2\mathbf{D}$ where the lower triangular elements of $\mathbf{D}$ are as follows.

$$\mathbf{D} = \begin{bmatrix} 0.19 & & & & & & \\ 0.14 & 0.24 & & & & & \\ 0.15 & 0.17 & 0.28 & & & & \\ 0.17 & 0.16 & 0.21 & 0.23 & & & \\ 0.12 & 0.10 & 0.06 & 0.20 & 0.52 & & \\ 0.21 & 0.13 & 0.13 & 0.17 & 0.16 & 0.26 & \\ 0.14 & 0.18 & 0.15 & 0.15 & 0.09 & 0.12 & 0.19 \end{bmatrix}$$

The simulation settings matches the simulations, with unstructured $Cov\left(\mathbf{u}_{t}\right)$, presented in Alam & Carling (2008). The performances of the fixed effects (FE) and PQL approaches were studied using probit and Poisson mixed models in Alam & Carling (2008). The results of this simulation study is therefore comparable with those in Alam & Carling (2008). We also use the same measures namely the mean sum of squared error (MSE) of β, $MSE\left(\beta\right) = \frac{1}{R}\sum_{r=1}^{R}\left(\widehat{\beta}_{r} - \beta\right)^{2}$, and the absolute relative error (ARE) of $\mathbf{D}$ where, $ARE\left(\mathbf{D}\right) = \frac{\sum_{r}\sum_{l=1}^{k(k+1)/2}|d_{l} - \widehat{d}_{lr}|}{R\sum_{l=1}^{k(k+1)/2}|d_{l}|}$ (d_{l} is the l^{th} lower triangular element of $\mathbf{D}$ and $\widehat{d}_{lr}$ is its estimate at r^{th} Monte-Carlo replication). We

k	T	n_{jt}	$\mathbf{D}$		$2\mathbf{D}$	
			$MSE(\beta)$	$RAE(\mathbf{D})$	$MSE(\beta)$	$RAE(\mathbf{D})$
7	10	10	0.008	1.01	0.009	0.76
		50	0.002	0.55	0.009	0.49
		200	0.001	0.47	0.003	0.44
	20	10	0.004	0.72	0.006	0.52
		50	0.001	0.40	0.004	0.36
		200	0.001	0.34	0.003	0.31

Table 1: MSE and RAE of 2PL Estimator for a Logistic Mixed Model

summarize our results in 200 iteration that is $R = 200$. The simulation results are given in Table 1.

Table 1 show that for an unstructured covariance matrix of the random effects $\mathbf{D} = \{d_{lm}\}$ with $\max\{d_{lm}\} < 0.53$ the fixed effects are almost consistently estimated (see the decreasing MSE with increasing sample sizes in Table 1). Though the MSE of β does not vanish at $n_{jt} = 200$ and $t = 20$, it is still very small and we observe a declining trend in the MSE as the sample sizes (n_{jt}) increase. The MSE's with $Cov(\mathbf{u}_t) = 2\mathbf{D}$ are larger than those with $Cov(\mathbf{u}_t) = \mathbf{D}$, but still they are not very large ($\max\left\{MSE\left(\widehat{\beta}\right)\right\} \leq 0.009$). The RAEs on the other hand are still large, compared to the MSEs of β, at $n_{jt} = 200$. Alam and Carling (2008) conducted a simulation where the probit link, $n_{jt} = 200$, $T = 20$, $Cov(\mathbf{u}_t) = \mathbf{D}$ and with all other parameters the same as the current simulation produced the RAE of fixed effects (FE) and mixed effects (PQL) approaches 0.33 and 0.31 respectively. Thus the RAE of the 2PL estimator at the same setting but with logistic link (RAE=0.34, Table 1) is slightly higher than, though insignificantly, those reported in Alam and Carling (2008). With $Cov(\mathbf{u}_t) = 2\mathbf{D}$, the RAE of 2PL at the same situation is 0.31 (see Table 1) which again comply with the results of Alam and Carling (2008). Thus we can conclude that with large sample sizes that is $n_{jt} \to \infty$, all the three methods, 2PL, FE and ME, provide the same results.

4 Application with credit default data

This section presents an application of the 2PL method with a real data set on credit defaults, collected from two major Swedish banks. This data set was first analyzed by Carling, Rönnegård & Roszbach (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 aim of the research was to derive a credit risk model by incorporating industry specific default correlation, thereby accounting for systematic risk. 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 whereas in this paper industries are defined by merging the SNI industries[1], at the first two digits level, in the way that closely resembles their industry definition. Since it was not possible to re-construct exactly the industry definition presented in Carling et al. (2004) by merging SNI industries, the industry definition used in this paper differs slightly from theirs. 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 for some of the industries, the frequency of defaults in some quarters are close to none which makes the GLM estimation impossible. For that reason, the number of industries was reduced from 7 to 6.

We propose a logistic mixed model for the credit default as

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

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

and

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

where, $k = 6$ (number of industries), $T = 25$ (number of quarters), $\mathbf{x}_{ijt}$ is the $(ijt)^{th}$ row of the observed design matrix, $\boldsymbol{\beta}$ is the vector of fixed effects parameters, $\mathbf{z}_{ijt}$ is the $(ijt)^{th}$ row of the design matrix associated with the random effects and contains a 1 in its j^{th} position and 0 otherwise and $\mathbf{u}_t$ is an *iid* realization of the random effect $\mathbf{u}$ at quarter t.

The model presented in Carling et al. (2004) was a complementary log-log mixed model with independent u_{kt}. To ensure comparability, Carling et al. (2004)'s model is re-analyzed with the logistic link and with the re-defined 6 industries. From here onwards, the logistic model with diagonal $\mathbf{D}$ will be denoted as PQLD which matches 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 the fixed effects is also estimated and that method is denoted FE (see also Alam & Carling (2008)).

The fixed effects parameter estimates from the three approaches are given in the Table 2. Regarding the statistical test of significance for the fixed effects parameters, PQLD uses t-test implemented through %GLIMMIX macro (Littell, Milliken, Stroup & D. 1996) in SAS 9.1 while PQLU and FE use Wald Chi-squared test implemented through SAS GENMOD procedure (Olsson 2002).

Table 2 shows that most of the fixed effects parameter estimates are similar between the three approaches, except for the coefficient associated with "Bank A". The big difference in estimates are found for the coefficients whose estimates are insignificant (see variables and their respective

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

Covariates	Models and Approaches					
	PQLD		PQLU		FE	
	Estimate	Std. Er.	Estimate	Std. Er.	Estimate	Std. Er.
Intercept	-4.44	0.214	-4.34	0.175	-3.90	0.378
Duration						
Credit Survived 1 Yr.	-0.18	0.137	-0.30	0.132	-0.18†	0.136
Credit Survived 2 Yr.	0.03†	0.137	0.05†	0.132	0.02†	0.137
Credit Survived 3 Yr.	0.13†	0.138	-0.16†	0.133	0.17†	0.138
Credit Survived 4 Yr.	0.28†	0.136	0.22†	0.132	0.28†	0.135
Credit Survived 5+ Yr.	0.30†	0.148	0.10†	0.141	0.31†	0.148
Loan specific						
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
Missing data indicators						
Account. data complete	-2.60	0.090	-2.53	0.088	-2.61	0.090
Acc. dat. Reported previously	0.73	0.087	0.74	0.084	0.73	0.087
Acc. dat. Reported afterward	-3.55	0.242	-3.44	0.240	-3.56	0.241
Acc. dat. Missing	0.00	NA	0.00	NA	0.00	NA
Sales data missing	0.85	0.120	0.85	0.115	0.86	0.120
Bank						
Bank A	-0.09	0.040	-0.18	0.035	-0.08†	0.041
Accounting						
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
Credit bureau remarks						
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
Macro-economic						
Output gap	-0.18	0.024	-0.19	0.010	NA	NA
Slope of yield curve	-0.23	0.060	-0.26	0.021	NA	NA

†*not significant at* 2.5% *level.*

No. of observations = 950693.

Table 2: A comparison of the fixed effects parameter estimates from PQLU, PQLD and FE approaches.

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

Table 3: Estimated covariance matrix of the random effects.

coefficient estimates in Table 2). This indicates that the fixed effects parameter estimates are not much sensitive to these three approaches.

The covariance matrix**,** estimated through PQLD is a diagonal matrix with the diagonal elements being $diag\{\mathbf{D}\}$ =(0.3577, 0.2592, 0.2514, 0.1097, 0.2312, 0.4203). The **D** matrix estimated by PQLU as an unstructured matrix and the estimates are presented in table 3. The covariance matrix is also estimated by using the predicted realization of the random effects obtained in PQLD and the estimates are also given in Table 3.

From Table 3 we see that the covariance parameters estimates are not much different between the two models except for the parameters regarding the fourth industry. The similarity in the estimates of the covariance matrix comes as no surprise in this case since the inverse of the Hessian matrix for **u** was pointwise very close to zero.

The estimation of PQLD implemented in SAS 9.1 using %GLIMMIX macro (Wolfinger & O'Connell 1993) required about 44 minutes 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 and reduced the computational time by about10 minutes. The 2PL estimation with the fixed effects parameters estimated through the FE approach took only about 14 minutes, thus reducing the computational time by 30 minutes compared to the PQL approach. However, it should be noted that in the FE approach we could not include two interesting variables, the output gap and the slope of the yield curve, in the model due to them not varying within industry and quarter. Therefore, the covariance parameters estimates of the FE approach may not be comparable with those in the PQL approach.

5 Conclusion

In the literature, several approximate likelihood methods, such as PQL and DEQL, are already available and they work well with simple GLMMs and with moderate sizes of data. However, with large sample sizes and complicated correlation structures of the random effects, such as the one presented in section 4, existing procedures are computationally too heavy. For those cases, we propose the 2PL approach which is computationally faster than the existing procedures in the more simple cases and works also for complicated ones. It is worth noting that we have had several attempts to estimate the desired credit risk model (see PQLU in section 4) by using %GLIMMIX (in SAS) and lmer (in R) which has gone in vain as the procedures did not converge in several hours. Hence, in such cases, we believe that the 2PL is the only available option to estimate these models in a reasonable time.

Through a simulation study we show that the 2PL estimates are reasonably precise, at least with the random effects having a small variance. How small should the covariance matrix of the random effects be to ensure the safe applicability of the 2PL? We do not have a general answer to that question yet. For a particular problem where 2PL is the only option, we suggest carrying out a small simulation study to check if the 2PL approach is reasonable in that situation.

In cases where the 2PL estimator may be imprecise due to large variance of the random effects, the fixed effects (FE) approach (Alam & Carling 2008) may be applied. The core idea of the FE approach is to estimate the $\mathbf{D}$ matrix by using the realizations of the random effects. Alam & Carling (2008) shows by simulations that with large cluster sizes along with a large number of clusters the $\mathbf{D}$ matrix may be estimated consistently by the FE approach. As a side-product of this paper we gave an analytical expression, during the derivation of $\widehat{\mathbf{D}}$, which can be used to check the preciseness of the FE approach. The condition can always be checked by using the expression, $H_h^{-1}(\widetilde{\mathbf{u}}) = \frac{1}{T}\left(\mathbf{Z}^T\widetilde{\mathbf{W}}\mathbf{Z}/\mathbf{a}(\phi) + \widehat{\mathbf{D}}^{-1}\right)^{-1}$ where closeness of $H_h^{-1}(\widetilde{\mathbf{u}})$ to zero determines the appropriateness of the FE estimate of $\mathbf{D}$.

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.

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

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.

Lee, Y. & Nelder, J. A. (2003), 'Extended-REML estimators', *Journal of Applied Statistics* **30**(8), 845–856.

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

Liang, K. Y. & Zeger, S. L. (1986), 'Longitudinal data analysis using generalized linear models', *Biometrika* **73**, 13–22.

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.

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.

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

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

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

Zeger, S. L., Liang, K. Y. & Albert, P. S. (1988), 'Models for longitudinal data: a genralized estimation equation approach', *Biometrics* **44**(4), 1049–1060.

A Appendix

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

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

where, $\mathbf{A}$ and $\mathbf{B}$ are the matrices of constants, ∂ denotes differential and the derivatives are taken w.r. to $\mathbf{D}$.

Chain Rule: Let, h be a composite function such that $h(\mathbf{X}) = g(F(\mathbf{X}))$ when $F(\mathbf{X}) = \mathbf{b}$ then $\mathcal{D}(h(\mathbf{X})) = (\mathcal{D}g(\mathbf{b}))\mathcal{D}F(\mathbf{X})$, where $\mathcal{D}$ operator stands for the matrix differentiation *i.e.* $\mathcal{D}F(\mathbf{X}) = \frac{\partial}{\partial vec(\mathbf{X})}F(\mathbf{X})$. 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 (8)

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 (10)

From equation (9) we have

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

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

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

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

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

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

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

A.3 Derivation of equation (11)

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

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

and

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

we have from A.1

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

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

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

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

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

Equation (11) can easily be obtained from the last equation given above.