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

Subharup Guha, Louise Ryan and Michele Morara*

November 16, 2009
Keywords: Backfitting algorithm, CAR model, collapsibility, epidemiology, iterative weighted least squares, large sample sizes.

Abstract

Generalized linear mixed models (GLMMs) are often fit by computational procedures such as penalized quasi-likelihood (PQL). Special cases of GLMMs are generalized linear models (GLMs), which are often fit using algorithms like iterative weighted least squares (IWLS). High computational costs and memory space constraints make it difficult to apply these iterative procedures to data sets having a very large number of records.

We propose a computationally efficient strategy based on the Gauss-Seidel algorithm that iteratively fits sub-models of the GLMM to collapsed versions of the data. 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. For Poisson and binomial regression models, the Gauss-Seidel approach is found to substantially outperform existing methods in terms of maximum analyzable sample size. Remarkably, for both models, the average time per iteration and the total time until convergence of the Gauss-Seidel procedure are less than 0.3% of the corresponding times for the IWLS algorithm. Platform-independent pseudo-code for fitting GLMS, as well as the source code used to generate and analyze the data sets in the simulation studies, are available online as supplemental materials.

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
mixture 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 postal codes, a Poisson GLMM with log link and normally distributed random effects is a reasonable model for the data (Clayton and Kaldor, 1987). Generalized linear models (GLMs) are often used 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. These challenges make exact approaches, such as fully Bayesian approaches, computationally prohibitive. There has been a large literature on the development of alternative algorithms for fitting GLMMs that are computationally simpler, but only approximate the maximum likelihood of the parameters. For instance, Breslow and Clayton (1993), also see Pinheiro and Bates (2000), proposed the penalized quasi-likelihood (PQL) as a computationally appealing alternative to full maximum likelihood and fully Bayesian approaches. A number of other 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). Wolfinger, Tobias and Sall (1994), and more recently Bates and DebRoy (2004), discuss computationally efficient methods for implementing PQL.

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 even standard PQL 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 \( y = (y_1, \ldots, y_n)^T \) 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 \( \eta_i \) is the linear predictor (McCullagh and Nelder, 1999).

Let \( \alpha = (\alpha_1, \ldots, \alpha_p)^T \) be the vector of fixed effects. The vector of random effects, \( b = (b_1, \ldots, b_q)^T \), is assumed to follow a multivariate normal distribution with mean zero and covariance matrix \( D(\theta) \), where \( \theta = (\theta_1, \ldots, \theta_r)^T \) 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 \( \eta_i = x_i^T \alpha + z_i^T b \), where \( x_i^T \) and \( z_i^T \) respectively denote the \( i^{th} \) rows of design matrices \( X \) and \( Z \) of dimensions \( n \times p \) and \( n \times q \). For most commonly used GLMMs, the variance \( \text{Var}(y_i) \) 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 \( \mathbf{b} = (b_1, \ldots, b_q)^T \) of region-specific random effects follows a \( q \)-variate normal distribution with mean zero and covariance matrix \( D(\theta) \). In many spatial epidemiology studies, \( Z \) is a 0-1 matrix assigning the responses to one of the \( q \) regions: for observation \( i = 1, \ldots, n \) and region \( s = 1, \ldots, q \), element \( z_{is} \) of the design matrix \( Z \) 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 \( R \) 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 \( \mathbf{b} \) have a multivariate normal distribution with mean \( \mathbf{0} \) and covariance matrix \( D \) whose Moore-Penrose generalized inverse is \( D^{-} = R/\sigma^2 \).

The variance component \( \theta \) of the intrinsic autoregressive model consists of a single parameter, the precision \( \sigma^{-2} \), 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 \( \theta = (\sigma^{-2}, \lambda) \). The covariance matrix \( D \) of the random effects is assumed to satisfy

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

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

### 1.3 Standard implementations of the PQL approach

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 \( \theta \) 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:

\[
H \cdot \begin{pmatrix} \alpha \\ b \end{pmatrix} = \begin{bmatrix} X^T W Y \\ Z^T W Y \end{bmatrix}, \tag{2}
\]
where the matrix
\[ H = \begin{bmatrix} X^T W X & X^T W Z \\ Z^T W X & Z^T W Z + D^{-1} \end{bmatrix}. \]  

(3)

The symbol \( Y = (Y_1, \ldots, Y_n) \) in the aforementioned expressions represents the vector of working values and not the data, \( y \). The \( n \times n \) diagonal matrix of working weights is denoted by \( W \). 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: \( Y_i = \eta_i + (y_i - \mu_i)\frac{\partial \eta_i}{\partial \mu_i} \). 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: \( \hat{\alpha} = (X^T V^{-1}X)^{-1} X^T V^{-1} y \) and \( \hat{b} = V^{-1} Z^T W (Y - X\hat{\alpha}) \).

**Step 2:** Assuming the fixed effects to be equal to their current estimate \( \hat{\alpha} \), we estimate the variance components. There are several possible approaches. A natural option is to maximize the quasi-likelihood
\[ ql_1(\hat{\alpha}, \theta) = -\frac{1}{2} \log |V| - \frac{1}{2} r^T V^{-1} r. \]  

(4)

where \( r = Y - X\hat{\alpha} \) and \( V = W^{-1} + ZDZ^T \) is an \( n \times n \) matrix. Alternatively, Breslow and Clayton (1993) recommend using the REML version of the quasi-likelihood:
\[ ql_2(\hat{\alpha}, \theta) = -\frac{1}{2} \log |V| - \frac{1}{2} \log |X^T V^{-1}X| - \frac{1}{2} r^T V^{-1} r. \]  

(5)

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

$$ql_3(\theta) = -\frac{1}{2} \log |V| - \frac{1}{2} \rho^T V^{-1} \rho$$

$$ql_4(\theta) = -\frac{1}{2} \log |V| - \frac{1}{2} \log |X^T V^{-1} X| - \frac{1}{2} \rho^T V^{-1} \rho$$

(6)

The value of $\theta$ that maximizes the objective functions in (4), (5) and (6) can be iteratively computed by either Fisher scoring or Newton-Raphson method. Further details of the calculation are available in Guha and Ryan (2006). The identity

$$V^{-1} = W - WZ (D^{-1} + Z^T WZ)^{-1} Z^T W$$

(7)

appearing in Wolfinger et al. (1994) is key. It avoids the inversion of the non-diagonal matrix $V$, an operation having $O(n^3)$ computational cost per iteration, and provides an alternative expression relying only on the submatrices in (3). It is therefore of $O(n)$ computational cost. Bates and DebRoy (2004) discuss an alternative framework for profiling-based approaches.

**IWLS algorithm.** For estimating the fixed effects $\alpha$ 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, $X^T W X \cdot \alpha = X^T W Y$. 

8
1.4 Computational challenges with large sample sizes

Our motivating example is the case study of Section 4. The outcome data consist of the number of emergency room visits diagnosed as ischemic heart disease (IHD) in New South Wales (NSW), Australia. The data are grouped by patient’s resident postcode, day of visit and age/gender category. With 591 postcodes, 1,826 days and 30 age/gender categories, there are more than 33 million observations. Available covariates from the census include the population at risk, population density and socioeconomic status associated with a postcode on a given day. The goal is to study the association between IHD and socioeconomic status, and investigate how the association varies with age and gender.

For data sets with 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 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 \(H\) in (2). 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 \(X^T V^{-1} X\). Applying identity (7), we have that

\[
X^T V^{-1} X = X^T W X - X^T W Z \left( D^{-1} + Z^T W Z \right)^{-1} Z^T W X. \tag{8}
\]

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

For the variance components, it can be shown that Newton-Raphson updates, profile likelihoods, and the REML quasi-likelihood (5), all rely on the matrix \(X^T V^{-1} X\) (e.g. see Guha and
Ryan, 2006). As a result, these techniques are infeasible for very large sample sizes.

Memory space can be a major concern. In the NSW example, the required memory for standard PQL updates (implemented in R in the functions `glmmPQL` and `lmer`) far exceeds the current capacity of most computers. This is true for even relatively simple models with no random effects. The memory requirements increase with model complexity, because of which many realistic models cannot be fit to large sample size data sets.

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 applying two strategies: (i) The fixed and random effects are separately updated in blocks by applying the Gauss-Seidel algorithm. (ii) Among the various options for updating the variance components, we maximize quasi-likelihood $q_1(\hat{\alpha}, \theta)$ via Fisher scoring. Innovation (i) can be implemented through standard model-fitting software. Additionally, it exploits the structure of the design matrix and special model properties, like the collapsibility property of GLMMs, to achieve further gains in computational efficiency.

The rest of the paper is organized as follows. Section 2 describes a procedure for updating the fixed effects via the blocked Gauss-Seidel algorithm. Section 2.3 theoretically specifies why this approach results in substantial savings in the computational cost per iteration and memory usage. These savings are typically obtained at the price of a slower convergence rate. The drawback can be largely overcome by a judicious choice of blocks, for which Section 2.4 recommends useful strategies. Section 2.5 proves that convergence of the iterative procedure is guaranteed for exponential family models with canonical link. In the special case of GLMs and GLMMs with known variance components, the algorithm converges to the global solution.

Section 4 applies these ideas to the NSW data and investigates the relationship between heart
disease and socioeconomic factors. The simulation studies of Section 3 compare the blocked Gauss-Seidel algorithm with the IWLS algorithm for Poisson and logistic regression models. The Gauss-Seidel approach is found to be substantially more effective, both in terms of average time per computation and total time until convergence. Furthermore, its efficient utilization of memory allows it to analyze much larger data sets than the IWLS algorithm. The Appendix provides platform-independent pseudo-code for the blocked Gauss-Seidel algorithm.

2 A COMPUTATIONALLY EFFICIENT ALGORITHM FOR LARGE SAMPLE SIZES

2.1 Gauss-Seidel (backfitting) procedure

In a general setting, let the vector $\psi$ represent the model parameters. Let $\psi^{(m)}$ be the estimates at the end of the $m^{th}$ iteration. The classic Gauss-Seidel algorithm (Givens and Hoeting, 2005, p. 43) computes an updated estimate, $\psi^{(m+1)}$, by performing a series of univariate maximizations of an objective function (like the log-likelihood, quasi-likelihood or Green’s PQL criterion) with respect to individual parameters, $\psi_j$, holding the remaining parameters fixed at their current estimated values. In other words, the $(m + 1)^{st}$ iteration updates the parameter $\psi_j$ 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 $(\psi_{j+1}^{(m)}, \ldots, \psi_p^{(m)})$. Daroch and Ratfield (1972) applied this technique to extend iterative proportional fitting to general Poisson regression settings. Refer to Ma and Hudson (1997) for an application involving high-dimensional problems in emission and transmission tomography.
A number of variations of the classic Gauss-Seidel algorithm 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).

### 2.2 Iterative procedure for GLMMs

Initial values for the model parameters are computed as described in Pinheiro and Bates (2000). 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. For large sample sizes, such a strategy often results in drastic reductions in computational costs as discussed below in detail. However, a potential problem is slow convergence due to correlated parameter estimates being updated in different blocks. Highly correlated fixed effects should therefore be jointly updated in a single block, after suitable reparametrization to reduce multicollinearity in the covariate matrix columns.

Formally, suppose that there are \( r \) blocks of fixed effects, \( \alpha_{\gamma_1}, \ldots, \alpha_{\gamma_r} \), where the index set \( \gamma_t \subset \{1, \ldots, p\} \) for every \( t = 1, \ldots, r \). Assume that the blocks are chosen in such a manner that each of the \( p \) fixed effects occurs in at least one block along with other fixed effects with which
it is strongly correlated. This implies that $\cup_{t=1}^{r} \gamma_t = \{1, \ldots, p\}$. Let $\alpha_{-\gamma_t}$ denote the vector of remaining fixed effects after excluding the block $\alpha_{\gamma_t}$ and let $G(\alpha, b, \theta)$ represent Green’s PQL criterion. For $t = 1, \ldots, r$, let $X_t$ ($X_{-t}$) be the matrix of covariates associated with the $\alpha_{\gamma_t}$ block ($\alpha_{-\gamma_t}$ block). To reduce within-block multicollinearity, we linearly transform the block of fixed effects:

$$
\phi_{\gamma_t} = P_t \cdot \alpha_{\gamma_t}
$$

(9)

where the invertible matrix $P_t$ is of order $|\gamma_t|$ and is chosen so that the columns of $X_t^* = X_t \cdot P_t$ are approximately uncorrelated. It is easy to verify that Green’s PQL criterion is invariant with respect to linear transformations of the fixed and random effects such as (9).

Let the vector of current estimates be $(\hat{\phi}_{\gamma_t}^{old}, \hat{\alpha}_{-\gamma_t}, \hat{b}, \hat{\theta})$ and let $\hat{\eta}^{old} = \eta(\hat{\phi}_{\gamma_t}^{old}, \hat{\alpha}_{-\gamma_t}, \hat{b})$ be the current value of the linear predictor. Applying the iterative approach described in Section 2.1 to the blocks and regarding Green’s PQL criterion as the objective function to be maximized, we maximize $G(\phi_{\gamma_t}, \hat{\alpha}_{-\gamma_t}, \hat{b}, \hat{\theta})$ with respect to $\phi_{\gamma_t}$ and then invert the transformation (9) to obtain updated estimates of $\alpha_{\gamma_t}$ for $t = 1, \ldots, r$. Typically, the maximization cannot be performed in closed form. Updates are then obtained by Fisher scoring, which for GLMMs with canonical links is equivalent to Newton-Raphson updates. The updated block is $\hat{\phi}_{\gamma_t} = \hat{\phi}_{\gamma_t}^{old} + \Delta_t$, where the increment $\Delta_t$ satisfies the equation $-\partial^2 G/\partial \phi_t^2 \cdot \Delta_t = \partial G/\partial \phi_t$ and the partial derivatives are evaluated using the vector of current estimates, $(\hat{\phi}_{\gamma_t}^{old}, \hat{\alpha}_{-\gamma_t}, \hat{b}, \hat{\theta})$.

Similar to the derivation of relation (2) in Breslow and Clayton (1993), we can show that the foregoing Gauss-Seidel update is equivalent to solving the normal equation:

$$
X_t^* W X_t^* \cdot \Delta_t = X_t^* W \tau,
$$

(10)
where the vector of differences $\tau = Y - \hat{\eta}^{\text{old}}$. The vector $Y$ of working values and the diagonal matrix $W$ of working weights are both computed using $\hat{\eta}^{\text{old}}$. After the increment $\Delta_t$ has been computed in (10), the updated linear predictor, $\hat{\eta} = \eta(\hat{\phi}_{\gamma_t}, \hat{\alpha}_{-\gamma_t}, \hat{b})$, can be easily computed as

$$\hat{\eta} = \hat{\eta}^{\text{old}} + X^*_t \Delta_t.$$ 

We observe that equation (10) represents the Fisher scoring updates for a GLM having the same likelihood and link function as the original model, but for which the block of transformed fixed effects, $\phi_{\gamma_t}$, are the unknown parameters, $X^*_t$ is the design matrix, and $\hat{e}_t = X_{-t}\hat{\alpha}_{-\gamma_t} + Z\hat{b}$ is the vector of offsets in the linear predictor, $\eta = \hat{e}_t + X^*_t\phi_{\gamma_t}$. This allows the use of standard model-fitting procedures (e.g. glm function in R) to update the blocks. The vector $\hat{e}_t$ of offsets for the nested GLM can be specified in R through the offset option. Notice that it is necessary to include only those rows of $X^*_t$ that do not correspond to all zeros, leading to a reduction in the dimension of the data set.

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. In R, the number of Newton-Raphson updates can be limited to one or two by setting the maxit option of the glm function. There is considerable amount of flexibility in grouping the fixed effects into blocks. The optimal blocks (with respect to the computational cost per iteration and the total time to convergence) depends on particulars of the model and 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.

We emphasize here that 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.
In fact, dramatic reductions in dimension are often possible if equation (10) is implemented by exploiting a special collapsibility property of GLMMs, an option not currently available in standard statistical software. The dimension reduction is achieved by collapsing the data over the unique, non-zero rows of the matrix $X_t^*$. The reduction can be especially high for factor effects and their interactions, as well as data sets having repeated covariate levels within blocks, because we only need to focus on observations where the covariates take a specific level.

**Collapsibility** GLMMs have a special collapsibility property that allows us to exploit any repeats in the associated covariates at a given stage and to work with summarized versions of the data and covariate matrices. In the examples of Sections 4 and 3, we worked with summaries consisting of approximately million records, rather than the 33 million records for the full data set. 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 $X_t^*$ associated with the block $\phi_{\gamma_t}$ has rows $x_{t1}^T, \ldots, x_{tn}^T$ of which only $n_t \leq n$ rows are non-zero and unique. Denote these unique, non-zero rows by $u_{t1}^T, \ldots, u_{tn_t}^T$ and let the matrix $U_t$ consist of only these $n_t$ number of rows. For $i = 1, \ldots, n_t$, let $w_i = \sum_{j: x_{ij}^* = u_{ti}^*} w_j$ ($\tau_i = \sum_{j: x_{ij}^* = u_{ti}^*} \tau_j$) represent the summed working weights (differences) corresponding to the covariate value of $u_{ti}^*$. These quantities depend on the current estimates of the remaining model parameters, and are computed on the fly as the algorithm proceeds. Define the vector $\tau_t^* = (\tau_1, \ldots, \tau_{n_t})$ and diagonal matrix $W_t^* = \text{diag}(w_1, \ldots, w_{n_t})$. Both quantities are of dimension $n_t \leq n$. It is easy to show that equation (10) is equivalent to the following normal equation:

$$U_t^T W_t^* U_t \cdot \Delta_t = U_t^T W_t^* \tau_t^*.$$  

(11)
Step 1b: Updating the random effects. The random effects are updated given current estimates of the remaining parameters.

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

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

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

Proof. The estimates of the random effects are chosen to maximize Green’s PQL criterion, $G(\alpha, b, \theta)$. Applying the Newton-Raphson method, the current values of the partial derivatives $\partial G / \partial b$ and $-\partial^2 G / \partial b^2$ are, respectively, $Z^T W \left( Y - \hat{\eta}^{old} \right) - D^{-1} \hat{b}^{old}$ and $Z^T W Z + D^{-1}$, and the iterative update, $\Delta \hat{b}$, has the afore-mentioned expression.

Remark: Many epidemiological studies consider models that assign each of the $n$ records to exactly one random effect. Examples include analyses of spatially varying disease rates. The design matrix $Z$ then contains only 0’s and 1’s. This special structure results in $Z^T W Z$ being a diagonal matrix in (12). The $j^{th}$ diagonal element is then easily computed by summing the working weights corresponding to the random effect $b_j$. The vector $Z^T W \left( Y - \hat{\eta}^{old} \right)$ is computed 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. As mentioned earlier, for data sets with large sample sizes, this involves a substantially lower computational cost compared to the other strategies discussed in Section 1.3.

The following proposition provides an expression for the iterative update of the variance components. The results follows from standard normal theory results and relation (7). We omit the proof for brevity.

Proposition 2.2 Suppose that the updated estimates $\hat{\theta}$ are computed by applying Fisher scoring to the quasi-likelihood function (4), so that $\hat{\theta} = \hat{\theta}^{old} + \Delta \hat{\theta}$. Then the iterative update is given by $\Delta \hat{\theta} = (I^{exp})^{-1} s$, where $s$ denotes the score vector and $I^{exp}$ is the expected information matrix. For $i, j = 1, \ldots, q$, the elements of these quantities are given by

$$
\begin{align*}
    s_i &= \frac{1}{2} \hat{b}^T D^{-1} \frac{\partial D}{\partial \theta_i} D^{-1} \hat{b} - \frac{1}{2} \text{tr} \left( (Z^T V^{-1} Z) \frac{\partial D}{\partial \theta_i} \right), \\
    I^{exp}_{ij} &= \frac{1}{2} \text{tr} \left( (Z^T V^{-1} Z) \frac{\partial D}{\partial \theta_i} (Z^T V^{-1} Z) \frac{\partial D}{\partial \theta_j} \right),
\end{align*}
$$

where $D = D(\hat{\theta}^{old})$ and

$$
Z^T V^{-1} Z = D^{-1} (Z^T W Z + D^{-1})^{-1} Z^T W Z.
$$

Remark: On comparing expressions (12) and (14), we find that that Step 1b and Step 2 have approximately the same computational cost. As mentioned in the Remark following Proposition 2.1, the updates are trivial to compute, even for large $n$, when the matrix $Z$ consists only of 0’s
and 1’s.

**Special case: CAR models.** 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 \( D^{-1} \) rather than \( D \), we apply relation (1) and evaluate \( \frac{\partial D^{-1}}{\partial \lambda} = R - I \) and \( \frac{\partial D^{-1}}{\partial \sigma^{-2}} = (1 - \lambda)I + \lambda R \). The identity \( \frac{\partial D}{\partial x} = -D \frac{\partial D^{-1}}{\partial x} D \) gives us the partial derivatives with respect to \( D \), which are substituted in (13) to compute the iterative update for the variance components.

### 2.3 Why is the Gauss-Seidel algorithm feasible 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 \( t = 1, \ldots, r \), suppose the block of fixed effects \( \alpha_{\gamma_t} \) consists of \( p_t = |\gamma_t| \) parameters. Let \( m_t \) be the number of non-zero rows of the covariate matrix \( X_t^* \) associated with the transformed block of fixed effects, \( \phi_{\gamma_t} = P_t \cdot \alpha_{\gamma_t} \). Let \( m_q \) be the number of non-zero rows of the design matrix \( Z \) associated with the random effects. While using standard subroutines for updating the fixed effects, a joint update of all \( p \) fixed effects typically requires the specification of the entire design matrix \( X \) consisting of \( np \) elements. For large sample sizes, the memory requirement often exceeds the capacity of most computers. In contrast, a Gauss-Seidel update of the transformed fixed effects \( \phi_{\gamma_t} \) relies only on the non-zero rows of the matrix \( X_t^* \). This consists of \( m_t p_t \) elements and requires only a fraction \( (m_t/n \cdot p_t/p) \) of the memory. Implementations of the algorithm that exploit collapsibility depend only on the matrix \( U_t \) consisting of \( n_t p_t \) elements, which further reduces the memory requirements by a factor of \( m_t/n_t \geq 1 \).
An additional problem with the Section 1.3 algorithm is the high cost per iteration associated with computing the matrix $H$, which is of the order $n(p+q)^2$. Gauss-Seidel updates of the fixed and random effects (Steps 1a and 1b of Section 2.2) require only the block diagonals of the matrix $H$. The costs per iteration are then of the order $m_t^2$ for block $\alpha_{\gamma_t}$ and order $m_q q^2$ for the random effects, both of which are bounded above by $n(p+q)^2$. Greater computational savings are achieved in implementations that exploit collapsibility, since the cost per iteration is of the order $n_t^2 \leq m_t^2$ for the block $\alpha_{\gamma_t}$. The two costs are equal if and only if the rows of the design matrix $X_t$ have no repeats for all $t = 1, \ldots, r$.

### 2.4 Selection of blocks

There are many ways in which the $p$ fixed effects could be grouped into $r$ blocks. Some groupings result in a more efficient algorithm than others. Relevant criteria for evaluating a given set of blocks include: (i) computational cost per iteration, and (ii) number of iterations to converge. Typically, success with respect to one criterion is achieved at the expense of the other. Consider 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 the number of iterations to converge (on a hypothetical machine with infinite resources) is usually relatively small. For comparison, consider a 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 standard PQL algorithm, convergence could be slow because of potential hem-stitching problems due to highly correlated estimates.

The above observation suggests some practical strategies for creating the blocks. The fixed
effects should be grouped by aligning their zeros in the design matrix $X$. This produces a large number of zero rows in the covariate matrices associated with the blocks, and as explained in Section 2.3, results in lower computational costs per iteration through dimension reduction. Simultaneously, in order to optimize the performance with respect to criterion (ii), the blocks should be chosen so that the fixed effects that are highly correlated are jointly updated within the same block. This can be done by initializing the parameters using the procedure recommended by Pinheiro and Bates (2000), and evaluating the submatrix $Z^TWZ$ for these initial parameter values. We would then group those fixed effects into common blocks that are highly correlated under $(Z^TWZ)^{-1}$, thereby ensuring a fast-converging Gauss-Seidel procedure. Subject to the budgeted cost per iteration, some of the small blocks can be combined into larger blocks resulting in a faster converging, but computationally more expensive, algorithm.

As a further refinement, a fixed effect may appear in two or more blocks, with suitable reparametrization, if it is expected to be strongly correlated with the parameters in those blocks. For example, suppose $\{\alpha_j\}_{j=1}^{p_1}$ and $\{\delta_j\}_{j=1}^{p_2}$ are two sets of factor effects with $\alpha_1 = \delta_1 = 0$ assumed for identifiability with the model intercept, $\beta_0$. Imagine that high computational costs necessitate that the two sets of factor effects be updated in different blocks. Due to multicollinearity, we expect the estimate of the model intercept, $\hat{\beta}_0$, to be highly correlated with all the factor effects estimates. An algorithm that does not include the intercept in both the blocks is likely to have hem-stitching problems and take many iterations to converge. A solution is to jointly update the intercept and factor effects in the two blocks after linear transformation: $\alpha'_j = \beta_0 + \alpha_j$ and $\delta'_k = \beta_0 + \delta_k$ where $j = 1, \ldots, p_1$ and $k = 1, \ldots, p_2$. In the examples of Sections 4 and 3, this strategy was applied to jointly update the model intercept, the population density coefficient, and the day of week effects in one block, and then update the transformed intercept and age/gender
effects in different blocks. Including the intercept in both the blocks reduced the required number
of iterations by a factor of several hundreds.

2.5 Convergence properties

Consider the following equivalent description of the Section 2.2 procedure. Let \( s = (\alpha, b, \theta) \) denote the vector of model parameters partitioned into \( r + 2 \) blocks: \( s_t = \alpha_t \) for \( t = 1, \ldots, r \), \( s_{r+1} = b \) and \( s_{r+2} = \theta \). Define the vector \( F(s) = (\partial G(s)/\partial \alpha, \partial G(s)/\partial b, \partial q_1(s)/\partial \theta) \), where as before, \( G(\cdot) \) represents Green’s PQL criterion and \( q_1(\cdot) \) is the quasi-likelihood function (4). Then the Section 2.2 procedure iteratively solves the non-linear system of equations, \( F(s) = 0 \), via Fisher scoring. With this additional notation, we have the following result.

**Proposition 2.3** For a GLMM with canonical link and unknown variance components, the Section 2.2 procedure converges to a local solution of \( F(s) = 0 \) 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 \( F(s) = 0 \). Global convergence is also guaranteed for a GLM with canonical link.

**Proof.** Define the vectors \( g^{(1)} \) and \( g^{(2)} \) and the matrix \( H^{(1)} \) having typical elements \( g_i^{(1)} = \text{tr} \left( Z^T V^{-1} Z \frac{\partial D}{\partial \theta_i} \right) \), \( g_i^{(2)} = -r^T V^{-1} Z \frac{\partial D}{\partial \theta_i} Z V^{-1} r \), and \( H_{ij}^{(1)} = -\text{tr} \left( Z^T V^{-1} Z \frac{\partial D}{\partial \theta_i} Z^T V^{-1} Z \frac{\partial D}{\partial \theta_j} \right) \). Using the most recent parameter estimates, define the matrix \( A_t = X_t^T W X_t \) and vector \( c_t = A_t \cdot \hat{\alpha}_t^{old} + (Y - \hat{\eta}^{old}) - D^{-1} \hat{b}_t^{old} \) for blocks \( t = 1, \ldots, r \). Let \( A_{r+1} = Z^T W Z + D^{-1} \), \( c_{r+1} = A_{r+1} \cdot \hat{b}^{old} + Z^T W (Y - \hat{\eta}^{old}) - D^{-1} \hat{b}_t^{old} \), \( A_{r+2} = H^{(1)} \) and \( c_{r+2} = g^{(1)} + g^{(2)} \). For a canonical link function, the system \( F(s) = 0 \) is equivalent to iteratively
solving the system \( A_t \cdot \dot{s}_t = c_t \) for \( t = 1, \ldots, (r + 2) \).

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

Consider a GLMM with known variance components. For a canonical link function, the matrix \( \partial F(s)/\partial s \) is symmetric and positive definite for all \( s \) (unlike GLMM's with unknown variance components for which \( \partial F(s)/\partial s \) 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 \( F(s) = 0 \). The same argument applies to a GLM consisting of only fixed effects. Hence shown.

As with other implementations of the PQL algorithm, convergence difficulties with the Gauss-Seidel algorithm may be encountered for non-canonical links, especially if the model does not provide a good fit to the data near a solution to \( F(s) = 0 \) (refer to Thisted, 1988).
3 SIMULATION STUDIES

3.1 Poisson regression

We performed a simulation study to compare the performance of the blocked Gauss-Seidel and IWLS algorithms for Poisson regression models. Both algorithms were implemented in C++ on a 2.4 GHz Intel processor with 1.99 GB of RAM. Data corresponding to 1,827 days and 30 age/gender categories, having similar structure to those of the NSW study, were generated from model (17).

Starting with $H = 1$ postcode, we generated increasingly large data sets of size $n = H \times 1827 \times 30$ records, by increasing $H$ in small increments up to a maximum of $H = 600$ postcodes. The $H \times 1827$ values of SEIFA were independently generated from the standard normal distribution, as were the re-centered population densities. The $n$ number of populations at risk, $N_{hjk}$, were independently sampled from the Poisson distribution with mean $\exp(-\beta_0)$, where the intercept $\beta_0 = -17$. The true values of the remaining model parameters were generated as follows: the 6 non-zero day of week effects (with Monday being the reference category) and the $K - 1 = 19$ non-zero social category effects (with social category 1 as the reference group) were independently generated from the distribution $U[0, 1]$. The age/gender category–SEIFA interactions, $\beta_{1k}$, were generated from the distribution $-U[0, 2]$. The coefficient $\beta_2$ was generated from the standard normal distribution.

The smaller amount of memory required by the Gauss-Seidel algorithm is a key advantage for handling large data sets. For the IWLS algorithm, a 32-bit machine was unable to allocate enough memory for more than $H = 60$ postcodes ($n \approx 3$ million records). In contrast, the Gauss-Seidel algorithm easily analyzed data sets exceeding $H = 600$ postcodes (nearly 33 million records).
For a target absolute change of $10^{-4}$ or less in the parameter estimates, the left panel of Figure 1 plots the total time until convergence for the two algorithms. The converged estimates obtained from both algorithms were within two standard errors of the true parameter values. The average time per iteration for the blocked Gauss-Seidel algorithm was less than 0.3% that of the IWLS algorithm. Moreover, the Gauss-Seidel algorithm converged in fewer number of iterations (13 versus 21 iterations for the IWLS algorithm) and the percentage gain in total convergence time exceeded 99.8%.

### 3.2 Logistic regression

Next, we investigated the performance of the blocked Gauss-Seidel algorithm for logistic regression models. Data similar in structure to the NSW study were generated from a binomial GLM with logit link:

\[
Y_{hjk} \sim \text{Bin}(N_{hjk}, p_{hjk}), \quad \text{where } p_{hjk} = \frac{e^{\eta_{hjk}}}{(e^{\eta_{hjk}} + 1)},
\]

\[
\eta_{hjk} = \delta_k + \alpha_w + \beta_0 + \beta_1 k \cdot \text{seifa}_{hj} + \beta_2 \cdot \text{dens}_{hj}, \quad (15)
\]

where $h = 1, \ldots, H$ postcodes, $j = 1, \ldots, 1827$ days and $K = 30$ social categories, totalling approximately 33 million records. The true values of the parameters were generated as in Section 3.1. We followed a similar procedure, starting with $H = 1$ postcode and generating increasingly large data sets of size $n = H \times 1827 \times 30$ records by increasing $H$ in increments up to a maximum of $H = 600$ postcodes.

For a target absolute change of $10^{-4}$ or less in the parameter estimates, the right panel of Figure 1 plots the total time until convergence for the IWLS and collapsed version of the
blocked Gauss-Seidel algorithms. The results are similar to those for Poisson regression models. Although the IWLS algorithm was unable to allocate enough memory on a 32-bit machine for more than $H = 60$ postcodes (approximately 3 million records), the Gauss-Seidel procedure easily analyzed data sets exceeding $H = 600$ postcodes (nearly 33 million records). All of the converged estimates were within two standard errors of the true parameter values. The percentage gains for the Gauss-Seidel procedure, relative to the IWLS algorithm, exceeded 99.8% for the average times per iteration as well as the total convergence times. This demonstrates the remarkable success and flexibility of the Gauss-Seidel algorithm in fitting data sets having a large number of records.

4 APPLICATION

The Spatial Environmental Epidemiology in New South Wales (SEE NSW) project yielded outcome data on ischemic heart disease (IHD). Approximately 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 15th 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.
Our objective is to explore the association of IHD with an index of socioeconomic disadvantage provided by ABS called SEIFA (Socio-Economic Indexes for Areas) index. Additionally, we wish to investigate how this association varies with age and gender. 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 $\text{seifa}_{hj}$ and $\text{dens}_{hj}$ 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:

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

where $\delta_k$ is the effect associated with the $k^{th}$ social category (with $\delta_1$ being the reference group whose effect is assumed to be zero), $\alpha_w$ is the effect of the $w^{th}$ day of the week (with $w = 1$ representing Mondays as the reference group), $\beta_0$ is the intercept, $\beta_{1k}$ is the interaction term between socioeconomic status (measured by the SEIFA index) and the $k^{th}$ social category, $\beta_2$ is the linear predictor associated with the population density, and $b_h$ is the random effect associated with postcode $h$. Available information includes the neighborhood structure of the 591 postcodes. To account for spatial dependence in the IHD hospitalization rates, we assume that the postcode-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 techniques to fit model (16). Many epidemiologists routinely handle very large databases by doing an age/sex
standardization and then collapsing over these strata. This approach usually leads to biased estimates. The proposed algorithm of Section 2 avoids the need for standardization, allowing us to perform a full epidemiological analysis that looks at the effect of SEIFA, while simultaneously adjusting for confounding factors such as age group, gender, day of week and time trend, and spatial effects of location.

4.1 Results

The fixed effects of model (16) were partitioned into 31 blocks. Blocks 1 through 30 consisted of the model intercept, main effects and interactions associated with the age/gender categories. To avoid multicollinearity problems, we estimated the sum of the intercept and age/gender category factor effects rather than the parameters separately. Block 31 consisted of the model intercept, population density coefficient and day of week effects. For each block, this strategy achieved a 30-fold reduction in the number of records by focussing on subsetted/summarized data consisting of approximately 1 million observations each. Our implementation of the Gauss-Seidel procedure in R took only 50 minutes to converge on a 4 GB Linux machine. Table 1 presents some of the estimated parameters of model (16). The parameters related to SEIFA have been scaled by a factor of $10^{-3}$. For comparison, we also fit the following nested GLM using a Gauss-Seidel procedure:

$$Y_{hjk} \sim Po(\mu_{hjk}), \text{ where } \log(\mu_{hjk}) = \log(N_{hjk}) + \delta_k + \alpha_w + \beta_0 + \beta_{1k} \cdot \text{seifa}_{hj} + \beta_2 \cdot \text{dens}_{hj}. \quad (17)$$

The algorithm converged in about 31 minutes. Some of the fitted values are displayed of the GLM in Table 1 for comparison. The remaining estimates were very similar to those of model (16),
and are discussed below.

Figure 2 plots the estimated main effects of the age/gender categories of model (16). 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 3. 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 4 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.

5 DISCUSSION

This paper proposes an efficient implementation of PQL estimation for GLMMs that is especially useful for analyzing large sample sizes. Convergence of the iterative procedure is theo-
retically guaranteed under mild conditions. The algorithm is most effective in implementations utilizing the special collapsibility property of GLMMs. Useful blocking strategies can be applied to further speed up algorithmic convergence and reduce computational costs.

The blocked Gauss-Seidel algorithm is applied to study the association between incidence of ischemic heart disease and socioeconomic status in NSW, Australia, using a Poisson regression model with spatially varying random effects. The outcome data consist of approximately 33 million records and cannot be fit using standard implementations of PQL. The Gauss-Seidel procedure was applied to quickly fit these data using R and perform a full epidemiological analysis that simultaneously adjusted for age, gender, day of week and spatial variability.

Simulation studies demonstrate the computational advantages of the procedure in fitting Poisson and logistic regression models to large data sets similar in structure and size as that of the NSW study. The IWLS algorithm is unable to analyze data sets exceeding 10% of its size. The Gauss-Seidel algorithm is found to substantially outperform existing techniques for PQL with respect to criteria such as maximum allowable number of observations, time per iteration, and total time until convergence. The algorithm is currently being implemented on a wider scale and will be made publicly available as an R library.

6 Supplemental Materials

All of the following supplemental files are contained in a single online archive and can be obtained via a single download:

**Appendix:** Platform-independent pseudo-code for fitting GLMs. (“Glm_appendix.pdf”)

**C++ Simulation Source Code:** The source code used to generate and fit test data-sets us-
ing Poisson and Binomial generalized linear models and both Gauss-Seidel and Raphson-Newton methods. (zip archive file “Glm.zip”)

REFERENCES


Figure 1: Total convergence time in seconds for the algorithms Gauss-Seidel (solid line) and IWLS (dashed line) for the Poisson regression model (left panel) and the logistic regression model (right panel) of Section 3. See the text for further explanation.

Figure 2: 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.
<table>
<thead>
<tr>
<th>Parameter</th>
<th>GLMM (16)</th>
<th></th>
<th></th>
<th>GLM (17)</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Estimate</td>
<td>SE</td>
<td>P-Value</td>
<td>Estimate</td>
<td>SE</td>
<td>P-Value</td>
</tr>
<tr>
<td>Tuesday</td>
<td>-0.061</td>
<td>0.0059</td>
<td>&lt; 0.001</td>
<td>-0.061</td>
<td>0.0059</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Wednesday</td>
<td>-0.072</td>
<td>0.006</td>
<td>&lt; 0.001</td>
<td>-0.072</td>
<td>0.0059</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Thursday</td>
<td>-0.064</td>
<td>0.0059</td>
<td>&lt; 0.001</td>
<td>-0.064</td>
<td>0.0059</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Friday</td>
<td>-0.059</td>
<td>0.0059</td>
<td>&lt; 0.001</td>
<td>-0.059</td>
<td>0.0059</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Saturday</td>
<td>-0.144</td>
<td>0.0062</td>
<td>&lt; 0.001</td>
<td>-0.144</td>
<td>0.0061</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Sunday</td>
<td>-0.149</td>
<td>0.0062</td>
<td>&lt; 0.001</td>
<td>-0.149</td>
<td>0.0062</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Intercept</td>
<td>-18.11</td>
<td>0.0023</td>
<td>&lt; 0.001</td>
<td>-18.07</td>
<td>0.0022</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>density/10^5</td>
<td>0.546</td>
<td>0.0934</td>
<td>&lt; 0.001</td>
<td>1.28</td>
<td>0.0908</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>σ^{-2}</td>
<td>2.739</td>
<td>0.1803</td>
<td>&lt; 0.001</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>λ</td>
<td>0.85</td>
<td>0.0559</td>
<td>&lt; 0.001</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>SEIFA interactions/10^4</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Men &lt; 20 yrs</td>
<td>-0.499</td>
<td>2.845</td>
<td>0.861</td>
<td>-1.027</td>
<td>2.783</td>
<td>0.712</td>
</tr>
<tr>
<td>Men 20-24 yrs</td>
<td>1.618</td>
<td>1.883</td>
<td>0.39</td>
<td>0.929</td>
<td>1.842</td>
<td>0.614</td>
</tr>
<tr>
<td>Men 25-29 yrs</td>
<td>-5.077</td>
<td>0.8471</td>
<td>&lt; 0.001</td>
<td>-5.435</td>
<td>0.8153</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Men 30-34 yrs</td>
<td>-4.029</td>
<td>0.522</td>
<td>&lt; 0.001</td>
<td>-4.407</td>
<td>0.5036</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Men 35-39 yrs</td>
<td>-3.534</td>
<td>0.3164</td>
<td>&lt; 0.001</td>
<td>-3.884</td>
<td>0.3062</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Men 40-44 yrs</td>
<td>-2.983</td>
<td>0.2199</td>
<td>&lt; 0.001</td>
<td>-3.363</td>
<td>0.2135</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Men 45-49 yrs</td>
<td>-2.838</td>
<td>0.1678</td>
<td>&lt; 0.001</td>
<td>-3.283</td>
<td>0.1635</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Men 50-54 yrs</td>
<td>-2.78</td>
<td>0.1357</td>
<td>&lt; 0.001</td>
<td>-3.279</td>
<td>0.1326</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Men 55-59 yrs</td>
<td>-2.143</td>
<td>0.1259</td>
<td>&lt; 0.001</td>
<td>-2.627</td>
<td>0.1234</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Men 60-64 yrs</td>
<td>-1.986</td>
<td>0.1188</td>
<td>&lt; 0.001</td>
<td>-2.43</td>
<td>0.1165</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Men 65-69 yrs</td>
<td>-1.143</td>
<td>0.1104</td>
<td>&lt; 0.001</td>
<td>-1.533</td>
<td>0.1081</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Men 70-74 yrs</td>
<td>-0.719</td>
<td>0.1032</td>
<td>&lt; 0.001</td>
<td>-1.112</td>
<td>0.101</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Men 75-79 yrs</td>
<td>-0.29</td>
<td>0.1085</td>
<td>0.008</td>
<td>-0.714</td>
<td>0.1062</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Men 80-84 yrs</td>
<td>-0.097</td>
<td>0.1261</td>
<td>0.442</td>
<td>-0.512</td>
<td>0.1235</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Men ≥ 85 yrs</td>
<td>1.396</td>
<td>0.1468</td>
<td>&lt; 0.001</td>
<td>0.905</td>
<td>0.1439</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Women &lt; 20 yrs</td>
<td>-1.898</td>
<td>3.161</td>
<td>0.548</td>
<td>-2.388</td>
<td>3.086</td>
<td>0.439</td>
</tr>
<tr>
<td>Women 20-24 yrs</td>
<td>-7.912</td>
<td>1.864</td>
<td>&lt; 0.001</td>
<td>-8.02</td>
<td>1.79</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Women 25-29 yrs</td>
<td>-5.727</td>
<td>1.603</td>
<td>&lt; 0.001</td>
<td>-6.015</td>
<td>1.537</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Women 30-34 yrs</td>
<td>-5.486</td>
<td>0.9349</td>
<td>&lt; 0.001</td>
<td>-5.795</td>
<td>0.8994</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Women 35-39 yrs</td>
<td>-5.965</td>
<td>0.5565</td>
<td>&lt; 0.001</td>
<td>-6.211</td>
<td>0.5374</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Women 40-44 yrs</td>
<td>-4.583</td>
<td>0.3751</td>
<td>&lt; 0.001</td>
<td>-4.919</td>
<td>0.3628</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Women 45-49 yrs</td>
<td>-4.764</td>
<td>0.2684</td>
<td>&lt; 0.001</td>
<td>-5.147</td>
<td>0.2598</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Women 50-54 yrs</td>
<td>-4.374</td>
<td>0.2144</td>
<td>&lt; 0.001</td>
<td>-4.866</td>
<td>0.2089</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Women 55-59 yrs</td>
<td>-3.569</td>
<td>0.1935</td>
<td>&lt; 0.001</td>
<td>-4.049</td>
<td>0.1892</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Women 60-64 yrs</td>
<td>-2.781</td>
<td>0.1679</td>
<td>&lt; 0.001</td>
<td>-3.195</td>
<td>0.1639</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Women 65-69 yrs</td>
<td>-2.118</td>
<td>0.1446</td>
<td>&lt; 0.001</td>
<td>-2.514</td>
<td>0.1411</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Women 70-74 yrs</td>
<td>-1.356</td>
<td>0.1235</td>
<td>&lt; 0.001</td>
<td>-1.783</td>
<td>0.1203</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Women 75-79 yrs</td>
<td>-0.516</td>
<td>0.1127</td>
<td>&lt; 0.001</td>
<td>-0.989</td>
<td>0.11</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Women 80-84 yrs</td>
<td>0.222</td>
<td>0.1119</td>
<td>0.047</td>
<td>-0.268</td>
<td>0.1095</td>
<td>0.014</td>
</tr>
<tr>
<td>Women ≥ 85 yrs</td>
<td>1.466</td>
<td>0.1053</td>
<td>&lt; 0.001</td>
<td>0.958</td>
<td>0.1031</td>
<td>&lt; 0.001</td>
</tr>
</tbody>
</table>

Table 1: Estimates of selected parameters of models (16) and (17).
Figure 3: Estimated day-of-week effects. Monday is the reference group. The lines represent intervals of two standard errors.

Figure 4: Estimated SEIFA interactions for the 15 age groups of the male subpopulation. The lines represent margins of two standard errors.