On the Use of Penalized Quasi-Likelihood Information Criterion for Generalized Linear Mixed Models

Francis K.C. Hui*[1]

[1]Research School of Finance, Actuarial Studies and Statistics, The Australian National University, Canberra, Australia

Abstract

Information criteria are a common approach for joint fixed and random effects selection in mixed models. While straightforward to implement, a major difficultly when applying information criteria is that they are typically based on maximum likelihood estimates, yet calculating such estimates for one, let alone multiple, candidate mixed models presents a major computational hurdle. To overcome this, we study penalized quasi-likelihood estimation and use it as the basis for performing fast joint selection. Under a general framework, we show that penalized quasi-likelihood estimation produces consistent estimates of the true parameters. Then, we propose a new penalized quasi-likelihood information criterion whose distinguishing feature is the way it accounts for model complexity in the random effects, since penalized quasi-likelihood estimation effectively treats the random effects as fixed. We demonstrate that the criterion asymptotically identifies the true set

*Corresponding author: francis.hui@anu.edu.au

of important fixed and random effects. Simulations show the quasi-likelihood information criterion performs competitively with and sometimes better than common maximum likelihood information criteria for joint selection, while offering substantial reductions in computation time.

Keywords: fixed effects; model selection; penalized quasi-likelihood; random effects; variable selection

1 Introduction

Generalized linear mixed models are ubiquitous in many fields of applied statistics. This article studies the problem of joint fixed and random effects selection in mixed models, which often plays an important role in inference given the need to identify covariates driving the overall mean response, and possible sources of heterogeneity around the overall mean. One common approach to joint selection is information criteria, given the relative ease with which they can be implemented, with popular choices being the Akaike information criterion (AIC, Akaike, 1974) and Bayesian information criterion (BIC, Schwarz, 1978). In the past decade, there has been growing research into modifying such criteria specifically for mixed models, including the development of conditional AIC for performing cluster-focused inference and efficient prediction (e.g., Vaida and Blanchard, 2005; Greven and Kneib, 2010; Donohue et al., 2011; Yu et al., 2013), and some investigation into consistent model selection using BIC-type criteria (e.g., Jones, 2011; Delattre et al., 2014; Yu et al., 2018).

While straightforward to implement, a major challenge when applying information criteria is their dependence on maximum likelihood estimates. Since the marginal likelihood does not possess a closed form for generalized linear mixed models, except for the case of normal responses and the identity link, calculating the maximum likelihood estimates for one, let alone multiple, candidate models presents a major computational challenge. For joint selection, we also need to consider different combinations of fixed and random effects. This makes the

number of candidate models much larger than the standard generalized linear model setting, although in some settings we may not necessarily want to consider all possible combinations (e.g., Cheng et al., 2010; Hui et al., 2017a).

The difficulties in maximum likelihood estimation for mixed models has spurred research into various approaches to overcome this hurdle, including the Expectation-Maximization algorithm or variants thereof (McCulloch, 1997; Booth and Hobert, 1999), quadrature methods and the Laplace approximation (Vonesh, 1996), and more recent approaches such as variational approximations and the Expectation-Propagation algorithm (Ormerod and Wand, 2012; Kim and Wand, 2018). Of these estimation approaches, by far the quickest and most straightforward is penalized quasi-likelihood estimation (Breslow and Clayton, 1993), which can be derived as an approximation to the Laplace approximation by omitting the nonlinear term involving a log determinant. The speed of penalized quasi-likelihood estimation lies in effectively treating the random effects as fixed, such that the resulting objective function resembles the log-likelihood function for a generalized linear model augmented with a ridge-like penalty. In turn, this can be maximized efficiently using techniques such as a modified iterative reweighted least squares.

In this article, we study the large sample properties of penalized quasi-likelihood estimation and use it as the basis for constructing a quasi-likelihood based information criterion for fast joint selection in generalized linear mixed models. Under a general framework encompassing independent cluster and multilevel mixed models, we show that penalized quasi-likelihood estimation produces consistent estimates of the fixed and random effects coefficients. The result expands on the previous works of Vonesh et al. (2002) and Hui et al. (2017b), who established consistency of penalized quasi-likelihood estimators for the special case of single level mixed models; see also the maximum posterior estimation method of Jiang et al. (2001). It is important to emphasize that, even though we develop results assuming a fixed number of fixed and random effect covariates, the total number of parameters when applying penalized quasi-likelihood estimation still grows with sample size due to the increasing number of random

effect coefficients, since the number of clusters in each set of random effects grows. To establish consistency then, we need to allow the cluster size in each set of random effects to also grow at a suitable rate. The need for growing cluster sizes is a requirement of all approximate likelihood estimation methods for generalized linear mixed models (Vonesh, 1996; Nie, 2007).

Having established estimation consistency, we then propose a new penalized quasi-likelihood information criterion (QIC) for fast joint selection. A distinguishing feature of QIC is the way it differentially accounts for model complexity in the fixed and random effects: because penalized quasi-likelihood estimation treats the random effects as fixed, then QIC accounts for model complexity using the number of random effects coefficients rather than the number of estimated elements in the corresponding random effects covariance matrix. This leads to an inherent penalization for model complexity, since the number of random effects coefficients automatically grows with sample size even if the number of random effects covariates remains fixed, as mentioned previously. Under suitable regularity conditions, and given an appropriate choice of the model complexity penalty, we show that QIC is selection consistent i.e., it asymptotically identifies the true set of important fixed and random effects. While the asymptotic properties of information criteria have been explored in several contexts, e.g., Shao (1997) for linear models and Mueller and Welsh (2009) for generalized linear models, this article is one of the first to establish a selection consistency type result for generalized linear mixed models, although see the recent work of Yu et al. (2018). Simulations demonstrate that the proposed QIC performs competitively with some commonly used information criteria for joint selection in mixed models, while offering substantial reductions in computation time.

2 Generalized Linear Mixed Models

We consider a framework for generalized linear mixed models encompassing common cases of nested e.g. longitudinal, and multilevel mixed models. Consider a set of N observations

consisting of a vector of responses $y = (y_1, \ldots, y_N)^\top$, an $N \times p_F$ model matrix X associated with fixed effect covariates, and an $N \times p_R$ model matrix Z associated with random effect covariates. We assume the first column of X is a vector of ones corresponding to a fixed intercept, which is not subject to selection. Suppose Z can be partitioned into J sets of independent random effects, where J is fixed as $N \to \infty$, such that we can write $Z = (Z_1 \ldots Z_J)$ where Z_j is the model matrix associated with the jth set. Furthermore, for set $j = 1, \ldots, J$ suppose the N observations can be regarded as comprising of n_j clusters each with size m_j, such that $N = n_j m_j$. The assumption of equal cluster sizes m_j is done purely to ease notation, and the developments below can be extended to handle n_j clusters of differing sizes. We can then write $Z_j = (Z_{j1} \ldots Z_{jn_j})$, where Z_{jk} is an $N \times p_{Rj}$ matrix with each column comprising only m_j non-zero entries, and may be interpreted as the model submatrix corresponding to cluster k in the jth set of random effects. The dimensions of Z_j are given by $\dim(Z_j) = N \times n_j p_{Rj}$, with $p_R = \sum_{j=1}^{J} n_j p_{Rj}$.

It is important to emphasize that our use of the word cluster in the above set-up is rather loose, with no requirement that the clusters are independent by design. Also, while we assume that both p_F and p_{Rj} for $j = 1, \ldots, J$ are fixed as $N \to \infty$, it is clear that p_R diverges with N due to its dependence on the n_j's. Put another way, p_{Rj} can be interpreted as the number of random effects for each cluster in set j, and although this is fixed the total number of random effects in set j i.e., the number of columns in Z_j, grows at rate $O(n_j)$.

To complete the formulation, conditional on the random effects coefficients, the responses y_i for $i = 1, \ldots, N$ are assumed to be independent observations from the exponential family of distributions, $f(y_i \mid \mu_i, \phi) = \exp[\phi^{-1}\{y_i\vartheta_i - a(\vartheta_i)\} + c(y_i, \phi)]$ where ϑ_i is the canonical parameter, $a(\cdot)$ and $c(\cdot)$ are known functions, and $\phi > 0$ is the dispersion parameter. Let $\mu = (\mu_1, \ldots, \mu_N)^\top$ denote the vector of conditional means. Then the conditional mean is modeled as $g(\mu) = \eta = X\beta + \sum_{j=1}^{J} Z_j b_j$ for a known link function $g(\cdot)$ applied element-wise to μ, where $\eta = (\eta_1, \ldots, \eta_N)^\top$ is the vector of linear predictors, β is the vector of fixed effect

coefficients, and $b_j = (b_{j1}^\top, \ldots, b_{jn_j}^\top)^\top$ is the $n_j p_{Rj}$-vector of random effect coefficients for random effect set j. Finally, for set $j = 1, \ldots, J$ and cluster $k = 1, \ldots, n_j$, the random effects b_{jk} are assumed to be drawn from a multivariate normal distribution with zero mean vector and covariance matrix Σ_j. That is, $b_{jk} \sim \mathcal{N}(0, \Sigma_j)$.

To offer some insight into the notation established above, we consider the following two examples.

Example 1. Independent cluster mixed models: *We have $J = 1$ with the n_1 clusters representing individuals, say, and m_1 representing number of measurements recorded for each individual. The matrix $Z_1 = (Z_{11} \ldots Z_{1n_1})$ can be constructed to have a block diagonal structure, such that in Z_{1k}, only rows $km_1 - m_1 + 1$ to km_1 have non-zero elements. As a concrete example, suppose we only have a random intercept for each individual. Then $Z_1 = I_{n_1} \otimes 1_{m_1}$ where I_{n_1} is an identity matrix of dimension n_1 and 1_{m_1} denotes an m_1 vector of ones. Furthermore, $b_{1k} \sim \mathcal{N}(0, \sigma_1^2)$ and $Z_1 b_1 = Diag(b_1) \otimes 1_{m_1}$ where $Diag(b_1)$ is a $n_1 \times n_1$ diagonal matrix formed from the vector b_1. One application of independent cluster mixed models is to longitudinal data, and the above example represents a longitudinal mixed model with equal number of measurements for each individual.*

Example 2. Nested Effects: *Nested effects mixed models extend the independent cluster case to incorporate several levels of nesting, leading to $J > 1$ sets of random effects e.g., extractions are nested within plots, which are nested within sites (Schielzeth and Nakagawa, 2013). Each lower level cluster occurs only in one higher level cluster, meaning that for all $j = 1, \ldots, J \geq 2$, the matrices Z_j can be simultaneously constructed to possess a block diagonal structure; see the Supplementary Material for a concrete example. The asymptotic properties of maximum likelihood estimation for independent cluster and nested mixed models have been studied in a number of papers (e.g., Richardson and Welsh, 1994).*

3 Penalized Quasi-Likelihood Estimation

3.1 Set-up

Let $\Lambda = \{\text{vech}(\Sigma_1)^\top, \ldots, \text{vech}(\Sigma_J)^\top\}^\top$, and $\Psi = (\beta^\top, \Lambda^\top, \phi)^\top \in \mathbb{R}^{p_F+1+2^{-1}\sum_{j=1}^J p_{Rj}(p_{Rj}+1)}$ denote all the parameters for the generalized linear mixed model outlined in Section 2. The marginal log-likelihood is then defined as $\ell(\Psi) = \log\left\{\prod_{i=1}^N \int f(y_i \mid \mu_i, \phi) \prod_{j=1}^J \prod_{k=1}^{n_j} f(b_{jk} \mid \Sigma_j) db_{jk}\right\}$, where $f(b_{jk} \mid \Sigma_j) = \mathcal{N}(0, \Sigma_j)$. The maximum likelihood estimates, denoted here as $\tilde{\Psi}$, are obtained by maximizing $\ell(\Psi)$. As reviewed in Section 1, maximum likelihood estimation presents a major computational challenge because the marginal likelihood does not possess a closed form in general. Therefore, joint selection of fixed and random effects can become impractical even when p_F and the p_{Rj}'s are not specially large.

As an alternative, we consider penalized quasi-likelihood estimation. This involves iterating between the following two steps until convergence. First, given Λ, we update $(\beta^\top, b_1^\top, \ldots, b_J^\top)^\top$ and ϕ, if unknown, by maximizing the quasi-likelihood function,

$$\ell_{\text{Q}}(\beta, b, \phi \mid \Lambda) = \sum_{i=1}^N \log\{f(y_i \mid \mu_i, \phi)\} - \frac{1}{2}\sum_{j=1}^J \sum_{k=1}^{n_j} b_{jk}^\top \Sigma_j^{-1} b_{jk}. \tag{1}$$

If necessary, Σ_j^{-1} can be replaced by a generalized inverse if the random effects become completely linearly dependent or the variance component is estimated to zero (Breslow and Clayton, 1993). Compared to $\ell(\Psi)$, maximizing equation (1) is computationally straightforward: the random coefficients are effectively treated as fixed effects, and so $\ell_{\text{Q}}(\beta, b, \phi \mid \Lambda)$ resembles the log-likelihood for a generalized linear model augmented with a ridge-like penalty on the b_{jk}'s. Estimation can utilize efficient algorithms such as penalized iterative reweighted least squares, for instance. Second, given values $(\beta^\top, b_1^\top, \ldots, b_J^\top, \phi)^\top$, we update the random effects covariance matrices Σ_j for $j = 1, \ldots, J$. Here, we adopt the approach of Hui et al. (2017b) and use the iterative update $\hat{\Sigma}_j \Leftarrow n_j^{-1}\sum_{k=1}^{n_j}[\{Z_{jk}^\top \hat{W}_{jk} Z_{jk} + (\hat{\Sigma}_j)^{-1}\}^{-1} + \hat{b}_{jk}\hat{b}_{jk}^\top]$ where $\hat{W}_{jk}$ is an $N \times N$ diagonal matrix with diagonal elements given by $\hat{\phi}a''(\hat{\vartheta}_i)$. While other approaches are

possible for updating the Σ_j's (see for instance, Breslow and Clayton, 1993; Jiang et al., 2001; Pan and Huang, 2014), the method above is straightforward and poses little computational hurdle in the overall estimation procedure. For independent cluster mixed models, Hui et al. (2017b) showed that the iterative update above arises from solving the Laplace approximated marginal log-likelihood with respect to the random effects covariance matrix; we also refer the reader to Vonesh et al. (2002) and Nie (2007) for broader connections between penalized quasi-likelihood estimation and the Laplace approximation approach.

3.2 Estimation Consistency

Let $\hat{\Psi} = (\hat{\beta}^\top, \hat{\Lambda}^\top, \hat{\phi})^\top$ and $\hat{b} = (\hat{b}_1^\top, \ldots, \hat{b}_J^\top)^\top$ denote the penalized quasi-likelihood estimates. In this section, we establish the consistency of these estimates. Let $\Psi_0 = (\beta_0^\top, \Lambda_0^\top, \phi_0)^\top$ denote the vector of true parameters. In the developments below, we will assume that ϕ_0 is known e.g., $\phi_0 = 1$ for binomial and Poisson distributions. All the developments can be extended to the case where ϕ is unknown and requires estimation. For $j = 1, \ldots, J$, define $b_{0j} = (b_{0j1}^\top, \ldots, b_{0jn_j}^\top)^\top$ as realizations from the true random effects distributions, $b_{0jk} \sim \mathcal{N}(0, \Sigma_{0j})$ for $k = 1, \ldots, n_j$. Note some of rows and columns in the Σ_{0j}'s may be entirely zero corresponding to truly unimportant random effects. In such case, it is implicitly assumed that $\mathcal{N}(0, \Sigma_{0j})$ is a degenerate normal distribution with the relevant elements of b_{0jk} set to zero.

Let $\|\cdot\|$ and $\|\cdot\|_\infty$ denote the L_2 and infinity norms, respectively. Also, for a matrix M denote its minimum and maximum eigenvalue by $\lambda_{\min}(M)$ and $\lambda_{\max}(M)$ respectively. We require the following regularity conditions.

(C1) The quantities p_F and p_{Rj} for all $j = 1, \ldots, J$ are fixed as $N \to \infty$, while $n_j \to \infty$ for all $j = 1, \ldots, J$ such that $p_R = \sum_{j=1}^J n_j p_{Rj} = o(N)$.

(C2) The vector $(\beta_0^\top, \phi_0)^\top$ is an interior point in the compact set $\Omega \in \mathbb{R}^{p_F+1}$. Furthermore, for all $j = 1, \ldots, J$, the matrices Σ_{0j} satisfy $0 < \lambda_{\max}(\Sigma_{0j}) < C_0 < \infty$ for a sufficiently

large constant C_0.

(C3) For every N, there exists a sufficiently large constant C_1 such that $\|X, Z\|_\infty < C_1 < \infty$. Furthermore, there exists a sufficiently small constant c_0 such that $0 < c_0 < \lambda_{\min}(N^{-1}X^\top X) < \lambda_{\max}(N^{-1}X^\top X) < c_0^{-1} < \infty$ and $0 < c_0 < \lambda_{\min}(m_j^{-1}Z_j^\top Z_j) < \lambda_{\max}(m_j^{-1}Z_j^\top Z_j) < c_0^{-1} < \infty$ for all $j = 1, \ldots, J$.

(C4) The function $a(\vartheta)$ is at least three times continuously differentiable in its domain. Furthermore, at the vector $(\beta_0^\top, b_{01}^\top, \ldots, b_{0J}^\top)^\top$ the functions $a''(\vartheta)$ and $\{a''(\vartheta)\}^{-1}$ are bounded in probability for all $i = 1, \ldots, N$.

(C5) There exists an open subset containing $(\beta_0^\top, b_{01}^\top, \ldots, b_{0J}^\top)^\top$ such that for all $i = 1, \ldots, N$ and every point in the subset, the function $a'''(\vartheta)$ is bounded in probability.

Condition (C1) formalizes the notion that while the number of fixed and random effect covariates is fixed, the total number of random effect coefficients i.e., the number of columns in Z, is permitted to grow at a slower rate than the total sample size. That is, consistent estimates can be achieved provided both $n_j \to \infty$ and $m_j \to \infty$ for all $j = 1, \ldots, J$, and we allow different sets of random effects to grow at different rates. This requirement for the cluster sizes to also grow differentiates penalized quasi-likelihood estimation from maximum likelihood estimation, which can achieve consistency even when $\max_{j=1,\ldots,J}(m_j)$ is bounded. There is no lower bound placed on the rate of growth of the cluster sizes m_j, although in practice finite sample performance of penalized quasi-likelihood estimation would be expected to deteriorate with smaller cluster sizes. It should be acknowledged that other approximation methods including Laplace's method and adaptive quadrature also require cluster size to grow in order to achieve consistency, although the rates of convergence differ depending on the exact nature of approximation (Liu and Pierce, 1994; Nie, 2007). Conditions (C2)-(C5) are used to guarantee smoothness of equation (1) and its regular behavior in the neighborhood of the true parameter value and true realizations of the random effects.

Theorem 1. *Assume Conditions (C1)-(C5) are satisfied. Given a set of matrices* Σ_j *for* $j = 1, \ldots, J$ *satisfying* $0 < \lambda_{\max}(\Sigma_j^{-1}) < C_2 < \infty$ *for some sufficiently large constant* C_2*, then the estimates satisfy* $\|\hat{\beta} - \beta_0\| = o_p(1)$*, and* $\|\hat{b}_{jk} - b_{0jk}\| = o_p(1)$ *for all* $j = 1, \ldots, J$ *and* $k = 1, \ldots, n_j$.

All proofs are provided in the Supplementary Material. The above result implies consistency of the penalized quasi-likelihood estimates for the fixed and random effect coefficients under a broad setting with multiple sets of random effects. It sets the foundation for constructing a penalized quasi-likelihood information criterion in the next section. Theorem 1 generalizes the result of Hui et al. (2017b), who demonstrated it for penalized quasi-likelihood estimates for independent cluster mixed models; see Example 1 above. The result is also consistent with Jiang et al. (2001) and Fan and Li (2012), whom demonstrated that under the right large sample scenarios estimation consistency of the fixed and random effect coefficients is unaffected by the choice of the Σ_j's. Moreover, Theorem 1 implies that the iterative update estimate of random effects covariance matrix is also consistent as follows. Let $\|\cdot\|_F$ denote the Frobenius norm.

Corollary 1. *For all* $j = 1, \ldots, J$*, the iterative update for* $\hat{\Sigma}_j$ *satisfies* $\|\hat{\Sigma}_j - \Sigma_{0j}\|_F = o_p(1)$.

4 A Penalized Quasi-Likelihood Information Criterion

Let $\ell_1(\beta, b_1, \ldots, b_J) = \sum_{i=1}^{N} \log\{f(y_i \mid \mu_i, \phi_0)\}$ denote the first term in equation (1) with ϕ_0 assumed to be known and noting selection is not performed on ϕ. We propose the following penalized quasi-likelihood information criterion for joint selection in generalized linear mixed models.

$$\text{QIC} = -2\ell_1(\hat{\beta}, \hat{b}_1, \ldots, \hat{b}_J) + \log(N)\dim(\hat{\beta}) + \sum_{j=1}^{J} \gamma_j \dim(\hat{b}_j), \tag{2}$$

where $\gamma_j > 0$ denotes a model complexity penalty for the j-th set of random effects. The second term can also be written as $\sum_{j=1}^{J} \gamma_j n_j \dim(\hat{b}_{jk})$, and note all the γ_j's are positive quantities. Because QIC only requires computation of penalized quasi-likelihood estimates based on (1), the criterion facilitates computationally efficient joint selection, in contrast to information criteria that require calculation of the maximum likelihood estimates.

An important feature of QIC is how it handles the model complexity penalties for the fixed and random effect coefficients. For the former, we use a penalty depending on the total sample size, $\log(N)$. The is the equivalent to the version of BIC implemented in the `R` package `lme4` (Bates et al., 2014), and is also the model complexity penalty advocated by Delattre et al. (2014) and used recently in Yu et al. (2018). For $J = 1$ however, it does differ from the $\log(n_1)$ penalty used in the `R` package `nlme` (Pinheiro et al., 2018) and in the SAS GLIMMIX module (SAS, 2011), for instance. In the context of penalized quasi-likelihood estimation, $\log(N)$ is a sensible choice given the estimation of β based on (1) utilizes the total sample size. Turning to the random effects, notice that even though p_{Rj} is fixed, we have $\dim(\hat{b}_j) = n_j \dim(\hat{b}_{jk}) = O(n_j)$. This means there is an inherent model complexity penalty built into this component of QIC due to the dependence on the number of clusters, and so logically the corresponding model complexity penalty γ_j does need not be specially large to avoid overfitting. In fact, the theoretical developments below show that strong performance can be achieved even if all the γ_j are chosen to be positive constants. This contrasts to most other settings where a growing model complexity is mandatory e.g., such as in BIC, to achieve selection consistency. More generally, because penalized quasi-likelihood estimation can be interpreted as treating the random effect coefficients as fixed, then constructing an information criterion based on the $\hat{b}_j$'s rather than the covariance matrices $\hat{\Sigma}_j$'s, which are considered nuisance parameters in equation (1), is reasonable. Indeed, using $\sum_{j=1}^{J} \gamma_j \dim(\hat{b}_j)$ differs from many other information criteria that count the number of elements in $\hat{\Lambda}$ e.g., the BIC used in Bates et al. (2014), SAS (2011), and Delattre et al. (2014). Finally, we allow for different model complexity penalties for each set

of random effects: this is useful to accommodate potentially differing rates of growth of the number of clusters and cluster sizes across the J sets of random effects, in line with Condition (C1). It includes the special case of setting all the penalties to be the same i.e., $\gamma_1 = \ldots = \gamma_J = \gamma > 0$.

5 Selection Consistency

Let $\mathcal{M}$ generically denote a model formed by taking a subset of the p_F candidate fixed effect covariates, plus subsets for each of the p_{Rj} candidate random effect covariates for every $j = 1, \ldots, J$. Then we can construct the corresponding submatrices $X(\mathcal{M})$ and, for $j = 1, \ldots, J$, $Z_j(\mathcal{M}) = \{Z_{j1}(\mathcal{M}), \ldots, Z_{jn_j}(\mathcal{M})\}$ where $Z_{jk}(\mathcal{M})$ is formed from the columns of Z_{jk}. Since the fixed intercept is assumed to be always included, then $X(\mathcal{M})$ always contains the first column of X. Also, if an entire set of random effects is removed from a candidate model, then implicitly we only form $Z_j(\mathcal{M})$'s for those sets of random effects which are included.

Let $\hat{\Psi}(\mathcal{M}) = \{\hat{\beta}(\mathcal{M})^\top, \hat{\Lambda}(\mathcal{M})^\top\}^\top$ and $\hat{b}(\mathcal{M}) = \{\hat{b}_1(\mathcal{M})^\top, \ldots, \hat{b}_J(\mathcal{M})^\top\}^\top$ denote the penalized quasi-likelihood estimates for model $\mathcal{M}$, and write $\text{QIC}(\mathcal{M}) = -2\ell_1\{\hat{\beta}(\mathcal{M}), \hat{b}_1(\mathcal{M}), \ldots, \hat{b}_J(\mathcal{M})\} + \log(N)\dim\{\hat{\beta}(\mathcal{M})\} + \sum_{j=1}^J \gamma_j \dim\{\hat{b}_j(\mathcal{M})\}$. Furthermore, let $\mathcal{M}_0$ denote the model containing only the truly important fixed effect (non-zero elements of β_0) and random effect (non-zero diagonal elements of Σ_{0j} for all $j = 1, \ldots, J$) covariates. To study the asymptotic behavior of QIC for joint selection, we require generalizations of Conditions (C2)–(C5) to ensure regular behavior at the candidate models. Given their similarity to the original conditions in Section 3.2, then we present these in the Supplementary Material. However, one aspect of these conditions that is noteworthy is the notion of pseudo-true parameters $\Psi_0(\mathcal{M}) = \{\beta_0(\mathcal{M})^\top, \Lambda_0(\mathcal{M})^\top\}^\top$, and realizations of random effects from the pseudo-true random effects distributions, $b_{0jk}(\mathcal{M}) \sim \mathcal{N}\{0, \Sigma_{0j}(\mathcal{M})\}$ for $j = 1, \ldots, J$ and $k = 1, \ldots, n_j$. The pseudo-true parameters can be interpreted as the minimizer of a Kullback-Leibler type divergence between

the true model $\mathcal{M}_0$ and a candidate model $\mathcal{M}$, and are defined as the limiting values to which the penalized quasi-likelihood estimates converge to. That is, $\|\hat{\beta}(\mathcal{M}) - \beta_0(\mathcal{M})\| = o_p(1)$, $\|\hat{b}_{jk}(\mathcal{M}) - b_{0jk}(\mathcal{M})\| = o_p(1)$ for $j = 1, \ldots, J$ and $k = 1, \ldots, n_j$, and $\|\hat{\Sigma}_j(\mathcal{M}) - \Sigma_{0j}(\mathcal{M})\|_F = o_p(1)$; we refer the reader to White (1982) among others for more on the idea of pseudo-true parameters. The following additional condition is required on the asymptotic behavior of differences in the penalized quasi-likelihood function.

(C6) Let $D_Q(\mathcal{M}||\mathcal{M}_0) = N^{-1}[\ell_Q\{\beta_0(\mathcal{M}), b_0(\mathcal{M}), \phi_0 \mid \Lambda_0(\mathcal{M})\} - \ell_Q\{\beta_0(\mathcal{M}_0), b_0(\mathcal{M}_0), \phi_0 \mid \Lambda_0(\mathcal{M}_0)\}]$. Then for all models $\mathcal{M}_0 \not\subset \mathcal{M}$, it holds that $D_Q(\mathcal{M}||\mathcal{M}_0) \to c_1(\mathcal{M}) > 0$ in probability.

Condition (C6) states that for any underfitted model, the mean difference in penalized quasi-likelihood function between the true and underfitted model is guaranteed to be positive. The condition is similar to those seen in Mueller and Welsh (2009), among others, for generalized linear models, and encapsulates the notion of underfitting for generalized linear mixed models using penalized quasi-likelihood estimation.

Lemma 1. *Assume Conditions (C1), (C2')–(C5'), and (C6) are satisfied.*

- *If $\gamma_j = o(m_j)$ for all $j = 1, \ldots, J$, then*

$$P\left(\min_{\mathcal{M}:\mathcal{M}_0 \not\subset \mathcal{M}} N^{-1}\left\{QIC(\mathcal{M}) - QIC(\mathcal{M}_0)\right\} > 0\right) \to 1.$$

- *Let $n_{\min}$ and $n_{\max}$ denote the clusters with the smallest and largest rate of growth, respectively. If $n_{\min}^{-1}\log(n_{max}) = o(\gamma_j)$ for all $j = 1, \ldots, J$, then*

$$P\left(\min_{\mathcal{M}:\mathcal{M}_0 \subset \mathcal{M}} \left\{QIC(\mathcal{M}) - QIC(\mathcal{M}_0)\right\} > 0\right) \to 1,$$

where $\subset$ denotes the proper subset operator.

The first and second parts of Lemma 1 show that the value of QIC for any underfitted and overfitted model, respectively, is asymptotically guaranteed to be greater than the value of QIC

at $\mathcal{M}_0$. In other words, with probability tending to one we can always pick $\mathcal{M}_0$ to produce a lower value of QIC. Not surprisingly, the two parts require the γ_j's not to grow too quickly, and (generally speaking) not tending too zero too slowly, respectively. This leads to the main result regarding QIC for joint selection in generalized linear mixed models.

Theorem 2. *Let $\hat{\mathcal{M}}$ be the model chosen by minimizing QIC in equation (2). Under the conditions of Lemma 1, then pr$(\hat{\mathcal{M}} = \mathcal{M}_0) \to 1$.*

The proof of the above result follows directly from Lemma 1, and so is omitted. Theorem 2 states that QIC will asymptotically select the true set of fixed and random effects covariates.

6 Simulation Study

6.1 General Design

We conducted a numerical study to evaluate the performance of QIC relative to some commonly used information criteria for variable selection in generalized linear mixed models. Let $\ell(\tilde{\Psi})$ denote the marginal log-likelihood evaluated at the maximum likelihood estimate. Then we consider the following criteria: I) QIC as defined in (2), where we set $\gamma_j = \log\log(m_j)$; II) AIC $= -2\ell(\tilde{\Psi}) + 2\dim(\tilde{\beta}) + 2\dim\{\text{vech}(\tilde{\Sigma})\}$; III) $\text{BIC}_{\text{SAS}} = -2\ell(\tilde{\Psi}) + \max_{j=1,\ldots,J}\{\log(n_j)\}\dim(\tilde{\beta}) + \sum_{j=1}^{J}\log(n_j)\dim\{\text{vech}(\tilde{\Sigma}_j)\}$, which is a generalization of the BIC used in SAS GLIMMIX for $J > 1$ sets of random effects; IV) $\text{BIC}_{\text{DLC}} = -2\ell(\tilde{\Psi}) + \log(N)\dim(\tilde{\beta}) + \sum_{j=1}^{J}\log(n_j)\dim\{\text{vech}(\tilde{\Sigma}_j)\}$, which is a generalization of the BIC proposed by Delattre et al. (2014) for $J > 1$; V) $\text{BIC}_{\text{lme4}} = -2\ell(\tilde{\Psi}) + \log(N)\dim(\tilde{\beta}) + \log(N)\sum_{j=1}^{J}\dim\{\text{vech}(\tilde{\Sigma}_j)\}$, which is used in the `lme4` package.

Method I is the only penalized quasi-likelihood information criterion. Moreover, our choice of $\gamma_j = \log\log(m_j)$ means the model complexity penalty varies for different sets of random effects (depending on their rate of growth), while also satisfying the requirements of Theorem 2. If the cluster sizes are not equal within a particular set of random effects, then we replace it with

the smallest observed cluster size within that set. We acknowledge that it is possible to empirically experiment with different choices of γ_j e.g., the optimal choice of the model complexity penalty will likely depend on structure of the random effects and the relative sizes of n_j versus m_j (although this is an issue relevant to all information criteria), or make it entirely data driven at the expense of substantially increased computation (e.g., Bai et al., 1999; Jiang et al., 2009). In turn, the choice of $\gamma_j = \log\log(m_j)$ should be viewed as more of a general recommendation, and is consistent with the use of default values for tuning constants in other variable selection methods e.g., regulariation penalties (Zhang et al., 2010) and other information criterion (Chen and Chen, 2008; Luo and Chen, 2013).

With regards to implementation, penalized quasi-likelihood estimation of the mixed models was performed via the `rpql` package (Hui, 2018), while maximum likelihood estimation was performed via `lme4`.We also conducted additional simulations using adaptive quadrature instead of the Laplace approximation to construct the maximum likelihood information criteria, but found very similar model selection performance and so have omitted them for brevity. We provide template `R` for performing the simulations in the Supplementary Material.

We considered two settings encompassing independent cluster and nested mixed effects models. We only present results from independent cluster mixed models in the main text, while the results for the nested effects mixed models are provided in the Supplementary Material, and present similar conclusions to those below. We focused on three types of non-normally distributed responses, as computationally there is usually little to gain when using penalized quasi-likelihood estimation for linear mixed models given the tractable form of the marginal likelihood in this case. For each simulation design, we generated 400 datasets, and assessed performance for each criterion based on the percentage of datasets with correctly chosen fixed and/or random effects structures. We also present the mean computation time in seconds used for each procedure.

6.2 Setting 1: Independent Cluster Mixed Models

We simulated data from an independent cluster mixed model as follows. For each of n_1 clusters, we first generated the cluster size, denoted here as m_{1k} for $k = 1, \ldots, n_1$, by sampling integer values in the range from $\lceil 2n_1^{0.3} \rceil$ to $\lceil 3n_1^{0.3} \rceil$ inclusive. This created a unbalanced design, noting we developed our theory using equal cluster sizes for ease of notation, while also ensuring that the cluster size grew with the number of clusters as required by Condition (C1) i.e., one may regard it as having asymptotically balanced cluster sizes. With $N = \sum_{k=1}^{n_1} m_{1k}$, we constructed a $N \times 13$ fixed effects model matrix X by setting first column equal to one, while the remaining eight elements in each row of X were simulated from a multivariate normal distribution $\mathcal{N}(0, D)$, where D is a matrix with elements $D_{rs} = 0.4^{|r-s|}$. The random effects covariates were then taken to be the first five columns of X, such that $p_{R1} = 5$. We set the vector of true fixed effect coefficients as $K^{-1}\beta_0 = (-0.1, 1, -1, 0, 0, 0.5, -0.5, 0, 0, 0, 0, 0, 0)^\top$, while the random effect coefficients were generated as $b_{1k} \sim \mathcal{N}(0, \Sigma_{01})$ where

$$K^{-1}\Sigma_{01} = \begin{pmatrix} 1 & 0.6 & 0.6 & 0 & 0 \\ 0.6 & 1 & 0.4 & 0 & 0 \\ 0.6 & 0.4 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix},$$

for a scalar constant $K > 0$. Thus covariates 1 to 3 were important fixed and random effects, while covariates 6, and 7 were important fixed effects only. Finally, the responses y_i were generated based on the mixed model formulation in Section 2 i.e., since $J = 1$ then $\eta = X\beta_0 + Z_1 b_1$, and we considered three types of non-normally distributed responses: 1) binomial responses with trial size equal to three, a logit link function, and $K = 1$; 2) Poisson counts with a log link and $K = 0.5$; 3) Bernoulli responses with a logit link and $K = 1$. The use of a smaller K in the Poisson count case makes the variable selection process more challenging.

We present results for the first two response types, while the results for Bernoulli responses are provided in the Supplementary Material and present similar conclusions to below. We varied the number of clusters as $n_1 = \{25, 50, 100, 200\}$.

With independent cluster mixed models applied for longitudinal data, it is common for covariates to be included as either a fixed effect only or as both fixed and random effects. To account for this in the variable selection process, we implemented the following selection procedure as recommended by Cheng et al. (2010): 1) starting with a saturated model including all candidate fixed and random effects, we performed backward elimination on the random effects while keeping all fixed effects in the model; 2) with the random effects selected in step 1 also included as fixed effects, we performed backward elimination on the remaining fixed effect covariates. These two steps ensure that no covariate will be selected as a random effect only. We also considered other selection algorithms, notably a two-step exhaustive subsets approach, but found that this procedure produced similar results to the backward selection approach presented here while taking considerably longer; for brevity these results are presented in the Supplementary Material.

Compared to the four maximum likelihood information criteria, QIC with $\gamma_j = \log\log(m_j)$ performed competitively for both Binomial and Poisson responses (Table 1), while offering considerable reductions in computation time (Figure 1). AIC, and to a lesser extent BIC_{SAS}, performed relatively poorly for fixed effects selection as it had a tendency to overfit due to the model complexity penalty not being severe enough. By contrast, the three information criteria which used $\log(N)$ as the model complexity penalty for the fixed effects i.e., QIC, BIC_{DLC}, and BIC_{lme4}, did not suffer from this overfitting issue as much. With regards to random effects selection, QIC was not necessarily the best performer, although the differences between it and the three BIC-type criterion were generally only slight. AIC performed worst at the larger sample sizes, which was expected given the criterion is not designed to be selection consistent. Finally, while QIC did not scale as well relative to their maximum likelihood counterparts (we

Table 1: Percentage of datasets with correctly chosen structure (joint fixed and random effects/fixed effects only/random effects only) for simulation Setting 1. The first method is the quasi-likelihood information criterion, while the remaining four methods are maximum likelihood information criteria.

n_1	QIC	AIC	BIC_{SAS}	BIC_{DLC}	BIC_{lme4}
			Binomial Responses		
25	48.75/56.52/78.00	16.75/18.75/79.00	33.00/45.75/68.75	47.75/65.25/68.75	24.25/59.75/35.75
50	80.50/80.75/97.50	20.50/20.50/92.25	64.25/64.75/98.75	88.25/89.25/98.75	85.50/89.50/95.50
100	92.50/92.50/100.00	24.75/24.75/89.00	77.00/77.00/100.00	94.00/94.00/100.00	94.00/94.00/100.00
200	94.50/92.75/100.00	22.50/22.50/91.75	86.25/86.25/100.00	94.75/94.75/100.00	94.75/94.75/100.00
			Poisson Responses		
25	51.00/51.50/92.75	21.25/21.50/91.25	46.00/46.75/94.75	63.00/64.25/94.75	57.50/60.50/88.50
50	85.75/83.75/97.50	17.50/17.50/92.25	62.25/62.25/99.50	88.50/88.50/99.50	89.00/89.00/100.00
100	93.25/93.25/100.00	22.50/22.50/91.25	76.25/76.25/100.00	92.50/92.50/100.00	92.50/92.50/100.00
200	95.75/95.50/100.00	17.00/17.00/90.00	80.00/80.00/100.00	95.25/95.25/100.00	95.25/95.25/100.00

conjecture this is due to the core algorithm of `lme4` being written in `C++`, while out penalized quasi-likelihood estimation is entirely written in `R`), there were nevertheless substantial reductions in computation time when using QIC for joint selection across all four sample sizes tested.

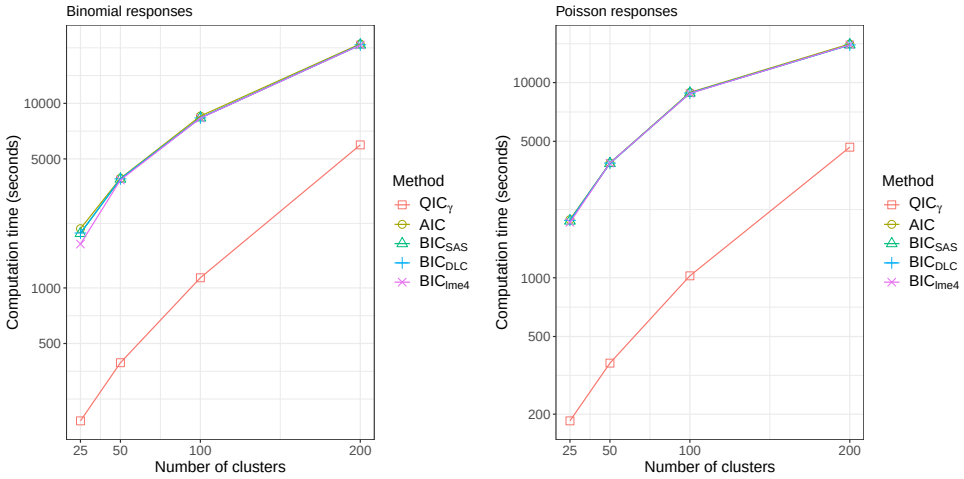
Acknowledgments

This research was supported by two Australian Research Council Discovery awards. Thanks to Samuel Mueller and Alan Welsh for useful discussions, and the Editors and reviewers for useful comments.

Supplementary Material

Supplementary material available online includes all proofs and additional simulation results, and template `R` code.

Figure 1: Mean computation time in seconds for simulation Setting 1. The first method is the penalized quasi-likelihood information criterion, while the remaining four methods are maximum likelihood information criteria.



References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19:716–723.

Bai, Z. D., Rao, C. R., and Wu, Y. (1999). Model selection with data-oriented penalty. *Journal of Statistical Planning and Inference*, 77:103–117.

Bates, D., Maechler, M., Bolker, B., and Walker, S. (2014). *lme4: Linear mixed-effects models using Eigen and S4*. R package version 1.1-7.

Booth, J. G. and Hobert, J. P. (1999). Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, 61:265–285.

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

Chen, J. and Chen, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika*, 95:759–771.

Cheng, J., Edwards, L. J., Maldonado-Molina, M. M., Komro, K. A., and Muller, K. E. (2010). Real longitudinal data analysis for real people: building a good enough mixed model. *Statistics in Medicine*, 29:504–520.

Delattre, M., Lavielle, M., and Poursat, M. (2014). A note on BIC in mixed-effects models. *Electronic Journal of Statistics*, 8:456–475.

Donohue, M., Overholser, R., Xu, R., and Vaida, F. (2011). Conditional Akaike information under generalized linear and proportional hazards mixed models. *Biometrika*, 98:685–700.

Fan, Y. and Li, R. (2012). Variable selection in linear mixed effects models. *Annals of Statistics*, 40:2043–2068.

Greven, S. and Kneib, T. (2010). On the behaviour of marginal and conditional AIC in linear mixed models. *Biometrika*, 97:773–789.

Hui, F. K. C. (2018). *rqpl: Regularized PQL for joint selection in GLMMs*. R package version 0.7.

Hui, F. K. C., Mueller, S., and Welsh, A. H. (2017a). Hierarchical Selection of Fixed and Random Effects in Generalized Linear Mixed Models. *Statistica Sinica*, 27:501–518.

Hui, F. K. C., Mueller, S., and Welsh, A. H. (2017b). Joint selection in mixed models using regularized PQL. *Journal of the American Statistical Association*, 112(519):1323–1333.

Jiang, J., Jia, H., and Chen, H. (2001). Maximum posterior estimation of random effects in generalized linear mixed models. *Statistica Sinica*, 11:97–120.

Jiang, J., Nguyen, T., and Rao, J. S. (2009). A simplified adaptive fence procedure. *Statistics & Probability Letters*, 79:625–629.

Jones, R. H. (2011). Bayesian information criterion for longitudinal and clustered data. *Statistics in medicine*, 30:3050–3056.

Kim, A. S. and Wand, M. P. (2018). On expectation propagation for generalised, linear and mixed models. *Australian & New Zealand Journal of Statistics*, 60:75–102.

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

Luo, S. and Chen, Z. (2013). Extended BIC for linear regression models with diverging number of relevant features and high or ultra-high feature spaces. *Journal of Statistical Planning and Inference*, 143:494–504.

McCulloch, C. E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American statistical Association*, 92:162–170.

Mueller, S. and Welsh, A. (2009). Robust model selection in Generalized Linear Models. *Statistica Sinica*, 19:1155–1170.

Nie, L. (2007). Convergence rate of MLE in generalized linear and nonlinear mixed-effects models: theory and applications. *Journal of Statistical Planning and Inference*, 137:1787–1804.

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

Pan, J. and Huang, C. (2014). Random effects selection in generalized linear mixed models via shrinkage penalty function. *Statistics and Computing*, 24:725–738.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., and R Core Team (2018). *nlme: Linear and Nonlinear Mixed Effects Models*. R package version 3.1-137.

Richardson, A. and Welsh, A. H. (1994). Asymptotic properties of restricted maximum likelihood (REML) estimates for hierarchical mixed linear models. *Australian Journal of statistics*, 36:31–43.

SAS (2011). *SAS/STAT 9.3 User's Guide: The GLIMMIX Procedure*, chapter 40, pages 4337–4435. SAS Institute.

Schielzeth, H. and Nakagawa, S. (2013). Nested by design: model fitting and interpretation in a mixed model era. *Methods in Ecology and Evolution*, 4:14–24.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6:461–464.

Shao, J. (1997). An asymptotic theory for linear model selection. *Statistica Sinica*, 7:221–264.

Vaida, F. and Blanchard, S. (2005). Conditional Akaike information for mixed-effects models. *Biometrika*, 92:351–370.

Vonesh, E. F. (1996). A note on the use of Laplace's approximation for nonlinear mixed-effects models. *Biometrika*, 83:447–452.

Vonesh, E. F., Wang, H., Nie, L., and Majumdar, D. (2002). Conditional second-order generalized estimating equations for generalized linear and nonlinear mixed-effects models. *Journal of the American Statistical Association*, 97:271–283.

White, H. (1982). Maximum Likelihood Estimation of Misspecified Models. *Econometrica*, 50:1–26.

Yu, D., Zhang, X., and Yau, K. K. (2013). Information based model selection criteria for generalized linear mixed models with unknown variance component parameters. *Journal of Multivariate Analysis*, 116:245–262.

Yu, D., Zhang, X., and Yau, K. K. W. (2018). Asymptotic properties and information criteria for misspecified generalized linear mixed models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80:817–836.

Zhang, C.-H. et al. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38:894–942.