Statistica Sinica Preprint No: SS-2023-0343	
Title	Asymptotic Results for Penalized Quasi-Likelihood Estimation in Generalized Linear Mixed Models
Manuscript ID	SS-2023-0343
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202023.0343
Complete List of Authors	Xu Ning, Francis Hui and Alan Welsh
Corresponding Authors	Xu Ning
E-mails	nicksonnz@hotmail.com

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0343

Asymptotic Results for Penalized Quasi-Likelihood Estimation in Generalized Linear Mixed Models

Xu Ning, Francis K.C. Hui, and A.H. Welsh

The Australian National University

Abstract: Generalized Linear Mixed Models (GLMMs) are widely used for analysing clustered data. One well-established method of overcoming the integral in the marginal likelihood function for GLMMs is penalized quasi-likelihood (PQL) estimation, although to date there are few asymptotic distribution results relating to PQL estimation for GLMMs in the literature. In this paper, we establish large sample results for PQL estimators of the parameters and random effects in independent-cluster GLMMs, when both the number of clusters and the cluster sizes go to infinity. This is done under two distinct regimes: conditional on the random effects (essentially treating them as fixed effects) and unconditionally (treating the random effects as random). Under the conditional regime, we show the PQL estimators are asymptotically normal around the true fixed and random effects. Unconditionally, we prove that while the estimator of the fixed effects is asymptotically normally distributed, the correct asymptotic distribution of the so-called prediction gap of the random effects may in fact be a normal scale-mixture distribution under certain relative rates of growth. A simulation study is used to verify the finite sample performance of our theoretical results.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0343

Key words and phrases: Asymptotic independence, Clustered data, Large sample distribution, Longitudinal data, Prediction.

1. Introduction

Generalized linear mixed models (GLMMs) are widely used in statistics to model relationships in clustered and correlated data (McCulloch and Searle, 2004). As the marginal likelihood function of GLMMs, except for normally distributed responses with the identity link, contains an intractable integral, many methods have been developed to estimate and perform inference for the parameters in a computationally efficient manner. These include the Laplace approximation, Gauss-Hermite quadrature, and variational approximations, among others (McCulloch and Searle, 2004; Ormerod and Wand, 2012; Brooks et al., 2017). A connected and well-established approach is penalized Quasi-Likelihood (PQL) estimation (Breslow and Clayton, 1993). As one of the first methods to circumvent the intractable integral, PQL estimation has seen a resurgence in modern statistics as a very computationally efficient method for high-dimensional multivariate GLMMs (e.g., Hui, 2020; Kidziński et al., 2022). However, despite its long history, large sample distributional results for PQL estimation in mixed models are scarce.

The most often studied asymptotic results for maximum likelihood es-

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0343

timators of GLMMs are based on increasing the number of clusters while keeping the size of each cluster fixed or bounded (McCulloch and Searle, 2004; Nie, 2007). Asymptotic results when both the cluster size and number of clusters grow are less developed, although some results for the maximum likelihood estimator as well as the empirical best linear unbiased predictor (EBLUP) for the linear mixed model (LMM) have been developed; see Lyu and Welsh (2021a,b) and references therein. Recently, Jiang et al. (2022) proved an asymptotic normality result for a maximum quasi-likelihood estimator of the fixed parameters, which is different from the PQL estimator, for independent-cluster GLMMs.

This work is distinct from the above results: compared to Lyu and Welsh (2021a,b) we consider a more general random effects structure that permits random slopes in GLMMs. Meanwhile, Jiang et al. (2022) considered GLMMs but not the case when cluster sizes grow faster than the number of clusters; nor did they present results for predictors of random effects, both of which are considered in this article. Furthermore, we establish results for the prediction gap in GLMMs, which are new to the literature and allow unconditional inference to be performed for the random effects (noting unconditional inference for random effects in LMMs has been considered previously in a very different way through the unconditional mean

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0343

squared error of prediction, Kackar and Harville, 1984; Prasad and Rao, 1990). Note for the PQL estimator specifically, Vonesh et al. (2002), Hui et al. (2017) and Hui (2020) demonstrated estimation consistency under increasing cluster size and number of clusters, but did not develop any large sample distributional results.

It is important to remark that when cluster sizes do not increase, PQL is known to be asymptotically biased (e.g., Breslow and Lin, 1995). As such, increasing both the number of clusters and cluster size is a necessary condition for the PQL estimator to be consistent. Indeed, increasing number of clusters and cluster size is necessary for the consistency of other estimators such as the Laplace estimator (Ogden, 2017; Hui, 2020; Ogden, 2021). With this in mind, we develop our large sample distributional results under this setting, with the precise rates of growth to be formalised later. We note this asymptotic framework is relevant for many applications with large cluster sizes e.g., educational studies with large numbers of students (units) grouped within schools (clusters), and medical studies with large groups (clusters) of patients (units) treated at different hospitals.

We derive our asymptotic results for the PQL estimator under two distinct sampling regimes. In the first, we condition on the random effects, i.e. treat them as fixed effects, although we will still refer to them as random

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0343

effects for consistency. In the second, unconditional regime, we allow the random effects to be random. Conditional inference is appropriate when hypothetical resampling occurs within the same observed clusters, while unconditional inference may be more appropriate when (new) clusters are sampled from some population. Importantly, we demonstrate the asymptotic distributional results for the two regimes differ markedly. Conditional on the random effects we show the PQL estimator is asymptotically normally distributed around the true parameter values, with a convergence rate of $N^{1/2}$ (square root of the total number of observations) for the fixed effects and $n_i^{1/2}$ (square root of the cluster size of the ith cluster) for the random effects (which are now also fixed parameters). We find that when a variable is included as both a fixed and random effect covariate, the PQL estimator is asymptotically normally distributed around a sum-to-zero reparametrized version of the estimand. Unconditionally, we demonstrate the asymptotic normality of the PQL estimator for the fixed effects around the true values, but with a slower convergence rate of $m^{1/2}$ (square root of the number of clusters). Furthermore, we demonstrate that the "prediction gap" i.e., the difference between the PQL estimator of and the true random effect, is not in general asymptotically normally distributed; instead, it follows a normal scale-mixture when m grows faster than n_i.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0343

Our results have important implications for inference in GLMMs. There is a choice of whether conditional or unconditional inference is desired, with different asymptotic distributions needing to be applied in each case. Also, the potential asymptotic non-normality of the prediction gap has consequences in practice, since normality is often assumed when constructing prediction intervals for the random effects in GLMMs (Bates et al., 2015; Brooks et al., 2017). The theoretical results in this paper offer an important step towards more formal, rigorous asymptotic inference using PQL estimation (and perhaps other similar estimators) for GLMMs.

The structure of the article is as follows. In Section 2, we introduce GLMMs and PQL estimation. In Sections 3 and 4, we present and develop our asymptotic framework and results for the conditional and unconditional regimes. In Section 5, we present results from a simulation study which empirically verify our large sample developments. Finally, in Section 6 we discuss the implications of our results.

2. Generalized Linear Mixed Models

We study the independent-cluster generalized linear mixed model defined as follows. Let y_{ij} denote the jth measurement of cluster i, $\boldsymbol{x}_{ij}$ denote a vector of p_f fixed effect covariates, and $\boldsymbol{z}_{ij}$ denote a vector of p_r random

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0343

effect covariates, for $j = 1, \ldots, n_i$, and $i = 1, \ldots, m$. Let $N = \sum_{i=1}^{m} n_i$, $n = m^{-1}N$, $n_L = \min_{1 \le i \le m} n_i$, and $n_U = \max_{1 \le i \le m} n_i$. The m clusters are independent of each other. Conditional on a p_r-vector of random effects $\dot{\boldsymbol{b}}_i$, where the dot above any quantity is used to denote its true value (or that it is evaluated at the true parameter values), the responses y_{ij} from cluster i are assumed to be independent observations from the exponential family with mean $\dot{\mu}_{ij}$ and dispersion parameter $\dot{\phi}$. That is, $f(y_{ij}|\dot{\boldsymbol{\beta}}, \dot{\boldsymbol{b}}_i, \dot{\phi}) = \exp\left[\dot{\phi}^{-1}\{y_{ij}\dot{\vartheta}_{ij} - a(\dot{\vartheta}_{ij})\} + c(y_{ij}, \dot{\phi})\right]$, for known functions $a(\cdot)$, $c(\cdot)$, and $g(\cdot)$ satisfying $g(\dot{\mu}_{ij}) = g\{a'(\dot{\vartheta}_{ij})\} = \dot{\eta}_{ij} = \boldsymbol{x}_{ij}^{\top}\dot{\boldsymbol{\beta}} + \boldsymbol{z}_{ij}^{\top}\dot{\boldsymbol{b}}_i$, where $\dot{\boldsymbol{\beta}}$ denotes a p_f-vector of true fixed effect coefficients, and $\dot{\eta}_{ij}$ the corresponding true linear predictor. For ease of development, we assume that the canonical link is used, so that $\dot{\vartheta} = \dot{\eta}$. Commonly used distributions within the exponential family include the normal, Poisson, binomial and gamma distributions. The true realised random effects $\dot{\boldsymbol{b}}_i$ are independently and identically distributed (i.i.d.), drawn from a multivariate normal distribution with zero mean vector and unstructured $p_r \times p_r$ random effects covariance matrix $\dot{\boldsymbol{G}}$. That is, $\dot{\boldsymbol{b}}_i \overset{i.i.d.}{\sim} N(\boldsymbol{0}, \dot{\boldsymbol{G}})$.

Write $\boldsymbol{X}_i = [\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{in_i}]^{\top}$, and $\boldsymbol{Z}_i = [\boldsymbol{z}_{i1}, \ldots, \boldsymbol{z}_{in_i}]^{\top}$, so we can concatenate the means across the measurements for each cluster to obtain $g(\dot{\boldsymbol{\mu}}_i) = \boldsymbol{X}_i\dot{\boldsymbol{\beta}} + \boldsymbol{Z}_i\dot{\boldsymbol{b}}_i$ for $\dot{\boldsymbol{\mu}}_i = (\dot{\mu}_{i1}, \ldots, \dot{\mu}_{in_i})^{\top}$, where $g(\dot{\boldsymbol{\mu}}_i)$ denotes applying

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0343

the link function $g(\cdot)$ to $\dot{\boldsymbol{\mu}}_i$ component-wise. We can further concatenate across clusters and write $g(\dot{\boldsymbol{\mu}}) = \boldsymbol{X}\dot{\boldsymbol{\beta}} + \boldsymbol{Z}\dot{\boldsymbol{b}}$, with $\dot{\boldsymbol{\mu}} = (\dot{\boldsymbol{\mu}}_1^\top, \ldots, \dot{\boldsymbol{\mu}}_m^\top)^\top$, $\boldsymbol{X} = [\boldsymbol{X}_1^\top, \ldots, \boldsymbol{X}_m^\top]^\top$, $\boldsymbol{Z} = \text{bdiag}(\boldsymbol{Z}_1, \ldots \boldsymbol{Z}_m)$, and $\dot{\boldsymbol{b}} = (\dot{\boldsymbol{b}}_1^\top, \ldots, \dot{\boldsymbol{b}}_m^\top)^\top$. Here, bdiag() is the block-diagonal matrix operator, $\boldsymbol{X}$ is of dimension $N \times p_f$, and $\boldsymbol{Z}$ is an $N \times mp_r$ sparse block-diagonal matrix, with at most p_r non-zero components per row, and at most n_U non-zero components per column.

Let $\boldsymbol{y}_i = (y_{11}, \ldots, y_{1n_i})^\top$ and $\boldsymbol{y} = (\boldsymbol{y}_1^\top, \ldots, \boldsymbol{y}_m^\top)^\top$. Then the marginal log-likelihood function for the independent-cluster GLMM is given by

$$\ln f(\boldsymbol{y}|\boldsymbol{\beta}, \phi, \boldsymbol{G}) = \sum_{i=1}^{m} \ln \int \left(\prod_{j=1}^{n_i} f(y_{ij}|\boldsymbol{\beta}, \boldsymbol{b}_i, \phi) \right) f(\boldsymbol{b}_i|\boldsymbol{G})\, d\boldsymbol{b}_i. \tag{2.1}$$

The above integral is not available analytically except in the special case of a normal response with an identity link function. Let $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \boldsymbol{b}^\top)^\top$ denote the full vector of fixed and random effects. Then for a given $\boldsymbol{G}$ and ϕ, the PQL objective function for an independent-cluster GLMM is defined as

$$Q(\boldsymbol{\theta}) = \sum_{i=1}^{m} \sum_{j=1}^{n_i} \ln f(y_{ij}|\boldsymbol{\beta}, \boldsymbol{b}_i, \phi) - \frac{1}{2} \sum_{i=1}^{m} \boldsymbol{b}_i^\top \boldsymbol{G}^{-1} \boldsymbol{b}_i, \tag{2.2}$$

and the PQL estimator is defined as $\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta})$. As there are no integrals in (2.2), the computational cost of PQL estimation is low relative to standard maximum likelihood estimation (Breslow and Clayton, 1993).

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0343

Note for normal linear mixed models, the integral in the likelihood already possesses an analytical solution when an identity link is used, and PQL estimation is equivalent to the mixed model equations of Henderson (1973) assuming the error variance is known.

The PQL procedure provides explicit estimators of both the fixed and random effects. The latter is practically useful since the random effects play an important implicit role in fitting and using the GLMM. For instance, the realised values of the random effects (or functions thereof) are often important in prediction problems such as small-area estimation (Jiang, 2003; Pfeffermann, 2013), while the empirical distribution of the random effects estimators is often examined in model diagnostics (Hui et al., 2021). On the other hand, (2.2) alone does not incorporate estimation of the random effects covariance matrix. From a theoretical standpoint, existing papers on large sample theory for PQL and related objective functions have assumed $\dot{\boldsymbol{G}}$ is known for the purposes of asymptotic development (e.g., Vonesh et al., 2002; Nie, 2007). Practically speaking, several approaches have been suggested to estimate $\dot{\boldsymbol{G}}$ when applying PQL, for example by using a working LMM (Breslow and Clayton, 1993), the Laplace objective function (Hui et al., 2017), or simply the sample covariance matrix of the estimated random effects (Jiang et al., 2001). Indeed, Jiang et al. (2001) and Hui et al.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0343

(2017) demonstrated that the sample covariance of the estimated random effects is a consistent estimator of $\dot{\boldsymbol{G}}$ under suitable regularity conditions.

In this article, we set $\boldsymbol{G} = \hat{\boldsymbol{G}}$ in (2.2), where $\hat{\boldsymbol{G}}$ is a symmetric positive definite matrix that is either non-stochastic or its inverse $\hat{\boldsymbol{G}}^{-1}$ is stochastically bounded. Importantly, our large sample developments do not require $\hat{\boldsymbol{G}}$ to necessarily be a consistent estimator of the true random effects covariance matrix $\dot{\boldsymbol{G}}$. For example, while we can use the estimators of $\dot{\boldsymbol{G}}$ mentioned above, our theory also permits setting $\hat{\boldsymbol{G}}$ to some fixed matrix e.g., the identity matrix, say. Intuitively, this is because we develop our large sample results for PQL estimation in such a way so as to do not depend on the value of $\hat{\boldsymbol{G}}$ itself (in a spirit similar to that of Jiang et al., 2001; Fan and Li, 2012); only the true random effects covariance matrix $\dot{\boldsymbol{G}}$ appears in our theorems.

We also adopt the above approach for the dispersion parameter in the GLMM. That is, we set $\phi = \hat{\phi}$ in (2.2), where $\hat{\phi}$ is a known constant or a stochastically bounded term. In the Poisson and binomial distributions, $\hat{\phi}$ is set to its known true value $\dot{\phi} = 1$. In cases where the true dispersion parameter is unknown, we can use either a constant or one of the suggested estimators of the dispersion parameter in the literature (e.g., a scaled sum of squared conditional Pearson residuals). For the remainder of this article,

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0343

and as discussed in Section 1, we focus on the fixed and random effects in GLMMs. We do not discuss inferential properties of $\dot{\phi}$ and $\dot{\boldsymbol{G}}$.

3. Conditional on Random Effects

In many applications of independent-cluster GLMMs e.g., for longitudinal data, covariates included as random effects are also included as fixed effects (Cheng et al., 2010). With this in mind, we develop our results under the setting where all covariates are partnered i.e. included as both fixed and random effects such that $\boldsymbol{x}_{ij} = \boldsymbol{z}_{ij}$ for all (i, j) and $p_f = p_r =: p$. Next, let $\boldsymbol{A}$ be a $q \times (m+1)p$ matrix with the finite selection property. That is, for any row of $\boldsymbol{A}$, there exists an $m_0 \in \mathbb{N}$ such that the $\{(m_0+1)p+1\}$th to $\{(m+1)p\}$th components of the row are zero for all $m > m_0$. All components of $\boldsymbol{A}$ must have a component-wise limit, with at least one of these limits being non-zero. We partition $\boldsymbol{A}$ into $[\boldsymbol{A}_f, \boldsymbol{A}_r]$, where $\boldsymbol{A}_f$ is of dimension $q \times p$ and $\boldsymbol{A}_r$ is of dimension $q \times mp$. Also, for an arbitrary matrix $\boldsymbol{C}$, let $\boldsymbol{C}_{[i:j,k:l]}$ denote the sub-matrix comprising the ith to jth row and kth to lth column of $\boldsymbol{C}$ and $\boldsymbol{C}_{[i,]}$ and $\boldsymbol{C}_{[,j]}$ denote the ith row and jth column respectively. Similarly, for a vector $\boldsymbol{c}$ we let $\boldsymbol{c}_{[i:j]}$ denote the sub-vector formed by taking the ith to jth components; the quantity $\boldsymbol{c}_{[i]}$ simply denotes the ith component of $\boldsymbol{c}$.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0343

Let $\boldsymbol{\mu}_i(\boldsymbol{\theta}) = \{a'(\eta_{i1}), \ldots, a'(\eta_{in_i})\}^\top$, $\boldsymbol{\mu}(\boldsymbol{\theta}) = \{a'(\eta_{11}), \ldots, a'(\eta_{mn_m})\}^\top$, $\dot{\boldsymbol{W}}_i = \hat{\phi}^{-1}\text{diag}\{a''(\dot{\eta}_{i1}), \ldots, a''(\dot{\eta}_{in_i})\}$ and $\dot{\boldsymbol{W}} = \hat{\phi}^{-1}\text{diag}\{a''(\dot{\eta}_{11}), \ldots, a''(\dot{\eta}_{mn_m})\}$. Furthermore, write $\dot{\mu}_{ij} = a''(\dot{\eta}_{ij})$, $\dot{\boldsymbol{\mu}}_i = \boldsymbol{\mu}_i(\dot{\boldsymbol{\theta}})$ and $\dot{\boldsymbol{\mu}} = \boldsymbol{\mu}(\dot{\boldsymbol{\theta}})$, and let $\otimes$ denote the Kronecker product operator, $\boldsymbol{I}_m$ denote the $m \times m$ identity matrix, and $\boldsymbol{1}_m$ denote a matrix or vector of ones, with dimension indicated by the relevant subscripts. Furthermore, let $\boldsymbol{D}_r = \text{diag}(n_1^{1/2}\boldsymbol{1}_p, \ldots, n_m^{1/2}\boldsymbol{1}_p)$, $\boldsymbol{D} = \text{bdiag}(N^{1/2}\boldsymbol{I}_p, \boldsymbol{D}_r)$, $\boldsymbol{D}^* = \text{bdiag}(m^{1/2}\boldsymbol{I}_p, \boldsymbol{D}_r)$, $\boldsymbol{D}^+ = m^{1/2}\boldsymbol{I}_{(m+1)p}$, and define the two limiting quantities

$$\boldsymbol{\Omega} = \lim_{m,n_L\to\infty} \frac{\dot{\phi}}{\hat{\phi}} \boldsymbol{A}\, \text{bdiag}\left\{ \frac{1}{m}\sum_{i=1}^{m} \frac{n}{n_i} \left(\frac{\boldsymbol{X}_i^\top \dot{\boldsymbol{W}}_i \boldsymbol{X}_i}{n_i} \right)^{-1}, \left(\frac{\boldsymbol{X}_1^\top \dot{\boldsymbol{W}}_1 \boldsymbol{X}_1}{n_1} \right)^{-1}, \ldots, \left(\frac{\boldsymbol{X}_m^\top \dot{\boldsymbol{W}}_m \boldsymbol{X}_m}{n_m} \right)^{-1} \right\} \boldsymbol{A}^\top,$$

$$\boldsymbol{\Omega}_r = \lim_{m,n_L\to\infty} \frac{\dot{\phi}}{\hat{\phi}} \boldsymbol{A}_r \boldsymbol{D}_r \left(\boldsymbol{Z}^\top \dot{\boldsymbol{W}} \boldsymbol{Z} \right)^{-1} \boldsymbol{D}_r^\top \boldsymbol{A}_r^\top.$$

Note $\boldsymbol{\Omega}$ and $\boldsymbol{\Omega}_r$ are not actually functions of $\hat{\phi}$, since $\hat{\phi}\dot{\phi}^{-1}\dot{\boldsymbol{W}}_i = \dot{\phi}^{-1}\text{diag}\{a''(\dot{\eta}_{i1}), \ldots, a''(\dot{\eta}_{in_i})\}$ and similarly for $\dot{\boldsymbol{W}}$.

We consider the setting where both the minimum cluster size n_L and the number of clusters m grow to infinity, such that $n_i = O(n_L)$ uniformly for $i = 1, \ldots, m$. This implies for any $i = 1, \ldots, m$, we have $n_i = O(n)$, $n = O(n_i)$, $N = O(mn_i)$, and $mn_i = O(N)$. This restriction on the growth rates of the cluster sizes is commonly employed in asymptotic analysis of PQL estimation (e.g., Vonesh et al., 2002). Additionally, we require the

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0343

following regularity conditions.

(C1) The function $a(\eta)$ is at least three times continuously differentiable in its domain, with $0 < c_0 \leq a''(\eta) \leq c_0^{-1} < \infty$ and $|a'''(\eta)| \leq c_0^{-1} < \infty$ for some sufficiently small constant c_0.

(C2) For every $i = 1, \ldots, m$ and $j = 1, \ldots, n_i$, there exists a sufficiently large constant C_1 such that $\|\boldsymbol{x}_{ij}\|_\infty < C_1$ where $\|\cdot\|_\infty$ is the maximum norm. Furthermore, denote $\dot{\boldsymbol{H}}_i = (n_i^{-1}\hat{\phi}\dot{\phi}^{-1}\boldsymbol{X}_i^\top \dot{\boldsymbol{W}}_i \boldsymbol{X}_i)^{-1}$. Then for all $i = 1, \ldots, m$, the matrices $\lim_{n_i \to \infty} \dot{\boldsymbol{H}}_i = \dot{\boldsymbol{K}}_i$ and $\lim_{m, n_L \to \infty} m^{-1} \sum_{i=1}^m n n_i^{-1} \dot{\boldsymbol{H}}_i = \dot{\boldsymbol{K}}$ are positive definite with minimum and maximum eigenvalues bounded from above and below by c_1^{-1} and c_1 respectively, for a sufficiently small constant c_1.

(C3) The vector of true parameters $\dot{\boldsymbol{\theta}} = (\dot{\boldsymbol{\beta}}^\top, \dot{\boldsymbol{b}}^\top)^\top$, where $\dot{\boldsymbol{b}} = (\dot{\boldsymbol{b}}_1^\top, \ldots, \dot{\boldsymbol{b}}_m^\top)^\top$, is an interior point in some compact set $\Theta \subset \mathbb{R}^{(m+1)p}$.

(C4) The working matrix $\hat{\boldsymbol{G}}$ is positive definite, and its inverse is $O_p(1)$. Also, the working quantity $\hat{\phi}$ is strictly positive and $O_p(1)$.

(C5) For all $i = 1, \ldots, m$ and $n_i \in \mathbb{N}$, it holds that $E([n_i^{1/2}(\boldsymbol{X}_i^\top \dot{\boldsymbol{W}}_i \boldsymbol{X}_i + \hat{\boldsymbol{G}}^{-1})^{-1}\{\hat{\phi}^{-1}\boldsymbol{X}_i^\top(\boldsymbol{y}_i - \dot{\boldsymbol{\mu}}_i) - \hat{\boldsymbol{G}}^{-1}\dot{\boldsymbol{b}}_i\}]^4) < \infty$, where the power and expectation are applied component-wise.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0343

Conditions (C1) - (C3) are needed to guarantee the existence and regular behavior of the asymptotic variance for the PQL estimating function, and to establish a Lindeberg condition needed to obtain a central limit theorem. Condition (C4) is required to ensure that the shrinkage of the random effects is asymptotically negligible, and formalises our discussion of $\hat{\boldsymbol{G}}$ and $\hat{\phi}$ at the end of Section 2. Condition (C5) is needed to bound the order of $\|\hat{\boldsymbol{\theta}} - \dot{\boldsymbol{\theta}}\|_\infty$, and is satisfied by many distributions e.g. Poisson and binomial, when the random effects are normally distributed (see also van de Geer and Müller, 2012).

For the remainder of this section, we consider the regime where we condition on the random effects, so that $\dot{\boldsymbol{\theta}}$ is a $(m+1)p$-vector of constants. The assumptions and conditions outlined above however will be applied to both the conditional and unconditional regime.

3.1 Main Result for the Conditional Regime

Let $\mathbf{1}_m^* = (-1, \mathbf{1}_m^\top)^\top$. Then we have the following:

Theorem 1. *Assume Conditions (C1) - (C5) are satisfied and* $mn_L^{-1} \to 0$. *Then as* $m, n_L \to \infty$, *and conditional on the true vector of random effects* $\dot{\boldsymbol{b}}$, *it holds that*

(a) $\|\hat{\boldsymbol{\theta}} - (\dot{\boldsymbol{\theta}} - \mathbf{1}_m^* \otimes m^{-1} \sum_{i=1}^m \dot{\boldsymbol{b}}_i)\|_\infty = o_p(1)$.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0343

(b) $\boldsymbol{AD}\{\hat{\boldsymbol{\theta}} - (\dot{\boldsymbol{\theta}} - \mathbf{1}_m^* \otimes m^{-1}\sum_{i=1}^m \dot{\boldsymbol{b}}_i)\} \xrightarrow{D} N(\mathbf{0}, \boldsymbol{\Omega})$.

The first part of the theorem establishes consistency for the PQL estimator around a sum-to-zero reparametrized version of the true parameters (see below for more discussion on the latter aspect). The block diagonal structure of $\boldsymbol{\Omega}$ in the second part of the theorem shows that conditional on true random effects vector, the corresponding estimators are asymptotically independent between clusters, and also asymptotically independent of the fixed effects estimators.

We illustrate a few special cases of Theorem 1 using specific selection matrices. First, suppose $\boldsymbol{A} = [\boldsymbol{I}_p, \mathbf{0}_{p\times mp}]$. If $\sum_{i=1}^m \dot{\boldsymbol{b}}_i = \mathbf{0}_p$, then we obtain $\boldsymbol{AD}(\hat{\boldsymbol{\theta}} - \dot{\boldsymbol{\theta}}) = N^{1/2}(\hat{\boldsymbol{\beta}} - \dot{\boldsymbol{\beta}}) \xrightarrow{D} N(\mathbf{0}, \dot{\boldsymbol{K}})$ conditional on the random effects, where $\dot{\boldsymbol{K}}$ is the limiting matrix defined in Condition (C2). Also, suppose $\boldsymbol{A} = [\mathbf{0}_p, \boldsymbol{I}_p, \mathbf{0}_{p\times(m-1)p}]$. Then from Theorem 1, we have $\boldsymbol{AD}(\hat{\boldsymbol{\theta}} - \dot{\boldsymbol{\theta}}) = n_1^{1/2}(\hat{\boldsymbol{b}}_1 - \dot{\boldsymbol{b}}_1) \xrightarrow{D} N(\mathbf{0}, \dot{\boldsymbol{K}}_1)$, conditional on the random effects. The analogous result holds for choosing any particular cluster. Finally, since the random effects exhibit a slower convergence rate than the fixed effects, and noting the asymptotic independence, then for an arbitrary p-dimensional constant $\boldsymbol{a}$ we obtain $n_i^{1/2}\boldsymbol{a}^\top(\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{b}}_i - \dot{\boldsymbol{\beta}} - \dot{\boldsymbol{b}}_i) \xrightarrow{D} N\left(\mathbf{0}, \boldsymbol{a}^\top \dot{\boldsymbol{K}}_i \boldsymbol{a}\right); \quad i = 1, \ldots, m,$ conditional on the random effects. As an example, if the linear predictor involves a fixed and random intercept and a fixed and random slope for

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0343

a single covariate, then we set $\boldsymbol{a} = (1, x_{ij})^\top$ and obtain $n_i^{1/2}(\hat{\eta}_{ij} - \dot{\eta}_{ij}) = n_i^{1/2}(\hat{\beta}_0 + \hat{b}_{i0} + \hat{\beta}_1 x_{ij} + \hat{b}_{i1} x_{ij} - \dot{\beta}_0 - \dot{b}_{i0} - \dot{\beta}_1 x_{ij} - \dot{b}_{i1} x_{ij}) \xrightarrow{d} N(0, \boldsymbol{a}^\top \dot{\boldsymbol{K}}_i \boldsymbol{a})$.

For statistical inference, we can appeal to Slutsky's Theorem and replace $\dot{\boldsymbol{K}}_i$ with $\hat{\boldsymbol{H}}_i$, and $\dot{\boldsymbol{K}}$ with $m^{-1}\sum_{i=1}^m n n_i^{-1} \hat{\boldsymbol{H}}_i$. Here $\hat{\boldsymbol{H}}_i$ is defined as $(n_i^{-1} \boldsymbol{X}_i^\top \hat{\boldsymbol{W}}_i \boldsymbol{X}_i)^{-1}$ where $\hat{\boldsymbol{W}}_i = \tilde{\phi}^{-1}\text{diag}\{a''(\hat{\eta}_{i1}), \ldots, a''(\hat{\eta}_{in_i})\}$ for some consistent estimator of the dispersion parameter $\tilde{\phi}$ e.g., based on the inverse scaled sum of squared conditional Pearson residuals. Theorem 1 then provides a straightforward way to construct confidence intervals, say, for all the parameters and combinations thereof. In fact, the forms of these intervals are similar to standard results in (penalized) GLMs(McCulloch and Searle, 2004): this is not surprising given we are working conditional on the true vector of random effects. say.

Finally, note the PQL estimator is consistent for a sum-to-zero reparametrized version of the true parameters. This occurs because the PQL estimators of the random effects must satisfy a sum-to-zero constraint regardless of the underlying true parameter values, and under a conditional regime, this induces an asymptotic bias $\mathbf{1}_m^* \otimes (m^{-1}\sum_{i=1}^m \dot{\boldsymbol{b}}_i)$ in Theorem 1, which can be interpreted as shifting the mean of the random effects into the corresponding fixed effects. We offer more discussion around this asymptotic bias in the supplementary material.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0343

4. Unconditional Regime

We now turn to establishing results under an unconditional regime i.e., treating $\dot{\boldsymbol{b}}_i$'s as random instead of conditioning on them. This has two main implications. First, in the unconditional setting the quantity $m^{-1}\sum_{i=1}^{m}\dot{\boldsymbol{b}}_i$ is no longer deterministic and should not be treated as a bias term. Instead, it is of order $O_p(m^{-1/2})$, and so competes with other leading terms in the relevant Taylor expansion to be the dominating term. This results in a reduction of the rate of convergence for the fixed effects estimator, from $N^{1/2}$ in the conditional regime to $m^{1/2}$ in the unconditional regime. Second, in contrast to the conditional regime, the observations within the same cluster are no longer independent. This has ramifications when applying the central limit theorem to establish asymptotic multivariate normality. In Section 4.1, we provide a simple but insightful example based on a Poisson random intercept model, which demonstrates that the prediction gap is not always asymptotically normally distributed.

The two approximations below, derived from the Taylor expansion of the PQL objective function, will be central to understanding the large sample developments we make on a more intuitive level. For a given $\hat{\phi}$, we have

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0343

$$\hat{\boldsymbol{\beta}} - \dot{\boldsymbol{\beta}} = m^{-1} \sum_{i=1}^{m} \dot{\boldsymbol{b}}_i + o_p(1) \tag{4.1}$$

$$\hat{\boldsymbol{b}} - \dot{\boldsymbol{b}} = -\mathbf{1}_m \otimes m^{-1} \sum_{i=1}^{m} \dot{\boldsymbol{b}}_i + (\boldsymbol{Z}^\top \dot{\boldsymbol{W}} \boldsymbol{Z})^{-1} \{\hat{\phi}^{-1} \boldsymbol{Z}^\top (\boldsymbol{y} - \dot{\boldsymbol{\mu}})\} + o_p(1). \tag{4.2}$$

We will refer to both equations in the discussion of the theorems to be presented later on.

4.1 Prediction Gap - Counterexample

We offer a motivating and insightful example to illustrate that the prediction gap is not, in general, asymptotically normally distributed. This example also offers a simple case where $\boldsymbol{X}_i \neq \boldsymbol{Z}_i$, and offers an interesting comparison to the theory established under the assumption of $\boldsymbol{X}_i = \boldsymbol{Z}_i$.

Consider a Poisson random intercept model with canonical log link. That is, the true model is given by $f(y_{ij}|\dot{b}_i) = \exp(y_{ij}\dot{\eta}_{ij} - \dot{\mu}_{ij})/(y_{ij}!)$ with $\ln(\dot{\mu}_{ij}) = \dot{\eta}_{ij} = \dot{b}_i$, and $\dot{b}_i \overset{i.i.d.}{\sim} N(0, \dot{\sigma}_b^2)$. Assume a working $\hat{\sigma}_b^2$, and apply PQL estimation to estimate the random effects b_i for $i = 1, \ldots, n$. For simplicity, we also assume a balanced design, such that $n_i = n$ for all $i = 1, \ldots, m$. Then it is possible to show (see the supplementary material for the formal derivation) that when $mn^{-2} \to 0$, the prediction gap of the

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0343

first cluster $\hat{b}_1 - \dot{b}_1$ satisfies

$$n^{1/2}(\hat{b}_1 - \dot{b}_1) = n^{-1/2} \sum_{j=1}^{n} \{y_{1j} \exp(-\dot{b}_1) - 1\} + o_p(1). \tag{4.3}$$

Therefore, we obtain $\hat{b}_1 = \dot{b}_1 + o_p(1)$, and similarly for each cluster $i = 1, \ldots, m$. When conditioned on $\dot{b}_1$, $n^{-1/2} \sum_{j=1}^{n} \{y_{1j} \exp(-\dot{b}_1) - 1\}$ is a normalised sum of independent random variables. Unconditionally however, the sum consists of an exchangeable collection of uncorrelated but dependent random variables with mean zero and finite non-zero variance. Using the central limit theorem for exchangeable random variables (Blum et al., 1958), it can be subsequently be shown that the quantity $n^{-1/2} \sum_{j=1}^{n} \{y_{1j} \exp(-\dot{b}_1) - 1\}$, and hence $n^{1/2}(\hat{b}_1 - \dot{b}_1)$, is not asymptotically normally distributed.

With the above example in mind, we now state the main results for the unconditional regime.

4.2 Fixed Effects

We have the following result for the PQL estimator of the fixed effects under an unconditional regime.

Theorem 2. *Assume Conditions (C1) - (C5) are satisfied, and* $mn_L^{-2} \to 0$. *Then as* $m, n_L \to \infty$ *and unconditional on the random effects* $\dot{\boldsymbol{b}}$, *it holds that* $m^{1/2}(\hat{\boldsymbol{\beta}} - \dot{\boldsymbol{\beta}}) \xrightarrow{D} N(\mathbf{0}, \dot{\boldsymbol{G}})$.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0343

This result should not be too surprising given the form of (4.1). Furthermore, the rate of convergence and asymptotic distribution coincides with the result obtained by Jiang et al. (2022) for the partnered fixed effects for the quasi-maximum likelihood estimator. More importantly, Theorem 2 allows practitioners to straightforwardly perform statistical inference for the fixed effects, so long as $mn_L^{-2} \to 0$. Although $\dot{\boldsymbol{G}}$ is not known, we can appeal to Slutsky's theorem and replace it with a consistent estimator (e.g., the sample covariance matrix of the estimated random effects). Theorem 2 contrasts with Theorem 1 derived under the conditional regime, where $mn_L^{-1} \to 0$ is required but the convergence rate is $N^{1/2}$. This reduction in the rate of convergence arises because the leading term in the Taylor expansion is different: in the unconditional regime, it is simply the normalised sum of random effects over all the clusters, and thus its variability is dominated by the term $m^{-1/2} \sum_{i=1}^{m} \dot{\boldsymbol{b}}_i$. However, this term is deterministic in the conditional regime, and serves to enforce a sum-to-zero constraint instead as discussed in Section 3. Generally speaking, the Taylor expansion can be interpreted as comprising terms which either capture the stochasticity in the random effects vector $\dot{\boldsymbol{b}}$, or the stochasticity in responses y_{ij} given the random effects. These terms compete with each other, and which one dominates depends on the relative rates of m and n_i. This intricacy in the

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0343

nature of the results will be made apparent in our results for the prediction gap in Section 4.4.

4.3 Estimators of the Random Effects

Next, we state a convergence result for the PQL estimators of the random effects under the unconditional regime.

Theorem 3. *Assume Conditions (C1) - (C5) are satisfied and* $mn_L^{-2} \to 0$. *Then as* $m, n_L \to \infty$ *and unconditional on the random effects* $\dot{\boldsymbol{b}}$, *it holds that* $\boldsymbol{A}_r(\hat{\boldsymbol{b}} - \dot{\boldsymbol{b}}) \xrightarrow{P} \mathbf{0}_q$.

Practically, Theorem 3 confirms the asymptotic distribution of a finite subset of the PQL estimators is the distribution of the random effects themselves. This can play a useful role for helping to validate the examination of the empirical distribution of PQL estimators $\hat{\boldsymbol{b}}$ as a model diagnostic tool. For instance, if the random effects are normally distributed and $\boldsymbol{A}_r$ only selects the first cluster, then we would expect $\hat{\boldsymbol{b}}_1$ to have an approximate $N(\mathbf{0}, \dot{\boldsymbol{G}})$ distribution. On the other hand, Theorem 3 does not help us when it comes to performing likelihood-based inference for the true random effects $\dot{\boldsymbol{b}}$, as this does not appear in the approximation $\hat{\boldsymbol{b}}_1 \sim N(\mathbf{0}, \dot{\boldsymbol{G}})$ itself.

As an aside, note the above means we can apply the continuous mapping theorem and show that $q^{-1}\sum_{i=1}^{q} \hat{\boldsymbol{b}}_i\hat{\boldsymbol{b}}_i^\top - q^{-1}\sum_{i=1}^{q} \dot{\boldsymbol{b}}_i\dot{\boldsymbol{b}}_i^\top \xrightarrow{P} 0$ for any $q \in \mathbb{N}$.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0343

Since $q^{-1}\sum_{i=1}^{q}\dot{\boldsymbol{b}}_i\dot{\boldsymbol{b}}_i^\top \xrightarrow{P} \dot{\boldsymbol{G}}$ as $q \to \infty$, this further reiterates the use of a sample covariance matrix of the estimated random effects as an estimator of $\dot{\boldsymbol{G}}$ (consistent with Jiang et al., 2001; Hui et al., 2017).

4.4 Prediction Gap

In this section, we present a result for the large sample distribution of a finite subset of the prediction gap, $\hat{\boldsymbol{b}} - \dot{\boldsymbol{b}}$, in the unconditional regime. As mentioned above, the asymptotic distribution as well as the convergence rate of the prediction gap depends on the relative rates of growth of m and n_i. This contrasts with the conditional regime, where there is no dependence on the relative rate and the PQL estimator of the random effects is always normally distributed with the convergence rate $n_i^{1/2}$.

We first introduce some terminology. Suppose we have two arbitrary continuous cumulative distribution functions (cdfs) F_1 and F_2 with supports in $\mathbb{R}^p$. Then we define the convolution of F_1 and F_2, denoted $F_1 * F_2$, as $(F_1 * F_2)(\boldsymbol{z}) = \int_{\mathbb{R}^p} F_1(\boldsymbol{z} - \boldsymbol{\tau})dF_2(\boldsymbol{\tau})$. Next, for a random variable $\boldsymbol{P}$, we say $\boldsymbol{P} \sim \text{mixN}\{\boldsymbol{\mu}(\boldsymbol{b}), \boldsymbol{\Sigma}(\boldsymbol{b}), F_{\boldsymbol{b}}\}$ if $\boldsymbol{P}|\boldsymbol{b} \sim N\{\boldsymbol{\mu}(\boldsymbol{b}), \boldsymbol{\Sigma}(\boldsymbol{b})\}$ and $F_{\boldsymbol{b}}$ is the cdf of $\boldsymbol{b}$, where the conditional mean vector $\boldsymbol{\mu}(\boldsymbol{b})$ and covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{b})$ may depend on $\boldsymbol{b}$. In other words, $\boldsymbol{P}$ has cdf $F_{\boldsymbol{P}}(\boldsymbol{p}) = \int \Psi_{\boldsymbol{P}|\boldsymbol{b}}(\boldsymbol{p})dF_{\boldsymbol{b}}(\boldsymbol{b})$, where $\Psi_{\boldsymbol{P}|\boldsymbol{b}}$ is the cdf of $N\{\boldsymbol{\mu}(\boldsymbol{b}), \boldsymbol{\Sigma}(\boldsymbol{b})\}$. A special case of this normal

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0343

scale-mixture distribution is when $\boldsymbol{\mu}(\boldsymbol{b})$ and $\boldsymbol{\Sigma}(\boldsymbol{b})$ do not depend on $\boldsymbol{b}$, so that $F_{\boldsymbol{P}}(\boldsymbol{p}) = \int \Psi_{\boldsymbol{P}|\boldsymbol{b}}(\boldsymbol{p}) dF_{\boldsymbol{b}}(\boldsymbol{b}) = \Psi_{\boldsymbol{P}|\boldsymbol{b}}(\boldsymbol{p}) \int dF_{\boldsymbol{b}}(\boldsymbol{b}) = \Psi_{\boldsymbol{P}|\boldsymbol{b}}(\boldsymbol{p})$; in other words, the normal scale-mixture distribution reduces to a normal distribution. Note estimators with asymptotic normal mixture distributions have arisen in previous literature, for instance, on results relating to local asymptotic normality and non-ergodic models (Cam and Yang, 1988; Basawa and Scott, 2012).

Using the above definition, we obtain the following results.

Theorem 4. *Assume Conditions (C1)-(C5) are satisfied and* $mn_L^{-2} \to 0$. *Then as* $m, n_L \to \infty$ *and unconditional on the random effects* $\dot{\boldsymbol{b}}$, *for each* $i = 1, \ldots, m$ *we have the following:*

(a) If $mn_i^{-1} \to \infty$, *then* $n_i^{1/2}(\hat{\boldsymbol{b}}_i - \dot{\boldsymbol{b}}_i) \xrightarrow{D} mixN(\boldsymbol{0}, \dot{\boldsymbol{K}}_i, F_{\dot{\boldsymbol{b}}_i})$.

(b) If $mn_i^{-1} \to \gamma_i \in (0, \infty)$, *then* $n_i^{1/2}(\hat{\boldsymbol{b}}_i - \dot{\boldsymbol{b}}_i) \xrightarrow{D} mixN(\boldsymbol{0}, \dot{\boldsymbol{K}}_i, F_{\dot{\boldsymbol{b}}_i}) * N(\boldsymbol{0}, \gamma_i^{-1}\dot{\boldsymbol{G}})$.

(c) If $mn_i^{-1} \to 0$, *then* $m^{1/2}(\hat{\boldsymbol{b}}_i - \dot{\boldsymbol{b}}_i) \xrightarrow{D} N(\boldsymbol{0}, \dot{\boldsymbol{G}})$.

Corollary 1. *Assume Conditions (C1)-(C5) are satisfied, and* $mn_L^{-2} \to 0$. *If* $mn_L^{-1} \to \infty$, *then as* $m, n_L \to \infty$ *and unconditional on the random effects* $\dot{\boldsymbol{b}}$, $\boldsymbol{A}_r \boldsymbol{D}_r(\hat{\boldsymbol{b}} - \dot{\boldsymbol{b}}) \xrightarrow{D} mixN(\boldsymbol{0}, \boldsymbol{\Omega}_r, F_{\dot{\boldsymbol{b}}})$.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0343

Theorems 3 and 4 bears some similarity to the results of Lyu and Welsh (2021a), who show for LMMs that the distribution of the EBLUP can asymptotically be written as the convolution between the distribution of the random effects and the distribution of a smaller order stochastic term. However, the above is the first to establish such results for GLMMs. Theorem 4 states that the correct asymptotic distribution to use when performing inference using the PQL estimate of the random effects depends on the relative growth rates of m and n_i. As hinted at previously, this is a consequence of there being two competing terms in the corresponding Taylor expansion (4.2): one term arising from the random effects, and the other term arising from the distribution of the responses given the random effects.

When $mn_i^{-1} \to \infty$ i.e., the number of clusters grows faster than the cluster size, the appropriate asymptotic distribution is given by the scale-mixture distribution $\text{mixN}\{\mathbf{0}, (\hat{\phi}\dot{\phi}^{-1}\boldsymbol{X}_i^\top \dot{\boldsymbol{W}}_i \boldsymbol{X}_i)^{-1}, F_{\boldsymbol{b}_i}\}$, noting again that $\hat{\phi}\dot{\phi}^{-1}\dot{\boldsymbol{W}}_i = \dot{\phi}^{-1}\text{diag}\{a''(\dot{\eta}_{i1}), \ldots, a''(\dot{\eta}_{in_i})\}$. Corollary 1 offers a slightly more general result than that given in Theorem 4 for the $mn_L^{-1} \to \infty$ case. Note in the linear case, the GLM iterative weights $\dot{\boldsymbol{W}}$ do not depend on the random effects $\dot{\boldsymbol{b}}$, and so the corresponding normal scale-mixture distribution reduces to a normal distribution, consistent with the asymptotic normality

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0343

result derived for the EBLUP in Lyu and Welsh (2021a). Practically, numerical techniques or simulation are required to compute the quantiles of the normal scale-mixture distribution for constructing prediction intervals. We use this approach in our simulations in Section 5.

When $mn_i^{-1} \to 0$ i.e., the cluster sizes grows faster than the number of clusters, Theorem 4 shows that the appropriate approximation to consider is the normal distribution $N(\mathbf{0}, n^{-1}\dot{\boldsymbol{G}})$. Note this is identical to the fixed effects result of Theorem 2, and yields relatively straightforward prediction intervals for $\dot{\boldsymbol{b}}_i$ as long as we have a consistent estimator for $\dot{\boldsymbol{G}}$. Intuitively, the asymptotic distribution here is identical to that derived in Theorem 2 because the dominating terms in the Taylor expansions in both cases are effectively the same. Finally, when $mn_i^{-1} \to \gamma \in (0, \infty)$, Theorem 4b states that the asymptotic distribution of the PQL estimates is given by the convolution of the two cases above, noting that these two leading terms in the Taylor expansion are asymptotically independent. Again, numerical techniques/simulations are needed to compute prediction intervals.

In summary, Theorem 4 offers an asymptotically valid way of computing prediction intervals for the realised random effects in the unconditional regime, when the random effects have a corresponding partnered fixed effect in the model. It implies that estimating the variance of the prediction

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0343

gap, and then naively assuming normality in order to construct prediction intervals for the random effects, will fail to yield asymptotically correct inference under the unconditional regime for PQL estimation.

4.5 Linear Predictor

Neither Theorems 2 nor 4 above derive the joint distribution of the fixed effects estimator and prediction gap, of which the linear predictor is a function. Below, to address this, we establish a separate result specifically for the sum of a random effect and its partnered fixed effect, given an arbitrary p-dimensional constant vector $\boldsymbol{a}$.

Theorem 5. *Assume Conditions (C1)-(C5) are satisfied,* $mn_L^{-2} \to 0$, *and* $mn_U^{-1/2} \to \infty$. *Then as* $m, n_L \to \infty$ *and unconditional on the random effects* $\dot{\boldsymbol{b}}$, *it holds for each* $i = 1, \ldots, m$ *that* $n_i^{1/2}\boldsymbol{a}^\top(\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{b}}_i - \dot{\boldsymbol{\beta}} - \dot{\boldsymbol{b}}_i) \xrightarrow{D} mixN(\boldsymbol{0}, \boldsymbol{a}^\top \dot{\boldsymbol{K}}_i \boldsymbol{a}, F_{\dot{\boldsymbol{b}}_i})$.

As an example, consider again a linear predictor involving a fixed and random intercept and a fixed and random slope for a single covariate. Then we set $\boldsymbol{a} = (1, x_{ij})^\top$ and obtain $n_i^{1/2}(\hat{\eta}_{ij} - \dot{\eta}_{ij}) = n_i^{1/2}(\hat{\beta}_0 + \hat{b}_{i0} + \hat{\beta}_1 x_{ij} + \hat{b}_{i1} x_{ij} - \dot{\beta}_0 - \dot{b}_{i0} - \dot{\beta}_1 x_{ij} - \dot{b}_{i1} x_{ij}) \xrightarrow{d} mixN(0, \boldsymbol{a}^\top \dot{\boldsymbol{K}}_i \boldsymbol{a}, F_{\dot{\boldsymbol{b}}_i})$. For performing inference on the linear predictor in a GLMM or functions thereof, Theorem 5 states that we again need to employ the normal scale-mixture distribution. This

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0343

result differs from the asymptotic normality of the linear predictor derived under the conditional regime in Section 3. Note we can also develop a similar result for the difference between the prediction gaps of two clusters, and we refer to the supplementary material for details of this result.

To conclude the section, we remark that Theorems 2-5 do not offer results on the joint distribution of the prediction gap and fixed effects. However, we know from the associated proof that the prediction gaps for each cluster are asymptotically independent from each other as well as from the fixed effects estimator when $mn_U^{-1} \to \infty$ and $mn_L^{-2} \to 0$, and so a joint distribution result can be derived from this.

5. Simulation Study

We performed a numerical study to assess the usefulness of our asymptotic results in finite samples. We simulated data from an independent-cluster GLMM with five fixed and random effect covariates, considering Poisson and Bernoulli responses, as follows. First, we set the first component of $\boldsymbol{x}_{ij} = \boldsymbol{z}_{ij}$ equal to one to represent a fixed/random intercept. The second and third components are simulated from a bivariate normal distribution with mean zero and standard deviation one, with correlation equal to 0.5. The fourth component is generated independently from a standard normal

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0343

distribution, and the last component is simulated from a Bernoulli distribution with a probability of success equal to 0.5. Next, we set the 5-vector of true fixed effect coefficients to either $\dot{\boldsymbol{\beta}} = (2, 0.1, -0.1, 0.1, 0.1)^\top$ for Poisson responses, or $\dot{\boldsymbol{\beta}} = (-0.1, 0.1, -0.1, 0.1, 0.1)^\top$ for Bernoulli responses and the 5×5 random effects covariance matrix in both cases to $\dot{\boldsymbol{G}} = \boldsymbol{I}_5$. Based on these true parameter values, we simulated the random effect coefficients $\dot{\boldsymbol{b}}_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \dot{\boldsymbol{G}})$. Finally, conditional on $\dot{\boldsymbol{b}}_i$ the responses y_{ij} were generated from either a Poisson distribution with log link, or a Bernoulli distribution with logit link. We varied the number of clusters as $m = \{25, 50, 100, 200, 400\}$ and the cluster sizes $n_i = n = \{25, 50, 100, 200, 400\}$, noting we assumed equal cluster sizes in the simulation design for simplicity For each combination of (m, n), we simulated 1000 datasets. For the conditional regime, we simulated $\dot{\boldsymbol{b}}$ only once and conditioned on this for all simulated datasets; for unconditional regime, we simulated a new $\dot{\boldsymbol{b}}$ for each simulated dataset.

For each simulated dataset, we fitted the corresponding GLMM using PQL estimation, where we use the sample covariance matrix of the estimated random effects as our update for $\hat{\boldsymbol{G}}$. That is, we iteratively maximize equation (2.2) with respect to $\boldsymbol{\beta}$ and $\boldsymbol{b}$ for a given $\hat{\boldsymbol{G}}$ (noting $\hat{\phi} = 1$ is known for both these distributions), and update $\hat{\boldsymbol{G}}$ as $m^{-1} \sum_{i=1}^{m} \hat{\boldsymbol{b}}_i \hat{\boldsymbol{b}}_i^\top$, until

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0343

convergence.

We assessed performance separately under the conditional and unconditional regimes. In the former, we examined the empirical coverage probability of 95% coverage intervals constructed for $\boldsymbol{\beta}$ and for $\boldsymbol{b}_1$ (the choice of the first cluster is arbitrary). The intervals were constructed based on Theorem 1, with the asymptotic covariance matrix $\boldsymbol{\Omega}$ computed using the true parameter values. We refer to such intervals as coverage intervals as opposed to confidence intervals. We also performed Shapiro-Wilk tests on the components of the (1000) realised PQL estimates of $\boldsymbol{\beta}$ and $\boldsymbol{b}_1$, in order to assess the asymptotic normality of their respective sampling distributions. For the unconditional regime, we examined the empirical coverage probability of 95% coverage intervals constructed from Theorems 2 and 4 respectively. Again, this was done for the fixed effect coefficients $\boldsymbol{\beta}$ and the random effects for the the first cluster $\boldsymbol{b}_1$. To construct all intervals, we used the true parameter values to compute the relevant asymptotic variance (this was done solely to reduce the computational burden of the numerical study), and, when required, obtained quantiles of relevant normal scale-mixture distributions by directly simulating 10,000 samples from them. We also performed Shapiro-Wilk tests on the components of the (1000) realised values of $\hat{\boldsymbol{\beta}} - \dot{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{b}}_1 - \dot{\boldsymbol{b}}_1$. Finally, we examined histograms for the third

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0343

components of $\hat{\boldsymbol{\beta}}-\dot{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{b}}_1-\dot{\boldsymbol{b}}_1$, which are representative of the histograms of the other components, as an additional method of assessing asymptotic normality of the corresponding sampling distributions.

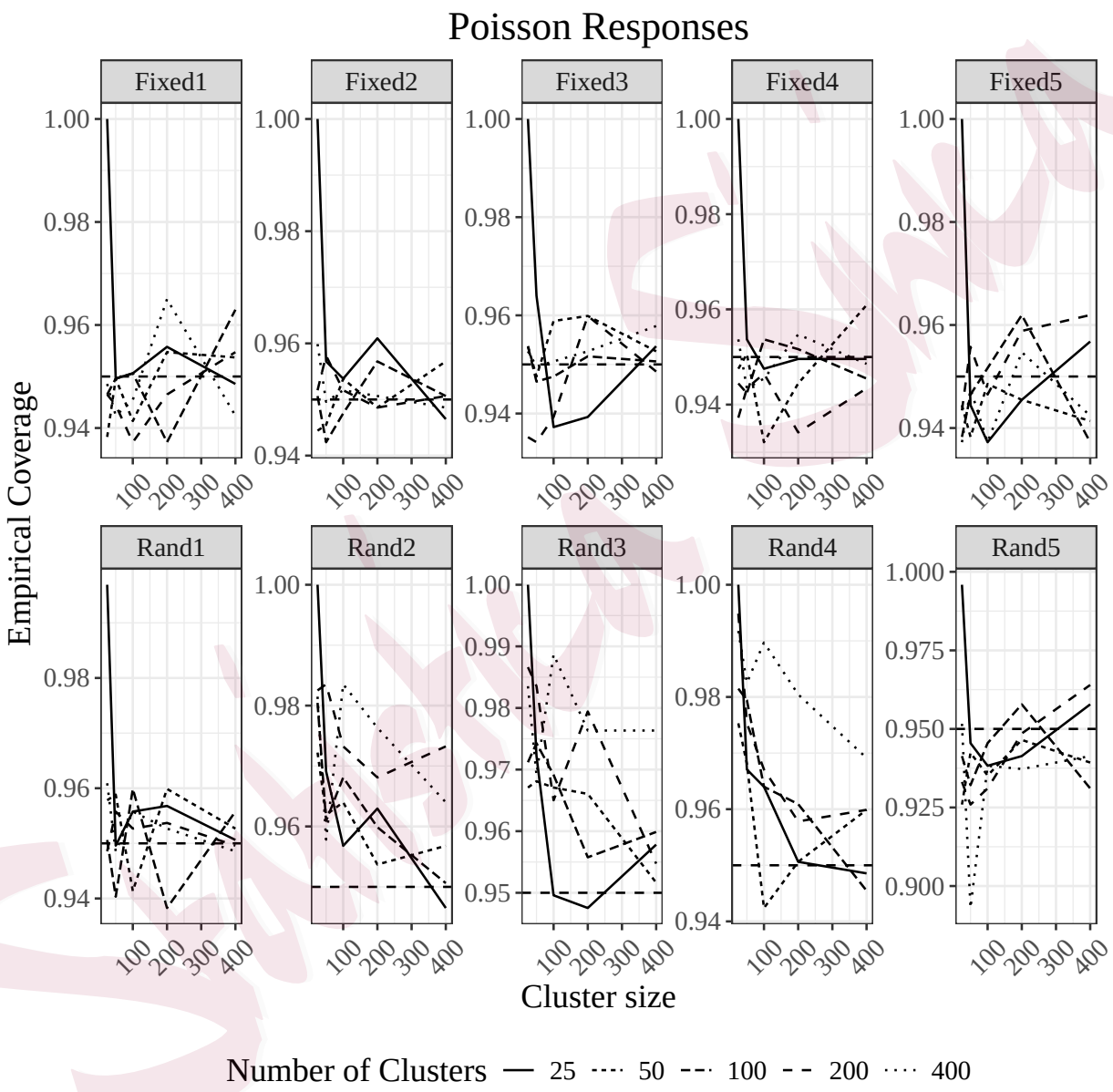
5.1 Simulation Results

For reasons of brevity, below we focus on results for the unconditional regime. Results for the conditional regime are presented in the supplementary material and largely support the use of Theorem 1 for inference.

For the unconditional regime, Figures 1 and 2 display the empirical coverage probabilities and results from applying the Shapiro-Wilk test, respectively. For the fixed effect coefficients, the coverage probabilities for the intervals obtained based on Theorem 2 were relatively accurate across most combinations of (m, n), with the exception of when $(m, n) = (25, 25)$. For the random effect coefficients, the coverage probabilities for intervals calculated based on Theorem 4 approached the nominal coverage rapidly as (m, n) increased for the Poisson response case, while for the Bernoulli case convergence was slightly slower due to the reduced amount of information per response.

The Shapiro-Wilk tests run were consistent with the conclusions reached in Theorems 2–4. Specifically, PQL estimates of the fixed effect coefficients

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0343



Statistica Sinica: Preprint
doi:10.5705/ss.202023.0343

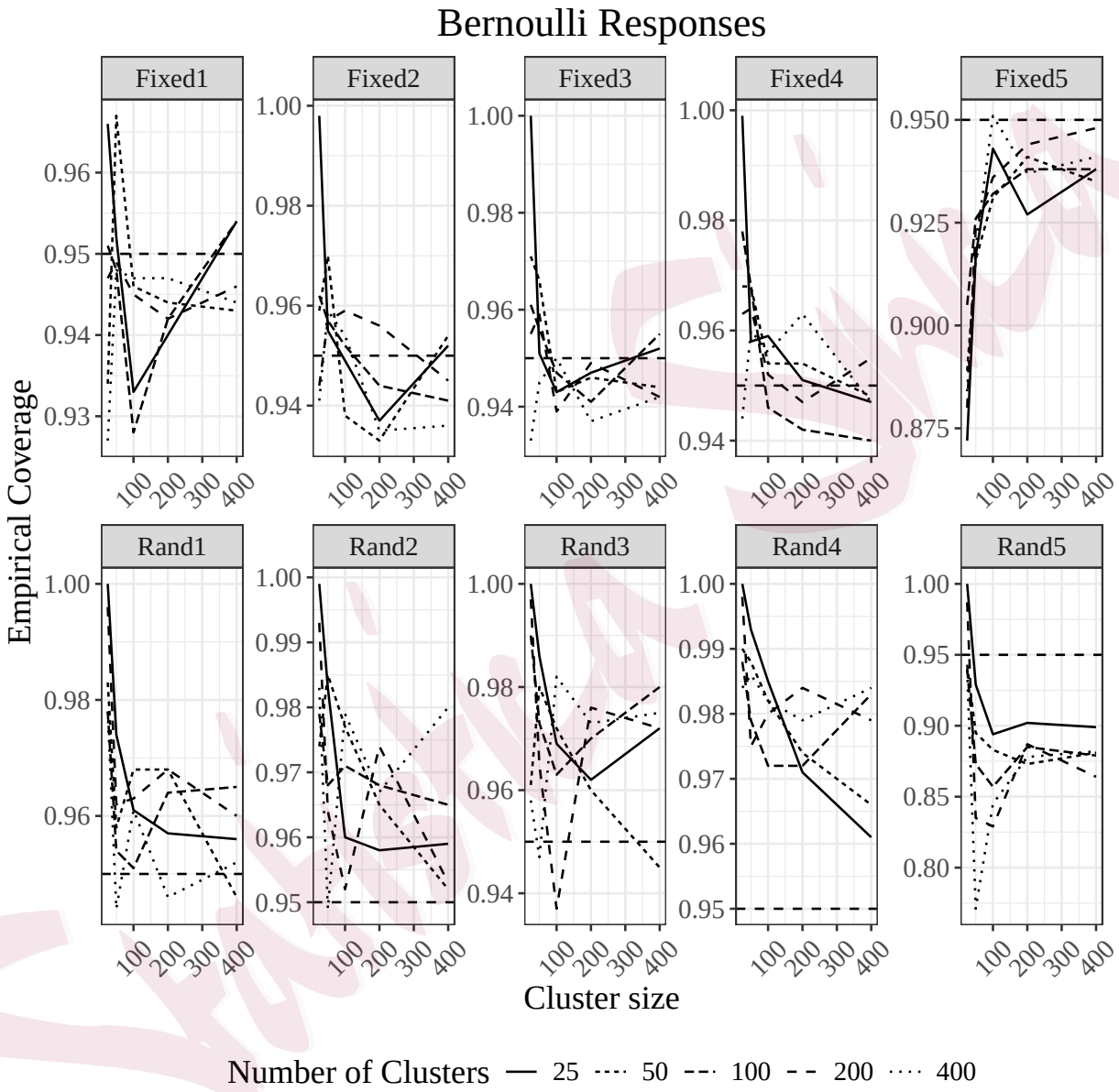


Figure 1: Empirical coverage probability of 95% coverage intervals for the fixed and random effects, obtained under the unconditional regime.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0343

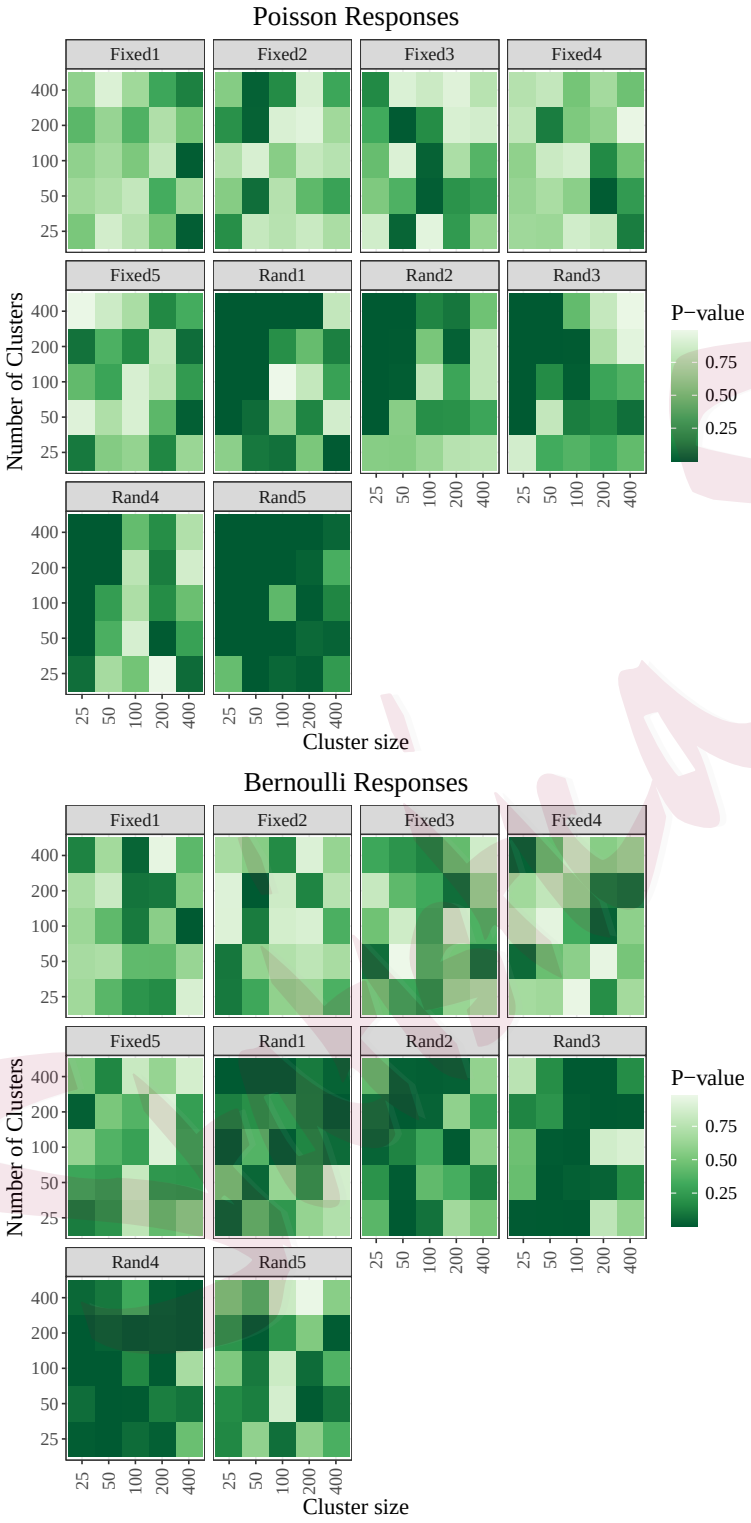
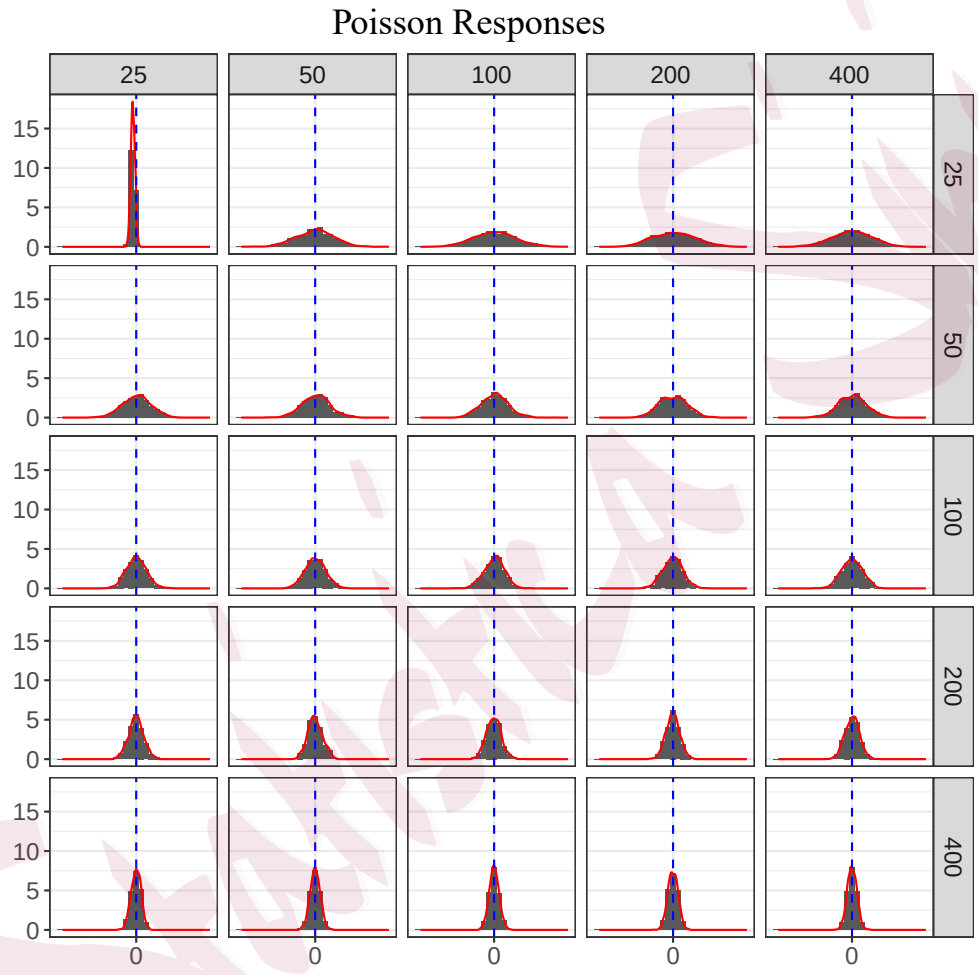
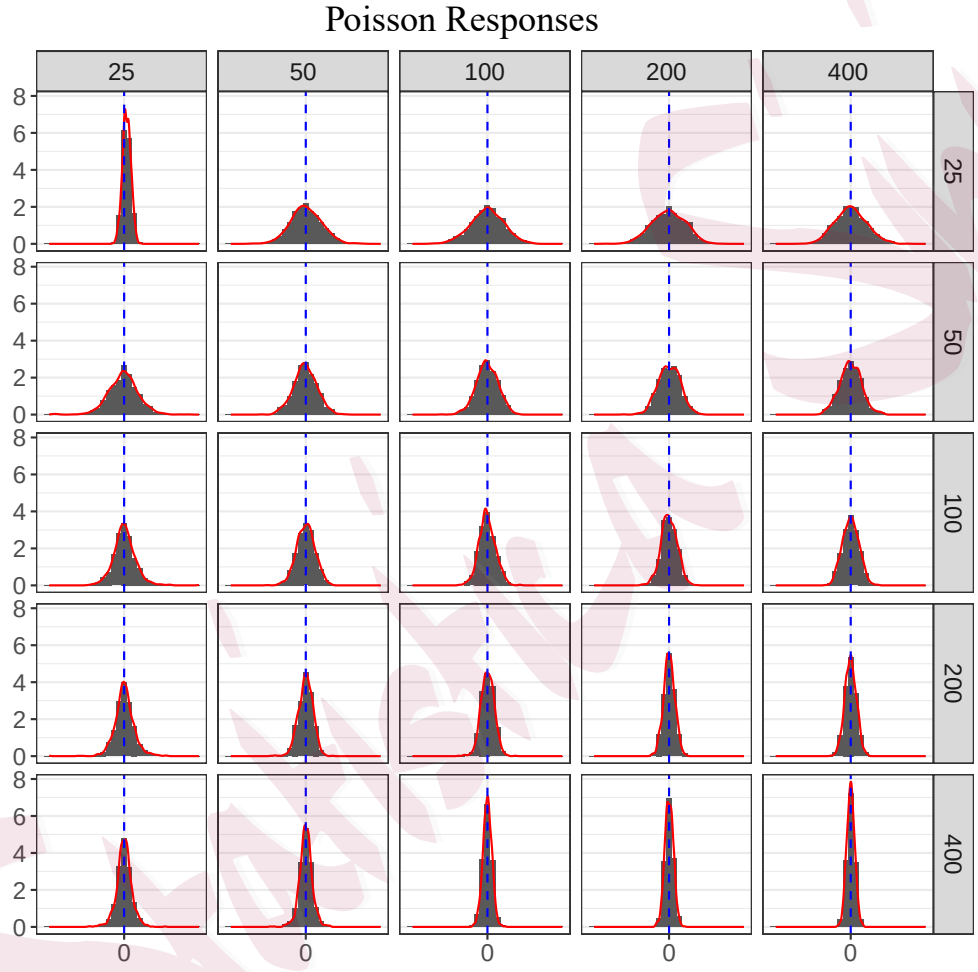


Figure 2: p-values from Shapiro-Wilk tests applied to the fixed and random effects estimates obtained using maximum PQL estimation, under the unconditional regime.

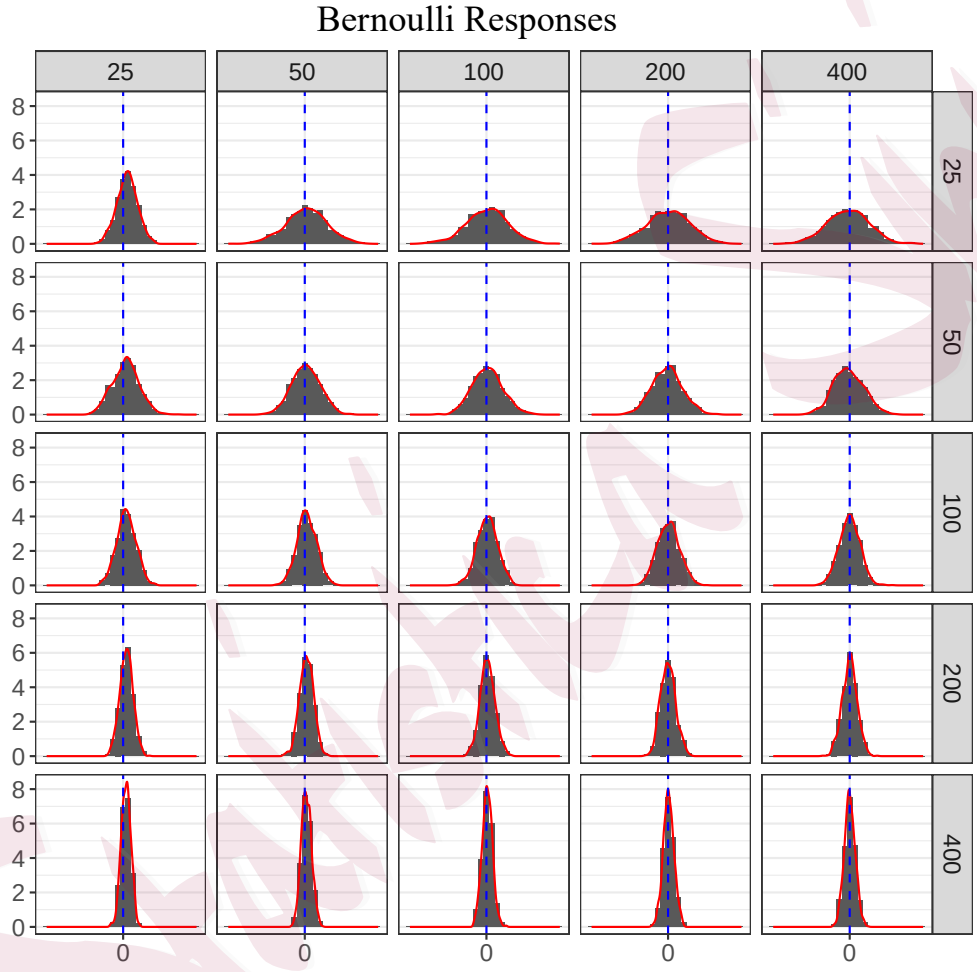
Statistica Sinica: Preprint
doi:10.5705/ss.202023.0343



Statistica Sinica: Preprint
doi:10.5705/ss.202023.0343



Statistica Sinica: Preprint
doi:10.5705/ss.202023.0343



Statistica Sinica: Preprint
doi:10.5705/ss.202023.0343

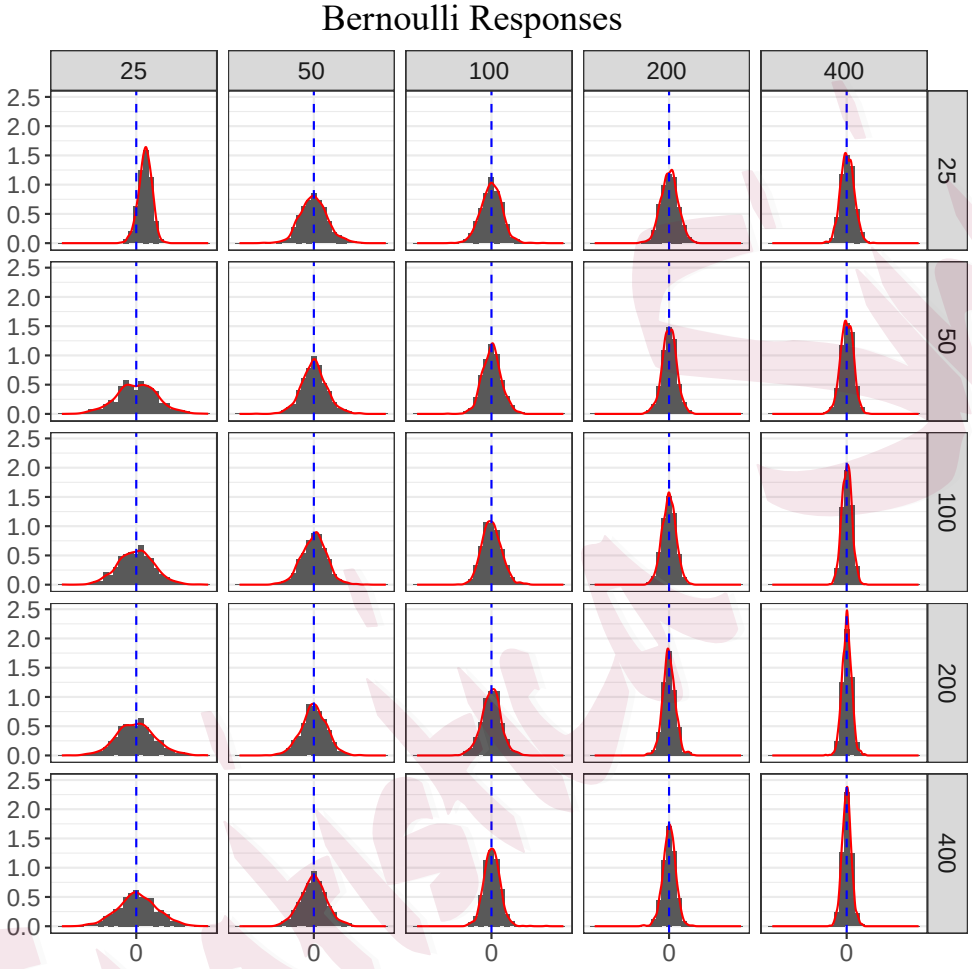


Figure 3: Histograms for the third component of $\hat{\boldsymbol{\beta}} - \dot{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{b}}_1 - \dot{\boldsymbol{b}}_1$, under the unconditional regime. Vertical facets represent the cluster sizes, while horizontal facets represent the number of clusters. The dotted line indicates zero, and the curve is a kernel density smoother.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0343

generally did not exhibit signs of non-normality, but the *difference* between the estimators and true random effects displayed evidence of non-normality except when n grew faster than m. This is also supported by the histograms in Figure 3 which show some evidence of higher kurtosis in the cases corresponding to small p-values in the Shapiro Wilk test. The histograms also suggest that both m and n need to grow for the estimators to be consistent for the true fixed and random effects, and in particular n needs to grow for the estimators to be unbiased. This is true especially for the Bernoulli responses, for which convergence was much slower and very large cluster sizes were needed for the estimators to be relatively unbiased.

In the supplementary material, we present additional results which showed that the sample covariance matrix of the estimated random effects became a better estimator of the true random effects covariance matrix $\dot{\boldsymbol{G}}$ as both m and n grew. Also, recall from our discussion in Section 2 that our asymptotic developments only require a working $\hat{\boldsymbol{G}}$, which need not be a consistent estimator of the true random effects covariance matrix. As a demonstration of this, we performed several additional simulations where in the PQL estimation procedure, we simply fix $\hat{\boldsymbol{G}}$ to a constant matrix and considered choices e.g., some constant multiplied by the identity matrix. Results in the supplementary material demonstrate that coverage probabil-

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0343

ities for our proposed intervals still tended to the nominal level as (m, n) increased, while corresponding Shapiro-Wilk tests and histograms were also consistent with our theory in large sample sizes and the empirical results presented above.

6. Discussion

In this article, we established new asymptotic distributional results for fixed effects, random effects, and the prediction gap, for an independent-cluster GLMM fitted using penalized quasi-likelihood estimation. Our results have important implications when it comes to inference and prediction for mixed-effects models. For the conditional regime, we establish asymptotic normality for any finite subset of the parameters. For random effects predictions and inference in the unconditional regime, we validate examining the empirical distribution of the estimated random effects as a diagnostic tool for assessing deviations away from the assumed random effects distribution (as is already commonly done in practice for GLMMs e.g., Hui et al., 2021). On the other hand, while the random effects estimators obtained using PQL are asymptotically normally distributed when the true random effects are normally distributed, we demonstrate that the difference between these two i.e., the prediction gap, need not be normally distributed. Our large sample

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0343

results thus suggest the use of a normal approximation when performing unconditional inference for the random effects, as is commonly done in practice (Bates et al., 2015; Brooks et al., 2017), can be potentially misleading.

An important avenue of future research is to establish rates of convergence, especially in the unconditional regime, when $\boldsymbol{x}_{ij}$ contains both $\boldsymbol{z}_{ij}$ plus additional components which are only included as purely fixed effects in the model. In the supplementary material, we develop some further results for such unpartnered fixed effects in the special cases of LMMs and GLMs. In both these cases, we see the convergence rate improves from $O_p(m^{1/2})$ to $O_p(N^{1/2})$, compared to the partnered fixed effects. On the other hand, for random effects without a partnered fixed effect, it is likely that the correct asymptotic distribution for the prediction gap will be the normal scale-mixture irrespective of the relative rates of m and n_i, as we saw in the motivating counterexample. Also, relaxing the canonical link assumption is an interesting and important extension to our work; we conjecture that non-canonical links could be accounted for by generalising the form of the GLM iterative weights, as is done in GLMs.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0343

Supplementary Material

The online Supplementary Material contains proofs of our theorems and extra simulation results.

Acknowledgments

Xu Ning was supported by the Australian Government Research Training Program Scholarship. Francis Hui and Alan Welsh were supported by an Australian Research Council Discovery Project DP230101908.

References

Basawa, I. V. and D. J. Scott (2012). *Asymptotic optimal inference for non-ergodic models*. Springer Science & Business Media.

Bates, D., M. Mächler, B. Bolker, and S. Walker (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software 67*(1), 1–48.

Blum, J., H. Chernoff, M. Rosenblatt, and H. Teicher (1958). Central limit theorems for interchangeable processes. *Canadian Journal of Mathematics 10*, 222–229.

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

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0343

REFERENCES

Breslow, N. E. and X. Lin (1995). Bias correction in generalised linear mixed models with a single component of dispersion. *Biometrika 82*, 81–91.

Brooks, M. E., K. Kristensen, K. J. van Benthem, A. Magnusson, C. W. Berg, A. Nielsen, H. J. Skaug, M. Maechler, and B. M. Bolker (2017). glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R Journal 9*, 378–400.

Cam, L. L. and G. L. Yang (1988). On the preservation of local asymptotic normality under information loss. *The Annals of Statistics 16*, 483–520.

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

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

Henderson, C. R. (1973). Sire evaluation and genetic trends. *Journal of Animal Science 1973*, 10–41.

Hui, F. K. C. (2020). On the use of a penalized quasilikelihood information criterion for generalized linear mixed models. *Biometrika 108*, 353–365.

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

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0343

REFERENCES

Hui, F. K. C., S. Müller, and A. H. Welsh (2021). Random effects misspecification can have severe consequences for random effects inference in linear mixed models. *International Statistical Review 89*, 186–206.

Jiang, J. (2003). Empirical best prediction for small-area inference based on generalized linear mixed models. *Journal of Statistical Planning and Inference 111* (1-2), 117–127.

Jiang, J., H. Jia, and H. Chen (2001). Maximum posterior estimation of random effects in generalized linear mixed models. *Statistica Sinica 11* (1), 97–120.

Jiang, J., M. P. Wand, and A. Bhaskaran (2022). Usable and precise asymptotics for generalized linear mixed model analysis and design. *Journal of the Royal Statistical Society: Series B 84*, 55–82.

Kackar, R. N. and D. A. Harville (1984). Approximations for standard errors of estimators of fixed and random effects in mixed linear models. *Journal of the American Statistical Association 79*, 853–862.

Kidziński, Ł., F. K. Hui, D. I. Warton, and T. J. Hastie (2022). Generalized matrix factorization: efficient algorithms for fitting generalized linear latent variable models to large data arrays. *The Journal of Machine Learning Research 23*, 13211–13239.

Lyu, Z. and A. H. Welsh (2021a). Asymptotics for EBLUPs: Nested error regression models. *Journal of the American Statistical Association 117*, 1–15.

Lyu, Z. and A. H. Welsh (2021b). Increasing cluster size asymptotics for nested error regression models. *Journal of Statistical Planning and Inference 217*, 52–68.

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0343

REFERENCES

McCulloch, C. E. and S. R. Searle (2004). *Generalized, linear, and mixed models*. John Wiley & Sons.

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.

Ogden, H. (2017). On asymptotic validity of naive inference with an approximate likelihood. *Biometrika 104*, 153–164.

Ogden, H. (2021). On the error in Laplace approximations of high-dimensional integrals. *Stat 10*, e380.

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

Pfeffermann, D. (2013). New important developments in small area estimation. *Statistical Science 28*, 40–68.

Prasad, N. N. and J. N. Rao (1990). The estimation of the mean squared error of small-area estimators. *Journal of the American statistical association 85*, 163–171.

van de Geer, S. and P. Müller (2012). Quasi-likelihood and/or robust estimation in high dimensions. *Statistical Science 27*, 469–480.

Vonesh, E. F., H. Wang, L. Nie, and D. Majumdar (2002). Conditional second-order generalized estimating equations for generalized linear and nonlinear mixed-effects models. *Journal of*

Statistica Sinica: Preprint
doi:10.5705/ss.202023.0343

the American Statistical Association 97, 271–283.

Research School of Finance, Actuarial Studies and Statistics, The Australian National University, Canberra, ACT 0200, Australia.

E-mail: xu.ning@anu.edu.au

Research School of Finance, Actuarial Studies and Statistics, The Australian National University, Canberra, ACT 0200, Australia.

E-mail: francis.hui@anu.edu.au

Research School of Finance, Actuarial Studies and Statistics, The Australian National University, Canberra, ACT 0200, Australia.

E-mail: alan.welsh@anu.edu.au