# Regularization and Classification of Linear Mixed Models via the Elastic Net Penalty with Application to the Good Judgment Project*

Jonathan Sidi[1], Ya'acov Ritov[1] and Lyle Ungar[2]

[1]Department of Statistics, Hebrew University in Jerusalem
[2]Computer and Information Statistics, University of Pennsylvania

July 29, 2015

Corresponding Author: Jonathan Sidi, Email: yoni.sidi@mail.huji.ac.il

**Keywords**: Linear mixed model; Elastic Net penalty; Variable selection; Probabilistic forecast

*This research was supported by a research contract to the University of Pennsylvania and the University of California from the Intelligence Advanced Research Projects Activity (IARPA) via the Department of Interior National Business Center contract number D11PC20061. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions expressed herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government.

## Abstract

Advances in the field of model selection and prediction via regularization has forged the ability of a variety of disciplines to classify and model large-scale data. Widely used methods which apply penalties in classification are the Least Absolute Shrinkage and Selection Operator (LASSO), the Adaptive LASSO and the Elastic Net. These methods have predominately been used to classify problems of Generalized Linear Models (GLM) in which the dependency of the covariance structure is assumed to be independent. This assumption is not commonly met in practical data and the ability to model such dependencies is integral in fitting the data correctly, such data is modeled using Linear Mixed Models (LMM). Recent research applying LASSO and Adaptive LASSO to LMM's has produced promising initial results of identifying both the random and fixed effects found in data, proving both consistency and an oracle optimality. However, an inherent drawback to those variable selection methods is their performance under high correlation between covariates. To overcome this we introduce the Elastic Net penalty to LMM selection. This penalty has been found to reduce the prediction error in data with high correlation between variables; such a characteristic can be utilized in more complex data designs while optimizing the LMM problem. Findings are tested through simulations and a case study using data accumulated in an longitudinal study where probabilistic forecasts are derived from crowd sentiment. The data structure consists repeated measures and a large number of fixed and random covariates.

# 1 Introduction

Generalized Linear Mixed models (GLMM) Breslow and Clayton (1993) have been applied in a variety of fields to study data designs with between-subject variation. Such designs include longitudinal, repeated measures and clustered data and have been studied thoroughly in the low dimension setting, e.g., Bates (2010) and Searle et al. (1992). In these settings the linear predictor contains in addition to fixed effects, found in Generalized Linear Models (GLM), latent random effects which capture the unique design pertaining to the data. These random effects usually are assumed to have a centered parametric distribution belonging to the exponential family.

Advances in the field of model selection and prediction via regularization, using different penalty terms, has forged the ability of a variety of disciplines to classify and model large-scale data. Widely used methods which apply penalties in classification are the Least Absolute Shrinkage and Selection Operator (LASSO), the Adaptive LASSO and the Elastic Net. These methods have predominately been used to classify problems of GLM, Friedman et al. (2010) and Van de Geer (2008), in which the dependency of the covariance structure is assumed to be independent. This assumption in practical data is not commonly met and the ability to model such dependencies is integral in fitting the data correctly, such data is modeled using Linear Mixed Models (LMM) and GLMMs.

Recent research Schelldorfer et al. (2011),Fan and Li (2012) and Bondell et al. (2010) apply the LASSO, SCAD and Adaptive LASSO respectively to LMMs and have produced promising initial results of identifying both the random and fixed effects found in data, proving both consistency and an oracle optimality. These articles apply a penalty to the LMM optimization, while achieving this each in a distinct approach. This paper, proposes a new algorithm, LMMEN that attempts to utilize the advantages of each method and submit a new type of penalty which better captures the design of the LMM.

## 1.1 Model

The GLMM is defined as having $m$ subjects in the sample. For the $i$th subject the response variable is denoted as $y_{ij}$ for the $j$th observation, where $j = 1 \dots n_i$ and let $N = \sum_{i=1}^{m} n_i$. The training data $\mathbf{X}$ can be defined as two groups of covariates: the fixed effects covariates vector denoted as $x_{ij}$ with dimensions $p \times 1$ and the random effects covariates vector denoted as $z_{ij}$ with dimensions $q \times 1$.

$y_{ij}$ are assumed to be conditionally independent given the subject-specific random effects, $b_i$, with a conditional mean $E[y_{ij}|b_i] = \mu_{ij}$ and a conditional variance $var(y_{ij}|b_i) = \phi\omega_{ij}^{-1}\nu(\mu_{ij})$. Where $\phi$ is a positive dispersion parameter, $\omega_{ij}$ is a pre-specified weight, and $\nu(\cdot)$ is the variance function. The relationship between $\mu_{ij}$ to $\mathbf{X}$ is defined as

$$g(\mu_{ij}) = x_{ij}'\beta + z_{ij}'b_i,$$

where $g(\cdot)$ is a strictly increasing link function, $\beta$ is the fixed effects coefficient vector for $x$ and $b_i$ is the subject-specific random effects for $z$. $y_{ij}$ are assumed to be independent and of the form $y_{ij}|b_i \sim F_y$ and $b_i$ is assumed to be of the form $b_i \sim F_b$. The distributions $F_y$ and $F_b$ are predominately assumed to be normal, i.e.:

$$\begin{aligned} F_y &\sim N(\mu_{ij}, \phi\omega_{ij}^{-1}\nu(\mu_{ij})) \\ F_b &\sim N(0, D(\psi)), \end{aligned}$$

where $\psi$ is a $c \times 1$ vector of variance components in the covariance matrix of the random effects $D$. Under the identity link function with normal distribution we define the LMM

$$\begin{aligned} & y_i = x'_{ij}\beta + z'_{ij}b_i + \epsilon_i, \\ & \epsilon_i \sim N(0, \sigma^2 I_{n_i}). \end{aligned} \tag{1}$$

McCulloch et al. (2011) state that distribution specification may be affected by basic characteristics of the random effects distribution, such as dependence on a covariate or the cluster sample size. For example, the mean or variance of $F_b$ depends on a covariate. When the mean of the random effects distribution depends on a covariate, a fundamental relationship is introduced between the covariate and the distribution, potentially creating a serious bias in estimating the form of the relationship between the covariate and the outcome. Heagerty and Kurland (2001) show that the impact of highly unequal variances can lead to substantial bias. Such bias from distribution specification can cause unintended inference when testing between and within cluster covariates.

Schelldorfer et al. (2011) defined the GLMMLASSO, which solves the log likelihood of the LMM problem via integral approximation (Laplace approximation) and submit the approximated function to numerical optimization. The advantage of integral approximation methods is to provide an actual objective function for optimization, which enables one to perform likelihood ratio tests among nested models and to compute likelihood-based fit statistics. The disadvantage of these methods is the difficulty of accommodating crossed random effects and multiple subject effects, and the inability to accommodate residual effect covariance structures, or even only residual effect over-dispersion. Moreover, the number of random effects should be small for integral approximation methods to be practically feasible. This disadvantage could potentially inhibit the estimation of random effects in a high dimensional data setting. The penalty term which is used on the approximated likelihood function is the $L_1$ penalty. The algorithm proposed penalizes only the fixed effects in the model, thereby estimating the parameters $\{\beta, \theta, \phi\}$ and predicting the random effects vector $b$ using those estimates. The size of the tuning parameter is calculated in two steps: first via the AIC criterion to generate a relevant set of variables and secondly via the BIC criterion to select the final set of active fixed effects which is an unbiased estimator of degrees of freedom in linear models.

Fan and Li (2012) introduce a class of variable selection methods for the fixed effects using a penalized profile likelihood, provided that the random effects vector has a nonsingular covariance matrix. This penalized profile likelihood is equivalent to the penalized quadratic loss function of fixed effects readily found in penalized least squared methods, such as LARS Efron et al. (2004). Random effects are selected under the constraint that the dimension of the fixed effect is smaller than the sample size. They describe an iterative solution for high dimensionality of both the fixed and random effect by which of selecting the fixed effects using the penalized least squares by ignoring all random effects to reduce the number of fixed effects to below sample size. Then in the second step, with the selected fixed effects, they select random effects and finally using the selected random effects refine the fixed effects selections.

Bondell et al. (2010) apply linearization (Taylor expansion) to solve the LMM which is more aptly suited in models with correlated errors, a large number of random effects, crossed random effects, and multiple types of subjects. The disadvantages of this approach include the absence of a true objective function for the overall optimization process and potentially biased estimates. The likelihood function is reparameterized via a modified Cholesky decomposition of the random effects covariance structure Chen and Dunson (2003). This augmentation allows for penalties on both the fixed and random effects. The penalty used in the optimization is the Adaptive Lasso, Zou (2006), which allows for 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 level of the tuning parameter is calculated using the BIC criterion.

# 2 Reparameterization of the Generalized Linear Mixed Model

This paper will utilize the reparameterization of the LMM model initially defined in Chen and Dunson (2003), and used in Bondell et al. (2010). The reparameterization offers a simple design which regularization penalties can be easily applied to the fixed and random effects simultaneously. The covariance matrix of the random effects $D$ is factorized as follows:

$$D = \Lambda\Gamma\Gamma'\Lambda, \tag{2}$$

where $\Lambda = \text{diag}(d_1, \ldots, d_q)$ is a $q \times q$ non-negative diagonal matrix with elements proportional to standard deviations of the random effects, and $\Gamma$ is a lower triangular matrix that relates to the correlations among the random effects with the $(l, m)$ elements denoted $\gamma_{lm}$. The elements of $\Lambda$ are defined as possibly equal zero, thus enabling a subset of random effects to be selected. $\Lambda$ and $\Gamma$ are identifiable due to the assumption that:

$$d_l \geq 0,\ \gamma_{ll} = 1, \text{ and } \gamma_{lm} = 0, \text{ for } l = 1, \ldots, q;\ m = l + 1, \ldots, q.$$

Applying the modified decomposition (2) to the LMM model (1) the reparameterized LMM is defined, where the covariance matrix of $b_i$ is a function of

$\Lambda, \Gamma$:

$$y_i = x_i'\beta + z_i'\Lambda\Gamma b_i + \epsilon_i.$$

## 3 Simultaneous Variable Selection and Estimation via Regularization Penalties

The Adaptive LASSO has been used as the penalty function on the modified LMM by Bondell et al. (2010) due to its oracle qualities. Although, there are drawbacks to its use, the primary disadvantage is that candidate covariates correlated to variables chosen in the active set are dropped from the final solution. This characteristic has been found to be a drawback in large scale data with grouped covariates. Moreover, when solving the likelihood of the LMM we can see that the fixed and random effect are dependent.

$$L(\phi|y,b) = -\frac{N+mq}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(||y - Z(I_m \otimes \Lambda)(I_m \otimes \Gamma)b - X\beta||^2 + b'b), \quad (3)$$

with $\otimes$ denoting the Kronker product, $Z$ is a block diagonal matrix of $Z_i$, $I_m$ is an identity matrix of dimension $m$.

To overcome these issues we apply a variation on the Elastic Net penalty to the reparameterized likelihood function, (3). The standard Elastic Net penalty denoted as $P$, Friedman et al. (2010), is designed to be applied on a fixed effects model where only $\beta$ is penalized, as seen in (4) below. In this formulation the problem of collinearity is addressed ($L_2$ penalty) in conjunction with shrinkage of redundant variables ($L_1$ penalty). The degree of grouping correlated variables is modulated by the parameter $\alpha$.

$$\begin{aligned}\hat{\beta} &= \min_{\beta \in R^{p+1}} \left[ \frac{1}{2N} \sum_{i=1}^{N} (y_i - x_i'\beta)^2 + P(\beta) \right] \\ P(\beta) &= \lambda_2 \sum_{j \in P} \beta_j^2 + \lambda_1 \sum_{j \in P} |\beta_j|].\end{aligned} \quad (4)$$

We augment (4), while keeping the overall structure and characteristics of the Elastic Net, i.e., the quadratic structure in the $L_2$ penalty. The reparameterization of the LMM allows the penalty function, $\tilde{P}_\alpha(\beta, d)$, to be dependant on both the fixed and random effects in the model.

In addition, correlated random effects can be included in the final model selection, whereas in the Adaptive LASSO settings this was not possible, thus overcoming the problematic testing of simultaneous random effects, Chen and Dunson (2003). The LMMEN is defined as the following:

$$\begin{aligned}Q(\phi|y,b) &= ||y - Z\Lambda\Gamma b - X\beta||^2 + \tilde{P}(\beta, d) \\ \tilde{P}(\beta, d) &= \lambda_2^f \sum_{i \in P} \beta_i^2 + \lambda_2^r \sum_{j \in Q} d_j^2 + \lambda_1^f \sum_{i \in P} |\beta_i| + \lambda_1^r \sum_{j \in Q} |d_j|.\end{aligned} \quad (5)$$

Where $\tilde{P}$ and $Q(\phi)$ denote the penalty applied to the likelihood and the penalized log-likelihood. When the final model is not a mixed effects model, but either a fixed effects or random effects model then the original form of $P_\alpha$ is applied.

# 4 Asymptotics

Assume that the data $\{(X_i, Z_i, y_i);\ i = 1...m\}$ is a random sample of $m$ subjects from a linear mixed-effects model with a probability density function $f(y_i|X_i, Z_i, \phi)$. Let $y_i$ be an $n_i \times 1$ response measurements for subject $i$, $X_i$ be an $n_i \times p$ design matrix of explanatory variables, and $Z_i$ be an $n_i \times q$ design matrix of random effects.

Let $\phi = (\beta', d', \gamma')'$ be a vector of size $k \times 1$, where $\beta \in \mathbb{R}^p$, $d \in \mathbb{R}^q$ and $\gamma$ is of the dimension $\frac{q(q-1)}{2}$. $p = m^\alpha$ is the number of fixed effects, and $q = m^\delta$ the number of the random effects to be estimated. Then number of free elements in the covariance matrix of the random effects, $\Phi$, is $\frac{q(q-1)}{2}$.

In Bondell et al. (2010) the hyperparameters satisfy $\alpha < 1$ and $\delta < 1$ giving a setup of $m > p$, $m > q$. The total number of unknown hyper parameters is $k = p + \frac{q(q+1)}{2} \ll m$. In this paper we are letting $\alpha > 1$ , $\delta < 1$ giving a framework of $m < p$ , $m > q$, i.e. a high-dimensional problem. The total number of unknown parameters that are estimated in this framework is $k = m^\alpha + \frac{m^\delta(m^\delta+1)}{2} \gg m$.

Let $L_i(\phi) = \log(f(y_i|X_i, Z_i, \phi))$ denote the contribution of observation $i$ to the log-likelihood function, given by:

$$L_i(\phi) = -\frac{1}{2}\log|\mathbf{V}_i| - \frac{1}{2}(\mathbf{y}_i - \mathbf{X}_i\beta)'\mathbf{V}_i^{-1}(\mathbf{y}_i - \mathbf{X}_i\beta), \tag{6}$$

where $\mathbf{V}_i = \sigma^2(Z_i\Lambda\Gamma\Gamma'\Lambda Z_i + I_{n_i})$. Denoting the true value of $\phi$ as

$$\phi_0 = (\varphi_{10}, \ldots, \varphi_{k0})' = (\phi_{10}', \phi_{20}')',$$

where $\phi_{10} = (\beta_{10}', d_{10}', \gamma_{10}')'$ is an $s \times 1$ vector whose components are non-zero and $\phi_{20}$ are the $(k-s)$ remaining components of $\phi_0$ such that $\phi_{20} = 0$. Accordingly, let $\phi = (\phi_1', \phi_2')'$. To present the theorems the following regularity conditions are imposed:

**C1** The Fisher information matrix $I(\phi_{10})$ knowing $\phi_{20} = 0$ is finite and positive definite.

**C2** There exists an open subset $\Theta$ of $\mathbb{R}^k$, containing the true parameter $\phi_0$ such that $L_i(\phi)$ given in (**??**) admits all third order derivatives, which are continuous and bounded. There exists a finite mean function $M_{jlm}(y_i, X_i, Z_i)$ such that

$$\left|\frac{\partial^3}{\partial\beta\partial\varphi_l\partial\varphi_m}L_i(\phi)\right| < M(y_i, X_i, Z_i).$$

We have:

**Theorem 1.** *Let* $\phi_0 = (\phi_{10}', 0')'$, *and the observations follow the LMM model satisfying conditions C1 and C2. If* $w_m m^{-1/2} \to \infty$, $(\lambda_1^f + \lambda_1^r)\sqrt{s}/mw_m \to 0$, $(\lambda_2^f + \lambda_2^r)/m \to 0$, *and* $(\lambda_2^f + \lambda_2^r)s/mw_m \to 0$, *then there exists a local maximizer* $\hat{\phi} = \begin{pmatrix} \hat{\phi}_1 \\ 0 \end{pmatrix}$ *of* $Q\left\{\begin{pmatrix} \hat{\phi}_1 \\ 0 \end{pmatrix}\right\}$ *such that* $\hat{\phi}_1$ *is* $w_m$ *consistent for* $\phi_{10}$.

**Theorem 2.** *Let the observations follow the LMM model satisfying conditions C1 and C2. If* $\lambda_m \to \infty$ *then with probability tending to 1 for any given* $\phi_1$ *satisfying* $||\phi_1 - \phi_{10}||_1 \leq Mm^{-1/2}$ *and some constant* $M > 0$,

$$Q\left\{\left(\begin{array}{c} \phi_1 \\ 0 \end{array}\right)\right\} = \max_{||\phi_2||_1 \leq Mm^{-1/2}} Q\left\{\left(\begin{array}{c} \phi_1 \\ \phi_2 \end{array}\right)\right\}.$$

# 5 Simulations

Simulation testing the model selection performance were carried out on five scenarios. In each scenario 100 data sets were simulated from a multivariate normal density.

$$y_i \sim N(X_i\beta, \sigma^2(Z_i \Psi Z_i' + I_{n_i}))$$

The true values of $(\beta_1, \beta_2) = (1, 1)$, and the true variance covariance matrix

$$\Psi = \begin{pmatrix} 9 & 4.8 & 0.6 \\ 4.8 & 4 & 1 \\ 0.6 & 1 & 1 \end{pmatrix}$$

The parameterization of the five scenarios are defined in Table 1.

| Scenario | Subjects | Obs per subject | Fixed Effects | Random Effects | Correlation |
|---|---|---|---|---|---|
| | m | $n_i$ | p | q | $\rho$ |
| 1 | 30 | 5 | 9 | 4 | no |
| 2 | 60 | 10 | 9 | 4 | no |
| 3 | 60 | 5 | 9 | 10 | no |
| 4 | 60 | 10 | 9 | 4 | Multicollin |
| 5 | 30 | 5 | 200 | 4 | p>n |

Table 1: Simulation Scenarios

The first three scenarios are taken from the Bondell et al. (2010) to test the LMMEN to its counterpart "Penalized Linear Mixed Effects Model" (Pen.LME). The true model under consideration in scenarios 1 and 2 is defined as model (7a) and scenario 3 where $X = Z$ as model (7b).

$$y_{ij} = b_{i1} + \beta_1 x_{ij1} + \beta_2 x_{ij2} + b_{i2} Z_{ij1} + b_{i3} Z_{ij2} + \epsilon_{ij} \qquad \epsilon_{ij} \sim N(0,1) \quad (7a)$$

$$y_{ij} = b_{i1} + (\beta_1 + b_{i2}) x_{ij1} + b_{i3} X_{ij3} + \epsilon_{ij} \qquad \epsilon_{ij} \sim N(0,1) \quad (7b)$$

Scenario 4 tests the model performance under settings that there is a high correlation between fixed variables. The scenario uses the same data as scenario

2 and replaces $X_3$ with a linear combination of $X_1$,$X_2$ where $X_3 = wX_1 + (1-w)X_2 + \epsilon$ where $\epsilon \sim N(0,\tau)$. This introduces high correlation in the first three fixed effects, in this setting the LASSO and Adaptive Lasso discard one of these fixed effects thus rendering the model selection inferior. The final scenario tests the performance in high dimension settings. The number of fixed effects is increased to 200 and the first 20 are real parameters while the remainder 180 are nuisance, the random effects remain as in the previous scenarios. This scenario can only be run under LMMEN since the initial values are not calculated using the solution of an unpenalized mixed model, as in the Pen.LME.

Figure 1 depicts the distribution of each parameter estimated within each scenario, where the fixed effects are on the left hand side and the standard deviations of the random effects are on the right hand side. The results of the LMMEN (black) is compared to the Pen.LME (grey). The first three scenarios' results are comparable between the two methods, where the real parameters are chosen consistently. In the fourth scenario the LMMEN selects all the covariates while distributing relatively equal weights to each one. The Pen.LME's adaptive lasso penalty can not discern between the highly correlated variables and sets the 3rd fixed effect near to zero a high percent of the time. In addition, we see that there is no loss of performance in the LMMEN in that it excludes the nuisance fixed effects and correctly estimating the random effects. In the fifth scenario the LMMEN selects the first 20 fixed parameters persistently while setting to zero the nuisance fixed effects, while correctly estimating the random effects.

Figure 2 shows the mean percent of variables correctly selected for the whole model, only the fixed effects and only the random effects for each scenario. This measures performance of model selection without the constraint of an oracle property. For example, in the first scenario of LMMEN 89% of the percent of the variables were correctly selected. We see that the two methods perform similarly, where in the second scenario the Pen.LME out-performed the LMMEN. In high dimension scenario the LMMEN selected 94% of the correct variables on average. Comparing the performance of selecting the all the parameters perfectly, oracle quality, we see that both selections methods results are tempered. In scenario one, the LMMEN selects the perfect model 38% of the time, while comparatively the Pen.LME perfectly selects 50% of the simulations. We see that in the multicollinearity scenario (4), the LMMEN out-performs the Pen.LME selecting all three correct fixed effects 25% of the time, while the Pen.LME 4%. As expected the in the fifth scenario the LMMEN was not able to select all 200 variables correctly in any of the simulations.

# 6 Case Study

The LMMEN algorithm was tested on high dimensional panel data accumulated as part the Good Judgment Project within the Aggregative Contingent Estimation (ACE) Program [1]. The aim of this program is *"to dramatically enhance*

[1] Sponsored by the U.S. Intelligence Advanced Research Projects Activity (IARPA).

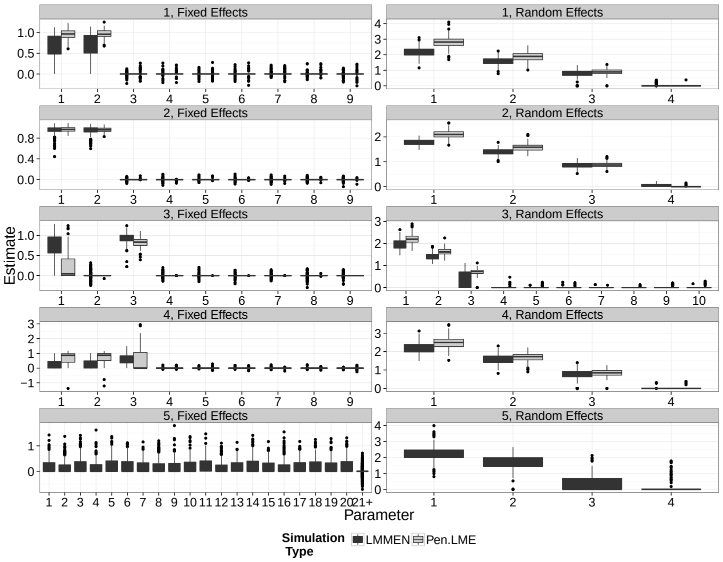

Figure 1: Simulation results of the five scenarios defined in Table 1. Each row is a different scenario, while the left hand side of the panel depict the distribution of the simulated estimated fixed effects parameters, the right hand side depicts the distribution of the estimated random effect parameters. The grey boxplots are the results of the Pen.LME and the black boxplots are the LMMEN.

*the accuracy, precision, and timeliness of forecasts for a broad range of event types, through the development of advanced techniques that elicit, weight, and combine the judgments of many intelligence analysts.".* The study is characterized as a longitudinal study where probabilistic forecasts are derived from crowd sentiment.

The Good Judgment team recruited approximately 3,000 users in the first year. Those users were randomly assigned to 12 groups. There were 75 active questions over the first year. Each user could answer an active question at any time until the question was closed and resolved. This design is a natural one for a repeated measures model with random effects, in which the questions are designated as subjects with random intercepts and for each group a random effect is estimated. In addition there are 40 fixed variables that contain demographic, psychological and past performance information. The data tested was 100 random samples of 50 answers from 20 randomly sampled questions, giving a block structure of 1,000 observations. The LMMEN with specifications for the design structure will be compared to the Elastic Net algorithm, within the GLMNET R library[2] Friedman et al. (2010), which assumes an unstructured covariance design and crossvalidated levels of scaling parameters. The simple mean is used

[2] `http://cran.r-project.org/web/packages/glmnet/index.html`. Version 2.0-2 was used in the simulations.

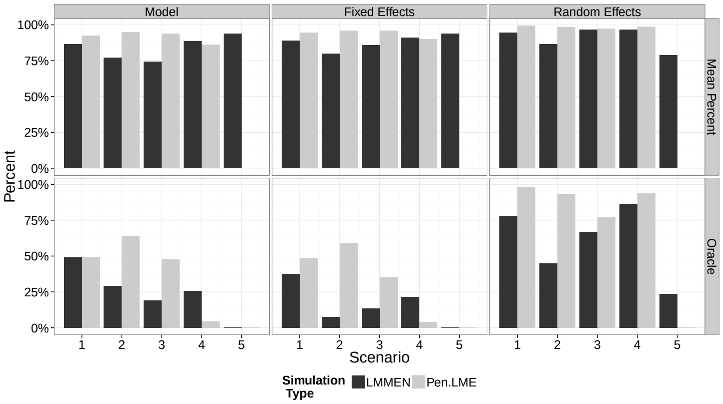

Figure 2: Model selection performance comparison of Pen.LME (grey) and the LMMEN (black) for the simulations defined in Table 1. The panel columns from left to right are the selection performance by entire model, the fixed effects and the random effects. The panel rows depict the performance statistic, the upper row is the mean percent of parameters selected correctly, the bottom row measures the oracle property of each model.

as the baseline aggregation method. Two levels of algorithm performance will be investigated, first is the model selection and second is the accuracy of the aggregated predictions. The statistic which will be used to test performance of the aggregated predictions is the Brier score. In this case study only binary events are taken under consideration and 6 questions are omitted under this constraint. Thus the Brier score equation is defined as

$$BS = \frac{1}{N}\sum_{t=1}^{N}(f_t - o)^2,$$

in which $f_t$ is the prediction at time $t$, $o$ is the question outcome, and $N$ is the number of prediction instances. First we compare the model selection between the two algorithms as seen in panel (a) of Figure 3. It can be seen that the LMMEN produces a higher level of sparsity than the GLMNET and the variables chosen are persistent in the simulation. The addition of the design structure allows us to select random effects found in the data. The groups of users are assumed to be distributed normally with a variation parameter. After applying the LMMEN the optimal solution produces a sparse covariance matrix with relation to the random effects. The results of this selection can be found in panel (b) of Figure 3. We see that the variance estimate of groups {1B,1C,4A,4B,4C} is equal to zero in a large percent of the simulations, thus concluding that there is no difference between the user responses in those groups.

The second level of performance investigated is the prediction accuracy. The estimated non-zero covariates after selection are used to aggregate out of sample

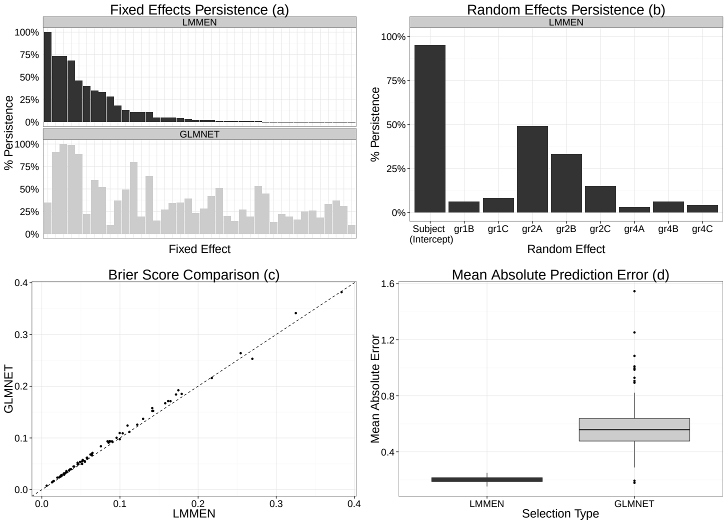

Figure 3: Model performance of GLMNET and LMMEN tested on 100 random samples of 1000 observations from the Good Judgment study. Panel (a) compares the distribution of fixed effects selection persistency between the two methods. Panel (b) depicts the random effects selection persistency of the LMMEN. Panel (c) compares the probabilistic forecast accuracy of the two methods using the Brier Score as the loss function. Panel (d) compares the distribution of the mean absolute prediction error of each method

user predictions of active questions. The results of the two selection methods can be found in panel (c) of Figure 3. We see that both methods have low brier scores, i.e., the aggregated predictions are close to the final outcome. The average brier scores for GLMNET and LMMEN are .089 and .085 respectively. This shows that there is no major loss of prediction accuracy between the two selection methods. Both methods out perform the benchmark aggregation (grand mean) which has an average brier score of 0.11. The more significant difference as seen in panel (d) of Figure 3 is the mean absolute prediction error, where in this case the average estimation error in the GLMNET is larger than the LMMEN. This reinforces the importance of modeling the structure of the data correctly in order to minimize errors in estimation.

# 7 Discussion

In the paper we have shown that fixed and random effects in high dimensional linear mixed models can be simultaneous selected. This selection method introduces the ability to select variables under conditions of multicollinearity both in the fixed and random effects. This method, LMMEN, furthers current variable selection of these models with the introduction of a ridge penalty into the optimization.

It was found through simulations that this method correctly selects fixed and random effects under sparse data designs. Simulations were carried out under the Gaussian assumption for both the conditional distribution and the distribution of the random effects. Further simulations will be carried out which relax the assumption of the conditional distribution. When testing the LMMEN in the case study the variable selection was comparable to a similar regularization method which does not model the covariance structure, GLMNET. The LMMEN gave further insight into the characteristics of groups of users, where a subset of them were found not have prediction difference within the groups.

This paper applies the Brier Score (L2 loss) as the loss function to tune the penalty parameters in the case study. One could calibrate the penalty parameters is the intraclass correlation (ICC) levels. The ICC is intrinsic to random effects models, and is regularly used for evaluating the level of correlation between different groups as defined by the model. Applying the LMMEN while calibrating to minimize the ICC could be a vital tool for correctly selecting candidate random effects to model the data design and will be assessed in future work.

# A Appendix: proofs

For the penalized log-likelihood in (5), let $\phi = (\phi_1', 0')'$ and let

$$L^1(\phi_1) \equiv L\left\{\begin{pmatrix}\phi_1\\ 0\end{pmatrix}\right\} \text{ and } Q^1(\phi_1) \equiv Q\left\{\begin{pmatrix}\phi_1\\ 0\end{pmatrix}\right\}$$

denote the log-likelihood and the penalized log-likelihood of the first $s$ components of $\phi$.

*Proof Theorem 1.* Consider the penalized log-likelihood $Q(\phi)$ given in (5) in the neighborhood of the true value $\phi_{10}$. Let $u \neq 0$, and $\phi_1 = \phi_{10} + w_m u$. Setting $\phi_2 = 0$, we show that for a small enough $\epsilon > 0$, there exists a large constant $C$ such that for a sufficiently large $m$,

$$P\Big(\sup_{\|u\|=C} Q^1(\phi_{10} + w_m u) < Q(\phi_{10})\Big) \geq 1 - \epsilon.$$

Thus, with probability $1-\epsilon$ the maximum is within the ball of radios $Cw_m$.

Note that

$$\begin{aligned} mD_m(u) &\equiv Q^1(\phi_1) - Q^1(\phi_{10}) \\ &= -\left[L^1(\phi_{10} + w_m u) - L^1(\phi_{10})\right] \\ &\quad + \lambda_1^f\big(\|\beta_0 + w_m u_\beta\|_1 - \|\beta_0\|_1\big) + \lambda_1^r\big(\|d_0 + w_m u_d\|_1 - \|d_0\|_1\big) \\ &\quad + \lambda_2^f\big(\|\beta_0 + w_m u_\beta\|_2^2 - \|\beta_0\|_2^2\big) + \lambda_2^r\big(\|d_0 + w_m u_d\|_2^2 - \|d_0\|_2^2\big), \end{aligned}$$

where we divided $u$ to its natural components $u_\beta \in R^p$ and $u_d \in R^q$. Using the Taylor series expansion we have

$$
\begin{aligned}
&D_m(u)\\
&= -w_m(m^{-1}\nabla L(\phi_{10}))'u - \frac{w_m^2}{2m}u'[\nabla^2 L(\phi_{10})]u + R_m\\
&\quad + m^{-1}\lambda_1^f\big(\|\beta_0 + w_m u_\beta\|_1 - \|\beta_0\|_1\big) + m^{-1}\lambda_1^r\big(\|d_0 + w_m u_d\|_1 - \|d_0\|_1\big)\\
&\quad + m^{-1}\lambda_2^f\big(\|\beta_0 + w_m u_\beta\|_2^2 - \|\beta_0\|_2^2\big) + m^{-1}\lambda_2^r\big(\|d_0 + w_m u_d\|_2^2 - \|d_0\|_2^2\big),
\end{aligned}
$$

where $\nabla L(\phi_{10}), \nabla^2 L(\phi_{10})$ denote the vector and matrix of the first and second order partial derivatives of $L(\phi_1)$ at $\phi_{10}$ respectively. $\nabla P(\beta, d), \nabla^2 P(\beta, d)$ denote the first and second derivatives of the penalty term at $(\beta_0, d_0)$. The remainder $R_n$ tends to zero as m $\to \infty$ since, by C2, $|R_m|$ can be bounded by

$$
\left(\frac{w_m^3\|u\|_2^3}{6m}\right)\sum_{i=1}^m M(y_i, X_i, Z_i) = O_P(w_m^3).
$$

The $j$th partial derivative for each corresponding $\beta_1, d_1, \gamma_1$ the $\nabla L(\phi_{10})$ satisfies $E\left\{\frac{\partial}{\partial \beta_j}L(\phi_1)\right\} = E\left\{\frac{\partial}{\partial d_j}L(\phi_1)\right\} = E\left\{\frac{\partial}{\partial \gamma_j}L(\phi_1)\right\} = 0$ and thus the corresponding empirical means are $O_p(m^{-1/2})$.

For $\nabla^2 L(\phi_{10})$ we have

$$
m^{-1}\nabla^2 L(\phi_{10}) \to_p -I(\phi_{10}),
$$

where $I(\phi_{10})$ is the Fisher information evaluated at $\phi_{10}$, which is positive definite by (C1). By choosing a sufficiently large $C$, the second term dominates the first term uniformly in $\|u\| = C$.

For the penalty term if $w_m P(\beta, d) \to 0$ as $m \to \infty$ it follows that $P(\beta, d) \to_p 0$, and thus also dominated by the second term. The absolute value of the penalty component of $D_m(u)$ is bounded by

$$
\begin{aligned}
&m^{-1}w_m\lambda_1^f\|u_\beta\|_1 + m^{-1}w_m\lambda_1^r\|u_d\|_1 + m^{-1}\lambda_2^f\big(2w_m\|\beta_0\|_2\|u_\beta\|_2 + w_m^2\|u_0\|_2^2\big)\\
&\qquad + m^{-1}\lambda_2^r\big(2w_m\|d_0\|_2\|u_d\|_2 + w_m^2\|u_d\|_2^2\big)\\
&\quad \le m^{-1}w_m C\big(\lambda_1^f\sqrt{s} + \lambda_1^r\sqrt{s} + \lambda_2^f(2\|\beta_0\|_2 + w_m C) + \lambda_2^r(2\|d_0\|_2 + w_m C)\big).
\end{aligned}
$$

which is dominated by the second term of $D_m(u)$. Therefore, by choosing a sufficiently large $C$ there exists a local maximum inside $\{\phi_{10} + w_m u : \|u\| < C\}$ with probability $1 - \epsilon$, thus there exists a local maximizer $\hat{\phi} = (\hat{\phi_1}', 0')'$ of $\phi_0 = (\phi_1', 0')'$ such that $\|\hat{\phi_1} - \phi_{10}\| = O_p(w_m)$. □

For the following proof we define $\phi = (\beta', d', \gamma')$ as a $k \times 1$ vector of unknown parameters of size $k = k_\beta + k_d + k_\gamma$. Let $\phi_2 = (\beta_2', d_2', \gamma_2')$ be a vector of size $k_2 = k - s$ corresponding to the true zero parameters, given $k_2 = k_{\beta_2} + k_{d_2} + k_{\gamma_2}$. Reminding that we defined earlier that the likelihood and the penalized log likelihood as

$$
L(\phi) = L\left\{\begin{pmatrix}\phi_1\\ \phi_2\end{pmatrix}\right\} \text{ and } Q(\phi) = Q\left\{\begin{pmatrix}\phi_1\\ \phi_2\end{pmatrix}\right\}.
$$

*Proof Theorem 2.* For $m \to \infty$ and any $\phi_1 : ||\phi_1 - \phi_{10}||_1 \leq Mm^{-1/2}$ and for $\epsilon_m = Mm^{-1/2}$ and for each $j = (s+1), \ldots, (k_{\beta_2} + k_{d_2})$ we have with probability tending to 1 that

$$\begin{aligned} &\frac{\partial}{\partial\varphi_j} Q(\phi) < 0 \text{ for } 0 < \varphi_j < \epsilon_m \\ &\frac{\partial}{\partial\varphi_j} Q(\phi) > 0 \text{ for } -\epsilon_m < \varphi_j < 0 \end{aligned} \tag{8}$$

The partial derivative of $Q(\phi)$ with respect to $\varphi_j$ is given by:

$$\frac{\partial}{\partial\varphi_j} Q(\phi) = \frac{\partial}{\partial\varphi_j} L(\phi) - (\lambda_1 \text{sgn}(\varphi_j) + 2\lambda_2 \varphi_j) \,,$$

noting that the penalty is dependent on whether $\varphi_j$ is $\beta$ or d.

One can verify (8) through the Taylor Series expansion of $\frac{\partial}{\partial\varphi_j} L(\phi) = \frac{\partial}{\partial\varphi_j} L(\phi)$ around $\phi_0$:

$$\begin{aligned} \frac{\partial}{\partial\varphi_j} Q(\phi) =& \frac{\partial}{\partial\varphi_j} L(\phi_0) - \sum_{l=1}^{k} \frac{\partial}{\partial\varphi_l} \left( \frac{\partial}{\partial\varphi_j} L(\phi_0) \right) (\varphi_l - \varphi_{l0}) \\ &+ \frac{1}{2} \sum_{i=1}^{m} \sum_{l=1}^{k} \sum_{g=1}^{k} \frac{\partial^2}{\partial\varphi_l \partial\varphi_g} \left( \frac{\partial}{\partial\varphi_j} L_i(\phi_*) \right) (\varphi_l - \varphi_{l0})(\varphi_g - \varphi_{g0}) \\ &- (\lambda_1 \text{sgn}(\varphi_j) + 2\lambda_2 \varphi_j) \,, \end{aligned} \tag{9}$$

where $\phi_*$ is on the interval connecting $\phi$ and $\phi_0$. Next we define the first order derivatives needed to numerically solve (9):

$$\begin{aligned} L_\beta =& \frac{\partial}{\partial\beta_j} L(\phi_0) = X_j' V^{-1}(y - X\beta) = O_p(m^{-1/2}) \\ L_d =& \frac{\partial}{\partial d_j} L(\phi_0) = \frac{1}{2} \left[ \text{Tr}(V^{-1}S^j) + (y - X\beta)'(V^{-1}S^j V^{-1})(y - X\beta) \right] = O_p(m^{-1/2}), \end{aligned}$$

where $S^j = Z(\frac{\partial}{\partial d_j} D\Gamma\Gamma' D)Z'$ and $\text{Tr}(A)$ is the trace operator on a given matrix A. We now define the second order derivatives which follow $\frac{1}{m}\nabla^2 L(\phi)|_{\phi=\phi_0} \to E_{\phi_1=\phi_{10}}[\nabla^2 L(\phi)]$, where

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

$$\begin{aligned} E[L_{\beta\beta}]_j =& -XV^{-1}X \\ E[L_{\beta d}]_j =& -E\left[ X_j'(V^{-1}S^j V^{-1})(y - X\beta) \right] |_{\phi=\phi_0} = 0 \\ E[L_{\beta\gamma}]_j =& -E\left[ X_j'(V^{-1}T^j V^{-1})(y - X\beta) \right] |_{\phi=\phi_0} = 0 \\ E[L_{dd}]_{jl} =& -\text{Tr}(V^{-1}S^j V^{-1}S^l)|_{\{j \geq (s+1), \phi_j = 0\}} = 0 \\ E[L_{\gamma\gamma}]_{jl} =& -\text{Tr}(V^{-1}T^j V^{-1}T^l)|_{\{j \geq (s+1), \phi_j = 0\}} = 0 \\ E[L_{d\gamma}]_{jl} =& -\text{Tr}(V^{-1}S^j V^{-1}T^l)|_{\{j \geq (s+1), \phi_j = 0\}} = 0, \end{aligned}$$

where $T^j = ZD(\frac{\partial}{\partial \gamma_j}\Gamma\Gamma')DZ'$.
Using these partial derivatives we solve (9) first for $\phi_j = \beta_j$ and then for $\phi_j = d_j$.

$$\begin{aligned}
&\frac{1}{\sqrt{m}}\left(\frac{\partial}{\partial \beta_j}Q(\phi)\right)\\
=&\frac{1}{\sqrt{m}}\Bigg[L_\beta - m\left(\sum_{l=1}^{k_\beta} L_{\beta\beta}(\beta_l-\beta_{l0}) + \sum_{l=k_{\beta+1}}^{k_d} L_{\beta d}(d_l-d_{l0}) + \sum_{l=k_{d+1}}^{k_\gamma} L_{\beta\gamma}(\gamma_l-\gamma_{l0})\right)\\
&+\sum_{i=1}^{m}\sum_{l=1}^{k_\beta}\sum_{g=k_\beta+1}^{k_d}\frac{\partial}{\partial \beta_g}L_{\beta d}(\beta_l-\beta_{l0})(d_g-d_{g0})\\
&+\sum_{i=1}^{m}\sum_{l=1}^{k_\beta}\sum_{g=k_d+1}^{k_\gamma}\frac{\partial}{\partial \beta_g}L_{\beta\gamma}(\beta_l-\beta_{l0})(\gamma_g-\gamma_{g0})\\
&+\sum_{i=1}^{m}\sum_{l=k_\beta+1}^{k_d}\sum_{g=k_{d+1}}^{k_\gamma}\frac{\partial}{\partial \gamma_g}L_{\beta d}(d_l-d_{l0})(\gamma_g-\gamma_{g0})\\
&+\frac{1}{2}\Bigg(\sum_{i=1}^{m}\sum_{l=k_\beta+1}^{k_d}\sum_{g=k_\beta+1}^{k_d}\frac{\partial}{\partial d_g}L_{\beta d}(d_l-d_{l0})(d_g-d_{g0})\\
&+\sum_{i=1}^{m}\sum_{l=k_d+1}^{k_\gamma}\sum_{g=k_{d+1}}^{k_\gamma}\frac{\partial}{\partial \gamma_g}L_{\beta\gamma}(\gamma_l-\gamma_{l0})(\gamma_g-\gamma_{g0})\Bigg) - \left(\lambda_1^f \operatorname{sgn}(\beta_j) + 2\lambda_2^f(\beta_j)\right)\Bigg],
\end{aligned}$$

given $||\phi-\phi_0||_1 \leq Mm^{-1/2}$ then we have

$$\frac{1}{\sqrt{m}}\left(\frac{\partial}{\partial \beta_j}Q(\phi)\right) = -\left(\lambda_1^f \operatorname{sgn}(\beta_j) + 2\lambda_2^f(\beta_j)\right) + O_p(1). \tag{10}$$

For $\beta_{j0} = 0$ and $\{\lambda_1^f, \lambda_2^f\} \to \infty$ the sign of the derivative is completely determined by $\beta_j$, more specifically:

$$\begin{array}{lll}
\text{if} & M > \beta_j > 0 & \text{then} \quad \frac{\partial}{\partial \beta_j}Q(\phi) < 0\\
\text{if} & -M < \beta_j < 0 & \text{then} \quad \frac{\partial}{\partial \beta_j}Q(\phi) > 0
\end{array}.$$

Similarly,

$$\begin{aligned}
&\frac{1}{\sqrt{m}}\left(\frac{\partial}{\partial d_j}Q(\phi)\right)\\
=&\frac{1}{\sqrt{m}}\Bigg[L_d - m\left(\sum_{l=1}^{k_\beta} L_{\beta\beta}(\beta_l-\beta_{l0}) + \sum_{l=k_{\beta+1}}^{k_d} L_{\beta d}(d_l-d_{l0}) + \sum_{l=k_{d+1}}^{k_\gamma} L_{\beta\gamma}(\gamma_l-\gamma_{l0})\right)\\
&+\sum_{i=1}^{m}\sum_{l=1}^{k_\beta}\sum_{g=1}^{k_\beta}\frac{\partial}{\partial \beta_g}L_{d\beta}(\beta_l-\beta_{l0})(\beta_g-\beta_{g0})
\end{aligned}$$

$$
\begin{aligned}
&+\sum_{i=1}^{m}\sum_{l=k_\beta+1}^{k_d}\sum_{g=k_d+1}^{k_\gamma}\frac{\partial}{\partial\gamma_g}L_{dd}(d_l-d_{l0})(\gamma_g-\gamma_{g0})\\
&+\sum_{i=1}^{m}\sum_{l=k_\beta+1}^{k_d}\sum_{g=k_{d+1}}^{k_\gamma}\frac{\partial}{\partial d_g}L_{d\beta}(\beta_l-\beta_{l0})(d_g-d_{g0})\\
&+\frac{1}{2}\Bigg(\sum_{i=1}^{m}\sum_{l=k_\beta+1}^{k_d}\sum_{g=k_\beta+1}^{k_d}\frac{\partial}{\partial d_g}L_{dd}(d_l-d_{l0})(d_g-d_{g0})\\
&+\sum_{i=1}^{m}\sum_{l=k_d+1}^{k_\gamma}\sum_{g=k_{d+1}}^{k_\gamma}\frac{\partial}{\partial\gamma_g}L_{d\gamma}(\gamma_l-\gamma_{l0})(\gamma_g-\gamma_{g0})\Bigg)-(\lambda_1^r\text{sgn}(d_j)+2\lambda_2^r(d_j))\Bigg],
\end{aligned}
$$

given $||\phi-\phi_0||_1 \leq Mm^{-1/2}$ then we have

$$\frac{1}{\sqrt{m}}\left(\frac{\partial}{\partial d_j}Q(\phi)\right)=-(\lambda_1^r\text{sgn}(d_j)+2\lambda_2^r(d_j))+O_p(1).$$

For $d_{j0}=0$ and $(\lambda_1^r,\lambda_2^r)\rightarrow\infty$ the sign of the derivative is completely determined by $d_j$, more specifically:

$$\begin{array}{llll} \text{if} & M > d_j > 0 & \text{then} & \frac{\partial}{\partial d_j}Q(\phi) < 0 \\ \text{if} & -M < d_j < 0 & \text{then} & \frac{\partial}{\partial d_j}Q(\phi) > 0 \end{array}.$$

□

# References

D. M. Bates. lme4: Mixed-effects modeling with r. *URL http://lme4. r-forge. r-project. org/book*, 2010.

H. D. Bondell, A. Krishna, and S. K. Ghosh. Joint variable selection for fixed and random effects in linear mixed-effects models. *Biometrics*, 66(4):1069–1077, 2010.

N. E. Breslow and D. G. Clayton. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88(421):9–25, 1993.

Z. Chen and D. B. Dunson. Random effects selection in linear mixed models. *Biometrics*, 59(4):762–769, 2003.

B. Efron, T. Hastie, I. Johnstone, R. Tibshirani, et al. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004.

Y. Fan and R. Li. Variable selection in linear mixed effects models. *Annals of statistics*, 40(4):2043, 2012.

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

P. J. Heagerty and B. F. Kurland. Misspecified maximum likelihood estimates and generalised linear mixed models. *Biometrika*, 88(4):973–985, 2001.

C. E. McCulloch, J. M. Neuhaus, et al. Misspecifying the shape of a random effects distribution: why getting it wrong may not matter. *Statistical Science*, 26(3):388–402, 2011.

J. Schelldorfer, P. Bühlmann, and S. Van de Geer. Estimation for high-dimensional linear mixed-effects models using l1-penalization. *Scandinavian Journal of Statistics*, 38(2):197–214, 2011.

S. R. Searle, G. Casella, and C. E. McCulloch. *Variance Components*, volume 242. John Wiley and Sons, New York, 1992.

S. A. Van de Geer. High-dimensional generalized linear models and the lasso. *The Annals of Statistics*, pages 614–645, 2008.

H. Zou. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429, 2006.