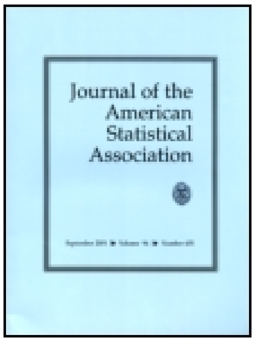


Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: http://www.tandfonline.com/loi/uasa20

Sparse Pairwise Likelihood Estimation for Multivariate Longitudinal Mixed Models

Francis K.C. Hui, Samuel Müller & A.H. Welsh

To cite this article: Francis K.C. Hui, Samuel Müller & A.H. Welsh (2017): Sparse Pairwise Likelihood Estimation for Multivariate Longitudinal Mixed Models, Journal of the American Statistical Association, DOI: 10.1080/01621459.2017.1371026

To link to this article: http://dx.doi.org/10.1080/01621459.2017.1371026

Sparse Pairwise Likelihood Estimation for Multivariate Longitudinal Mixed Models

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

It is becoming increasingly common in longitudinal studies to collect and analyze data on multiple responses. For example, in the social sciences we may be interested in uncovering the factors driving mental health of individuals over time, where mental health is measured using a set of questionnaire items. One approach to analyzing such multi-dimensional data is multivariate mixed models, an extension of the standard univariate mixed model to handle multiple responses. Estimating multivariate mixed models presents a considerable challenge however, let alone performing variable selection to uncover which covariates are important in driving each response. Motivated

*Corresponding author: fhui28@gmail.com; Mathematical Sciences Institute, The Australian National University, 2601, Canberra, ACT, Australia

by composite likelihood ideas, we propose a new approach for estimation and fixed effects selection in multivariate mixed models, called Approximate Pairwise Likelihood Estimation and Shrinkage (APLES). The method works by constructing a quadratic approximation to each term in the pairwise likelihood function, and then augmenting this approximate pairwise likelihood with a penalty that encourages both individual and group coefficient sparsity. This leads to a relatively fast method of selection, as we can utilize coordinate ascent type methods to then construct the full regularization path for the model. Our method is the first to extend penalized likelihood estimation to multivariate generalized linear mixed models. We show that the APLES estimator attains a composite likelihood version of the oracle property. We propose a new information criterion for selecting the tuning parameter, which employs a dynamic model complexity penalty to facilitate aggressive shrinkage, and demonstrate that it asymptotically leads to selection consistency i.e., leads to the true model being selected. A simulation study demonstrates that the APLES estimator outperforms several univariate selection methods based on analyzing each outcome separately.

Keywords: composite likelihood, LASSO, mixed models, multivariate longitudinal data, pairwise fitting, penalized likelihood, variable selection

1 Introduction

In longitudinal studies, it is increasingly common to collect data on multiple responses for each individual. Such multivariate longitudinal data is becoming the rule rather than the exception, given that many of these studies are conducted on a large scale, following hundreds of individuals over many years. Therefore it is more cost-effective as well as informative to collect multiple responses. One example of this kind of study is the Household Income and Labour Dynamics in Australia (HILDA) survey (Watson and Wooden, 2012),

ACCEPTED MANUSCRIPT

a nationally representative panel survey collected annually in Australia since 2001. As part of the survey, data on an individual's mental health are collected to study how mental health changes over time in response to various personal and environmental factors (Leach et al., 2014). Mental health data are fundamentally multivariate in nature: the HILDA survey for instance uses a set of five items designed to quantify mental health based on experiences in the last month e.g., How much time in the past 4 weeks have you been a very nervous person? For each item, the individual provides a rating from 1 to 6 (none to all of the time).

One approach for analyzing multivariate longitudinal data is to extend the Generalized Linear Mixed Model (GLMM) commonly used for a single repeated measure to handle multiple responses. In this article, we refer to such models as multivariate GLMMs (see Verbeke et al., 2014, for a detailed review). A key benefit of this approach is that it allows us to borrow strength across responses: we can use the joint information across multiple responses to both better inform the overall population's trajectory over time, and capture the association between outcomes by modeling the cross response correlation of the random effects. On the other hand, analyzing such data using multivariate GLMMs presents formidable challenges in both estimation and variable selection.

When the responses are not all assumed to be normally distributed, the marginal likelihood does not have a closed analytic form, and we have potentially a quite high-dimensional integral to deal with. One approach for getting around this is to instead maximize a composite likelihood (Varin et al., 2011). In the context of multivariate GLMMs, this was first considered by Fieuws and Verbeke (2006), who propose a pairwise fitting estimation method based on separately fitting all possible bivariate response GLMMs and then averaging the maximum likelihood estimates obtained from the separate fits. Pairwise fitting has been subsequently studied in many articles for estimation and hypothesis testing in multivariate GLMMs (e.g., Fieuws et al., 2006; Faes et al., 2008; Ivanova et al., 2015). As

ACCEPTED MANUSCRIPT

Faes et al. (2008) emphasize however, estimation based on pairwise fitting is different from maximizing the pairwise likelihood directly, since the former involves *post-hoc* averaging of separate estimates of the same parameter (see also Vasdekis et al., 2014, who extended the approach to consider weighted means in order to improve efficiency). The more general issue of inference using the pairwise fitting method however remains largely unexplored.

On the issue of variable selection, in most applications of mixed models, the number of candidate fixed effects can often be considerably more than the number of random effects e.g., often only a random intercept and random slope for time is included, as in our example in Section 6. With multivariate GLMMs, we can also expect the mean structures to vary considerably between the responses: some covariates may be uninformative for all responses, in which case its vector of fixed effect coefficients (one for each response) will be equal to zero simultaneously, while other covariates may be partially informative, in which case some of the elements of the vector are non-zero. In light of the possible range in mean structures between responses, we argue that standard methods of variable selection such as all subsets and forward/backward selection using information criteria are impractical. Instead, we use penalized likelihood methods as a computationally feasible method of selection. Note that while penalized selection has been heavily studied for GLMs (e.g., Zou, 2006; Bondell and Reich, 2008), their use in univariate GLMMs dates back only to Bondell et al. (2010). We refer to Müller et al. (2013) for a comprehensive review of variable selection in linear mixed models, and Hui et al. (2017a) for an example of penalized likelihood in univariate GLMMs based on maximum likelihood estimation.

In this article, we propose a new, computationally efficient approach for fixed effects selection in multivariate GLMMs, called Approximate Pairwise Likelihood Estimation and Shrinkage (APLES). The method works by first taking the pairwise composite likelihood and applying a quadratic approximation to each of the bivariate likelihood terms. The resulting approximate pairwise likelihood bears some resemblance to the one step sparse estimate of Zou and Li (2008) and the unified least squares approximation method of Wang and Leng (2012) for generalized linear models, although such an approach has not been considered before for univariate let alone multivariate GLMMs. More relevant is the link between the approximate pairwise likelihood function and the pairwise fitting estimation of Fieuws and Verbeke (2006) and others reviewed above, since maximizing each of the component bivariate likelihoods is analogous to what is done in pairwise fitting. Rather than directly averaging the separate estimates however, we augment the approximate pairwise likelihood with a penalty in order to achieve sparse fixed effect coefficients. Specifically, we propose a penalty which encourages group sparsity across responses, such that the fixed effect coefficients for a covariate can be shrunk to zero simultaneously. To our knowledge, this article is the first to extend penalized likelihood methods to multivariate GLMMs, and thus presents an important advance in both the mixed model and composite likelihood literatures.

It is important to highlight the difference between the APLES estimator and penalizing the pairwise likelihood directly to achieve sparse estimates. Specifically, because most parameters including the fixed effect coefficients occur in multiple bivariate log-likelihood terms, the score equation for a penalized pairwise likelihood will involve a sum of several separate score equations, each of which generally does not possess a tractable form unless both responses are normally distributed. By contrast, calculating the APLES estimator is straightforward precisely because the approximate pairwise likelihood is a sum of quadratic forms, with each term resembling the quadratic form seen in the log-likelihood function for a multivariate normal distribution. Therefore, we can utilize coordinate ascent type methods to obtain closed form updates and efficiently construct the full regularization path.

Under general regularity conditions, we show that APLES is selection consistent and

ACCEPTED MANUSCRIPT

achieves a composite likelihood version of the oracle property, i.e., the APLES method asymptotically performs as well as if the true fixed effects structure is known in advance and estimated using pairwise likelihood. This leads to asymptotic normality with covariance equal to the inverse of the Godambe information matrix (Varin et al., 2011). Although we work in the setting where the number of candidate fixed effects is bounded as the sample size grows, the multivariate nature of the response means there is still a large number of coefficients up for selection and therefore it is a large dimensional problem. Furthermore, the selection consistency and oracle property are, to our knowledge, the first such asymptotic results to be proven in the multivariate GLMMs and composite likelihood estimation literatures. For tuning parameter selection, we propose a new information criterion which utilizes the approximate pairwise likelihood as the goodness of fit function, and show that it leads to selection consistency i.e., the criterion asymptotically chooses a tuning parameter corresponding to the true model. While tuning parameter selection for penalized likelihood in generalized linear models has been studied extensively (e.g., Gunes and Bondell, 2012), it has been much less explored in mixed models (see Groll and Tutz, 2014; Hui et al., 2016, for some examples in univariate GLMMs). Likewise, there is relatively little literature on composite likelihood based information criteria, with two notable exceptions being Varin and Vidoni (2005) and Gao and Song (2010), who consider composite likelihood versions of the Akaike and Bayesian information criterion respectively. Establishing the theoretical properties of our proposed criterion presents an important advance to both the multivariate GLMM and composite likelihood literatures. We also point out that our proposed information criterion differs from many other criteria that have been proposed in that it uses a dynamic model complexity penalty similar to that of Hui et al. (2015b). This in turn leads to more aggressive shrinkage compared to the Bayesian information criteria, and empirically we found that it resulted in better selection performance.

ACCEPTED MANUSCRIPT

A simulation study demonstrates that the APLES estimator in conjunction with the proposed information criterion performed well compared to the standard approach of performing model selection on each response separately to select the fixed effects structure. We apply the APLES estimator to the HILDA survey to uncover some of the social and environmental drivers behind changes to different aspects of mental health over time. We provide template `R` code for calculating the APLES estimator and for performing the simulations in the Supplementary Material.

To summarize, the main contributions of this article are as follows: 1) We propose APLES, a method for fixed effects selection in multivariate GLMMs, which combines composite likelihood ideas with a penalty for inducing (possibly group) sparsity in the fixed effects, 2) We establish estimation consistency and the oracle property of the proposed variable selection method, 3) We propose a method of selecting the tuning parameter which satisfies the conditions necessary for selection consistency, 4) Simulations demonstrate the strong empirical performance of the APLES estimator and the proposed tuning parameter selection method, over the standard approaches of performing variable selection on each response separately, 5) Application of the APLES estimator to the HILDA datasets uncovers many of the important factors driving the mental health of individuals over time.

2 Multivariate GLMMs

For individual $i = 1, \ldots, n$, let y_{ijk} denote the measurement of response $k = 1, \ldots, K$ at time point $j = 1, \ldots, n_i$. Along with the responses, let $\boldsymbol{x}_{ij}$ denote a vector of p_f covariates to be included in the model as fixed effects, and $\boldsymbol{z}_{ij}$ a vector of p_r random effect covariates. Unless stated otherwise, both $\boldsymbol{x}_{ij}$ and $\boldsymbol{z}_{ij}$ contain an intercept term as the first element. The multivariate GLMM is defined as follows (Verbeke et al., 2014). Conditional on a vector

ACCEPTED MANUSCRIPT

of random effects $\boldsymbol{b}_i = (\boldsymbol{b}_{i1}^T, \ldots, \boldsymbol{b}_{iK}^T)^T$, the responses y_{ijk} are assumed to be independent observations from the exponential family with mean μ_{ijk} and response-specific dispersion parameter ϕ_k, the latter of which may or may not be known. While it is possible for the K sets of responses to be of mixed type e.g., a combination of continuous and binary responses, for simplicity we assume all the responses come from the same distributional form, as in the case for our motivating example where all the responses are ordinal. For a known link function $g(\cdot)$, the mean is related to the covariates as $g(\mu_{ijk}) = \eta_{ijk} = \boldsymbol{x}_{ij}^T\boldsymbol{\beta}_k + \boldsymbol{z}_{ij}^T\boldsymbol{b}_{ik}$, where $\boldsymbol{\beta}_k$ and $\boldsymbol{b}_{ik}$ are the fixed and random effect coefficients for response k. We assume that the random effects are drawn from a multivariate normal distribution with zero mean vector and unstructured random effects covariance matrix $\boldsymbol{\Sigma}$ of dimension $Kp_r \times Kp_r$ i.e., $f(\boldsymbol{b}_i|\boldsymbol{\Sigma}) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$. Here $\boldsymbol{\Sigma}$ carries information pertaining to both the temporal correlation within a response, found on the K blocks of $p_r \times p_r$ submatrices lying on the diagonal of $\boldsymbol{\Sigma}$, and the cross-correlations between responses, found on the submatrices lying away from the diagonal.

Let $\boldsymbol{y}_{ik} = (y_{i1k}, \ldots, y_{in_ik})$ denote the vector of responses for individual i and outcome k, and $\boldsymbol{\Psi} = (\boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_K^T, \phi_1, \ldots, \phi_K, \text{vech}(\boldsymbol{\Sigma})^T)^T$ the vector of parameters. Then assuming the individuals $i = 1, \ldots, n$ are independent, the marginal log-likelihood of the multivariate GLMM is given by

$$\ell(\boldsymbol{\Psi}) = \sum_{i=1}^{n} \ell_i(\boldsymbol{y}_{i1}, \ldots, \boldsymbol{y}_{iK}|\boldsymbol{\Psi}) = \sum_{i=1}^{n} \log\left(\int \prod_{j=1}^{n_i}\prod_{k=1}^{K} f(y_{ijk}|\boldsymbol{b}_{ik}, \boldsymbol{\beta}_k, \phi_k) f(\boldsymbol{b}_i|\boldsymbol{\Sigma}) d\boldsymbol{b}_i\right), \quad (1)$$

where $f(y_{ijk}|\boldsymbol{b}_{ik}, \boldsymbol{\beta}_k, \phi_k)$, the conditional distribution of the responses, belongs to the exponential family. The dimension of $\boldsymbol{b}_i$ makes the integral in the marginal likelihood function potentially of a large dimension. This has motivated alternative, computationally less burdensome approaches to estimation as we discuss in Section 3.

By far the most commonly studied case of multivariate GLMMs is when all K responses are normally distributed. Here, if we let $\boldsymbol{y}_{ik} = (y_{i1k}, \ldots, y_{in_ik})$, then $f(y_{ijk}|\boldsymbol{b}_{ik}, \boldsymbol{\beta}_k, \phi_k) = \mathcal{N}(\boldsymbol{x}_{ij}^T\boldsymbol{\beta}_k + \boldsymbol{z}_{ij}^T\boldsymbol{b}_{ik}, \phi_k)$ where ϕ_k is the response-specific variance. In our motivating dataset, the K responses are ordinal i.e., ratings by individuals to items related to mental health. Therefore, we use a cumulative logit model defined as follows (McCullagh, 1980). Let response k be an ordinal variable with L_k levels, such that $y_{ijk} \in \{1, \ldots, L_k\}$. It is assumed that the number of levels, L_k, does not change over time. Define a proxy response variable y^*_{ijkl} such that $y^*_{ijkl} = 1$ if $y_{ijk} = l$ and zero otherwise. Then for individual i, we have the multinomial distribution $f(\boldsymbol{y}^*_{ijkl}|\boldsymbol{b}_{ik}, \boldsymbol{\beta}_k, \phi_k) = \prod_{l=1}^{L_k} \left\{F(\nu_{kl} - \eta_{ijk}) - F(\nu_{k(l-1)} - \eta_{ijk})\right\}^{y^*_{ijkl}}$, where $F(x) = \{1 + \exp(-x)\}^{-1}$ is the logit link and $\eta_{ijk} = \boldsymbol{x}_{ij}^T\boldsymbol{\beta}_k + \boldsymbol{z}_{ij}^T\boldsymbol{b}_{ik}$. Note the negative sign in front of the linear predictor: this is standard in ordinal regression, so that larger values of η_{ijk} correspond to a higher probability of the response y_{ijk} being in a higher category. The parameters $\{\nu_{kl}; k = 1, \ldots, K; l = 0, \ldots, L_k\}$ are the response-specific cutoffs, constrained to be in ascending order i.e., $\nu_{k0} = -\infty < \nu_{k1} < \ldots < \nu_{kL_k} = \infty$, with the constraint $\nu_{k1} = 0$ for all $k = 1, \ldots, K$ to ensure parameter identifiability (since the first element of $\boldsymbol{x}_{ij}$ is a fixed intercept term). Finally, note that in the case of $L_k = 2$ for all k, the above reduces to a logistic multivariate GLMM.

3 Fixed Effects Selection using APLES

Given the computational burden involved in trying to maximizing $\ell(\boldsymbol{\Psi})$ in (1), especially if K and/or p_r is not small, we replace the marginal likelihood by a composite likelihood as the objective function. Specifically, we consider the pairwise likelihood function $\ell_{\text{PL}}(\boldsymbol{\Psi}) = \sum_{i=1}^n \ell_i(\boldsymbol{y}_{i1}, \boldsymbol{y}_{i2}|\boldsymbol{\Psi}_{12}) + \ldots + \sum_{i=1}^n \ell_i(\boldsymbol{y}_{i1}, \boldsymbol{y}_{iK}|\boldsymbol{\Psi}_{1K}) + \sum_{i=1}^n \ell_i(\boldsymbol{y}_{i2}, \boldsymbol{y}_{i3}|\boldsymbol{\Psi}_{23}) + \ldots + \sum_{i=1}^n \ell_i(\boldsymbol{y}_{i(K-1)}, \boldsymbol{y}_{iK}|\boldsymbol{\Psi}_{(K-1)K}) = \sum_{r=1}^{K-1} \sum_{s=r+1}^{K} \sum_{i=1}^n \ell_i(\boldsymbol{y}_{ir}, \boldsymbol{y}_{is}|\boldsymbol{\Psi}_{rs})$, where $\ell_i(\boldsymbol{y}_{ir}, \boldsymbol{y}_{is}|\boldsymbol{\Psi}_{rs})$ is the bivariate log-likelihood function for response pair (r, s) for individual i, and is given by

$$\ell_i(\boldsymbol{y}_{ir}, \boldsymbol{y}_{is}|\boldsymbol{\Psi}_{rs}) = \log\left(\int \prod_{j=1}^{n_i} f(y_{ijr}|\boldsymbol{b}_{ir}, \boldsymbol{\beta}_r, \phi_r) f(y_{ijs}|\boldsymbol{b}_{is}, \boldsymbol{\beta}_s, \phi_r) f(\boldsymbol{b}_{ir}, \boldsymbol{b}_{is}|\boldsymbol{\Sigma}_{rs}) d\boldsymbol{b}_{ir} d\boldsymbol{b}_{is}\right).$$

Each of the bivariate likelihoods depends only on a subset of the full parameter vector, specifically, $\boldsymbol{\Psi}_{rs}$ involves only the fixed effect coefficients and the submatrix of $\boldsymbol{\Sigma}$ that describe the random effects covariances for response pair (r, s). While maximizing the pairwise likelihood is easier than the full marginal likelihood, involving only integrals of dimension $2p_r$, it still presents a considerable challenge since many of the parameters are found in several of the bivariate likelihood terms comprising $\ell_{\text{PL}}(\boldsymbol{\Psi})$. This motivated Fieuws and Verbeke (2006) and others to propose the pairwise fitting method, where each of the $K(K-1)/2$ bivariate likelihoods $\sum_{i=1}^{n} \ell_i(\boldsymbol{y}_{ir}, \boldsymbol{y}_{is}|\boldsymbol{\Psi}_{rs})$ is maximized separately and the parameters are then *post-hoc* averaged to obtain a unique set of estimates.

Motivated by the goal of fixed effects selection in multivariate mixed models, in this article we propose an alternate approach which we shall see is actually closely linked to the pairwise fitting method. Consider a quadratic expansion of $\sum_{i=1}^{n} \ell_i(\boldsymbol{y}_{ir}, \boldsymbol{y}_{is}|\boldsymbol{\Psi}_{rs})$ about its maximum, $\sum_{i=1}^{n} \ell_i(\boldsymbol{y}_{ir}, \boldsymbol{y}_{is}|\boldsymbol{\Psi}_{rs}) = \sum_{i=1}^{n} \ell_i(\boldsymbol{y}_{ir}, \boldsymbol{y}_{is}|\tilde{\boldsymbol{\Psi}}_{rs}) - 2^{-1}(\boldsymbol{\Psi}_{rs} - \tilde{\boldsymbol{\Psi}}_{rs})^T \boldsymbol{H}(\tilde{\boldsymbol{\Psi}}_{rs})(\boldsymbol{\Psi}_{rs} - \tilde{\boldsymbol{\Psi}}_{rs})$, where $\boldsymbol{H}(\boldsymbol{\Psi}_{rs}) = -\sum_{i=1}^{n} \partial^2 \ell_i(\boldsymbol{y}_{ir}, \boldsymbol{y}_{is}|\boldsymbol{\Psi}_{rs})/\partial\boldsymbol{\Psi}_{rs}\partial\boldsymbol{\Psi}_{rs}^T$ is the negative Hessian, and $\tilde{\boldsymbol{\Psi}}_{rs}$ is the maximizer of the bivariate log-likelihood satisfying $\sum_{i=1}^{n} \partial\ell_i(\boldsymbol{y}_{ir}, \boldsymbol{y}_{is}|\tilde{\boldsymbol{\Psi}}_{rs})/\partial\boldsymbol{\Psi}_{rs} = \boldsymbol{0}$. By applying the quadratic expansion to each of the terms in $\ell_{\text{PL}}(\boldsymbol{\Psi})$, we propose the approximate pairwise likelihood function

$$\ell_{\text{APL}}(\boldsymbol{\Psi}) = -\frac{1}{2}\sum_{r=1}^{K-1}\sum_{s=r+1}^{K} (\boldsymbol{\Psi}_{rs} - \tilde{\boldsymbol{\Psi}}_{rs})^T \boldsymbol{H}(\tilde{\boldsymbol{\Psi}}_{rs})(\boldsymbol{\Psi}_{rs} - \tilde{\boldsymbol{\Psi}}_{rs}). \tag{2}$$

The approximate pairwise likelihood will be used shortly as the basis for sparse fixed effects selection by augmenting it with a penalty. Without penalization though, we can show that maximizing (2) leads to an estimator which takes the form of a weighted mean of the individual maximizers $\tilde{\boldsymbol{\Psi}}_{rs}$. Specifically, without loss of generality, suppose we augment each vector $\tilde{\boldsymbol{\Psi}}_{rs}$ by inserting zeros in the appropriate positions so that it is of the same length as $\boldsymbol{\Psi}$. We denote these augmented vectors as $\tilde{\boldsymbol{\Psi}}_{rs,\text{full}}$. The positions of the zeros in $\tilde{\boldsymbol{\Psi}}_{rs,\text{full}}$ thus correspond to the elements in $\boldsymbol{\Psi}$ not associated with response pair (r, s). Likewise, we augment $\boldsymbol{H}(\tilde{\boldsymbol{\Psi}}_{rs})$ by inserting rows and columns of zeros so that it is a $\dim(\boldsymbol{\Psi}) \times \dim(\boldsymbol{\Psi})$ symmetric matrix, and denote this as $\boldsymbol{H}(\tilde{\boldsymbol{\Psi}}_{rs,\text{full}})$. The zero rows and columns in $\boldsymbol{H}(\tilde{\boldsymbol{\Psi}}_{rs,\text{full}})$ again correspond to second and cross derivatives not associated with response pair (r, s). Then we can write (2) as $\ell_{\text{APL}}(\boldsymbol{\Psi}) = -2^{-1} \sum_{r=1}^{K-1} \sum_{s=r+1}^{K} (\boldsymbol{\Psi} - \tilde{\boldsymbol{\Psi}}_{rs,\text{full}})^T \boldsymbol{H}(\tilde{\boldsymbol{\Psi}}_{rs,\text{full}})(\boldsymbol{\Psi} - \tilde{\boldsymbol{\Psi}}_{rs,\text{full}})$. Solving for $\partial \ell_{\text{APL}}(\boldsymbol{\Psi})/\partial \boldsymbol{\Psi} = \mathbf{0}$, we obtain the weighted mean estimator $\tilde{\boldsymbol{\Psi}}_{\text{wm}} = \left\{ \sum_{r=1}^{K-1} \sum_{s=r+1}^{K} \boldsymbol{H}(\tilde{\boldsymbol{\Psi}}_{rs,\text{full}}) \right\}^{-1} \sum_{r=1}^{K-1} \sum_{s=r+1}^{K} \boldsymbol{H}(\tilde{\boldsymbol{\Psi}}_{rs,\text{full}}) \tilde{\boldsymbol{\Psi}}_{rs,\text{full}}$ (see also Vasdekis et al., 2014). Thus we see that the approximate pairwise likelihood naturally facilitates an estimator based on a weighted mean of the separate bivariate GLMM estimates, and contrasts with the pairwise fitting approach which uses an unweighted mean. It would be of interest to compare the efficiency of this weighted mean estimator with the unweighted mean, although given our focus here is on variable selection we do not pursue this issue further (see Vasdekis et al., 2014, for relevant work).

To perform fixed effects selection, we combine (2) with a penalty on the $\boldsymbol{\beta}_k$'s. Instead of separately penalizing each coefficient, we make use of the inherent grouping of the fixed effects across responses, on a per-covariate basis. For $k = 1, \ldots, K$, let $\boldsymbol{\beta}_k = (\beta_{k1}, \ldots, \beta_{kp_f})$. Then the set $\{\beta_{kd}; k = 1, \ldots, K\}$ represents the cluster of K coefficients associated with the d^{th} fixed effect covariate. If a covariate d is uninformative for all fixed effects, it is appealing to use a penalty which can simultaneously shrink all these coefficients to zero, thereby removing the covariate from the model entirely. On the other hand, when covariates are partially informative for a subset of the K outcomes, we want a penalty capable of setting only some of the effects to zero. In summary, we seek penalties with both group and individual coefficient sparsity, the former for removing completely uninformative covariates and the latter for handling partially informative covariates. Such penalties which reflect the natural structure of the covariates have been proposed in other settings e.g., composite penalties for generalized linear models (Zhao et al., 2009; Huang et al., 2012), finite mixture models (Hui et al., 2015a), and in applied settings such as microbial data (Garcia et al., 2014), and indeed there are a number of ways by which penalties can be constructed to possess both group and individual coefficient sparsity. To our knowledge however, this article is the first to consider them for longitudinal mixed models with multiple outcomes.

In light of the above discussion, we propose the following Approximate Pairwise Likelihood Estimator and Shrinkage (APLES) method for fixed effects selection in multivariate GLMMs.

Definition 1. *Let $\tilde{\beta}_{kd,wm}$ be the estimate of β_{kd} obtained from the weighted mean estimator $\tilde{\mathbf{\Psi}}_{wm}$. For a single tuning parameter $\lambda > 0$, the APLES estimator for multivariate longitudinal GLMMs is defined as*

$$\hat{\mathbf{\Psi}} = \arg\max_{\mathbf{\Psi}} \; \ell_{APL}(\mathbf{\Psi}) - n\lambda \sum_{d=2}^{p_f} \sum_{k=1}^{K} w_{kd}|\beta_{kd}| - n\lambda \sum_{d=2}^{p_f} v_d \left(\sum_{k=1}^{K} \beta_{kd}^2 \right)^{1/2},$$

where $\ell_{APL}(\mathbf{\Psi})$ is given by (2), $w_{kd} = |\tilde{\beta}_{kd,wm}|^{-2}$ and $v_d = \left(\sum_{k=1}^{K} \tilde{\beta}_{kd,wm}^2 \right)^{-1}$ are pre-defined adaptive weights.

The summation in both components of the penalty begins at $d = 2$, as we assume the first element in each $\boldsymbol{\beta}_k$ corresponds to a fixed intercept that we do not penalize. If there are other fixed effect covariates in any of the K outcomes that are not to be penalized, then the corresponding adaptive weights w_{kd} and/or v_d may be set to zero accordingly. One noteworthy example of this is covariates included as both fixed and random effects, where we may not necessarily want to penalize the fixed effects because it can lead to undesirable cases of non-hierarchical shrinkage i.e., the covariate ends up in one or more of the K mean structures model as a random effect only (see Hui et al., 2017a, when $K = 1$). Otherwise, the adaptive weights are constructed from the weighted mean estimator $\tilde{\boldsymbol{\Psi}}_{\text{wm}}$. The inclusion of pre-defined weights facilitates flexible, adaptive penalization with only a single tuning parameter. Also, we fix the powers on the pre-defined weights in order to facilitate development of the asymptotic results in Section 4, while at the same time easing computation as we only have to search over a single tuning parameter instead of two.

The APLES estimator achieves flexible fixed effects selection by combining an adaptive LASSO with an adaptive group LASSO, linked by a common tuning parameter, to achieve both group and individual coefficient sparsity. In particular, the second component of the penalty is applied across responses on a per-covariate basis: the L_2 norm encourages group sparsity where all K fixed effects for a covariate d are set equal to zero simultaneously, thus removing the covariate from all components of the multivariate GLMM. This is combined with the K individual group sparsity events encouraged by the first component of the penalty. This combination of an overall group sparsity event along with K individual sparsity events means we can remove fixed effect covariates from all K responses simultaneously, or remove it from only a subset of the responses.

3.1 Estimation

The APLES estimator is straightforward to calculate, and not surprisingly the most challenging and computationally intensive part of Definition 1 lies at the beginning of the calculations where we have to construct the approximate pairwise likelihood function. Each of the estimates $\tilde{\boldsymbol{\Psi}}_{rs}$ is obtained by maximizing the marginal log-likelihood of the bivariate GLMM, which is often done using the Expectation-Maximization algorithm or adaptive quadrature. The Hessian matrix can then be obtained through Louis's method (Louis, 1982) for instance. However, once the approximate pairwise likelihood is built, construction of the regularization path for $\ell_{\text{APL}}(\boldsymbol{\Psi})$ is comparably fast and straightforward. For a value of λ, we first apply a local linear approximation (Zou and Li, 2008) to the penalty function. Suppose at iteration t we have current estimates $\hat{\boldsymbol{\Psi}}^{(t)}$. Then for covariate d we can approximate the second component of the penalty as $v_d \left(\sum_{k=1}^{K} \beta_{kd}^2 \right)^{1/2} \approx v_d \left(\sum_{k=1}^{K} (\hat{\beta}_{kd}^{(t)})^2 \right)^{1/2} + v_d \left(\sum_{k=1}^{K} (\hat{\beta}_{kd}^{(t)})^2 \right)^{-1/2} \sum_{k=1}^{K} \hat{\beta}_{kd}^{(t)} \left(|\beta_{kd}| - |\hat{\beta}_{kd}^{(t)}| \right)$. As previously, suppose we augment all the vectors $\tilde{\boldsymbol{\Psi}}_{rs}$ by inserting zeros in the appropriate positions so that they are of the same length as $\boldsymbol{\Psi}$, and analogously we augment each of the $\boldsymbol{H}(\tilde{\boldsymbol{\Psi}}_{rs})$ by inserting rows and columns of zeros so that it is a square symmetric matrix of dimension $\dim(\boldsymbol{\Psi})$. Again, we denote these augmented quantities as $\tilde{\boldsymbol{\Psi}}_{rs,\text{full}}$ and $\boldsymbol{H}(\tilde{\boldsymbol{\Psi}}_{rs,\text{full}})$ respectively. Let $\Psi_{[u]}$ denote element u in $\boldsymbol{\Psi}$, and $H(\tilde{\boldsymbol{\Psi}}_{rs,\text{full}})_{[uv]}$ denote element (u, v) in $\boldsymbol{H}(\tilde{\boldsymbol{\Psi}}_{rs,\text{full}})$. Then for element $u = 1, \ldots, \dim(\boldsymbol{\Psi})$, after applying the local linear approximation above, setting the score equation to zero from Definition 1 leads to the equation $-\sum_{r=1}^{K-1} \sum_{s=r+1}^{K} H(\tilde{\boldsymbol{\Psi}}_{rs,\text{full}})_{[uu]}(\Psi_{[u]} - \tilde{\Psi}_{rs,\text{full}[u]}) = \sum_{r=1}^{K-1} \sum_{s=r+1}^{K} \sum_{v \neq u} H(\tilde{\boldsymbol{\Psi}}_{rs,\text{full}})_{[uv]}(\Psi_{[v]} - \tilde{\Psi}_{rs,\text{full}[v]}) + n\lambda \hat{\xi}_{[u]}^{(t)} \text{sign}(\Psi_{[u]})$, where $\hat{\boldsymbol{\xi}}^{(t)}$ is a vector of length $\dim(\boldsymbol{\Psi})$ defined as follows: for $k = 1, \ldots, K$ and $d = 1, \ldots, p_f$, we set $\hat{\xi}_{[kp_f - p_f + d]}^{(t)} = w_{kd} + v_d \hat{\beta}_{kd}^{(t)} \left(\sum_{k=1}^{K} (\hat{\beta}_{kd}^{(t)})^2 \right)^{-1/2}$ corresponding to the penalized fixed effect coefficients. Note the set $\{kp_f - p_f + d; k = 1, \ldots, K; d = 1, \ldots, p_f\}$ covers elements 1 to Kp_f. For elements $u = Kp_f + 1, \ldots, \dim(\boldsymbol{\Psi})$, we set $\hat{\xi}_{[u]}^{(t)} = 0$ corresponding to the dispersion parameters $\phi_1, \ldots, \phi_K$ (cutoffs if ordinal responses) and the random effects covariance matrix vech$(\boldsymbol{\Sigma})$, all of which are not penalized. Define $S(a, c)$ as the soft-thresholding operator, such that $S(a, c) = \text{sign}(a)(|a| - c)_+$. From the above score equation we obtain the

ACCEPTED MANUSCRIPT

closed form solution

$$\hat{\Psi}_{[u]} = \frac{S\left(r_{[u]},\ n\lambda\hat{\xi}^{(t)}_{[u]}\right)}{\sum_{r=1}^{K-1}\sum_{s=r+1}^{K} H(\tilde{\mathbf{\Psi}}_{rs,\mathrm{full}})_{[uu]}}, \tag{3}$$

where $r_{[u]} = \sum_{r=1}^{K-1}\sum_{s=r+1}^{K}\left\{H(\tilde{\mathbf{\Psi}}_{rs,\mathrm{full}})_{[uu]}\tilde{\Psi}_{rs,\mathrm{full}[u]} - \sum_{v\neq u} H(\tilde{\mathbf{\Psi}}_{rs,\mathrm{full}})_{[uv]}(\Psi_{[v]} - \tilde{\Psi}_{rs,\mathrm{full}[v]})\right\}$. The above results suggest the following update algorithm: for a given λ, 1) apply the local linear approximation to the second part of the penalty, 2) use (3) to cycle through $u = 1,\ldots,\dim(\mathbf{\Psi})$ and produce updated estimates $\hat{\mathbf{\Psi}}^{(t+1)}$, 3) iterate between steps 1 and 2 until convergence. Once completed we can then move on to the next value of λ on the regularization path, using warm starts i.e., the estimates based on the previous value of λ as starting points. Finally, we point out that while the estimation procedure is itself not particularly new, bearing similarity to other coordinate-wise methods (e.g., Friedman et al., 2010), the novelty of the estimation procedure comes precisely from the methodological developments necessary in order to reach a stage where we *can* adapt coordinate-wise optimization methods i.e., the proposal of the approximate pairwise likelihood in (2) as the loss function in combination with the adaptive LASSO and adaptive group LASSO penalties.

4 Theoretical Properties

In this section, we establish the following large sample properties. First, we show that under mild regularity conditions on the likelihood function, tuning parameter, and adaptive weights, the APLES estimator in Definition 1 satisfies a composite likelihood version of the oracle property. Second, we propose a new information criterion for choosing the tuning parameter and show that it attains selection consistency.

Let the number of fixed and random effect covariates p_f and p_r be bounded as the number of clusters $n \to \infty$. The number of responses K is also assumed to be constant with n. Note also that even though p_f and p_r are bounded, it nevertheless presents a large dimensional selection problem with Kp_f fixed effect coefficients in consideration.

4.1 Oracle Property

Let $\boldsymbol{\Psi}^0 = (\boldsymbol{\beta}_1^{0T}, \ldots, \boldsymbol{\beta}_K^{0T}, \phi_1^0, \ldots, \phi_K^0, \text{vech}(\boldsymbol{\Sigma}^0)^T)^T$ denote the true parameter point. Without loss of generality, for $k = 1, \ldots, K$ let $\boldsymbol{\beta}_k^0 = (\boldsymbol{\beta}_{k1}^{0T}, \boldsymbol{\beta}_{k2}^{0T} = \mathbf{0}^T)^T$, where $\boldsymbol{\beta}_{k1}^0$ are the truly non-zero fixed effects for response k. In turn, we can write $\boldsymbol{\Psi}^0 = (\boldsymbol{\Psi}_1^{0T}, \boldsymbol{\Psi}_2^{0T} = \mathbf{0}^T)^T$ where $\boldsymbol{\Psi}_1^0 = (\boldsymbol{\beta}_{11}^{0T}, \ldots, \boldsymbol{\beta}_{K1}^{0T}, \phi_1^0, \ldots, \phi_K^0, \text{vech}(\boldsymbol{\Sigma}^0)^T)^T$ and $\boldsymbol{\Psi}_2^0 = (\boldsymbol{\beta}_{12}^{0T}, \ldots, \boldsymbol{\beta}_{K2}^{0T})$. Likewise, the APLES estimator in Definition 1 can be written as $\hat{\boldsymbol{\Psi}} = (\hat{\boldsymbol{\Psi}}_1^T, \hat{\boldsymbol{\Psi}}_2^T)^T$ and $\hat{\boldsymbol{\beta}}_k = (\hat{\boldsymbol{\beta}}_{k1}^T, \hat{\boldsymbol{\beta}}_{k2}^T)$ for all $k = 1, \ldots, K$. The following regularity conditions are required to study the asymptotic behavior of the APLES estimator.

(C1) The true parameter point $\boldsymbol{\Psi}^0$ is an interior point of the parameter space Ω, and the model is identifiable at $\boldsymbol{\Psi}^0$. Furthermore, there exists a constant κ_1 satisfying $0 < \kappa_1 < \min\{|\beta_{kd}^0|; \beta_{kd}^0 \neq 0\} < \infty$.

(C2) For all $r, s = 1, \ldots, K$ and $\boldsymbol{\Psi}_{rs} \in \Omega_{rs}$ where $\Omega_{rs} \in \Omega$ is an open subset, the log-likelihood $\ell_1(\boldsymbol{y}_{1r}, \boldsymbol{y}_{1s}|\boldsymbol{\Psi}_{rs})$ has common support and is at least three times differentiable on $\boldsymbol{\Psi}_{rs}$.

(C3) Let $\ell_{\text{PL1}}(\boldsymbol{\Psi}) = \sum_{r=1}^{K-1} \sum_{s=r+1}^{K} \ell_1(\boldsymbol{y}_{1r}, \boldsymbol{y}_{1s}|\boldsymbol{\Psi}_{rs})$. Then (a) The first derivative satisfies $\text{E}\left(\partial \ell_{\text{PL1}}(\boldsymbol{\Psi}^0)/\partial \boldsymbol{\Psi}\right) = \mathbf{0}$, and (b) the sensitivity matrix $\mathcal{I}(\boldsymbol{\Psi}) = \text{E}\left\{-\partial^2 \ell_{\text{PL1}}(\boldsymbol{\Psi})/\partial \boldsymbol{\Psi} \partial \boldsymbol{\Psi}^T\right\}$ and variability matrix $\mathcal{J}(\boldsymbol{\Psi}) = \text{E}\left\{\left(\partial \ell_{\text{PL1}}(\boldsymbol{\Psi})/\partial \boldsymbol{\Psi}\right)\left(\partial \ell_{\text{PL1}}(\boldsymbol{\Psi})/\partial \boldsymbol{\Psi}\right)^T\right\}$ are both finite and positive definite at $\boldsymbol{\Psi}^0$.

(C4) There exists an open subset $\Omega^* \in \Omega$ containing $\boldsymbol{\Psi}^0$ such that for all $\boldsymbol{\Psi} \in \Omega^*$,

ACCEPTED MANUSCRIPT

there exist integrable functions $Q_{uvw}(\boldsymbol{v}_1)$ such that for all $r = 1, \ldots, K-1$ and $s = r+1, \ldots, K$, $|\partial^3 \ell_1(\boldsymbol{y}_{1r}, \boldsymbol{y}_{1s}|\boldsymbol{\Psi}_{rs})/\partial\Psi_u\partial\Psi_v\Psi_w| < Q_{uvw}(\boldsymbol{v}_1)$ for all $u, v, w = 1, \ldots, \dim(\boldsymbol{\Psi})$ and $\boldsymbol{v}_1$ denoting the data i.e., the responses $\boldsymbol{y}_{1k}$ and covariates $\boldsymbol{x}_{1j}$ and $\boldsymbol{z}_{1j}$, collected for the first individual, and which satisfy $\mathrm{E}\{Q^2_{uvw}(\boldsymbol{v}_1)\} < \infty$.

(C5) The tuning parameter satisfies $n^{1/2}\lambda \to 0$ and $n\lambda \to \infty$.

Conditions (C1)-(C4) are general assumptions typically made when studying composite likelihood theory (see for instance, Section 9.2, Molenberghs and Verbeke, 2006), and are required to ensure that the pairwise likelihood for the multivariate GLMM is sufficiently smooth and well-defined in the neighborhood of the true parameter point. They can be thought of as analogs to the conditions made when studying full maximum likelihood estimation in univariate GLMMs (e.g., Ibrahim et al., 2011), with the only major difference being condition (C3) where the sensitivity and variability matrices are not the same due to the use of a pairwise likelihood. Condition (C1) implies all the diagonal elements of the $\boldsymbol{\Sigma}^0$ (and hence all elements of $\boldsymbol{\Psi}^0_1$) are non-zero. That is, all the random effects included in the model are truly important. This has been done to simplify the theoretical derivations, although the condition could actually be relaxed to allow some of the diagonal elements of $\boldsymbol{\Sigma}^0$ to be zero i.e., the saturated model overfits the random effects for one or more of the responses. We do not pursue this extension here however, given the focus of the APLES estimator is on producing sparse fixed effect coefficients. Conditions (C2)-(C4) are defined in terms of the likelihood contribution of the first individual, who serves as an arbitrary representative as the n individuals are assumed to be independent clusters.

To assess the large sample properties of the APLES estimator, we first need the following result regarding the weighted mean estimator.

ACCEPTED MANUSCRIPT

Lemma 1. *Under conditions (C1)-(C4), and as $n \to \infty$, it holds that* $\|\tilde{\mathbf{\Psi}}_{wm} - \mathbf{\Psi}^0\| = O_p(n^{-1/2})$.

The proofs of all results are provided in the Supplementary Material. Lemma 1 demonstrates the $n^{1/2}$-consistency of the weighted mean estimator. This in turn allows us to gauge the behavior of the adaptive weights w_{kd} and v_d. We now present the main result concerning the oracle property of the proposed penalized likelihood estimator.

Theorem 1. *Under conditions (C1)-(C5), and as $n \to \infty$, the APLES estimator given by Definition 1 satisfies the oracle property:*

(a) Asymptotic normality: $n^{1/2}(\hat{\mathbf{\Psi}}_1 - \mathbf{\Psi}_1^0) \xrightarrow{d} \mathcal{N}\left\{\mathbf{0}, \boldsymbol{G}_1^{-1}(\mathbf{\Psi}^0)\right\}$,

(b) Selection consistency: For $k = 1, \ldots, K$, it holds that $P(\hat{\boldsymbol{\beta}}_{k2} = \mathbf{0}) \to 1$.

where $\boldsymbol{G}(\mathbf{\Psi}^0) = \mathcal{I}(\mathbf{\Psi}^0)\mathcal{J}^{-1}(\mathbf{\Psi}^0)\mathcal{I}(\mathbf{\Psi}^0)$ and $\boldsymbol{G}_1(\mathbf{\Psi}^0)$ denotes the $\dim(\mathbf{\Psi}_1^0) \times \dim(\mathbf{\Psi}_1^0)$ submatrix of $\boldsymbol{G}(\mathbf{\Psi}^0)$ associated with $\mathbf{\Psi}_1^0$.

The proof of the theorem is similar to that of the oracle property in Zou and Li (2008), but involves additional complications arising from the asymptotic behavior of the pairwise and approximate pairwise likelihoods. Theorem 1a implies asymptotically that the estimates of the truly non-zero parameters are normally distributed with covariance equal to the inverse of the Godambe information matrix. Theorem 1b ensures that with probability tending to one, the APLES estimator selects only the truly non-zero fixed effect coefficients. That is, even though a covariate d is included *a-priori* as a fixed effect in all K responses, in large samples it will only be selected for the (subset of) responses for which it is truly informative. Overall, the theorem presents a composite likelihood version of the oracle property i.e., asymptotically we perform as well as if we know the true multivariate GLMM and estimated it using pairwise likelihood.

4.2 Tuning Parameter Selection

Given a particular dataset, we adapt the Extended Regularized Information Criterion (ERIC) of Hui et al. (2015b), who considered GLMs with the adaptive LASSO penalty, for use in choosing the tuning parameter in Definition 1.

$$\text{ERIC}_{\text{APL}}(\lambda) = -2\ell_{\text{APL}}(\hat{\boldsymbol{\Psi}}) - \log(\lambda)\,\hat{p}_f(\lambda), \tag{4}$$

where $\ell_{\text{APL}}(\hat{\boldsymbol{\Psi}})$ is the approximate pairwise likelihood evaluated at the APLES estimate and $\hat{p}_f(\lambda) = \sum_{k=1}^{K}\sum_{d=2}^{p_f} \mathbb{1}_{\hat{\beta}_{kd}\neq 0}$ counts the number of estimated non-zero fixed effects coefficients (see Müller and Welsh, 2010, for a review of information criteria for model selection).

Equation (4) differs from the form found in Hui et al. (2015b) in three ways: 1) the most important difference is that we use the approximate pairwise likelihood as the goodness of fit function. This is sensible here given it is the loss function when calculating the APLES estimator in Definition 1; 2) the form of ERIC proposed by Hui et al. (2015b) included an additional parameter in the model complexity term to control the severity of penalization, which they argued was necessary in settings where the number of covariates grew with sample size. In our situation, with both p_f and p_r fixed, we choose to omit this, although we acknowledge that future research should explore the potential inclusion of this term; 3) finally, our model complexity penalty takes the form $-\log(\lambda) = \log(1/\lambda)$ whereas the form of ERIC in Hui et al. (2015b) uses $\log(n/\lambda)$. This difference is simply due to different parameterizations of the tuning parameter used i.e., λ versus $n\lambda$.

The key feature of ERIC is its *dynamic* model complexity penalty which depends on the tuning parameter itself (Hui et al., 2015b). This contrasts with the *static* complexity penalties in Bayesian Information Criterion (BIC) used previously for other penalized likelihood methods e.g., for univariate GLMMs Ibrahim et al. (2011) used the $\log(n)$ penalty

ACCEPTED MANUSCRIPT

while Lin et al. (2013) used the $\log(\sum_{i=1}^{n} n_i)$ penalty. For a given dataset, these criteria penalize a fixed amount for every coefficient entered into the model. By contrast, the degree of penalization induced by ERIC differs depending on how complex the model is already, as captured by λ. In particular, the quantity $-\log(\lambda)$ becomes more severe the smaller λ is i.e., the faster λ tends to zero. Since small values of λ correspond to larger models, this implies ERIC's dynamic model complexity penalty leads to more aggressive fixed effects shrinkage, resulting in less overfitting and sparser models. Based on extensive simulations, some of which are presented in Section 5, we found that this aggressive shrinkage enforced by ERIC lead to better finite sample performance compared to using BIC to choose the tuning parameter. We can also compare (4) to information criteria proposed for variable selection with composite likelihoods. Specifically, Varin and Vidoni (2005) and Gao and Song (2010) studied the asymptotic behavior of composite likelihood based Akaike and Bayesian Information Criteria respectively. One interesting difference between these criteria and ERIC is that the former count the number of non-zero parameters (effective degrees of freedom) based on the trace of $\mathcal{I}^{-1}(\mathbf{\Psi})\mathcal{J}(\mathbf{\Psi})$. Apart from saving additional, possibly burdensome computation, we choose to use the number of penalized estimates not shrunk to zero as the aforementioned trace form is challenging to extend to the approximate pairwise likelihood setting e.g., simply summing traces built from the $K(K+1)/2$ pairwise likelihoods comprising (2) is likely to over count the number of parameters.

We now demonstrate that (4) asymptotically selects a λ satisfying condition (C5) and therefore leads to estimators with the oracle property. Consider a solution path for the penalized likelihood estimator indexed by the interval of tuning parameters $\lambda \in [0, \lambda_{\max}]$, where $\lambda_{\max} = O(1)$ corresponds to a multivariate GLMM where all the penalized fixed effects are shrunk to zero. Every value of λ in the interval then defines a model containing a subset of the fixed effects, based on the non-zero elements in the APLES estimate, and we denote this model by $\mathcal{M}_\lambda$. Furthermore, for every submodel we can calculate an unpenalized estimate based on maximizing the submodel analogue of (2), and we denote that estimate here as $\hat{\boldsymbol{\Psi}}(\mathcal{M}_\lambda)$. Finally, the true model, defined by the non-zero elements $\boldsymbol{\Psi}_1^0$, is denoted as $\mathcal{M}_0$.

Partition $[0, \lambda_{\max}]$ into three sets: 1) $\Lambda_0 = \{\lambda : \mathcal{M}_\lambda = \mathcal{M}_0\}$, which is the set of λ values that select the true model, 2) $\Lambda_- = \{\lambda : \mathcal{M}_\lambda \not\supset \mathcal{M}_0\}$, which is the set defining underfitted models i.e., models missing at least one truly non-zero fixed effect in one or more of the responses, 3) $\Lambda_+ = \{\lambda : \mathcal{M}_\lambda \supset \mathcal{M}_0\}$, which is the set defining overfitted models i.e., models containing the true model and at least one truly zero fixed effect in one or more of the responses. Let $\lambda_0 = n^{-1}\log(n)$ be a tuning parameter satisfying condition (C5), and hence $\mathrm{P}(\mathcal{M}_{\lambda_0} = \mathcal{M}_0) \to 1$ by Theorem 1. We remark that λ_0 is constructed for theoretical purposes only, and need not be the tuning parameter chosen by minimizing (4). The following result outlines the large sample behavior of ERIC for overfitted and underfitted models compared to when it is evaluated at λ_0.

Lemma 2. *Under conditions (C1)-(C5), and as $n \to \infty$, it holds that $P(\inf_{\lambda \in \Lambda_- \cup \Lambda_+} ERIC_{APL}(\lambda) - ERIC_{APL}(\lambda_0) > 0) \to 1$.*

Lemma 2 ensures that any tuning parameter that selects an overfitted or underfitted multivariate GLMM will asymptotically produce a larger value of the approximate pairwise likelihood ERIC compared to the model chosen using λ_0. This leads to the following result.

Theorem 2. *Define $\hat{\lambda}$ as the tuning parameter chosen by minimizing $ERIC_{APL}(\lambda)$ in (4). Then under conditions (C1)-(C5), and as $n \to \infty$, it holds that $P(\mathcal{M}_{\hat{\lambda}} = \mathcal{M}_0) \to 1$.*

Theorem 2 implies that, with probability tending to one, using ERIC to choose the tuning parameter leads to selection consistency. The proof follows immediately from Lemma 2 and the fact that $\mathrm{P}(\mathcal{M}_{\lambda_0} = \mathcal{M}_0) \to 1$, and is therefore omitted.

5 Simulation Study

We performed a simulation to compare the APLES estimator to other methods of model selection based on fitting separate GLMMs to each response, with the aim being to assess whether jointly modeling the responses offered any empirical advantage (in terms of fixed effects selection) compared to analyzing each response separately. We chose to focus primarily on Gaussian responses, as much of the research and available software for selection in univariate GLMMs has been developed for this case, thereby giving us the opportunity to compare APLES to a range of alternative approaches. Simulations involving multivariate binary and ordinal responses are provided in the Supplementary Material.

We compared the APLES estimator to the following separate response method: 1) a backward elimination approach in univariate GLMMs based on hypothesis testing as implemented in the `R` package `lmerTest` with default settings (Kuznetsova et al., 2016); 2) a backward elimination approach in univariate GLMMs based on BIC with model complexity penalty $\log\left(\sum_{i=1}^{n} n_i\right)$, which is the version of BIC implemented in the `lme4` package (Bates et al., 2015); 3) a special case of the joint penalties in Bondell et al. (2010) and Lin et al. (2013), such that only fixed effects selection is performed in univariate GLMMs using an adaptive LASSO penalty, and the recommended BIC is used for choosing the tuning parameter. We also considered using ERIC to choosing the tuning parameter in this case i.e., for the adaptive LASSO penalty in univariate GLMMs, but found that its performance was on par with or worse than the recommended BIC of Bondell et al. (2010) and Lin et al. (2013) and thus have omitted its results below. Besides, ERIC so far has not been considered for use in mixed models overall, with this article being the first, and we consider its application to univariate GLMMs specifically as an avenue of future research. In addition to the above three methods, we considered the `glmmLasso` package (Groll and Tutz, 2014), which performs fixed effects selection in univariate GLMMs using the unweighted LASSO penalty. However due to its poor performance compared to the other methods, its results have been omitted below. Finally, we employed two methods for choosing the tuning parameter in the APLES estimator: 1) ERIC as given in (4), 2) a BIC-type criterion with model complexity based on the total sample size, $\text{BIC2}(\lambda) = -2\ell_{\text{APL}}(\hat{\boldsymbol{\Psi}}) + \log\left(\sum_{i=1}^{n} n_i\right)\hat{p}_f(\lambda)$; see also additional simulation results in the Supplementary Material where we compared three information criteria for choosing the tuning parameter in APLES.

We simulated data from a true multivariate GLMM, with $K = 5$ Gaussian responses, $p_f = 16$ fixed effect covariates, and $p_r = 3$ random effect covariates. We generated covariates $\boldsymbol{x}_{ij}$ by setting the first element equal to one for a fixed intercept, and simulating the remaining 15 elements from a multivariate Gaussian distribution with mean zero and covariance $\text{Cov}(x_{ijr}, x_{ijs}) = 0.5^{|r-s|}$. The random effect covariates $\boldsymbol{z}_{ij}$ were then set as the first three elements of $\boldsymbol{x}_{ij}$. Next, let $\boldsymbol{B}^0$ denote the 5×16 matrix of true fixed effect coefficients, where row k is the vector of true coefficients for response k. We set the first column of $\boldsymbol{B}^0$ equal to $(-2, -1, 0, 1, 2)$, and simulated the remaining $5 \times 15 = 75$ elements from a standard normal distribution. To make the fixed effects sparse, we then randomly selected 40% of the elements from columns 4 to 16 (26 elements) in $\boldsymbol{B}^0$ and set them to zero. We also set all the elements in columns 10, 15, and 16 to zero. The above procedure ensures that no coefficients in the first three columns of $\boldsymbol{B}^0$ are zero, reflecting the fact these columns correspond to covariates that are included as random effects. In summary, elements 10, 15, and 16 in $\boldsymbol{x}_{ij}$ are completely uninformative covariates for all 5 outcomes, while some other elements in $\boldsymbol{x}_{ij}$ correspond to partially informative covariates. To complete the true model, we constructed a $Kp_r \times Kp_r = 15 \times 15$ random effects covariance matrix by simulating from a Wishart distribution with 16 degrees of freedom and a scale matrix set to a diagonal matrix with elements 0.1. The vector of 16 true random effect coefficients for each individual was then simulated from a multivariate Gaussian distribution with mean zero and the above covariance matrix. Finally, conditional on $\boldsymbol{b}_i$, the responses y_{ijk} for $k = 1, \ldots, 5$ were generated from a Gaussian distribution with mean $\boldsymbol{x}_{ij}^T \boldsymbol{\beta}_k + \boldsymbol{z}_{ij}^T \boldsymbol{b}_{ik}$ and variance equal to one.

We considered combinations of $n = \{25, 50, 100\}$ and equal cluster sizes $n_i = \{10, 20\}$. For each combination of n and n_i, we simulated 200 datasets. For all methods considered, the true random effects structure was assumed to be known, i.e., $\boldsymbol{z}_{ij}$ and the first three elements in $\boldsymbol{x}_{ij}$ were included and no selection performed on them. Therefore only elements 4 to 16 in $\boldsymbol{x}_{ij}$, a total of 65 coefficients, were available for selection. Performance was assessed based on the mean number of false positives i.e., truly zero coefficients that are not shrunk to zero, the mean number of false negatives i.e., true non-zero coefficients that are shrunk to zero, and the mean squared error of the estimated coefficients, $\mathrm{E}\left(\|\hat{\boldsymbol{B}} - \boldsymbol{B}^0\|^2\right)$ where $\hat{\boldsymbol{B}}$ is the estimated matrix of fixed effect coefficients. For these measures of performance, the means and expectations are performed empirically over the 200 simulated datasets. As a single measure of selection performance, we also calculated the F-measure defined as $F = 2 \times \text{true positives}/(2 \times \text{true positives} + \text{false positives} + \text{false negatives})$ (Powers, 2011). The F-measure lies between 0 and 1, with values closer to one indicative of better classification (between non-zero and zero coefficients).

From Table 1, the APLES estimator in combination with ERIC performed the best overall: in all settings it attained the highest F-measure and lowest or second lowest mean squared error compared to the separate response selection based methods. That both the APLES estimator approaches have a lower number of false negatives (indicative of underfitting) is perhaps suggestive of the improved power in using a joint estimation and selection approach: by borrowing strength across responses, APLES has improved efficiency and power to better detect truly non-zero fixed effect coefficients compared to analyzing

Table 1: Simulation results for fixed effects selection, comparing methods (from left to right): 1) the APLES estimator with ERIC, 2) the APLES estimator with BIC2(λ), 3) backward elimination using `lmerTest`, 4) backward elimination using BIC, 5) the adaptive LASSO. Performance was assessed using the mean number of false positives (FP) and false negatives (FN), mean squared error (MSE), and F-measure (F_1).The method with the highest F-measure in each setting is highlighted in bold.

(n, n_i)	APLES$_{\text{ERIC}}$ FP/ FN/ MSE/ F_1	APLES$_{\text{BIC2}}$ FP/ FN/ MSE/ F_1	`lmerTest` FP/ FN/ MSE/ F_1	BIC FP/ FN/ MSE/ F_1	Adapt. LASSO FP/ FN/ MSE/ F_1
$(25, 10)$	2.77/ 0.58/ 0.98/ **0.95**	8.77/ 0.17/ 1.10/ 0.87	2.96/ 4.13/ 1.58/ 0.93	1.91/ 5.04/ 1.81/ 0.88	1.94/ 4.40/ 1.57/ 0.89
$(25, 20)$	1.40/ 0.46/ 0.77/ **0.97**	6.44/ 0.01/ 0.78/ 0.91	2.71/ 3.62/ 1.28/ 0.93	1.43/ 4.82/ 1.64/ 0.89	1.06/ 4.49/ 1.42/ 0.91
$(50, 10)$	1.50/ 0.46/ 0.50/ **0.97**	6.90/ 0.03/ 0.53/ 0.90	2.93/ 2.38/ 0.72/ 0.92	1.48/ 3.55/ 0.93/ 0.92	1.18/ 2.81/ 0.84/ 0.93
$(50, 20)$	1.09/ 0.15/ 0.35/ **0.98**	5.36/ 0.00/ 0.35/ 0.92	2.63/ 2.08/ 0.55/ 0.92	1.31/ 3.22/ 0.79/ 0.92	0.75/ 2.74/ 0.72/ 0.94
$(100, 10)$	1.13/ 0.32/ 0.25/ **0.98**	5.02/ 0.01/ 0.24/ 0.92	2.73/ 1.03/ 0.30/ 0.94	1.32/ 1.91/ 0.39/ 0.95	0.80/ 1.66/ 0.39/ 0.96
$(100, 20)$	1.06/ 0.01/ 0.18/ **0.98**	4.53/ 0.00/ 0.19/ 0.93	2.90/ 1.08/ 0.25/ 0.94	1.28/ 1.94/ 0.35/ 0.95	0.49/ 1.71/ 0.33/ 0.96

each response separately. The `lmerTest`, which used backward elimination based on hypothesis testing, had the highest number of false positives (indicative of overfitting), while backward elimination using BIC underfitted the most. Table 1 also shows the strong performance of ERIC for choosing the tuning parameter in the APLES estimator: BIC2(λ) overfitted substantially compared to ERIC, while there was little difference in the extent of underfitting between two criteria. This suggests that the dynamic, aggressive shrinkage of the latter works better when applied to the APLES estimator. Finally, as expected all methods performed better when the number of clusters n and/or cluster size n_i increased.

In the Supplementary Material, we present two additional simulations: the first involves multivariate ordinal responses resembling that of the HILDA mental health data analyzed in Section 6, and is designed to compare different methods of choosing the tuning parameter in the APLES estimator. The second simulation involves multivariate binary responses and is designed similarly to the Gaussian response case above. Overall, these results also present evidence favoring the use of the APLES estimator in combination with ERIC for performing fixed effects selection.

6 Example: Mental Health

We applied our proposed APLES estimator to the longitudinal HILDA survey introduced in Section 1, with the aim of uncovering the important factors driving an individuals' mental health response over time. The responses consisted of $K = 5$ questionnaire items comprising the Mental Health Inventory 5 (Leach et al., 2014) and were as follows: How much of the time during the past four weeks 1) have you been a nervous person? 2) have you felt so down in the dumps that nothing could cheer you up? 3) have you felt calm and peaceful? 4) have you felt down? 5) have you been a happy person? For each question, an individual gave a score from 1 (All of the time) to 6 (None of the time). We used data collected from 2006 inclusive onwards, so that information on a person's height and weight were available (used in calculating the person's body mass index); such data were not collected before 2006. This lead to $n_i = 9$ waves of data for our analyses, from 2006 to 2014. For illustration purposes, we also subset the data to only focus on individuals with no missing data on any of the five outcomes or any of the predictors across the nine waves. This resulted in a dataset with $n = 221$ individuals (clusters) and a total sample size of 1989 observations.

As fixed effects covariates, we considered $p_f = 31$ possible predictors spanning both personal e.g., age, marital status, and social e.g., frequency of social outings, predictors. The full list of 31 predictors is reported in Figure 1. All continuous predictors were standardized to have mean zero and unit variance, and all categorical predictors were converted into dummy variables. For the random effects structure, we included a random intercept and random slope for age for each individual, leading to a 10×10 random effects covariance matrix in the multivariate GLMM. Aside from age, which was not penalized since it was also included as a random slope, the other 30 covariates were available for selection for each of the five responses. We assumed a cumulative logit model for each of the five

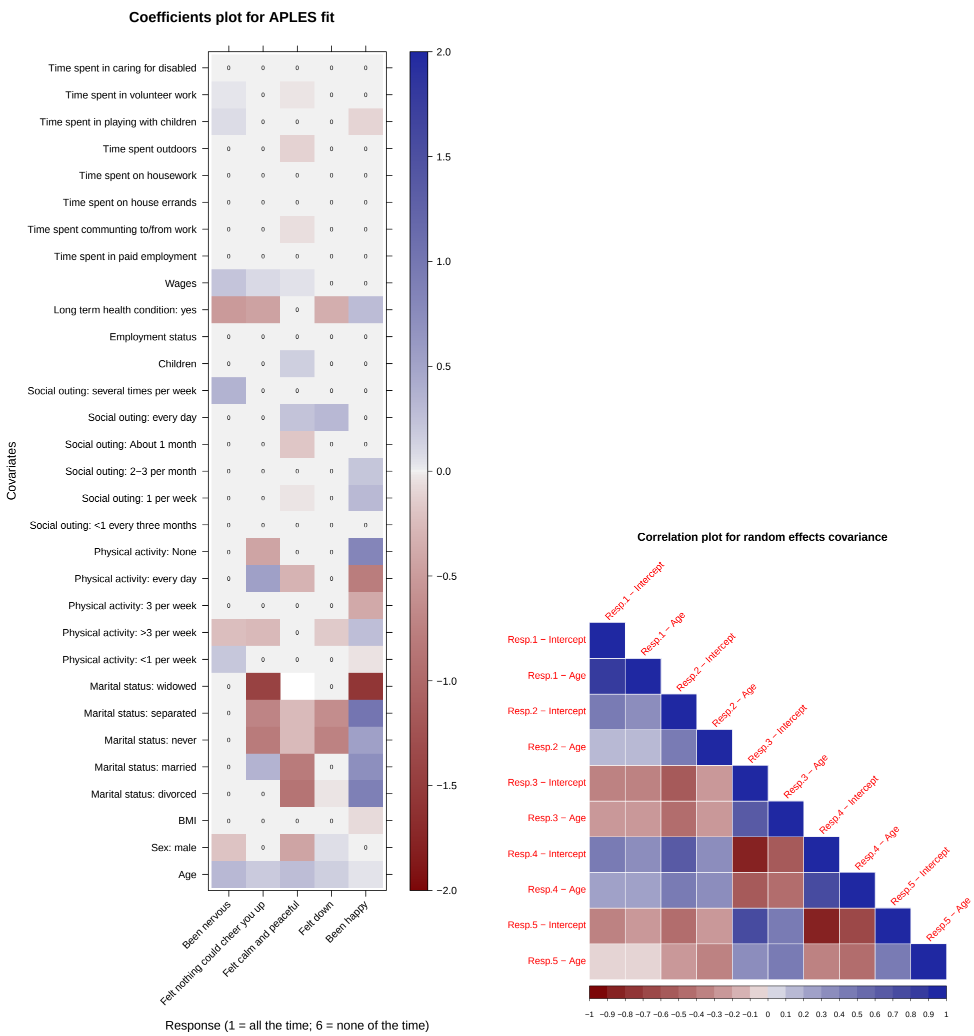
ACCEPTED MANUSCRIPT

outcomes, as described at the end of Section 2.

The results from applying the proposed APLES estimator, in conjunction with ERIC to choose the tuning parameter, are summarized in Figure 1. From the left panel, we can see that the APLES estimator produced a relatively sparse fixed effects structure: six covariates including employment status and time spent per week on housework were removed from all five outcomes, while other covariates were only important for one outcome e.g., body mass index only showed a non-zero effect only for the fifth outcome, with a higher index associated with reduced happiness. In fact, the only covariate that presented a non-zero fixed effect for all five outcomes was age, which was not penalized. Overall however, we were able to draw some important conclusions based on the selected model, as seen in the left panel of Figure 1: 1) there were weak positive associations between improved mental health and age, as well as general associations between improved mental health and being married, 2) increased physical activity was generally associated with better mental health, although the effects were strongest when physical activity was undertaken more than three times per week, 3) the presence of a long term health condition was an important predictor of poor mental health, 4) both social outings and the division of time per week to various activities were weakly associated with an individual's mental health over time.

From the right panel of Figure 1, we observe that within each response there are strong positive correlations between the random intercept and random slope for age, implying those with a healthier baseline mental health profile tended to experience more positive responses with age. Across responses however, there are some distinct correlation patterns; e.g, there were moderate negative correlations between the random intercepts/slopes for outcome 3 and those of outcomes 1 and 2, and positive correlations between outcomes 3 and 5. These pronounced cross-response correlation patterns are a reflection of strong underlying correlations between the different responses e.g., an individual who has been

Figure 1: Results from applying the APLES estimator to the mental health dataset. The left panel presents a coefficients plot with each column being one of the $K = 5$ outcomes and each row one of the $p_f = 31$ covariates. The coefficients are color-coded based on sign and magnitude, with many coefficients shrunk to zero (as indicated by a zero). The right panel presents a correlation plot based on the estimated random effects covariance matrix. The plot is ordered in terms of the responses e.g., Resp. 1 represents the first outcome (Have you been a nervous person?), with "Intercept" denoting the random intercept and "Age" being the random slope of age.



feeling calm and peaceful (outcome 3) little to none of the time would likely be feeling nervous (outcome 1) and down (outcome 2) a lot of the time, and feeling happy little of the time (outcome 5). This is also consistent with substantial between-individual variability in both baseline mental health profiles (random intercept) as well as their trajectories with age (random slope).

We also compared the results obtained from the APLES model fit to those obtained by separately fitting a ordinal GLMM to each of the five responses, using the `R` package `ordinal` (Christensen, 2015). Results for the latter are found in the Supplementary Material, and show that the separate model approach produced an even sparser model than the APLES model fit e.g., from the separate fitting approach, both social outings and marital status have very little association with mental health in general. While the reasons behind the differences in results between the two approaches are complex, we speculate that one reason might be due to the joint modeling approach having improved power and efficiency at detecting truly non-zero coefficients across the five outcomes, as compared to the separate response approach. This reasoning would be consistent with the simulation results in Section 5, where we saw that APLES underfitted less than the separate response approaches.

7 Discussion

As the collection of multivariate longitudinal data continues to grow, there is an increased demand for statistical methods capable of jointly analyzing and performing inference on such data. In this article, motivated by data following individuals' mental health over time, we focused on the challenge of selecting the important fixed effects in multivariate mixed models. We propose APLES, a joint estimation and selection approach based on constructing an approximate pairwise likelihood function and augmenting it with a penalty capable of removing fixed effects from the mean of all (or some) outcomes simultaneously. Along with proposing a new information criterion for choosing the tuning parameter that promotes aggressive shrinkage, we showed that the APLES estimator attains the oracle property, and in finite sample studies performs better than some current selection methods which analyze each response separately.

While the focus of this article has been on fixed effects selection, and we have implicitly assumed the number of random effects included in the model is not too large, a natural question to ask in future research is whether the APLES estimator can be employed to efficiently perform joint selection of fixed and random effects in multivariate GLMMs, especially if the number of fixed and/or effects is diverging. This is currently being explored; in particular, the penalty to use should exploit both the clustering of coefficients on a per-covariate basis as well as respect the hierarchical principle of fixed and random effects in longitudinal GLMMs (see Hui et al., 2017a). Another extension, particularly with motivating data on mental health, is to extended APLES to factor analytic multivariate mixed models, where a factor analysis of the outcomes is used to reduce the dimension of the responses to a small number of latent variables (potentially representing an underlying mental health score) and a multivariate GLMM fitted to this latent variables (see Verbeke et al., 2014, for a review of such models in the literature).

The approximate pairwise likelihood is an attractive basis for estimating and doing inference with multivariate mixed models. There is however much to explore in this area. For instance, one challenge with APLES was to measure the degrees of freedom with the approximate pairwise likelihood: as seen in Gao and Song (2010), simply using the number of non-zero coefficients may not work well if the number of covariates is large or exceeds sample size. More broadly, the issue of estimating the degrees of freedom and measures of model complexity for random effects remains an open and active problem in statistics (see for instance the recent research by You et al., 2016, for linear mixed models). Of course, use of the approximate pairwise likelihood presumes the bivariate models can be fitted efficiently using maximum likelihood. If not however, then perhaps alternative, faster methods of estimation could be used instead (e.g., using variational approximations, Hui et al., 2017b). The implications of using these alternative methods for estimating the bivariate models on the asymptotic and finite sample performance of the APLES estimator present an interesting challenge to explore.

Acknowledgements

This research was supported by the Australian Research Council discovery project grant DP140101259. Thanks to Wade Blanchard, Robert Clark, and Cristiano Varin for useful discussions.

Supplementary Material

Proofs of Lemmas 1 and 2, Theorem 1, additional simulation results for multivariate binary responses, and extra results for application to the mental health data may be found in the Supplementary Material. We also provide template `R` code for calculating the APLES estimator and for performing the simulations in the Supplementary Material.

References

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

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.

Bondell, H. D. and Reich, B. J. (2008). Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with OSCAR. *Biometrics*, 64:115–123.

Christensen, R. H. B. (2015). ordinal—Regression Models for Ordinal Data. R package version 2015.6-28. http://www.cran.r-project.org/package=ordinal/.

Faes, C., Aerts, M., Molenberghs, G., Geys, H., Teuns, G., and Bijnens, L. (2008). A high-dimensional joint model for longitudinal outcomes of different nature. *Statistics in Medicine*, 27:4408–4427.

Fieuws, S. and Verbeke, G. (2006). Pairwise fitting of mixed models for the joint modeling of multivariate longitudinal profiles. *Biometrics*, 62:424–431.

Fieuws, S., Verbeke, G., Boen, F., and Delecluse, C. (2006). High dimensional multivariate mixed models for binary questionnaire data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 55:449–460.

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

Gao, X. and Song, P. X.-K. (2010). Composite likelihood bayesian information criteria for model selection in high-dimensional data. *Journal of the American Statistical Association*, 105:1531–1540.

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.

ACCEPTED MANUSCRIPT

Groll, A. and Tutz, G. (2014). Variable selection for generalized linear mixed models by ℓ_1–penalized estimation. *Statistics and Computing*, 24:137–154.

Gunes, F. and Bondell, H. D. (2012). A confidence region approach to tuning for variable selection. *Journal of Computational and Graphical Statistics*, 21:295–314.

Huang, J., Breheny, P., and Ma, S. (2012). A selective review of group selection in high-dimensional models. *Statistical Science*, 27:481–499.

Hui, F. K. C., Mueller, S., and Welsh, A. H. (2016). Joint selection in mixed models using regularized PQL. *Journal of the American Statistical Association*, In press.

Hui, F. K. C., Mueller, S., and Welsh, A. H. (2017a). Hierarchical selection of fixed and random effects in generalized linear mixed models. *Statistica Sinica*, 27:501–518.

Hui, F. K. C., Warton, D. I., and Foster, S. D. (2015a). Multi-species distribution modeling using penalized mixture of regressions. *The Annals of Applied Statistics*, 9:866–882.

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

Hui, F. K. C., Warton, D. I., Ormerod, J. T., Haapaniemi, V., and Taskinen, S. (2017b). Variational approximations for generalized linear latent variable models. *Journal of Computational and Graphical Statistics*, 26:35–43.

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.

Ivanova, A., Molenberghs, G., and Verbeke, G. (2015). Fast and highly efficient pseudo-likelihood methodology for large and complex ordinal data. *Statistical Methods in Medical Research*.

Kuznetsova, A., Bruun Brockhoff, P., and Haubo Bojesen Christensen, R. (2016). *lmerTest: Tests in Linear Mixed Effects Models*. R package version 2.0-30.

Leach, L. S., Olesen, S. C., Butterworth, P., and Poyser, C. (2014). New fatherhood and psychological distress: a longitudinal study of australian men. *American Journal of Epidemiology*, 180:582–589.

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.

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44:226–233.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 42:109–142.

Molenberghs, G. and Verbeke, G. (2006). *Models for Discrete Longitudinal Data*. Springer Series in Statistics. Springer New York.

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

Müller, S. and Welsh, A. H. (2010). On model selection curves. *International Statistical Review*, 78:240–256.

Powers, D. M. W. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness & correlation. *Journal of Machine Learning Technologies*, 2:37–63.

Varin, C., Reid, N., and Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica*, 21:5–42.

Varin, C. and Vidoni, P. (2005). A note on composite likelihood inference and model selection. *Biometrika*, 92:519–528.

Vasdekis, V. G. S., Rizopoulos, D., and Moustaki, I. (2014). Weighted pairwise likelihood estimation for a general class of random effects models. *Biostatistics*, 15:677–689.

Verbeke, G., Fieuws, S., Molenberghs, G., and Davidian, M. (2014). The analysis of multivariate longitudinal data: A review. *Statistical Methods in Medical Research*, 23:42–59.

Wang, H. and Leng, C. (2012). Unified LASSO estimation by least squares approximation. *Journal of the American Statistical Association*, 102:1039–1048.

Watson, N. and Wooden, M. P. (2012). The HILDA survey: a case study in the design and development of a successful household panel survey. *Longitudinal and Life Course Studies*, 3:369–381.

You, C., Müller, S., and Ormerod, J. T. (2016). On generalized degrees of freedom with application in linear mixed models selection. *Statistics and Computing*, 26:199–210.

Zhao, P., Rocha, G., and Yu, B. (2009). The composite absolute penalties family for grouped and hierarchical variable selection. *The Annals of Statistics*, 37:3468–3497.

Zou, H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association*, 101:1418–1429.

ACCEPTED MANUSCRIPT

Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *The Annals of Statistics*, 36:1509–1533.