JMLR: Workshop and Conference Proceedings 44 (2015) 277-291 NIPS 2015

The 1st International Workshop "Feature Extraction: Modern Questions and Challenges"

Covariance Selection in the Linear Mixed Effect Model

Jonathan P Williams JPWILL@LIVE.UNC.EDU
Department of Statistics and Operations Research
University of North Carolina at Chapel Hill
Chapel Hill, NC 27514, USA

Ying Lu YING.LU@NYU.EDU
Department of Humanities and Social Sciences
New York University
New York, NY 10003, USA

Editor: Dmitry Storcheus

Abstract

This paper improves and extends the two-step penalized iterative estimation procedure for the linear mixed effect model (LMM) by explicitly penalizing the off-diagonal components of the covariance matrix of random effects. To explicitly penalize the off-diagonal terms in the covariance matrix of random effects, glasso is incorporated in the penalized LMM approach. The paper also provides theoretical justification and a computational algorithm for the provided approach. Empirical analysis using random simulated data shows that explicitly penalizing the off-diagonal covariance components can greatly improve the model selection procedure.

Keywords: covariance matrix selection, model selection, linear mixed effects model, random effects, penalized least squares, sparsity

1. Introduction

Peng and Lu (2012) proposed a two-step penalized iterative estimation procedure to select and estimate a Linear Mixed Effect Model (LMM). Their procedure is improved and extended in this paper by proposing a method which explicitly penalizes the off-diagonal components of the covariance matrix of random effects. Such an extension is crucial for being able to estimate the structure of a sparse covariance matrix, but the procedure in Peng and Lu (2012) is only able to explicitly penalize the diagonal components of the covariance matrix of random effects.

Traditionally, one uses Maximum Likelihood (ML), or Restricted Maximum Likelihood (REML) to estimate the LMM. Given normality, the maximum likelihood method yields the most efficient estimates. The ML estimation problem reduces to a constrained optimization problem which is very computationally intensive, and relies heavily on a normality assumption for the conditional likelihood function of the data. Another downside is that when the number of random effects variables is large, the ML method has computational difficulties in estimating the random effects covariance matrix. Further, the ML approach, by itself does not address the issue of model selection.

Fan and Li (2001) show, both theoretically and empirically, that PLS estimates simultaneously select and estimate simple linear regression models as if the true model were known, that is, PLS is shown to have Oracle properties. This method of model selection and estimation for the simple linear regression model can be extended to simultaneously select and estimate the LMM, and to a great advantage. A recent overview of the literature on model selection in LMM, over the past decade, is given by Muller, Scealy, and Welsh (2013). Other methods of reducing the problem of model selection include dimensionality reduction. See Mohri, Rostamizadeh, and Storcheus (2015) for a more recent paper on this topic.

As an alternative to the traditional methods, Peng and Lu (2012) proposed their two-step penalized iterative estimation procedure which simultaneously selects and estimates a LMM, and which is more robust than the ML approach because it does not rely on a normality assumption of the data. Their method also avoids the constrained optimization problem in the ML approach, and therefore allows for greater computational stability.

Use is made in this paper of the penalized *glasso* algorithm in R, from Friedman, Hastie, and Tibshirani (2008) to improve the two-step penalized iterative estimate of the covariance matrix of random effects. The resulting covariance structure can also be applied, in the form of constraints, to the classical ML/REML method, to estimate the LMM with reduced computational burden.

The outline of the paper is as follows. Section 2 provides motivation, and a brief overview of the theory of penalized least squares. It then goes on to discuss covariance estimation and selection, and the *glasso* algorithm from Friedman, Hastie, and Tibshirani (2008) which is relied on heavily in this paper. Then section 3 establishes the theory from Peng and Lu (2012) for the two-step penalized iterative estimation procedure for estimating a LMM, and finishes by outlining the extension method proposed in this paper. Lastly, brief simulation results of the proposed method are presented in section 4, and concluding remarks follow. However, the reader is strongly encouraged to see Appendix A for the full empirical analysis of this paper.

2. Penalized Least Squares for Model Selection and Parameter Estimation

Let p be the number of parameters to be estimated in a simple linear regression model, and let n be the number of observations. In the classical linear regression, $p << n$. The motivation for PLS is the scenario in which p is large, or even $p > n$, but that the model with all p parameters is believed to be sparse. This is a very viable assumption in situations in which many explanatory variables are arguably relevant to the model. Traditional model selection methods such as stepwise deletion, and AIC and BIC criterion can be time consuming, and they do not behave consistently. In contrast, PLS is a very effective model selection procedure for reducing the complexity of a simple linear regression model while still retaining desirable properties for estimating nonzero coefficients.

The basic idea of PLS is to penalize 'small' coefficient estimates on becoming larger in the minimization procedure. In such a manner, 'small' coefficients have, in some sense, an incentive for shrinking in magnitude to zero. The definition of 'small' will become clear in the sections which follow, but for now it can be said that it depends on a data-driven

thresholding parameter. Two well-known thresholding functions are LASSO and SCAD. LASSO, with it's L_1 penalty function is effective for handling 'small' coefficients estimates with PLS estimation, but it inherently shrinks 'large' coefficient estimates. Conversely, the SCAD thresholding function is able to treat 'small' coefficient estimates in the same desirable way that LASSO does, but also leaves 'large' estimates unbiased.

Fan and Li (2001) give the following three properties which estimators using a good penalty function should have.

(1) Unbiasedness of the large, truly nonzero parameter estimates.

(2) Sparsity of the estimates. That is, 'small' parameter estimates are set to zero to reduce model complexity.

(3) Continuity of the estimator over the parameter space.

To understand these properties and their origins, see the discussion in Fan and Li (2001).

Perhaps the most well known and simple penalty function is the L_1 penalty function used for LASSO to estimate penalized least squares. However with the exception of simplicity, the SCAD penalty function is more desirable than the L_1 penalty function for use in estimating PLS. The SCAD penalty function satisfies all three of the properties of a good penalty function from the previous section, whereas the L_1 penalty function only satisfies two; it does not satisfy the unbiasedness property for large parameter estimates. For this reason, and since SCAD is represented less extensively in the literature, the major emphasis of this study will rely on PLS estimates using the SCAD penalty function. For a more complete discussion of penalty functions see Fan and Li (2001).

2.1 Covariance Estimation and Selection

The estimation of the covariance matrix of random effects is particulary important in estimating the LMM. In section 3.2, within the iterative procedure of Peng and Lu (2012), a penalty function is used to estimate both the fixed effects, and the covariance of random effects. However, to estimate this covariance matrix, only the variance (diagonal) terms are explicitly being penalized, and the covariance (off-diagonal) terms between each pair of variables are only set to zero if the variances of either of the corresponding variables is set to zero. Thus, a limitation of the penalized two-step iterative method is that the covariance (off-diagonal) terms are never explicitly penalized. To remedy this limitation, and thereby extend the penalized estimation approach taken in Peng and Lu (2012), the aim of this paper is ultimately to consider options for penalizing the nonzero covariance (off-diagonal) terms.

A method for covariance selection and estimation is proposed in Friedman, Hastie, and Tibshirani (2008). Namely, they propose the Graphical LASSO algorithm, or GLASSO, and their paper also includes an R package/function which they built for the algorithm, called *glasso*. For simplicity, throughout the rest of this paper, *glasso* will denote the R package/function, and will also refer directly to the proposed method. More work remains to better understand the *glasso* algorithm, but here, the basic ideas will be presented.

Under the assumption of normality, the precision matrix can be found by minimizing the penalized log-likelihood of the function of the data, given by

$$\log\det\left(\Sigma^{-1}\right) - \mathrm{tr}\left(S\Sigma^{-1}\right) - \rho\left\|\Sigma^{-1}\right\|_1.$$

The L_1 norm is the sum of the absolute values of the matrix, and ρ is a thresholding parameter associated with the penalty function. Alternatively, this optimization problem can be solved for Σ rather than Σ^{-1} by optimizing over each row and corresponding column of $W = \widehat{\Sigma}$ (some estimate of Σ) in the following block coordinate descent fashion (Friedman, Hastie, and Tibshirani (2008)). Partition

$$\widehat{\Sigma} = W = \begin{pmatrix} W_{11} & w_{12} \\ w_{12}^T & w_{22} \end{pmatrix}, \quad S = \begin{pmatrix} S_{11} & s_{12} \\ S_{12}^T & s_{22} \end{pmatrix}.$$

The solution for w_{12} is given by the y which satisfies

$$\min_y \left\{ y^T W_{11}^{-1} y : \|y - s_{12}\|_\infty \leq \rho \right\},$$

which is shown to be equivalent to $w_{12} = W_{11}\beta$, where β satisfies

$$\min_\beta \left\{ \frac{1}{2}\left\| W_{11}^{\frac{1}{2}}\beta - W_{11}^{-\frac{1}{2}} s_{12} \right\|^2 + \rho\|\beta\|_1 \right\}. \tag{1}$$

All of the off-diagonal terms can be solved for by permuting the rows and columns of W and S. The actual *glasso* algorithm begins with $W = S + \rho I_p$.

The LASSO problem in equation (1) is solved for all permutations, and the solutions $w_{12} = W_{11}\widehat{\beta}$ are filled into the corresponding row and column of $W = \widehat{\Sigma}$. This procedure is continued until convergence. The reader is invited to consult Friedman, Hastie, and Tibshirani (2008) for further details. Their paper advocates for the algorithm because it is a simple and remarkably fast algorithm for estimating a sparse inverse covariance matrix. It is used in this paper to refine the estimate of the covariance matrix of random effects – from the penalized two-step iterative estimation procedure which will be presented in the sections that follow.

3. Linear Mixed Effect Model and Penalized Linear Mixed Effect Model

The LMM will be established and estimated using the distribution-free, iterative procedure from Peng and Lu (2012). For each cluster i, $1 \leq i \leq m$ as

$$Y_i = X_i\vec{\beta} + Z_i\vec{b}_i + \varepsilon_i.$$

Each Y_i is a $n_i \times 1$ column vector, X_i is a $n_i \times p$ design matrix corresponding to the fixed effects, $\vec{\beta}$ is a $p \times 1$ vector of fixed effects parameters, Z_i is a $n_i \times q$ design matrix corresponding to random effects, $\vec{b}_i$ is a $q \times 1$ vector of random effects parameters, and ε_i is a $n_i \times 1$ vector of random errors which are independent of X_i, Z_i, and $\vec{b}_i$. The distributional assumptions of the model are the following.

$$\varepsilon_i \sim N(0, \sigma^2 I_{n_i})$$
$$\vec{b}_i \sim N(0, \sigma^2 D)$$

where D is a $q \times q$ symmetric nonnegative definite matrix. Thus,

$$Y_i \sim N\left(X_i\vec{\beta}, \sigma^2(Z_i D Z_i^T + I_{n_i})\right).$$

3.1 Two-Step Iterative Estimation Procedure

To describe the estimation procedure for the fixed and random effects using the two-step iterative procedure, as in Peng and Lu (2012) first let $Y = (Y_1^T, Y_2^T, \ldots, Y_m^T)^T$, $\vec{b} = (\vec{b_1}^T, \vec{b_2}^T, \ldots, \vec{b_m}^T)^T$, $\varepsilon = (\varepsilon_1^T, \varepsilon_2^T, \ldots, \varepsilon_m^T)^T$ be $n \times 1$ column vectors, $X = (X_1^T, X_2^T, \ldots, X_m^T)^T$ a $n \times p$ matrix, and let $Z = \text{diag}\{Z_1, Z_2, \ldots, Z_m\}$ be a $n \times qm$ block diagonal matrix, where $n = \sum_{i=1}^{m} n_i$. For an initial estimate, ignore the random effects and begin with the simple linear regression model. The initial estimate of the fixed effects coefficients is then given by OLS, which can be expressed using the clustered structure of the data in the following from.

$$\widehat{\vec{\beta}} = \vec{\beta}_0 = (X^T X)^{-1} X^T Y = \left(\sum_{i=1}^{m} X_i^T X_i \right)^{-1} \sum_{i=1}^{m} X_i^T Y_i$$

With this initial estimate, recall that

$$Y_i = X_i \vec{\beta} + Z_i \vec{b}_i + \varepsilon_i \quad \Longrightarrow \quad \underbrace{Y_i - X_i \vec{\beta}}_{=: u_i} = Z_i \vec{b}_i + \varepsilon_i. \tag{2}$$

Then estimate the random effect for each i, $1 \leq i \leq m$ by least squares as

$$\begin{aligned} &\widehat{\vec{b}}_i = (Z_i^T Z_i)^{-1} Z_i^T \widehat{u}_i \ \text{ and} \\ &\widehat{\varepsilon}_i = \widehat{u}_i - Z_i \widehat{\vec{b}}_i. \end{aligned} \tag{3}$$

With theses estimates, Peng and Lu (2012) propose the following estimators of σ^2 and D.

$$\widehat{\sigma}^2 = \frac{\sum_{i=1}^{m} \widehat{\varepsilon}_i^T \widehat{\varepsilon}_i}{n - qm}. \tag{4}$$

To obtain the estimator for D, observe that $D = \text{E}\Big[\frac{1}{m\sigma^2} \sum_{i=1}^{m} \vec{b}_i \vec{b}_i^T\Big] \approx \frac{1}{m\sigma^2} \sum_{i=1}^{m} \vec{b}_i \vec{b}_i^T$. So for an estimator of D, use $\widehat{\vec{b}}_i$ as an estimator of $\vec{b}_i$, $1 \leq i \leq m$. However, recall the equation

$$\widehat{\vec{b}}_i = (Z_i^T Z_i)^{-1} Z_i^T u_i = \vec{b}_i + (Z_i^T Z_i)^{-1} Z_i^T \varepsilon_i.$$

So

$$\begin{aligned}
\sum_{i=1}^{m} \widehat{\vec{b}}_i \widehat{\vec{b}}_i^T &= \sum_{i=1}^{m} \big(\vec{b}_i + (Z_i^T Z_i)^{-1} Z_i^T \varepsilon_i\big)\big(\vec{b}_i + (Z_i^T Z_i)^{-1} Z_i^T \varepsilon_i\big)^T \\
&= \sum_{i=1}^{m} \big[\vec{b}_i \vec{b}_i^T + (Z_i^T Z_i)^{-1} Z_i^T \varepsilon_i \vec{b}_i^T \\
&\qquad + \vec{b}_i \varepsilon_i^T Z_i (Z_i^T Z_i)^{-1} + (Z_i^T Z_i)^{-1} Z_i^T \varepsilon_i \varepsilon_i^T Z_i (Z_i^T Z_i)^{-1}\big] \\
&\approx \sum_{i=1}^{m} \big[\vec{b}_i \vec{b}_i^T + (Z_i^T Z_i)^{-1} Z_i^T \varepsilon_i \varepsilon_i^T Z_i (Z_i^T Z_i)^{-1}\big] \quad \text{Peng and Lu (2012)} \\
\implies \sum_{i=1}^{m} \vec{b}_i \vec{b}_i^T &\approx \sum_{i=1}^{m} \big[\widehat{\vec{b}}_i \widehat{\vec{b}}_i^T - (Z_i^T Z_i)^{-1} Z_i^T \varepsilon_i \varepsilon_i^T Z_i (Z_i^T Z_i)^{-1}\big] \\
&\approx \sum_{i=1}^{m} \big[\widehat{\vec{b}}_i \widehat{\vec{b}}_i^T - \sigma^2 (Z_i^T Z_i)^{-1}\big].
\end{aligned}$$

Thus,

$$D \approx \widehat{D} = \frac{\sum_{i=1}^{m} \widehat{\vec{b}}_i \widehat{\vec{b}}_i^T}{m\widehat{\sigma}^2} - \frac{\sum_{i=1}^{m} (Z_i^T Z_i)^{-1}}{m}. \tag{5}$$

With this estimate of D, re-estimate the fixed effects using weighted least squares. That is,

$$\widehat{\vec{\beta}} = \big(X^T W X\big)^{-1} X^T W Y, \tag{6}$$

where $W = \text{diag}\big\{(Z_1 \widehat{D} Z_1^T + I_{n_1})^{-1}, \ldots, (Z_m \widehat{D} Z_m^T + I_{n_m})^{-1}\big\}$. Finally, iterate between (2)-(6) until convergence in $\widehat{\vec{\beta}}, \widehat{\sigma}^2$, and $\widehat{D}$.

3.2 Penalized Two-Step Iterative Estimation Procedure

In this section, the two-step iterative estimation procedure (in the previous section) will be modified as in Peng and Lu (2012) to use PLS estimation. The two-step iterative estimation procedure is a well-behaved and robust estimation procedure for the LMM, but it is not a method for model selection. However, it will be shown here that by modifying this estimation procedure to use PLS estimation it becomes a great procedure for estimation as well as selection of the LMM. See Fan and Li (2001) for the heart of the PLS theory which will be relied on here. The iterative estimation procedure for the LMM will remain largely unchanged from the previous section, but the estimation solutions will be different. Recall that to get an estimate of D (the covariance matrix of random effects) in the previous section, first each of the vectors of random effects parameters, $\vec{b}_i$, are explicitly estimated from the data for every cluster, $1 \leq i \leq m$. To estimate the vectors of random effects parameters, recall from equation (2), that

$$u_i = Z_i \vec{b}_i + \varepsilon_i, \quad \text{where} \quad u_i = Y_i - X_i \vec{\beta}.$$

So the vector of random effects parameters can be obtained from a simple linear regression model that regresses the model errors, assuming no random effects, on the random effect

variables. Assuming the ε_i are well behaved, this is an OLS estimation problem, and thus PLS can be implemented at this stage of the iterative estimation procedure to penalize the random effect coefficient estimates. As a result of using PLS to estimate the random effect parameters, corresponding 'small' variance terms in the estimated covariance matrix $\widehat{D}$ will be shrunk (in magnitude) to zero. Additionally, if it happens through the estimation that one of diagonal components of $\widehat{D}$ is negative in value, then this term will be set equal to zero because the diagonal components of a covariance matrix correspond to variances which can only be nonnegative in value. As mentioned in (Pend and Lu 2012), the only case when there is a nonzero probability for one of the diagonal components of $\widehat{D}$ to be negative valued is when the corresponding true parameter in D is equal to zero. In fact, when the variance of any of the random effects parameters (the diagonal components of D) is estimated to be zero, the corresponding row and column of $\widehat{D}$ will be set equal to zero. The reason being that if a particular random effect has no variance, then it is no longer considered random, and thus it's covariance with the other random effects is meaningless. Accordingly, the actual implementation of the penalty function will be on the diagonal components of $\widehat{D}$.

In similar notation to Peng and Lu (2012) let $c_k = \sqrt{D_{kk}}$, where D_{kk} is the kth diagonal component of D, and obtain the PLS estimates by minimizing

$$G(\vec{b}_i) = \frac{1}{2}(u_i - Z_i\vec{b}_i)^T(u_i - Z_i\vec{b}_i) + n\sum_{k=1}^{q} p_\xi(c_k),$$

for $1 \le i \le m$. Observe that each c_k is a function of the b_i because

$$D \approx \frac{1}{m\sigma^2}\sum_{i=1}^{m} \vec{b}_i\vec{b}_i^T.$$

Here, for the reasons discussed in section Fan and Li (2001), the SCAD penalty function will be chosen. Then as in Fan and Li (2001) and Peng and Lu (2012), the following second order approximation is used for the SCAD penalty function where for some initial estimate $c_{0k} \approx c_k$

$$p_\xi(c_k) \approx p_\xi(c_{0k}) + \frac{p'_\xi(c_{0k})}{2c_{0k}}\big(c_k^2 - c_{0k}^2\big),$$

and the second order approximation for G is made as

$$G(\vec{b}_i) \approx f(\vec{b}_{0i}) + \nabla f(\vec{b}_{0i})(\vec{b}_i - \vec{b}_{0i}) + \frac{1}{2}(\vec{b}_i - \vec{b}_{0i})^T\nabla^2 f(\vec{b}_{0i})(\vec{b}_i - \vec{b}_{0i}) + \underbrace{\frac{n}{2}\vec{b}_i^T\Sigma_\lambda(\vec{c}_0)\vec{b}_i}_{(\bigstar)},$$

where $\vec{b}_{0i}$ is an initial estimate in a neighborhood of $\vec{b}_i$, and $\vec{c}_0$ is the vector of initial estimates c_{0k}, $1 \le k \le q$. In this case, $f(\vec{b}_i) = \frac{1}{2}(u_i - Z_i\vec{b}_i)^T(u_i - Z_i\vec{b}_i)$. The difference in minimizing G for random effects versus for fixed effects is in the last term of the approximation. That is,

$$(\bigstar)\quad n\sum_{k=1}^{q} p_\xi(c_k) \approx n\sum_{k=1}^{q} \frac{p'_\xi(c_{0k})}{2c_{0k}}c_k^2 \quad \text{(up to a constant)}.$$

Since $G(\vec{b}_i)$ must be minimized with respect to $\vec{b}_i$, it will be helpful to express c_k in terms of $\vec{b}_i$. By construction,

$$\begin{aligned} c_k^2 = D_{kk} \approx \left[\frac{1}{m\sigma^2}\sum_{i=1}^{m}\vec{b}_i\vec{b}_i^T\right]_{kk} &= \left[\frac{1}{m\sigma^2}\sum_{i=1}^{m}(b_{i1},\ldots,b_{iq})^T(b_{i1},\ldots,b_{iq})\right]_{kk} \\ &= \left[\frac{1}{m\sigma^2}\sum_{i=1}^{m}\begin{pmatrix} b_{i1}^2 & b_{i1}b_{i2} & \cdots & b_{i1}b_{iq} \\ b_{i2}b_{i1} & b_{i2}^2 & \cdots & b_{i2}b_{iq} \\ \vdots & \vdots & \ddots & \vdots \\ b_{iq}b_{i1} & b_{iq}b_{i2} & \cdots & b_{iq}^2 \end{pmatrix}\right]_{kk} \\ \implies \quad c_k^2 &\approx \frac{1}{m\sigma^2}\sum_{i=1}^{m}b_{ik}^2, \end{aligned}$$

so

$$(\bigstar) \quad n\sum_{k=1}^{q}p_\xi(c_k) \approx n\sum_{k=1}^{q}\left(\frac{p'_\xi(c_{0k})}{2m\sigma^2 c_{0k}}\sum_{i=1}^{m}b_{ik}^2\right) = \sum_{i=1}^{m}n\sum_{k=1}^{q}\frac{p'_\xi(c_{0k})}{2m\sigma^2 c_{0k}}b_{ik}^2.$$

Then since $G(\vec{b}_i)$ will be differentiated with respect to a given $i \in \{1,\ldots,m\}$, the $\vec{b}_j$ terms for $j \neq i$ will vanish. Hence, the following simplification is appropriate.

$$\begin{aligned} (\bigstar) \quad & n\sum_{k=1}^{q}p_\xi(c_k) \approx n\sum_{k=1}^{q}\frac{p'_\xi(c_{0k})}{2m\sigma^2 c_{0k}}b_{ik}^2 = \frac{n}{2}\left(\frac{p'_\xi(c_{01})}{m\sigma^2 c_{01}}b_{i1}^2 + \cdots + \frac{p'_\xi(c_{0q})}{m\sigma^2 c_{0q}}b_{iq}^2\right) \\ \implies \quad & n\sum_{k=1}^{q}p_\xi(c_k) \approx \frac{n}{2}\vec{b}_i^T\underbrace{\operatorname{diag}\left\{\frac{p'_\xi(c_{01})}{m\sigma^2 c_{01}},\ldots,\frac{p'_\xi(c_{0q})}{m\sigma^2 c_{0q}}\right\}}_{=\Sigma_\xi(\vec{c}_0)}\vec{b}_i. \end{aligned}$$

It now follows that the closed form PLS solution can be solved for using the Newton-Raphson algorithm. That is,

$$\vec{b}_i^* = \left(Z_i^T Z_i + n\Sigma_\xi(\vec{c}_0)\right)^{-1}Z_i^T\widehat{u}_i, \quad \text{for } 1 \leq i \leq m,$$

can be iterated until convergence to yield the penalized random effects coefficient estimates. Note that $\widehat{u}_i$ can be estimated initially using an OLS estimate for $\vec{\beta}$, say $\vec{\beta}_{LS}$. And, as in the previous section, $\widehat{D}$ can be updated as in (5), which is

$$D^* = \frac{\sum_{i=1}^{m}\vec{b}_i^*(\vec{b}_i^*)^T}{m\sigma^{2*}} - \frac{\sum_{i=1}^{m}(Z_i^T Z_i)^{-1}}{m}.$$

Recall that $Y_i \sim N\left(X_i\vec{\beta}, \sigma^2(Z_i D Z_i^T + I_{n_i})\right)$. To take advantage of the improved estimate of D, and thus the improved estimate of $\operatorname{cov}(Y_i)$, update $\vec{\beta}_{LS}$, as in the penalized setting, using Weighted Least Squares. Accordingly,

$$\vec{\beta}^* = \left(X^T W X + n\Sigma_\lambda(\vec{\beta}_0)\right)^{-1}X^T W Y,$$

where $W = \operatorname{diag}\left\{(Z_1 D^* Z_1^T + I_{n_1})^{-1},\ldots,(Z_m D^* Z_m^T + I_{n_m})^{-1}\right\}$. With these improved estimates of $\vec{\beta}^*$ and D^*, iterate until convergence, as in the previous section.

Now that PLS estimation has been discussed within the LMM setting, what remains is to discuss the selection of the tuning parameters λ and ξ. See the appendices for this discussion.

3.3 Incorporate *glasso* to Select Off-Diagonal Covariance Terms

The main objective of this paper is to extend the penalized LMM approach to explicitly penalize the off-diagonal terms in the covariance matrix of random effects. Here the estimated covariance matrix of random effects from the penalized two-step iterative procedure, $\widehat{D}$, is used to construct the initial covariance matrix estimate, S, as in section (2.1). S is given by the nonzero sub-matrix of $\widehat{D}$. Since rows and columns of the penalized covariance estimate are simultaneously shrunk to zero when their corresponding diagonal component is shrunk to zero, by sub-matrix of $\widehat{D}$ it is meant $\widehat{D}$ with the zero rows/columns removed. Beginning with S, the backward procedure is to start with all off-diagonal elements of S, and then drop the elements component-wise, whose absence do not significantly diminish the fit of the model.

In a penalized framework, the backward procedure seems most appropriate, and instead of dropping elements component-wise, the off-diagonal terms could rather be shrunk iteratively using a penalty thresholding rule such as SCAD or LASSO. Accordingly, to extend the penalized LMM method, the *glasso* algorithm is applied to the nonzero sub-matrix of $\widehat{D}$. A normality assumption is used for the data to accommodate the *glasso* algorithm, but recall that the penalized LMM does not rely on this assumption.

4. Simulation Results

Presented here are simulation results from estimating the LMM using the penalized two-step iterative estimation procedure, and using the *glasso* algorithm to refine the estimated covariance matrix of random effects. The procedure is repeated for 200 samples of randomly generated data for a LMM with the following true vector of fixed effects coefficients, and true covariance matrix of random effects.

$$\beta = (-3.3, 1, -7, 4.6, 0, 0, 0, 0, 0, 0, 0, 0)^T, \quad \text{and} \quad D = \begin{pmatrix} 9 & 5 & 0 & 0 & 0 & 0 \\ 5 & 7 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & -1.5 & 0 & 0 \\ 0 & 0 & -1.5 & 4 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

To best accommodate the estimation procedure the covariance has a block diagonal structure, $m = 60$, and $n_i = 14$ for all $1 \leq i \leq 60$. The focus here is on the performance of the estimation procedure in more of an ideal situation. If there is no evidence of desirable results in an ideal situation, then there would be little motivation to further study the proposed estimation procedure.

Table 1 displays a measure of the added improvement of explicitly penalizing the covariance (off-diagonal) terms of the covariance matrix. Most notably, the proportion of off-diagonal elements which are incorrectly estimated to be nonzero is reduced by just over twenty compared to not using *glasso*, regardless of the criterion used. Recall that, unless

the variance (diagonal) term has been estimated to be zero in the penalized estimation procedure, the covariance (off-diagonal) terms are not shrunk to zero. However, about ten percent of the time, the *glasso* algorithm will incorrectly shrink off-diagonal terms to zero, regardless of the criterion used.

The reader is strongly encouraged to see Appendix A for the full empirical analysis carried out in this paper, in which the improvements of this procedure are more fully illustrated.

Tuning	FPR	glFPR	FNR	glFNR
GCV	0.31	0.09	0.00	0.09
AIC	0.32	0.08	0.00	0.09
BIC	0.31	0.09	0.00	0.07

Table 1: Performance of the penalized two-step iterative estimation procedure in selecting the correct random effects covariance structure. 'FPR' is the false positive rate, that is, the proportion of off-diagonal coefficients that are incorrectly estimated to be nonzero. 'FNR' is the false negative rate, that is, the proportion of off-diagonal coefficients that are incorrectly estimated to be zero. The labels 'glFPR', and 'glFNR' denote the corresponding rate using the glasso algorithm from Friedman, Hastie, and Tibshirani (2008).

5. Conclusion

The two-step penalized iterative estimation procedure proposed by Peng and Lu (2012) does not explicitly select the off-diagonal terms of the covariance matrix of random effects in a the LMM. A method was proposed in this paper to improve and extend their procedure by applying the *glasso* algorithm to select the off-diagonal terms of the covariance matrix. The empirical evidence suggests that implementation of the penalized *glasso* algorithm in R, from Friedman, Hastie, and Tibshirani (2008) greatly improves the two-step penalized iterative estimate of the covariance matrix of random effects from Peng and Lu (2012) when the true covariance matrix has a block diagonal form.

It is of interest in further research to better understand theoretically how these procedures perform when estimating a LMM of which both the number of fixed and random effects covariates are greater than the smallest cluster size in the data, that is, when $p > \min_{1\leq i\leq m} n_i$ and $q > \min_{1\leq i\leq m} n_i$. Currently, the procedure from Peng and Lu (2012) can only handle the case when $p > \min_{1\leq i\leq m} n_i$, but requires $q < \min_{1\leq i\leq m} n_i$. However, theory suggests that their penalized LMM procedure should be able to handle the case when $p > \min_{1\leq i\leq m} n_i$ and $q > \min_{1\leq i\leq m} n_i$.

References

E. Demidenko. *Mixed Models: Theory and Applications*. John Wiley and Sons, Inc., 2004. ISBN 9780471601616.

A. P. Dempster. Covariance selection. *Biometrics*, 28(1):157–175, 1972.

J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properities. *American Statistical Association*, 96(456):1348–1360, 2001.

J. Fan, Y. Feng, and Y. Wu. Network exploration via the adaptive lasso and scad penalties. *The Annals of Applied Statistics*, 3(2):521–541, 2009.

J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

J. S. Hodges and D. J. Sargent. Counting degrees of freedom in hierarchical and other richly-parameterized models. *Biometrika*, 88(2):367–379, 2001.

C. Lam and J. Fan. Sparsistency and rates of convergence in large covariance matrix estimation. *The Annals of Statistics*, 37(6B):4254–4278, 2009.

M. Mohri, A. Rostamizadeh, and D. Storcheus. Foundations of coupled nonlinear dimensionality reduction. *arXiv preprint arXiv:1509.08880v2*, 2015.

S. Muller, J. L. Scealy, and A. H. Welsh. Model selection in linear mixed models. *Statistical Science*, 28(2):135–167, 2013.

H. Peng and Y. Lu. Model selection in linear mixed effect models. *Journal of Multivariate Analysis*, 109(11):109–129, 2012.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, 58(1):267–288, 1996.

H. Wang, R. Li, and C. L. Tsai. Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika*, 94(3):553–568, 2007.

Appendix A. Simulation Results

The purpose of this section is to provide a more full empirical analysis of the proposed methods and procedures presented in this paper. Simulated data for the LMM is randomly generated using the random number generators in R. m clusters of $n_i \times p$ design matrices $X_1, \ldots, X_m$ are generated whose elements come from the Uniform$[-2, 2]$ distribution. The columns of the X_i are the fixed effect explanatory variables. The random effect variables compose the columns of the $n_i \times q$ matrices, $Z_1, \ldots, Z_m$. In this setup, the q columns of the Z_i taken as the first q columns of the corresponding X_i, if $q <= p$, else the remaining $q - p$ columns are generated the same as they were for the fixed effect covariates. To include an intercept term, the first column of each X_i and Z_i is replaced with a column of ones. For simplicity rename $p = p + 1$ and $q = q + 1$. Then, for $1 \leq i \leq m$, a $q \times 1$ column vector

of random effect coefficients is generated as $b_i \sim N(0, \sigma^2 D)$, and a $n_i \times 1$ column vector of random errors is generated as $\varepsilon_i \sim N(0, \sigma^2 I_{ni})$, where $\sigma^2 = 1$. Thus, for each cluster, $1 \leq i \leq m$, the observed dependent variable data is given as

$$Y_i = X_i\beta + Z_i b_i + \varepsilon_i,$$

which implies that the true model structure of the data is a LMM.

Recall that the penalized two-step iterative estimation procedure is a model selection and estimation procedure for the LMM. While it is able to select the random effect covariates in a LMM, it does this by only penalizing the variance (diagonal) terms of the covariance matrix of random effects. It is not, however, able to select the full covariance structure in the matrix. To do so, there must be a means of penalizing the covariance (off-diagonal) terms separately in the covariance matrix of random effects. As in Section 2.1, a simple extension of the penalized iterative procedure, worth exploring, is to apply the *glasso* algorithm to the nonzero submatrix of the estimated covariance matrix. As discussed in Section 2.1, this method is a step in the direction of effectively selecting the covariance structure of the random effects in the LMM. Currently, more work needs to be done to further understand the procedures used in the *glasso* R package, but as discussed in the package documentation, the algorithm used works best for covariance matrices with a block diagonal structure.

To give a sense of the performance of the model selection in the simulation setup of section 4, before the *glasso* algorithm is applied, see Table 2. The penalized two-step iterative estimation procedure does a very nice job of model selection for both the fixed and random effects.

Tuning	Correct	CF	CR
GCV	0.88	0.88	1.00
AIC	0.84	0.85	0.97
BIC	0.90	0.90	1.00

Table 2: Determining the proportion of times that the penalized two-step iterative estimation procedure selects the correct model. 'Correct' is the proportion of times that both the correct fixed and random effects were selected. 'CF' is the proportion of times that the correct fixed effects were selected. 'CR' is the proportion of times that the correct random effects were selected. Compare to Table 2 in Peng and Lu (2012).

To more directly evaluate the estimation properties of the penalized two-step iterative estimation procedure, consider Table 3 which presents the biases and Median Absolute Deviation (MAD) of the estimated fixed effects coefficients and random effects variances (for the nonzero parameters). In the table, the biases and MAD's can be compared to that of the non-penalized two-step iterative estimation procedure, and the ML estimates (using the *lmer* function of the *lme4* package for R), where only the data corresponding to the true model is used for both of these estimators. That is, the estimates from the penalized estimation procedure, under the sparse model, are compared to those of the other estimation procedures, under the true model. In this fashion, the penalized estimation procedure is not given a handicap, and is subject to more rigorous evaluation of it's estimation performance.

Note, that the *glasso* algorithm has no effect on these biases and MAD's. It appears overwhelmingly clear from Table 3 that no estimation method has out-performed all others, in this 200 sample simulation. This provides strong empirical evidence that, at least in a larger sized sample setting, the estimation properties of the penalized two-step iterative estimation procedure, under the sparse model, perform on par with the ML estimates, under the true model. Recall that the ML estimates are the most efficient estimates. This evidence, along with the evidence in Table 2 suggesting very good performance of model selection, strongly supports the theory that the penalized estimation procedure does an excellent job of model selection, and does not compromise any estimation performance, in a large sample setting. These results support the results of Peng and Lu (2012).

Once the penalized two-step iterative estimate of the covariance matrix of random effects has been estimated, for each of the 200 randomly generated samples of data, the *glasso* algorithm is then used to further refine the estimates of the covariance (off-diagonal) terms of the matrix. The algorithm should shrink 'small' estimates to zero, and it requires looping over a one-dimensional grid of tuning parameters (a range of six values is used here). The last two tables (in addition to table 1 in section 4) evaluate the additional selection performance of the *glasso* algorithm.

	Bias					MAD				
	GCV	AIC	BIC	IterO	MLEO	GCV	AIC	BIC	IterO	MLEO
β_1	-0.07	-0.09	-0.06	-0.04	-0.04	0.31	0.32	0.31	0.31	0.31
β_2	-0.06	-0.09	-0.05	-0.02	-0.02	0.24	0.24	0.23	0.23	0.23
β_3	0.01	0.01	0.01	0.01	0.01	0.14	0.14	0.13	0.13	0.13
β_4	0.02	0.02	0.02	0.02	0.02	0.21	0.21	0.21	0.21	0.21
D_{11}	-0.15	-0.11	-0.16	-0.16	-0.06	1.07	1.07	1.06	1.06	1.11
D_{22}	0	0.05	-0.01	-0.02	0.08	0.93	0.95	0.92	0.9	0.83
D_{33}	-0.01	-0.01	-0.01	-0.01	0.02	0.29	0.3	0.29	0.29	0.29
D_{44}	-0.05	-0.04	-0.05	-0.05	0	0.54	0.54	0.54	0.54	0.51

Table 3: Observe the average bias and the Median Absolute Deviation (MAD) of the estimated fixed effects coefficients, and random effects variances. 'IterO' refers to the non-penalized two-step iterative estimation procedure under the true model, and 'MLEO' refers to the MLE estimates under the true model. Compare to Table 4 in Peng and Lu (2012).

Table 4 reports the proportion of times that the exact structure of the true covariance matrix of random effects has been selected. That is, the position and number of zero and nonzero terms in the estimated matrix is identical to that of the true matrix. Since the penalized estimation procedure, alone, has no way of shrinking covariance (off-diagonal) terms to zero, it is never able to select the exact structure of the covariance matrix. However, with the refined matrix estimates, using the *glasso* algorithm, the exact structure of the covariance matrix of random effects is chosen for just over half of the 200 randomly generated samples. In such a large sample setting, it would be hoped that this performance would be much better, especially given that the true covariance matrix was chosen to be block diagonal to accommodate the *glasso* algorithm. But still, these results are promising, and

much more work is needed to better understand this algorithm from Friedman, Hastie, and Tibshirani (2008).

Tuning	Correct	glCorrect
GCV	0.00	0.55
AIC	0.00	0.56
BIC	0.00	0.56

Table 4: 'Correct' is the proportion of times that the correct covariance structure (off-diagonal) of the random effects was selected. 'glCorrect' is the proportion of times that the correct covariance structure (off-diagonal) of the random effects was selected, using the glasso algorithm from Friedman, Hastie, and Tibshirani (2008).

	GCV	AIC	BIC	glGCV	glAIC	glBIC	IterO	MLEO
Prediction Error	0.71	0.71	0.71	0.71	0.71	0.71	0.71	53.54

Table 5: Average prediction error for each criterion. The prediction error is computed as the sum of squared error divided by the sample size. The labels 'glGCV','glAIC', and 'glBIC' denote the corresponding criterion statistic using the glasso algorithm from Friedman, Hastie, and Tibshirani (2008).

Finally, Table 5 displays the prediction error of the penalized two-step iterative estimation procedure with and without the application of the *glasso* algorithm, and, under the true model, of the non-penalized two-step iterative estimation procedure and of the ML estimates. The *glasso* algorithm has no effect on the predictive performance of the estimates because it does not directly effect the b_i vectors of random effect coefficients which have been explicitly estimated prior to implementing *glasso*. In future research, it will be useful to incorporate the penalized off-diagonal estimate of the covariance matrix of random effects into the procedure for estimating the b_i. A major advantage of explicitly estimating the b_i is observed dramatically, by contrast to the much larger prediction error for the ML estimates (which were obtained using only data from the true model). This sharp improvement in prediction power over that of the ML estimates is a very favorable property of the penalized iterative procedure, although, only the large sample setting is considered here. Observe also that the penalized procedure under the sparse model performs just as well, predictively speaking, as the non-penalized procedure under the true model.

Appendix B. Tuning Paramerters

For the fixed effects tuning parameter, λ, Theorem 3.1 from Peng and Lu (2012) states that under certain regularity conditions, if

$$\lambda_n \longrightarrow 0 \ \text{ and } \ \sqrt{n}\lambda_n \longrightarrow \infty \ \text{ as } \ n \to \infty, \tag{7}$$

where λ_n is the optimal fixed effects tuning parameter for a sample of size n, then, given a $\sqrt{n}$-consistent estimator of D, there is a local minimizer of the fixed effects estimates $\vec{\beta}^*$ which satisfies,

(i) $\|\vec{\beta} - \vec{\beta}^*\| = O\left(\frac{1}{\sqrt{n}}\right)$.

(ii) Sparsity; $\vec{\beta}^*$ does not over-fit the fixed effects structure of the model.

(iii) Asymptotic normality;

$$\sqrt{n}(\vec{\beta} - \vec{\beta}^*) \xrightarrow{D} N(0, \Sigma_{\vec{\beta}}), \text{ where } \Sigma_{\vec{\beta}} = \lim_{m\to\infty} \sigma^2 \left(\frac{1}{n} \sum_{i=1}^{m} X_i^T \left(Z_i D^* Z_i^T + I_{n_i} \right)^{-1} X_i \right)^{-1}.$$

Note that in property (iii), the X_i and Z_i are assumed to be the data given by the true model. Similarly, for the random effects tuning parameter, ξ, Theorem 3.2 also from Peng and Lu (2012) states that under certain regularity conditions, given a $\sqrt{n}$-consistent estimator of $\vec{\beta}$, if

$$\xi_n \sqrt{\frac{n}{\log(n)}} \longrightarrow O(1) \quad \text{as} \quad n \to \infty, \tag{8}$$

where ξ_n is the optimal random effects tuning parameter for a sample of size n, then there is a local minimizer of the random effects estimates $\vec{b}_i^*$ for $1 \leq i \leq m$, which satisfies

(i) Sparsity; $\text{diag}(D^*)$ does not over-fit the random effects structure of the model.

(ii) Asymptotic normality and consistency of $\text{diag}(D^*)$.

The GCV, AIC, and BIC are used as the selection criterion. When deciding the range of λ and ξ values to consider in selecting and estimating the LMM, adhering to their asymptotic behavior can be used as a starting point. $\lambda = \frac{\log(n)}{\sqrt{n}}$ is an appropriate choice. And from equation (8) it follows that $\xi = \sqrt{\frac{\log(n)}{n}}$ is an appropriate choice since

$$\xi \sqrt{\frac{n}{log(n)}} = \sqrt{\frac{\log(n)}{n}} \sqrt{\frac{n}{\log(n)}} = 1 = O(1).$$