# Joint Selection in Mixed Models using Regularized PQL

Francis K.C. Hui*[1], Samuel Müller[2], and A.H. Welsh[1]

[1]Mathematical Sciences Institute, The Australian National University, Canberra, Australia

[2]School of Mathematics and Statistics, University of Sydney, Sydney, Australia

**Abstract**

The application of generalized linear mixed models present some major challenges for both estimation, due to the intractable marginal likelihood, and model selection, as we usually want to jointly select over both fixed and random effects. We propose to overcome these challenges by combining penalized quasi-likelihood (PQL) estimation with sparsity inducing penalties on the fixed and random coefficients. The resulting approach, referred to as regularized PQL, is a computationally efficient method for performing joint selection in mixed models. A key aspect of regularized PQL involves the use of a group based penalty for the random effects: sparsity is induced such

*Corresponding author: fhui28@gmail.com; Mathematical Sciences Institute, The Australian National University, 0200, Canberra, ACT, Australia

that all the coefficients for a random effect are shrunk to zero simultaneously, which in turns leads to the random effect being removed from the model. Despite being a quasi-likelihood approach, we show that regularized PQL is selection consistent, i.e. it asymptotically selects the true set of fixed and random effects, in the setting where the cluster size grows with the number of clusters. Furthermore, we propose an information criterion for choosing the single tuning parameter and show that it facilitates selection consistency. Simulations demonstrate regularized PQL outperforms several currently employed methods for joint selection even if the cluster size is small compared to the number of clusters, while also offering dramatic reductions in computation time.

**Keywords:** fixed effects, generalized linear mixed models, lasso, penalized likelihood, quasi-likelihood, variable selection

# 1 Introduction

Generalized linear mixed models (GLMMs) are a powerful class of models for analyzing correlated, non-normal data. Like all regression problems however, model selection is a difficult but critical part of inference. The problem is especially difficult for mixed models for two reasons: 1) fitting these models is computationally challenging, and 2) we often want to jointly select over both the fixed and random effects. Regarding the first problem, the marginal likelihood for a GLMM has no analytic form except with normal responses and the identity link, and so numerous estimation methods exist to overcome this difficulty. These range from approximation methods such as penalized quasi-likelihood (PQL, Breslow and Clayton, 1993), Laplace's method (Tierney and Kadane, 1986), and numerical quadrature (Rabe-Hesketh et al., 2002), to exact methods such as the Expectation-Maximization algorithm (EM algorithm, McCulloch, 1997). Of these approaches, PQL is the simplest to implement, as it effectively treats the random effects as "fixed" and estimates them in a similar manner to other fixed effects as in a generalized linear model (GLM). Furthermore, when the cluster size grows with the number of clusters, PQL estimates have been shown to be estimation consistent (Vonesh et al., 2002).

For jointly selecting fixed and random effects in GLMMs, proposed methods range from modifications of information criteria (e.g., Vaida and Blanchard, 2005) to more recent advances such as the fence (Jiang et al., 2008); see Müller et al. (2013) for an overview of model selection for LMMs specifically. These methods however are computationally burdensome to implement, especially since the number of candidate models in the GLMM context is considerably larger than the GLM context when performing joint selection. One appealing approach is to use penalized likelihood methods, although their application to mixed models has only recently been explored. For LMMs, Bondell et al. (2010) proposed adaptive lasso penalties for selecting the fixed and random effects, while Peng and Lu (2012) and Lin et al. (2013) proposed two-stage methods that separate out the fixed and random effect selection. For GLMMs, Ibrahim et al. (2011) proposed a modified version of the penalty in Bondell et al. (2010), and employed a Monte Carlo EM algorithm for estimation. This approach however is computationally intensive, with Ibrahim et al. (2011) limiting their simulations to LMMs only. Focusing solely on computational aspects, Schelldorfer et al. (2014) and Groll and Tutz (2014) proposed algorithms for fixed effects selection only using the lasso penalty in high-dimensional GLMMs, while Pan and Huang (2014) investigated random effects selection only. The large sample properties of these algorithms however remain to be determined.

In this article, we propose a new approach to joint selection in GLMMs using regularized PQL estimation, and a method to choose the associated tuning parameter. Rather than working with the marginal likelihood, we propose combining the PQL with adaptive lasso and adaptive group lasso penalties to select the fixed and random effect coefficients respectively. The group lasso is used to exploit the grouped structure inherent in the random effects: for any random intercept or slope, the coefficients across all clusters are shrunk to zero at the same time, which leads to the corresponding row and column of the random effect covariance matrix being shrunk to zero. Such a group penalty approach to random effects selection has been used previously in linear mixed models by Fan and Li (2012), but this article is the first to apply it to GLMMs by regularizing the PQL. Another difference between this article and Fan and Li (2012) is that the latter separate the fixed and random effects selection into two stages, with different likelihoods and tuning parameters at each stage, whereas we perform fixed and random effects selection simultaneously using a single tuning parameter. Compared to the Monte Carlo EM method of Ibrahim et al. (2011), joint selection using regularized PQL is extremely fast: it can be viewed as a specific type of penalized GLM, and the full regularization path can be constructed without the need for integration.

In the setting where the cluster size grows at a slower rate than the number of clusters, we show that the regularized PQL estimates are estimation and selection consistent. This is an important advance on Vonesh et al. (2002). For the critical choice of the tuning parameter, we propose a new information criterion which we show leads to selection consistency. This information criterion combines a BIC-type penalty for the fixed effects with an AIC-type penalty for the random effects. Over the past decade, numerous BIC-type criteria have been proposed for choosing the tuning parameter in penalized GLMs, particularly in the high-dimensional setting, with results establishing their selection consistency (e.g., Zhang et al., 2010; Hui et al., 2015). Analogous results however do not exist for mixed models, with the exception of Ibrahim et al. (2011) whose proposed approach involves at least two tuning parameters. A key contribution of this article is showing that in the case of regularized PQL, differential penalization of the fixed and random effects is needed to achieve selection consistency.

For many applications where the cluster size is small, we propose a hybrid estimator to improve finite sample performance, i.e. regularized PQL is used for model selection only, and the final submodel is estimated using maximum likelihood. Simulations demonstrate that regularized PQL, in conjunction with the proposed information criterion, outperforms several currently available methods for joint selection in GLMMs, while offering dramatic reductions in computation time. We illustrate the application of regularized PQL estimation on a longitudinal dataset for determining the predictors of forest health over time.

To summarize, the main contributions of this article are as follows: 1) we propose a computationally efficient method of performing joint selection in GLMMs, which combines the PQL with adaptive (group) lasso penalties to regularize the fixed and random effect coefficients; 2) we develop an information criterion for choosing the tuning parameter in regularized PQL estimation, that involves differing model complexity terms on the fixed and random effects; 3) we demonstrate estimation and selection consistency properties for regularized PQL estimation, and show that the proposed information criterion asymptotically chooses a tuning parameter that leads to selection consistency; 4) we perform simulations to demonstrate the computational speed and strong performance of regularized PQL, relative to other penalized likelihood methods, even when the cluster size is relatively small compared to the number of clusters.

# 2 Generalized Linear Mixed Models

We focus on the independent cluster model with random intercepts and slopes. Let $y_{ij}$ denote the $j^{\text{th}}$ measurement for the $i^{\text{th}}$ cluster, where $i = 1, \ldots, n$ and $j = 1, \ldots, m_i$.

Note that we allow for unequal cluster sizes. Let $\boldsymbol{x}_{ij}$ be a vector of $p_f$ covariates corresponding to fixed effects, and $\boldsymbol{z}_{ij}$ be a vector of $p_r$ covariates corresponding to random effects. Both $\boldsymbol{x}_{ij}$ and $\boldsymbol{z}_{ij}$ may contain an intercept term as their first element. We assume that $p_f$ and $p_r$ are fixed, with $p_r < \min_i(m_i)$ where $\min_i(\cdot)$ denotes the minimum over $i = 1, \ldots, n$. Conditional on the random effects $\boldsymbol{b}_i$, the responses $y_{ij}$ are assumed to come from a distribution in the exponential family, with density function $f(y_{ij}|\boldsymbol{\beta}, \boldsymbol{b}_i, \phi) = \exp[\{y_{ij}\vartheta_{ij} - a(\vartheta_{ij})\}/\phi + c(y_{ij}, \phi)]$ for known functions $a(\cdot)$ and $c(\cdot)$ and dispersion parameter $\phi$. The mean, $\mu_{ij}$, is modeled as $g(\mu_{ij}) = \eta_{ij} = \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \boldsymbol{z}_{ij}^T\boldsymbol{b}_i$, for a known link function $g(\cdot)$. For simplicity, we assume the canonical link is used, so $g(\mu_{ij}) = \vartheta_{ij} = \eta_{ij}$ and $\mu_{ij} = a'(\eta_{ij})$. The random effects are normally distributed, $\boldsymbol{b}_i \sim \mathcal{N}_{p_r}(\boldsymbol{0}, \boldsymbol{D})$, where $\boldsymbol{D}$ is the random effect covariance matrix.

For the $i^{\text{th}}$ cluster, we have an $m_i$-vector $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{im_i})$, a $m_i \times p_f$ matrix $\boldsymbol{X}_i = (\boldsymbol{x}_{i1} \ldots \boldsymbol{x}_{im_i})^T$ of fixed effect covariates, and a $m_i \times p_r$ matrix $\boldsymbol{Z}_i = (\boldsymbol{z}_{i1} \ldots \boldsymbol{z}_{im_i})^T$ of random effect covariates. In turn, we can write $g(\boldsymbol{\mu}_i) = \boldsymbol{\eta}_i = \boldsymbol{X}_i^T\boldsymbol{\beta} + \boldsymbol{Z}_i^T\boldsymbol{b}_i$, where $g(\cdot)$ is applied component-wise, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{p_f})$, $\boldsymbol{b}_i = (b_{i1}, \ldots, b_{ip_r})$, $\boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{im})$ and similarly for $\boldsymbol{\eta}_i$. Finally, let $\boldsymbol{b} = (\boldsymbol{b}_1^T, \ldots, \boldsymbol{b}_n^T)^T$ be the $np_r$-vector of all random effects, and $\boldsymbol{\Psi} = \{\boldsymbol{\beta}^T, \text{vech}(\boldsymbol{D})^T\}^T$ where vech$(\cdot)$ denotes the half-vectorization operator. Note that each $\boldsymbol{b}_i$ is of fixed dimension $p_r$, while $\dim(\boldsymbol{b})$ grows with linearly with $n$.

For the GLMM above, the marginal log-likelihood is

$$\ell(\boldsymbol{\Psi}) = -\frac{n}{2}\log\det(\boldsymbol{D}) + \sum_{i=1}^{n}\log\left(\int \exp\left(\sum_{j=1}^{m_i}\log f(y_{ij}|\boldsymbol{\beta}, \boldsymbol{b}_i) - \frac{1}{2}\boldsymbol{b}_i^T\boldsymbol{D}^{-}\boldsymbol{b}_i\right) d\boldsymbol{b}_i\right),$$

where $\det(\boldsymbol{D})$ is the determinant of $\boldsymbol{D}$. Aside from linear mixed models, the integral in the marginal log-likelihood does not have an analytic form, and this complicates maximum likelihood estimation. A popular, alternative estimation method is PQL estimation, which involves maximizing the quasi-likelihood function

$$\ell_{\text{PQL}}(\boldsymbol{\Psi}, \boldsymbol{b}) = \sum_{i=1}^{n} \sum_{j=1}^{m_i} \log f(y_{ij}|\boldsymbol{\beta}, \boldsymbol{b}_i) - \frac{1}{2} \sum_{i=1}^{n} \boldsymbol{b}_i^T \boldsymbol{D}^{-} \boldsymbol{b}_i, \tag{1}$$

where $\boldsymbol{D}^{-}$ denotes the Moore-Penrose generalized inverse of $\boldsymbol{D}$. The use of a generalized inverse here, as opposed to the standard matrix inverse, allows us to deal with cases where the covariance matrix is singular (see Breslow and Clayton, 1993). This is necessary when we establish asymptotic properties in Section 4, where the true random effects are assumed to be sparse.

There is a close link between PQL estimation and Laplace's method for GLMMs. Specifically, for a fixed $\boldsymbol{\Psi}$, let $\tilde{\boldsymbol{b}} = (\tilde{\boldsymbol{b}}_1^T, \ldots, \tilde{\boldsymbol{b}}_n^T)^T$ denote the maximizer of (1). Then the Laplace approximated log-likelihood is defined as

$$\ell_{\text{LA}}(\boldsymbol{\Psi}) = \sum_{i=1}^{n} \sum_{j=1}^{m_i} \log f(y_{ij}|\boldsymbol{\beta}, \tilde{\boldsymbol{b}}_i) - \frac{1}{2} \sum_{i=1}^{n} \tilde{\boldsymbol{b}}_i^T \boldsymbol{D}^{-} \tilde{\boldsymbol{b}}_i - \frac{1}{2} \sum_{i=1}^{n} \log \det(\boldsymbol{Z}_i^T \tilde{\boldsymbol{W}}_i \boldsymbol{Z}_i \boldsymbol{D} + \boldsymbol{I}_{p_r}),$$

where $\tilde{\boldsymbol{W}}_i$ is a $m_i \times m_i$ diagonal weight matrix with elements $\tilde{\boldsymbol{W}}_{i,jj} = a''(\tilde{\eta}_{ij})/\phi$, $\tilde{\eta}_{ij} = \boldsymbol{x}_{ij}^T \boldsymbol{\beta} + \boldsymbol{z}_{ij}^T \tilde{\boldsymbol{b}}_i$, and $\boldsymbol{I}_{p_r}$ is an identity matrix of dimension $p_r$. The key difference between PQL and the Laplace approximation lies in the last term, which is a non-linear function of $\boldsymbol{\beta}$ and $\boldsymbol{b}$. By assuming the weights in $\boldsymbol{W}_i$ vary slowly with the mean, Breslow and Clayton (1993) proposed ignoring this last term, from which the PQL follows. Note that when the minimum cluster size $\min_i(m_i)$, and hence all $m_i$, are large, the estimates from PQL and Laplace's method should be close to each other, since the last term in $\ell_{\text{LA}}(\boldsymbol{\Psi})$ is of a smaller order than the first term (see also Demidenko, 2013, Section 7.3). For normal responses, $a''(\eta_{ij}) = 1$, so the estimates of $\boldsymbol{\beta}$ based on $\ell_{\text{LA}}(\boldsymbol{\Psi}, \boldsymbol{b})$ and $\ell_{\text{PQL}}(\boldsymbol{\Psi}, \boldsymbol{b})$ coincide, noting that the Laplace approximation is exact for normal linear mixed models.

Compared to maximizing the marginal and Laplace approximated log-likelihoods, PQL estimation is straightforward: equation (1) resembles the log-likelihood for a GLM combined with a generalized ridge penalty, where $\boldsymbol{b}$ is also treated as a fixed effect vector, and so modifications of standard optimization routines such as iteratively reweighted least squares can be used for maximization. This in turn motivates us to consider using the PQL as a loss function for penalized joint selection in GLMMs.

# 3 Regularized PQL Estimation

We propose regularized PQL estimation to perform selection over both the fixed and random effects in GLMMs.

**Definition.** *For a given $\boldsymbol{D}$, the regularized PQL estimates of the fixed and random effect coefficients are given by*

$$(\hat{\boldsymbol{\beta}}_\lambda, \hat{\boldsymbol{b}}_\lambda) = \arg\max_{\boldsymbol{\beta},\boldsymbol{b}} \ell_p(\boldsymbol{\Psi}, \boldsymbol{b}) = \arg\max_{\boldsymbol{\beta},\boldsymbol{b}} \ell_{PQL}(\boldsymbol{\Psi}, \boldsymbol{b}) - \lambda \sum_{k=1}^{p_f} v_k |\beta_k| - \lambda \sum_{l=1}^{p_r} w_l \|\boldsymbol{b}_{\bullet l}\|,$$

*where $v_k$ and $w_k$ are adaptive weights based on preliminary estimates of $\beta_k$ and $\boldsymbol{D}$ respectively, $\boldsymbol{b}_{\bullet l} = (b_{il}, \ldots, b_{nl})$ denotes all the coefficients corresponding to the $l^{th}$ random effect, and $\| \cdot \|$ denotes its $L_2$ norm.*

We use an adaptive lasso penalty with weights $v_k$ for the fixed effects, and an adaptive group lasso penalty with weights $w_l$ for the random effects, linked by one tuning parameter $\lambda > 0$. Specifically, let $\tilde{\beta}$ and $\tilde{\boldsymbol{D}}$ denote the unpenalized, maximum likelihood estimates of the fixed effect coefficients and random effect covariance matrix respectively from fitting the full GLMM. This fitting could be performed, for example, by applying the EM algorithm, or via recent advances in maximum likelihood estimation for GLMMs such as data cloning (Lele et al., 2010). Then we choose $v_k = |\tilde{\beta}_k|^{-\kappa}$ and $w_l = \tilde{\boldsymbol{D}}_{ll}^{-\kappa}$, where $\tilde{\boldsymbol{D}}_{ll}$ is the $l^{\text{th}}$ diagonal element of $\tilde{\boldsymbol{D}}$ and $\kappa > 0$ is a common power parameter. Note that while the penalty involves $\boldsymbol{b}$, the adaptive weights for the random effects require only an initial estimate of $\boldsymbol{D}$. Also, in the case where the fixed intercept term is included but not penalized, the adaptive lasso penalty is summed from $k = 2$ to $p_f$.

The adaptive weights mean that a single tuning parameter, as opposed to using different $\lambda$'s for the fixed and random effects, is able to achieve consistency of the regularized PQL estimates. In Section 3.2, we discuss how to select the tuning parameter. Of course, having to select over multiple $\lambda$'s also presents a considerable computational challenge (see for instance, Garcia et al., 2014). Note that due to the concavity of both $\ell_{\text{PQL}}(\boldsymbol{\Psi}, \boldsymbol{b})$ and the lasso penalties, if there exists a maximizer to $\ell_p(\boldsymbol{\Psi}, \boldsymbol{b})$ then it is also the unique, regularized PQL estimate (see also Lemma 2.1, Jiang et al., 2001).

Regularized PQL performs joint selection of the fixed and random effects in mixed models. The adaptive group lasso penalizes random slopes across clusters, thereby utilizing the grouped structure inherent in the random effects. For a sufficiently large value of $\lambda$, maximizing the regularized PQL shrinks $\|\boldsymbol{b}_{\bullet l}\| = 0$, that is, all the coefficients corresponding to the $l^{\text{th}}$ random slope (or the random intercept) are shrunk to zero. This implies that the $l^{\text{th}}$ row and column of $\boldsymbol{D}$ are also set to zero (see Section 3.1). This method of penalizing the coefficients $\boldsymbol{b}$ explicitly differs from the random effects penalties that shrink one or more elements of $\boldsymbol{D}$ or a decomposition of $\boldsymbol{D}$ to zero (Bondell et al., 2010; Ibrahim et al., 2011). In fact, the potential to penalize $\boldsymbol{b}$ arises precisely because the PQL is a function of the $\boldsymbol{b}$'s.

Since $\ell_p(\boldsymbol{\Psi}, \boldsymbol{b})$ does not require integrating over the random effects, the solution path for the regularized PQL estimates is easily constructed. Conditional on $\boldsymbol{D}$ and $\boldsymbol{b}$, estimates of the fixed effects $\boldsymbol{\beta}$ are obtained by fitting a GLM with the adaptive lasso penalty across all clusters, with $\boldsymbol{z}_{ij}^T \boldsymbol{b}_i$ as an offset. Then conditional on $\boldsymbol{D}$ and $\boldsymbol{\beta}$, estimates of the random effects $\boldsymbol{b}$ are obtained by fitting a GLM with an adaptive elastic net penalty, with $\boldsymbol{x}_{ij}^T \boldsymbol{\beta}_i$ as an offset. In the simulations and application, we used a local quadratic approximation (Fan and Li, 2001) to calculate the regularized PQL estimates, and this was already considerably faster than methods involving the marginal likelihood. Utilizing more sophisticated methods for estimation (e.g., coordinate descent, Friedman et al., 2010) will further reduce computation time.

## 3.1 Estimation of the Covariance Matrix

For a given $\boldsymbol{D}$, regularized PQL provides estimates of the fixed and random effect coefficients $(\hat{\boldsymbol{\beta}}_\lambda^T, \hat{\boldsymbol{b}}_\lambda^T)^T$. With these estimates, we can update the random effect covariance matrix in a number of ways (e.g., Breslow and Clayton, 1993; Vonesh et al., 2002). We propose substituting $(\hat{\boldsymbol{\beta}}_\lambda^T, \hat{\boldsymbol{b}}_\lambda^T)^T$ back into $\ell_{\text{LA}}(\boldsymbol{\Psi})$, and then maximizing to obtain an estimate of $\boldsymbol{D}$. Straightforward algebra (see Appendix A) shows that an estimate of the covariance matrix can be obtained via the following iterative equation: At the $t^{\text{th}}$ iteration,

$$\hat{\boldsymbol{D}}_\lambda^{(t)} = \frac{1}{n} \sum_{i=1}^{n} \left\{ \left( \boldsymbol{Z}_i^T \hat{\boldsymbol{W}}_{\lambda i} \boldsymbol{Z}_i + (\hat{\boldsymbol{D}}_\lambda^{(t-1)})^- \right)^{-1} + \hat{\boldsymbol{b}}_{\lambda i} \hat{\boldsymbol{b}}_{\lambda i}^T \right\}, \tag{2}$$

where $\hat{\boldsymbol{b}}_\lambda^T = (\hat{\boldsymbol{b}}_{\lambda 1}^T, \ldots, \hat{\boldsymbol{b}}_{\lambda n}^T)^T$ and $\hat{\boldsymbol{W}}_{\lambda i}$ is the weight matrix for subject $i$ evaluated at $(\hat{\boldsymbol{\beta}}_\lambda^T, \hat{\boldsymbol{b}}_{\lambda i}^T)^T$. Note that when $\|\boldsymbol{b}_{\bullet l}\|$ is shrunk to zero, it makes sense to set the $l^{\text{th}}$ row and column of $\boldsymbol{D}$ to zero, reflecting the removal of this covariate from the random effects component of the model. In such a case, the iterative formula above is applied only to the submatrix of $\boldsymbol{D}$ with non-zero rows and columns. Finally, we point out that this update of the covariance matrix is only used in the context of regularized PQL estimation; as we discuss in Section 3.3, we propose using a hybrid estimator to calculate the final parameter estimates.

## 3.2 Tuning Parameter Selection

As with all penalized likelihood methods, both the finite sample and asymptotic performance of regularized PQL depend critically on being able to choose an appropriate value of the tuning parameter. For the GLM framework, there has been considerable research into choosing $\lambda$ using, most commonly, cross validation or information criteria (e.g., Zhang et al., 2010), and we focus on the latter method. Specifically, we consider tuning parameters within the range $[\lambda_{\min}, \lambda_{\max}]$, where $\lambda_{\min}$ leads to the full model containing all the candidate fixed and random effects, and $\lambda_{\max}$ is the smallest $\lambda$ that leads to the null model. A solution path is constructed by considering a sequence of $\lambda$'s over this range, and selecting the value of $\lambda$ (hence the best submodel) by minimizing the information criterion

$$\text{IC}(\lambda) = -\frac{2}{N}\ell_{\text{PQL}}(\hat{\boldsymbol{\Psi}}_\lambda, \hat{\boldsymbol{b}}_\lambda) + \frac{\log(n)}{N}\dim(\hat{\boldsymbol{\beta}}_\lambda) + \frac{2}{N}\dim(\hat{\boldsymbol{b}}_\lambda), \tag{3}$$

where $\dim(\hat{\boldsymbol{\beta}}_\lambda)$ and $\dim(\hat{\boldsymbol{b}}_\lambda)$ are the number of non-zero estimated fixed and random effect coefficients respectively and, importantly, $\dim(\hat{\boldsymbol{b}}_\lambda) = n\dim(\hat{\boldsymbol{b}}_{\lambda 1})$. Note that division by total sample size $N$ is often used when studying information criteria for tuning parameter selection (e.g., Zhang et al., 2010).

A key feature of $\text{IC}(\lambda)$, which sets it apart from standard information criteria used for tuning parameter selection in other penalties for mixed models (e.g., Ibrahim et al., 2011; Lin et al., 2013), is its use of different model complexity penalties. Specifically, a BIC-type penalty of '$\log(n)$' is used for the fixed effects, and an AIC-type penalty of '2' is used for the random effects. The latter arises because the model complexity is already taken into account by $\dim(\hat{\boldsymbol{b}}_\lambda)$, which grows linearly with $n$ regardless of the number of random effects in the model. Put another way, overfitting of the $\hat{\boldsymbol{b}}_\lambda$'s is inherently prevented by the information criterion, since the removal of one random effect from the model amounts to the removal of $n$ coefficients in regularized PQL by the group sparsity of $\|\hat{\boldsymbol{b}}_{\lambda \bullet l}\|$. By contrast, $\dim(\hat{\boldsymbol{\beta}}_\lambda)$ is always of order $O_p(1)$, and so the BIC-type penalty of $\log(n)$ is necessary to properly account for model complexity in the fixed effects and prevent overfitting (see Shao, 1997, for related work on the use of differing model complexities in the linear regression context). In Section 4.1, we show that using IC$(\lambda)$ to choose the tuning parameter selection leads to selection consistency in regularized PQL.

## 3.3 Hybrid Estimation Approach

In real finite sample settings, regularized PQL can produce biased estimates of the fixed effects and the random effect covariance matrix. The bias is related to the well known finite sample bias for unpenalized PQL estimation when the cluster sizes are not large compared to the number of clusters (Lin and Breslow, 1996). Moreover, as shown in Theorem 1, we establish consistency of the regularized PQL estimates where the convergence rate depends on the rate of growth of the cluster sizes $m_i$. Thus compared to the unpenalized maximum likelihood estimates, which are $n^{1/2}$-consistent, the regularized PQL estimates are not as efficient if all the $m_i$'s are smaller than $n$, which is typically the case with longitudinal studies.

To improve finite sample performance, we propose a hybrid estimation approach in which we use regularized PQL *only* for joint selection of the fixed and random effects, and use maximum likelihood estimation of the selected submodel to obtain the final estimates $\boldsymbol{\beta}$ and vech$(\boldsymbol{D})$, as well as to construct predictions of the random effects based on posterior modes (for instance). Hybrid estimation approaches have been used previously (e.g., Hui et al., 2015), although the purpose there was to reduce the bias introduced by penalization, while we use the hybrid approach to address both the relative lack of efficiency and finite sample bias of the regularized PQL estimates. Of course, since the hybrid approach is applied on the submodel chosen by regularized PQL estimation, it also inherits the selection consistency property encapsulated in the second part of Theorem 1. In the simulations in Section 5, we empirically evaluate the performance of the hybrid estimation approach compared to just using the estimates from regularized PQL.

# 4 Asymptotic Properties

We study the large sample properties of regularized PQL estimation in the setting where the cluster sizes grow with the number of clusters. Without loss of generality, suppose that the clusters are labeled such that the first cluster grows at the slowest rate, and the last cluster grows at the largest rate. That is, $m_1 = O(m_k)$ for all $k = 2, \ldots, n$, and $m_l = O(m_n)$ for all $l = 1, \ldots, n_1$, so the rates of growth of the cluster sizes are bounded below by the order of $m_1$ and above by the order of $m_n$. Note this includes the case where all cluster sizes are constrained to grow at the same rate. It is also worth pointing out that no restriction is made directly on whether the cluster sizes are balanced or not; Instead, the assumptions made concern the rate of growth of the cluster sizes. We assume that $m_n/n \to 0$, such that all cluster sizes grow at a smaller rate than number of clusters. This setting arises commonly in longitudinal studies in epidemiology (for instance), where the number of measurements recorded for each cluster increases slowly as more subjects are recruited into the study.

To aid our theoretical development, write the random effect covariance matrix as $\boldsymbol{D} = \boldsymbol{\Gamma}\boldsymbol{\Gamma}^T$, where $\boldsymbol{\Gamma} = \boldsymbol{Q}\boldsymbol{\Lambda}^{1/2}$ with $\boldsymbol{Q}$ the orthogonal matrix of normalized eigenvectors and $\boldsymbol{\Lambda}$ the diagonal matrix whose entries are the eigenvalues of $\boldsymbol{D}$. Note if the $l^{\text{th}}$ row of $\boldsymbol{\Gamma}$ is equal to zero, then it implies that both the $l^{\text{th}}$ row and column of $\boldsymbol{D}$ are zero. Consequently, for the remainder of this section, we redefine the parameter vector as $\mathbf{\Psi} = \{\boldsymbol{\beta}^T, \text{vec}(\mathbf{\Gamma})^T\}^T \in \Re^{p_f+p_r^2}$, replacing $\text{vech}(\boldsymbol{D})$ by $\text{vec}(\mathbf{\Gamma})$. This parameterization is used only in the theoretical development, as it avoids the true parameter point being on the boundary of the parameter space (see Condition C4 below) and is not employed in the estimation process.

Let $\mathbf{\Psi}_0 = \{\boldsymbol{\beta}_0^T, \text{vec}(\mathbf{\Gamma}_0^T)\}^T$ be the true parameter point and, without loss of generality, write $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{01}^T, \boldsymbol{\beta}_{02}^T = \mathbf{0}^T)^T$ and $\text{vec}(\mathbf{\Gamma}_0) = (\text{vec}(\mathbf{\Gamma}_{01}^T), \text{vec}(\mathbf{\Gamma}_{02}^T) = \mathbf{0}^T)^T$. Let $p_{0f} = \dim(\boldsymbol{\beta}_{01})$ denote the number of truly non-zero fixed effects, and $p_{0r}$ the number of rows in $\mathbf{\Gamma}_{01}$. Also, for $i = 1, \ldots, n$, let $\boldsymbol{b}_{0i}$ denote a realization from the true random effects distribution; the first $p_{0r}$ elements of $\boldsymbol{b}_{0i}$ are drawn from a multivariate normal distribution with mean zero and covariance matrix $\boldsymbol{D}_{01} = \mathbf{\Gamma}_{01}\mathbf{\Gamma}_{01}^T$, and $b_{0il} = 0$ for $l = p_{0r}+1, \ldots, p_r$. Finally, let $N = \sum_{i=1}^{n} m_i$ be the total number of observations. The following regularity conditions are required.

(C1) The function $a(\eta)$ is three times continuously differentiable in its domain, with $a''(\eta) \geq c_0 > 0$ for some sufficiently small constant $c_0$.

(C2) For every $i = 1, \ldots, n$ and $j = 1, \ldots, m_i$, there exists a sufficiently large constant $C$ such that $\|\boldsymbol{x}_{ij}\|_\infty < C$ and $\|\boldsymbol{z}_{ij}\|_\infty < C$ where $\|\cdot\|_\infty$ is the maximum norm. Furthermore, the matrices $m_i^{-1}\boldsymbol{X}_i^T\boldsymbol{X}_i$ and $m_i^{-1}\boldsymbol{Z}_i^T\boldsymbol{Z}_i$ are positive definite with minimum and maximum eigenvalues bounded from above and below by $1/c_1$ and $c_1$ respectively, where $c_1$ is some positive constant.

(C3) Let $\ell_1(\boldsymbol{\beta}, \boldsymbol{b}) = \sum_{i=1}^{n}\sum_{j=1}^{m_i} \log f(y_{ij}|\boldsymbol{\beta}, \boldsymbol{b}_i)$, and $\boldsymbol{H}(\boldsymbol{\beta}, \boldsymbol{b}) = -N^{-1}\nabla^2\ell_1(\boldsymbol{\beta}, \boldsymbol{b})$. Then there exists a $\varepsilon > 0$ such that for $n$ and all $m_i$ sufficiently large, the minimum eigenvalue of $\boldsymbol{H}(\boldsymbol{\beta}, \boldsymbol{b})$ is bounded away from zero for all $\|(\boldsymbol{\beta}^T, \boldsymbol{b}^T)^T - (\boldsymbol{\beta}_0^T, \boldsymbol{b}_0^T)^T\|_\infty \leq \varepsilon$.

(C4) $\mathbf{\Psi}_0$ is a interior point in the compact set $\mathbf{\Omega} \in \Re^{p_f+p_r^2}$.

(C5) The tuning parameter $\lambda$ satisfies (a) $\lambda m_1^{1/2}/N \to 0$ and (b) $\lambda m_1^{1/2} n^{\kappa/2}/N \to \infty$, where $m_n/n \to 0$.

Conditions (C1) and (C2) ensure the observed information matrices based on $\ell_{\text{PQL}}(\boldsymbol{\Psi}, \boldsymbol{b})$ are positive definite, and imply that the expectations $\text{E}_{\boldsymbol{b}}\{a''(\eta)\}$ and $\text{E}_{\boldsymbol{b}}\{a'''(\eta)\}$, where the expectations are with respect to the true random effects distribution, are finite. Condition (C3) extends this to a small neighborhood around the true parameters. Along with the independence of the responses $\boldsymbol{y}_i$ for each cluster, conditions (C1)-(C4) are sufficient to ensure the maximum likelihood estimate of $\boldsymbol{\Psi}$, i.e. the maximizer $\ell(\boldsymbol{\Psi})$, exists and is $n^{1/2}$-consistent (Lehmann, 1983). Condition (C5) imposes restriction on the rate at which the tuning parameter can grow.

We now present a result on the large sample consistency of the regularized PQL estimates.

**Theorem 1.** *Under conditions (C1)-(C5a), as $n, m_i \to \infty$ for all $i$ and $m_n/n \to 0$, the regularized PQL estimator satisfies $\|\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}_0\| = O_p(m_1^{-1/2})$ and $\|\hat{\boldsymbol{b}}_{\lambda i} - \boldsymbol{b}_{0i}\| = O_p(m_1^{-1/2})$ for all $i = 1, \ldots, n$. If condition (C5b) is also satisfied, then $P(\hat{\boldsymbol{\beta}}_{\lambda 02} = \mathbf{0}) \to 1$ and $P(\|\hat{\boldsymbol{b}}_{\lambda \bullet l}\| = 0) \to 1$ for all $l = p_{0r}+1, \ldots, p_r$, where $\hat{\boldsymbol{b}}_{\lambda \bullet l} = (\hat{b}_{\lambda 1 l}, \ldots, \hat{b}_{\lambda n l})$ denotes all the estimated coefficients corresponding to the $l^{th}$ random effect.*

Note that even though $p_r$ is fixed, each $\hat{\boldsymbol{b}}_{\lambda \bullet l}$ is growing at the same rate as the number of clusters, $n$, so the proof of Theorem 1 has to be developed in a high-dimensional setting. Outlines of the proofs of all theorems are given in Appendix B, with detailed proofs provided in the Supplementary Material.

The $m_1^{1/2}$-consistency of the fixed effects agrees with the result of Vonesh et al. (2002), who showed $\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}_0 = O_p\left(\max\{m^{-1/2}, n^{-1/2}\}\right)$ in the case where all cluster sizes were equal to $m$, and $n \to \infty$ and $m \to \infty$. The $m_1^{1/2}$-consistency for each $\hat{\boldsymbol{b}}_{\lambda i}$ is also reason-able, since regularized PQL treats the $\boldsymbol{b}$ as fixed effects and the estimation of $\boldsymbol{b}_{\lambda i}$ depends only on the $m_i$ observations within the $i^{\text{th}}$ cluster. Estimation consistency for all random effect coefficients is thus governed by the smallest rate of growth of the cluster sizes, $m_1$. The second part of Theorem 1 states that the regularized PQL estimators asymptotically select only the truly non-zero fixed and random effects in the GLMM. Together with its computational simplicity, this presents a strong argument for the use of regularized PQL for joint selection.

## 4.1 Consistency of IC($\lambda$)

In this section, we show that using the tuning parameter chosen by minimizing IC($\lambda$) asymptotically identifies the true model. For any value of $\lambda \in [\lambda_{\min}, \lambda_{\max}]$, let $\alpha$ denote the submodel (subset of fixed and random effects) selected by regularized PQL estimation. Clearly $\alpha$ depends on $\lambda$, but for ease of notation we have suppressed this dependence. Next, let $(\tilde{\boldsymbol{\Psi}}_\alpha^T, \tilde{\boldsymbol{b}}_\alpha^T)^T$ denote the unregularized PQL estimator for this submodel, obtained by maximizing the PQL in (1) along with the iterative update of the covariance matrix in (2). Finally, let $\lambda_0$ be a sequence of tuning parameters that satisfy condition (C5) and hence selects the true model, which we denote here as $\alpha_0$.

For the development below, we require an additional condition. Let '$\supset$' denote the proper superset relation.

(C6) There exists a constant $c_2$ such that $E\left\{\ell_o(\boldsymbol{\Psi}_0) - \ell_o(\boldsymbol{\Psi}_\alpha^*)\right\} \geq c_2 > 0$ for all models $\alpha \not\supset \alpha_0$, where $\ell_o(\boldsymbol{\Psi})$ denotes the marginal log-likelihood of a GLMM for a single observation, and $\boldsymbol{\Psi}_\alpha^*$ denotes the pseudo-true parameters for model $\alpha$ which minimize $E\left\{-\ell_o(\boldsymbol{\Psi}_\alpha^*)\right\}$.

Conditions like (C6) are imposed in theoretical developments on selection consistency e.g., see condition (viii) in Müller and Welsh (2009) for robust selection on GLMs, and condition (C4) in Zhang et al. (2010) in the setting of penalized GLMs. It amounts to requiring that the Kullback-Leibler distance between any underfitted GLMM, with pseudo-true parameters $\mathbf{\Psi}^*_\alpha$, and the true GLMM is positive; see the Supplementary Material and White (1982) for further discussion of pseudo-true parameters.

We now define a proxy information criterion based on these unregularized PQL estimates,

$$\text{IC}_{\text{proxy}}(\alpha) = -\frac{2}{N}\sum_{i=1}^{n}\sum_{j=1}^{m_i}\log f(y_{ij}|\tilde{\boldsymbol{\beta}}_\alpha, \tilde{\boldsymbol{b}}_{\alpha i}) + \frac{\log(n)}{N}\dim(\tilde{\boldsymbol{\beta}}_\alpha) + \frac{2}{N}\dim(\tilde{\boldsymbol{b}}_\alpha).$$

Note the loss function for this proxy criterion involves only the first part of the PQL. The reason for introducing this proxy criterion is to simplify the theoretical development: since $\text{IC}_{\text{proxy}}(\alpha)$ does not involve penalized estimates, we can focus on establishing its asymptotic behavior when $\alpha$ represents an underfitted or overfitted model without having to deal with the effects of $\lambda$. We then have the following result.

**Lemma 1.** *Under conditions (C1)-(C4) and (C6), and as $n, m_i \to \infty$ for all $i$ and $m_n/n \to 0$, the proxy information criterion satisfies* $P\left\{\min_{\alpha\neq\alpha_0} IC_{proxy}(\alpha) > IC_{proxy}(\alpha_0)\right\} \to 1.$

Lemma 1 guarantees that asymptotically, all underfitted (at least one truly non-zero coefficient is missing from the model) and overfitted (all truly non-zero coefficients and one or more zero coefficient are included in the model) models estimated using unregularized PQL will have values of $\text{IC}_{\text{proxy}}(\alpha)$ greater than the value attained at the true model $\alpha_0$. From these results, we are able to infer the large sample properties of $\text{IC}(\lambda)$ for choosing the tuning parameter.

**Theorem 2.** *Let $\hat{\alpha}$ be the model chosen by minimizing $IC(\lambda)$ defined in (3). Then under conditions (C1)-(C4) and (C6), and as $n, m_1 \to \infty$, it holds that* $P(\hat{\alpha} = \alpha_0) \to 1.$

The above guarantees that the model chosen by minimizing IC($\lambda$) is asymptotically equal to the model chosen by $\lambda_0$. Since $\lambda_0$ satisfies condition (C5) and selects the true model, it follows immediately that choosing the tuning parameter based on IC($\lambda$) leads to consistent model selection using regularized PQL.

# 5 Simulation Study

We performed an empirical study to assess the performance of regularized PQL estimation and IC($\lambda$) for three commonly applied forms of GLMMs, namely the linear mixed model, Bernoulli, and Poisson GLMMs. For brevity, we only present the first two sets of results here; the Poisson GLMM results are presented in the Supplementary Material. For simplicity, we restrict our simulations to cases where the cluster sizes are the same, $m_1 = \ldots = m_n = m$. In all three settings, 200 datasets were generated for each combination of $n$ and $m$. We focused on settings where $m$ is small compared to $n$, to test the scope of the theory in Section 4. For all simulations, the power parameter was fixed at $\kappa = 2$, while the hybrid estimator was obtained by refitting the selected submodel using adaptive quadrature via the `R` package `lme4` (Bates et al., 2015).

For each setting, performance was assessed by the percentage of correctly chosen overall models, fixed effects, and random effects, as well as several measures of fit. Let $\hat{\boldsymbol{\Psi}}_{\text{method}}$ and $\hat{\boldsymbol{b}}_{\text{method}}$ generically denote the parameter estimates and predicted random effects obtained directly from regularized PQL or the hybrid estimation approach discussed in Section 3.3. Then for both estimation methods we calculated the following four quantities: mean absolute bias of the estimates $\text{E}\left(\|\hat{\boldsymbol{\Psi}}_{\text{method}} - \boldsymbol{\Psi}_0\|_1\right)$ where $\|\cdot\|_1$ denotes the $L_1$ norm, total variance of the estimates $\sum_{l=1}^{\dim(\boldsymbol{\Psi})} \text{Var}(\hat{\boldsymbol{\Psi}}_{\text{method},l})$, mean squared prediction error for random effects $\text{E}(\|\hat{\boldsymbol{b}}_{\text{method}} - \boldsymbol{b}_0\|^2)$, and the mean predicted log-likelihood $\text{E}\{\ell_{\text{pred}}(\hat{\boldsymbol{\Psi}}_{\text{method}})\}$

evaluated using a validation dataset. For all four quantities, the expectations and variances were calculated empirically across the simulated datasets. Afterwards, for each quantity we constructed a ratio comparing the hybrid estimation approach to estimates directly from regularized PQL, such that ratios less than one imply the hybrid estimator has lower absolute bias/total variance/prediction error/predicted log-likelihood relative to regularized PQL.

## 5.1 Normal Responses

We replicated the design of Bondell et al. (2010), which was subsequently used by Fan and Li (2012) and Lin et al. (2013), so we can compare our method with other recently proposed penalized likelihood methods for linear mixed models. Datasets were generated based on the true model $y_{ij} \sim N(\boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \boldsymbol{z}_{ij}^T\boldsymbol{b}_i, \sigma^2)$, where $p_f = 9$ fixed effects with fixed intercept, $p_r = 4$ random effects including a random intercept, and $\sigma^2 = 1$. The vector of true fixed effects parameters was set to $\boldsymbol{\beta}_0 = (1, 1, 0, \ldots, 0)$, while the true $4 \times 4$ random effect covariance matrix is given by vech$(\boldsymbol{D}_0) = (9, 4.8, 0.6, 0, 4, 1, 0, 1, 0, 0)$. In other words, there were seven uninformative fixed effects and one uninformative random effect. All the elements of $\boldsymbol{x}_{ij}$ and the last three elements $\boldsymbol{z}_{ij}$ were generated from the uniform distribution $U[-2, 2]$, with the first element of $\boldsymbol{z}_{ij}$ set equal to one. Four penalized likelihood methods were compared: 1) regularized PQL estimation (rPQL), 2) the SCAD-P approach of Fan and Li (2012) using the SCAD penalty, 3) the M-ALASSO approach of Bondell et al. (2010) using an adaptive lasso, and 4) the two-stage ALASSO approach of Lin et al. (2013). The results for methods 2 to 4 were taken from their respective papers.

Regularized PQL performed strongly overall; it was the best at selecting both the correct overall model and fixed effects in the small sample case, while in the large sample case there was little difference between it and SCAD-P, which correctly identified the best model

Table 1: Results from simulation Setting 1 for linear mixed models. The methods are: regularized PQL (rPQL), SCAD-P (Fan and Li, 2012), M-ALASSO (Bondell et al., 2010), and ALASSO (Lin et al., 2013). Performance was assessed in terms of percentage datasets with correctly chosen overall models (%C), fixed effects (%CF), and random effects (%CR), as well as the ratios of mean absolute bias (Bias) and total variance (Var) of the estimates, mean squared prediction error (PSE), and predicted log-likelihood (PL).

| $(n, m)$ | Method | %C | % CF | % CR | Bias/Var/PSE/PL |
|---|---|---|---|---|---|
| $(30, 5)$ | rPQL | 88 | 98 | 88 | $0.84/1.02/0.88/0.97$ |
| | SCAD-P | - | 90 | 86 | - |
| | M-ALASSO | 71 | 73 | 79 | - |
| | ALASSO | 79 | 81 | 96 | - |
| $(60, 10)$ | rPQL | 98 | 99 | 98 | $0.99/1.03/0.97/0.95$ |
| | SCAD-P | 100 | 100 | 100 | - |
| | M-ALASSO | 83 | 83 | 89 | - |
| | ALASSO | 95 | 96 | 99 | - |

in all simulated datasets (Table 1). The fact that the performance of regularized PQL was closer to SCAD-P than the other two penalized methods was not surprising, as regularized PQL and SCAD-P adopt a similar approach to group penalizing the random effect coefficients, while M-ALASSO and ALASSO instead penalize the Cholesky decomposition of the random effect covariance matrix.

All the ratios were relatively close to one, suggesting that there was no substantial differences between the hybrid estimation approach compared to regularized PQL. This was not surprising, given that for linear mixed models, PQL estimation does produce asymptotically unbiased and consistent estimators even in the setting where $m$ is fixed (Breslow and Clayton, 1993). On computation time, regularized PQL took an average of 26 and 59 seconds to fit the $(n = 30, m = 5)$ and $(n = 60, m = 10)$ settings respectively. We believe these times are competitive, while acknowledging that further reductions could have been made if we had used more sophisticated methods of optimization.

## 5.2 Bernoulli Responses

We simulated datasets from a Bernoulli GLMM using a logit link, with $p_f = p_r = 9$ covariates, both including an intercept term. For $i = 1, \ldots, n$, vectors of fixed effect covariates $\boldsymbol{x}_{ij}$ were constructed with a one in the first term and the remaining terms generated from a multivariate normal distribution $N_8(\mathbf{0}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma}_{rs} = 0.5^{|r-s|}$. The random effect covariates $\boldsymbol{z}_{ij}$ were set equal to $\boldsymbol{x}_{ij}$. The vector of true fixed effects parameters was set to $\boldsymbol{\beta}_0 = (-0.1, 1, -1, 1, -1, 0, \ldots, 0)$, while the true random effect covariance matrix was a $9 \times 9$ diagonal matrix with the first three diagonal elements set to $(3, 2, 1)$ and the remaining diagonal entries zero.

We are not aware of any available software for penalized joint selection in GLMMs. For comparison then, we write our own code to implement the following two penalized likelihood methods: 1) extending the M-ALASSO penalty of Bondell et al. (2010) to the case of non-Gaussian responses, with the tuning parameter chosen using their recommended BIC, 2) the adaptive lasso penalty of Ibrahim et al. (2011), with the tuning parameters chosen using their proposed IC$_Q$ criterion. Estimation for both methods was performed using a penalized Monte-Carlo EM algorithm and, due to their heavy computational load, considered a sequence (grid) of 100 values (combinations) of the tuning parameter. Aside from these two penalties, we also applied the `glmmLasso` package (Groll and Tutz, 2014), which performs fixed effects selection *only* in GLMMs using the unweighted lasso penalty. Since `glmmLasso` only performs fixed effects selection, we assumed that the random effects component was known, i.e. only the first three elements of $\boldsymbol{z}_{ij}$ were included in the random effects structure. As recommended by Groll and Tutz (2014), BIC was used to select the tuning parameter in `glmmLasso`.

Finally, as an alternative to penalized likelihood, we included for comparison a two stage, forward selection method using BIC$(\alpha) = -2\ell(\tilde{\Psi}_\alpha) + \log(N)\dim(\tilde{\boldsymbol{\Psi}}_\alpha)$, where $\tilde{\boldsymbol{\Psi}}_\alpha$

denotes the maximum likelihood estimates for for submodel $\alpha$. At the first stage, a saturated fixed effects structure was assumed and forward selection performed on the random effects. At the second stage, all random effects chosen in the first stage were entered into the model as fixed effects also, and forward selection was used on the remaining covariates to select them as fixed effects only. Compared to all subsets selection, the two stage approach is not only computationally more efficient, but also preserves the hierarchy of the covariates present in longitudinal GLMMs (Hui et al., 2016).

Regularized PQL performed best at selecting both fixed and random effects, with performance improving with $m$ and $n$ (Table 2). Comparing the hybrid and regularized PQL estimation methods, we see that the hybrid estimator produces considerably less biased but more variable estimates. This is consistent with the effects of penalization, that is, shrinkage of the fixed and random effects will reduce the variability of the estimates at the expense of increased bias. On the other hand, both the ratios for mean squared error prediction and predictive log-likelihood are less than one, particularly when $m$ is small compared to $n$, suggesting that the hybrid estimator did have improved predictive performance compared to directly using the regularized PQL estimates. The M-ALASSO penalty, `glmmlasso`, and forward selection using BIC all performed slightly poorer than regularized PQL at selecting the fixed effects, while on random effects selection M-ALASSO and forward selection using BIC had a tendency to overfit. Finally, the penalty of Ibrahim et al. (2011) performed poorly in this simulation, with subsequent investigation revealing that $\text{IC}_Q$ almost always chose the smallest possible set of tuning parameters (leading to the saturated model being selected). It also tended to behave erratically e.g., the loss function component of $\text{IC}_Q$ did not vary monotonically with model complexity. It should be noted that $\text{IC}_Q$ criterion was, in fact, *not* recommended for use by the authors in an earlier paper (Ibrahim et al., 2008).

Table 2: Results from simulation Setting 2 for Bernoulli GLMMs. The methods are: regularized PQL (rPQL), M-ALASSO (Bondell et al., 2010), I-ALASSO (Ibrahim et al., 2011), `glmmLasso` (Groll and Tutz, 2014), and forward selection (Forward Sel.) using BIC($\alpha$). Performance was assessed in terms of the percentage of datasets with correctly chosen overall models (%C), fixed effects (%CF), and random effects (%CR), as well as ratios of mean absolute bias (Bias) and total variance (Var) of the estimates, mean squared prediction error (PSE), and predicted log-likelihood (PL). Finally, the mean computation time for each method was also recorded, with standard deviations in parentheses.

| $(n, m)$ | Method | %C | % CF | % CR | Comp. time | Bias/Var/PSE/PL |
|---|---|---|---|---|---|---|
| | rPQL | 67 | 93 | 67 | 238(65) | 0.21/18.28/0.73/0.71 |
| | M-ALASSO | 12 | 56 | 17 | $\approx 10^4$ | - |
| $(50, 10)$ | I-ALASSO | 0 | 0 | 0 | 7309(884) | - |
| | `glmmLasso` | - | 73 | - | 908(109) | - |
| | Forward Sel. | 9 | 94 | 10 | 192(67) | - |
| | rPQL | 86 | 94 | 90 | 256(33) | 0.27/4.70/0.63/0.85 |
| | M-ALASSO | 37 | 86 | 44 | $\approx 10^4$ | - |
| $(50, 20)$ | I-ALASSO | 0 | 10 | 0 | 8748(1115) | - |
| | `glmmLasso` | - | 89 | - | 1301(162) | - |
| | Forward Sel. | 74 | 98 | 76 | 1686(391) | - |
| | rPQL | 78 | 96 | 81 | 390(73) | 0.08/6.34/0.69/0.76 |
| | M-ALASSO | 12 | 83 | 17 | $\approx 20^4$ | - |
| $(100, 10)$ | I-ALASSO | 0 | 1 | 0 | $\approx 10^4$ | - |
| | `glmmLasso` | - | 78 | - | 3226(231) | - |
| | Forward Sel. | 33 | 93 | 36 | 1614(326) | - |
| | rPQL | 95 | 97 | 98 | 501(98) | 0.21/4.07/0.70/0.86 |
| | M-ALASSO | 43 | 92 | 45 | $\approx 20^4$ | - |
| $(100, 20)$ | I-ALASSO | 0 | 18 | 0 | $\approx 10^4$ | - |
| | `glmmLasso` | - | 94 | - | 5738(264) | - |
| | Forward Sel. | 95 | 97 | 98 | 6493(1031) | - |

Except for $(n = 50, m = 10)$ where forward selection using BIC was slightly faster, regularized PQL was also the fastest method at performing joint selection, with computation time typically an order of magnitude smaller than the four competing approaches (Table 2). The long computation times of M-ALASSO and the penalty of Ibrahim et al. (2011) could be attributed to the need for a penalized Monte-Carlo EM algorithm, in contrast to regularized PQL which does not involve any integration. Finally, computation times for forward selection using BIC scaled the worst with $n$ and $m$ e.g., doubling the cluster size from $m = 10$ to 20 led to at least four-fold increase in estimation time.

Simulation results for the Poisson GLMMs are presented in the Supplementary Material, and present similar trends to those seen in the Bernoulli GLMM design above. That is, regularized PQL performed competitively in jointly selecting the fixed and random effects, while taking much less time to fit the solution path than competing methods. Also presented in the Supplementary Material are results based on using forward selection with other types of information criteria, which performed worse than BIC$(\alpha)$ shown above, as well as simulation designs where $m$ explicitly grows as a function of $n$, which empirically confirmed the estimation and selection consistency established in Section 4.

# 6 Application to Forest Health Monitoring

We applied regularized PQL estimation to a longitudinal dataset on the health status of beech trees at plots located across northern Bavaria, Germany. The aim of the analysis was to uncover important baseline and time varying covariates influencing the probability of a tree experiencing defoliation.

Table 3: Nine baseline (time independent) and two time varying covariates available for selection in the forest health dataset.

| Covariates | Brief description |
|---|---|
| | *Baseline covariates* |
| Alkali | Proportion of base alkali ions; categorical (very low, low, high, very high) |
| Canopy | Forest canopy density; continuous (%) |
| Elevation | Elevation above sea level; continuous (meters) |
| Fertilization | Fertilization applied; binary (yes, no) |
| Humus | Humus layer thickness; ordinal (five levels) |
| Inclination | Slope inclination; continuous (%) |
| Moisture | Soil moisture level; categorical (moderately dry, moderately moist, moist) |
| Soil | Soil layer depth; continuous (centimeters) |
| Stand | Stand type; categorical (deciduous, mixed) |
| | *Time varying covariates* |
| Age | Age of observation stands; continuous (years) |
| pH | Soil pH at 0–2 centimeters; continuous (centimeters) |

Different versions of the data, i.e. with differing predictors and response type, have been considered previously by Kneib et al. (2009), who focused on the spatial effects, and Groll and Tutz (2014), who examined this data to illustrate high-dimensional GLMMs. In particular, Groll and Tutz (2014) also had the goal of identifying important predictors of tree defoliation, and we will compare our results with theirs. The version of the dataset we used is the `ForestHealth` dataset in the `R2BayesX` package (Belitz et al., 2015). The dataset consists of $n = 78$ trees with $m = 22$ measurements for all trees, with a binary response $y_{ij} = 1$ indicating that defoliation exceeding 12.5% and $y_{ij} = 0$ otherwise. As displayed in Table 3, nine baseline and two time varying covariates were recorded. All continuous covariates were standardized prior to analysis, while dummy variables were created for the categorical variables.

We fitted a Bernoulli GLMM with all covariates included as fixed effects. Furthermore, to account for any potential non-linear relationship between age and the probability of defoliation on the logit scale, we included polynomial terms for age as fixed effects up to the fourth power, similar to Groll and Tutz (2014). For the random effects, we included a random intercept to account for heterogeneity in the overall health of the trees at baseline, and random slopes for age and pH to capture the variability between trees in their response to these covariates over time. We chose not to include any polynomial terms as random effects for ease of interpretation. We first fitted a saturated model to construct adaptive lasso weights. Then we used regularized PQL with the IC($\lambda$) in (3) to perform model selection, where IC($\lambda$) was used to select both $\lambda$ and $\kappa$, the latter chosen from the range $\{1, 2\}$. This resulted in the model

$$\begin{aligned}
\text{logit}(\mu_{ij}) &= 0.528 + 0.364\text{Age}_{ij} - 1.235\text{Canopy}_i - 0.101\text{pH}_{ij} \\
&\quad + b_{0i} + b_{1i}\text{Age}_{ij} + b_{2i}\text{pH}_{ij}; \quad i = 1, \ldots, 78, j = 1, \ldots, 22 \\
\widehat{\text{Cov}}(\boldsymbol{b}_i) &= \begin{pmatrix} 5.042 & 2.822 & 1.024 \\ - & 3.427 & 0.928 \\ - & - & 0.839 \end{pmatrix}.
\end{aligned}$$

Not surprisingly, older trees, increased soil acidity (lower pH), and denser forest canopy cover were all associated with increased probability of defoliation. There was substantial heterogeneity in the baseline status of the trees (remembering the continuous covariates were standardized), as well as in their responses to age and pH. Regularized PQL shrunk all the polynomial terms of age to zero, suggesting that perhaps any truly non-linear effect of age was masked by the large variability between trees in their linear responses and/or that the non-linear effects were comparatively small compared to the between-tree variability. To confirm this, we fitted the selected submodel in the `R` package `lme4` using Laplace's approximation, and compared it to a GLMM that included polynomial terms for age up the fourth power. The resulting bootstrap likelihood ratio test confirmed that these polynomial fixed effects for age were indeed not significant (P-value = 0.11). Finally, all the off-diagonal terms in the estimated random effect covariance matrix were positive, indicating that large effects for one predictor tended to occur with large effects in the other predictors, e.g., the higher the baseline probability of defoliation, the worse the effect of increasing age and soil acidity on the the tree's health.

The results obtained here differ from those in Groll and Tutz (2014), who applied the `glmmLasso` package to a very similar version of this dataset, in some important ways: 1) Groll and Tutz (2014) did not have pH as a predictor in their analysis, whereas we found that, based on regularized PQL, pH was both an important fixed and random effect; 2) the method of Groll and Tutz (2014) identified an important fixed, quadratic effect of age, although the magnitude of the coefficient was very close to zero; 3) regularized PQL identified canopy cover as an important predictor, whereas Groll and Tutz (2014) identified stand type as important. Perhaps the driving reason behind these differences was that `glmmLasso` selects only fixed effects, and Groll and Tutz (2014) only included a random intercept in the model. By contrast, regularized PQL performs joint selection so we could include and select on Age and pH as random slopes, and indeed both these covariates were identified as being significant effects.

# 7 Discussion

Joint selection of fixed and random effects in mixed models is a challenging problem, due to both the intractability of the marginal likelihood and the large number of candidate models. In this article, we proposed regularized PQL estimation to overcome these problems. By combining the PQL with adaptive lasso penalties for selecting the fixed and random effects, regularized PQL offers a attractive method of computing the solution path. We showed that regularized PQL is selection consistent. This is an important result given PQL was originally motivated by Breslow and Clayton (1993) as a fast but approximate method of estimating GLMMs. With regularized PQL, we have a computationally fast approach of joint selection that asymptotically selects the true set of fixed and random effects. We proposed an information criterion for choosing the tuning parameter in regularized PQL which leads to selection consistency. The criterion combines a BIC-type model complexity penalty for the fixed effects with a AIC-type penalty for the random effects. This is a reflection of the differing degrees of model complexity needed for the fixed coefficients, which grows at rate $O(1)$, versus the random coefficients, which grows at rate $O(n)$. In the linear regression and penalized GLM contexts respectively, Shao (1997) and Zhang et al. (2010) investigated the impacts of differing degrees of model complexity, and our criterion can be regarded as an extension of such results to the GLMM context using regularized PQL estimation.

Simulations demonstrate the selection consistency of regularized PQL in conjunction with the proposed information criterion, showing that it can outperform other methods of joint selection while offering considerable reductions in computation time. The use of a hybrid estimation method further helps to reduce finite sample bias and improve prediction. Indeed, using regularized PQL for fast model selection only mirrors other works in the GLM context, where penalized likelihood approaches have been proposed purely as a means of computationally efficient model selection (e.g., Schelldorfer et al., 2014; Hui et al., 2015). Of course, we acknowledge further simulations are required to fully assess the robustness of regularized PQL selection e.g., how it performs when the truly non-zero coefficients and hence signal to noise ratio is small, and that there are other methods of joint selection in GLMMs which were not included in our study e.g., the predictive shrinkage selection method of Hu et al. (2015) designed specifically for Poisson mixture models in the context of network analysis.

One obvious extension to make to regularized PQL estimation is to high-dimensional GLMMs, where the number of fixed and random effects grows with the number of clusters and/or cluster size; see for example the recent works of Fan and Li (2012) and Groll and Tutz (2014). For the case where $p_r$ remains fixed but $p_f$ is permitted to grow, we believe the estimation and selection consistency results established in this article will continue to hold, provided the conditions on the tuning parameter are altered slightly. In a more general setting where $p_r$ grows with $m_1$ and $n$, some of the results established for high-dimensional penalized GLMs (see the overview by Fan and Lv, 2010) may in principle be adapted to GLMMs, especially since the PQL treats the random effects as if they are fixed coefficients. Another possible extension which is especially useful for longitudinal studies is to modify rPQL so that the penalties select covariates in a hierarchical manner, such that all covariates in the model are chosen as either fixed effects only or composite (fixed and random) effects (see for instance, Hui et al., 2016). This reflects the notion that covariates in longitudinal GLMMs should not be included in the model as random slopes only.

# Acknowledgements

This research was supported by the Australian Research Council discovery project grant DP140101259. Thanks to Andreas Groll and Gordana Popovic for useful discussions.

# References


Bates, D., Maechler, M., Bolker, B., and Walker, S. (2015). *lme4: Linear mixed-effects models using Eigen and S4*. R package version 1.1-8.

Belitz, C., Brezger, A., Kneib, T., Lang, S., and Umlauf, N. (2015). *BayesX: Software for Bayesian inference in structured additive regression models*. Version 1.0.

Bondell, H. D., Krishna, A., and Ghosh, S. K. (2010). Joint variable selection for fixed and random effects in linear mixed-effects models. *Biometrics*, 66:1069–1077.

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

Demidenko, E. (2013). *Mixed Models: Theory and Applications with R*. Wiley Series in Probability and Statistics. Wiley.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96:1348–1360.

Fan, J. and Lv, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, 20:101–148.

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

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33:1–22.

Garcia, T. P., Müller, S., Carroll, R. J., and Walzem, R. L. (2014). Identification of important regressor groups, subgroups and individuals via regularization methods: application to gut microbiome data. *Bioinformatics*, 30:831–837.

Groll, A. and Tutz, G. (2014). Variable selection for generalized linear mixed models by $\ell_1$–penalized estimation. *Statistics and Computing*, 24:137–154.

Hu, K., Choi, J., Sim, A., and Jiang, J. (2015). Best predictive generalized linear mixed model with predictive lasso for high-speed network data analysis. *International Journal of Statistics and Probability*, 4:132–148.

Hui, F. K. C., Mueller, S., and Welsh, A. H. (2016). Hierarchical Selection of Fixed and Random Effects in Generalized Linear Mixed Models. *Statistica Sinica*, Accepted for publication.

Hui, F. K. C., Warton, D. I., and Foster, S. D. (2015). Tuning parameter selection for the adaptive lasso using ERIC. *Journal of the American Statistical Association*, 110:262–269.

Ibrahim, J. G., Zhu, H., Garcia, R. I., and Guo, R. (2011). Fixed and random effects selection in mixed effects models. *Biometrics*, 67:495–503.

Ibrahim, J. G., Zhu, H., and Tang, N. (2008). Model selection criteria for missing-data problems using the EM algorithm. *Journal of the American Statistical Association*, 103:1648–1658.

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., Rao, J. S., Gu, Z., Nguyen, T., et al. (2008). Fence methods for mixed model selection. *The Annals of Statistics*, 36:1669–1692.

Kneib, T., Hothorn, T., and Tutz, G. (2009). Variable selection and model choice in geoadditive regression models. *Biometrics*, 65:626–634.

Lehmann, E. (1983). *Theory of Point Estimation*. Wiley.

Lele, S. R., Nadeem, K., and Schmuland, B. (2010). Estimability and likelihood inference for generalized linear mixed models using data cloning. *Journal of the American Statistical Association*, 105:1617–1625.

Lin, B., Pang, Z., and Jiang, J. (2013). Fixed and random effects selection by REML and pathwise coordinate optimization. *Journal of Computational and Graphical Statistics*, 22:341–355.

Lin, X. and Breslow, N. E. (1996). Bias correction in generalized linear mixed models with multiple components of dispersion. *Journal of the American Statistical Association*, 91:1007–1016.

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

Müller, S., Scealy, J. L., and Welsh, A. H. (2013). Model selection in linear mixed models. *Statistical Science*, 28:135–167.

Müller, S. and Welsh, A. (2009). Robust model selection in generalized linear models. *Statistica Sinica*, 19:1155–1170.

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

Peng, H. and Lu, Y. (2012). Model selection in linear mixed effect models. *Journal of Multivariate Analysis*, 109:109–129.

Rabe-Hesketh, S., Skrondal, A., and Pickles, A. (2002). Reliable estimation of generalized linear mixed models using adaptive quadrature. *The Stata Journal*, 2:1–21.

Schelldorfer, J., Meier, L., and Bühlmann, P. (2014). Glmmlasso: an algorithm for high-dimensional generalized linear mixed models using $\ell_1$-penalization. *Journal of Computational and Graphical Statistics*, 23:460–477.

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

Tierney, L. and Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, 81:82–86.

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

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.

Zhang, Y., Li, R., and Tsai, C. (2010). Regularization parameter selections via generalized information criterion. *Journal of the American Statistical Association*, 105:312–323.

# A Derivation of Covariance Matrix update in equation (2)

Consider the Laplace approximation log-likelihood $\ell_{\text{LA}}(\mathbf{\Psi})$ given in Section 2 of the main text. Substituting in the regularized PQL estimates $\hat{\boldsymbol{\beta}}_\lambda$ and $\hat{\boldsymbol{b}}_\lambda$, we obtain

$$\ell(\boldsymbol{D}) = -\frac{n}{2}\log\det(\boldsymbol{D}) - \frac{1}{2}\sum_{i=1}^{n}\log\det(\boldsymbol{Z}_i^T\hat{\boldsymbol{W}}_{\lambda i}\boldsymbol{Z}_i + \boldsymbol{D}^-) + \sum_{i=1}^{n}\sum_{j=1}^{m_i}\log f(y_{ij}|\hat{\boldsymbol{\beta}}_\lambda, \hat{\boldsymbol{b}}_{\lambda i})$$
$$- \frac{1}{2}\sum_{i=1}^{n}\hat{\boldsymbol{b}}_{\lambda i}^T\boldsymbol{D}^-\hat{\boldsymbol{b}}_{\lambda i},$$

where for $i = 1, \dots, n$, $\hat{\boldsymbol{b}}_{\lambda i}$ are the regularized PQL estimates of the random effects and $\hat{\boldsymbol{W}}_{\lambda i,jj} = (\text{Var}(y_{ij})g'(\hat{\mu}_{\lambda ij})^2)^-$. Differentiating $\ell(\boldsymbol{D})$ with respect to $\boldsymbol{D}$, we have

$$\frac{\partial\ell(\boldsymbol{D})}{\partial\text{vech}(\boldsymbol{D})} = -\frac{n}{2}\text{vech}(\boldsymbol{D}^-) + \frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{D}^-\otimes\boldsymbol{D}^-)\text{vech}\{(\boldsymbol{Z}_i^T\hat{\boldsymbol{W}}_{\lambda i}\boldsymbol{Z}_i + \boldsymbol{D}^-)^{-1}\}$$
$$+ \frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{D}^-\otimes\boldsymbol{D}^-)\text{vech}(\hat{\boldsymbol{b}}_{\lambda i}\hat{\boldsymbol{b}}_{\lambda i}^T)$$
$$= \mathbf{0}$$

It follows that $n(\boldsymbol{D}^-\otimes\boldsymbol{D}^-)^-\text{vech}(\boldsymbol{D}^-) = n\ \text{vech}(\boldsymbol{D}) = \sum_{i=1}^{n}\text{vech}\{(\boldsymbol{Z}_i^T\hat{\boldsymbol{W}}_{\lambda i}\boldsymbol{Z}_i + \boldsymbol{D}^-)^{-1} + \hat{\boldsymbol{b}}_{\lambda i}\hat{\boldsymbol{b}}_{\lambda i}^T\}$, from which the formula in (2) of the main text follows. □

# B Outlines of Proofs

Full derivations are found in the Supplementary Material; here we provide an outline for each of these proofs.

Proof of Theorem 1: We consider the objective function $\ell_p(\boldsymbol{\beta}, \mathbf{\Gamma}, \boldsymbol{b}) = \ell_{\text{PQL}}(\boldsymbol{\beta}, \mathbf{\Gamma}, \boldsymbol{b}) - \lambda\sum_{k=1}^{p_f} v_k|\beta_k| - \lambda\sum_{l=1}^{p_r} w_l\|\boldsymbol{b}_{\bullet l}\|$ and define

$\Delta = n^{-1}\{\ell_p(\boldsymbol{\beta}_0 + \alpha_m \boldsymbol{u}_1, \boldsymbol{\Gamma}, \boldsymbol{b}_0 + \alpha_m \boldsymbol{u}_2) - \ell_p(\boldsymbol{\beta}_0, \boldsymbol{\Gamma}, \boldsymbol{b}_0)\}$ for a vector $\boldsymbol{u}$ of appropriate length and $\alpha_m = m_1^{-1/2}$. Under conditions (C1)-(C2), (C4) and (C5a), we show that this difference is asymptotically dominated by a quadratic term of form $-(\alpha_m^2/2)\boldsymbol{u}^T\{-n^{-1}\nabla^2\ell_1(\boldsymbol{\beta}, \boldsymbol{b})\}\boldsymbol{u}$, which is negative. This implies that with probability tending to one there exists a local maximum at $(\boldsymbol{\beta}_0, \boldsymbol{b}_0)$, which we then show to be a global maximum.

Given the $m_1^{1/2}$-consistency from the first part of the theorem, to prove selection consistency of the regularized PQL estimates we need only show that for truly zero fixed and random effects, the signs of the score equations $\partial\ell_p(\boldsymbol{\Psi}, \boldsymbol{b})/\partial\beta_k|_{\hat{\boldsymbol{\Psi}}_\lambda, \hat{\boldsymbol{b}}_\lambda}$ and $\partial\ell_p(\boldsymbol{\Psi}, \boldsymbol{b})/\partial b_{il}|_{\hat{\boldsymbol{\Psi}}_\lambda, \hat{\boldsymbol{b}}_\lambda}$ depend asymptotically only on the sign of the estimated coefficients. This is proved by expanding the score equations about the true parameter values and, in particular, using condition (C5b) to show that the derivative of the adaptive (group) lasso penalty dominates all the terms in the score equations.

Proof of Lemma 1: We consider separately the cases of underfitted and overfitted models. In the first case, we utilize condition (C6) to show that the difference in the loss function $-2\sum_{i=1}^{n}\sum_{j=1}^{m_i}\log f(y_{ij}|\tilde{\boldsymbol{\beta}}_\alpha, \tilde{\boldsymbol{b}}_{\alpha i})$ between any underfitted model and the true model is positive and asymptotically dominates all differences in the model complexity. In the second case, Condition (C3) is utilized to show that the difference in the loss function between any overfitted model and the true model is asymptotically dominated by the difference in the model complexity penalties $\log(n)\dim(\tilde{\boldsymbol{\beta}}_\alpha) + 2n\dim(\tilde{\boldsymbol{b}}_{\alpha i})$, which by definition is greater than zero when overfitting.

Proof of Theorem 2: Under conditions (C1)-(C2), (C4), we prove the result $N^{-1}\ell_{\text{PQL}}(\hat{\boldsymbol{\Psi}}_{\lambda_0}, \hat{\boldsymbol{b}}_{\lambda_0}) = N^{-1}\ell_1(\tilde{\boldsymbol{\beta}}_{\alpha_0}, \tilde{\boldsymbol{b}}_{\alpha_0}) + o_p(1)$. We then show for any tuning parameter $\lambda$ producing an underfitted or overfitted model $\alpha$, it holds that $\{\text{IC}(\lambda) - \text{IC}(\lambda_0)\} \geq \{\text{IC}_{\text{proxy}}(\alpha) - \text{IC}_{\text{proxy}}(\alpha_0)\}$. Since the right hand side is positive with probability tending to one by Lemma 1, the result follows.