Harvard University
Harvard University Biostatistics Working Paper Series

Year 2005 *Paper* 33

Gauss-Seidel Estimation of Generalized Linear Mixed Models with Application to Poisson Modeling of Spatially Varying Disease Rates

Subharup Guha* Louise Ryan†

*Harvard University, sguha@hsph.harvard.edu

†Harvard School of Public Health and Dana-Farber Cancer Institute, lryan@hsph.harvard.edu

Gauss-Seidel Estimation of Generalized Linear Mixed Models with Application to Poisson Modeling of Spatially Varying Disease Rates

Subharup Guha and Louise Ryan

Abstract

Generalized linear mixed models (GLMMs) provide an elegant framework for the analysis of correlated data. Due to the non-closed form of the likelihood, GLMMs are often fit by computational procedures like penalized quasi-likelihood (PQL). Special cases of these models are generalized linear models (GLMs), which are often fit using algorithms like iterative weighted least squares (IWLS). High computational costs and memory space constraints often make it difficult to apply these iterative procedures to data sets with very large number of cases.

This paper proposes a computationally efficient strategy based on the Gauss-Seidel algorithm that iteratively fits sub-models of the GLMM to subsetted versions of the data. Additional gains in efficiency are achieved for Poisson models, commonly used in disease mapping problems, because of their special collapsibility property which allows data reduction through summaries. Convergence of the proposed iterative procedure is guaranteed for canonical link functions. The strategy is applied to investigate the relationship between ischemic heart disease, socioeconomic status and age/gender category in New South Wales, Australia, based on outcome data consisting of approximately 33 million records. A simulation study demonstrates the algorithm's reliability in analyzing a data set with 12 million records for a (non-collapsible) logistic regression model.

Gauss-Seidel Estimation of Generalized Linear Mixed Models with Application to Poisson Modeling of Spatially Varying Disease Rates

Subharup Guha and Louise Ryan*

September 13, 2006

*Subharup Guha is Postdoctoral Research Fellow, Department of Biostatistics, Harvard School of Public Health, 655 Huntington Avenue, SPH2, 4th Floor, Boston, MA 02115 (email: sguha@hsph.harvard.edu); and Louise Ryan is Professor of Biostatistics, Harvard School of Public Health, 655 Huntington Avenue, SPH2, 4th Floor, Boston, MA 02115 (email: lryan@hsph.harvard.edu).

Keywords: Backfitting algorithm, CAR model, collapsibility, epidemiology, iterative weighted least squares, large sample sizes.

Abstract

Generalized linear mixed models (GLMMs) provide an elegant framework for the analysis of correlated data. Due to the non-closed form of the likelihood, GLMMs are often fit by computational procedures like penalized quasi-likelihood (PQL). Special cases of these models are generalized linear models (GLMs), which are often fit using algorithms like iterative weighted least squares (IWLS). High computational costs and memory space constraints often make it difficult to apply these iterative procedures to data sets with very large number of cases.

This paper proposes a computationally efficient strategy based on the Gauss-Seidel algorithm that iteratively fits sub-models of the GLMM to subsetted versions of the data. Additional gains in efficiency are achieved for Poisson models, commonly used in disease mapping problems, because of their special collapsibility property which allows data reduction through summaries. Convergence of the proposed iterative procedure is guaranteed for canonical link functions. The strategy is applied to investigate the relationship between ischemic heart disease, socioeconomic status and age/gender category in New South Wales, Australia, based on outcome data consisting of approximately 33 million records. A simulation study demonstrates the algorithm's reliability in analyzing a data set with 12 million records for a (non-collapsible) logistic regression model.

1 INTRODUCTION

A common subject of investigation in epidemiology is the study of environmental effects and geographical variation in the incidence or mortality rates of rare diseases. Generalized linear mixed models (GLMMs) are often used because they can account for overdispersion and spatial correlation in the data. When the available data consist of counts aggregated over small areas like counties or postcodes, a Poisson GLMM with log link and normally distributed random effects is a reasonable model

for the data (Clayton and Kaldor, 1987). Computational techniques are necessary for fitting such models, especially due to the non-closed form of the likelihood function. A number of authors have discussed general estimation procedures for GLMMs, including Wolfinger and O'Connell (1993), Engel and Keen (1992), Waclawiw and Liang (1993) and Davidian and Giltinan (1993). Breslow and Clayton (1993), also see Pinheiro and Bates (2000), propose an approximate solution, the so-called penalized quasi-likelihood (PQL). Wolfinger, Tobias and Sall (1994) and more recently Bates and DebRoy (2004) discuss computationally efficient methods for implementing PQL.

Generalized linear models (GLMs) are often when the data can be modeled as independent outcomes. GLMs can be regarded as a special case of GLMMs where the random effects are set equal to zero. Except for the simplest of models for which the parameter estimates are available in closed form, computational procedures like the iterative weighted least squares algorithm (McCullagh and Nelder, 1999, p. 40) are necessary for fitting GLMs.

Fitting GLMs and GLMMs can be particularly challenging for large sample sizes. In the study that motivates this work, investigators wish to explore environmental, spatial and temporal variation in the rates of ischemic heart disease in New South Wales (NSW), Australia. Daily disease counts over a period of five years from 591 postal areas, stratified by age and gender, lead to a data set of approximately 33 million observations. A number of computational challenges including insufficient memory and high computational costs make it difficult to apply the standard model fitting algorithms to these data. We propose a Gauss-Seidel algorithm that allows the fitting of GLMMs to data sets with very large sample sizes via PQL. The iterative procedure performs blockwise updates of the fixed effects, followed by the random effects and the variance components. Focussing only on the block Gauss-Seidel updates of the fixed effects, we obtain a computationally efficient alternative to the IWLS algorithm for fitting GLMs. The application is motivated by the NSW data set and focuses on the analysis of lattice data where spatial location is defined according to an irregular grid.

1.1 Modeling framework for GLMMs

Let $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ be the vector of observations. Conditional on an unobserved random effect, the observation y_i is assumed to be independently distributed in the exponential family with mean $\mu_i = g(\eta_i)$, where $g(\cdot)$ is the link function and η_i is the linear predictor (McCullagh and Nelder, 1999).

Let $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)^T$ be the vector of fixed effects. The vector of random effects, $\boldsymbol{b} = (b_1, \ldots, b_q)^T$, is assumed to follow a multivariate normal distribution with mean zero and covariance matrix $\boldsymbol{D(\theta)}$, where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_r)^T$ represents the variance components. Conditional on the random effects, a generalized linear mixed model (GLMM) in canonical form assumes a linear predictor of the form $\eta_i = \boldsymbol{x_i^T \alpha} + \boldsymbol{z_i^T b}$, where $\boldsymbol{x_i^T}$ and $\boldsymbol{z_i^T}$ respectively denote the i^{th} rows of design matrices $\boldsymbol{X}$ and $\boldsymbol{Z}$ of dimensions $n \times p$ and $n \times q$. For most commonly used GLMMs, the variance $Var(y_i)$ can be written as the product of a variance function $F(\mu_i)$ and a term that is constant over the n observations (refer to Table 2.1 of McCullagh and Nelder, 1999).

1.2 Spatial variation of incidence or mortality rates of rare diseases

In disease mapping problems, the outcome data are disease or mortality counts observed over a geographical area consisting of q regions. A Poisson GLMM with log link is a commonly used model in these problems. The vector $\boldsymbol{b} = (b_1, \ldots, b_q)^T$ of region-specific random effects follows a q-variate normal distribution with mean zero and covariance matrix $\boldsymbol{D(\theta)}$. In many spatial epidemiology studies, $\boldsymbol{Z}$ is a 0-1 matrix assigning the responses to one of the q regions: for case $i = 1, \ldots, n$ and region $s = 1, \ldots, q$, element z_{is} of the design matrix $\boldsymbol{Z}$ equals one if the observation y_i corresponds to region s, and equals zero otherwise.

Modeling spatial dependence. Besag, York and Mollié (1991) discuss a number of models that account for spatial dependence among the random effects. Let $s \sim t$ represent the event that regions

s and t are neighbors. Let n_s be the number of neighbors of region s. Define the q by q matrix $\boldsymbol{R}$ as follows:

$$R_{s,t} = \begin{cases} n_s, & \text{if } s = t, \\ -I(s \sim t), & \text{if } s \neq t, \end{cases}$$

for $s, t = 1, \ldots, q$. The intrinsic autoregressive model of Besag et al. (1991) assumes that the random effects $\boldsymbol{b}$ have a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{D}$ whose Moore-Penrose generalized inverse is $\boldsymbol{D}^- = \boldsymbol{R}/\sigma^2$.

The variance component $\boldsymbol{\theta}$ of the intrinsic autoregressive model consists of a single parameter, the precision σ^{-2}, accounting for both spatial dependence and overdispersion. A number of authors have suggested extensions to the basic model, including Besag et al. (1991) and Cressie (1991). Leroux, Lei and Breslow (1998) augment the variance component to $\boldsymbol{\theta} = (\sigma^{-2}, \lambda)$. The covariance matrix $\boldsymbol{D}$ of the random effects is assumed to satisfy

$$\sigma^2 \boldsymbol{D}^{-1} = (1 - \lambda)\boldsymbol{I} + \lambda \boldsymbol{R}, \tag{1}$$

where $0 \leq \lambda < 1$. The matrix $\boldsymbol{D}$ is invertible because the parameter space of λ excludes the value 1. This assumption is usually reasonable; for most data, the best-fitting λ belongs to the interior of the interval $(0, 1)$. When $\lambda = 0$, we obtain the independence model, for which $\boldsymbol{D} = \sigma^2 \boldsymbol{I}$.

1.3 Some standard implementations of PQL

Penalized quasi-likelihood (PQL) (Breslow and Clayton 1993; also see Pinheiro and Bates, 2000) is a well-known procedure for analyzing GLMMs with unknown variance components. Initial values are assigned to the model parameters as described in Pinheiro and Bates (2000). The standard implementation of PQL iteratively updates the estimates by the following steps until relative changes in the estimates become sufficiently small.

Step 1: Assuming the variance components $\boldsymbol{\theta}$ to be known and equal to their current estimates, the fixed and random effects are updated by maximizing Green's (1987) PQL criterion. Applying Fisher's scoring, which is equivalent to the Newton-Raphson method for GLMMs with canonical link functions, the estimates are computed as the iterative solution to the following system of equations:

$$\boldsymbol{H} \cdot \begin{pmatrix} \boldsymbol{\alpha} \\ \boldsymbol{b} \end{pmatrix} = \begin{bmatrix} \boldsymbol{X}^T\boldsymbol{W}\boldsymbol{Y} \\ \boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{Y} \end{bmatrix}, \tag{2}$$

where the matrix

$$\boldsymbol{H} = \begin{bmatrix} \boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X} & \boldsymbol{X}^T\boldsymbol{W}\boldsymbol{Z} \\ \boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{X} & \boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{Z} + \boldsymbol{D}^{-1} \end{bmatrix}. \tag{3}$$

The symbol $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ in equation (2) represents the vector of working values and not the data, $\boldsymbol{y}$. The $n \times n$ diagonal matrix of working weights is denoted by $\boldsymbol{W}$. For $i = 1, \ldots, n$, the working value Y_i is related to the data y_i and to the current parameter estimates, as follows: $Y_i = \eta_i + (y_i - \mu_i)\partial\eta_i/\partial\mu_i$. The working weight is given by $w_i = \{F(\mu_i)\}^{-1} (\partial\mu_i/\partial\eta_i)^2$.

An equivalent implementation (Breslow and Clayton, 1993) separates updates of the fixed and random effects as follows: $\hat{\boldsymbol{\alpha}} = \left(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{Y}$ and $\hat{\boldsymbol{b}} = \boldsymbol{V}^{-1}\boldsymbol{Z}^T\boldsymbol{W}\left(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\alpha}}\right)$.

Step 2: Assuming the fixed effects to be equal to their current estimate $\hat{\boldsymbol{\alpha}}$, we estimate the variance components. There are several possible approaches. A natural option is to maximize the quasi-likelihood

$$ql_1\left(\hat{\boldsymbol{\alpha}}, \boldsymbol{\theta}\right) = -\frac{1}{2}\log|\boldsymbol{V}| - \frac{1}{2}\boldsymbol{r}^T\boldsymbol{V}^{-1}\boldsymbol{r}. \tag{4}$$

where $\boldsymbol{r} = \boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\alpha}}$ and $\boldsymbol{V} = \boldsymbol{W}^{-1} + \boldsymbol{Z}\boldsymbol{D}\boldsymbol{Z}^T$ is an n by n matrix. Alternatively, Breslow and Clayton (1993) recommend using the REML version of the quasi-likelihood:

$$ql_2\left(\hat{\boldsymbol{\alpha}}, \boldsymbol{\theta}\right) = -\frac{1}{2}\log|\boldsymbol{V}| - \frac{1}{2}\log\left|\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X}\right| - \frac{1}{2}\boldsymbol{r}^T\boldsymbol{V}^{-1}\boldsymbol{r}. \tag{5}$$

The value of $\boldsymbol{\theta}$ that maximizes (4) or (5) can be recursively computed by either Fisher scoring or Newton-Raphson method.

Wolfinger, Tobias and Sall (1994) discuss *profile likelihood* methods for estimating the variance components of linear mixed models. Their work is directly applicable here. Given the variance components $\boldsymbol{\theta}$, the fixed effects can be estimated as $\boldsymbol{a}(\boldsymbol{\theta}) = \left(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{Y}$, where the matrix $\boldsymbol{V}$ and vector $\boldsymbol{Y}$ of working values depend on $\boldsymbol{\theta}$. After analytical substitution of the fixed effects estimates, we obtain the following maximum likelihood and REML objective functions for profiling:

$$\begin{aligned} ql_3(\boldsymbol{\theta}) &= -\frac{1}{2}\log|\boldsymbol{V}| - \frac{1}{2}\boldsymbol{\rho}^T\boldsymbol{V}^{-1}\boldsymbol{\rho} \\ ql_4(\boldsymbol{\theta}) &= -\frac{1}{2}\log|\boldsymbol{V}| - \frac{1}{2}\log\left|\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X}\right| - \frac{1}{2}\boldsymbol{\rho}^T\boldsymbol{V}^{-1}\boldsymbol{\rho} \end{aligned} \tag{6}$$

where $\boldsymbol{\rho} = \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{a}(\boldsymbol{\theta})$. Profiling reduces the dimension of the parameter space resulting in more efficient estimation. However, as we shall later see, it is associated with greater computational costs per iteration which makes it infeasible for very large sample sizes.

In the remainder of this subsection, we outline the iterative updates of the variance components using the four above-mentioned strategies. We rely on the analytical framework of Wolfinger et al. (1994). Bates and DebRoy (2004) provide an alternative framework for implementing the profiling-based approaches via the Newton-Raphson method. The identity

$$\boldsymbol{V}^{-1} = \boldsymbol{W} - \boldsymbol{W}\boldsymbol{Z}\left(\boldsymbol{D}^{-1} + \boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}^T\boldsymbol{W} \tag{7}$$

which appears in Wolfinger et al. (1994) is key. It avoids the inversion of the non-diagonal matrix $\boldsymbol{V}$, an operation having $O(n^3)$ computational cost per iteration, and provides an alternative expression relying only on the submatrices of $\boldsymbol{H}$ defined in (3). It is therefore of $O(n)$ computational

cost. Let the matrix $\boldsymbol{C}$ be a square-root of $\left(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}$ and let $\boldsymbol{X}^* = \boldsymbol{X}\boldsymbol{C}$. Define

$$\begin{aligned}
&\boldsymbol{W}(\boldsymbol{Z},\boldsymbol{Z}) = \boldsymbol{Z}^T\boldsymbol{V}^{-1}\boldsymbol{Z}, \qquad \boldsymbol{W}(\boldsymbol{X},\boldsymbol{Z}) = \boldsymbol{X}^{*T}\boldsymbol{V}^{-1}\boldsymbol{Z} = \boldsymbol{W}(\boldsymbol{Z},\boldsymbol{X})^T,\\
&\boldsymbol{W}(\boldsymbol{X},\boldsymbol{X}) = \boldsymbol{X}^{*T}\boldsymbol{V}^{-1}\boldsymbol{X}^*, \qquad \boldsymbol{W}(\boldsymbol{\rho},\boldsymbol{X}) = \boldsymbol{\rho}^T\boldsymbol{V}^{-1}\boldsymbol{X}^* = \boldsymbol{W}(\boldsymbol{X},\boldsymbol{\rho})^T,\\
&\boldsymbol{W}(\boldsymbol{\rho},\boldsymbol{Z}) = \boldsymbol{\rho}^T\boldsymbol{V}^{-1}\boldsymbol{Z} = \boldsymbol{W}(\boldsymbol{Z},\boldsymbol{\rho})^T, \qquad \boldsymbol{W}(\boldsymbol{r},\boldsymbol{X}) = \boldsymbol{r}^T\boldsymbol{V}^{-1}\boldsymbol{X}^* = \boldsymbol{W}(\boldsymbol{X},\boldsymbol{r})^T,\\
&\boldsymbol{W}(\boldsymbol{r},\boldsymbol{Z}) = \boldsymbol{r}^T\boldsymbol{V}^{-1}\boldsymbol{Z} = \boldsymbol{W}(\boldsymbol{Z},\boldsymbol{r})^T
\end{aligned}$$

Using the foregoing matrices and for $i,j = 1,\ldots,r$, define the vectors $\boldsymbol{k}_1^{(i)}$ and $\boldsymbol{k}_5^{(i)}$, constants $k_2^{(i,j)}$ and $k_6^{(i,j)}$, and matrices $\boldsymbol{K}_3^{(i)}$ and $\boldsymbol{K}_4^{(i,j)}$, as follows:

$$\begin{aligned}
\boldsymbol{k}_1^{(i)} &= \boldsymbol{W}(\boldsymbol{X},\boldsymbol{Z})\,\frac{\partial \boldsymbol{D}}{\partial\theta_i}\,\boldsymbol{W}(\boldsymbol{Z},\boldsymbol{\rho}),\\
k_2^{(i,j)} &= 2\boldsymbol{W}(\boldsymbol{\rho},\boldsymbol{Z})\,\frac{\partial \boldsymbol{D}}{\partial\theta_i}\,\boldsymbol{W}(\boldsymbol{Z},\boldsymbol{Z})\,\frac{\partial \boldsymbol{D}}{\partial\theta_j}\boldsymbol{W}(\boldsymbol{Z},\boldsymbol{\rho}) - \boldsymbol{W}(\boldsymbol{\rho},\boldsymbol{Z})\,\frac{\partial^2 \boldsymbol{D}}{\partial\theta_i\partial\theta_j}\,\boldsymbol{W}(\boldsymbol{Z},\boldsymbol{\rho}),\\
\boldsymbol{K}_3^{(i)} &= \boldsymbol{W}(\boldsymbol{X},\boldsymbol{Z})\,\frac{\partial \boldsymbol{D}}{\partial\theta_i}\,\boldsymbol{W}(\boldsymbol{Z},\boldsymbol{X}),\\
\boldsymbol{K}_4^{(i,j)} &= 2\boldsymbol{W}(\boldsymbol{X},\boldsymbol{Z})\,\frac{\partial \boldsymbol{D}}{\partial\theta_i}\,\boldsymbol{W}(\boldsymbol{Z},\boldsymbol{Z})\,\frac{\partial \boldsymbol{D}}{\partial\theta_j}\boldsymbol{W}(\boldsymbol{Z},\boldsymbol{X}) - \boldsymbol{W}(\boldsymbol{X},\boldsymbol{Z})\,\frac{\partial^2 \boldsymbol{D}}{\partial\theta_i\partial\theta_j}\,\boldsymbol{W}(\boldsymbol{Z},\boldsymbol{X}),\\
\boldsymbol{k}_5^{(i)} &= \boldsymbol{W}(\boldsymbol{X},\boldsymbol{Z})\,\frac{\partial \boldsymbol{D}}{\partial\theta_i}\,\boldsymbol{W}(\boldsymbol{Z},\boldsymbol{r}),\\
k_6^{(i,j)} &= 2\boldsymbol{W}(\boldsymbol{r},\boldsymbol{Z})\,\frac{\partial \boldsymbol{D}}{\partial\theta_i}\,\boldsymbol{W}(\boldsymbol{Z},\boldsymbol{Z})\,\frac{\partial \boldsymbol{D}}{\partial\theta_j}\boldsymbol{W}(\boldsymbol{Z},\boldsymbol{r}) - \boldsymbol{W}(\boldsymbol{r},\boldsymbol{Z})\,\frac{\partial^2 \boldsymbol{D}}{\partial\theta_i\partial\theta_j}\,\boldsymbol{W}(\boldsymbol{Z},\boldsymbol{r})
\end{aligned} \tag{8}$$

Now define the vectors $\boldsymbol{g}^{(1)}$, $\boldsymbol{g}^{(2)}$ and $\boldsymbol{g}^{(3)}$ of length r and square matrices $\boldsymbol{H}^{(1)}$, $\boldsymbol{H}^{(2)}$ and $\boldsymbol{H}^{(3)}$ of order r, having typical elements

$$\begin{aligned}
g_i^{(1)} &= tr\left(\boldsymbol{W}(\boldsymbol{Z},\boldsymbol{Z})\,\frac{\partial \boldsymbol{D}}{\partial\theta_i}\right), \quad g_i^{(2)} = -\boldsymbol{W}(\boldsymbol{\rho},\boldsymbol{Z})\frac{\partial \boldsymbol{D}}{\partial\theta_i}\boldsymbol{W}(\boldsymbol{Z},\boldsymbol{\rho}), \quad g_i^{(3)} = -tr(\boldsymbol{K}_3^{(i)})\\
g_i^{(4)} &= -\boldsymbol{W}(\boldsymbol{r},\boldsymbol{Z})\frac{\partial \boldsymbol{D}}{\partial\theta_i}\boldsymbol{W}(\boldsymbol{Z},\boldsymbol{r}),\\
H_{ij}^{(1)} &= -tr\left(\boldsymbol{W}(\boldsymbol{Z},\boldsymbol{Z})\,\frac{\partial \boldsymbol{D}}{\partial\theta_i}\,\boldsymbol{W}(\boldsymbol{Z},\boldsymbol{Z})\,\frac{\partial \boldsymbol{D}}{\partial\theta_j}\right) + tr\left(\boldsymbol{W}(\boldsymbol{Z},\boldsymbol{Z})\,\frac{\partial^2 \boldsymbol{D}}{\partial\theta_i\partial\theta_j}\right),\\
H_{ij}^{(2)} &= k_2^{(i,j)} - 2\boldsymbol{k}_1^{(i)^T}\boldsymbol{k}_1^{(j)}, \quad H_{ij}^{(3)} = tr\left(\boldsymbol{K}_4^{(i,j)} - \boldsymbol{K}_3^{(i)}\boldsymbol{K}_3^{(j)}\right),\\
H_{ij}^{(4)} &= k_5^{(i,j)} - 2\boldsymbol{k}_4^{(i)^T}\boldsymbol{k}_4^{(j)},
\end{aligned} \tag{9}$$

Quasi-likelihood: $\max_{\boldsymbol{\theta}} ql_1(\hat{\boldsymbol{\alpha}}, \boldsymbol{\theta})$		
Technique	*Gradient*	*Hessian*
Newton-Raphson	$\boldsymbol{g}^{(1)} + \boldsymbol{g}^{(4)}$	$\boldsymbol{H}^{(1)} + \boldsymbol{H}^{(4)}$
Fisher scoring	$\boldsymbol{g}^{(1)} + \boldsymbol{g}^{(4)}$	$-\boldsymbol{H}^{(1)}$
REML quasi-likelihood: $\max_{\boldsymbol{\theta}} ql_2(\hat{\boldsymbol{\alpha}}, \boldsymbol{\theta})$		
Technique	*Gradient*	*Hessian*
Newton-Raphson	$\boldsymbol{g}^{(1)} + \boldsymbol{g}^{(4)} + \boldsymbol{g}^{(3)}$	$\boldsymbol{H}^{(1)} + \boldsymbol{H}^{(4)} + \boldsymbol{H}^{(3)}$
Fisher scoring	$\boldsymbol{g}^{(1)} + \boldsymbol{g}^{(4)} + \boldsymbol{g}^{(3)}$	$-\boldsymbol{H}^{(1)} + \boldsymbol{H}^{(4)}$

Table 1: Derivatives for the iterative updates of the variance components by quasi-likelihood maximization.

where $i, j = 1, \ldots, r$.

Tables 1 and 2 summarize the expressions for the afore-mentioned strategies for estimating the variance components. In every case, an updated estimate is obtained as the current value minus the product of the inverse Hessian matrix and the gradient vector.

Special case. Suppose the random effects are spatially varying and are distributed according to the CAR model (1) for which the variance components are $\boldsymbol{\theta} = (\lambda, \sigma^{-2})$. Since the natural parametrization is in terms of $\boldsymbol{D}^{-1}$ rather than $\boldsymbol{D}$, we first apply relation (1) and evaluate $\frac{\partial \boldsymbol{D}^{-1}}{\partial \lambda} = \boldsymbol{R} - \boldsymbol{I}$ and $\frac{\partial \boldsymbol{D}^{-1}}{\partial \sigma^{-2}} = (1-\lambda)\boldsymbol{I} + \lambda \boldsymbol{R}$. The identity $\frac{\partial \boldsymbol{D}}{\partial x} = -\boldsymbol{D}\frac{\partial \boldsymbol{D}^{-1}}{\partial x}\boldsymbol{D}$ gives us the partial derivatives with respect to $\boldsymbol{D}$, which are then substituted in (9) to compute the Hessian matrices and gradient vectors that appear in Tables 1 and 2.

1.4 IWLS algorithm for GLMs

For estimating the fixed effects $\boldsymbol{\alpha}$ of a GLM, the IWLS algorithm (McCullagh and Nelder, 1999, p. 40) is obtained from Step 1 of Section 1.3 as the solution to the normal equation, $\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X} \cdot \boldsymbol{\alpha} = \boldsymbol{X}^T\boldsymbol{W}\boldsymbol{Y}$.

Profile maximum likelihood: $\max_{\boldsymbol{\theta}} ql_3(\boldsymbol{\theta})$		
Technique	*Gradient*	*Hessian*
Newton-Raphson	$\boldsymbol{g}^{(1)} + \boldsymbol{g}^{(2)}$	$\boldsymbol{H}^{(1)} + \boldsymbol{H}^{(2)}$
Fisher scoring	$\boldsymbol{g}^{(1)} + \boldsymbol{g}^{(2)}$	$-\boldsymbol{H}^{(1)}$
Profile REML: $\max_{\boldsymbol{\theta}} ql_4(\boldsymbol{\theta})$		
Technique	*Gradient*	*Hessian*
Newton-Raphson	$\boldsymbol{g}^{(1)} + \boldsymbol{g}^{(2)} + \boldsymbol{g}^{(3)}$	$\boldsymbol{H}^{(1)} + \boldsymbol{H}^{(2)} + \boldsymbol{H}^{(3)}$
Fisher scoring	$\boldsymbol{g}^{(1)} + \boldsymbol{g}^{(2)} + \boldsymbol{g}^{(3)}$	$-\boldsymbol{H}^{(1)} + \boldsymbol{H}^{(3)}$

Table 2: Derivatives for the iterative updates of the variance components by profiling.

1.5 Computational challenges with large sample sizes

As a motivating example, consider the case study of Section 3 where the outcome data consist of the number of emergency room visits in New South Wales (NSW), Australia, that were diagnosed as ischemic heart disease (IHD) in postcode h, on day j and belonging to age/gender category k, out of the N_{hjk} people at risk. With $h = 1, \ldots, 591$, $j = 1, \ldots, 1826$ and $k = 1, \ldots, 30$, there are more than 33 million observations. Available covariates from the census include the population density and socioeconomic status of postcode h on day j. The goal is to study the association between IHD, socioeconomic status and age/gender category.

For such large sample sizes, a major computational challenge is the joint update of the fixed and random effects (Step 1 of Section 1.3). This procedure is cost-intensive because the large number of outcome data and covariates have to be processed $(p+q)^2$ number of times every iteration to compute the matrix $\boldsymbol{H}$ in (2). This significantly slows down calculations when n is very large. An equivalent strategy described in Section 1.3 first updates the fixed effects and then estimates the random effects as the BLUP in the underlying linear mixed model. This procedure is also cost-intensive for large n

because it relies on the matrix $\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X}$; applying identity (7), we have that

$$\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X} = \boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X} - \boldsymbol{X}^T\boldsymbol{W}\boldsymbol{Z}\left(\boldsymbol{D}^{-1} + \boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{X}. \tag{10}$$

Since the right hand side involves all elements of the matrix $\boldsymbol{H}$, this technique has the approximately same computational cost as a joint update of the fixed and random effects.

For the variance components updates in Step 2 of Section 1.3, consider the profiling-based strategies presented in Table 2. All four strategies involve the vector $\boldsymbol{g}^{(2)}$, which requires the computation of the matrix $\boldsymbol{W}(\boldsymbol{Z}, \boldsymbol{\rho})$; see definition (9). But since $\boldsymbol{\rho} = \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{a}(\boldsymbol{\theta})$ where $\boldsymbol{a}(\boldsymbol{\theta}) = \left(\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{Y}$, all of these strategies rely on the matrix $\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X}$. This rules out profiling-based strategies for large sample sizes. Consider now the quasi-likelihood–based strategies presented in Table 1. The matrix $\boldsymbol{H}^{(4)}$ is needed in three of the strategies. We find from relation (9) that $\boldsymbol{H}^{(4)}$ involves the matrix $\boldsymbol{W}(\boldsymbol{Z}, \boldsymbol{X})$, which depends through $\boldsymbol{C}^*$ on the matrix $\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X}$. For large n, this rules out the quasi-likelihood procedures that depend on either the Newton-Raphson or REML approaches.

Limitations of memory space are also important in problems involving a large number of cases. In the NSW example, the required memory for the standard PQL updates of the fixed and random effects are far beyond the current capacity of most computers. This is true for even relatively simple models having no random effects or interaction terms. Moreover, the memory requirements typically increase with model complexity, because of which most realistic models cannot be fit to these data. While using standard R functions like `glmmPQL` or `lmer` (for GLMMs) and `glm` (for GLMs), the functions are found to quit with an 'insufficient memory space' message.

This is the motivation for the efficient algorithm presented in Section 2. The proposed implementation of PQL achieves a drastic reduction in the computational cost and memory requirements by three strategies: *(i)* An initialization procedure quickly provides good starting values. *(ii)* The fixed and random effects are separately updated in *blocks* by applying the Gauss-Seidel algorithm. *(iii)* Among the various alternatives listed in Tables 1 and 2 for updating the variance components, we recommend

the maximization of quasi-likelihood $ql_1(\hat{\boldsymbol{\alpha}}, \boldsymbol{\theta})$ via Fisher scoring because it is well-suited for processing large number of observations.

Innovation *(ii)* allows the use of standard software like the `glm` function in R. Additionally, it exploits the structure of the design matrix and special properties of the model, like the collapsibility property of Poisson GLMMs with log link, to achieve further gains in computational efficiency.

1.6 The Gauss-Seidel (backfitting) algorithm

In a general setting, let the vector $\boldsymbol{\psi}$ represent the model parameters. Let $\boldsymbol{\psi}^{(m)}$ be the estimated parameters at the end of the m^{th} iteration. The classic Gauss-Seidel algorithm (Givens and Hoeting, 2005, p. 43) computes an updated estimate, $\boldsymbol{\psi}^{(m+1)}$, by performing a series of univariate maximizations of an objective function (like the log-likelihood in the case of GLMs, and Green's PQL criterion in the case of GLMMs) with respect to individual parameters, ψ_j, holding the remaining parameters fixed at their current estimated values. In other words, the $(m+1)^{st}$ iteration updates the parameter ψ_j by setting $(\psi_1, \ldots, \psi_{j-1})$ equal to $(\psi_1^{(m+1)}, \ldots, \psi_{j-1}^{(m+1)})$ and setting $(\psi_{j+1}, \ldots, \psi_p)$ equal to $(\psi_{j+1}^{(m)}, \ldots, \psi_p^{(m)})$.

A number of variations are discussed in Ortega and Rheinboldt (2000). For example, if the maximizations cannot be performed in closed form, they can be replaced by a fixed number (say, m) of Newton-Raphson updates. An m-step Gauss-Seidel iteration method generally converges at the same asymptotic rate as the corresponding infinite-step method (Thisted, 1988, p. 191). A blocked version of the Gauss-Seidel algorithm, which jointly updates groups of two or more parameters, is also an option (Ortega and Rheinboldt, 2000, p. 225).

The Section 1.3 procedure could be regarded as a blocked Gauss-Seidel algorithm. This is because it iteratively updates the fixed and random effects given current estimates of the variance components, and updates the variance components given estimates of the fixed effects. In Section 2, we generalize this idea to iteratively update smaller blocks of the fixed effects. As formally shown in Section 2.2, this

results in substantial savings in the cost per iteration, allowing the analysis of large number of cases using realistic models that cannot be otherwise fit using the Section 1.3 procedure. The computational savings are typically obtained at the price of a slower convergence rate, although this drawback can be overcome to a large extent by a judicious choice of initial parameter values and blocks, for which Sections 2.3 and 2.4 recommend useful strategies. Section 2.5 proves that convergence of the iterative procedure is guaranteed for models with canonical link. In the special case of GLMMs with known variance components and GLMs, the algorithm finds the unique global solution.

Section 3 applies these ideas to investigate the relationship between heart disease and socioeconomic factors in NSW, Australia. Section 4 investigates the performance of the algorithm for models lacking the computational benefit of collapsibility. The algorithm is found to perform well, although it is less computationally efficient than with Poisson models.

2 EFFICIENT PQL ESTIMATION FOR LARGE SAMPLE SIZES

2.1 Main iterative procedure

Initial values for the model parameters are computed as described in Section 2.3. The following cycle of steps is then repeated until relative changes in the estimates between successive iterations become sufficiently small:

Step 1a: Blockwise update of the fixed effects. Assuming the remaining parameters to be known and equal to their current estimates, the fixed effects are successively updated in blocks by Gauss-Seidel steps. It is sufficient to perform only one or two updates of each block. This step can be implemented using standard model-fitting procedures like the `glm` function in R.

More formally, let the p fixed effects be partitioned into r blocks, $\boldsymbol{\alpha_1}, \ldots, \boldsymbol{\alpha_r}$, where $r \leq p$. Let $\boldsymbol{\alpha}_{-t}$ denote the vector of remaining fixed effects after excluding the block $\boldsymbol{\alpha}_t$ and let $G(\boldsymbol{\alpha}, \boldsymbol{b}, \boldsymbol{\theta})$ represent

Green's PQL criterion. For $t = 1, \ldots, r$, let $\boldsymbol{X}_t$ ($\boldsymbol{X}_{-t}$) be the matrix of covariates associated with the $\boldsymbol{\alpha}_t$ block ($\boldsymbol{\alpha}_{-t}$ block). Applying the Section 1.6 procedure to the blocks and regarding Green's PQL criterion as the objective function to be maximized, we have that for $t = 1, \ldots, r$, the block $\boldsymbol{\alpha_t}$ is updated by maximizing $G(\boldsymbol{\alpha}_t, \hat{\boldsymbol{\alpha}}_{-t}, \hat{\boldsymbol{b}}, \hat{\boldsymbol{\theta}})$ with respect to $\boldsymbol{\alpha}_t$. Since this maximization cannot typically be performed in closed form, updates can be obtained by Fisher scoring, which for GLMMs with canonical link functions is equivalent to Newton-Raphson updates. The updated block is then $\hat{\boldsymbol{\alpha}}_t = \hat{\boldsymbol{\alpha}}_t^{old} + \boldsymbol{\Delta}_t$, where the increment $\boldsymbol{\Delta}_t$ satisfies the equation $-\partial^2 G/\partial \boldsymbol{\alpha}_t^2 \cdot \boldsymbol{\Delta}_t = \partial G/\partial \boldsymbol{\alpha}_t$ and the partial derivatives are evaluated using the vector of current estimates, $(\hat{\boldsymbol{\alpha}}_t^{old}, \hat{\boldsymbol{\alpha}}_{-t}, \hat{\boldsymbol{b}}, \hat{\boldsymbol{\theta}})$.

Similar to the derivation of relation (2) given in Breslow and Clayton (1993), we can show that the foregoing Gauss-Seidel update is equivalent to solving this normal equation in the increment $\boldsymbol{\Delta}_t$:

$$\boldsymbol{X}_t^T \boldsymbol{W} \boldsymbol{X}_t \cdot \boldsymbol{\Delta}_t = \boldsymbol{X}_t^T \boldsymbol{W} \left(\boldsymbol{Y} - \hat{\boldsymbol{\eta}}^{old} \right), \tag{11}$$

where the linear predictor is $\hat{\boldsymbol{\eta}}^{old} = \boldsymbol{\eta}(\hat{\boldsymbol{\alpha}}_t^{old}, \hat{\boldsymbol{\alpha}}_{-t}, \hat{\boldsymbol{b}})$. The vector $\boldsymbol{Y}$ of working values and the diagonal matrix $\boldsymbol{W}$ of working weights are both computed using $\hat{\boldsymbol{\eta}}^{old}$. After the increment $\boldsymbol{\Delta}_t$ has been computed in (11), the updated linear predictor $\hat{\boldsymbol{\eta}} = \boldsymbol{\eta}(\hat{\boldsymbol{\alpha}}_t, \hat{\boldsymbol{\alpha}}_{-t}, \hat{\boldsymbol{b}})$ can be easily computed as $\hat{\boldsymbol{\eta}} = \hat{\boldsymbol{\eta}}^{old} + \boldsymbol{X}_t \boldsymbol{\Delta}_t$.

We observe that equation (11) represents the Fisher scoring updates for a GLM having the same likelihood and link function as the original model, but having the block of fixed effects $\boldsymbol{\alpha}_t$ as its unknown parameters, $\boldsymbol{X}_t$ as its design matrix, and $\hat{\boldsymbol{e}}_t = \boldsymbol{X}_{-t} \hat{\boldsymbol{\alpha}}_{-t} + \hat{\boldsymbol{b}}$ as the known vector of offsets in its linear predictor, $\boldsymbol{\eta} = \hat{\boldsymbol{e}}_t + \boldsymbol{X}_t \boldsymbol{\alpha}_t$. This allows the use of standard model-fitting procedures like the `glm` function in R to update the blocks, although the availability of standard software is not required but just convenient; the key issue being that at each iteration, we are solving a lower-dimensional problem. The vector $\hat{\boldsymbol{e}}_t$ of offsets for the nested GLM can be specified in R by the `offset` option.

An m-step Gauss-Seidel method converges at the same asymptotic rate as the corresponding infinite step method. Because of this, it is sufficient to perform only one or two Newton-Raphson updates of each block. The number of Newton-Raphson updates can be limited to one or two by setting the `maxit`

option.

Unlike the remaining steps of the algorithm, there is a considerable amount of flexibility in the implementation of Step 1a because of the number of ways in which the fixed effects can be grouped to form the blocks. The blocks that result in higher reductions in the computational cost and quicker convergence would depend on the particular model and the data. For example, a model with separable spatial and temporal main effects would be updated differently than a model where the spatial and temporal terms interact. Section 2.4 provides some guidelines in a general situation.

A special advantage of Poisson models with log link. Further computational efficiency is possible in disease mapping applications that assume a Poisson GLMM with log link. These models have a special collapsibility property (Elliot et al. 2001), not shared by other models like logistic regression models, which allows parameters assumed to be known at any given step to be included in the adjusted population at risk. Additionally, it allows us to exploit any repeats in the covariates associated with the unknown parameters at any given stage to work with summarized versions of the data. The data summaries can be computed during a pre-processing step. In the NSW example of Section 3, summarized data consisting of approximately million cases (rather than the 33 million cases of the full data set) were analyzed using the `glm` function to estimate the fixed effects of the GLMM. Most computers are currently able to handle computations on this reduced scale.

For a more precise description of how collapsibility works, suppose that the matrix of covariates $\boldsymbol{X}_t$ associated with the block $\boldsymbol{\alpha}_t$ has rows $\boldsymbol{x}_{t1}^T, \ldots, \boldsymbol{x}_{tn}^T$ of which only $n_t \leq n$ rows are non-zero and unique. Denote these unique, non-zero rows by $\boldsymbol{u}_{t1}^T, \ldots, \boldsymbol{u}_{tn_t}^T$ and let the matrix $\boldsymbol{U}_t$ consist of only these n_t number of rows. For $i = 1, \ldots, n_t$, let $y_{i\cdot} = \sum_{j:\boldsymbol{x}_{tj}=\boldsymbol{u}_{ti}} y_j$ ($\hat{e}_{i\cdot} = \sum_{j:\boldsymbol{x}_{tj}=\boldsymbol{u}_{ti}} \hat{e}_{tj}$) represent the summed responses (offsets) corresponding to the covariate value of $\boldsymbol{u}_{ti}$. It is easy to show that for a Poisson model with log link, the nested GLM that updates the block $\boldsymbol{\alpha}_t$ further simplifies to the model:

$y_{i.} \sim Po(\kappa_i)$, where $\log(\kappa_i) = \hat{e}_{i.} + \boldsymbol{U}_t\boldsymbol{\alpha}_t$ for $i = 1, \ldots, n_t$. This GLM involves only the summarized data consisting of $n_t \leq n$ cases. The summarized responses $\{y_{i.}\}_{i=1}^{n_t}$ can be computed during a one-time pre-processing step. The summarized offsets $\{\hat{e}_{i.}\}_{i=1}^{n_t}$ depend on the current estimates of the remaining model parameters, and must be computed on the fly as the algorithm proceeds.

Step 1b: Updating the random effects. The following Proposition describes how the estimated random effects are updated given current estimates of the fixed effects and the variance components.

Proposition 2.1 *Suppose the fixed effects and variance components are known and are equal to their respective current estimates, $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\theta}}$. Assume that $\boldsymbol{D} = \boldsymbol{D}(\hat{\boldsymbol{\theta}})$. Let $\hat{\boldsymbol{b}}^{old}$ denote the current estimate of the random effects and the corresponsing linear predictor be $\hat{\boldsymbol{\eta}}^{old} = \boldsymbol{\eta}(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{b}}^{old})$. The updated random effects are given by $\hat{\boldsymbol{b}} = \hat{\boldsymbol{b}}^{old} + \Delta\hat{\boldsymbol{b}}$, where the iterative adjustment has the expression*

$$\Delta\hat{\boldsymbol{b}} = \left(\boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{Z} + \boldsymbol{D}^{-1}\right)^{-1}\left(\boldsymbol{Z}^T\boldsymbol{W}\left(\boldsymbol{Y} - \hat{\boldsymbol{\eta}}^{old}\right) - \boldsymbol{D}^{-1}\hat{\boldsymbol{b}}^{old}\right). \tag{12}$$

The vector $\boldsymbol{Y}$ of working values and the set of working values in the diagonal matrix $\boldsymbol{W}$ in (12) are computed using $\hat{\boldsymbol{\eta}}^{old}$. The updated linear predictor $\hat{\boldsymbol{\eta}} = \boldsymbol{\eta}(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{b}})$ is given by $\hat{\boldsymbol{\eta}} = \hat{\boldsymbol{\eta}}^{old} + \boldsymbol{Z}\Delta\hat{\boldsymbol{b}}$.

Proof. The estimates of the random effects are chosen to maximize Green's PQL criterion, $G(\boldsymbol{\alpha}, \boldsymbol{b}, \boldsymbol{\theta})$. The current values of the partial derivatives $\partial G/\partial \boldsymbol{b}$ and $-\partial^2 G/\partial \boldsymbol{b}^2$ are, respectively, $\boldsymbol{Z}^T\boldsymbol{W}\left(\boldsymbol{Y} - \hat{\boldsymbol{\eta}}^{old}\right) - \boldsymbol{D}^{-1}\hat{\boldsymbol{b}}^{old}$ and $\boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{Z} + \boldsymbol{D}^{-1}$. The random effects are updated by the Newton-Raphson method as claimed.

Remark: In many epidemiological studies, the design matrix $\boldsymbol{Z}$ contains only 0's and 1's that assign each of the n observations to exactly one of the q random effects. This special structure results in $\boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{Z}$ being a diagonal matrix in (12). The j^{th} diagonal element is then computed simply by summing the

working weights corresponding to the random effect b_j. The vector $\boldsymbol{Z}^T\boldsymbol{W}\left(\boldsymbol{Y}-\hat{\boldsymbol{\eta}}^{old}\right)$ is also computed very quickly in a similar manner.

Step 2: Updating the variance components. Using the current estimates of the model parameters, the variance components are updated by maximizing quasi-likelihood (4) via Fisher scoring. For data sets with large sample sizes, this involves a substantially lower computational cost compared to the other strategies presented in Tables 1 and 2. From Table 1, we find that the recommended procedure relies on three quantities: matrix $\boldsymbol{H}^{(1)}$ and the vectors $\boldsymbol{g}^{(1)}$ and $\boldsymbol{g}^{(4)}$. The first two quantities involve the matrix $\boldsymbol{W}(\boldsymbol{Z},\boldsymbol{Z})$ defined in (9). On applying identity (7), we obtain

$$\boldsymbol{W}(\boldsymbol{Z},\boldsymbol{Z}) = \boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{Z} - \boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{Z}\left(\boldsymbol{D}^{-1}+\boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{Z},$$

which depends only on the lower block-diagonal of matrix $\boldsymbol{H}$ in (3). As described earlier, it is trivial to compute in many epidemiological problems where the matrix $\boldsymbol{Z}$ consists only of 0's and 1's. The third quantity, $\boldsymbol{g}^{(4)}$, relies on $\boldsymbol{W}(\boldsymbol{Z},\boldsymbol{r})$. This depends on the lower block-diagonal of $\boldsymbol{H}$ and the vector $\boldsymbol{Z}^T\boldsymbol{W}\left(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\alpha}}\right)$, which is easy to compute for even n very large.

2.2 Why is the proposed algorithm efficient for large sample sizes?

We return to some of the computational challenges with large sample sizes discussed in Section 1.3 and show why the proposed algorithm outperforms standard implementations of PQL. For $t = 1,\ldots,r$, let the block of fixed effects $\boldsymbol{\alpha_t}$ consist of p_t parameters, so that $\sum_t p_t = p$ is the total number of fixed effects. Let m_t be the number of non-zero rows of the covariate matrix $\boldsymbol{X}_t$ associated with the block $\boldsymbol{\alpha_t}$ and m_q be the number of non-zero rows of the design matrix $\boldsymbol{Z}$ associated with the random effects.

As mentioned earlier, a major computational burden in Step 1 of Section 1.3 is the joint update of the fixed and random effects, because the evaluation of the matrix $\boldsymbol{H}$ has a cost per iteration of the order $n(p+q)^2$. The cost can be prohibitively high when n is large. In contrast, Gauss-Seidel updates

of the fixed and random effects (Steps 1a and 1b of Section 2.1) require only the block diagonals of the matrix $\boldsymbol{H}$ and so have a smaller cost per iteration of the order $\sum_{t=1}^{r} m_t p_t^2 + m_q q^2 \leq n(p+q)^2$. Greater computational savings are achieved for Poisson models with log link, for which the cost per iteration is of the order $\sum_{t=1}^{r} n_t p_t^2 + nq^2$. This is bounded above by $\sum_{t=1}^{r} m_t p_t^2 + m_q q^2$, the cost per iteration for non-collapsible models. The two costs are equal if and only if the rows of the design matrix $\boldsymbol{X}_t$ have no repeats for all blocks $\boldsymbol{\alpha_t}$.

When using standard subroutines like the `glm` function in R for updating the fixed effects, a further advantage is the smaller amount of required memory for the proposed algorithm. A joint update of all p fixed effects using the `glm` function requires the specification of the entire design matrix $\boldsymbol{X}$ consisting of np elements. For large sample sizes, the memory requirement often exceeds the total capacity of the computer. However, a Gauss-Seidel update of the block $\boldsymbol{\alpha_t}$ relies only on the non-zero rows of $\boldsymbol{X}_t$. This consists of $m_t p_t$ elements and requires only a fraction $(m_t/n \cdot p_t/p)$ of the memory. There are further savings for a Poisson model with log link, which depend only on the matrix $\boldsymbol{U}_t$ consisting of $n_t p_t$ elements. This further reduces the memory requirement by a factor of $m_t/n_t \geq 1$.

The recommended strategy of updating the variance components is more cost-effective than the other methods presented in Tables 1 and 2, because they are associated with essentially the same computational cost as Step 1 of Section 1.3 (see Section 1.5). The recommended procedure relies only the lower block-diagonal of matrix $\boldsymbol{H}$ in (3) and on the vector $\boldsymbol{Z}^T\boldsymbol{W}\,(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\alpha}})$. Both quantities can be quickly computed, as explained in Section 2.1, with the cost benefits being especially high in a typical epidemiological problem where matrix $\boldsymbol{Z}$ consists of only 0's and 1's.

2.3 Initialization

Based on simulation results, we recommend the following technique for providing good starting values for the main iterative procedure. A similar strategy is described in Pinheiro and Bates (2000). Any

error in these initial estimates is subsequently eliminated by the main iterative procedure.

Initializing the fixed effects. On setting all the random effects equal to zero, the GLMM reduces to a GLM. Starting with arbitrary initial values, we iterate Step 1a of Section 2.1 until the estimated fixed effects converge.

Initializing the random effects. Assume the fixed effects to be equal to their initialized values. The random effects are estimated by the following iterative procedure which is approximately valid for a large number of cases.

Proposition 2.2 *Let* $\hat{\boldsymbol{b}}^{old}$ *be the estimated random effects and the estimated linear predictor be* $\hat{\boldsymbol{\eta}}^{old} = \boldsymbol{\eta}(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{b}}^{old})$. *The updated random effects are obtained as* $\hat{\boldsymbol{b}} = \hat{\boldsymbol{b}}^{old} + \Delta\hat{\boldsymbol{b}}$, *where* $\Delta\hat{\boldsymbol{b}}$ *denotes the iterative adjustment. For a large number of cases, conditions commonly satisfied by the covariates and the sampling scheme guarantee that the iterative adjustment is approximately given by*

$$\Delta\hat{\boldsymbol{b}} \approx \left(\boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}^T\boldsymbol{W}\left(\boldsymbol{Y} - \hat{\boldsymbol{\eta}}^{old}\right),$$

where the vector $\boldsymbol{Y}$ *of working values and the set of working values in the diagonal matrix* $\boldsymbol{W}$ *are both computed using* $\hat{\boldsymbol{\eta}}^{old}$. *The updated linear predictor* $\hat{\boldsymbol{\eta}} = \boldsymbol{\eta}(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{b}})$ *is then computed as* $\hat{\boldsymbol{\eta}} = \hat{\boldsymbol{\eta}}^{old} + \boldsymbol{Z}\Delta\hat{\boldsymbol{b}}$.

As a special case, suppose the model is a Poisson GLMM with log link. Assume also that the design matrix $\boldsymbol{Z}$ *consists of only 0's and 1's. When* n *is large, the random effects can be approximately estimated by:*

$$\hat{b}_j \approx \log\left(\frac{\sum_i y_i\, I(z_{ij}=1)}{\sum_i w_i\, \exp(-\boldsymbol{z}_i^T\hat{\boldsymbol{b}}^{old})\, I(z_{ij}=1)}\right), \qquad \text{for } j = 1, \ldots, q. \tag{13}$$

Proof. The result can be proved from a Bayesian perspective. Assuming a flat prior for the fixed effects and conditional on the variance components, it is easy to show that Green's PQL criterion is equal to the joint log-posterior density of the fixed/random effects plus a constant. Maximization of Green's PQL

criterion is therefore equivalent to computing the posterior mode of the fixed/random effects. We apply a Taylor approximation to obtain the following distribution of the working values: $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\alpha} + \boldsymbol{Z}\boldsymbol{b} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim N(0, \boldsymbol{W}^{-1})$ and prior $\boldsymbol{b} \sim N(0, \boldsymbol{D})$. Conditional on $\hat{\boldsymbol{\alpha}}$ and $\boldsymbol{b}$, the vector $\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\alpha}}$ is distributed as $N(\boldsymbol{Z}\boldsymbol{b}, \boldsymbol{D})$. By Lindley and Smith (1971), the posterior distribution of $\boldsymbol{b}$, conditional on $\hat{\boldsymbol{\alpha}}$ and the working values $\boldsymbol{Y}$, is $N\left(\left(\boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{Z} + \boldsymbol{D}^{-1}\right)^{-1}\boldsymbol{Z}^T\boldsymbol{W}\left(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\alpha}}\right), \left(\boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{Z}\right)^{-1} + \boldsymbol{D}^{-1}\right)$. As the sample size n grows, any prior information is progressively washed out by the data under mild conditions. Hence, as $n \to \infty$, the limiting distribution of $\boldsymbol{b}$ conditional on $\hat{\boldsymbol{\alpha}}$ and the working values $\boldsymbol{Y}$ is $N(\left(\boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}^T\boldsymbol{W}\left(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\alpha}}\right), \left(\boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{Z}\right)^{-1})$.

Therefore, for large n, the estimated random effects are $\hat{\boldsymbol{b}} \approx \left(\boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}^T\boldsymbol{W}\left(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\alpha}}\right)$, which is the Gauss-Seidel step for $\boldsymbol{b}$ given $\hat{\boldsymbol{\alpha}}$, if $\boldsymbol{b}$ were a fixed effect. Similar to the argument presented in McCullagh and Nelder (1989, p. 42), it can be shown that this is equivalent to the iterative scheme $\hat{\boldsymbol{b}} = \hat{\boldsymbol{b}}^{old} + \Delta\hat{\boldsymbol{b}}$ specified in the Proposition.

For a Poisson GLMM with log link and categorical random effects, the above procedure computes the MLE $\hat{\boldsymbol{b}}$ for the GLM: $y_i \sim Po(N_i e^{\boldsymbol{z}_i^T\boldsymbol{b}})$, where the known "adjusted populations at risk" are $N_i = \exp\left(\boldsymbol{x}_i^T\hat{\boldsymbol{\alpha}}\right) = w_i \exp\left(-\boldsymbol{z}_i^T\hat{\boldsymbol{b}}^{old}\right)$. The MLE $\hat{b}_j$ has the stated closed-form expression.

Initializing the variance components. Using the initialized estimates for the random effects, the following procedure provides starting values for the variance components. For the CAR model of Leroux et al. (1998), the procedure simplifies as described in Corollary 2.4. The proof relies on standard normal theory results and has been omitted.

Proposition 2.3 *Assume that the random effects are equal to their current estimates, $\hat{\boldsymbol{b}}$. Estimates of the variance components $\boldsymbol{\theta}$ can be obtained by maximizing the profile log-likelihood, $L(\boldsymbol{\theta}) = c + \frac{1}{2}\log|\boldsymbol{D}| - \frac{1}{2}\hat{\boldsymbol{b}}^T\boldsymbol{D}^{-1}\hat{\boldsymbol{b}}$, where c is a constant that does not depend on $\boldsymbol{\theta}$. The score vector and expected information*

matrix of the iterative procedure are given by:

$$s_i = \frac{1}{2}\hat{\boldsymbol{b}}^T \boldsymbol{D}^{-1}\frac{\partial \boldsymbol{D}}{\partial \theta_i}\boldsymbol{D}^{-1}\hat{\boldsymbol{b}} - \frac{1}{2}tr\left(\boldsymbol{D}^{-1}\frac{\partial \boldsymbol{D}}{\partial \theta_i}\right), \tag{14}$$

$$I_{ij}^{exp} = \frac{1}{2}tr\left(\boldsymbol{D}^{-1}\frac{\partial \boldsymbol{D}}{\partial \theta_i}\boldsymbol{D}^{-1}\frac{\partial \boldsymbol{D}}{\partial \theta_j}\right).$$

Corollary 2.4 *Define the matrix* $\boldsymbol{C}_\lambda = (1-\lambda)\boldsymbol{I} + \lambda\boldsymbol{R}$ *and function* $L(\lambda) = \log|\boldsymbol{C}_\lambda| - \log\left(\hat{\boldsymbol{b}}^T\boldsymbol{C}_\lambda\hat{\boldsymbol{b}}\right)$. *For the CAR model (1), we have* $\hat{\lambda} = \arg\max_\lambda L(\lambda)$ *and* $\hat{\sigma}^2 = \hat{\boldsymbol{b}}^T\boldsymbol{C}_{\hat{\lambda}}\hat{\boldsymbol{b}}/q$.

2.4 Selection of the blocks

There are many ways in which the p fixed effects can be grouped into r blocks. Some groupings result in a more efficient algorithm than others. Relevant criteria for evaluating these groupings include: *(i)* computational cost per iteration and *(ii)* convergence rate. Typically, one criterion is achieved at the expense of the other. Consider for example the PQL algorithm of Section 1.3, where the fixed effects are jointly updated with the random effects. The cost per iteration is often infeasible for large sample sizes, but convergence (on a hypothetical machine with infinite resources) is usually rapid. For a second example, consider an implementation of the block Gauss-Seidel algorithm where each fixed effect is its own block, so that $r = p$. Although it has a much smaller cost per iteration than the previous example, convergence could be slow because of potential "hem-stitching" problems caused by highly correlated estimates.

This suggests that the blocks should be created so that the fixed effects whose estimates are highly correlated are jointly updated within the same block. Simultaneously, the fixed effects should be grouped by aligning their zeros in the design matrix $\boldsymbol{X}$. This produces a large number of zero rows in the matrices $\boldsymbol{X}_t$ and as explained in Section 2.2, lower computational costs per iteration. These ideas are more formally stated below.

The following strategy achieves small per iteration computational costs: For the s^{th} fixed effect, let the set C_s represent the rows i of the design matrix $\boldsymbol{X}$ for which the covariate x_{is} is non-zero. We create

a partition P of the fixed effects depending on whether or not the fixed effects have the same elements in C_s. That is, the s^{th} and v^{th} fixed effect are placed in the same group if and only if $C_s = C_v$. This strategy aligns zeros in the covariates (if there are any) to create a set of blocks for which the covariate matrices $\boldsymbol{X}_t$ have fewer non-zero rows (m_t). This results in a smaller cost per iteration since it is of the order $\sum_{t=1}^{r} m_t p_t^2 + m_q q^2$ (see Section 2.2). It also results in a smaller amount of required memory, since it is the fraction $m_t/n \cdot p_t/p$ of that required for a joint update of the fixed effects (see Section 2.2).

The following strategy results in a faster converging algorithm by avoiding hem-stitching problems. It is motivated by the fact that for large sample sizes, the p by p submatrix $\boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{Z}$ in (3), evaluated at the true parameter values, is approximately equal to the posterior precision of the fixed effects. We initialize the parameters using the procedure of Section 2.3 and evaluate the submatrix $\boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{Z}$ for these initial parameter values. We then group into common blocks the fixed effects that are highly correlated under $\left(\boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{Z}\right)^{-1}$ ensuring a fast-converging Gauss-Seidel procedure.

As a further refinement for Poisson models with canonical link, for which the computational savings achieved by the Gauss-Seidel step are often very high, some of the blocks with small number of parameters may be combined into larger blocks within the budgeted per iteration cost of procedure. The advantage of an algorithm based on a coarser partition is that it typically has an improved convergence rate. More formally, a sequence of blocks $\boldsymbol{\alpha}_{t_j}$ may be combined into a single block if the additional cost per iteration of the order $n_t \left(p_t + \sum_{t=j}^{r} p_{t_j}\right)^2 - \sum_{t=j}^{r} n_{t_j} p_{t_j}^2 - n_t p_t^2$ is affordable. In the NSW example of Section 3.1, this strategy was used to jointly update the model intercept, the population density coefficient and the SEIFA interactions in Sub-step B.

2.5 Convergence

An equivalent description of the block Gauss-Seidel procedure. Let $\boldsymbol{s} = (\boldsymbol{\alpha}, \boldsymbol{b}, \boldsymbol{\theta})$ denote the vector of model parameters partitioned into $(r+2)$ blocks: $\boldsymbol{s}_t = \boldsymbol{\alpha}_t$ for $t = 1, \ldots, r$, $\boldsymbol{s}_{r+1} = \boldsymbol{b}$ and

$\boldsymbol{s}_{r+2} = \boldsymbol{\theta}$. Define the vector $\boldsymbol{F}(\boldsymbol{s}) = (\partial G(\boldsymbol{s})/\partial \boldsymbol{\alpha}, \partial G(\boldsymbol{s})/\partial \boldsymbol{b}, \partial q_1(\boldsymbol{s})/\partial \boldsymbol{\theta})$, where as before, $G(\cdot)$ represents Green's PQL criterion and $q_1(\cdot)$ is the quasi-likelihood (4). The Section 2.1 procedure iteratively solves the non-linear system of equations, $\boldsymbol{F}(\boldsymbol{s}) = \boldsymbol{0}$.

Proposition 2.5 *For a GLMM with canonical link and unknown variance components, the Section 2.1 procedure converges to a local solution of* $\boldsymbol{F}(\boldsymbol{s}) = \boldsymbol{0}$ *starting from any initial value in the parameter space. For the special case of a GLMM with known variance components, the algorithm converges* globally *to the unique solution of* $\boldsymbol{F}(\boldsymbol{s}) = \boldsymbol{0}$. *Global convergence is also guaranteed for a GLM with canonical link.*

Proof. We briefly outline the proof here. Using the most recent parameter estimates, define the matrix $\boldsymbol{A}_t = \boldsymbol{X}_t^T \boldsymbol{W} \boldsymbol{X}_t$ and vector $\boldsymbol{c}_t = \boldsymbol{A}_t \cdot \hat{\boldsymbol{\alpha}}_t^{old} + \left(\boldsymbol{Y} - \hat{\boldsymbol{\eta}}^{old}\right) - \boldsymbol{D}^{-1}\hat{\boldsymbol{b}}^{old}$ for blocks $t = 1, \ldots, r$. Also define $\boldsymbol{A}_{r+1} = \boldsymbol{Z}^T \boldsymbol{W} \boldsymbol{Z} + \boldsymbol{D}^{-1}$, $\boldsymbol{c}_{r+1} = \boldsymbol{A}_{r+1} \cdot \hat{\boldsymbol{b}}^{old} + \boldsymbol{Z}^T \boldsymbol{W} \left(\boldsymbol{Y} - \hat{\boldsymbol{\eta}}^{old}\right) - \boldsymbol{D}^{-1}\hat{\boldsymbol{b}}^{old}$, $\boldsymbol{A}_{r+2} = \boldsymbol{H}^{(1)}$ and $\boldsymbol{c}_{r+2} = \boldsymbol{g}^{(1)} + \boldsymbol{g}^{(4)}$. For a canonical link function, the system $\boldsymbol{F}(\boldsymbol{s}) = \boldsymbol{0}$ is equivalent to iteratively solving the system $\boldsymbol{A}_t \cdot \hat{\boldsymbol{s}}_t = \boldsymbol{c}_t$ for $t = 1, \ldots, (r+2)$.

Let $\boldsymbol{L}_t(\boldsymbol{s})$ be the lower triangular part (including the diagonal) of matrix $\boldsymbol{A}_t$. Let $\boldsymbol{U}_t(\boldsymbol{s})$ be the strictly upper triangular part of matrix $\boldsymbol{A}_t$. Let the spectral radius of matrix $\boldsymbol{M}$ be denoted by $\rho(\boldsymbol{M})$. If the current state $\boldsymbol{s}^{(0)}$ is close to a local solution, say $\boldsymbol{s}^*$ satisfying $\boldsymbol{F}(\boldsymbol{s}^*) = \boldsymbol{0}$, then the Gauss-Seidel procedure converges to $\boldsymbol{s}^*$ provided the spectral radii $\rho\left(-\boldsymbol{L}_t^{-1}(\boldsymbol{s}^*)\boldsymbol{U}_t(\boldsymbol{s}^*)\right) < 1$ for all $t = 1, \ldots, (r+2)$. The condition is satisfied because the matrices $\boldsymbol{A}_t$ are symmetric and positive definite.

Consider a GLMM with known variance components. For a canonical link function, the matrix $\partial \boldsymbol{F}(\boldsymbol{s})/\partial \boldsymbol{s}$ is symmetric and positive definite for all $\boldsymbol{s}$ (unlike GLMM's with unknown variance components for which $\partial \boldsymbol{F}(\boldsymbol{s})/\partial \boldsymbol{s}$ need not be symmetric). Applying the Global SOR Theorem (Ortega and Rheinboldt, 2000), we conclude that for any initial value, the Gauss-Seidel procedure converges to the unique solution of $\boldsymbol{F}(\boldsymbol{s}) = \boldsymbol{0}$. The same argument applies to a GLM having only fixed effects as the parameters. Hence shown.

Remark: As with other implementations of the PQL algorithm (refer to Thisted, 1988), convergence difficulties are possible for non-canonical links, especially if the model does not provide a good fit to the data near a solution to $\boldsymbol{F}(\boldsymbol{s}) = \mathbf{0}$.

3 APPLICATION

Burden et al. (2005) analyze data from the Spatial Environmental Epidemiology in New South Wales (SEE NSW) project. Due to the computational demands of fitting a large administrative database, they consider CAR models applied to data collapsed over time, age and gender. In this section, we apply the framework developed in Section 2 to do a full analysis.

Outcome data on ischemic heart disease (IHD) consisting of nearly 33 million observations were abstracted from daily separation records from all public and private hospitals in New South Wales during the period July 1, 1996 to June 30, 2001. Patient reported residential postcode was used to assign the geographical location of hospitalization for IHD. In addition to postcode of residence, available data included the date of hospitalization, patient age and patient gender. Patient age was grouped into one of the following categories: younger than 20 years, 13 different 5-year intervals (20-24 years, 24-29 years etc. up to 80-84 years), plus a 15^{th} category of 85 years or older. Population data were obtained from census information collected by the Australian Bureau of Statistics (ABS) and inter-censal estimates, called Estimated Residential Populations (ERPs), provided for July 1st of each non-census year.

The objective is to explore the association of IHD with an index of socioeconomic disadvantage, the SEIFA (Socio-Economic Indexes for Areas) index, provided by the Australian Bureau of Statistics. It is convenient to adopt a subscript notation with h denoting area (postcode), j indexing time and k indexing social categories consisting of unique combinations of age and gender. Let Y_{hjk} denote the number of IHD hospitalizations in the h^{th} postcode ($h = 1, \ldots, 591$), among the N_{hjk} subjects at risk on day j ($j = 1, \ldots, 1826$) and belonging to the k^{th} social category ($k = 1, \ldots, 30$). Let seifa_{hj} and

dens_{hj} respectively denote the SEIFA index and population density of postcode h on day j. Because IHD is relatively rare, we assume the Poisson approximation to the binomial and fit the model:

$$Y_{hjk} \sim Po(\mu_{hjk}), \text{ where } \log(\mu_{hjk}) = \log(N_{hjk}) + \delta_k + \alpha_w + \beta_0 + \beta_{1k} \cdot \text{seifa}_{hj} + \beta_2 \cdot \text{dens}_{hj} + b_h, \quad (15)$$

where δ_k is the effect associated with the k^{th} social category (with δ_1 being the reference group whose effect is assumed to be zero), α_w is the effect of the w^{th} day of the week (with $w = 1$ representing Mondays as the reference group), β_0 is the intercept, β_{1k} is the interaction term between socioeconomic status (measured by the SEIFA index) and the k^{th} social category, β_2 is the linear predictor associated with the population density, and b_h is the random effect associated with postcode h. Intercept β_0 is estimable because of the identifiability assumptions on the factor effects.

Available information also includes the neighborhood structure of the 591 postcodes. To account for overdispersion and spatial dependence in the IHD hospitalization rates, we assume that the area-specific random effects are distributed according to the CAR model (1).

The size of the data set ($n \approx 33$ million) makes it impossible to apply standard software to fit the above model. Many epidemiologists will routinely handle very large databases by doing an age/sex standardization and then collapsing over these strata. We used the method discussed in Section 2 to compute PQL estimates of the model (15) parameters. Our proposed algorithm allowed us to perform a full epidemiological analysis that looks at the effect of SEIFA, while simultaneously adjusting for age group, gender, day of week and time trend, while at the same time incorporating the spatial effects of location. If we worked with aggregated data instead, it would be necessary to assume that some of these effects were unimportant. By avoiding the need for collapsing, our approach allows us to explore the possible interaction effects of age and gender with the SEIFA variable and find some compelling interactions described in Section 3.2.

3.1 Updating the fixed effects of model (15)

Assume that the random effects in model (15) are equal to their current estimate, $\hat{\boldsymbol{b}}$. The fixed effects are updated as follows using blocked Gauss-Seidel steps. The blocks were created using the technique described in Section 2.4. The partition P (see Section 2.4) had 38 blocks, consisting of the SEIFA interaction term β_{11}, 29 blocks containing the two fixed effects (β_{1k} and δ_k) related to the social categories $k = 2, \ldots, 30$, and eight more blocks comprised of the six non-zero day of week effects, population density coefficient (β_2) and the model intercept (β_0). The first 30 blocks are related to the social categories and updated as described below in Sub-step A. The remaining eight blocks were recombined into a single block (see the end of Section 2.4) and updated as in Sub-step B.

Sub-step A: Treat the random effects and fixed effects β_0, β_2 and $\{\alpha_w\}$ as known and equal to their current estimates. For social category $k = 1, \ldots, 30$, we have:

$$\begin{aligned} Y_{hjk} &\sim Po\left(u_{hj}^{(k)} \cdot \exp\left(\delta_k + \beta_{1k} \cdot \text{seifa}_{hj}\right)\right), \\ \text{where the offset } u_{hj}^{(k)} &= N_{hjk} \cdot \exp\left(\hat{\beta}_0 + \hat{\alpha}_w + \hat{\beta}_2 \cdot \text{dens}_{hj} + \hat{b}_h\right). \end{aligned}$$

For the sub-model indexed by $k = 1, \ldots, 30$, the estimates of parameters δ_k and β_{1k} are updated using the IWLS algorithm and the subsetted data $\{Y_{hjk} \mid k \text{ fixed}\}$ consisting of about a million records. It is sufficient to perform just one or two iterations of the IWLS algorithm at this step. The required number of iterations can be specified in `R` by the `maxit` option of the `glm` command.

Sub-step B: Treating all parameters except those related to the social categories as known and equal to their current estimates, yields:

$$\begin{aligned} Y_{hj.} &\sim Po\left(s_{hj} \cdot \eta_{hj}\right), \\ \text{where } \eta_{hj} &= \exp\left(\beta_0 + \alpha_w + \beta_2 \cdot \text{dens}_{hj}\right) \text{ and } s_{hj} = \exp\left(\hat{b}_h\right) \cdot \sum_k N_{hjk} \cdot \exp\left(\hat{\delta}_k + \hat{\beta}_{1k}\text{seifa}_{hj}\right). \end{aligned}$$

In the above expression, $Y_{hj.}$ equals $\sum_k Y_{hjk}$ and s_{hj} represents the known model offset. The parameters of the sub-model, β_0, β_2 and $\{\alpha_w\}$, are updated using a standard Poisson regression model and the collapsed data $\{Y_{hj.}\}$ consisting of about 1 million cases. A fast computer with a decent amount of memory can be easily used to fit the sub-model.

3.2 Results

The blocked Gauss-Seidel procedure (including the initialization step) took about 50 minutes to converge on a 4 GB Linux machine. Table 3 presents some of the estimated parameters of model (15). The parameters related to SEIFA have been scaled by a factor of 10^{-3}. We also fit the following GLM:

$$Y_{hjk} \sim Po(\mu_{hjk}), \text{ where } \log(\mu_{hjk}) = \log(N_{hjk}) + \delta_k + \alpha_w + \beta_0 + \beta_{1k} \cdot \text{seifa}_{hj} + \beta_2 \cdot \text{dens}_{hj}. \quad (16)$$

This model is nested within GLMM (15). Starting with arbitrary initial values for the fixed effects and setting the random effects identically equal to zero, model (16) was recursively fit using Step 1a of the Section 2.1 algorithm. The procedure converged in about 31 minutes. Notice that the initialization step is unnecessary for GLMs. Some of the fitted values are displayed in Table 3 for comparison. The remaining parameters of model (16) were very similar to those of GLMM (15) and are discussed below.

Figure 1 plots the estimated main effects of the age/gender categories of model (15). Recall that the youngest age group of the male subpopulation is the reference group. The age/gender category effects are found to have similar trends for both genders. As expected, the IHD rates increase sharply with age. Males have generally higher rates than females of the same age. The day-of-week effects are shown in Figure 2. The effect for Monday (the reference group) is found to be significantly higher. There is a sharp decrease in the estimated effects over the weekend. This is a commonly observed pattern of emergency room visits for certain diseases. A possible explanation is that people tend to ignore warning health signs during the weekends, resulting in a high number of emergency room visits on Mondays. Job-related stress might be another reason for the higher effects on weekdays.

Figure 3 displays the estimated interactions of age/gender category and SEIFA for males. The estimates for females had a very similar trend. Most of the interactions have negative estimates indicating that people with high social disadvantage (smaller SEIFA values) are more likely to develop IHD. The interactions increase with age until they are significantly positive for the oldest age group in both the genders. The increase reflects the greater vulnerability to IHD with age, suggesting that as people grow older, socioeconomic status plays a less important role in determining the risk of heart disease.

4 SIMULATION: NON-COLLAPSIBLE MODELS

We investigated the performance of the algorithm with non-Poisson models, for which the collapsibility property cannot be exploited to simplify calculations. Data similar in structure to the NSW study were generated from a binomial GLM with logit link:

$$\begin{aligned} Y_{hjk} \sim \text{Bin}(N_{hjk},\, p_{hjk}), \qquad & \text{where } p_{hjk} = e^{\eta_{hjk}} / \left(e^{\eta_{hjk}} + 1\right), \\ & \text{and } \eta_{hjk} = \delta_k + \alpha_w + \beta_0 + \beta_{1k} \cdot \text{seifa}_{hj} + \beta_2 \cdot \text{dens}_{hj}, \end{aligned} \tag{17}$$

where $h = 1, \ldots, 591$ postcodes, $j = 1, \ldots, 1000$ days and $K = 20$ social categories, totalling approximately 12 million cases. The SEIFA indices and population densities of the postcodes were assumed to be constant over time and were respectively generated from the distributions $N(0, 25^2)$ and exponential with mean 1. The parameters in (17) were generated as follows: the 6 non-zero day of week effects (with Monday as the reference category) were sampled from $U[0,\, 1]$. The $K - 1 = 19$ non-zero social category effects (with social category 1 as the reference group) were generated from $U[0.1,\, 1]$. The $K = 20$ social category–SEIFA interactions were computed as $\beta_{1k} = U_k/10^3 - 1.5 \times 10^{-3}$, where $U_k \overset{iid}{\sim} U[0,\, 1]$. Finally, we set the intercept $\beta_0 = 0.1$ and the population density effect $\beta_2 = 0.25$.

When we tried fitting model (17) to the approximately 12 million cases using the `glm` function of R, it quit with an "insufficient memory" message. We applied the block Gauss-Seidel algorithm of

Section 2.1. For a target relative change of 10^{-8} or less, the algorithm converged in 30 iterations taking about 78 minutes on a 4 GB Linux machine. Logarithms of the estimated and true parameter values are plotted in Figure 4. The logarithms rather than the actual values have been displayed to accommodate the differences in scale. The points lie very close to the 45° line through the origin, indicating that the converged estimates are very close to the values used to generate the data.

Although it is a reasonable model for the NSW case study, model (17) could not be fit to the full data set of 33 million cases using the algorithm for non-collapsible models.

5 DISCUSSION

The paper proposes an efficient implementation of PQL estimation for GLMMs that is especially useful for analyzing large sample sizes. An initialization step provides good starting values for the main iterative procedure, which then performs block-wise Gauss-Seidel updates of the model parameters. Convergence is guaranteed under mild theoretical conditions. The algorithm is most effective for the Poisson models often used in disease mapping problems because of their special collapsibility property. Useful blocking strategies are applied to speed up algorithmic convergence and further reduce computational costs.

The proposed algorithm is applied to study the association between incidence of ischemic heart disease and socioeconomic status in NSW, Australia, using a Poisson model with spatially varying random effects. The outcome data consisting of approximately 33 million records cannot be fit using standard implementations of the PQL procedure. A simulation study further demonstrates the substantial benefits of the algorithm by fitting a non-collapsible, logistic regression model to generated data that are similar in structure to that of the NSW case study and consist of approximately 12 million records.

REFERENCES

Banerjee, S., Carlin, B. and Gelfand, A. (2004). *Hierarchical Modeling and Analysis for Spatial Data*, Chapman & Hall/CRC, Florida.

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

Besag, J., Mollie, A. and York, J. (1991), "Bayesian image restoration, with two applications in spatial statistics," *Annals of the Institute of Statistical Mathematics*, 43, 1–20.

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

Burden S., Guha, S. Morgan, G., Ryan, L. and Young L. (2005), "Spatio-temporal Analysis of Ischemic Heart Disease in NSW, Australia," *Environmental and Ecological Statistics*, 12, 427–448.

Clayton, D. and Kaldor, J. (1987), "Empirical Bayes estimates of age-standardized relative risks for use in disease mapping," *Biometrics*, 43, 671–681.

Elliott, P., Wakefield, J. C., Best, N. G. et al. (2001), *Spatial Epidemiology: Methods and Applications Studies*, Oxford University Press, Oxford, UK.

Givens, G. H. and Hoeting, J. A. (2005), *Computational Statistics*, John Wiley & Sons, Inc, Hoboken, NJ.

Lindley, D. V. and Smith, A. F. M. (1972), "Bayes estimates for the linear model," *Journal of the Royal Statistical Society, Series B*, 34, 1–41.

Leroux, B. G., Lei, X. and Breslow, N. (2000), "Estimation of spatial disease rates in small areas: a new mixed model for spatial dependence," *The IMA Volumes in Mathematics and its Applications, Statistical Models in Epidemiology, the Environment, and Clinical Trials* (Halloran, M. E. and Berry, D. eds.), 116, 179–191, Springer-Verlag, New York.

McCullagh, P. and Nelder, J. A. (1999), *Generalized Linear Models* (2nd ed.), CRC Press LLC, Boca

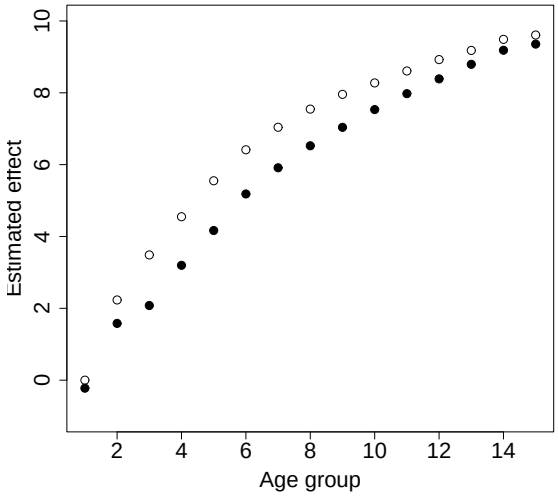

Figure 1: Estimated effects of social category for the 15 age groups of both genders. The open circles represent males and the solid circles represent females. Standard errors are omitted for ease of presentation.

Raton, Florida.

Ortega, J. M. and Rheinboldt, W. C. (2000), *Iterative Solution of Nonlinear Equations in Several Variables*, Society for Industrial and Applied Mathematics.

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

Thisted, R. A. (1988), *Elements of Statistical Computing*, CRC Press LLC, Boca Raton, Florida.

Trewin, D. (2003), "Socio-Economic Indexes for Areas, Australia, 2001," Information Paper ABS Catalogue No: 2039.0, Australian Bureau of Statistics, Canberra, Australia.

Wolfinger, R., Tobias, R. and Sall, J. (1994), "Computing Gaussian Likelihoods and Their Derivatives for General Linear Mixed Models," *SIAM Journal of Statistical Computing*, 15, 1294–1310.

	GLMM (15)			GLM (16)		
Parameter	**Estimate**	**SE**	**P-Value**	**Estimate**	**SE**	**P-Value**
Tuesday	-0.061	0.0059	< 0.001	-0.061	0.0059	< 0.001
Wednesday	-0.072	0.006	< 0.001	-0.072	0.0059	< 0.001
Thursday	-0.064	0.0059	< 0.001	-0.064	0.0059	< 0.001
Friday	-0.059	0.0059	< 0.001	-0.059	0.0059	< 0.001
Saturday	-0.144	0.0062	< 0.001	-0.144	0.0061	< 0.001
Sunday	-0.149	0.0062	< 0.001	-0.149	0.0062	< 0.001
intercept	-18.11	0.0023	< 0.001	-18.07	0.0022	< 0.001
density/10^5	0.546	0.0934	< 0.001	1.28	0.0908	< 0.001
σ^{-2}	2.739	0.1803	< 0.001	–	–	–
λ	0.85	0.0559	< 0.001	–	–	–
	SEIFA interactions/10^3					
Men $<$ 20 yrs	-0.499	2.845	0.861	-1.027	2.783	0.712
Men 20-24 yrs	1.618	1.883	0.39	0.929	1.842	0.614
Men 25-29 yrs	-5.077	0.8471	< 0.001	-5.435	0.8153	< 0.001
Men 30-34 yrs	-4.029	0.522	< 0.001	-4.407	0.5036	< 0.001
Men 35-39 yrs	-3.534	0.3164	< 0.001	-3.884	0.3062	< 0.001
Men 40-44 yrs	-2.983	0.2199	< 0.001	-3.363	0.2135	< 0.001
Men 45-49 yrs	-2.838	0.1678	< 0.001	-3.283	0.1635	< 0.001
Men 50-54 yrs	-2.78	0.1357	< 0.001	-3.279	0.1326	< 0.001
Men 55-59 yrs	-2.143	0.1259	< 0.001	-2.627	0.1234	< 0.001
Men 60-64 yrs	-1.986	0.1188	< 0.001	-2.43	0.1165	< 0.001
Men 65-69 yrs	-1.143	0.1104	< 0.001	-1.533	0.1081	< 0.001
Men 70-74 yrs	-0.719	0.1032	< 0.001	-1.112	0.101	< 0.001
Men 75-79 yrs	-0.29	0.1085	0.008	-0.714	0.1062	< 0.001
Men 80-84 yrs	-0.097	0.1261	0.442	-0.512	0.1235	< 0.001
Men $\geq$ 85 yrs	1.396	0.1468	< 0.001	0.905	0.1439	< 0.001
Women $<$ 20 yrs	-1.898	3.161	0.548	-2.388	3.086	0.439
Women 20-24 yrs	-7.912	1.864	< 0.001	-8.02	1.79	< 0.001
Women 25-29 yrs	-5.727	1.603	< 0.001	-6.015	1.537	< 0.001
Women 30-34 yrs	-5.486	0.9349	< 0.001	-5.795	0.8994	< 0.001
Women 35-39 yrs	-5.965	0.5565	< 0.001	-6.211	0.5374	< 0.001
Women 40-44 yrs	-4.583	0.3751	< 0.001	-4.919	0.3628	< 0.001
Women 45-49 yrs	-4.764	0.2684	< 0.001	-5.147	0.2598	< 0.001
Women 50-54 yrs	-4.374	0.2144	< 0.001	-4.866	0.2089	< 0.001
Women 55-59 yrs	-3.569	0.1935	< 0.001	-4.049	0.1892	< 0.001
Women 60-64 yrs	-2.781	0.1679	< 0.001	-3.195	0.1639	< 0.001
Women 65-69 yrs	-2.118	0.1446	< 0.001	-2.514	0.1411	< 0.001
Women 70-74 yrs	-1.356	0.1235	< 0.001	-1.783	0.1203	< 0.001
Women 75-79 yrs	-0.516	0.1127	< 0.001	-0.989	0.11	< 0.001
Women 80-84 yrs	0.222	0.1119	0.047	-0.268	0.1095	0.014
Women $\geq$ 85 yrs	1.466	0.1053	< 0.001	0.958	0.1031	< 0.001

Table 3: Estimates of selected parameters of models (15) and (16).

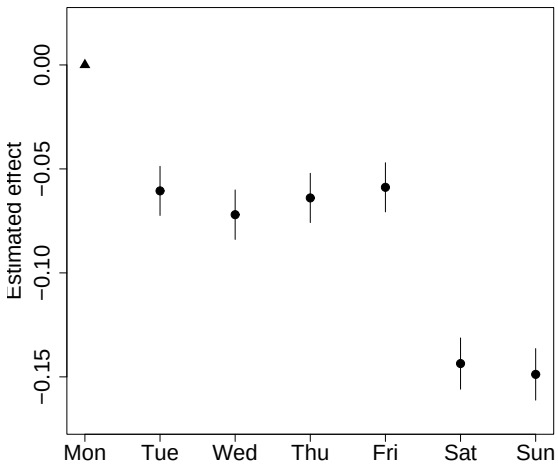

Figure 2: Estimated day-of-week effects. Monday is the reference group. The lines represent intervals of two standard errors.

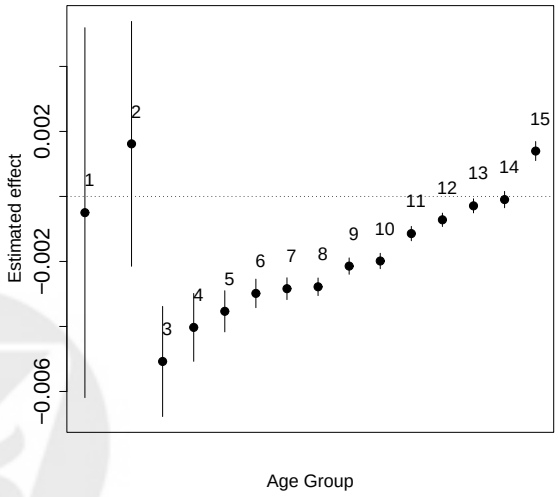

Figure 3: Estimated SEIFA interactions for the 15 age groups of the male subpopulation. The lines represent margins of two standard errors.

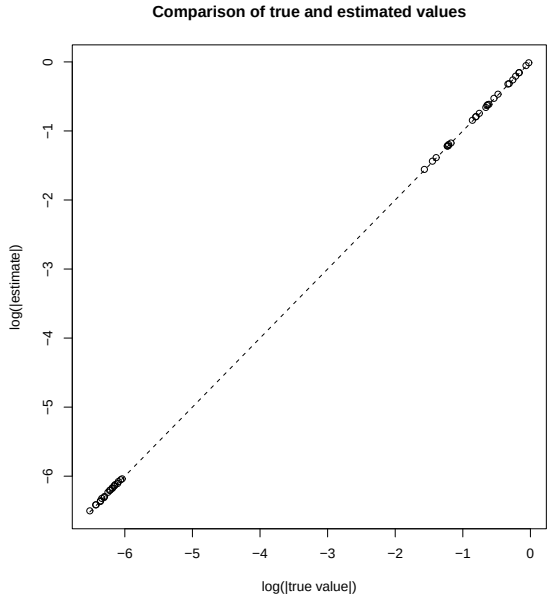

Figure 4: Logarithm of the absolute values of the true and estimated effects for the data generated from a binomial GLM with logit link. The data were fit using the blocked Gauss-Seidel algorithm. A 45 degree line through the origin is displayed for comparison.