



Published in final edited form as:

*Biometrics*. 2010 December ; 66(4): 1069–1077. doi:10.1111/j.1541-0420.2010.01391.x.

## 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.

### 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,

---

Correspondence Author: Howard D. Bondell bondell@stat.ncsu.edu Telephone: (919)515-1914; Fax: (919)515-1169.

Supplementary Materials

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

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  $i = 1, 2, \dots, m$  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  $\mathbf{y}_i$  be an  $n_i \times 1$  response for subject  $i$ . Let  $\mathbf{X}_i$  be the  $n_i \times p$  design matrix of explanatory variables, and  $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)'$  be the regression parameter vector. Let  $\mathbf{b}_i^* = (b_{i1}^*, \dots, b_{iq}^*)'$  be a  $q \times 1$  vector of subject-specific random effects with  $\mathbf{b}_i^* \sim N(0, \sigma^2 \boldsymbol{\Psi})$ , and assumed independent across subjects. Denote  $\mathbf{Z}_i$  as the  $n_i \times q$  design matrix corresponding to the random effects. Often one sets  $\mathbf{Z}_i = \mathbf{X}_i$ , but it is not necessary. Then, a general class of LME models can be written as

$$\mathbf{y}_i = \mathbf{X}_i \boldsymbol{\beta} + \mathbf{Z}_i \mathbf{b}_i^* + \boldsymbol{\varepsilon}_i, \quad (1.1)$$

where the errors  $\boldsymbol{\varepsilon}_i$ 's are independently distributed  $N(0, \sigma^2 \mathbf{I}_{n_i})$  and independent of the  $\mathbf{b}_i^*$ '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  $\Psi$  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  $\Psi$  (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; Pinheiro 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,  $\Psi$ , of the random effects as  $\Psi = \mathbf{D}\mathbf{\Gamma}\mathbf{\Gamma}'\mathbf{D}$ , where  $\mathbf{D} = \text{diag}(d_1, d_2, \dots, d_q)$  is a diagonal matrix, and  $\mathbf{\Gamma}$ , whose  $(l, r)^{th}$  element is denoted by  $\gamma_{lr}$ , is a  $q \times q$  lower triangular matrix with 1's on the diagonal. This decomposition in terms of  $\mathbf{D}$  and  $\mathbf{\Gamma}$  is unique, and leads to a non-negative definite matrix  $\Psi$ . Given the decomposition, the re-parameterized LME model can be written as

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{D}\mathbf{\Gamma}\mathbf{b}_i + \boldsymbol{\varepsilon}_i, \quad (2.1)$$

where we assume  $y_i$  has been centered and the predictors have been standardized so that,  $\mathbf{X}_i'\mathbf{X}_i$  and  $\mathbf{Z}_i'\mathbf{Z}_i$  represent the correlation matrices, and  $\mathbf{b}_i = (b_{i1}, \dots, b_{iq})'$  is a  $q \times 1$  vector of independent  $N(0, \sigma^2\mathbf{I}_q)$ . The covariance matrix of  $\mathbf{b}_i^*$ , is now expressed as a function of  $\mathbf{d} = (d_1, d_2, \dots, d_q)'$ , and the  $q(q-1)/2$  free elements of  $\mathbf{\Gamma}$ , denoted by the vector  $\boldsymbol{\gamma} = (\gamma_{lr} : l = 1, \dots, q : r = l+1, \dots, q)'$ . We denote  $\boldsymbol{\phi} = (\boldsymbol{\beta}', \mathbf{d}', \boldsymbol{\gamma}')$ , 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  $l^{\text{th}}$  column and  $l^{\text{th}}$  row of  $\Psi$  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  $X_i$  and  $Z_i$ , the distribution of  $y_i$  follows a normal distribution with mean  $X_i\beta$ , and variance  $V_i = \sigma^2 (Z_i D \Gamma \Gamma' D Z_i' + I_{n_i})$ . Dropping constant terms, the log-likelihood function is given by

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

where  $\tilde{V} = \text{Diag}(V_1, \dots, V_m)$  a block diagonal matrix of  $V_i$ 's, and  $\mathbf{y} = (\mathbf{y}'_1, \dots, \mathbf{y}'_m)'$ ,  $\mathbf{X} = [X'_1, \dots, X'_m]$  are the stacked  $y_i$  and  $X_i$ , respectively.

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

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

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

We now maximize the conditional expectation of (2.3) along with a penalty function on  $\beta$  and  $\mathbf{d}$ , to decide whether to include or exclude a predictor. Dropping out the terms which do not involve either  $\beta$  or  $\mathbf{d}$  is then equivalent to minimizing the conditional expectation of

$$\|\mathbf{y} - \mathbf{Z} \tilde{D} \tilde{\Gamma} \mathbf{b} - \mathbf{X}\beta\|^2 \text{ plus the penalty term.}$$

## 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

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

where  $\lambda_m$  is a non-negative regularization parameter,  $\bar{w}_j$  are adaptive weights, typically  $\bar{w}_j = 1/|\hat{\beta}_j|$  with  $\hat{\beta}_j$  the ordinary least squares estimate. As  $\lambda_m$  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

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

Here  $\tilde{\beta}$  is the generalized least squares estimate of  $\beta$ , and  $\tilde{d}$  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

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

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

### 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  $y_i$  plus the unobserved random parameters. We adopt the EM algorithm, in that, we compute the conditional expectation of  $Q_c(\phi|y, b)$  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  $b$  given  $\phi$  and  $y$  is,  $b|y, \phi \sim N(\hat{b}, U)$  where the mean and variance are given by,

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

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

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

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

$$g(\phi|\phi^{(\omega)}) = E_{b|y, \phi^{(\omega)}} \left\{ \|y - X\beta - Z \text{Diag}(\tilde{\Gamma} b) (1_q \otimes I_m) d\|^2 \right\} + \lambda_m \left( \sum_{j=1}^p \frac{|\beta_j|}{|\beta_j|} + \sum_{j=1}^q \frac{|d_j|}{|d_j|} \right). \quad (3.5)$$

Then, for the M-step, we minimize the objective function,  $g(\phi|\phi^{(\omega)})$  with respect to  $(\beta', d', \gamma)'$ . This optimization within the M-step is done by iterating between  $\gamma$  and the vector  $(\beta', d')'$ . The optimization iteration for  $\gamma$  is closed form, while the iteration for  $(\beta', d')'$  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  $(\beta^{(\omega+1)'}, d^{(\omega+1)'}, \gamma^{(\omega+1)'})'$  and also  $\sigma^{2(\omega+1)}$ , and return to the E-step. Upon convergence of the full EM algorithm, we obtain our final estimates  $\hat{\phi} = (\hat{\beta}', \hat{d}', \hat{\gamma}')'$ .

**3.2.2 Choice of tuning parameter**—The EM algorithm described above applies to a fixed value of  $\lambda_m$ . In practice,  $\lambda_m$  is chosen on a grid and the solution is obtained for each  $\lambda_m$ . Next, we must choose from among the candidate values of  $\lambda_m$  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

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

where  $L(\hat{\phi})$  is the obtained value of  $L(\phi)$ , as in (2.2), using the estimate  $\hat{\phi}$  obtained for that value of  $\lambda_m$ . We take the degrees of freedom,  $df_{\lambda_m}$ , as the number of non-zero coefficients in  $\hat{\phi}$ . 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  $BIC_{\lambda_m}$  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  $\phi = (\beta', d', \gamma)'$  and let  $\bar{\phi}$  denote an initial  $\sqrt{m}$  consistent estimator of  $\phi$ . Let  $Q(\phi)$  denote the penalized log-likelihood function with  $L(\phi)$  is as given in (2.2), then

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

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

(4.1)

Denote the true value of  $\phi$  as

$$\phi_0 = (\phi_{10}, \dots, \phi_{k0})' = (\phi'_{10}, \phi'_{20})'$$
(4.2)

where  $\phi_{10} = (\beta'_{10}, d'_{10}, \gamma'_{10})'$  is an  $s \times 1$  vector whose components are non-zero and let  $\phi_{20}$  be the  $(k - s)$  remaining components of  $\phi_0$ , so that  $\phi_{20} = 0$ . In a similar manner we also

decompose  $\phi$  itself as  $\phi = (\phi'_1, \phi'_2)'$ . 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  $\phi = (\phi'_1, 0)'$ , that is fixing  $\phi_2 = 0$ . Let  $L(\phi_1)$ ,  $Q(\phi_1)$  denote the log-likelihood and the penalized log-likelihood of the first  $s$  components of  $\phi$  given by

$$L(\phi_1) \equiv L \left\{ \begin{pmatrix} \phi_1 \\ 0 \end{pmatrix} \right\} = -\frac{1}{2} \log |\tilde{V}_{(1)}| - \frac{1}{2} (\mathbf{y} - \mathbf{X}_{(1)}\beta)' (\tilde{V}_{(1)})^{-1} (\mathbf{y} - \mathbf{X}_{(1)}\beta),$$

$$Q(\phi_1) \equiv Q \left\{ \begin{pmatrix} \phi_1 \\ 0 \end{pmatrix} \right\} = L(\phi_1) - \lambda_m \sum_{j=1}^s \bar{w}_j (|\phi_j|),$$
(4.3)

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

**Theorem 1.** Let  $\phi = (\phi'_1, 0)'$ , and the observations follow the LME model (2.1) satisfying conditions (i) – (iv) given in Web Appendix A. If  $\lambda_m / \sqrt{m} \rightarrow 0$ , then there exists a local

maximizer  $\hat{\phi} = \begin{pmatrix} \hat{\phi}_1 \\ 0 \end{pmatrix}$  of  $Q \left\{ \begin{pmatrix} \phi_1 \\ 0 \end{pmatrix} \right\}$  such that  $\hat{\phi}_1$  is  $\sqrt{m}$  consistent for  $\phi_{10}$ .

**Theorem 2.** Let the observations follow the LME model (2.1) satisfying conditions (i) – (iv) given in Web Appendix A. If  $\lambda_m \rightarrow \infty$ , then with probability tending to 1 for any given  $\phi_1$  satisfying  $\|\phi_1 - \phi_{10}\| \leq Mm^{-1/2}$  and some constant  $M > 0$ ,

$$Q \left\{ \begin{pmatrix} \phi_1 \\ 0 \end{pmatrix} \right\} = \max_{\|\phi_2\| \leq Mm^{-1/2}} Q \left\{ \begin{pmatrix} \phi_1 \\ \phi_2 \end{pmatrix} \right\}.$$
(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  $\hat{\phi}_2 = 0$ . 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  $\lambda_m \rightarrow \infty$  and  $\lambda_m/\sqrt{m} \rightarrow 0$ , we have

$$\sqrt{m}I(\phi_{10})\left\{\widehat{\phi}_1 - \phi_{10} + \left(\frac{\lambda_m}{m}\right)I^{-1}(\phi_{10})\mathbf{h}(\phi_{10})\right\} \rightarrow_d N(0, I(\phi_{10})) \quad (4.5)$$

where  $\mathbf{h}(\phi_{10}) = (\bar{w}_1 \text{sgn}(\phi_{10}), \dots, \bar{w}_s \text{sgn}(\phi_{s0}))'$  an  $s \times 1$  vector, and  $I(\phi_{10})$  is the Fisher information knowing that  $\phi_2 = 0$ .

*Remark 2.* From Theorem 2 and 3 as  $\lambda_m \rightarrow \infty$  and  $\lambda_m/\sqrt{m} \rightarrow 0$ , we can say that our penalized estimator enjoys the oracle property in that asymptotically it performs as well as the oracle estimators, knowing  $\phi_2 = 0$ . In particular, to first order,

$$\sqrt{m}(\widehat{\phi}_1 - \phi_{10}) \rightarrow N(0, I^{-1}(\phi_{10})).$$

## 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  $\phi_1$  knowing  $\phi_2 = 0$ .

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

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

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

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

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

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



such that there are 7 unimportant predictors for the fixed effects and 1 unimportant predictor for the random effects. The covariates  $x_{ijk}$  for,  $k = 1, \dots, 9$  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  $k = 1, \dots, 9$ , are generated from a uniform  $(-2, 2)$  distribution. We set  $\mathbf{Z}_i = \mathbf{X}_i$  plus a random intercept term. The true model is then given by

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

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 comparisons, model comparisons and validation are made based on the Kullback-Leibler discrepancy (Kullback and Leibler, 1951) given by

$$KLD = E \left\{ \log f(Y, \mathbf{X}, \mathbf{Z} | \phi_0) - \log f(Y, \mathbf{X}, \mathbf{Z} | \hat{\phi}) \right\}. \quad (5.5)$$

The joint density  $f(Y, \mathbf{X}, \mathbf{Z} | \phi_0)$  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  $f(Y, \mathbf{X}, \mathbf{Z} | \hat{\phi})$  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 (Tibshirani, 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 ( $\text{TNO}_3$ ) 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 ( $\text{NO}_2$ ) 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  $\log(\text{TNO}_3)$  rather than  $\text{TNO}_3$ , as in previous analyses. To build the relationship, we consider the following variables within the mixed model framework: sulfate ( $\text{SO}_4$ ), ammonium ( $\text{NH}_4$ ), ozone ( $\text{O}_3$ ), 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  $\log(\text{TNO}_3)$  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  $s_j(t) = \sin\left(\frac{2\pi jt}{12}\right)$  and  $c_j(t) = \cos\left(\frac{2\pi jt}{12}\right)$  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  $\mathbf{Z}_i = \mathbf{X}_i$  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  $\varepsilon_i \sim N(0, \sigma^2 \Sigma_i)$  for some  $\Sigma_i$ . 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  $\tilde{\Sigma}$  be the block diagonal matrix of  $\Sigma_i$ , for known  $\tilde{\Sigma}$ , we replace  $\frac{N+mq}{2} \log \sigma^2$  and  $\|y - Z \tilde{D} \tilde{\Gamma} b - X\beta\|^2$  in (2.5) by  $\left\{ \frac{N+mq}{2} \log \sigma^2 + (1/2) \log |\tilde{\Sigma}| \right\}$  and  $\|\tilde{\Sigma}^{-1/2} y - \tilde{\Sigma}^{-1/2} Z \tilde{D} \tilde{\Gamma} b - \tilde{\Sigma}^{-1/2} X\beta\|^2$  respectively. After this transformation, the remainder follows by redefining  $(X, Z, y)$ . If there are unknown parameters in  $\Sigma$ , the process must then iterate between estimation of the fixed and random effect parameters given  $\Sigma$ , and estimation of the parameters in  $\Sigma$  given the fixed and random effects. This must be done separately for each tuning parameter  $\lambda_m$ . 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.

## Acknowledgments

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. Petrov, BN.; Csaki, F., editors. Information theory and an extension of maximum likelihood principle.. Second international symposium on information theory. 1973.
- Breiman L. Heuristics of instability and stabilization in model selection. *Annals of Statistics*. 1996; 24:2350–2383.
- Bondell HD, Reich BJ. Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with OSCAR. *Biometrics*. 2008; 64:115–123. [PubMed: 17608783]
- Chen Z, Dunson D,B. Random effects selection in linear mixed models. *Biometrics*. 2003; 59:762–769. [PubMed: 14969453]
- Diggle, P.; Liang, K.; Zeger, S. *Analysis of Longitudinal Data*. Oxford Press; 1994.
- Efron B, Hastie T, Johnstone I, Tibshirani R. Least angle regression. *Annals of Statistics*. 2004; 32:407–499.
- Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*. 2001; 96:1348–1360.
- Ghosh SK, Bhave PV, Davis JM, Lee H. Spatio-temporal analysis of total nitrate concentrations using dynamic statistical models. *Journal of the American Statistical Association*. 2009 To appear.
- Hannan EJ, Quinn BG. The determination of the order of an autoregression. *Journal of the Royal Statistical Society B*. 1979; 41:190–195.
- Jiang J, Rao JS. Consistent procedures for mixed linear model selection. *Sankhya A*. 2003; 65:23–42.
- Jiang J, Rao JS, Gu Z, Nguyen T. Fence methods for mixed models selection. *Annals of Statistics*. 2008; 36:1669–1692.
- Kinney S,K, Dunson D,B. Fixed and random effects selection in linear and logistic models. *Biometrics*. 2007; 63:690–698. [PubMed: 17403104]
- Kullback S, Leibler RA. On information and sufficiency. *Annals of Mathematical Statistics*. 1951; 22:72–86.
- Laird NM, Ware JL. Random-effects models for longitudinal data. *Biometrics*. 1982; 38:963–974. [PubMed: 7168798]
- Laird NM, Lange N, Stram D. Maximum likelihood computations with repeated measures: application of the EM algorithm. *Journal of the American Statistical Association*. 1987; 82:97–105.
- Lange N, Laird NM. The effect of covariance structures on variance estimation in balance growth-curve models with random parameters. *Journal of the American Statistical Association*. 1989; 84:241–247.
- Lee H, Ghosh SK. A re-parametrization approach for dynamic space-time models. *Journal of Statistical Theory and Practice*. 2008; 2:1–14. [PubMed: 21593998]
- Lindstrom MJ, Bates DM. Newton-Raphson and EM algorithms for linear mixed-effects models for repeated measures data. *Journal of the American Statistical Association*. 1988; 83:1014–1022.
- Malm WC, Schichtel BA, Pitchford ML, Ashbaugh LL, Eldred RA. Spatial and monthly trends in speciated fine particle concentration in the United States. *Journal of Geophysical Research*. 2004; 109:D03306.
- Miller, A. *Subset Selection in Regression*. 2nd ed.. Chapman & Hall/ CRC; 2002.
- Morell CH, Pearson JD, Brant LJ. Linear transformations of linear mixed-effects models. *The American Statistician*. 1997; 51:338–343.
- Nie L. Convergence rate of MLE in generalized linear and nonlinear mixed-effects models: Theory and applications. *Journal of Statistical Planning and Inference*. 2007; 137:1787–1804.
- Niu F, Pu P. Selecting mixed-effects models based on generalized information criterion. *Journal of Multivariate Analysis*. 2006; 97:733–758.
- Pinheiro J, Bates D. Unconstrained parameterizations for variance-covariance matrices. *Statistics and Computing*. 1996; 6:289–286.
- Rao C,R, Wu Y. A strongly consistent procedure for model selection in regression problems. *Biometrika*. 1989; 76:369–374.
- Schwarz G. Estimating the dimension of a model. *Annals of Statistics*. 1978; 6:461–464.
- Shao J. An asymptotic theory for linear model selection. *Statistica Sinica*. 1997; 7:221–264.

- Smith M, Kohn R. Parsimonious covariance matrix estimation for longitudinal data. *Journal of the American Statistical Association*. 2002; 97:1141–1153.
- Tibshirani R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B*. 1996; 58:267–288.
- Vaida F, Blanchard S. Conditional Akaike information for mixed-effects models. *Biometrika*. 2005; 92:351–370.
- Wolfinger RD. Covariance structure selection in general mixed models. *Communications in Statistics, Simulation and Computation*. 1993; 22:1079–1106.
- Yang Y. Can the strengths of AIC and BIC be shared? A conflict between model identification and regression estimation. *Biometrika*. 2005; 92:937–950.
- Yu S, Dennis R, Roselle S, Nenes A, Walker J, Eder B, Schere K, Swall J, Robarge W. An assessment of the ability of three-dimensional air quality models with current thermodynamic equilibrium models to predict aerosol NO<sub>3</sub>. *Journal of Geophysical Research*. 2005; 110:D07S13.
- Zou H. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*. 2006; 101:1418–1429.
- Zou H, Hastie T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society B*. 2005; 67:301–320.
- Zou H, Hastie T, Tibshirani R. On the degrees of freedom of the lasso. *Annals of Statistics*. 2007; 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       | <i>KLD</i> (S.E.) | R.E. | %Correct | %CF | %CR |
|----------|----------|--------------|-------------------|------|----------|-----|-----|
|          | Oracle   | -            | 9.70 (0.343)      | -    | -        | -   | -   |
| 1        | M-ALASSO | BIC          | 10.94(0.475)      | 0.88 | 71       | 73  | 79  |
|          | EGIC     | BIC          | 13.91(0.583)      | 0.69 | 47       | 56  | 52  |
|          | REMLIC   | AIC          | 15.51(0.567)      | 0.63 | 19       | 21  | 62  |
|          | REMLIC   | BIC          | 12.48(0.642)      | 0.77 | 59       | 59  | 68  |
|          | STEPWISE | AIC          | 16.01(0.611)      | 0.60 | 13       | 15  | 62  |
|          | 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   |
| 2        | 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  |
|          | REMLIC   | AIC          | 11.93(0.432)      | 0.72 | 31       | 34  | 74  |
|          | REMLIC   | BIC          | 10.18(0.415)      | 0.77 | 77       | 79  | 81  |
|          | STEPWISE | AIC          | 12.87(0.501)      | 0.61 | 26       | 28  | 74  |
|          | 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  |

| Example | Method   | Tuning | $KLD(S.E.)$  | R.E. | %Correct | %CF | %CR |
|---------|----------|--------|--------------|------|----------|-----|-----|
| 3       | REMLIC   | AIC    | 28.87(2.231) | 0.46 | 12       | 14  | 68  |
|         | REMLIC   | BIC    | 23.39(1.872) | 0.57 | 53       | 54  | 73  |
|         | STEPWISE | AIC    | 29.58(2.893) | 0.47 | 8        | 11  | 68  |
|         | 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  |



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.

**Table 3**

| Variables | Int  | SO <sub>4</sub> | NH <sub>4</sub> | O <sub>3</sub> | T | T <sub>d</sub> | RH    | SR | WS | P     | I(t)  | s <sub>1</sub> (t) | c <sub>1</sub> (t) | s <sub>2</sub> (t) | c <sub>2</sub> (t) | s <sub>3</sub> (t) | c <sub>3</sub> (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                  | 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                  | 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).

| Method   | Tuning | Variables Selected   |                                     | CV Value |
|----------|--------|--|-------------------------------------|----------|
|          |        | Fixed  | Random                              |          |
| 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 <sub>4</sub> | NH <sub>4</sub> | O <sub>3</sub> | T | T <sub>d</sub> | RH    | SR | WS | P     | I(t)  | s <sub>1</sub> (t) | c <sub>1</sub> (t) | s <sub>2</sub> (t) | c <sub>2</sub> (t) | s <sub>3</sub> (t) | c <sub>3</sub> (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                  |