Joint Variable Selection for Fixed and Random Effects in Linear Mixed-Effects Models

Howard D. Bondell, Arun Krishna, and Sujit K. Ghosh

Department of Statistics, North Carolina State University

Raleigh, NC 27695-8203, U.S.A.

Correspondence Author: Howard D. Bondell

email: bondell@stat.ncsu.edu

Telephone: (919)515-1914; Fax: (919)515-1169

Abstract

It is of great practical interest to simultaneously identify the important predictors that correspond to both the fixed and random effects components in a linear mixed-effects model. Typical approaches perform selection separately on each of the fixed and random effect components. However, changing the structure of one set of effects can lead to different choices of variables for the other set of effects. We propose simultaneous selection of the fixed and random factors in a linear mixed-effects model using a modified Cholesky decomposition. Our method is based on a penalized joint log-likelihood with an adaptive penalty for the selection and estimation of both the fixed and random effects. It performs model selection by allowing fixed effects or standard deviations of random effects to be exactly zero. A constrained EM algorithm is then used to obtain the final estimates. It is further shown that the proposed penalized estimator enjoys the Oracle property, in that, asymptotically it performs as well as if the true model was known beforehand. We demonstrate the performance of our method based on a simulation study and a real data example.

Keywords: Adaptive lasso; Constrained EM algorithm; Linear mixed model; Modified Cholesky decomposition; Penalized likelihood; Variable selection.

1 Introduction

Linear mixed-effects (LME) models (Laird and Ware, 1982) are a class of statistical models used to describe the relationship between the response and covariates, based on clustered data. Examples of clustered data are repeated measures and nested designs. By introducing subject-specific random effects, the LME model allows flexibility to model the means as well as the covariance structure.

As a motivating example, consider a recent study of the association between total nitrate concentration in the atmosphere and a set of measured predictors (Lee and Ghosh, 2008; Ghosh et al., 2009). Nitrate is one of the major components of fine particulate matter ($PM_{2.5}$) across the United States (Malm et al., 2004). However, it is one of the most difficult components to simulate accurately using numerical air quality models (Yu et al., 2005). An alternate approach is to identify empirical relationships that exist between nitrate concentrations and a set of observed variables that can act as surrogates for the different nitrate formation and loss pathways (Ghosh et al., 2009). To formulate these relationships, data obtained from the U.S. EPA Clean Air Status and Trends Network (CASTNet) sites are used. The CASTNet dataset consists of multiple sites with repeated measurements of pollution and meteorological variables on each site. The data enables us to identify these relationships which can allow for more accurate simulation of air quality. Further details of this data and associated analysis using our methods are described in Section 6.

To fix notation, denote the number of subjects by m, with response from each subject $i = 1, 2 \ldots, m$ measured n_i times, and let $N = \sum_{i=1}^{m} n_i$. For the CASTNet data described above, each site is considered as a subject. Let $\boldsymbol{y}_i$ be an $n_i \times 1$ response for subject i. Let $\boldsymbol{X}_i$ be the $n_i \times p$ design matrix of explanatory variables, and $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)'$ be the regression parameter vector. Let $\boldsymbol{b}_i^* = (b_{i1}^*, \cdots, b_{iq}^*)'$ be a $q \times 1$ vector of subject-specific random effects with $\boldsymbol{b}_i^* \sim N(0, \sigma^2 \boldsymbol{\Psi})$, and assumed independent across subjects. Denote $\boldsymbol{Z}_i$ as the $n_i \times q$ design matrix corresponding to the random effects. Often one sets $\boldsymbol{Z}_i = \boldsymbol{X}_i$, but it is not

necessary. Then, a general class of LME models can be written as

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i^* + \boldsymbol{\varepsilon}_i, \tag{1.1}$$

where the errors $\boldsymbol{\varepsilon}_i$'s are independently distributed $N(0, \sigma^2\boldsymbol{I}_{n_i})$ and independent of the $\boldsymbol{b}_i^*$'s.

Lange and Laird (1989) showed that underfitting the covariance structure would lead to bias in the estimated variance of the fixed effects. On the other hand, including unnecessary random effects could lead to a near singular random effect covariance matrix. The main goal of this paper is to simultaneously identify the subsets of important predictors that correspond to the fixed and the random components, respectively.

The problem of selecting variables has received considerable attention over the years, and a large number of methods have been proposed (see for example, Miller, 2002, for a review). Traditional methods, such as forward selection and backward elimination can be unstable due to the inherent discreteness (Breiman, 1996). More recently, penalized regression has emerged as a successful method to tackle this problem, for examples see Tibshirani (1996), Fan and Li (2001), Efron, Hastie, Johnstone and Tibshirani (2004), Zou and Hastie (2005), Zou (2006), Bondell and Reich (2008). However, the selection of random effects together with the fixed effects in the LME model has received little attention. Typical methods select fixed effects with the random effect structure unchanged. Few procedures have been proposed to select the random effects as well. Model selection criteria such as AIC (Akaike, 1973), BIC (Schwartz, 1978), GIC (Rao and Wu, 1989), and conditional AIC (Vaida and Blanchard, 2005) have been used to compare a list of models. However, the number of possible models increases exponentially with the number of predictors, as it is given by 2^{p+q}, which for the CASTNet data is over 4 billion models.

To reduce computational demand, Pu and Niu (2006) proposed the Extended GIC (EGIC), while Wolfinger (1993) and Diggle, Liang and Zeger (1994) proposed the Restricted Information Criteria, where selection is first performed on either the mean or the covariance structure while fixing the other at the full model. This results in the number of possible sub-models

considered to be $2^p + 2^q$, which may still be large, as for the CASTNet data this gives over 130,000 models. Jiang and Rao (2003) also proposed an alternative two-stage procedure. Forward or Backward selection can avoid enumerating all possible models (Morell, Pearson and Brant, 1997) however the discrete nature makes them unstable. More recently, Jiang, Rao, Gu and Nguyen (2008) proposed a 'fence' method to select predictors in a general mixed model. Although these methods may avoid the need to search through the entire model space, it may remain computationally intensive when the number of predictors are large. A Bayesian method was proposed by Chen and Dunson (2003) and Kinney and Dunson (2007) by selecting a prior with mass at zero for the random effect variances.

A difficulty in defining a shrinkage approach to random effects selection is that an entire row and column of $\mathbf{\Psi}$ must be eliminated to successfully remove a random effect. This leads to complications in how to perform the shrinkage appropriately. In this article we propose a new method for simultaneously selecting the fixed and the random effects parameters, in which the selection is done for both types of effects in a combined penalized procedure. Our proposed method is based on a re-parametrization of the LME model obtained by a modified Cholesky decomposition of $\mathbf{\Psi}$ (Chen and Dunson, 2003). This modified factorization aids us in the selection of the random effects by dropping out the random effects terms which have zero variance.

The SCAD (Fan and Li, 2001) and the adaptive LASSO (Zou, 2006) showed that asymptotically penalized estimators can perform as well as the 'Oracle' estimators which knows the true model beforehand. Motivated by the oracle properties of the adaptive LASSO estimates, we use an adaptive penalty on the re-parameterized model that simultaneously selects the fixed and the random effects.

The remainder of the paper is structured as follows. In Section 2, we describe the re-parameterized linear mixed models and its properties. Section 3, describes our method which selects the important variables for the fixed as well as the random effects. In Section 4 we

show that our penalized estimators possess the asymptotic Oracle property. We illustrate the performance of our method with a simulation study in Section 5. The proposed approach is applied to the U.S. EPA CASTNet data in Section 6 and compared to other selection methods. Finally, in Section 7 we conclude with a discussion. All proofs are given in the Web Appendix.

2 The Re-parameterized Linear Mixed Effects Model

The Cholesky decomposition has been extensively used as a computational tool for estimating the covariance matrix of the random effects (Lindstrom and Bates, 1988; Pinhiero and Bates, 1996; Smith and Kohn, 2002). However the parameters in the Cholesky decomposition does not allow for elimination of random effects. This is due to the fact that the covariance matrix depends on all of these parameters from the decomposition. To alleviate this drawback, we adopt a modified Cholesky decomposition as in Chen and Dunson (2003), where we factorize the covariance matrix, $\boldsymbol{\Psi}$, of the random effects as $\boldsymbol{\Psi} = \boldsymbol{D\Gamma\Gamma'D}$, where $\boldsymbol{D} = \text{diag}(d_1, d_2, \cdots, d_q)$ is a diagonal matrix, and $\boldsymbol{\Gamma}$, whose $(l, r)^{th}$ element is denoted by γ_{lr}, is a $q \times q$ lower triangular matrix with $1's$ on the diagonal. This decomposition in terms of $\boldsymbol{D}$ and $\boldsymbol{\Gamma}$ is unique, and leads to a non-negative definite matrix $\boldsymbol{\Psi}$. Given the decomposition, the re-parameterized LME model can be written as

$$\boldsymbol{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{D\Gamma b}_i + \boldsymbol{\varepsilon}_i, \tag{2.1}$$

where we assume $\boldsymbol{y}_i$ has been centered and the predictors have been standardized so that, $\boldsymbol{X}_i'\boldsymbol{X}_i$ and $\boldsymbol{Z}_i'\boldsymbol{Z}_i$ represent the correlation matrices, and $\boldsymbol{b}_i = (b_{i1}, \cdots, b_{iq})'$ is a $q \times 1$ vector of independent $N(0, \sigma^2\boldsymbol{I}_q)$. The covariance matrix of $\boldsymbol{b}_i^*$, is now expressed as a function of $\boldsymbol{d} = (d_1, d_2, \cdots, d_q)'$, and the $q(q-1)/2$ free elements of $\boldsymbol{\Gamma}$, denoted by the vector $\boldsymbol{\gamma} = (\gamma_{lr} : l = 1, \cdots, q : r = l+1, \ldots, q)'$. We denote $\boldsymbol{\phi} = (\boldsymbol{\beta}', \boldsymbol{d}', \boldsymbol{\gamma}')'$, a $k \times 1$ vector of unknown parameters, where $k = p + \frac{q(q+1)}{2}$.

With this convenient decomposition, setting $d_l = 0$ is equivalent to setting all the elements in the l^{th} column and l^{th} row of $\boldsymbol{\Psi}$ to zero and creating a new sub-matrix by removing the corresponding row and column. Hence a single parameter controls the inclusion/exclusion of the random effects.

2.1 The Likelihood

For the re-parameterized linear model, conditioning on $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$, the distribution of $\boldsymbol{y}_i$ follows a normal distribution with mean $\boldsymbol{X}_i\boldsymbol{\beta}$, and variance $\boldsymbol{V}_i = \sigma^2(\boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{\Gamma}\boldsymbol{\Gamma}'\boldsymbol{D}\boldsymbol{Z}_i' + \boldsymbol{I}_{n_i})$. Dropping constant terms, the log-likelihood function is given by

$$L(\boldsymbol{\phi}) = -\frac{1}{2}\log|\tilde{\boldsymbol{V}}| - \frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'\tilde{\boldsymbol{V}}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}), \tag{2.2}$$

where $\tilde{\boldsymbol{V}} = \text{Diag}(\boldsymbol{V}_1, \ldots, \boldsymbol{V}_m)$ a block diagonal matrix of $\boldsymbol{V}_i$'s, and $\boldsymbol{y} = (\boldsymbol{y}_1', \ldots, \boldsymbol{y}_m')'$, $\boldsymbol{X} = [\boldsymbol{X}_1', \ldots, \boldsymbol{X}_m']'$ are the stacked $\boldsymbol{y}_i$ and $\boldsymbol{X}_i$, respectively.

By treating $\boldsymbol{b} = (\boldsymbol{b}_1', \ldots, \boldsymbol{b}_m')'$ as observed, and again dropping constants, we can write the complete data log-likelihood function as

$$L_c(\boldsymbol{\phi}|\boldsymbol{y}, \boldsymbol{b}) = -\frac{N + mq}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\left(||\boldsymbol{y} - \boldsymbol{Z}\tilde{\boldsymbol{D}}\tilde{\boldsymbol{\Gamma}}\boldsymbol{b} - \boldsymbol{X}\boldsymbol{\beta}||^2 + \boldsymbol{b}'\boldsymbol{b}\right), \tag{2.3}$$

where $\boldsymbol{Z}$ represents a block diagonal matrix of $\boldsymbol{Z}_i$ and $\tilde{\boldsymbol{D}} = I_m \otimes \boldsymbol{D}$ and $\tilde{\boldsymbol{\Gamma}} = I_m \otimes \boldsymbol{D}$, with $\otimes$ representing the Kronecker product.

We now maximize the conditional expectation of (2.3) along with a penalty function on $\boldsymbol{\beta}$ and $\boldsymbol{d}$, to decide whether to include or exclude a predictor. Dropping out the terms which do not involve either $\boldsymbol{\beta}$ or $\boldsymbol{d}$ is then equivalent to minimizing the conditional expectation of $||\boldsymbol{y} - \boldsymbol{Z}\tilde{\boldsymbol{D}}\tilde{\boldsymbol{\Gamma}}\boldsymbol{b} - \boldsymbol{X}\boldsymbol{\beta}||^2$ plus the penalty term.

3 Penalized Selection and Estimation for the Re-parameterized LME model

3.1 The Shrinkage Penalty

Recently, Zou (2006) proposed the Adaptive LASSO where adaptive weights are used to penalize different regression coefficients in the L_1 penalty. That is, we wish to have a large amount of shrinkage applied to the zero-coefficients while smaller amounts are used for the non-zero ones which then results in an estimator with improved efficiency and selection properties. The Adaptive LASSO estimate for the linear regression model is defined as

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||^2 + \lambda_m \sum_{j=1}^{p} \bar{w}_j |\beta_j|, \tag{3.1}$$

where λ_m is a non-negative regularization parameter, $\bar{w}_j$ are adaptive weights, typically $\bar{w}_j = 1/|\bar{\beta}_j|$, with $\bar{\beta}_j$ the ordinary least squares estimate. As λ_m increases, the coefficients are continuously shrunk towards zero and, due to the L_1 form, some coefficients can be exactly shrunk to zero. We adopt this adaptive penalty, coupled with the re-parameterization, to perform the selection.

Given the LME model (2.1) and the complete data log-likelihood (2.3), we can define our penalized criterion under the L_1 penalty with the adaptive weights, jointly for the fixed and random effects as

$$\boldsymbol{Q}_c(\boldsymbol{\phi}|\boldsymbol{y}, \boldsymbol{b}) = ||\boldsymbol{y} - \boldsymbol{Z}\tilde{\boldsymbol{D}}\tilde{\boldsymbol{\Gamma}}\boldsymbol{b} - \boldsymbol{X}\boldsymbol{\beta}||^2 + \lambda_m \left(\sum_{j=1}^{p} \frac{|\beta_j|}{|\bar{\beta}_j|} + \sum_{j=1}^{q} \frac{|d_j|}{|\bar{d}_j|} \right). \tag{3.2}$$

Here $\bar{\boldsymbol{\beta}}$ is the generalized least squares estimate of $\boldsymbol{\beta}$, and $\bar{\boldsymbol{d}}$ is obtained by decomposition of the estimated covariance matrix obtained by restricted maximum likelihood, using the unpenalized likelihood with standard software.

Rearranging the terms, the equation given in (3.2) can instead be written as

$$\boldsymbol{Q}_c(\boldsymbol{\phi}|\boldsymbol{y}, \boldsymbol{b}) = ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\mathrm{Diag}(\tilde{\boldsymbol{\Gamma}}\boldsymbol{b})(\boldsymbol{1}_q \otimes I_m)\boldsymbol{d}||^2 + \lambda_m \left(\sum_{j=1}^{p} \frac{|\beta_j|}{|\bar{\beta}_j|} + \sum_{j=1}^{q} \frac{|d_j|}{|\bar{d}_j|} \right), \tag{3.3}$$

where $\mathbf{1}_q$ denotes a column vector of ones of length q. From (3.3), we have a quadratic form in $(\boldsymbol{\beta}', \boldsymbol{d}')'$.

3.2 Computation and Tuning

3.2.1 The Constrained EM Algorithm

Laird and Ware (1982) and Laird, Lange and Stram (1987) used the Expectation-Maximization (EM) algorithm in the context of the LME model, where the complete data consists of $\boldsymbol{y}_i$ plus the unobserved random parameters. We adopt the EM algorithm, in that, we compute the conditional expectation of $\boldsymbol{Q}_c(\boldsymbol{\phi}|\boldsymbol{y}, \boldsymbol{b})$ assuming the random effects are unobserved (E-step). Then we minimize to obtain the updated penalized likelihood estimates of our parameters (M-step). This process is repeated iteratively until convergence.

Given (2.3), the conditional distribution of $\boldsymbol{b}$ given $\boldsymbol{\phi}$ and $\boldsymbol{y}$ is, $\boldsymbol{b}|\boldsymbol{y}, \boldsymbol{\phi} \sim N(\hat{\boldsymbol{b}}, \boldsymbol{U})$ where the mean and variance are given by,

$$\begin{aligned} \hat{\boldsymbol{b}}^{(\omega)} &= (\tilde{\boldsymbol{\Gamma}}'^{(\omega)} \tilde{\boldsymbol{D}}^{(\omega)} \boldsymbol{Z}' \boldsymbol{Z} \tilde{\boldsymbol{D}}^{(\omega)} \tilde{\boldsymbol{\Gamma}}^{(\omega)} + \boldsymbol{I})^{-1} (\boldsymbol{Z} \tilde{\boldsymbol{D}}^{(\omega)} \tilde{\boldsymbol{\Gamma}}^{(\omega)})' (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^{(\omega)}) \\ \text{and} \quad \boldsymbol{U}^{(\omega)} &= \sigma^{2(\omega)} (\tilde{\boldsymbol{\Gamma}}'^{(\omega)} \tilde{\boldsymbol{D}}^{(\omega)} \boldsymbol{Z}' \boldsymbol{Z} \tilde{\boldsymbol{D}}^{(\omega)} \tilde{\boldsymbol{\Gamma}}^{(\omega)} + \boldsymbol{I})^{-1}, \end{aligned} \tag{3.4}$$

respectively. Here, ω indexes the iterations and $\omega = 0$ refers to the initial estimates, chosen to be the REML estimates. The updated estimate for σ^2 at iteration ω is given by

$$\sigma^{2(\omega)} = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^{(\omega)})' \left(\boldsymbol{Z} \tilde{\boldsymbol{D}}^{(\omega)} \tilde{\boldsymbol{\Gamma}}^{(\omega)} \tilde{\boldsymbol{\Gamma}}'^{(\omega)} \tilde{\boldsymbol{D}}^{(\omega)} \boldsymbol{Z}' + \boldsymbol{I}_N \right)^{-1} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^{(\omega)}) \Big/ N.$$

Let $\boldsymbol{\phi}^{(\omega)}$ be the estimate of $\boldsymbol{\phi}$ at the ω^{th} iteration. We first compute the E-step by taking the conditional expectation of $\boldsymbol{Q}_c(\boldsymbol{\phi}|\boldsymbol{y}, \boldsymbol{b})$,

$$\boldsymbol{g}(\boldsymbol{\phi}|\boldsymbol{\phi}^{(\omega)}) = \mathrm{E}_{\boldsymbol{b}|\boldsymbol{y}, \boldsymbol{\phi}^{(\omega)}} \left\{ ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\mathrm{Diag}(\tilde{\boldsymbol{\Gamma}}\boldsymbol{b})(\mathbf{1}_q \otimes I_m)\boldsymbol{d}||^2 \right\} + \lambda_m \left(\sum_{j=1}^{p} \frac{|\beta_j|}{|\bar{\beta}_j|} + \sum_{j=1}^{q} \frac{|d_j|}{|\bar{d}_j|} \right). \tag{3.5}$$

Then, for the M-step, we minimize the objective function, $\boldsymbol{g}(\boldsymbol{\phi}|\boldsymbol{\phi}^{(\omega)})$ with respect to $(\boldsymbol{\beta}', \boldsymbol{d}', \boldsymbol{\gamma}')'$. This optimization within the M-step is done by iterating between $\boldsymbol{\gamma}$ and the vec-

tor $(\boldsymbol{\beta}', \boldsymbol{d}')'$. The optimization iteration for $\boldsymbol{\gamma}$ is closed form, while the iteration for $(\boldsymbol{\beta}', \boldsymbol{d}')'$ will be a quadratic programming problem. Further details for computing the expectation explicitly and performing the M-step can be found in Web Appendix B. Once we have convergence within the M-step, we have the updated $\left(\boldsymbol{\beta}^{(\omega+1)'}, \boldsymbol{d}^{(\omega+1)'}, \boldsymbol{\gamma}^{(\omega+1)'}\right)'$ and also $\sigma^{2(\omega+1)}$, and return to the E-step. Upon convergence of the full EM algorithm, we obtain our final estimates $\hat{\boldsymbol{\phi}} = (\hat{\boldsymbol{\beta}}', \hat{\boldsymbol{d}}', \hat{\boldsymbol{\gamma}}')'$.

3.2.2 Choice of tuning parameter

The EM algorithm described above applies to a fixed value of λ_m. In practice, λ_m is chosen on a grid and the solution is obtained for each λ_m. Next, we must choose from among the candidate values of λ_m and obtain the final solution. This can be accomplished via minimizing a criterion such as AIC, BIC, GIC, Generalized Cross-Validation (GCV), or via k-fold Cross-Validation. It is known that under general conditions, BIC is consistent for model selection if the true model belongs to the class of models considered, while although AIC is minimax optimal, it is not consistent for selection (Shao, 1997; Yang, 2005; Pu and Niu, 2006). We use a BIC-type criterion given by

$$BIC_{\lambda_m} = -2L(\hat{\boldsymbol{\phi}}) + \log(N) \times (df_{\lambda_m}) \tag{3.6}$$

where $L(\hat{\boldsymbol{\phi}})$ is the obtained value of $L(\boldsymbol{\phi})$, as in (2.2), using the estimate $\hat{\boldsymbol{\phi}}$ obtained for that value of λ_m. We take the degrees of freedom, df_{λ_m}, as the number of non-zero coefficients in $\hat{\boldsymbol{\phi}}$. For the linear model this is an unbiased estimate of the degrees of freedom (Zou, Hastie and Tibshirani, 2007), and we adopt it for this setting as well. We then choose the solution that minimizes the BIC_{λ_m} criterion.

Note that in the BIC-type criterion, we use the total sample size, N, although in the mixed model situation this is not the effective sample size as pointed out by Pauler (1998), Jiang and Rao (2003), and Jiang et al. (2008). This criterion has worked well for tuning in our simulations (both reported and unreported), as well as the data example. This implementation

of the BIC-type criterion was also used by Pu and Niu (2006). We also compared tuning via AIC and HQIC (Hannan and Quinn, 1979) and found the best performance using the proposed BIC-type criterion.

4 Asymptotic Properties

Consider again $\boldsymbol{\phi} = (\boldsymbol{\beta}', \boldsymbol{d}', \boldsymbol{\gamma}')'$ and let $\bar{\boldsymbol{\phi}}$ denote an initial $\sqrt{m}$ consistent estimator of $\boldsymbol{\phi}$. Let $\boldsymbol{Q}(\boldsymbol{\phi})$ denote the penalized log-likelihood function with $L(\boldsymbol{\phi})$ is as given in (2.2), then

$$\boldsymbol{Q}(\boldsymbol{\phi}) = L(\boldsymbol{\phi}) - \lambda_m \sum_{j=1}^{k} \bar{w}_j(|\phi_j|)$$

$$\text{where, } \bar{w}_j = \begin{cases} 0, & \text{for, } \phi_j \in \boldsymbol{\gamma} \\ 1/\bar{\phi}_j, & \text{Otherwise} \end{cases}. \tag{4.1}$$

Denote the true value of $\boldsymbol{\phi}$ as

$$\boldsymbol{\phi}_0 = (\phi_{10}, \ldots, \phi_{k0})' = (\boldsymbol{\phi}_{10}', \boldsymbol{\phi}_{20}')' \tag{4.2}$$

where $\boldsymbol{\phi}_{10} = (\boldsymbol{\beta}_{10}', \boldsymbol{d}_{10}', \boldsymbol{\gamma}_{10}')'$ is an $s \times 1$ vector whose components are non-zero and let $\boldsymbol{\phi}_{20}$ be the $(k-s)$ remaining components of $\boldsymbol{\phi}_0$, so that $\boldsymbol{\phi}_{20} = 0$. In a similar manner we also decompose $\boldsymbol{\phi}$ itself as $\boldsymbol{\phi} = (\boldsymbol{\phi}_1', \boldsymbol{\phi}_2')'$. We shall next state our theorems, while the proofs and regularity conditions are given in Web Appendix A.

For the penalized log-likelihood given in (4.1), let $\boldsymbol{\phi} = (\boldsymbol{\phi}_1', \boldsymbol{0}')'$, that is fixing $\boldsymbol{\phi}_2 = 0$. Let $L(\boldsymbol{\phi}_1)$, $\boldsymbol{Q}(\boldsymbol{\phi}_1)$ denote the log-likelihood and the penalized log-likelihood of the first s components of $\boldsymbol{\phi}$ given by

$$\begin{aligned} L(\boldsymbol{\phi}_1) &\equiv L\left\{\begin{pmatrix} \boldsymbol{\phi}_1 \\ 0 \end{pmatrix}\right\} = -\frac{1}{2}\log|\tilde{\boldsymbol{V}}_{(1)}| - \frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}_{(1)}\boldsymbol{\beta}_1)'(\tilde{\boldsymbol{V}}_{(1)})^{-1}(\boldsymbol{y} - \boldsymbol{X}_{(1)}\boldsymbol{\beta}_1), \\ \boldsymbol{Q}(\boldsymbol{\phi}_1) &\equiv \boldsymbol{Q}\left\{\begin{pmatrix} \boldsymbol{\phi}_1 \\ 0 \end{pmatrix}\right\} = L(\boldsymbol{\phi}_1) - \lambda_m \sum_{j=1}^{s} \bar{w}_j(|\phi_j|), \end{aligned} \tag{4.3}$$

where $\tilde{\boldsymbol{V}}_{(1)} = \boldsymbol{Z}_{(1)}\tilde{\boldsymbol{D}}_1\tilde{\boldsymbol{\Gamma}}_1, \tilde{\boldsymbol{\Gamma}}_1'\tilde{\boldsymbol{D}}_1\boldsymbol{Z}_{(1)}' + \boldsymbol{I}$, is the block diagonal matrix corresponding to the non-zero components $(\boldsymbol{d}_1', \boldsymbol{\gamma}_1')'$ and $\boldsymbol{X}_{(1)}$ and $\boldsymbol{Z}_{(1)}$ are the corresponding design matrices.

Theorem 1. *Let $\boldsymbol{\phi} = (\boldsymbol{\phi}_1', \mathbf{0}')'$, and the observations follow the LME model (2.1) satisfying conditions $(i) - (iv)$ given in Web Appendix A. If $\lambda_m/\sqrt{m} \to 0$, then there exists a local maximizer $\hat{\boldsymbol{\phi}} = \begin{pmatrix} \hat{\boldsymbol{\phi}}_1 \\ \mathbf{0} \end{pmatrix}$ of $\boldsymbol{Q}\left\{\begin{pmatrix} \boldsymbol{\phi}_1 \\ \mathbf{0} \end{pmatrix}\right\}$ such that $\hat{\boldsymbol{\phi}}_1$ is $\sqrt{m}$ consistent for $\boldsymbol{\phi}_{10}$.*

Theorem 2. *Let the observations follow the LME model (2.1) satisfying conditions $(i) - (iv)$ given in Web Appendix A. If $\lambda_m \to \infty$, then with probability tending to 1 for any given $\boldsymbol{\phi}_1$ satisfying $||\boldsymbol{\phi}_1 - \boldsymbol{\phi}_{10}|| \leq Mm^{-1/2}$ and some constant $M > 0$,*

$$\boldsymbol{Q}\left\{\begin{pmatrix} \boldsymbol{\phi}_1 \\ 0 \end{pmatrix}\right\} = \max_{||\boldsymbol{\phi}_2|| \leq Mm^{-1/2}} \boldsymbol{Q}\left\{\begin{pmatrix} \boldsymbol{\phi}_1 \\ \boldsymbol{\phi}_2 \end{pmatrix}\right\}. \tag{4.4}$$

Remark 1. From Theorem 1 we see that we are able to get into a $\sqrt{m}$ neighborhood, while Theorem 2 shows that, with probability tending to 1, there exists a local maximizer in that neighborhood with $\hat{\boldsymbol{\phi}}_2 = \mathbf{0}$. Hence, combining the two, we see that our penalized likelihood estimator can identify the true model with probability tending to one.

Theorem 3. *Let the observations follow the LME model (2.1) satisfying conditions $(i) - (iv)$ given in Web Appendix A. Then as $\lambda_m \to \infty$ and $\lambda_m/\sqrt{m} \to 0$, we have*

$$\sqrt{m}I(\phi_{10})\left\{(\hat{\boldsymbol{\phi}}_1 - \boldsymbol{\phi}_{10}) + \left(\frac{\lambda_m}{m}\right) I^{-1}(\phi_{10})\boldsymbol{h}(\phi_{10})\right\} \to_d N\left(0, I(\phi_{10})\right) \tag{4.5}$$

where $\boldsymbol{h}(\boldsymbol{\phi}_{10}) = (\bar{w}_1 sgn(\phi_{10}), \ldots, \bar{w}_s sgn(\phi_{s0}))'$ an $s \times 1$ vector, and $I(\phi_{10})$ is the Fisher information knowing that $\boldsymbol{\phi}_2 = 0$.

Remark 2. From Theorem 2 and 3 as $\lambda_m \to \infty$ and $\lambda_m/\sqrt{m} \to 0$, we can say that our penalized estimator enjoys the oracle property in that asymptotically it performs as well as the oracle estimators, knowing $\boldsymbol{\phi}_2 = 0$. In particular, to first order, $\sqrt{m}(\hat{\boldsymbol{\phi}}_1 - \boldsymbol{\phi}_{10}) \to N\left(\mathbf{0}, I^{-1}(\boldsymbol{\phi}_{10})\right)$.

5 Simulation Study

In order to avoid complete enumeration of all possible (2^{p+q}) models, Wolfinger (1993) and Diggle, Liang and Zeger (1994) recommended the Restricted Information Criterion (denoted by REML.IC), in that, by using the most complex mean structure, selection is first performed on the variance-covariance structure by computing the AIC and/or BIC. Given the best covariance structure, selection is then performed on the fixed effects. Alternatively, Pu and Niu (2006) proposed the EGIC (Extended GIC), where using the BIC, selection is first performed on the fixed effects by including all of the random effects into the model. Once the fixed effect structure is chosen, selection is then performed on the random effects.

In this section, we compare our proposed method to the REML.IC as well as the EGIC. Given the selected random effects model by using the REML.IC, further comparisons are also shown for selection on the fixed effects performed using the LASSO, adaptive LASSO, and the stepwise selection procedure which allows movement in either the forward or backward directions. We evaluate the performance by comparing them to the 'Oracle' model which knows beforehand the true underlying model, i.e. the REML estimate of $\boldsymbol{\phi}_1$ knowing $\boldsymbol{\phi}_2 = 0$.

Three scenarios are considered. In each example, 200 datasets were simulated from a multivariate normal density

$$\boldsymbol{y}_i \sim N\left(\boldsymbol{X}_i\boldsymbol{\beta}, \sigma^2(\boldsymbol{Z}_i\boldsymbol{\Psi}\boldsymbol{Z}_i' + \boldsymbol{I}_{n_i})\right). \tag{5.1}$$

The three scenarios are given by:

1. Example 1: $m = 30$ subjects and $n_i = 5$ observations per subject, where $p = 9$ and $q = 4$. We consider the true model

$$y_{ij} = b_{i1} + \beta_1 x_{ij1} + \beta_2 x_{ij2} + b_{i2} z_{ij1} + b_{i3} z_{ij2} + \epsilon_{ij}, \quad \epsilon_{ij} \sim N(0,1), \tag{5.2}$$

with true values $(\beta_1, \beta_2) = (1, 1)$ and true variance-covariance matrix

$$\boldsymbol{\Psi} = \begin{bmatrix} 9 & 4.8 & 0.6 \\ 4.8 & 4 & 1 \\ 0.6 & 1 & 1 \end{bmatrix}, \tag{5.3}$$

such that there are 7 unimportant predictors for the fixed effects and 1 unimportant predictor for the random effects. The covariates x_{ijk} for, $k = 1, \ldots, 9$ and z_{ijl}, for $l = 1, 2, 3$ are generated from a uniform $(-2, 2)$ distribution, along with a vector of $\mathbf{1}$'s for the subject-specific intercept.

2. Example 2: The setup for the second scenario is the same as the first, except we increase the number of observation to $m = 60$ subjects and $n_i = 10$ observations per subject. This allows us to investigate the performance in a larger sample.

3. Example 3: We now set $m = 60$ subjects and $n_i = 5$ observations per subject, for a particular case where $p = 9$ and $q = 10$. The covariates x_{ijk} for $k = 1, \ldots, 9$, are generated from a uniform$(-2, 2)$ distribution. We set $\boldsymbol{Z}_i = \boldsymbol{X}_i$ plus a random intercept term. The true model is then given by

$$y_{ij} = b_{i1} + (\beta_1 + b_{i2})x_{ij1} + b_{i3}x_{ij2} + \beta_3 x_{ij3} + \epsilon_{ij}, \quad \epsilon_{ij} \sim N(0, 1), \tag{5.4}$$

with $(\beta_1, \beta_3) = (1, 1)$, and the true covariance matrix is the same as example one.

For the simulation study, in addition to selection comparsions, model comparisons and validation are made based on the Kullback-Leibler discrepancy (Kullback and Leibler, 1951) given by

$$KLD = E\left\{\log f(Y, \mathbf{X}, \mathbf{Z}|\boldsymbol{\phi}_0) - \log f(Y, \mathbf{X}, \mathbf{Z}|\hat{\boldsymbol{\phi}})\right\}. \tag{5.5}$$

The joint density $f(Y, \mathbf{X}, \mathbf{Z}|\boldsymbol{\phi}_0)$ is given by the conditional in (5.1) evaluated at the true parameters, along with the marginals of $\mathbf{X}$ and $\mathbf{Z}$. The density $f(Y, \mathbf{X}, \mathbf{Z}|\hat{\boldsymbol{\phi}})$ uses the estimate obtained by each method. The expectation is taken with respect to the true model.

Table 1 compares our proposed method (denoted by M-ALASSO) tuned via the BIC to 5 variable selection algorithms: EGIC (Pu and Niu, 2006), REML.IC (Wolfinger, 1993; Diggle, Liang and Zeger, 1994), stepwise procedure (denoted by STEPWISE), LASSO (Tibshrani, 1996) and the adaptive LASSO (Zou, 2006), all of which are tuned using either AIC and/or BIC. Note that, the LASSO, adaptive LASSO and STEPWISE are used to perform selection only on the fixed effects, given the random effects selected using REML.IC. Comparisons are also shown for the true model (denoted by Oracle) and the full model (denoted by Full).

Column 4 lists the median Kullback-Leibler discrepancy (KLD) along with its bootstrap standard errors over 200 simulations. In column 5 we report the relative-efficiency (RE), i.e the ratio of the median KLD of the 'Oracle' to the median KLD obtained for each method. We see that for all three scenarios performance of the proposed method is the closest to the 'Oracle' with a RE upward of 0.75. We also notice that as the sample increases, the relative KLD between our method and the 'Oracle' model becomes smaller, as the theoretical results suggest. Column six (% Correct) in Table 1 gives the percentage of times the true model (fixed and random effects combined) is selected, while columns seven (% CF) and column eight (% CR) correspond to the percentage of times the correct fixed and the correct random effects are selected by each method, respectively. In all examples we see that our method outperforms the competing method by correctly identifying the true model most often.

The simulation study demonstrates that the performance of the typical methods that select the fixed and random components separately is not as good as the proposed method which simultaneously selects both the fixed and random components. For example, in the first setup, using BIC to select the random effects while keeping the full fixed effect structure only selects the correct set of random effects 68% of the time. Now this incorrect structure will affect the 2nd step, i.e. the fixed effect selection, regardless of how selection is done in this next step, whether it be enumerating all possible models, or using a LASSO or adaptive LASSO. For the method that selects the fixed effects while leaving the random effects at the

full model (EGIC) it only selects the correct fixed effect structure 56% of the time, and this will of course carry over to the 2nd step of selecting the random effects.

6 Analysis of the U.S. EPA data

As discussed in the introduction, the U.S. EPA CASTNet (Clean Air Status and Trend Network) data has been widely used in air quality models to inter-relate levels of various air pollutants in the atmosphere. Recently Ghosh et al. (2009) used this data to capture the relationship between total nitrate concentration (TNO_3) and a set of measured predictors. We used a subset of the data obtained by selecting 15 relevant sites in the eastern portion of the United States. These sites were selected to overlap spatially with major sources of nitric oxide (NO) and nitrogen dioxide (NO_2) emission. The data and sites are further described in Web Appendix C. We use data from the years 2000-2004 averaged to create monthly observations. The sites vary in the number of observations that they have over a 5 year period, yielding a total of 826 observations. The response is taken as $\log(TNO_3)$ rather than TNO_3, as in previous analyses. To build the relationship, we consider the following variables within the mixed model framework: sulfate (SO_4), ammonium (NH_4), ozone (O_3), temperature (T), dew point temperature (T_d), relative humidity (RH), solar radiation (SR), wind speed (WS), and precipitation (P). The responses have been centered and the predictors have been standardized, hence the fixed intercept can be removed.

Plots of the $\log(TNO_3)$ concentration for each site as a function of time (Web Appendix C) shows seasonality. In order to allow for this periodic effect, we include trigonometric functions $s_j(t) = \text{Sin}(\frac{2\pi jt}{12})$ and $c_j(t) = \text{Cos}(\frac{2\pi jt}{12})$ and $j = 1, 2, 3$, as potential predictors to capture the seasonal effects. In addition there seems to be an overall downward trend over the 5 year period. The data now consists of 9 quantitative predictors, 6 constructed predictors, plus a covariate (denoted by $l(t)$) to capture a linear trend, making it a total of 16 variables (see Web Appendix C for a description of the dataset and some additional diagnostic plots).

This data is of specific interest due to the possible heterogeneity among the 15 sites. To achieve this, we fit a linear mixed-effects model by setting $\boldsymbol{Z}_i = \boldsymbol{X}_i$ along with a random intercept. This model specification allows each regression coefficient as well as the intercept to vary across the sites. Table 2 compares the variables selected using the different methods, for both fixed and random effects. The selected models were compared via a 5-fold cross-validation method. We randomly omitted 1/5 of the data and estimated the coefficients via REML based on the structure that was chosen by each method. The likelihood using those parameter estimates was then evaluated on the omitted data. This was repeated for 50 random splits of the data and the deviance was averaged and reported in Table 2. We note that the cross-validated deviance is smallest for our method. Table 3 lists the penalized likelihood estimates for the fixed effect regression coefficients and the random effect standard deviations using the proposed method.

Although, the proposed method allows for the possibility of performing selection among all possible choices of random slope, as in our analysis, in many applications, the practitioner only considers a small number of possible random effects. As a second analysis, we allowed only for random variation in the seasonal trends across the sites, and kept the slopes of the meteorological variables to be only fixed effects. Tables 4 and 5 show the corresponding results for that analysis. One thing to note is that the cross-validated likelihood remains best for the model chosen by the proposed method from the original analysis which allowed for random effects from each covariate. Using the proposed procedure, the original analysis included 7 fixed and 6 random effects, while the analysis that restricted to only a random seasonality included 9 fixed and 5 random effects.

7 Discussion

In this paper we have shown that the re-parameterized LME model using the modified Cholesky decomposition of the covariance matrix aids us in the efficient selection of the

random effects. By using simulated and real data we have illustrated that the proposed penalized method can outperform the commonly used methods with respect to both selection and estimation. By jointly selecting the fixed and random effects, performance is improved over performing selection in a two-stage manner. Much of the improvement can be attributed to the reliance of the two-stage procedure on the selection of the structure in the first step. Variation in the structure of one part of the model can greatly affect the selection for the other part. For example, by using the full fixed effect model and performing selection on the random effects under that model, additional noise is added by the irrelevant fixed effects. This can hamper the selection on the random effects, which will then also carry over to the 2nd step of selecting the fixed effects under the chosen random effect structure. We have also shown both theoretically as well as empirically that our penalized likelihood estimators asymptotically performs as well as the 'Oracle' model.

Note that the proposed method can be applied to any fixed covariance structure on the errors, in that one may have the within subject error structure as $\boldsymbol{\varepsilon}_i \sim N(\mathbf{0}, \sigma^2 \boldsymbol{\Sigma}_i)$ for some $\boldsymbol{\Sigma}_i$. An example would be longitudinal data where one may place an autoregressive structure on the correlation. Estimation would proceed as in the case for the unpenalized estimation procedure. Letting $\tilde{\boldsymbol{\Sigma}}$ be the block diagonal matrix of $\boldsymbol{\Sigma}_i$, for known $\tilde{\boldsymbol{\Sigma}}$, we replace $\frac{N+mq}{2} \log \sigma^2$ and $||\boldsymbol{y} - \boldsymbol{Z}\tilde{\boldsymbol{D}}\tilde{\boldsymbol{\Gamma}}\boldsymbol{b} - \boldsymbol{X}\boldsymbol{\beta}||^2$ in (2.5) by $\left\{ \frac{N+mq}{2} \log \sigma^2 + (1/2) \log |\tilde{\boldsymbol{\Sigma}}| \right\}$ and $||\tilde{\boldsymbol{\Sigma}}^{-1/2}\boldsymbol{y} - \tilde{\boldsymbol{\Sigma}}^{-1/2}\boldsymbol{Z}\tilde{\boldsymbol{D}}\tilde{\boldsymbol{\Gamma}}\boldsymbol{b} - \tilde{\boldsymbol{\Sigma}}^{-1/2}\boldsymbol{X}\boldsymbol{\beta}||^2$ respectively. After this transformation, the remainder follows by redefining $(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{y})$. If there are unknown parameters in $\boldsymbol{\Sigma}$, the process must then iterate between estimation of the fixed and random effect parameters given $\boldsymbol{\Sigma}$, and estimation of the parameters in $\boldsymbol{\Sigma}$ given the fixed and random effects. This must be done separately for each tuning parameter λ_m. To save on computational burden, iterating only one or two steps will typically suffice.

Since the approach is based on the assumption of normality for both the conditional distribution as well as the distribution of the random effects, it may suffer from a lack of

robustness to deviations from this assumption. These robustness issues have been studied in the context of the unpenalized LME framework. Modifications to the penalized approach to account for robustness to non-normality in the random effects deserve investigation, but are beyond the scope of this paper.

Supplementary Materials

The Web Appendix is available under the paper information link at the Biometrics website http://www.tibs.org/biometrics.

Acknowledgment and Disclaimer

We thank Steven Howard at the United States Environmental Protection Agency (EPA) for processing and formatting the CASTNet data for our application. A portion of the CASTNet data set has been used only for illustrative purpose of our methodology. The U.S. EPA through its Office of Research and Development is not responsible for the content of this document or its implications. Bondell is partially supported by NSF grant number DMS-0705968. Ghosh is partially supported by NIH grant number 5R01ES014843-02. The authors would like to thank the editor, associate editor and two anonymous referees for their help in improving this manuscript.

References

Akaike, H. (1973). Information theory and an extension of maximum likelihood principle. In *Second international symposium on information theory*, eds. Petrov, B. N. and Csaki, F.

Breiman, L. (1996). Heuristics of instability and stabilization in model selection. *Annals of Statistics* **24**, 2350-2383.

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

Chen, Z. and Dunson, D., B. (2003). Random effects selection in linear mixed models. *Biometrics* **59**, 762-769.

Diggle, P; Liang, K; Zeger, S. (1994). Analysis of Longitudinal Data, Oxford Press.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression. *Annals of Statistics* **32**, 407-499.

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

Ghosh, S. K., Bhave, P. V., Davis, J. M., and Lee, H. (2009). Spatio-temporal analysis of total nitrate concentrations using dynamic statistical models. *Journal of the American Statistical Association*, To appear.

Hannan, E. J. and Quinn, B. G. (1979). The determination of the order of an autoregression. *Journal of the Royal Statistical Society* B **41**, 190-195.

Jiang, J. and Rao, J. S. (2003). Consistent procedures for mixed linear model selection. *Sankhya A* **65**, 23-42.

Jiang, J., Rao, J. S., Gu, Z. and Nguyen, T. (2008). Fence methods for mixed models selection. *Annals of Statistics* **36**, 1669-1692.

Kinney, S., K. and Dunson, D., B. (2007). Fixed and random effects selection in linear and logistic models. *Biometrics* **63**, 690-698.

Kullback, S. and Leiber, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics* **22**, 72-86.

Laird, N. M. and Ware, J. L. (1982). Random-effects models for longitudinal data. *Biometrics* **38**, 963-974.

Laird, N. M. and Lange, N. and Stram, D. (1987). Maximum likelihood computations with repeated measures: application of the EM algorithm. *Journal of the American Statistical Association* **82**, 97-105.

Lange, N. and Laird, N. M. (1989). The effect of covariance structures on variance estimation in balance growth-curve models with random parameters. *Journal of the American Statistical Association* **84**, 241-247.

Lee, H. and Ghosh, S. K. (2008). A re-parametrization approach for dynamic space-time models. *Journal of Statistical Theory and Practice* **2**, 1-14.

Lindstrom, M. J. and Bates, D. M. (1988). Newton-Raphson and EM algorithms for linear mixed-effects models for repeated measures data. *Journal of the American Statistical Association* **83**, 1014-1022.

Malm, W. C., Schichtel, B. A., Pitchford, M. L., Ashbaugh, L. L., and Eldred, R. A. (2004). Spatial and monthly trends in speciated fine particle concentration in the United States. *Journal of Geophysical Research* **109**, D03306.

Miller, A. (2002). Subset Selection in Regression. Chapman & Hall/ CRC, 2^{nd} ed.

Morell, C. H., Pearson, J. D. and Brant, L. J. (1997). Linear transformations of linear mixed-effects models. *The American Statistician* **51**, 338-343.

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

Niu, F. and Pu, P. (2006). Selecting mixed-effects models based on generalized information criterion. *Journal of Multivariate Analysis* **97**, 733-758.

Pinhiero, J. and Bates, D. (1996). Unconstrained parameterizations for variance-covariance matrices. *Statistics and Computing* **6**, 289-286.

Rao, C., R. and Wu, Y. (1989). A strongly consistent procedure for model selection in regression problems. *Biometrika* **76**, 369-374.

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

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

Smith, M. and Kohn, R. (2002). Parsimonious covariance matrix estimation for longitudinal data. *Journal of the American Statistical Association* **97**, 1141-1153.

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

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

Wolfinger, R.D. (1993). Covariance structure selection in general mixed models. *Communications*

in Statistics, Simulation and Computation **22**, 1079- 1106.

Yang, Y. (2005). Can the strengths of AIC and BIC be shared? A conflict between model identification and regression estimation. *Biometrika* **92**, 937-950.

Yu, S., Dennis, R., Roselle, S., Nenes, A., Walker, J., Eder, B., Schere, K., Swall, J., and Robarge, W. (2005). An assessment of the ability of three-dimensional air quality models with current thermodynamic equilibrium models to predict aerosol NO-3. *Journal of Geophysical Research* **110**, D07S13.

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

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society B* **67**, 301-320.

Zou, H., Hastie, T., and Tibshirani, R. (2007). On the degrees of freedom of the lasso. *Annals of Statistics* **35**, 2173-2192.

Table 1: Comparing the median Kullback-Leibler discrepancy (KLD) from the true model, along with the percentage of the times the true model was selected (%Correct) for each method, across 200 datasets. R.E. represents the relative efficiency compared to the oracle model. %CF, %CR corresponds to the percentage of times the correct fixed and random effects were selected, respectively.

Example	Method	Tuning	$KLD(S.E.)$	R.E.	%Correct	%CF	%CR
	Oracle	-	9.70 (0.343)	-	-	-	-
	M-ALASSO	BIC	10.94(0.475)	0.88	71	73	79
	EGIC	BIC	13.91(0.583)	0.69	47	56	52
	REML.IC	AIC	15.51(0.567)	0.63	19	21	62
	REML.IC	BIC	12.48(0.642)	0.77	59	59	68
	STEPWISE	AIC	16.01(0.611)	0.60	13	15	62
1	STEPWISE	BIC	12.91(0.584)	0.75	51	53	68
	LASSO	AIC	13.52(0.489)	0.71	17	21	62
	LASSO	BIC	12.87(0.414)	0.75	45	47	68
	ALASSO	AIC	13.03(0.399)	0.74	21	24	62
	ALASSO	BIC	12.12(0.414)	0.80	62	63	68
	Full	-	20.71 (0.513)	0.47	0	0	0
	Oracle	-	7.84(0.326)	-	-	-	-
	M-ALASSO	BIC	7.98(0.341)	0.98	83	83	89
	EGIC	BIC	12.55(0.581)	0.63	48	59	53
	REML.IC	AIC	11.93(0.432)	0.72	31	34	74
	REML.IC	BIC	10.18(0.415)	0.77	77	79	81
	STEPWISE	AIC	12.87(0.501)	0.61	26	28	74
2	STEPWISE	BIC	10.71(0.438)	0.73	68	69	81
	LASSO	AIC	12.53(0.388)	0.63	29	29	74
	LASSO	BIC	11.44(0.419)	0.69	59	61	81
	ALASSO	AIC	11.12(0.443)	0.69	39	41	74
	ALASSO	BIC	9.41 (0.420)	0.83	74	75	81
	Full	-	15.47 (0.476)	0.51	0	0	0
	Oracle	-	13.34(0.912)	-	-	-	-
	M-ALASSO	BIC	17.45(0.961)	0.76	61	63	84
	EGIC	BIC	24.89(2.013)	0.53	41	43	59
	REML.IC	AIC	28.87(2.231)	0.46	12	14	68
	REML.IC	BIC	23.39(1.872)	0.57	53	54	73
	STEPWISE	AIC	29.58(2.893)	0.47	8	11	68
3	STEPWISE	BIC	25.66(2.011)	0.52	38	40	73
	LASSO	AIC	22.97(1.031)	0.58	11	15	68
	LASSO	BIC	21.08(1.176)	0.63	22	25	73
	ALASSO	AIC	21.69(0.958)	0.62	27	29	68
	ALASSO	BIC	20.23(0.961)	0.66	52	55	73
	Full	-	38.52(2.172)	0.27	0	0	0

Table 2: Variables selected for the fixed and the random components for the CASTNet data allowing for a random intercept and all possible random slopes. The last column corresponds to the value of the cross-validated deviance via 5-fold CV and the methods are ordered by that value (smaller is better).

		Variables Selected		
Method	Tuning	Fixed	Random	CV Value
M-ALASSO	BIC	$x_2, x_3, x_6, x_9, l(t), s_1(t), c_1(t)$	Int, $x_1, x_2, l(t), s_1(t), c_1(t)$	-161.17
STEPWISE	BIC	$x_1, x_2, x_3, x_6, x_7, x_9, l(t), s_1(t), c_1(t)$	Int, $x_1, x_2, x_3, l(t), s_1(t), c_1(t)$	-160.73
ALASSO	AIC	$x_1, x_2, x_3, x_6, x_7, x_9, l(t), s_1(t), c_1(t), s_2(t), s_3(t)$	Int, $x_1, x_2, x_3, l(t), s_1(t), c_1(t)$	-160.09
STEPWISE	AIC	$x_1, x_2, x_3, x_6, x_7, x_9, l(t), s_1(t), c_1(t), s_2(t)$	Int, $x_1, x_2, x_3, l(t), s_1(t), c_1(t)$	-159.32
ALASSO	BIC	$x_1, x_2, x_3, x_6, x_7, x_9, l(t), s_1(t), c_1(t), s_2(t)$	Int, $x_1, x_2, x_3, l(t), s_1(t), c_1(t)$	-159.32
LASSO	AIC	$x_1, x_2, x_3, x_5, x_6, x_7, x_8, x_9, l(t), s_1(t), c_1(t), s_2(t), c_2(t), s_3(t)$	Int, $x_1, x_2, x_3, l(t), s_1(t), c_1(t)$	-157.85
LASSO	BIC	$x_1, x_2, x_3, x_5, x_6, x_7, x_9, l(t), s_1(t), c_1(t), s_2(t), c_2(t), s_3(t)$	Int, $x_1, x_2, x_3, l(t), s_1(t), c_1(t)$	-157.55

Table 3: Penalized Likelihood estimates for fixed effect regression coefficients and the random effect standard deviations using the proposed method, allowing for a random intercept and all possible random slopes.

Variables	Int	SO_4	NH_4	O_3	T	T_d	RH	SR	WS	P	$l(t)$	$s_1(t)$	$c_1(t)$	$s_2(t)$	$c_2(t)$	$s_3(t)$	$c_3(t)$
Fixed	-	0	3.19	2.71	0	0	-0.58	0	0	-0.38	-1.07	4.84	7.23	0	0	0	0
Random	0.24	1.28	1.90	0	0	0	0	0	0	0	0.38	0.98	1.34	0	0	0	0

Table 4: Variables selected for the fixed and the random components for the CASTNet data allowing for only the random intercept and time trend. The last column corresponds to the value of the cross-validated deviance via 5-fold CV and the methods are ordered by that value (smaller is better).

		Variables Selected		
Method	Tuning	Fixed	Random	CV Value
M-ALASSO	BIC	$x_1, x_2, x_3, x_6, x_9, l(t), s_1(t), c_1(t), s_2(t)$	Int, $l(t), s_1(t), c_1(t), c_2(t)$	-160.61
STEPWISE	AIC	$x_1, x_2, x_3, x_6, x_7, x_9, l(t), s_1(t), c_1(t), s_2(t), s_3(t)$	Int, $l(t), s_1(t), c_1(t), c_2(t)$	-160.53
ALASSO	AIC	$x_1, x_2, x_3, x_6, x_7, x_9, l(t), s_1(t), c_1(t), s_2(t), s_3(t)$	Int, $l(t), s_1(t), c_1(t), c_2(t)$	-160.53
STEPWISE	BIC	$x_1, x_2, x_3, x_6, x_7, x_9, l(t), s_1(t), c_1(t), s_2(t)$	Int, $l(t), s_1(t), c_1(t), c_2(t)$	-160.08
ALASSO	BIC	$x_1, x_2, x_3, x_6, x_7, x_9, l(t), s_1(t), c_1(t), s_2(t)$	Int, $l(t), s_1(t), c_1(t), c_2(t)$	-160.08
LASSO	BIC	$x_1, x_2, x_3, x_5, x_6, x_7, x_9, l(t), s_1(t), c_1(t), s_2(t), c_2(t), s_3(t), c_3(t)$	Int, $l(t), s_1(t), c_1(t), c_2(t)$	-159.98
LASSO	AIC	$x_1, x_2, x_3, , x_5, x_6, x_7, x_8, x_9, l(t), s_1(t), c_1(t), s_2(t), c_2(t), s_3(t), c_3(t)$	Int, $l(t), s_1(t), c_1(t), c_2(t)$	-159.83

Table 5: Penalized Likelihood estimates for fixed effect regression coefficients and the random effect standard deviations using the proposed method, allowing for only the random slope and time trend.

Variables	Int	SO_4	NH_4	O_3	T	T_d	RH	SR	WS	P	$l(t)$	$s_1(t)$	$c_1(t)$	$s_2(t)$	$c_2(t)$	$s_3(t)$	$c_3(t)$
Fixed	-	-2.28	6.20	2.65	0	0	-0.84	0	0	-0.64	-1.07	4.90	6.67	-0.14	0	0	0
Random	0.22	-	-	-	-	-	-	-	-	-	0.51	1.18	1.14	0	0.46	0	0

Web Appendix for Joint Variable Selection for Fixed and Random Effects in Linear Mixed-Effects Models

Howard D. Bondell, Arun Krishna, and Sujit K. Ghosh

APPENDIX A

A.1 Regularity Conditions

Assume that the data $\{(\boldsymbol{X}_i, \boldsymbol{Z}_i, \boldsymbol{y}_i)\ ;\ i = 1, \ldots, m\}$ is a random sample from a linear mixed-effects model (2.2) with probability density $f(\boldsymbol{y}_i|\boldsymbol{X}_i, \boldsymbol{Z}_i, \boldsymbol{\phi})$ where $\boldsymbol{\phi} = (\boldsymbol{\beta}', \boldsymbol{d}', \boldsymbol{\gamma}')'$ is a $k \times 1$ vector of unknown parameters. Let $L_i(\boldsymbol{\phi}) = \log\left(f(\boldsymbol{y}_i|\boldsymbol{X}_i, \boldsymbol{Z}_i, \boldsymbol{\phi})\right)$ denote the contribution of observation i to the log-likelihood function, and is given by

$$L_i(\boldsymbol{\phi}) = -\frac{1}{2}\log|\boldsymbol{V}_i| - \frac{1}{2}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})'(\boldsymbol{V}_i)^{-1}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}), \tag{A.1}$$

where $\boldsymbol{V}_i = \sigma^2(\boldsymbol{Z}_i\boldsymbol{D}\boldsymbol{\Gamma}\boldsymbol{\Gamma}'\boldsymbol{D}\boldsymbol{Z}_i' + \boldsymbol{I}_{n_i})$. Let $L(\boldsymbol{\phi}) = \sum_{i=1}^m L_i(\boldsymbol{\phi})$ and $Q(\boldsymbol{\phi})$ denote the log-likelihood and the penalized log-likelihood as given in (2.4) and (4.1), respectively. To present the proof of the theorems the following regularity conditions are imposed:

(i) Each cluster size $1 \leq n_i \leq K$, for some $K < \infty$ and $i = 1, ..., m$.

(ii) Let $v_i = I\{Z_i \text{ is full rank}\}$, where $I\{A\}$ denotes the indictor function of the event A. Assume that $\sum_{i=1}^m v_i/m \to c$, for some $0 < c \leq 1$.

(iii) The Fisher information matrix $I(\boldsymbol{\phi}_{10})$ knowing $\boldsymbol{\phi}_{20} = 0$ is finite and positive definite.

(iv) There exists a subset Θ of $\mathbb{R}^k$, containing the true parameter $\boldsymbol{\phi}_0$ such that $L_i(\boldsymbol{\phi})$ given in (A.1) admits all third order derivatives. Specifically, for $\phi_j = \beta_j$ and $(\phi_l, \phi_m) = \{(d_l, \gamma_m), (d_l, \gamma_m), (\gamma_l, \gamma_m)\}$, there exists a function $M_{jlm}(\boldsymbol{y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i)$ such that

$$\left|\frac{\partial^3}{\partial\beta_j\partial\phi_l\partial\phi_m}L_i(\boldsymbol{\phi})\right| = \left|\boldsymbol{X}_{ij}'\frac{\partial\boldsymbol{V}_i^{-1}}{\partial\phi_l\partial\phi_m}(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta})\right| < M_{jlm}(\boldsymbol{y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i),$$

for all $\boldsymbol{\phi} \in \Theta$, and $E_{\boldsymbol{\phi}_0}[M_{jlm}(\boldsymbol{y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i)] < \infty$. For $(\phi_j, \phi_l) = (\beta_j, \beta_l)$ and ϕ_m is either d_m or γ_m there exists a function $N_{jlm}(\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{Z})$ such that

$$\left.\begin{array}{l}\left|\frac{\partial^3}{\partial\beta_j\partial\beta_l\partial d_m}L_i(\boldsymbol{\phi})\right| = |\boldsymbol{X}_{ij}'(\boldsymbol{V}_i^{-1}\boldsymbol{S}_i^m\boldsymbol{V}_i^{-1})\boldsymbol{X}_{il}| \\ \left|\frac{\partial^3}{\partial\beta_j\partial\beta_l\partial\gamma_m}L_i(\boldsymbol{\phi})\right| = |\boldsymbol{X}_{ij}'(\boldsymbol{V}_i^{-1}\boldsymbol{T}_i^m\boldsymbol{V}_i^{-1})\boldsymbol{X}_{il}|\end{array}\right\} < N_{jlm}(\boldsymbol{y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i),$$

for all $\boldsymbol{\phi} \in \Theta$, and $E_{\boldsymbol{\phi}_0}[N_{jlm}(\boldsymbol{y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i)] < \infty$. Here $\boldsymbol{S}_i^m$ and $\boldsymbol{T}_i^m$ denote the partial derivatives of $\boldsymbol{V}_i$ with respect to d_m and γ_m, respectively, and are given by

$$\boldsymbol{S}_i^m = \boldsymbol{Z}_i \left\{ \frac{\partial}{\partial d_m} (\boldsymbol{D}\boldsymbol{\Gamma}\boldsymbol{\Gamma}'\boldsymbol{D}) \right\} \boldsymbol{Z}_i', \;\; \boldsymbol{T}_i^m = \boldsymbol{Z}_i \boldsymbol{D} \left\{ \frac{\partial}{\partial d \gamma_m} (\boldsymbol{\Gamma}\boldsymbol{\Gamma}') \right\} \boldsymbol{D}\boldsymbol{Z}_i'. \tag{A.2}$$

For $\phi_j = d_j$ and $(\phi_l, \phi_m) = \{(d_l, d_m), (d_l, \gamma_m), (\gamma_l, \gamma_m)\}$,

$$\left| \frac{\partial^3}{\partial d_j \partial \phi_l \partial \phi_m} L_i(\boldsymbol{\phi}) \right| < P_{jlm}(\boldsymbol{y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i),$$

for all $\boldsymbol{\phi} \in \Theta$, and $E_{\boldsymbol{\phi}_0}[P_{jlm}(\boldsymbol{y}_i, \boldsymbol{X}_i, \boldsymbol{Z}_i)] < \infty$.

Although it must be that $d_j \geq 0$ for all j, we allow the estimates to fall outside the boundary of the parameter space by using the maximum likelihood (ML) method as opposed to the REML.

Note that condition (i) can be relaxed to allow the cluster sized to also increase without bound. However, this can lead to a faster convergence rate for the fixed effects than that for the random effects (see, for example, Nie, 2007). Appropriate modifications to the theory presented here is then possible, but beyond the scope of the paper.

Condition (ii) is a sufficient condition that allows for full information regarding each random effect to grow at order m, that will typically hold in practice. However, less strict conditions can be derived.

A.2 Proof of Theorem 1

Proof. Consider the penalized log-likelihood given in (4.1) in a neighborhood of the true value $\boldsymbol{\phi}_{10}$. Let $\alpha_m = m^{-1/2}$ with $\boldsymbol{u} \neq 0$, and $\boldsymbol{\phi}_1 = \boldsymbol{\phi}_{10} + \alpha_m \boldsymbol{u}$. Fixing $\boldsymbol{\phi}_2 = 0$ we show that for a small enough $\epsilon > 0$ there exists a large constant C such that for sufficiently large m,

$$P\left\{ \sup_{||\boldsymbol{u}||=C} \boldsymbol{Q} \begin{pmatrix} \boldsymbol{\phi}_{10} + \alpha_m \boldsymbol{u} \\ 0 \end{pmatrix} < \boldsymbol{Q} \begin{pmatrix} \boldsymbol{\phi}_{10} \\ 0 \end{pmatrix} \right\} \geq 1 - \epsilon.$$

Note that

$$\begin{aligned} D_m(\boldsymbol{u}) &\equiv \boldsymbol{Q}(\boldsymbol{\phi}_1) - \boldsymbol{Q}(\boldsymbol{\phi}_{10}) \\ &= \{L(\boldsymbol{\phi}_{10} + \alpha_m \boldsymbol{u}) - L(\boldsymbol{\phi}_{10})\} - \lambda_m \left[\sum_{j=1}^{s} \bar{w}_j (|\phi_{j0} + \alpha_m u_j| - |\phi_{j0}|) \right]. \end{aligned}$$

Using a Taylor series expansion we have

$$\begin{aligned} D_m(\boldsymbol{u}) &= \alpha_m (\nabla L(\boldsymbol{\phi}_{10}))' \boldsymbol{u} + \frac{1}{2} \boldsymbol{u}' [\nabla^2 L(\boldsymbol{\phi}_{10})] \boldsymbol{u} \alpha_m^2 - \lambda_m \sum_{j=1}^{s} \bar{w}_j \mathrm{sgn}(\phi_{j0}) \alpha_m u_j \\ &= \frac{1}{\sqrt{m}} (\nabla L(\boldsymbol{\phi}_{10}))' \boldsymbol{u} + \frac{1}{2m} \boldsymbol{u}' [\nabla^2 L(\boldsymbol{\phi}_{10})] \boldsymbol{u} - \frac{\lambda_m}{\sqrt{m}} \sum_{j=1}^{s} \bar{w}_j \mathrm{sgn}(\phi_{j0}) u_j, \end{aligned} \tag{A.3}$$

where $\nabla L(\boldsymbol{\phi}_{10}), \nabla^2 L(\boldsymbol{\phi}_{10})$ denote the vector and matrix of the first and second order partial derivatives of $L(\boldsymbol{\phi}_1)$, respectively, evaluated at $\boldsymbol{\phi}_{10}$. From regularity condition (iv) it follows that

$$\left(\frac{1}{6m^{3/2}}\right)\sum_{i=1}^{m}\frac{\partial}{\partial\phi_j\partial\phi_l\partial\phi_m}L_i(\boldsymbol{\phi})\Bigg|_{\boldsymbol{\phi}_1=\boldsymbol{\phi}_{10}} \to 0, \;\; as\; m\to\infty,$$

hence the remainder term vanishes. For $\nabla L(\boldsymbol{\phi}_{10})$ the j^{th} partial derivative for each corresponding $\boldsymbol{\beta}_1, \boldsymbol{d}_1$ and $\boldsymbol{\gamma}_1$ satisfies

$$\left.\begin{array}{l} E\left\{\frac{\partial}{\partial\beta_j}L(\boldsymbol{\phi}_1)\right\} = E\left[\boldsymbol{X}'_{(1)j}\tilde{\boldsymbol{V}}^{-1}_{(1)}(\boldsymbol{y}-\boldsymbol{X}_{(1)}\boldsymbol{\beta}_1)\right] \\ E\left\{\frac{\partial}{\partial d_j}L(\boldsymbol{\phi}_1)\right\} = E\left[\frac{1}{2}[Tr(\tilde{\boldsymbol{V}}^{-1}_{(1)}\tilde{\boldsymbol{S}}^{j}_{(1)}) + (\boldsymbol{y}-\boldsymbol{X}_{(1)}\boldsymbol{\beta}_1)'(\tilde{\boldsymbol{V}}^{-1}_{(1)}\tilde{\boldsymbol{S}}^{j}_{(1)}\tilde{\boldsymbol{V}}^{-1}_{(1)})(\boldsymbol{y}-\boldsymbol{X}_{(1)}\boldsymbol{\beta}_1)]\right] \\ E\left\{\frac{\partial}{\partial\gamma_j}L(\boldsymbol{\phi}_1)\right\} = E\left[\frac{1}{2}[-Tr(\tilde{\boldsymbol{V}}^{-1}_{(1)}\tilde{\boldsymbol{T}}^{j}_{(1)}) + (\boldsymbol{y}-\boldsymbol{X}_{(1)}\boldsymbol{\beta}_1)'(\tilde{\boldsymbol{V}}^{-1}_{(1)}\tilde{\boldsymbol{T}}^{j}_{(1)}\tilde{\boldsymbol{V}}^{-1}_{(1)})(\boldsymbol{y}-\boldsymbol{X}_{(1)}\boldsymbol{\beta}_1)]\right] \end{array}\right| \Bigg\}_{\boldsymbol{\phi}_1=\boldsymbol{\phi}_{10}} = 0,$$

where $\boldsymbol{X}_{(1)j}$ corresponds to the j^{th} column of stacked matrix $\boldsymbol{X}_{(1)}$, and $\tilde{\boldsymbol{S}}^{j}_{(1)}$ and $\tilde{\boldsymbol{T}}^{j}_{(1)}$ are block diagonal matrices of the partial derivatives of $\tilde{\boldsymbol{V}}_{(1)}$ and are given by

$$\tilde{\boldsymbol{S}}^{j}_{(1)} = \boldsymbol{Z}_{(1)}\left\{\frac{\partial}{\partial d_j}(\tilde{\boldsymbol{D}}_1\tilde{\boldsymbol{\Gamma}}_1\tilde{\boldsymbol{\Gamma}}'_1\tilde{\boldsymbol{D}}_1)\right\}\boldsymbol{Z}'_{(1)} \;\text{ and }\; \tilde{\boldsymbol{T}}^{j}_{(1)} = \boldsymbol{Z}_{(1)}\tilde{\boldsymbol{D}}_1\left\{\frac{\partial}{\partial\gamma_j}(\tilde{\boldsymbol{\Gamma}}_1\tilde{\boldsymbol{\Gamma}}'_1)\right\}\tilde{\boldsymbol{D}}_1\boldsymbol{Z}'_{(1)}.$$

From standard arguments we have

$$\left.\begin{array}{l} \frac{1}{\sqrt{m}}\boldsymbol{X}'_{(1)j}\tilde{\boldsymbol{V}}^{-1}_{(1)}(\boldsymbol{y}-\boldsymbol{X}_{(1)}\boldsymbol{\beta}_1) \\ \frac{1}{2\sqrt{m}}[Tr(\tilde{\boldsymbol{V}}^{-1}_{(1)}\tilde{\boldsymbol{S}}^{j}_{(1)}) + (\boldsymbol{y}-\boldsymbol{X}_{(1)}\boldsymbol{\beta}_1)'(\tilde{\boldsymbol{V}}^{-1}_{(1)}\tilde{\boldsymbol{S}}^{j}_{(1)}\tilde{\boldsymbol{V}}^{-1}_{(1)})(\boldsymbol{y}-\boldsymbol{X}_{(1)}\boldsymbol{\beta}_1)] \\ \frac{1}{2\sqrt{m}}[-Tr(\tilde{\boldsymbol{V}}^{-1}_{(1)}\tilde{\boldsymbol{T}}^{j}_{(1)}) + (\boldsymbol{y}-\boldsymbol{X}_{(1)}\boldsymbol{\beta}_1)'(\tilde{\boldsymbol{V}}^{-1}_{(1)}\tilde{\boldsymbol{T}}^{j}_{(1)}\tilde{\boldsymbol{V}}^{-1}_{(1)})(\boldsymbol{y}-\boldsymbol{X}_{(1)}\boldsymbol{\beta}_1)] \end{array}\right| \Bigg\}_{\boldsymbol{\phi}_1=\boldsymbol{\phi}_{10}} = \boldsymbol{O}_p(1). \quad \text{(A.4)}$$

For $\nabla^2 L(\boldsymbol{\phi}_1)$ we have

$$\frac{1}{m}\nabla^2 L(\boldsymbol{\phi}_{10}) \to_p -I(\boldsymbol{\phi}_{10}), \quad \text{(A.5)}$$

where $I(\boldsymbol{\phi}_{10})$ is the Fisher information evaluated at $\boldsymbol{\phi}_{10}$. Using (A.4) and (A.5) the expansion in (A.3) becomes

$$D_m(\boldsymbol{u}) = \boldsymbol{O}_p(1)\boldsymbol{u} - \frac{1}{2}\boldsymbol{u}'\{I(\boldsymbol{\phi}_{10}) + o_p(1)\}\boldsymbol{u} - \frac{\lambda_m}{\sqrt{m}}\sum_{j=1}^{s}\bar{w}_j\mathrm{sgn}(\phi_{j0})u_j.$$

Since $I(\boldsymbol{\phi}_{10})$ is finite and positive definite (condition i), hence choosing a sufficiently large C, the second term dominates the first term uniformly in $||\boldsymbol{u}|| = C$. For the penalty term, if $\lambda_m/\sqrt{m} \to 0$ as $m\to\infty$, and since $\bar{w}_j = 1/\hat{\phi}_j \to 1/\phi_j$, it follows that

$$\frac{\lambda_m}{\sqrt{m}}\sum_{j=1}^{s}\bar{w}_j\mathrm{sgn}(\phi_{j0})u_j \to_p 0,$$

and thus is also dominated by the second term. Hence by choosing a sufficiently large C there exists a local maximum in the ball $\{(\boldsymbol{\phi}_{10} + \alpha_m\boldsymbol{u}, 0)' : ||\boldsymbol{u}|| \le C\}$ with probability with $1-\epsilon$, and hence there exists a local maximizer $\hat{\boldsymbol{\phi}} = (\hat{\boldsymbol{\phi}}_1, 0)$ of $\boldsymbol{\phi}_0 = (\boldsymbol{\phi}_{10}, 0)$ such that $||\hat{\boldsymbol{\phi}}_1 - \boldsymbol{\phi}_{10}|| = O_p(m^{-1/2})$. □

A.3 Proof of Theorem 2

Let $\boldsymbol{\phi} = (\boldsymbol{\beta}', \boldsymbol{d}', \boldsymbol{\gamma}')'$ denote the $k \times 1$ vector of unknown parameters, where $k = k_\beta + k_d + k_\gamma$, the sum of the lengths corresponding to each parameter. Let $\boldsymbol{\phi}_2 = (\boldsymbol{\beta}_2', \boldsymbol{d}_2', \boldsymbol{\gamma}_2')'$ be a vector of length $k_2 = k - s$, corresponding to the true zero values, where $k_2 = k_{\beta_2} + k_{d_2} + k_{\gamma_2}$.

Proof. It is sufficient to show that with probability tending to 1 as $m \to \infty$, for any $\boldsymbol{\phi}_1$ satisfying $||\boldsymbol{\phi}_1 - \boldsymbol{\phi}_{10}|| \leq Mm^{-1/2}$ and for some small $\epsilon_m = Mm^{-1/2}$ and for each $j = (s+1), \ldots, (k_{\beta_2} + k_{d_2})$, we have that

$$\begin{aligned} \frac{\partial}{\partial \phi_j} \boldsymbol{Q}(\boldsymbol{\phi}) < 0 \quad &\text{for} \quad 0 < \phi_j < \epsilon_m, \\ \frac{\partial}{\partial \phi_j} \boldsymbol{Q}(\boldsymbol{\phi}) > 0 \quad &\text{for} \quad -\epsilon_m < \phi_j < 0. \end{aligned} \tag{A.6}$$

Note that

$$\frac{\partial}{\partial \phi_j} \boldsymbol{Q}(\boldsymbol{\phi}) = \frac{\partial}{\partial \phi_j} L(\boldsymbol{\phi}) - \lambda_m \bar{w}_j \text{sgn}(\phi_j).$$

To show (A.6) consider the Taylor series expansion about $\partial L(\boldsymbol{\phi})/\partial \phi_j$, we have

$$\begin{aligned} \frac{\partial}{\partial \phi_j} \boldsymbol{Q}(\boldsymbol{\phi}) = \quad & \frac{\partial}{\partial \phi_j} L(\boldsymbol{\phi}_0) - \sum_{l=1}^{k} \frac{\partial}{\partial \phi_j \partial \phi_l} L(\boldsymbol{\phi}_0)(\phi_l - \phi_{l0}) \\ & + \frac{1}{2} \sum_{i=1}^{m} \sum_{l=1}^{k} \sum_{m=1}^{k} \frac{\partial^3}{\partial \phi_j \partial \phi_l \partial \phi_m} L_i(\boldsymbol{\phi}_*)(\phi_l - \phi_{l0})(\phi_m - \phi_{m0}) - \lambda_m \bar{w}_j \text{sgn}(\phi_j), \end{aligned} \tag{A.7}$$

where $\boldsymbol{\phi}_*$ lies between $\boldsymbol{\phi}$ and $\boldsymbol{\phi}_0$. Again the first order partial derivative for j^{th} term for each $\boldsymbol{\beta}$ and $\boldsymbol{d}$ are given by

$$\begin{aligned} \tfrac{1}{\sqrt{m}} \tfrac{\partial}{\partial \beta_j} L(\boldsymbol{\phi}_0) \quad &= \frac{1}{\sqrt{m}} \boldsymbol{X}_j' \boldsymbol{V}_0^{-1} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_0) = \boldsymbol{O}_p(1), \\ \tfrac{1}{\sqrt{m}} \tfrac{\partial}{\partial d_j} L(\boldsymbol{\phi}_0) \quad &= 0. \end{aligned}$$

where $\boldsymbol{X}_j$ corresponds to the j^{th} column of the stacked matrix $\boldsymbol{X}$. The second order derivatives in (A.7) follows

$$\left(\frac{1}{m}\right) \nabla^2 L(\boldsymbol{\phi}) \Big|_{\boldsymbol{\phi}=\boldsymbol{\phi}_0} \;\to\; E(\nabla^2 L(\boldsymbol{\phi})) \Big|_{\boldsymbol{\phi}=\boldsymbol{\phi}_0},$$

where $E(\nabla^2 L(\boldsymbol{\phi}))$ is given as

$$E\left(\nabla^2 L(\boldsymbol{\phi})\right) = E \begin{bmatrix} L_{\boldsymbol{\beta}\boldsymbol{\beta}} & L_{\boldsymbol{\beta}\boldsymbol{d}} & L_{\boldsymbol{\beta}\boldsymbol{\gamma}} \\ L'_{\boldsymbol{\beta}\boldsymbol{d}} & L_{\boldsymbol{d}\boldsymbol{d}} & L_{\boldsymbol{d}\boldsymbol{\gamma}} \\ L'_{\boldsymbol{\beta}\boldsymbol{\gamma}} & L'_{\boldsymbol{d}\boldsymbol{\gamma}} & L_{\boldsymbol{\gamma}\boldsymbol{\gamma}} \end{bmatrix},$$

where

$$E(L_{\boldsymbol{\beta\beta}}) = -\boldsymbol{X}'\tilde{\boldsymbol{V}}^{-1}\boldsymbol{X},$$

and $E(L_{\boldsymbol{\beta d}}), E(L_{\boldsymbol{\beta\gamma}})$ has j^{th} column

$$\left.\begin{array}{l} E\left\{L_{\boldsymbol{\beta d}}\right\}_j = -E[\boldsymbol{X}_j'(\tilde{\boldsymbol{V}}^{-1}\tilde{\boldsymbol{S}}^j\tilde{\boldsymbol{V}}^{-1})(\boldsymbol{y}-\boldsymbol{X\beta})] \\ E\left\{L_{\boldsymbol{\beta\gamma}}\right\}_j = -E[\boldsymbol{X}_j'(\tilde{\boldsymbol{V}}^{-1}\tilde{\boldsymbol{T}}^j\tilde{\boldsymbol{V}}^{-1})(\boldsymbol{y}-\boldsymbol{X\beta})] \end{array}\right| _{\boldsymbol{\phi}=\boldsymbol{\phi}_0} \Bigg\} = 0,$$

where $\tilde{\boldsymbol{S}}^j$ and $\tilde{\boldsymbol{T}}^j$ are block diagonal matrices of $\boldsymbol{S}_i$ and $\boldsymbol{T}_i$ given in (A.2). The expectation for the second order partial derivatives for $\boldsymbol{d}$ and $\boldsymbol{\gamma}$ has $(j,l)^{th}$ term

$$\begin{aligned} E\left\{L_{\boldsymbol{dd}}\right\}_{jl} &= -\mathrm{Tr}(\tilde{\boldsymbol{V}}^{-1}\tilde{\boldsymbol{S}}^j\tilde{\boldsymbol{V}}^{-1}\tilde{\boldsymbol{S}}^l) \\ E\left\{L_{\boldsymbol{\gamma\gamma}}\right\}_{jl} &= -Tr(\tilde{\boldsymbol{V}}^{-1}\tilde{\boldsymbol{T}}^j\tilde{\boldsymbol{V}}^{-1}\tilde{\boldsymbol{T}}^l) \\ E\left\{L_{\boldsymbol{d\gamma}}\right\}_{jl} &= -\mathrm{Tr}(\tilde{\boldsymbol{V}}^{-1}\tilde{\boldsymbol{S}}^j\tilde{\boldsymbol{V}}^{-1}\tilde{\boldsymbol{T}}^l), \end{aligned} \tag{A.8}$$

for $j = s+1,\ldots,(k_\beta+k_d)$, it can be shown that $\tilde{\boldsymbol{S}}^j$ or $\tilde{\boldsymbol{T}}^j$ when evaluated at $\phi_j = 0$ are zero matrices and the set of equations given in (A.8) simplifies to zero.

First, consider $\phi_j = \beta_j$ the expansion given in (A.7) yields

$$\begin{aligned} &\frac{1}{\sqrt{m}}\left(\frac{\partial}{\partial\beta_j}\boldsymbol{Q}(\boldsymbol{\phi})\right) = \frac{1}{\sqrt{m}}\Bigg(O_p(m^{1/2}) - m\sum_{l=1}^{k_\beta}\{\boldsymbol{X}_j'\boldsymbol{V}_0^{-1}\boldsymbol{X}_l + o_p(1)\}(\beta_l-\beta_{l0}) - m\sum_{l=k_\beta+1}^{k_d} o_p(1)(d_l-d_{l0}) - m\sum_{l=k_d+1}^{k_\gamma} o_p(1)(\gamma_l-\gamma_{l0}) \\ &+\sum_{i=1}^{m}\sum_{l=1}^{k_\beta}\sum_{m=k_\beta+1}^{k_d}\boldsymbol{X}_{ij}'(\boldsymbol{V}_{i*}^{-1}\boldsymbol{S}_{i*}^m\boldsymbol{V}_{i*}^{-1})\boldsymbol{X}_{il}(\beta_l-\beta_{l0})(d_m-d_{m0}) + \sum_{i=1}^{m}\sum_{l=1}^{k_\beta}\sum_{m=k_d+1}^{k_\gamma}\boldsymbol{X}_{ij}'(\boldsymbol{V}_{i*}^{-1}\boldsymbol{T}_{i*}^m\boldsymbol{V}_{i*}{}^{-1})\boldsymbol{X}_{il}(\beta_l-\beta_{l0})(\gamma_m-\gamma_{m0}) \\ &+\frac{1}{2}\sum_{i=1}^{m}\sum_{l=k_\beta+1}^{k_d}\sum_{l=k_\beta+1}^{k_d}\boldsymbol{X}_{ij}\frac{\partial\boldsymbol{V}_i^{-1}}{\partial d_l\partial d_m}(\boldsymbol{y}_i-\boldsymbol{X}_i\boldsymbol{\beta}_*)(d_l-d_{l0})(d_m-d_{m0}) \\ &+\frac{1}{2}\sum_{i=1}^{m}\sum_{l=k_d+1}^{k_\gamma}\sum_{m=k_d+1}^{k_\gamma}\boldsymbol{X}_{ij}\frac{\partial\boldsymbol{V}_i^{-1}}{\partial\gamma_l\partial\gamma_m}(\boldsymbol{y}_i-\boldsymbol{X}_i\boldsymbol{\beta}_*)(\gamma_l-\gamma_{l0})(\gamma_m-\gamma_{m0}) \\ &+\sum_{i=1}^{m}\sum_{l=k_\beta+1}^{k_d}\sum_{m=k_d+1}^{k_\gamma}\boldsymbol{X}_{ij}\frac{\partial\boldsymbol{V}_i^{-1}}{\partial d_l\partial\gamma_m}(\boldsymbol{y}_i-\boldsymbol{X}_i\boldsymbol{\beta}_*)(d_l-d_{l0})(\gamma_m-\gamma_{m0}) - \lambda_m\bar{w}_j\mathrm{sgn}(\beta_j)\Bigg), \end{aligned} \tag{A.9}$$

where $||\boldsymbol{\phi}_* - \boldsymbol{\phi}_0|| \le ||\boldsymbol{\phi}-\boldsymbol{\phi}_0||$. Since we are considering $||\boldsymbol{\phi}-\boldsymbol{\phi}_0|| \le Mn^{-1/2}$, (A.9) gives

$$\frac{1}{\sqrt{m}}\frac{\partial}{\partial\beta_j}\boldsymbol{Q}(\boldsymbol{\phi}) = -\lambda_m\frac{\bar{w}_j}{\sqrt{m}}\mathrm{sgn}(\beta_j) + O_p(1).$$

Since for $\beta_{j0} = 0$, we have $\bar{w}_j/\sqrt{m} = |\sqrt{m}\bar{\beta}_j|^{-1} = O_p(1)$, and $\lambda_m \to \infty$, the sign of the derivative is completely determined by that of β_j.

Now consider $\phi_j = d_j$. The Taylor series expansion in (A.7) gives

$$\frac{1}{\sqrt{m}}\left(\frac{\partial}{\partial d_j}\boldsymbol{Q}(\boldsymbol{\phi})\right) = \frac{1}{\sqrt{m}}\left(\mathbf{0} - m\sum_{l=1}^{k_\beta} o_p(1)(\beta_l - \beta_{l0}) - m\sum_{l=k_\beta+1}^{k_d} o_p(1)(d_l - d_{l0}) - m\sum_{l=k_d+1}^{k_\gamma} o_p(1)(\gamma_l - \gamma_{l0})\right.$$
$$+\frac{1}{2}\sum_{i=1}^{m}\sum_{l=k_\beta+1}^{k_d}\sum_{m=k_\beta+1}^{k_d}\frac{\partial L_i(d_j)}{\partial d_l \partial d_m}(d_l - d_{l0})(d_m - d_{m0}) + \frac{1}{2}\sum_{i=1}^{m}\sum_{l=k_d+1}^{k_\gamma}\sum_{m=k_d+1}^{k_\gamma}\frac{\partial L_i(d_j)}{\partial \gamma_l \partial \gamma_m}(\gamma_l - \gamma_{l0})(\gamma_m - \gamma_{m0})$$
$$+\sum_{i=1}^{m}\sum_{l=1}^{k_\beta}\sum_{m=1}^{k_\beta}\boldsymbol{X}'_{il}(\boldsymbol{V}_{i*}^{-1}\boldsymbol{S}_{i*}^{j}\boldsymbol{V}_{i*}^{-1})\boldsymbol{X}_{il}(\beta_l - \beta_{l0})(\beta_m - d\beta_{m0}) + \sum_{i=1}^{m}\sum_{l=k_\beta+1}^{k_d}\sum_{m=k_d+1}^{k_\gamma}\frac{\partial L_i(d_j)}{\partial d_l \partial \gamma_m}(d_l - d_{l0})(\gamma_m - \gamma_{m0})$$
$$\left.+\sum_{i=1}^{m}\sum_{l=1}^{k_\beta}\sum_{m=k_d+1}^{k_\gamma}\frac{\partial L_i(d_j)}{\partial \beta_l \partial \gamma_m}(\beta_l - \beta_{l0})(\gamma_m - \gamma_{m0}) + \sum_{i=1}^{m}\sum_{l=1}^{k_\beta}\sum_{m=k_\beta+1}^{k_d}\frac{\partial L_i(d_j)}{\partial \beta_l \partial d_m}(\beta_l - \beta_{l0})(d_m - d_{m0})\right), \qquad \text{(A.10)}$$

where

$$L_i(d_j) = \frac{\partial}{\partial d_j}L_i(\boldsymbol{\phi}_*) = \frac{1}{2}[-Tr(\boldsymbol{V}_{i*}^{-1}\boldsymbol{S}_{i*}^{j}) + (\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}_*)'(\boldsymbol{V}_{i*}^{-1}\boldsymbol{S}_{i*}^{j}\boldsymbol{V}_{i*}^{-1})(\boldsymbol{y}_i - \boldsymbol{X}_i\boldsymbol{\beta}_*)],$$

and $\boldsymbol{\phi}_*$ lies between $\boldsymbol{\phi}$ and $\boldsymbol{\phi}_0$. As above, (A.10) simplifies to

$$\frac{1}{\sqrt{m}}\frac{\partial}{\partial d_j}\boldsymbol{Q}(\boldsymbol{\phi}) = -\lambda_m\frac{\bar{w}_j}{\sqrt{m}}\text{sgn}(d_j).$$

Hence, for $d_{j0} = 0$, as $\bar{w}_j/\sqrt{m} = |\sqrt{m}\bar{d}_j|^{-1} = O_p(1)$, the sign of the derivative is again completely determined by that of d_j. This completes the proof. □

A.4 Proof of Theorem 3

Proof. We have from Theorem 1 shown that there exists a $\hat{\boldsymbol{\phi}}_1$ that is a local maximizer of $\boldsymbol{Q}(\boldsymbol{\phi}_1)$ such that $||\hat{\boldsymbol{\phi}}_1 - \boldsymbol{\phi}_{10}|| = O_p(m^{-1/2})$, and satisfies the set of penalized likelihood equations

$$\frac{\partial}{\partial \boldsymbol{\phi}_1}\boldsymbol{Q}(\boldsymbol{\phi})\Big|_{\boldsymbol{\phi}=(\hat{\boldsymbol{\phi}}_1,0)'} = \frac{\partial}{\partial \boldsymbol{\phi}_1}L(\boldsymbol{\phi})\Big|_{\boldsymbol{\phi}=(\hat{\boldsymbol{\phi}}_1,0)'} - \lambda_m\boldsymbol{h}(\hat{\boldsymbol{\phi}}_1) = 0,$$

where $\boldsymbol{h}(\hat{\boldsymbol{\phi}}_1) = (\bar{w}_1\text{sgn}(\hat{\phi}_1), \ldots, \bar{w}_s\text{sgn}(\hat{\phi}_s))'$ an $s \times 1$ vector where $\bar{w}_j = 0$ for $\phi_j = \gamma_j$. Using the Taylor series expansion and multiplying throughout by $1/m$, we have

$$\tfrac{1}{m}\nabla L(\boldsymbol{\phi}_{10}) - \{I(\boldsymbol{\phi}_{10}) + o_p(1)\}(\hat{\boldsymbol{\phi}}_1 - \boldsymbol{\phi}_{10}) - \tfrac{\lambda_m}{m}\boldsymbol{h}(\phi_{10}) = 0$$
$$\sqrt{m}\left\{(\hat{\boldsymbol{\phi}}_1 - \boldsymbol{\phi}_{10})I(\boldsymbol{\phi}_{10}) + \tfrac{\lambda_m}{m}\boldsymbol{h}(\phi_{10})\right\} = \tfrac{1}{\sqrt{m}}\nabla L(\boldsymbol{\phi}_{10}) \qquad .$$

Since $\text{E}\{\nabla L(\boldsymbol{\phi}_1)\} = 0$ as in the proof of Theorem 1, it follows from the multivariate central theorem that

$$\frac{1}{\sqrt{m}}\nabla L(\boldsymbol{\phi}_{10}) \to_d N(0, I(\boldsymbol{\phi}_{10})),$$

where $I(\boldsymbol{\phi}_{10})$ is as given in the proof of Theorem 1. Therefore

$$\sqrt{m}I(\phi_{10})\{(\hat{\boldsymbol{\phi}}_1 - \boldsymbol{\phi}_{10}) + \frac{\lambda_m}{m}I(\phi_{10})^{-1}\boldsymbol{h}(\phi_{10})\} \rightarrow_d N\{0, I(\phi_{10})\}.$$

This completes the proof. □

APPENDIX B

B.1 Further computational details on the EM algorithm

Omitting terms that do not involve $\boldsymbol{\phi}$, we may rewrite the expression in (3.3) as

$$\begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{d} \end{bmatrix}' \begin{bmatrix} \boldsymbol{X}'\boldsymbol{X} & \boldsymbol{X}'\boldsymbol{Z}\text{Diag}(\tilde{\boldsymbol{\Gamma}}\boldsymbol{b})(\mathbf{1}_q \otimes I_m) \\ (\mathbf{1}_q \otimes I_m)'\text{Diag}(\tilde{\boldsymbol{\Gamma}}\boldsymbol{b})\boldsymbol{Z}'\boldsymbol{X} & (\mathbf{1}_q \otimes I_m)'\text{Diag}(\tilde{\boldsymbol{\Gamma}}\boldsymbol{b})\boldsymbol{Z}'\boldsymbol{Z}\text{Diag}(\tilde{\boldsymbol{\Gamma}}\boldsymbol{b})(\mathbf{1}_q \otimes I_m) \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{d} \end{bmatrix}$$
$$-2\, \boldsymbol{y}' \begin{bmatrix} \boldsymbol{X} & \boldsymbol{Z}\text{Diag}(\tilde{\boldsymbol{\Gamma}}\boldsymbol{b})(\mathbf{1}_q \otimes I_m) \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{d} \end{bmatrix} + \lambda_m \left\{ \sum_{j=1}^{p} \frac{|\beta_j|}{|\bar{\beta}_j|} + \sum_{j=1}^{q} \frac{|d_j|}{|\bar{d}_j|} \right\} \quad \text{(B.1)}$$

After some matrix manipulation, part of the lower right block of the matrix in the quadratic form above can be written as

$$\text{Diag}(\tilde{\boldsymbol{\Gamma}}\boldsymbol{b})\boldsymbol{Z}'\boldsymbol{Z}\text{Diag}(\tilde{\boldsymbol{\Gamma}}\boldsymbol{b}) = \boldsymbol{W} \bullet \tilde{\boldsymbol{\Gamma}}\text{Diag}(\boldsymbol{b})\mathbf{1}\mathbf{1}'\text{Diag}(\boldsymbol{b})\tilde{\boldsymbol{\Gamma}}', \quad \text{(B.2)}$$

where $\bullet$ represents the Hadamard (element by element) product operator, $\mathbf{1}$ represents an $mq \times 1$ vector of ones, and $\boldsymbol{W} = \boldsymbol{Z}'\boldsymbol{Z}$ is a symmetric block diagonal matrix. Computing the outer product $\text{Diag}(\boldsymbol{b})\mathbf{1}\mathbf{1}'\text{Diag}(\boldsymbol{b})$, the expression given in (B.2) further simplifies to $\boldsymbol{W} \bullet \tilde{\boldsymbol{\Gamma}}\boldsymbol{b}\boldsymbol{b}'\tilde{\boldsymbol{\Gamma}}'$.

Using this simplification, and taking the conditional expectation of (B.1) yields a penalized quadratic objective function for $(\boldsymbol{\beta}, \boldsymbol{d})$ as

$$\boldsymbol{g}(\boldsymbol{\beta}, \boldsymbol{d}|\boldsymbol{\phi}^{(\omega)}) =$$

$$\begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{d} \end{bmatrix}' \begin{bmatrix} \boldsymbol{X}'\boldsymbol{X} & \boldsymbol{X}'\boldsymbol{Z}\text{Diag}(\tilde{\boldsymbol{\Gamma}}\hat{\boldsymbol{b}}^{(\omega)})(\mathbf{1}_q \otimes I_m) \\ (\mathbf{1}_q \otimes I_m)'\text{Diag}(\tilde{\boldsymbol{\Gamma}}\hat{\boldsymbol{b}}^{(\omega)})\boldsymbol{Z}'\boldsymbol{X} & (\mathbf{1}_q \otimes I_m)'\left(\boldsymbol{W} \bullet \tilde{\boldsymbol{\Gamma}}\hat{\boldsymbol{G}}^{(\omega)}\tilde{\boldsymbol{\Gamma}}'\right)(\mathbf{1}_q \otimes I_m) \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{d} \end{bmatrix}$$
$$-2\, \boldsymbol{y}' \begin{bmatrix} \boldsymbol{X} & \boldsymbol{Z}\text{Diag}(\tilde{\boldsymbol{\Gamma}}\hat{\boldsymbol{b}}^{(\omega)})(\mathbf{1}_q \otimes I_m) \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{d} \end{bmatrix} + \lambda_m \left\{ \sum_{j=1}^{p} \frac{|\beta_j|}{|\beta_j|} + \sum_{j=1}^{q} \frac{|d_j|}{|d_j|} \right\},$$

where $\hat{\boldsymbol{G}}^{(\omega)} = E(\boldsymbol{b}\boldsymbol{b}') = \boldsymbol{U}^{(\omega)} + \hat{\boldsymbol{b}}^{(\omega)}\hat{\boldsymbol{b}}^{\prime(\omega)}$, and $\boldsymbol{U}^{(\omega)}$ and $\hat{\boldsymbol{b}}^{(\omega)}$ are as given in (3.4). For a fixed $\boldsymbol{\gamma}$, we now minimize the objective function (B.3) to obtain the updated estimates for $(\boldsymbol{\beta}, \boldsymbol{d})$.

The next step is, for a fixed $(\boldsymbol{\beta}, \boldsymbol{d})$, to obtain a closed form expression for the estimate of $\boldsymbol{\gamma}$, the vector that relates to the correlation between the random effect parameters. Note that, if $\boldsymbol{\gamma} = \mathbf{0}$, then the random effects are mutually independent with the random effect covariance matrix $\boldsymbol{\Psi}$ being reduced to a simple diagonal form. Furthermore, if $\boldsymbol{d}_l = 0$ then $\gamma_{lr} = 0$ for $r = l+1, \ldots, q$. Hence, the elements of $\boldsymbol{\Gamma}$ and $\boldsymbol{d}$ are functionally related. Fixing $(\boldsymbol{\beta}, \boldsymbol{d})$ at its most recent update, we first rewrite $||\boldsymbol{y} - \boldsymbol{Z}\tilde{\boldsymbol{D}}\tilde{\boldsymbol{\Gamma}}\boldsymbol{b} - \boldsymbol{X}\boldsymbol{\beta}||^2$ in a quadratic form for $\boldsymbol{\gamma}$ and then compute its conditional expectation. After a bit of matrix manipulation, and omitting terms not involving $\boldsymbol{\gamma}$, we have that the objective function for $\boldsymbol{\gamma}$ is given by

$$g(\boldsymbol{\gamma}|\boldsymbol{\phi}^{(\omega)}) = \boldsymbol{\gamma}'\boldsymbol{P}^{(\omega)}\boldsymbol{\gamma} - 2\left\{(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'\boldsymbol{R}^{(\omega)} - \boldsymbol{T}^{(\omega)}\right\}\boldsymbol{\gamma}, \tag{B.3}$$

where $\boldsymbol{P}^{(\omega)} = E_{\boldsymbol{b}|\boldsymbol{y},\boldsymbol{\phi}^{(\omega)}}\{\boldsymbol{A}'\boldsymbol{A}\}$, $\boldsymbol{T}^{(\omega)} = E_{\boldsymbol{b}|\boldsymbol{y},\boldsymbol{\phi}^{(\omega)}}(\boldsymbol{A}'\boldsymbol{Z}\tilde{\boldsymbol{D}}\boldsymbol{b})$, and $\boldsymbol{R}^{(\omega)} = E_{\boldsymbol{b}|\boldsymbol{y},\boldsymbol{\phi}^{(\omega)}}(\boldsymbol{A})$. Where $\boldsymbol{A} = [\boldsymbol{A}_1', \ldots, \boldsymbol{A}_m']'$ represents a stacked matrix of $\boldsymbol{A}_i$, with each $\boldsymbol{A}_i$ an $n_i \times q(q-1)/2$ matrix, whose elements in each row are given by $A_{ij} = (b_{il}d_r z_{ijr} : l = 1, \ldots, (q-1), r = l+1, \ldots, q)$. Here, A_{ij} denotes the j^{th} row of the i^{th} matrix, which contains $q(q-1)/2$ elements. The appropriate minimizer of (B.3) is then given by

$$\boldsymbol{\gamma}_* = \left(\boldsymbol{P}^{(\omega)}\right)^{-}\left\{\boldsymbol{R}'^{(\omega)}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) - \boldsymbol{T}^{(\omega)}\right\}, \tag{B.4}$$

where A^- denotes the Moore-Penrose generalized inverse of A. Note that the matrices $\boldsymbol{P}^{(\omega)}$, $\boldsymbol{T}^{(\omega)}$ and $\boldsymbol{R}^{(\omega)}$ only involve first and second moments, i.e. $\hat{\boldsymbol{b}}^{(\omega)}$ and $\boldsymbol{U}^{(\omega)}$.

The optimization problem is now solved by minimizing the quadratic form (B.3), along with explicit solution (B.4) iteratively to complete the M-step. The final penalized likelihood estimates, $\hat{\boldsymbol{\phi}} = (\hat{\boldsymbol{\beta}}', \hat{\boldsymbol{d}}', \hat{\boldsymbol{\gamma}}')'$ are obtained by successive EM steps.

B.2 Computation of the M-step

Recently, Efron, Hastie, Johnstone and Tibshirani (2004) proposed the LARS (Least Angle Regression) algorithm, and showed that it can be used to obtain the entire solution path for LASSO estimates, while being computationally efficient. Zou (2006) showed that with minor changes to the design matrix, the LARS algorithm can be implemented to obtain the estimates for the regression coefficients under the adaptive LASSO penalty. Although we

can use the LARS algorithm to minimize the penalized quadratic form in our M-step via a pseudo "design matrix", it is not as advantageous to obtain the entire solution path here. This is due to the fact that the "design matrix" changes with every iterative step of the EM algorithm. Hence, we propose the use of a standard quadratic programming technique to obtain the penalized likelihood estimates for our parameters at each iterative step.

Given $\boldsymbol{\phi} = \boldsymbol{\phi}^{(\omega)}$ and a tuning parameter λ_m, we write $\boldsymbol{\beta} = \boldsymbol{\beta}^+ - \boldsymbol{\beta}^-$ with both $\boldsymbol{\beta}^+$ and $\boldsymbol{\beta}^-$ being non-negative, and only one is non-zero, and $|\boldsymbol{\beta}| = \boldsymbol{\beta}^+ + \boldsymbol{\beta}^-$, the optimization problem for $(\boldsymbol{\beta}, \boldsymbol{d})$ given in (B.3) is equivalent to

$$\begin{aligned}
&\text{minimize} \quad \begin{bmatrix} \boldsymbol{\beta}^+ \\ \boldsymbol{\beta}^- \\ \boldsymbol{d} \end{bmatrix}' \begin{bmatrix} \boldsymbol{X}_*'\boldsymbol{X}_* & \boldsymbol{X}_*'\boldsymbol{Z}\mathrm{Diag}(\tilde{\boldsymbol{\Gamma}}\hat{\boldsymbol{b}}^{(\omega)})(\mathbf{1}_q \otimes I_m) \\ (\mathbf{1}_q \otimes I_m)'\mathrm{Diag}(\tilde{\boldsymbol{\Gamma}}\hat{\boldsymbol{b}}^{(\omega)})\boldsymbol{Z}'\boldsymbol{X}_* & (\mathbf{1}_q \otimes I_m)'\left(\boldsymbol{W} \bullet \tilde{\boldsymbol{\Gamma}}\hat{\boldsymbol{G}}^{(\omega)}\tilde{\boldsymbol{\Gamma}}'\right)(\mathbf{1}_q \otimes I_m) \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}^+ \\ \boldsymbol{\beta}^- \\ \boldsymbol{d} \end{bmatrix} \\
&-2\left(\boldsymbol{y}' \begin{bmatrix} \boldsymbol{X}_* & \boldsymbol{Z}\mathrm{Diag}(\tilde{\boldsymbol{\Gamma}}\hat{\boldsymbol{b}}^{(\omega)})(\mathbf{1}_q \otimes I_m) \end{bmatrix} + \lambda_m \begin{bmatrix} \frac{1}{|\beta_1|}, & \cdots & , \frac{1}{|\beta_p|}, \frac{1}{|\beta_1|}, \cdots, \frac{1}{|\beta_p|}, \frac{1}{d_1}, & \cdots & , \frac{1}{d_q} \end{bmatrix}\right) \begin{bmatrix} \boldsymbol{\beta}^+ \\ \boldsymbol{\beta}^- \\ \boldsymbol{d} \end{bmatrix} \\
&\qquad\qquad \text{subject to} \\
&\qquad\qquad \boldsymbol{\beta}^+ \geq 0, \;\; \boldsymbol{\beta}^- \geq 0, \;\; \boldsymbol{d} \geq 0. \qquad\qquad \text{(B.5)}
\end{aligned}$$

where the matrix $\boldsymbol{X}_* = [\boldsymbol{X} \;\; -\boldsymbol{X}]$.

The minimization with respect to the expanded parameter $(\boldsymbol{\beta}^+, \boldsymbol{\beta}^-, \boldsymbol{d})$ is now a direct quadratic programming problem with $2p + q$ total parameters and $2p + q$ total linear constraints.

APPENDIX C

C.1 Brief Description of CASTNet Data

A complete description of the data can be found on the EPA website: **http://www.epa.gov/castnet**. The data used here is a subset of the complete data and consists of 826 observation from 15 relevant sites from 2000 to 2004 across the eastern United States. The map shown in Figure 1 marks the relevant sites we have used for this analysis. Here is a list describing briefly our response variable and the 16 predictors.

Y	LOG(TNO)_3, Log of Total Nitrate Concentration (μmol/m^3)
x_1	SO_4, Sulphate Concentration (μmol/m^3)
x_2	NH_4, Ammonia Concentration (μmol/m^3)
x_3	O_3, Maximum Ozone (ppb, parts per billion)
x_4	T, Average Temperature ($^\circ C$)
x_5	T_d, Average Due Point Temperature ($^\circ C$)
x_6	RH, Average Relative Humidity (%)
x_7	SR, Average Solar Radiation (W/m^2)
x_8	WS, Average Wind Speed (m/sec)
x_9	P, Total Precipitation ($mm/month$)
$l(t)$	Time of measurement in months $(1, \ldots, 60)$ from 2000-2004
$s_j(t)$	$\text{Sin}(\frac{2\pi jt}{12})$ where $j = 1, 2, 3$
$c_j(t)$	$\text{Cos}(\frac{2\pi jt}{12})$ where $j = 1, 2, 3$

Figure 3 plots the predicted mean overlayed with the observed values of $\log(\text{TNO}_3)$ using our penalized likelihood estimates for the fixed effects, for 4 specific sites of interest. Though it seems to fit well, we see that in some sites it tends to underestimate (DCP 114) while overestimating (PNF 126) in others. This further reiterates that the random effects are important to account for heterogeneity between the sites. We can also see from Figure 2 that a single cycle $(s_1(t), c_1(t))$ seems to be sufficient to describe the seasonal trend.

Figure 1: The location of the 15 Sites that were used for our analysis. The ∇ represents the 4 sites used for the overlay plots in Figure 3.

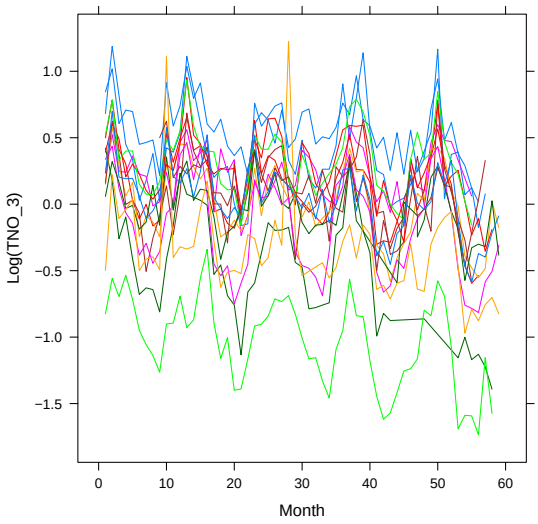

Figure 2: Site (individual) profile plot to assess the seasonal trend in Nitrate concentration over each 12 month period, for the CASTNet dataset.

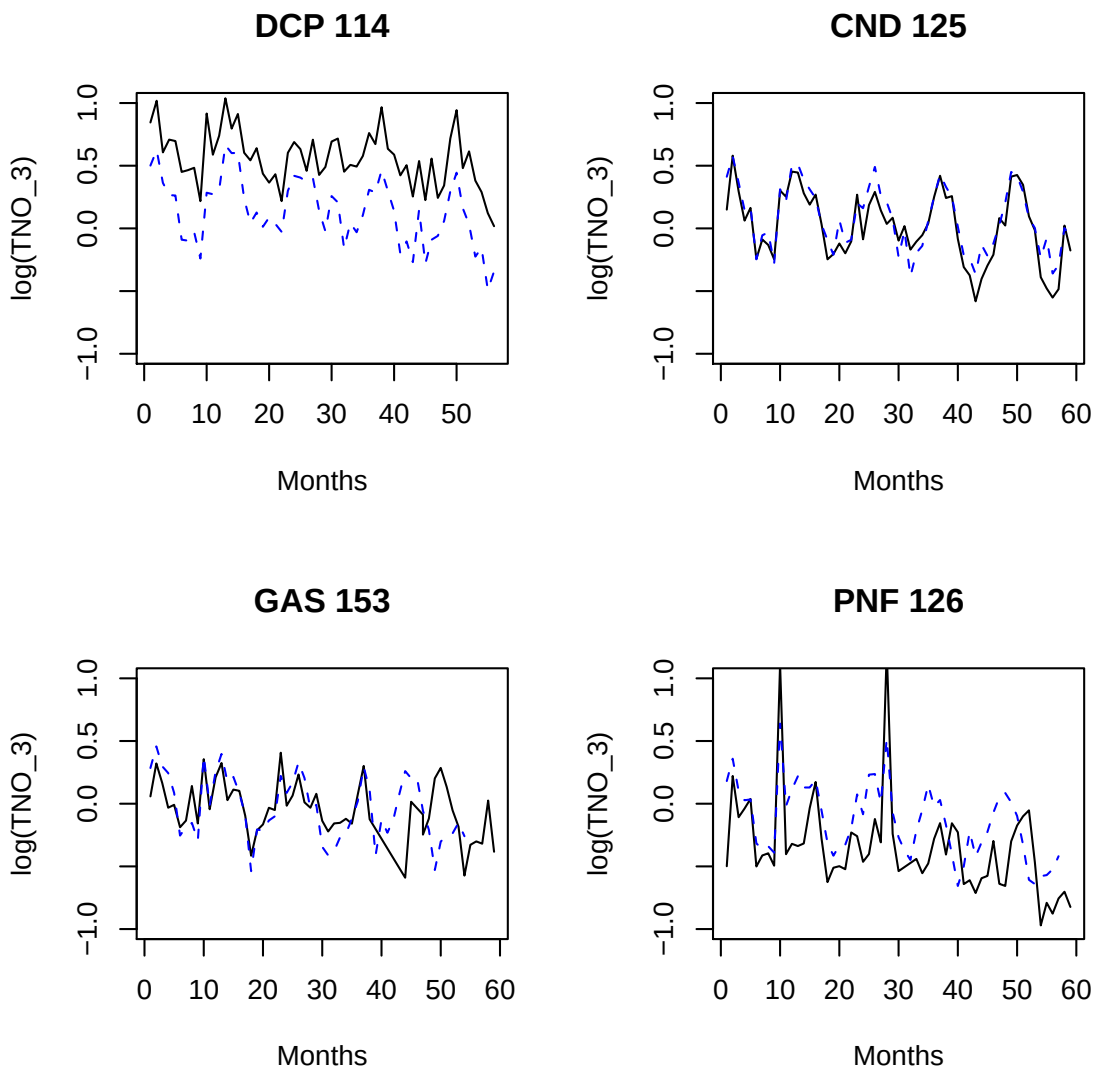

Figure 3: Plot of the observed $\mathrm{LOG(TNO)}_3$ concentration represented by the solid line, overlayed with the fitted values for the fixed effects model selected using our proposed method for 4 centers, for the CASTNet dataset.