Biometrics 000, 000–000 DOI: 000
000 0000

# Joint Selection in Generalized Linear Mixed Model Analysis: A Confidence Distribution Approach

**Shou-En Lu[1,2*], Sinae Kim[1,3], Jerry Q. Cheng[4], Changfa Lin[5], Sharad Goyal[6], Salma Jabbour[7]**

[1]Department of Biostatistics and Epidemiology, Rutgers University, Piscataway, New Jersey

[2]Biostatistics Shared Resource, Rutgers Cancer Institution of New Jersey, New Brunswick, New Jersey

[3]Global Biometrics and Data Sciences, Bristol Myers Squibb, Lawrence, New Jersey

[4]Department of Computer Science, New York Institute of Technology, New York, New York

[5]Deloitte, Parsippany, New Jersey

[6]Radiation Oncology, George Washington University Hospital, Washington, DC

[7]Radiation Oncology, Rutgers Cancer Institution of New Jersey, New Brunswick, New Jersey

Summary: Generalized linear mixed models (GLMMs) are commonly used to describe relationships between correlated responses and covariates. In this paper, we propose a regularized method to select both fixed and random effects in GLMMs. In contrast to using the observed data likelihood functions, we propose to construct the objective functions using the confidence distribution of model parameters based on the joint and separate marginal asymptotic distributions of the fixed effect and random effect parameter estimators to perform effect selections. With a proper choice of regularization parameters in the adaptive LASSO framework, we show the consistency and oracle properties of the proposed regularized estimators. Simulation studies have been conducted to assess the performance of the proposed estimators and demonstrate computational efficiency. Our method has also been applied to two longitudinal cancer studies to identify demographic and clinical factors associated with health outcomes after cancer therapies.

This paper has been submitted for consideration for publication in *Biometrics*

## 1. Introduction

Generalized linear mixed models (GLMMs) are a commonly used class of models to describe the relationship between correlated responses and covariates. Researchers often want to determine fixed effects and (or) random effects of the outcome variables from a pool of covariates using the variable selection approach. Our study is motivated by two longitudinal cancer studies. The first study wanted to identify demographic and clinical covariates that may be associated with tumor size in lung cancer patients. The second study aimed to relate a set of covariates to the incidence of common mammographic sequelae after breast conserving surgery and radiation therapy (tian2016comparison). For both studies, the investigators wanted to identify important covariates that may predict the outcomes as fixed effects. They also wanted to include random effects to assess if some covariates exhibit heterogeneous effect.

In the statistical literature, several approaches have been proposed for variable selection. For instance, the selection using the information criterion (e.g., Akaike information criterion (AIC) or Bayesian information criterion (BIC), etc) has been commonly used to determine the final model after fitting a number of candidate models (Keselman et al. (1998); Gurka (2006); Claeskens and Consentino (2008); Ibrahim, Zhu, and Tang (2008); Liang, Wu and Zou (2008), among others). For the popular regularized estimation method, some proposed methods to select both fixed and random effects (e.g., Bondell, Krishna, and Ghosh (2010), Peng and Lu (2012), and Lin, Pang and Jiang (2013) for linear mixed models (LMMs), and Ibrahim et al. (2011), and Hui, Mueller and Welsh (2017) for generalized linear mixed models (GLMMs), among others), some focused on the computational algorithms for fixed effects selection only (e.g., Schelldorfer, Meier, and Bühlmann (2014)), and some focused on random effects selection only (Pan and Huang (2014)), etc. In general, the computa-

tion of these methods are extensive. For some methods, the statistical properties of the estimators remain to be determined. Recently, Hui, Mueller and Welsh (2017) proposed a penalized quasi-likelihood (PQL) estimation with sparsity inducing penalties on both fixed and random coefficients and demonstrated improved computational efficiency. In spite of these recent advancements, variable selection in GLMMs remains challenging, primarily due to the computational difficulty and complexity associated with the integral related to the random effects in the likelihood functions. It may also affect the performance of the statistical inference. In this paper, we propose a regularized estimation method using the confidence distribution approach.

The seed idea of a confidence distribution could be traced back to Bayes (1763) and Fisher (1922). However, the concept and its applications have been developed extensively in recent years (e.g., Xie, Singh, and Strawderman (2011), Tian et al. (2011), Liu, Liu and Xie (2015), Wang et al. (2021); also see Xie and Singh (2013) and the references therein for more detailed review). The confidence distribution can be viewed as a sample dependent distribution function, and used to estimate and provide statistical inference for a parameter of interest (Cox (2013) and Wang et al. (2021)). Typically, the regularized estimation is performed based on the likelihood function constructed from the observed data. In this paper, we propose to perform regularized estimation by optimizing the objective functions constructed from the confidence distributions of model parameters. Specifically, we show that the conventional objective functions using the observed data likelihood function can be approximated by the objective function constructed based on the confidence distributions of the model parameters. With proper choices for the regularization parameters, it can be shown that the proposed estimators possess the estimation consistency, selection consistency and the oracle property. Moreover, we demonstrate that the proposed method is computationally efficient and can be easily implemented using existing standard software packages.

The rest of this paper is organized as follows. In Section 2, we provide a brief review of the statistical inference in the generalized linear mixed models. In Section 3, we provide the rationale of the proposed regularized estimation approach using confidence distribution and establish the statistical properties of the proposed regularized estimators. In Section 4, we discuss the implementation of the optimization method and the determination of tuning parameters. In Sections 5 and 6, we present simulation results and conduct the real data analysis to illustrate our methods. We conclude this paper with a discussion in Section 7.

## 2. Generalized Linear Models and Proposed Regularized Estimation

Consider a sample of $n$ independent clusters. Let $\boldsymbol{y}_i = (y_{i1}, y_{i2}, \cdots, y_{im_i})^T$ and $y_{ij}$ denote the $j$th measurement of the $i$th cluster, where $i = 1, 2, \cdots, n$, and $j = 1, 2, \cdots, m_i$. Let $\boldsymbol{x}_{ij}$ be a vector of $p_f$ covariates correpsonding to fixed effects, and $\boldsymbol{z}_{ij}$ be a vector of $p_r$ covariates corresponding to random effects. Both $\boldsymbol{x}_{ij}$ and $\boldsymbol{z}_{ij}$ include 1 for intercepts. Typically, $\boldsymbol{z}_{ij}$ is a subset of $\boldsymbol{x}_{ij}$. Conditional on the random effects $b_i$, we assume that the responses $y_{ij}'s$ follow a distribution of the exponential family with conditional mean $\mu_{ij}$ depending on $b_i$ through the link function $g(\cdot)$ given by

$$g(\boldsymbol{\mu}_i) = \boldsymbol{\eta}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{\Gamma}\boldsymbol{b}_i, \tag{1}$$

where $\boldsymbol{\beta}$ is the fixed effect regression coefficients, $\boldsymbol{b}_i$ is the random effects assumed to follow a multivariate normal distribution $\boldsymbol{N}_{p_r}(\boldsymbol{0}, \boldsymbol{I}_{p_r})$ with $\boldsymbol{I}_{p_r}$ being a $p_r \times\ p_r$ identity matrix, and $\boldsymbol{\Gamma}$ is a $p_r \times p_r$ Cholesky decomposition lower triangular matrix depending on parameter $\boldsymbol{\gamma}$ such that $\boldsymbol{\Gamma}\boldsymbol{b}_i$ follows $\boldsymbol{N}_{p_r}(\boldsymbol{0}, \boldsymbol{D})$ and $\boldsymbol{D} = \boldsymbol{\Gamma}\boldsymbol{\Gamma}^T$. For simplicity, we assume the canonical link such that $g(\boldsymbol{\mu}) = \boldsymbol{\eta}_i$. Consider finite dimensions of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, i.e., $p_f < \infty$ and $p_r < \infty$. The model parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\gamma}^T, \phi)^T$ can be estimated by maximizing the marginal likelihood of $y$ through integrating out $\boldsymbol{b}_i$,

$$\mathcal{L}(\boldsymbol{\theta}; \boldsymbol{y}) = \prod_{i=1}^{n} \int f_{\boldsymbol{y}|\boldsymbol{b}}(\boldsymbol{y}_i|\boldsymbol{b}_i; \boldsymbol{\theta}) f(\boldsymbol{b}_i|\boldsymbol{\theta}) d\boldsymbol{b}_i, \tag{2}$$

where $\phi$ is the dispersion parameter, $f_{\boldsymbol{y}|\boldsymbol{b}}(\boldsymbol{y}_i|\boldsymbol{b}_i;\boldsymbol{\theta})$ denotes the conditional density function of $\boldsymbol{Y}_i|\boldsymbol{b}_i$, and $f(\boldsymbol{b}_i|\boldsymbol{\theta})$ denotes the marginal density of $\boldsymbol{b}_i$. Note that the parameters of interest are $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$. Define the maximum likelihood estimator (MLE) of $\boldsymbol{\theta}$ by

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \log \mathcal{L}(\boldsymbol{\theta};\boldsymbol{y}),$$

Let $\boldsymbol{\theta}_0$ denote the true value of $\boldsymbol{\theta}$. Under mild regularity conditions, $\hat{\boldsymbol{\theta}}$ is consistent and $\sqrt{n}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0) \to^D \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{\Sigma}(\boldsymbol{\theta}))$, where $\boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}) = \lim_{n\to\infty} \boldsymbol{\mathcal{I}}(\boldsymbol{\theta})$, $\boldsymbol{\mathcal{I}}(\boldsymbol{\theta}) = -n^{-1}\partial^2 \log \mathcal{L}(\boldsymbol{\theta};\boldsymbol{y})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T$, and $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ is consistently estimated by $\hat{\boldsymbol{\Sigma}} = \boldsymbol{\mathcal{I}}^{-1}(\hat{\boldsymbol{\theta}})$ (e.g., Pan and Lin (2005)).

## 3. Proposed Regularized Estimations

### 3.1 *Construction of Objective Function*

Variable selection using regularized approach has achieved much success in recent decades. Typically, the objective function is constructed from the observed data likelihood function plus penalty functions. Let

$$Q^o(\boldsymbol{\theta}) = -\log \mathcal{L}(\boldsymbol{\theta};\boldsymbol{y}) + n\kappa_\rho^o(\boldsymbol{\beta}) + n\kappa_\tau^o(\boldsymbol{\gamma}),$$

and define the regularized estimator $\hat{\boldsymbol{\theta}}^o_{\rho\tau}$=$\arg\min_{\boldsymbol{\theta}} Q^o(\boldsymbol{\theta})$, where $\kappa_\rho^o(\boldsymbol{\beta})$ and $\kappa_\tau^o(\boldsymbol{\gamma})$ are penalty terms that control the sparsity for the estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ to select appropriate fixed effects and random effects, respectively. Because the integral in $\mathcal{L}(\boldsymbol{\theta};\boldsymbol{y})$ generally does not have a closed-form solution, various approaches have been proposed to tackle this computational challenge to obtain the MLE, score functions and the information matrix for making statistical inference (e.g., Wolfinger (1993), Pinheiro and Bates (1995), Westfall (1997), Lange (1999), Pinheiro and Chao (2006), among others). With the addition of penalty terms, obtaining $\hat{\boldsymbol{\theta}}^o_{\rho\tau}$ can be even more computationally challenging. Many studies have been proposed to derive the regularized estimators for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ (e.g., Bondell, Krishna, and Ghosh (2010), Peng and Lu (2012), Lin, Pang and Jiang (2013), Ibrahim et al. (2011), Schelldorfer, Meier, and Bühlmann (2014), and Pan and Huang (2014), etc), but they are

generally computationally complicated and extensive. To ease the complexity in deriving the regularized estimators for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, we propose to construct the objective function based on the confidence distribution of the MLE.

Inference based on confidence distribution has been discussed in the statistical literature (e.g., Efron (1993), Efron (1998), Singh, Xie and Strawderman (2007), Xie, Singh, and Strawderman (2011), Xie and Singh (2013)). A confidence density is the density function representation of a confidence distribution. Based on the asymptotic distribution of $\hat{\boldsymbol{\theta}}$ and Singh, Xie and Strawderman (2007), we write the confidence density of the parameter $\boldsymbol{\theta}$ as

$$h(\boldsymbol{\theta}) = \frac{1}{(2\pi)^{p/2}\left\{\det\left(n^{-1}\hat{\boldsymbol{\Sigma}}\right)\right\}^{1/2}} \exp\left\{-\frac{1}{2}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})^T\left(n^{-1}\hat{\boldsymbol{\Sigma}}\right)^{-1}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})\right\}, \tag{3}$$

where $p$ denotes the length of $\boldsymbol{\theta}$ and $\det(C)$ is the determinant of a matrix $C$. Take the logarithm of $h(\boldsymbol{\theta})$ such that $log[h(\boldsymbol{\theta})] = -(1/2)(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})^T\left(n^{-1}\hat{\boldsymbol{\Sigma}}\right)^{-1}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}) + c$, where $c$ is some constant free of $\boldsymbol{\theta}$. Consider the following approximation. It can be seen that

$$\begin{aligned} n^{-1}\log\mathcal{L}(\boldsymbol{\theta};\boldsymbol{y}) &\approx n^{-1}\log\mathcal{L}(\hat{\boldsymbol{\theta}};\boldsymbol{y}) + n^{-1}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^T\left\{\frac{\partial}{\partial\boldsymbol{\theta}^T}\log\mathcal{L}(\boldsymbol{\theta})\right\}|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \\ &\quad +1/2(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^T\left\{n^{-1}\frac{\partial^2}{\partial\boldsymbol{\theta}^T\partial\,\boldsymbol{\theta}}\log\mathcal{L}(\boldsymbol{\theta})\right\}|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}) \\ &= n^{-1}\mathcal{L}(\hat{\boldsymbol{\theta}}) + 1/2(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^T\hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}) \\ &= -\,n^{-1}\log[h(\boldsymbol{\theta})] + c' \end{aligned} \tag{4}$$

since $\partial\log\mathcal{L}(\boldsymbol{\theta};\boldsymbol{y})/\partial\boldsymbol{\theta}^T|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = 0$ and constant $c' = n^{-1}\log\mathcal{L}(\hat{\boldsymbol{\theta}};\boldsymbol{y}) + c$ is free of $\boldsymbol{\theta}$. This motivates us to propose using the confidence density -log$[h(\boldsymbol{\theta})]$ to approximate $\log\mathcal{L}(\boldsymbol{\theta};\boldsymbol{y})$ to perform the regularized estimation. Specifically, we construct the following objective function, after adding the penalty terms $\kappa_\rho(\boldsymbol{\beta})$ and $\kappa_\tau(\boldsymbol{D}_{\boldsymbol{\gamma}})$ and dropping constant terms:

$$Q(\boldsymbol{\theta}) = (\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})^T\left(n^{-1}\hat{\boldsymbol{\Sigma}}\right)^{-1}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}) + n\kappa_\rho(\boldsymbol{\beta}) + n\kappa_\tau(\boldsymbol{\gamma}). \tag{5}$$

Define the regularized estimator $\hat{\boldsymbol{\theta}}_{\rho\tau}$=arg min$_{\boldsymbol{\theta}}\, Q(\boldsymbol{\theta})$. It is easy to see the relationship between $Q(\boldsymbol{\theta})$ and $Q^o(\boldsymbol{\theta})$ satisfying $Q(\boldsymbol{\theta}) = 2Q^o(\boldsymbol{\theta}) + o_p(1)$, provided that $\kappa_\rho(\boldsymbol{\beta}) = 2\kappa^o_\rho(\boldsymbol{\beta})$, and

$\kappa_\tau(\boldsymbol{\gamma}) = 2\kappa_\tau^o(\boldsymbol{\gamma})$. Due to the asymptotic normality property of the MLE, we notice that the objective function $Q(\boldsymbol{\theta})$ takes the similar form as the objective function based on the least squares approximation proposed by Wang and Leng (2007) for the generalized linear models. In optimizing $Q(\boldsymbol{\theta})$, the computational burden in numerically approximating the integration in $\log \mathcal{L}(\boldsymbol{\theta}; \boldsymbol{y})$ only occurs in deriving $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\Sigma}}$, which can be achieved by using existing software packages. Once $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\Sigma}}$ are obtained, optimizing $Q(\boldsymbol{\theta})$ to obtain $\hat{\boldsymbol{\theta}}_{\rho\tau}$ no longer involves the numerical integration of $\log \mathcal{L}(\boldsymbol{\theta}; \boldsymbol{y})$, thus greatly improves the computational efficiency. In Sections 4 and 5, we discuss how to perform the optimization of $Q(\boldsymbol{\theta})$ using existing software packages and demonstrate the computational efficiency of our method and assess the performance of $\hat{\boldsymbol{\theta}}_{\rho\tau}$ using simulation studies. Note that, the construction of $Q(\boldsymbol{\theta})$) is based on the asymptotic normality of the MLE. Based on the

### 3.2 *Statistical Properties of the Proposed Estimator*

To facilitate statistical inference, the penalty functions $\kappa_\rho(\cdot)$ and $\kappa_\tau(\cdot)$ could be the adaptive LASSO (Zou, 2006), or the smoothly clipped absolute deviation (SCAD) (Fan and Li, 2001). In this paper, we focus on the adaptive LASSO framework. Specifically, for the fixed effects selection, we use the adaptive LASSO choosing $\kappa_\rho(\boldsymbol{\beta}) = \sum_{f=1}^{p_f} \rho_f |\beta_f|$, where $\rho_f' s$ are the tuning parameters that control the penalty with respect to $|\beta_f|$, for $f = 1, 2, ..., p_f$. For the random-effect selection, we propose to use the adaptive group LASSO, following the rationale outlined in He, Tu and Wang et al. (2015): Let $\boldsymbol{\gamma}_m$ denote the $m$th row of $\boldsymbol{\Gamma}$, then $\boldsymbol{\gamma}_m \boldsymbol{\gamma}_m^T = \boldsymbol{D}_{mm}$ which is the $m$th variance component of the random effects $\boldsymbol{\Gamma b}_i$. Note that $\boldsymbol{\gamma}_m = 0 \Leftrightarrow \boldsymbol{D}_{mm} = \boldsymbol{D}_{mh} = \boldsymbol{D}_{hm} = 0$ for all $h$; that is, if $\boldsymbol{\gamma}_m = 0$, then the variance and covariance elements of $\boldsymbol{\Gamma b}_i$ involving $(\boldsymbol{\Gamma b}_i)_m$ are also 0. As a result, if a row vector $\boldsymbol{\gamma}_m$ is not selected, the random effect $(\boldsymbol{\Gamma b}_i)_m$ and the corresponding component in $\boldsymbol{z}$ are excluded from the model and the positive-definitiveness of $\boldsymbol{D}$ is preserved. Thus, we choose the adaptive group LASSO penalty $\kappa_\tau(\boldsymbol{\gamma}) = \sum_{m=2}^{p_r} \tau_m ||\boldsymbol{\gamma}_m||$ and $\tau_m' s$ are the tuning parameters corresponding to $||\boldsymbol{\gamma}_m||$ , where $||\cdot||$

denotes the $L_2$ norm of a vector. Note that the summation starts from $m = 2$ to keep the random intercept and preserve the within-subject correlation.

Denote the true value of $\boldsymbol{\theta}$ by $\boldsymbol{\theta}_0$, and the true non-zeros and true zeros in $\boldsymbol{\theta}$ by $\boldsymbol{\theta}_{0a}$ and $\boldsymbol{\theta}_{0b}$, respectively. Obviously, $Q(\boldsymbol{\theta})$ is strickly convex in $\boldsymbol{\theta}$. We establish the consistency and oracle properties of $\hat{\boldsymbol{\theta}}_{\rho\tau}$ in Theorem 1.

THEOREM 1: *Let* $a_{f,n} = \max\{\rho_j, j \leqslant f_0\}$, $b_{f,n} = \min\{\rho_j, j > f_0\}$, $a_{r,n} = \max\{\tau_j, j \leqslant r_0\}$, *and* $b_{r,n} = \min\{\tau_j, j > r_0\}$. *Then the regularized estimator* $\hat{\boldsymbol{\theta}}_{\rho\tau}$ *satisfies the following as* $n \to \infty$*:*

*(1) (Estimation Consistency) If* $n^{1/2}a_{f,n} \xrightarrow{P} 0$ *and* $n^{1/2}a_{r,n} \xrightarrow{P} 0$, $\hat{\boldsymbol{\theta}}_{\rho\tau} \xrightarrow{P} \boldsymbol{\theta}_0$*;*

*(2) (Selection Consistency) If* $n^{1/2}a_{f,n} \xrightarrow{P} 0$, $n^{1/2}a_{r,n} \xrightarrow{P} 0$, $n^{1/2}b_{f,n} \xrightarrow{P} \infty$, *and* $n^{1/2}b_{r,n} \xrightarrow{P} \infty$, $Pr(\hat{\boldsymbol{\theta}}_{\rho\tau,b} = \mathbf{0}) \to 1$, *where* $\hat{\boldsymbol{\theta}}_{\rho\tau,b}$ *denotes the components in* $\hat{\boldsymbol{\theta}}_{\rho\tau}$ *corresponding to* $\boldsymbol{\theta}_{0b}$*.*

*(3) (Oracle Property) If* $n^{1/2}a_{f,n} \xrightarrow{P} 0$, $n^{1/2}a_{r,n} \xrightarrow{P} 0$, $n^{1/2}b_{f,n} \xrightarrow{P} \infty$, *and* $n^{1/2}b_{r,n} \xrightarrow{P} \infty$, $n^{1/2}(\hat{\boldsymbol{\theta}}_{\rho\tau a} - \boldsymbol{\theta}_{0a}) \xrightarrow{D} \mathcal{N}(\mathbf{0}, \left[(\boldsymbol{\Sigma}^{-1})_{\boldsymbol{\theta}_{0a}}\right]^{-1})$, *where* $(\boldsymbol{\Sigma}^{-1})_{\boldsymbol{\theta}_{0a}}$ *is the submatrix of* $\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}$ *corresponding to true non-zero* $\boldsymbol{\theta}_{0a}$*.*
*The variance* $\left[(\boldsymbol{\Sigma}^{-1})_{\boldsymbol{\theta}_{0a}}\right]^{-1}$ *can be consistently estimated by* $\left[(\hat{\boldsymbol{\Sigma}}^{-1})_{\boldsymbol{\theta}_{0a}}\right]^{-1}$, *where* $(\hat{\boldsymbol{\Sigma}}^{-1})_{\boldsymbol{\theta}_{0a}}$ *is the submatrix of* $\hat{\boldsymbol{\Sigma}}^{-1}$ *corresponding to* $\boldsymbol{\theta}_{0a}$*.*

Sketch of the proof is provided in the supplementary material. Recall that the objective function $Q(\boldsymbol{\theta})$ is built by the confidence density $h(\boldsymbol{\theta})$, according to the joint asymptotic distribution of $\hat{\boldsymbol{\theta}}$. Notice that $h(\boldsymbol{\theta})$ is multivariate normal density. The true values of the means in the joint distribution are the same as those in the marginal distributions. Therefore, we propose an alternative estimation based on the marginal density of $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\gamma}}$, with respect to $h(\boldsymbol{\theta})$, to separately estimate $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$. These marginal confidence densities also correspond to the marginal asymptotic distributions of the MLEs $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\gamma}}$, respectively. We refer the previous estimation as the CD-joint estimation, and the following estimation as the CD-

separate estimation. To proceed, we propose the separate objective functions as follows:

$$\begin{aligned} Q_f(\boldsymbol{\beta}) &= (\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})^T[n^{-1}\hat{\boldsymbol{\Sigma}}_{\beta\beta}]^{-1}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})+n\kappa_\rho(\boldsymbol{\beta}), \\ Q_r(\boldsymbol{\gamma}) &= (\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma})^T\left[n^{-1}\hat{\boldsymbol{\Sigma}}_{\gamma\gamma}\right]^{-1}(\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma})+n\kappa_\tau(\boldsymbol{\gamma}), \end{aligned}$$

where $\hat{\boldsymbol{\Sigma}}_{\beta\beta}$ and $\hat{\boldsymbol{\Sigma}}_{\gamma\gamma}$ are submatrices of $\hat{\boldsymbol{\Sigma}}$ corresponding to the marginal variance-covariance of $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\gamma}}$, respectively; the penalty terms $\kappa_\rho(\boldsymbol{\beta})$ and $\kappa_\tau(\boldsymbol{\gamma})$ remain the same as those in $Q(\boldsymbol{\theta})$. Define the regularized estimators as $\hat{\boldsymbol{\beta}}^s_\rho$=arg min$_{\boldsymbol{\beta}}\, Q_f(\boldsymbol{\beta})$ and $\hat{\boldsymbol{\gamma}}^s_\tau$=arg min$_{\boldsymbol{\gamma}}\, Q_r(\boldsymbol{\gamma})$. As noted previously, the true values of the underlying parameters $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ in the joint distribution of $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\gamma}}$ in $h(\boldsymbol{\theta})$ are the same as those in the individual marginal distributions of $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\gamma}}$, respectively. Therefore, the true values of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ for the estimators based on the CD-joint estimation and CD-separate estimation are the same.

Denote the true values of $\boldsymbol{\beta}$ by $\boldsymbol{\beta}_0$, and $\boldsymbol{\gamma}$ by $\boldsymbol{\gamma}_0$. Let $\boldsymbol{\beta}_{0a}$ and $\boldsymbol{\beta}_{0b}$ denote the true non-zeros and true zeros of $\boldsymbol{\beta}$, respectively, and $\boldsymbol{\gamma}_{0a}$ and $\boldsymbol{\gamma}_{0b}$ denote true non-zeros and true zeros of $\boldsymbol{\gamma}$, respectively. We establish the consistency and oracle properties for $\hat{\boldsymbol{\beta}}^s_\rho$ and $\hat{\boldsymbol{\gamma}}^s_\tau$ as follows.

THEOREM 2: *Let* $a_{f,n}=\max\{\rho_j, j\leqslant f_0\}$, *and* $b_{f,n}=\min\{\rho_j, j>f_0\}$. *Then the regularized estimator* $\hat{\boldsymbol{\beta}}^s_\rho$ *satisfies the following as* $n\to\infty$*:*

*(1) (Estimation Consistency) If* $n^{1/2}a_{f,n}\xrightarrow{P}0$, $\hat{\boldsymbol{\beta}}^s_\rho\xrightarrow{P}\boldsymbol{\beta}_0$*;*

*(2) (Selection Consistency) If* $n^{1/2}a_{f,n}\xrightarrow{P}0$, *and* $n^{1/2}b_{f,n}\xrightarrow{P}\infty$, $Pr(\hat{\boldsymbol{\beta}}^s_{\rho,b}=\mathbf{0}\,)\to 1$*; where* $\hat{\boldsymbol{\beta}}^s_{\rho,b}$ *denotes the components in* $\hat{\boldsymbol{\beta}}^s_\rho$ *corresponding to* $\boldsymbol{\beta}_{0b}$*.*

*(3) (Oracle Property) If* $n^{1/2}a_{f,n}\xrightarrow{P}0$, *and* $n^{1/2}b_{f,n}\xrightarrow{P}\infty$, $n^{1/2}\left(\hat{\boldsymbol{\beta}}^s_{\rho,a}-\boldsymbol{\beta}_{0a}\right)\xrightarrow{D}\mathcal{N}(\mathbf{0},\left[(\boldsymbol{\Sigma}^{-1}_{\beta\beta})_{\beta_{0a}}\right]^{-1})$, *where* $(\boldsymbol{\Sigma}^{-1}_{\beta\beta})_{\beta_{0a}}$ *is the submatrix of* $\boldsymbol{\Sigma}^{-1}_{\beta\beta}$ *composed of elements (variance and covariance) corresponding to true non-zero* $\boldsymbol{\beta}_{0a}$*.*

THEOREM 3: *Let* $a_{r,n}=\max\{\tau_j, j\leqslant r_0\}$, *and* $b_{r,n}=\min\{\tau_j, j>r_0\}$. *Then the regularized estimator* $\hat{\boldsymbol{\gamma}}^s_\tau$ *satisfies the following as* $n\to\infty$*:*

*(1) (Estimation Consistency) If* $n^{1/2}a_{r,n}\xrightarrow{P}0$, $\hat{\boldsymbol{\gamma}}^s_\tau\xrightarrow{P}\boldsymbol{\gamma}_0$*;*

*(2) (Selection Consistency) If* $n^{1/2}a_{r,n} \xrightarrow{P} 0$, *and* $n^{1/2}b_{r,n} \xrightarrow{P} \infty$, $Pr(\hat{\boldsymbol{\gamma}}^s_{\tau,b} = \mathbf{0}\ ) \to 1$*; where* $\hat{\boldsymbol{\gamma}}^s_{\tau,b}$ *denotes the components in* $\hat{\boldsymbol{\gamma}}^s_\tau$ *corresponding to* $\boldsymbol{\gamma}_{0b}$.

*(3) (Oracle Property) If* $n^{1/2}a_{r,n} \xrightarrow{P} 0$, *and* $n^{1/2}b_{r,n} \xrightarrow{P} \infty$, $n^{1/2}\left(\hat{\boldsymbol{\gamma}}^s_{\tau,a} - \boldsymbol{\gamma}_{0a}\right) \xrightarrow{D} \mathcal{N}(\mathbf{0}, \left[(\boldsymbol{\Sigma}^{-1}_{\gamma\gamma})_{\gamma_{0a}}\right]^{-1})$, *where* $(\boldsymbol{\Sigma}^{-1}_{\gamma\gamma})_{\boldsymbol{\gamma}_{0a}}$ *is the submatrix of* $\boldsymbol{\Sigma}^{-1}_{\gamma\gamma}$ *composed of elements (variance and covariance) corresponding to true non-zero* $\boldsymbol{\gamma}_{0a}$.

The proof for Theorem 2 and Theorem 3 are similar to that for Theorem 1, thus are omitted.

Although the dispersion parameter $\phi$ is often treated as a nuisance parameter, it can also be included in the separate estimation approach. Combining $\phi$ and $\boldsymbol{\gamma}$, we modify $Q_r(\boldsymbol{\gamma})$ by $Q^*_r(\boldsymbol{\gamma},\varphi)$ given below, based on the (marginal) joint distribution of $\hat{\boldsymbol{\gamma}}$ and $\hat{\phi}$ by (3):

$$Q^*_r(\boldsymbol{\gamma},\varphi) = \begin{pmatrix} \hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma} \\ \hat{\phi} - \phi \end{pmatrix}^T [n^{-1}\hat{\boldsymbol{\Sigma}}_{\gamma\phi}]^{-1} \begin{pmatrix} \hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma} \\ \hat{\phi} - \phi \end{pmatrix} + n\kappa_\tau(\boldsymbol{\gamma}),$$

where $\hat{\boldsymbol{\Sigma}}_{\gamma\phi}$ is the variance-covariance of $\hat{\boldsymbol{\gamma}}$ and $\hat{\phi}$.

For linear mixed models, $\hat{\boldsymbol{\beta}}$ is uncorrelated with $\hat{\boldsymbol{\gamma}}$ and $\hat{\phi}$ (e.g., Wang and Merkle (2018)). Thus $Q(\boldsymbol{\theta}) = Q_f(\boldsymbol{\beta}) + Q^*_r(\boldsymbol{\gamma}, \phi)$, implying that

$$\begin{aligned} \arg\min_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}) &= \arg\min_{\boldsymbol{\theta}} \{Q_f(\boldsymbol{\beta}) + Q^*_r(\boldsymbol{\gamma}, \phi)\} \\ &= \arg\min_{\boldsymbol{\beta}} Q_f(\boldsymbol{\beta}) + \arg\min_{\boldsymbol{\gamma},\phi} Q^*_r(\boldsymbol{\gamma}, \phi), \end{aligned}$$

the CD-joint estimation and CD-separate estimation are identical.

## 4. Optimization and determination of tuning parameters.

For joint estimation of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, we follow the method of Zhang and Lu (2007) and rewrite the objective function $Q(\boldsymbol{\theta})$ as

$$Q(\boldsymbol{\theta}) = (\Lambda\boldsymbol{\theta} - \Psi)^T (\Lambda\boldsymbol{\theta} - \Psi) + n\kappa_\rho(\boldsymbol{\beta}) + n\kappa_\tau(\boldsymbol{\gamma}), \tag{6}$$

where $\Psi = \Lambda\hat{\boldsymbol{\theta}}$ , and $\Lambda$ can be obtained using the singular value decomposition such that $\left(n^{-1}\hat{\boldsymbol{\Sigma}}\right)^{-1} = \Lambda^T\Lambda$. Then the function in (6) is a typical convex optimization problem and

can be solved by standard software packages, for instance, the R packages `glmnet` (Friedman et al., 2010) and `gglasso` (Yang and Zou, 2015). Same approach can also be applies to optimize $Q_f(\boldsymbol{\beta})$, $Q_r(\boldsymbol{\gamma})$ and $Q_r^*(\boldsymbol{\gamma},\varphi)$. Let $\Psi_{\boldsymbol{\beta}} = \Lambda_{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}$ and $\Psi_{\boldsymbol{\gamma}} = \Lambda_{\boldsymbol{\gamma}}\hat{\boldsymbol{\gamma}}$, where $\Lambda_{\boldsymbol{\beta}}$ and $\Lambda_{\boldsymbol{\gamma}}$ can be obtained using the singular value decomposition such that $[n^{-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta\beta}}]^{-1} = \Lambda_{\boldsymbol{\beta}}^T\Lambda_{\boldsymbol{\beta}}$ and $[n^{-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\gamma\gamma}}]^{-1} = \Lambda_{\boldsymbol{\gamma}}^T\Lambda_{\boldsymbol{\gamma}}$. As a result, $Q_f(\boldsymbol{\beta}) = \left(\Lambda_{\boldsymbol{\beta}}\boldsymbol{\beta} - \Psi_{\boldsymbol{\beta}}\right)^T\left(\Lambda_{\boldsymbol{\beta}}\boldsymbol{\beta} - \Psi_{\boldsymbol{\beta}}\right) + n\kappa_\rho(\boldsymbol{\beta})$ and $Q_\gamma(\boldsymbol{\gamma}) = \left(\Lambda_{\boldsymbol{\gamma}}\boldsymbol{\gamma} - \Psi_{\boldsymbol{\gamma}}\right)^T\left(\Lambda_{\boldsymbol{\gamma}}\boldsymbol{\gamma} - \Psi_{\boldsymbol{\gamma}}\right)$, respectively.

Typically, the tuning parameters can be chosen using the approaches of cross validation (CV) or generalized cross validation (GCV). But these methods can be computationally extensive. With the simple solution suggested by Zou (2006), we consider $\rho_f = \rho|\hat{\beta}_f|^{-\varphi_f}$ and $\tau_m = \tau||\hat{\boldsymbol{\gamma}}_m||^{-\varphi_r}$, for $f = 1, 2, ..., p_f$ and $m = 2, 3, ..., p_r$, where $\hat{\beta}_f$ and $\hat{\boldsymbol{\gamma}}_m$ are the maximum likelihood estimates for $\beta_f$ and $\gamma_m$, and $\varphi_f$ and $\varphi_r$ are pre-specified positive number. In our simulations and data analysis, we chose $\varphi_f = \varphi_r = 1$ for simplicity.

To determine the tuning parameters $\rho$ and $\tau$, we consider to minimize BIC, per recommendations by prior research (e.g., Wang and Leng (2007) and Wang and Leng (2008)). For CD-joint estimation, define the BIC as: $BIC_{\rho,\tau} = n(\hat{\boldsymbol{\theta}}_{\rho\tau} - \hat{\boldsymbol{\theta}})^T\hat{\boldsymbol{\Sigma}}^{-1}(\hat{\boldsymbol{\theta}}_{\rho\tau} - \hat{\boldsymbol{\theta}}) + (\log n)(df_\rho + df_\tau)$, where $df_\rho$ is the number of nonzero coefficients in $\hat{\boldsymbol{\beta}}_\rho$, and $df_\tau$ is the number of groups with non-zero within-group coefficients. For CD-separate estimation, we define $BIC_{f,\rho} = n(\hat{\beta}_\rho^s - \hat{\boldsymbol{\beta}})^T\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta\beta}}^{-1}(\hat{\beta}_\rho^s - \hat{\boldsymbol{\beta}}) + (\log n)df_\rho$ for $Q_f(\boldsymbol{\beta})$, $BIC_{r,\tau} = n(\hat{\boldsymbol{\gamma}}_\tau^s - \hat{\boldsymbol{\gamma}})^T\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\gamma\gamma}}^{-1}(\hat{\boldsymbol{\gamma}}_\tau^s - \hat{\boldsymbol{\gamma}}) + (\log n)df_\tau$ for $Q_r(\boldsymbol{\gamma})$ and $BIC_{r,\tau}^* = \begin{pmatrix} \hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma} \\ \hat{\phi} - \phi \end{pmatrix}^T [n^{-1}\hat{\boldsymbol{\Sigma}}_{\gamma\phi}]^{-1} \begin{pmatrix} \hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma} \\ \hat{\phi} - \phi \end{pmatrix} + (\log n)df_\tau$ for $Q_r^*(\boldsymbol{\gamma},\varphi)$. For linear mixed models, $BIC_{\rho,\tau} = BIC_{r,\rho} + BIC_{r,\tau}^*$. The $\rho$ and $\tau$ that minimize $BIC_{\rho,\tau}$ should also minimize $BIC_{r,\rho}$ and $BIC_{r,\tau}^*$, respectively, and vice versa.

## Simulation Studies

We conducted simulation studies to examine the performance of the proposed method and compared it with the method of Hui, Mueller and Welsh (2017) for its better performance

in variable selection and computation efficiency than some existing methods. Data were simulated under 3 scenarios according to model (1) with $p_f = 16$ fixed effects and $p_r = 4$ random effects, both including intercepts. In Scenario 1, $y'_{ij}s$ were simulated from the linear mixed model such that $y_{ij}|\boldsymbol{b}_i$ is following $N(\boldsymbol{x}_{ij}^T\boldsymbol{\beta}+\boldsymbol{z}_{ij}^T\boldsymbol{\Gamma}\boldsymbol{b}_i, \sigma^2)$ with $\phi = \sigma^2 = 1$. In Scenario 2, we simulated binary data from the random effects logistic regression model. In Scenario 3, we simulated count data from the random effects Poisson model using a log link. The true value for $\boldsymbol{\beta}$ was $\boldsymbol{\beta}_0 = (\mathbf{1}_6, \mathbf{0}_{10})$ for Scenarios 1 and 2, and $\boldsymbol{\beta}_0 = (-1, \mathbf{1}_5, \mathbf{0}_{10})$ for Scenario 3. The true $4 \times 4$ random effect covariance matrix $\boldsymbol{D}$ is given by $vech(\boldsymbol{D}) = (9, 4.8, 0.6, 0; 4, 0.9, 0; 1, 0; 0)$ for Scenario 1, $vech(\boldsymbol{D}) = (3, 1.2, 0.8, 0; 2, 0.5, 0; 1, 0; 0)$ for Scenarios 2 and 3, i.e., only the first three components of $\boldsymbol{z}_{ij}$, including the random intercept, are true predictors. As a result, we express the corresponding Cholesky decomposition lower triangular matrix as $\boldsymbol{\Gamma} = (3; 1.60, 0.20; 1.20, 0.57, 0.80; \mathbf{0}_4)$ for Scenario 1 and $\boldsymbol{\Gamma} = (1.73; 0.69, 1.23; 0.46, 0.15, 0.88; \mathbf{0}_4)$ for Scenarios 2 and 3. In each scenario, we considered varying number of clusters $n$ and cluster size $m$. All elements of $\boldsymbol{x}_{ij}$ were generated from the standard normal distribution. The first 3 elements of $\boldsymbol{z}_{ij}$ equal the first 3 elements of $\boldsymbol{x}_{ij}$, including intercepts. For the proposed methods, we refer CD-joint and CD-separate estimations as CD-J and CD-S, respectively. We refer the method of Hui et al. by rPQL. We assessed the performance of the biasedness, empirical standard error (ESE), coverage probability of 95% confidence intervals (CovP), percentage of selection (% Sel) and the average computation time (Time (mins)). When we calculated computation time, we excluded the time fitting GLMM models to derive the MLEs, and only calculated the time of the regularization process, including the determination of tuning parameters, when we applied either the proposed approaches or the rPQL approach. Results were summarized in Tables 1 to 3, based on 1000 simulation runs.

For the linear mixed models (Scenario 1), the first 3 fixed effect estimates in $\hat{\beta}_\rho^s$ showed some bias from the true values when the number of clusters $n$ is moderately small (e.g.,

$n = 60$). Note that the corresponding covariates are also associated with random effects. Similar situation also happened to the rPQL estimates. These CD estimates associated with larger values in the random effect $\boldsymbol{D}$ elements (e.g., 9, 4.8, 4 in $\boldsymbol{D}$) tended to have larger bias. The bias diminished as $n$ increased. When $n = 500$, the bias is nearly minimal. For the fixed effect estimates whose corresponding covariates are not associated with random effects, the bias is minimal even for small $n$. The empirical variance, as expected, decreased with $n$. The coverage probability of 95% confidence interval (CI) is close but slightly under the nominal 95% level. The coverage probability improved when $n$ was large. For the performance of variable selection, the selection of true covariates was close to 100%. The noise covariates were selected but at a very low rate, especially for large $n$. Compared to the rPQL method, the proposed methods resulted in slightly smaller false positive selection rate. For random effect selections, results were similar. In general, the random effect parameter estimates of the proposed CD method were close to the ture values; the empirical variance decreases with $n$, and the coverage probability of 95%CI is close to but slightly under the nominal 95% level. The selection of true covariates was close to 100%. The noise covariates were selected but at a very lower rate, compared to the rPQL method. In terms of computation time, the proposed method generally took $< 1$ minute. Therefore, when the $n$ is moderate to large, the proposed CD method can be an attractive and competitive approach.

For the random effects logistic regression models (Scenario 2), results are similar in the sense that the bias and ESE decreased with $n$, and the coverage probability improved as $n$ increased. The performance of the separate estimation (CD-S) outperformed the joint estimation (CD-J) with respect to bias, coverage probability of 95% CI, and the percentage of selection. This might be caused by the constraint of $\rho = \tau$ we set when we optimized $Q(\boldsymbol{\theta})$ using the `gglasso` function in R. Some perturbations in the relationship of $\rho$ and $\tau$ (e.g., $\rho = a\tau, a > 0$) that relax this condition might improve the performance of the joint

estimation. We also noted that the proposed CD approaches also outperformed the rPQL method, in terms of bias, percentage of selection and computation time.

Results for Scenario 3, the random effects Poisson regression models, are also similar. However, we noted that the coverage probabilities based on CD-J and CD-S approaches were not improved when $n = 200$ increased to $n = 400$, but were similar to the coverage probabilities based on the inference of MLEs. To improve the coverage of confidence intervals, one may consider to replace $\hat{\mathbf{\Sigma}}$ in the inference by the Huber-White sandwich estimator (Freedman (2006) and Wang and Merkle (2018)).

[Table 1 about here.]

[Table 2 about here.]

[Table 3 about here.]

## 5. Data Analysis

We applied the proposed CD approach to the two longitudinal cancer studies that motivated our method. The first study wanted to identify demographic and clinical covariates that may be associated with the tumor size in lung cancer patients. The second study aimed to relate a set of covariates to the incidence of common mammographic sequelae after breast conserving surgery and radiation therapy (tian2016comparison). For both studies, we wanted to identify important covariates that may predict the outcomes as fixed effects. We also wanted to include random effects to assess if some covariates exhibit heterogeneous effect.

Data Example 1. We fitted the linear mixed model and applied the proposed CD approach to identify covariates that may predict the tumor size using repeatedly measured data from 101 lung cancer patients. The outcome variable was log-transformed tumor size (unit?). Potential predictors for the fixed effect selection included Weeks (weeks since lung cancer diagnosis or treatment?), Age (age at diagnosis?), gender (females yes/no), MLDGy

(how much radiation normal lung tissues receives radiation?), LungV20 (portions of lung volume receives 20 Gy of ? ), and smoking (yes/no). Potential predictors for the random effect selection included Weeks, Age, MLDGy and LungV20. All continuous variables were standardized to mean 0 and variance 1. In order to test the performance of variable selection of our proposed method, we added 3 unrelated and randomly generated standard normal noise variables into the fixed effects selection. Results were summarized in Table 4. For fixed effect selection, Weeks, Gender, MLDGy and LungV20 were selected. Results suggested that the mean tumor size decreased with weeks ($\hat{\beta}_\rho^s = -0.117$, 95%CI: $3.554, 4.197$), and was bigger in men than women ($\hat{\beta}_\rho^s = 0.434$, 95%CI: $0.016, 0.852$). The mean tumor size also increased with MLDGy ($\hat{\beta}_\rho^s = 0.912$, 95%CI: $0.299, 1.526$) and decreased with LungV20 ($\hat{\beta}_\rho^s = -0.789$, 95%CI: $-1.399, -0.180$). For random effects, Weeks was selected ($\hat{\Gamma}_{21} = 0.001$, 95%CI: $0.001, 0.002$, and $\hat{\Gamma}_{22} = 0.062, 95\%CI : 0.061, 0.063$), in addition to the intercept ($\hat{\Gamma}_{11} = 1.233$, 95%CI: $1.231, 1.235$), suggesting a positive within-person correlation and a heterogeneous effect by Weeks. Adopting the commonly used refit approach (reference), we refitted the model that only included the selected fixed and random effects using the proposed CD approach. Results were similar.

Data Example 2. The second dataset included data from 89 breast cancer patients to identify covariates associated with the incidence of common mammographic sequelae after breast conserving surgery and radiation therapy. A total of 605 longitudianlly measured observations were in the data analysis. We fitted a random effects logistic regression model with calcification (yes/no) as the dependent vairable, and applied the proposed CD separate estimation approach to select fixed effects from the following covariates: age, years from radiation therapy, African Americans (yes/no), Her2/neu positive (yes/no), adjuvant chemo and/or hormonal therapy (yes/no), smoking (yes/no), bilateral disease (yes/no) and 2 unrelated and randomly generated standard normal noise variables. Potential covariates for the

random effect selection included the intercept, age, and 1 unrelated noise variable that was also included in the fixed effect selection. All continuous variables were standardized to mean 0 and variance 1. Results were summarized in Table 5. For fixed effects, age , years from radiation therapy, African Americans and Her2/neu positive were selected. Results suggested the risk of calcification increased with age ($\hat{\beta}_\rho^s = 4.298$, 95%CI: $2.449, 6.148$), Years since radiation therapy ($\hat{\beta}_\rho^s = 1.411$, 95%CI: $1.048, 1.773$) and was lower in African Americans ($\hat{\beta}_\rho^s = -5.878$, 95%CI: $-10.503, -1.253$) and Her2/neu positive ($\hat{\beta}_\rho^s = -1.617$, 95%CI: $-2.899, -0.335$). For random effects, Age ($\hat{\Gamma}_{21} = 7.292$, 95%CI: $0.175, 14.410$, and $\hat{\Gamma}_{22} = 6.318$,95%CI: $2.912, 9.725$) was selected, in addition to the random intercept ($\hat{\Gamma}_{11} = 4.372$, 95%CI: $2.228, 6.518$),suggesting a positive within-person correlation and a heterogeneous effect by Age. After refitting the random effects logistic model that only included the selected fixed and random effects selected using the proposed CD separate estimation approach, results were similar.

[Table 4 about here.]

[Table 5 about here.]

## 6. Discussion

In this paper, we propose a regularized estimation approach to select both fixed and random effects in GLMMs. Specifically, we show that the log marginal likelihood function, after integrating out the random effects, can be approximated by the log confidence density of the model parameters based on the asymptotic distribution of the MLE. As a result, we proposed to construct the objective functions using the confidence density as opposed to using the observed data likelihood function. Specifically, we proposed to estimation methods: One is referred as the CD-joint estimation, based on the joint asymptotic distribution of the MLE of the fixed effect and random effect parameters; the other is referred as the CD-

separate estimation, based on the individual marginal distribution of the MLES of the fixed effect and random effect parameters, respectively. Due to the asymptotic normality property of the MLE, we notice that the CD-based objective functions takes a similar form to the objective function based on the least squares approximation proposed by Wang and Leng (2007) for the generalized linear models. With a proper choice of regularization parameters in the format of adaptive LASSO framework, we show that the proposed estimators have the consistency and oracle properties. For practical use, the construction of the objective functions can take advantage of existing software packages that provide ready solutions of $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\Sigma}}$. Thus, optimizing the proposed objective functions to obtain the regularized estimators can bypass the computational complexity in numerically approximating the integral in the log marginal likelihood function and escalates computational efficiency. In the simulation studies, we demonstrated the performance of the proposed estimators, in terms of bias, coverage probability of 95% confidence intervals, variable selection and computational efficiency. Results showed that the separate estimation outperformed the joint estimation.

The key principle of our method is to approximate the log likelihood function by the log confidence density of the model parameters based on the asymptotic distribution of the maximum likelihood estimator. In this paper, our proposed method is designed for GLMM. In many real life problems, some complicated modeling approaches, such as mixture regression models and joint analysis of survival and longitudinal data analysis, may apply and the statistical inference is largely based on the likelihood based approach. The variable selection for these models might be challenging due to the complexity in these models and computation. We plan to extend our approach and apply it to perform variable selection to these models as future work.

**Web Appendix: Proof of Theorem 1**

(1) Since the objective function $Q(\boldsymbol{\theta})$ is a strictly convex function for $\boldsymbol{\theta}$, a local consistent minimizer is the global consistent min imizer. Therefore, it suffices to show the existence of a local consistent minimizer, then the estimation consistency follows immediately. Following **?** and letting $\mathbf{u} = (u_1, u_2, \ldots, u_d)^T$, where $d$ =length($\boldsymbol{\theta}$), the existence of a local consistent minimizer is implied by the fact that for any given $\epsilon > 0$, there exists a large constant $C$ such that

$$\lim_{n\to\infty} P\left\{\inf_{\mathbf{u}\in\mathbf{R}^d:||\mathbf{u}||_2=C} Q(\boldsymbol{\theta}_0 + n^{-1/2}\mathbf{u}) > Q(\boldsymbol{\theta}_0)\right\} > 1-\epsilon, \tag{7}$$

where $||\boldsymbol{a}||_2 = (\boldsymbol{a}^T\boldsymbol{a})^{1/2}$ for a column vector $\boldsymbol{a}$.

To show this, consider

$$\begin{aligned}
&Q(\boldsymbol{\theta}_0 + n^{-1/2}\mathbf{u}) - Q(\boldsymbol{\theta}_0)\\
&\quad= \mathbf{u}^T\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{u} + 2\mathbf{u}^T\hat{\boldsymbol{\Sigma}}^{-1}\left\{n^{1/2}(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}})\right\} + n\sum_{f=1}^{p_f}\rho_f(|\beta_{0f} + n^{-1/2}u_f| - |\beta_{0f}|)\\
&+n\sum_{m=2}^{p_r}\tau_m(||\boldsymbol{\gamma}_{0m} + n^{-1/2}u_{p_f+m}|| - ||\boldsymbol{\gamma}_{0m}||)\\
&\quad\geqslant \mathbf{u}^T\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{u} + 2\mathbf{u}^T\hat{\boldsymbol{\Sigma}}^{-1}\left\{n^{1/2}(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}})\right\} + n\sum_{\{f:\beta_{0f}\neq 0\}}\rho_f(|\beta_{0f} + n^{-1/2}u_f| - |\beta_{0f}|)\\
&+n\sum_{\substack{m=2\\ \{m:\boldsymbol{\gamma}_{0m}\neq 0\}}}\tau_m(||\boldsymbol{\gamma}_{0m} + n^{-1/2}u_{p_f+m}|| - ||\boldsymbol{\gamma}_{0m}||)\\
&\quad\geqslant \mathbf{u}^T\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{u} + 2\mathbf{u}^T\hat{\boldsymbol{\Sigma}}^{-1}\left\{n^{1/2}(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}})\right\} - n\sum_{\{f:\beta_{0f}\neq 0\}}\rho_j|n^{-1/2}u_j| - n\sum_{\substack{m=2\\ \{m:\boldsymbol{\gamma}_{0m}\neq 0\}}}\tau_m||n^{-1/2}u_{p_f+m}|| \qquad (8)\\
&\quad\geqslant \mathbf{u}^T\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{u} + 2\mathbf{u}^T\hat{\boldsymbol{\Sigma}}^{-1}\left\{n^{1/2}(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}})\right\} - (n^{1/2}f_0a_{f,n} + n^{1/2}r_0a_{r,n})||\mathbf{u}||, \qquad (9)
\end{aligned}$$

followed by the triangle inequality, $a_{f,n} = max\{\rho_j, j \leqslant f_0\}$ and $a_{r,n} = max\{\tau_j, j \leqslant r_0\}$. According to the conditions $n^{1/2}a_{f,n} \xrightarrow{P} 0$ and $n^{1/2}a_{r,n} \xrightarrow{P} 0$, the third term in (**??**) is $o_p(1)$. The first term in **??** converges in probability to $\mathbf{u}^T\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{u}$. The second term in (**??**) is bounded by $2C||\hat{\boldsymbol{\Sigma}}^{-1}n^{1/2}(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}})||$, which is linear in $C$ with a coefficient $2||\hat{\boldsymbol{\Sigma}}^{-1}n^{1/2}(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}})|| = O_p(1)$. As the variance $\boldsymbol{\Sigma}$ and its estimate $\hat{\boldsymbol{\Sigma}}$ are positive semidefinite, the first term in (**??**) is larger

than $\mu_{min}(\hat{\mathbf{\Sigma}}^{-1})C^2 \xrightarrow{P} \mu_{min}(\hat{\mathbf{\Sigma}}^{-1})C^2$, where $\mu_{min}(.)$ refers to the minimal eigenvalue. It follows that, with probability going to 1, the first term in (**??**) is larger than $\mu_{min}(\hat{\mathbf{\Sigma}}^{-1})C^2$ which is quadratic in $C$. By choosing a sufficiently large $C$, the first term dominates the other two terms with arbitrarily large probability. Hence, by choosing a sufficiently large $C$, (**??**) holds and the proof of estimation consistency is completed.

(2) The selection consistency can be shown by contradiction. To show $Pr(\hat{\boldsymbol{\theta}}_{\rho\tau,b} = \mathbf{0}\ ) \to 1$, we show that $Pr(\hat{\beta}_{\rho\tau.j} = 0) \to 1$ for any $f_0 < j \leqslant p_f$ and $Pr(\hat{\boldsymbol{\gamma}}_{\rho\tau,m} = \mathbf{0}) \to 1$ for any $r_0 < m \leqslant p_r$. Suppose $\hat{\beta}_{\rho\tau.j} \neq 0$ for some $f_0 < j \leqslant p_f$, then by definition

$$n^{-1/2} \left.\frac{\partial Q(\boldsymbol{\theta})}{\partial \beta_j}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{\rho\tau}} = 2\hat{\mathbf{\Sigma}}^{-1}_{(\beta_j)} n^{1/2}(\hat{\boldsymbol{\theta}}_{\rho\tau} - \hat{\boldsymbol{\theta}}) + n^{1/2}\rho_j sgn(\hat{\beta}_{\rho\tau.j}) = 0, \tag{10}$$

where $\hat{\mathbf{\Sigma}}^{-1}_{(\beta_j)}$ represents the row vector of $\hat{\mathbf{\Sigma}}^{-1}$ corresponding to the position of $\beta_j$ and $sgn(.)$ is the sign function. It can be shown that the first term on the right hand side of (10) is $O_p(1)$. Based on the condition $n^{1/2}b_{f,n} \xrightarrow{P} \infty$, we have $n^{1/2}\rho_j \geqslant n^{1/2}b_{f,n} \xrightarrow{P} \infty$. Then to satisfy (10), with probability tending to 1, $\hat{\beta}_{\rho\tau.j} = 0$, which contradicts the assumed condition that $\hat{\beta}_{\rho\tau.j} \neq 0$. As a result, with probability tending to 1, $\hat{\beta}_{\rho\tau.j} = 0$ for any $f_0 < j \leqslant p_f$. Similarly, suppose $\hat{\boldsymbol{\gamma}}_{\rho\tau,m} \neq 0$ for some $r_0 < m \leqslant p_r$, then by definition

$$n^{-1/2} \left.\frac{\partial Q(\boldsymbol{\theta})}{\partial \boldsymbol{\gamma}_m}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_{\rho\tau}} = 2\hat{\mathbf{\Sigma}}^{-1}_{(\boldsymbol{\gamma}_m)} n^{1/2}(\hat{\boldsymbol{\theta}}_{\rho\tau} - \hat{\boldsymbol{\theta}}) + n^{1/2}\tau_m \frac{\hat{\boldsymbol{\gamma}}_{\rho\tau,m}}{\left\|\hat{\boldsymbol{\gamma}}_{\rho\tau,m}\right\|} = 0, \tag{11}$$

where $\hat{\mathbf{\Sigma}}^{-1}_{(\boldsymbol{\gamma}_m)}$ represents the submatrix consisting of of the row vectors of $\hat{\mathbf{\Sigma}}^{-1}$corresponding to the position of $\boldsymbol{\gamma}_m$. It can be shown that the first term on the right hand side of (10) is $O_p(1)$. Based on the condition $n^{1/2}b_{r,n} \xrightarrow{P} \infty$, we have $n^{1/2}\tau_m \geqslant n^{1/2}b_{r,n} \xrightarrow{P} \infty$. Then to satisfy (10), with probability tending to 1, $\hat{\boldsymbol{\gamma}}_{\rho\tau,m} = 0$, which contradicts the assumed condition that $\hat{\boldsymbol{\gamma}}_{\rho\tau,m} \neq 0$. As a result, with probability tending to 1, $\hat{\boldsymbol{\gamma}}_{\rho\tau,m} = 0$ for any $r_0 < m \leqslant p_r$. This completes the proof of selection consistency.

(3) To prove the oracle property, we first ease the notation: Without loss of generality, we use $\boldsymbol{\theta}^r$ to denote the re-arranged $\boldsymbol{\theta}$ such that the first $f_0 + r_0(r_0+1)/2 + 1$ elements of $\boldsymbol{\theta}^r_0$ are the true non-zero parameters in the order of $\boldsymbol{\beta}_{0a}, \boldsymbol{\gamma}_{0a}$, and $\phi$, and the remaining elements

are $\boldsymbol{\beta}_{0b}$ and $\boldsymbol{\gamma}_{0a}$. In the same order, we use $\hat{\boldsymbol{\theta}}^r$, $\hat{\boldsymbol{\theta}}^r_{\rho\tau}$ , and $\hat{\boldsymbol{\Sigma}^r}^{-1}$ to denote the re-arranged $\hat{\boldsymbol{\theta}}$ , $\hat{\boldsymbol{\theta}}_{\rho\tau}$ and $\hat{\boldsymbol{\Sigma}}^{-1}$, resepctively. We also decompose $(\boldsymbol{\Sigma}^r)$ and $(\boldsymbol{\Sigma}^r)^{-1}$ into block matrices:

$$\boldsymbol{\Sigma}^r = \begin{pmatrix} \boldsymbol{\Sigma}^r_{aa} & \boldsymbol{\Sigma}^r_{ab} \\ \boldsymbol{\Sigma}^r_{ba} & \boldsymbol{\Sigma}^r_{bb} \end{pmatrix}, \quad (\boldsymbol{\Sigma}^r)^{-1} = \boldsymbol{\Omega} = \begin{pmatrix} \boldsymbol{\Omega}_{aa} & \boldsymbol{\Omega}_{ab} \\ \boldsymbol{\Omega}_{ba} & \boldsymbol{\Omega}_{bb} \end{pmatrix},$$

where $\boldsymbol{M}_{aa}$ is the leading $a \times a$ submatrix of $\boldsymbol{M}$. Decomposing $Q(\boldsymbol{\theta})$, we have

$$\begin{aligned} & Q(\boldsymbol{\theta}^r) \\ = \;& (\hat{\boldsymbol{\theta}}^r - \boldsymbol{\theta}^r)^T [n^{-1}\hat{\boldsymbol{\Omega}}^r]^{-1} (\hat{\boldsymbol{\theta}}^r - \boldsymbol{\theta}^r) + n\kappa_\rho(\boldsymbol{\beta}) + n\kappa_\tau(\boldsymbol{\gamma}) \\ = \;& n \left\{ \begin{pmatrix} \boldsymbol{\theta}^r_a \\ \boldsymbol{\theta}^r_b \end{pmatrix} - \begin{pmatrix} \hat{\boldsymbol{\theta}}^r_a \\ \hat{\boldsymbol{\theta}}^r_b \end{pmatrix} \right\}^T \begin{pmatrix} \hat{\boldsymbol{\Omega}}_{aa} & \hat{\boldsymbol{\Omega}}_{ab} \\ \hat{\boldsymbol{\Omega}}_{ba} & \hat{\boldsymbol{\Omega}}_{bb} \end{pmatrix} \left\{ \begin{pmatrix} \boldsymbol{\theta}^r_a \\ \boldsymbol{\theta}^r_b \end{pmatrix} - \begin{pmatrix} \hat{\boldsymbol{\theta}}^r_a \\ \hat{\boldsymbol{\theta}}^r_b \end{pmatrix} \right\} \\ & + n \sum_{j=1}^{f_0} \rho_j |\beta_j| + n \sum_{m=2}^{r_0} \tau_m \left\{ \hat{\boldsymbol{\gamma}}_{\tau_m} \right\} + n \sum_{j=f_0+1}^{p_f} \rho_j |\beta_j| + n \sum_{m=r_0+1}^{p_r} \tau_m \left\{ \hat{\boldsymbol{\gamma}}_{\tau_m} \right\} \end{aligned}$$

Taking partial derivative of $Q(\boldsymbol{\theta}^r)$ and evaluating at the global minimizers, by definition, we have

$$\left. \frac{\partial Q(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^{rT}_a} \right|_{\boldsymbol{\theta} = \begin{pmatrix} \hat{\boldsymbol{\theta}}^r_{\rho\tau,a} \\ \mathbf{0} \end{pmatrix}} = 2n\hat{\boldsymbol{\Omega}}_{aa}(\hat{\boldsymbol{\theta}}^r_{\rho\tau,a} - \hat{\boldsymbol{\theta}}^r_a) + 2n\hat{\boldsymbol{\Omega}}_{ab}(\mathbf{0} - \hat{\boldsymbol{\theta}}^r_b) + n\boldsymbol{D}(\hat{\boldsymbol{\theta}}^r_{\rho\tau,a}) = 0, \tag{12}$$

where $D(\hat{\boldsymbol{\theta}}^r_{\rho\tau,a}) = (\rho_1 sgn(\hat{\beta}_{\rho\tau,1}), \rho_2 sgn(\hat{\beta}_{\rho\tau,2}), \ldots, \rho_{f_0} sgn(\hat{\beta}_{\rho\tau,f_0}), 0, \tau_2 \frac{\hat{\boldsymbol{\gamma}}^T_{\rho\tau,2}}{\|\hat{\boldsymbol{\gamma}}_{\rho\tau,2}\|}, \ldots, \tau_{r_0} \frac{\hat{\boldsymbol{\gamma}}^T_{\rho\tau,r_0}}{\|\hat{\boldsymbol{\gamma}}_{\rho\tau,r_0}\|}, 0)^T$. Reorganize (12), we have $\hat{\boldsymbol{\theta}}^r_{\rho\tau,a} = \hat{\boldsymbol{\theta}}^r_a + (\boldsymbol{\Omega}_{aa})^{-1} \boldsymbol{\Omega}_{ab} \hat{\boldsymbol{\theta}}^r_b - 1/2 \, (\boldsymbol{\Omega}_{aa})^{-1} \boldsymbol{D}(\hat{\boldsymbol{\theta}}^r_{\rho\tau,a})$, which leads to

$$n^{1/2}(\hat{\boldsymbol{\theta}}^r_{\rho\tau,a} - \boldsymbol{\theta}^r_{0a}) = n^{1/2}(\hat{\boldsymbol{\theta}}^r_a - \boldsymbol{\theta}^r_{0a}) + (\boldsymbol{\Omega}_{aa})^{-1} \boldsymbol{\Omega}_{ab} \hat{\boldsymbol{\theta}}^r_b - 1/2 \, (\boldsymbol{\Omega}_{aa})^{-1} \boldsymbol{D}(\hat{\boldsymbol{\theta}}^r_{\rho\tau,a}). \tag{13}$$

According to the condition $n^{1/2} a_{f,n} \xrightarrow{P} 0$ and $n^{1/2} a_{r,n} \xrightarrow{P} 0$, we have $n^{1/2}\rho_j \leqslant n^{1/2} a_{f,n} \xrightarrow{P} 0$ and $n^{1/2}\tau_m \leqslant n^{1/2} a_{r,n} \xrightarrow{P} 0$. Thus the third term in (13) is $o_p(1)$. Then, we can rewrite (13) as

$$n^{1/2}(\hat{\boldsymbol{\theta}}^r_{\rho\tau,a} - \boldsymbol{\theta}^r_{0a}) = \left\{1, (\boldsymbol{\Omega}_{aa})^{-1} \boldsymbol{\Omega}_{ab}\right\} \cdot n^{1/2} \begin{pmatrix} \hat{\boldsymbol{\theta}}^r_a - \boldsymbol{\theta}^r_{0a} \\ \hat{\boldsymbol{\theta}}^r_b - \mathbf{0} \end{pmatrix} + o_p(1). \tag{14}$$

Given that

$$n^{1/2} \begin{pmatrix} \hat{\boldsymbol{\theta}}^r_a - \boldsymbol{\theta}^r_{0a} \\ \hat{\boldsymbol{\theta}}^r_b - \mathbf{0} \end{pmatrix} \xrightarrow{D} \mathcal{N}\left( \mathbf{0}, \begin{pmatrix} \boldsymbol{\Sigma}^r_{aa} & \boldsymbol{\Sigma}^r_{ab} \\ \boldsymbol{\Sigma}^r_{ba} & \boldsymbol{\Sigma}^r_{bb} \end{pmatrix} \right),$$

and that $\boldsymbol{\Omega}_{aa} \xrightarrow{P} \boldsymbol{\Omega}_{aa}$, $\boldsymbol{\Omega}_{ab} \xrightarrow{P} \boldsymbol{\Omega}_{ab}$, (14) can be derived into

$$n^{1/2}(\hat{\boldsymbol{\theta}}_a^r - \boldsymbol{\theta}_{0a}^r) \xrightarrow{D} \mathcal{N}\left(\mathbf{0}, \left\{1, (\boldsymbol{\Omega}_{aa})^{-1}\boldsymbol{\Omega}_{ab}\right\}\begin{pmatrix}\boldsymbol{\Sigma}_{aa}^r & \boldsymbol{\Sigma}_{ab}^r \\ \boldsymbol{\Sigma}_{ba}^r & \boldsymbol{\Sigma}_{bb}^r\end{pmatrix}\left\{1, [\boldsymbol{\Omega}^r]_{aa}^{-1}[\boldsymbol{\Omega}^r]_{ab}^{-1}\right\}^T\right).$$

Providing the fact that

$$\boldsymbol{\Omega} = \begin{pmatrix}\boldsymbol{\Omega}_{aa} & \boldsymbol{\Omega}_{ab} \\ \boldsymbol{\Omega}_{ba} & \boldsymbol{\Omega}_{bb}\end{pmatrix} = \begin{pmatrix}\boldsymbol{A} & -\boldsymbol{A}\boldsymbol{\Sigma}_{ab}^r(\boldsymbol{\Sigma}_{bb}^r)^{-1} \\ -(\boldsymbol{\Sigma}_{bb}^r)^{-1}\boldsymbol{\Sigma}_{ba}^r\boldsymbol{A} & (\boldsymbol{\Sigma}_{bb}^r)^{-1} + (\boldsymbol{\Sigma}_{bb}^r)^{-1}\boldsymbol{\Sigma}_{ba}^r\boldsymbol{A}\boldsymbol{\Sigma}_{\boldsymbol{ab}}^{\boldsymbol{r}}(\boldsymbol{\Sigma}_{bb}^r)^{-1}\end{pmatrix},$$

where $\boldsymbol{A} = (\boldsymbol{\Sigma}_{aa}^r - \boldsymbol{\Sigma}_{ab}^r(\boldsymbol{\Sigma}_{bb}^r)^{-1}\boldsymbol{\Sigma}_{ba}^r)^{-1}$. It then follows that $\boldsymbol{\Omega}_{aa}^{-1}\boldsymbol{\Omega}_{ab} = -\boldsymbol{\Sigma}_{ab}^r(\boldsymbol{\Sigma}_{bb}^r)^{-1}$. Then the proof of the oracle property is completed by verifying that

$$\begin{aligned}
&\left\{1, (\boldsymbol{\Omega}_{aa})^{-1}\boldsymbol{\Omega}_{ab}\right\}\begin{pmatrix}\boldsymbol{\Sigma}_{aa}^r & \boldsymbol{\Sigma}_{ab}^r \\ \boldsymbol{\Sigma}_{ba}^r & \boldsymbol{\Sigma}_{bb}^r\end{pmatrix}\left\{1, (\boldsymbol{\Omega}_{aa})^{-1}\boldsymbol{\Omega}_{ab}\right\}^T \\
= \;& \{\boldsymbol{\Sigma}_{aa}^r + (\boldsymbol{\Omega}_{aa})^{-1}\boldsymbol{\Omega}_{ab}\boldsymbol{\Sigma}_{ba}^r, \boldsymbol{\Sigma}_{ab}^r + (\boldsymbol{\Omega}_{aa})^{-1}\boldsymbol{\Omega}_{ab}\boldsymbol{\Sigma}_{bb}^r\}\begin{Bmatrix}1 \\ (\boldsymbol{\Omega}_{aa})^{-1}\boldsymbol{\Omega}_{ab}\end{Bmatrix} \\
= \;& \boldsymbol{\Sigma}_{aa}^r - \boldsymbol{\Sigma}_{ab}^r(\boldsymbol{\Sigma}_{bb}^r)^{-1}\boldsymbol{\Sigma}_{ba}^r - \boldsymbol{\Sigma}_{ab}^r\boldsymbol{\Sigma}_{ab}^r(\boldsymbol{\Sigma}_{bb}^r)^{-1} + \boldsymbol{\Sigma}_{ab}^r\boldsymbol{\Sigma}_{ab}^r(\boldsymbol{\Sigma}_{bb}^r)^{-1} \\
= \;& \boldsymbol{\Sigma}_{aa}^r - \boldsymbol{\Sigma}_{ab}^r(\boldsymbol{\Sigma}_{bb}^r)^{-1}\boldsymbol{\Sigma}_{ba}^r \\
= \;& \left(\left[(\boldsymbol{\Sigma}^r)^{-1}\right]_{aa}\right)^{-1} = \left(\left[(\boldsymbol{\Sigma})^{-1}\right]_{(\beta_{0a},\gamma_{0a},\phi)}\right)^{-1}
\end{aligned}$$

**Acknowledgements**

The research of SL, SK, and SJ was partially supported by NIH/NCI CCSG Grant 3P30CA072720.

**Data Availability Statement**

No external data are used.

**References**

Kowalchuck, R. K., Keselman, H. J., Algina, J., and Wolfinger, R. D. (2004). The analysis of repeated measurements with mixed-model adjusted F tests. *E*ducational and Psychological Measurement **64,** 224-242.

Liang, H., Wu, H., and Zou, G. (2008). A note on conditional AIC for linear mixed effects-models. *B*iometrika **95**, 773-778.

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *P*hil. Trans. Roy. Soc. **53**, 370418; Reprinted in *B*iometrika, **45** (1958), 293-315.

Bondell, H. D., Krishna, A., and Ghosh, S. K. (2010), Joint Variable Selection for Fixed and Random Effects in Linear Mixed-Effects Models, *B*iometrics **66**, 1069-1077.

Claeskens, G. and Consentino, F. (2008). Variable selection with incomplete covariate data. *B*iometrics **64**, 1062-1096.

Cox, D. R. (2013). Discussion. *I*nternational Statistical Review **81**, 40-41.

Efron, B. (1993). Bayes and likelihood calculations from confidence intervals. *B*iometrika **80**, 3-26.

Efron, B. (1998). R.A. Fisher in the 21st century. *S*tat. Sci. **13**, 95-122.

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

Fisher, R.A. (1922). On the mathematical foundations of theoretical statistics. *P*hilos. Trans. R. Soc. Lond. A, **222**, 309-368.

Freedman DA (2006). On the So-Called "Huber Sandwich Estimator" and "Robust Standard Errors". *T*he American Statistician, **60**, 299302.

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

Garon, E., Rizvi, N., Hui, R., and et al. (2015). Pembrolizumab for the treatment of non-small-cell lung cancer. *N*ew England Journal of Medicine **372,** 2018–2028.

Groll, A., and Tutz, G. (2014). Variable Selection for Generalized Linear Mixed Models by ↕1Penalized Estimation. *S*tatistics and Computing **24**, 137-154.

Gurka, M. J. (2006). Selecting the best linear mixed model under REML. *A*merican

Statistician **60**, 19-26.

He, Z., Tu, W., Wang, S., Fu, H., and Yu, Z. (2015). Simultaneous variable selection for joint models of longitudinal and survival outcomes. *B*iometrics **71**, 178-187.

Hui, F.K.C., Mueller, S., and Welsh, A.H. (2017). Joint Selection in Mixed Models using Regularized PQL. *J*ournal of the American Statistical Association **112**, 1323-1333.

Ibrahim, J. G., Zhu, H., and Tang, N. (2008). Model selection criteria for missing-data problems using the EM algorithm. *J*ournal of the American Statistical Association **103**, 1648-1658.

Ibrahim, J. G., Zhu, H., Garcia, R. I., and Guo, R. (2011). Fixed and Random Effects Selection in Mixed Effects Models. *B*iometrics **67**, 495-503.

Keselman, H. J., Algina, J., Kowalchuk, R. K., and Wolfinger, R. D. (1998). A comparison of two approaches for selecting covariance structures in the analysis of repeated measurements. *C*ommunications in StatisticsSimulation and Computation **27**, 591-604.

Kowalchuck, R. K., Keselman, H. J., Algina, J., and Wolfinger, R. D. (2004). The analysis of repeated measurements with mixed-model adjusted F tests. *E*ducational and Psychological Measurement **64,** 224-242.

Lange, K. (1999). Numerical Analysis for Statisticians, *N*ew York: Springer-Verlag.

Lehmann, E. L. (1993). The Fisher, Neyman-Pearson Theories of Testing Hypotheses: One Theory or Two? *J*ournal of the American Statistical Association **88**, 1242-1249.

Liang, H., Wu, H., and Zou, G. (2008). A note on conditional AIC for linear mixed effects-models. *B*iometrika **95**, 773-778.

Lin, B., Pang, Z., and Jiang, J. (2013) Fixed and Random Effects Selection by REML and Pathwise Coordinate Optimization. *J*ournal of Computational and Graphical Statistics **22**, 341-355.

Lindstrom, M. J., and Bates, D. M. (1988). Newton-Raphson and EM Algorithms for

Linear Mixed-Effects Models for Repeated-Measurements Data. *J*ournal of the American Statistical Association **83**, 1014-1022.

Lindstrom, M. J., and Bates, D. M. (1990). Nonlinear Mixed-Effects Models for Repeated-Measures Data. *B*iometrics **46**, 673-687.

Littell, R. C, Milliken, G. A., Stroup, W W, and Wolfinger, R. D. (1996). SAS System for Mixed Models. Cary, NC: SAS Institute, Inc.

Liu, D., Liu, R. Y., and Xie, M. (2015). Multivariate meta-analysis of heterogeneous studies using only summary statistics: Efficiency and robustness. *J*ournal of the American Statistical Association **110**, 326-340.

Pan, J., and Huang, C. (2014). Random Effects Selection in Generalized Linear Mixed Models via Shrinkage Penalty Function. *S*tatistics and Computing **24** 725-738.

Pan, Z., and Lin, D. Y. (2005). Goodness-of-fit methods for generalized linear mixed models. *B*iometrics **61**, 1000-1009.

Peng, H., and Lu, Y. (2012). Model Selection in Linear Mixed Effect Models. *J*ournal of Multivariate Analysis **109**, 109-129.

Pinheiro J C, Bates D M. (1995). Approximations to the log-likelihood function in the nonlinear mixed-effects model. *J*ournal of computational and Graphical Statistics, **4**, 12–35.

Pinheiro, J. C., and Chao, E. C. (2006). Efficient Laplacian and adaptive Gaussian quadrature algorithms for multilevel generalized linear mixed models. *J*ournal of Computational and Graphical Statistics **15**, 58-81.

Schelldorfer, Meier, and Bühlmann (2014) Glmmlasso: An Algorithm for High-Dimensional Generalized Linear Mixed Models Using L1-Penalization. *J*ournal of Computational and Graphical Statistics **23**, 460-477.

Singh, K., Xie, M., and Strawderman, W. E. (2007). Confidence distribution (CD) distribu-

tion estimator of a parameter. In Complex datasets and inverse problems (pp. 132-150). Institute of Mathematical Statistics.

Tian, L., Wang, R., Cai, T., and Wei, L.-J. (2011). The highest confi

dence density region and its usage for joint inferences about constrained parameters. *B*iometrics **67**, 604-610.

Vaida, F., and Blanchard, S. (2005). Conditional Akaike Information for Mixed-Effects Models. *B*iometrika **92**, 351-370.

Vonesh, E. F., Wang, H., Nie, L., and Majumdar, D. (2002). Conditional Second-Order Generalized Estimating Equations for Generalized Linear and Nonlinear Mixed-Effects Models. *J*ournal of the American Statistical Association **97**, 271-283.

Wang, W., Lu, S.-E., Cheng, J. Q., Xie, M., and Kostis, J. B. (2021). Multivariate survival analysis in big data: A divide-and-combine approach. *B*iometrics 2021 Apr 13. doi: 10.1111/biom.13469.

Wang, H., and Leng, C. (2007). Unified LASSO Estimation by Least Squares Approximation. *J*ournal of the American Statistical Association **102**, 1039–1048.

Wang, H., and Leng, C. (2008). A note on adaptive group lasso. *C*omputational Statistics and Data Analysis, **52**, 5277–5286.

Wang, T., and Merkle, E. C. (2018). merDeriv: Derivative Computations for Linear Mixed Effects Models with Application to Robust Standard Errors. *J*ournal of Statistical Software, Code Snippets, **87**, 1-16.

Westfall, P. H. (1997). Multiple Testing of General Contrasts Using Logical Constraints and Correlations. *J*ournal of the American Statistical Association, **92**, 299306.

Wolfinger, R. (1993). Laplace's approximation for nonlinear mixed models. *B*iometrika **80**, 791–795.

Xie, M., Singh, K., and Strawderman, W. E. (2011). Confidence distributions and a unifying

framework for meta-analysis. Journal of the American Statistical Association **106**, 320-333.

Xie, M. and Singh, K. (2013). Confidence distribution, the frequentist distribution estimator of a parameter: A review. *I*nternational Statistical Review **81**, 3-39.

Yang, Y. and Zou, H. (2015). A Fast Unified Algorithm for Computing Group-Lasso Penalized Learning Problems, *S*tatistics and Computing **25**, 1129-1141

Zhang, H. H. and Lu, W. (2007). Adaptive Lasso for Cox's proportional hazards model. *B*iometrika **94**, 691–703.

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

Table 1: Linear Mixed Model: Performance of the proposed regularized estimators

| Fixed Effects | | Method | $\boldsymbol{\beta}_0$ | 1 | 1 | 1 | 1 | 1 | 1 | $\mathbf{0}_{10}$ | Time |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $(n, m)$ | (60,10) | CD | $\hat{\boldsymbol{\beta}}^s_\rho$ | 0.853 | 0.880 | 0.955 | 0.990 | 0.992 | 0.992 | 0.000 | 0.001±0.000 |
| | | | $ESE$ | 0.492 | 0.331 | 0.149 | 0.046 | 0.044 | 0.046 | 0.013 | |
| | | | CovP(%) | 87.8 | 85.7 | 89.3 | 91.8 | 92.1 | 91.5 | - | |
| | | | % Sel | 100.0 | 95.3 | 100.0 | 100.0 | 100.0 | 100.0 | 3.4 | |
| | | rPQL | $\hat{\boldsymbol{\beta}}^{rPQL}$ | 0.867 | 0.717 | 0.812 | 0.991 | 0.989 | 0.990 | 0.000 | 0.219±0.035 |
| | | | $ESE$ | 0.458 | 0.527 | 0.352 | 0.046 | 0.046 | 0.044 | 0.013 | |
| | | | % Sel | 100.0 | 82.1 | 96.2 | 100.0 | 100.0 | 100.0 | 5.2 | |
| | (120,6) | CD | $\hat{\boldsymbol{\beta}}^s_\rho$ | 0.926 | 0.933 | 0.966 | 0.993 | 0.992 | 0.994 | 0.000 | 0.001±0.000 |
| | | | $ESE$ | 0.320 | 0.216 | 0.110 | 0.046 | 0.045 | 0.045 | 0.012 | |
| | | | CovP(%) | 90.7 | 90.3 | 90.6 | 92.5 | 93.2 | 92.2 | - | |
| | | | % Sel | 100.0 | 99.7 | 100.0 | 100.0 | 100.0 | 100.0 | 2.5 | |
| | | rPQL | $\hat{\boldsymbol{\beta}}^{rPQL}$ | 0.983 | 0.971 | 0.978 | 0.997 | 1.000 | 0.999 | 0.000 | 0.931±0.107 |
| | | | $ESE$ | 0.302 | 0.277 | 0.181 | 0.045 | 0.043 | 0.045 | 0.016 | |
| | | | % Sel | 100.0 | 96.1 | 99.30 | 100.0 | 100.0 | 100.0 | 11.4 | |
| | (500,6) | CD | $\hat{\boldsymbol{\beta}}^s_\rho$ | 0.970 | 0.975 | 0.989 | 0.999 | 0.996 | 0.997 | 0.000 | $0.006 \pm 0.001$ |
| | | | $ESE$ | 0.149 | 0.101 | 0.051 | 0.022 | 0.023 | 0.022 | 0.004 | |
| | | | CovP(%) | 91.9 | 91.6 | 93.3 | 94.1 | 94.0 | 93.8 | - | |
| | | | % Sel | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 0.8 | |
| | | rPQL | $\hat{\boldsymbol{\beta}}^{rPQL}$ | 0.995 | 0.995 | 0.998 | 1.000 | 1.000 | 0.999 | 0.000 | 81.162±12.143 |
| | | | $ESE$ | 0.144 | 0.095 | 0.049 | 0.021 | 0.021 | 0.021 | 0.004 | |
| e | | | % Sel | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 3.5 | |
| Random Effects | | | $\boldsymbol{\gamma}_0$ | 3 | 1.6 | 0.2 | 1.2 | 0.57 | 0.8 | $\mathbf{0}_4$ | |
| | (60,10) | CD | $\hat{\boldsymbol{\gamma}}^s_\rho$ | 2.962 | 1.562 | 0.183 | 1.170 | 0.544 | 0.772 | 0.001 | |
| | | | $ESE$ | 0.318 | 0.236 | 0.136 | 0.131 | 0.128 | 0.096 | 0.015 | |
| | | | CovP(%) | 88.8 | 90.7 | 91.6 | 90.9 | 92.1 | 89.1 | - | |
| | | | % Sel | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 3.0 | |
| | | rPQL | $\hat{\boldsymbol{\gamma}}^{rPQL}$ | 2.996 | 1.620 | 0.215 | 1.164 | 0.642 | 0.558 | 0.012 | |
| | | | $ESE$ | 0.280 | 0.235 | 0.149 | 0.330 | 0.226 | 0.114 | 0.062 | |
| | | | % Sel | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 19.5 | |
| | (120,6) | CD | $\hat{\boldsymbol{\gamma}}^s_\rho$ | 3.011 | 1.592 | 0.192 | 1.200 | 0.559 | 0.791 | 0.001 | |
| | | | $ESE$ | 0.243 | 0.174 | 0.103 | 0.111 | 0.107 | 0.086 | 0.030 | |
| | | | CovP(%) | 88.1 | 90.2 | 92.7 | 88.4 | 91.9 | 87.9 | - | |
| | | | % Sel | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 2.8 | |
| | | rPQL | $\hat{\boldsymbol{\gamma}}^{rPQL}$ | 3.000 | 1.580 | 0.199 | 1.007 | 0.682 | 0.799 | 0.015 | |
| | | | $ESE$ | 0.195 | 0.161 | 0.108 | 0.502 | 0.289 | 0.152 | 0.062 | |
| | | | % Selection | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 19.5 | |
| | (500,6) | CD | $\hat{\boldsymbol{\gamma}}^s_\rho$ | 3.007 | 1.608 | 0.203 | 1.203 | 0.567 | 0.794 | 0.001 | |
| | | | $ESE$ | 0.119 | 0.083 | 0.049 | 0.052 | 0.051 | 0.041 | 0.008 | |
| | | | CovP(%) | 88.2 | 93.4 | 92.1 | 94.3 | 94.2 | 91.0 | - | |
| | | | % Sel | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 1.4 | |
| | | rPQL | $\hat{\boldsymbol{\gamma}}^{rPQL}$ | 3.019 | 1.592 | 0.200 | 1.098 | 0.635 | 0.828 | 0.031 | |
| | | | $ESE$ | 0.097 | 0.077 | 0.049 | 0.410 | 0.237 | 0.108 | 0.084 | |
| | | | % Sel | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 0.121 | |

Table 2: Random Effects Logistic Model: Performance of the proposed regularized estimators

| Fixed Effects | | Method | $\boldsymbol{\beta}_0$ | 1 | 1 | 1 | 1 | 1 | 1 | $\mathbf{0}_{10}$ | Time |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $(n,m)$ | (200,10) | CD-J | $\hat{\boldsymbol{\beta}}_{\rho\tau}$ | 0.880 | 0.858 | 0.873 | 0.893 | 0.896 | 0.902 | 0.000 | 1.08±1.38 |
| | | | $ESE$ | 0.155 | 0.139 | 0.119 | 0.091 | 0.090 | 0.089 | 0.024 | |
| | | | CovP | 85.1 | 76.0 | 72.4 | 68.6 | 65.6 | 67.5 | - | |
| | | | % Sel | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 5.6 | |
| | | CD-S | $\hat{\boldsymbol{\beta}}_{\rho}^{s}$ | 0.937 | 0.918 | 0.928 | 0.944 | 0.939 | 0.931 | 0.000 | 0.00±0.00 |
| | | | $ESE$ | 0.158 | 0.139 | 0.117 | 0.089 | 0.089 | 0.088 | 0.020 | |
| | | | CovP | 92.3 | 89.2 | 88.9 | 87.6 | 86.7 | 88.0 | - | |
| | | | % Sel | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 2.4 | |
| | | rPQL | $\hat{\boldsymbol{\beta}}^{rPQL}$ | 0.646 | 0.641 | 0.659 | 0.705 | 0.705 | 0.706 | 0.000 | 2.95±0.39 |
| | | | $ESE$ | 0.103 | 0.103 | 0.103 | 0.058 | 0.057 | 0.056 | 0.039 | |
| | | | % Sel | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 40.7 | |
| | (500,10) | CD-J | $\hat{\boldsymbol{\beta}}_{\rho\tau}$ | 0.939 | 0.929 | 0.931 | 0.946 | 0.942 | 0.943 | 0.000 | 1.06±1.38 |
| | | | $ESE$ | 0.100 | 0.091 | 0.076 | 0.058 | 0.057 | 0.057 | 0.012 | |
| | | | CovP | 87.4 | 80.8 | 75.1 | 72.2 | 70.0 | 72.0 | - | |
| | | | % Sel | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 2.9 | |
| | | CD-S | $\hat{\boldsymbol{\beta}}_{\rho}^{s}$ | 0.971 | 0.964 | 0.963 | 0.971 | 0.965 | 0.965 | 0.000 | 0.00±0.00 |
| | | | $ESE$ | 0.102 | 0.091 | 0.076 | 0.058 | 0.055 | 0.057 | 0.009 | |
| | | | CovP | 94.0 | 91.3 | 90.3 | 90.7 | 90.1 | 89.9 | - | |
| | | | % Sel | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 1.2 | |
| | | rPQL | $\hat{\boldsymbol{\beta}}^{rPQL}$ | 0.647 | 0.646 | 0.658 | 0.702 | 0.701 | 0.706 | 0.071 | 63.00±10.69 |
| | | | $ESE$ | 0.066 | 0.055 | 0.046 | 0.036 | 0.034 | 0.035 | 0.027 | |
| | | | % Sel | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 54.0 | |
| Random effects | | | $\boldsymbol{\gamma}_0$ | 1.73 | 0.69 | 1.23 | 0.46 | 0.15 | 0.87 | $\mathbf{0}_4$ | |
| $(n,m)$ | (200,10) | CD-J | $\hat{\boldsymbol{\gamma}}_{\rho\tau}$ | 1.472 | 0.583 | 0.999 | 0.377 | 0.111 | 0.639 | 0.000 | |
| | | | $ESE$ | 0.173 | 0.151 | 0.148 | 0.124 | 0.127 | 0.150 | 0.000 | |
| | | | CovP | 52.8 | 90.5 | 51.0 | 90.4 | 95.1 | 46.6 | 0.0 | |
| | | | % Sel | 100.0 | 100.0 | 100.0 | 99.9 | 99.9 | 99.9 | 0.0 | |
| | | CD-S | $\hat{\boldsymbol{\gamma}}_{\tau}^{s}$ | 1.666 | 0.670 | 1.147 | 0.445 | 0.136 | 0.772 | 0.000 | |
| | | | $ESE$ | 0.191 | 0.170 | 0.161 | 0.143 | 0.144 | 0.159 | 0.009 | |
| | | | CovP | 90.3 | 94.8 | 86.5 | 93.1 | 94.1 | 84.5 | - | |
| | | | % Sel | 100.0 | 100.0 | 100.0 | 99.9 | 99.9 | 99.9 | 0.2 | |
| | | rPQL | $\hat{\gamma}^{rPQL}$ | 1.077 | 0.411 | 0.271 | 0.203 | 0.528 | 0.188 | 0.093 | |
| | | | $ESE$ | 0.102 | 0.108 | 0.091 | 0.319 | 0.307 | 0.307 | 0.105 | |
| | | | % Sel | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 69.7 | |
| | (500,10) | CD-J | $\hat{\boldsymbol{\gamma}}_{\rho\tau}$ | 1.562 | 0.626 | 1.094 | 0.409 | 0.123 | 0.731 | 0.000 | |
| | | | $ESE$ | 0.112 | 0.100 | 0.097 | 0.081 | 0.084 | 0.086 | 0.000 | |
| | | | CovP | 49.6 | 88.2 | 53.0 | 91.4 | 95.2 | 49.7 | - | |
| | | | % Sel | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 0.0 | |
| | | CD-S | $\hat{\boldsymbol{\gamma}}_{\tau}^{s}$ | 1.657 | 0.670 | 1.168 | 0.414 | 0.122 | 0.739 | 0.000 | |
| | | | $ESE$ | 0.117 | 0.106 | 0.102 | 0.086 | 0.089 | 0.089 | 0.000 | |
| | | | CovP | 88.7 | 93.7 | 84.8 | 94.2 | 94.3 | 82.3 | - | |
| | | | % Sel | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 0.0 | |
| | | rPQL | $\hat{\gamma}^{rPQL}$ | 1.074 | 0.414 | 0.274 | 0.191 | 0.544 | 0.186 | 0.102 | |
| | | | $ESE$ | 0.060 | 0.068 | 0.036 | 0.313 | 0.306 | 0.159 | 0.092 | |
| | | | % Sel | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 72.3 | |

Table 3: Random Effects Poisson Regression Model: Performance of the proposed regularized estimators

| Fixed Effects | | Method | $\boldsymbol{\beta}_0$ | $-1$ | 1 | 1 | 1 | 1 | 1 | $\mathbf{0}_{10}$ | Time |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $(n,m)$ | (200,10) | CD-J | $\hat{\boldsymbol{\beta}}_{\rho\tau}$ | -1.028 | 0.954 | 0.974 | 1.000 | 1.000 | 1.000 | 0.000 | 0.00±0.01 |
| | | | $ESE$ | 0.133 | 0.118 | 0.085 | 0.006 | 0.005 | 0.006 | 0.002 | |
| | | | CovP | 87.5 | 83.4 | 84.3 | 87.5 | 89.5 | 88.6 | - | |
| | | | % Sel | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 8.2 | |
| | | CD-S | $\hat{\boldsymbol{\beta}}^s_{\rho}$ | -0.996 | 0.997 | 0.999 | 1.000 | 1.000 | 1.000 | 0.000 | 0.00±0.03 |
| | | | $ESE$ | 0.132 | 0.110 | 0.082 | 0.007 | 0.006 | 0.006 | 0.002 | |
| | | | CovP | 90.9 | 90.2 | 88.8 | 88.8 | 92.3 | 93.7 | - | |
| | | | % Sel | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 6.4 | |
| | | rPQL | $\hat{\boldsymbol{\beta}}^{rPQL}$ | -0.977 | 0.998 | 0.993 | 1.000 | 0.999 | 0.999 | 0.000 | 2.84±0.49 |
| | | | $ESE$ | 0.156 | 0.112 | 0.078 | 0.011 | 0.009 | 0.008 | 0.002 | |
| | | | % Sel | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 12.9 | |
| | (400,10) | CD-J | $\hat{\boldsymbol{\beta}}_{\rho\tau}$ | -1.008 | 0.976 | 0.978 | 1.000 | 1.000 | 1.000 | 0.000 | 0.01±0.08 |
| | | | $ESE$ | 0.095 | 0.082 | 0.056 | 0.003 | 0.003 | 0.003 | 0.001 | |
| | | | CovP | 86.4 | 83.5 | 84.6 | 87.5 | 87.0 | 88.6 | - | |
| | | | % Sel | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 7.8 | |
| | | CD-S | $\hat{\boldsymbol{\beta}}^s_{\rho}$ | -0.995 | 0.998 | 0.991 | 1.000 | 1.000 | 1.000 | 0.000 | 0.00±0.01 |
| | | | $ESE$ | 0.097 | 0.081 | 0.051 | 0.005 | 0.003 | 0.003 | 0.001 | |
| | | | CovP | 86.2 | 87.5 | 88.3 | 87.8 | 87.0 | 89.1 | - | |
| | | | % Sel | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 7.0 | |
| | | rPQL | $\hat{\boldsymbol{\beta}}^{rPQL}$ | -0.990 | 0.999 | 0.094 | 0.999 | 0.999 | 0.999 | 0.000 | 27.84±4.67 |
| | | | $ESE$ | 0.094 | 0.074 | 0.053 | 0.006 | 0.004 | 0.004 | 0.002 | |
| | | | % Sel | 100.0 | 99.8 | 100.0 | 100.0 | 100.0 | 100.0 | 16.5 | |
| | | MLE | $\hat{\boldsymbol{\beta}}$ | -0.997 | 0.997 | 0.996 | 1.000 | 1.000 | 1.000 | - | |
| | | | CovP | 85.2 | 85.2 | 88.7 | 90.4 | 90.0 | 89.9 | - | |
| Random effects | | | $\boldsymbol{\gamma}_0$ | 1.73 | 0.69 | 1.23 | 0.46 | 0.15 | 0.87 | $\mathbf{0}_4$ | |
| $(n,m)$ | (200,10) | CD-J | $\hat{\boldsymbol{\gamma}}_{\rho\tau}$ | 1.669 | 0.666 | 1.200 | 0.447 | 0.134 | 0.849 | 0.000 | |
| | | | $ESE$ | 0.091 | 0.093 | 0.069 | 0.080 | 0.070 | 0.057 | 0.000 | |
| | | | CovP | 84.6 | 91.6 | 86.0 | 89.5 | 91.6 | 84.6 | - | |
| | | | % Sel | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 0.1 | |
| | | CD-S | $\hat{\boldsymbol{\gamma}}^s_{\tau}$ | 1.701 | 0.689 | 1.215 | 0.465 | 0.139 | 0.862 | 0.000 | |
| | | | $ESE$ | 0.091 | 0.098 | 0.069 | 0.083 | 0.072 | 0.057 | 0.001 | |
| | | | CovP | 90.9 | 90.9 | 87.4 | 90.9 | 90.9 | 86.0 | - | |
| | | | % Sel | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 1.4 | |
| | | rPQL | $\hat{\gamma}^{rPQL}$ | 1.591 | 0.643 | 0.429 | 1.195 | 0.156 | 0.841 | 0.000 | |
| | | | $ESE$ | 0.092 | 0.106 | 0.076 | 0.101 | 0.095 | 0.067 | 0.015 | |
| | | | % Sel | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 0.4 | |
| | (400,10) | CD-J | $\hat{\boldsymbol{\gamma}}_{\rho\tau}$ | -1.008 | 0.976 | 0.978 | 1.000 | 1.000 | 1.000 | 0.000 | |
| | | | $ESE$ | 0.095 | 0.082 | 0.056 | 0.003 | 0.003 | 0.003 | 0.001 | |
| | | | CovP | 82.2 | 87.2 | 86.2 | 87.8 | 88.3 | 88.6 | - | |
| | | | % Sel | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 2.4 | |
| | | CD-S | $\hat{\boldsymbol{\gamma}}^s_{\tau}$ | 1.716 | 0.688 | 1.229 | 0.466 | 0.149 | 0.868 | 0.000 | |
| | | | $ESE$ | 0.075 | 0.073 | 0.050 | 0.057 | 0.051 | 0.038 | 0.001 | |
| | | | CovP | 84.0 | 88.6 | 88.0 | 87.8 | 87.5 | 88.0 | - | |
| | | | % Sel | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 2.4 | |
| | | rPQL | $\hat{\gamma}^{rPQL}$ | 1.596 | 0.642 | 0.426 | 1.210 | 0.157 | 0.854 | 0.000 | |
| | | | $ESE$ | 0.066 | 0.074 | 0.054 | 0.046 | 0.047 | 0.039 | 0.000 | |
| | | | % Sel | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 0.0 | |
| | | MLE | $\hat{\boldsymbol{\gamma}}$ | 1.716 | 0.689 | 1.229 | 0.466 | 0.149 | 0.868 | - | |
| | | | CovP | 87.3 | 88.6 | 88.0 | 87.8 | 87.5 | 88.0 | - | |

Table 4: Linear Mixed Model Analysis of Lung Cancer Data

| Variable Selection using CD approach | | | Maximum Likelihood Estimates after Refit | | |
|---|---|---|---|---|---|
| | Fixed Effects | | | Fixed Effects | |
| Intercept | 3.875 | (3.554, 4.197) | Intercept | 3.845 | (3.490, 4.199) |
| Weeks | -0.117 | (-0.129, -0.105) | Weeks | -0.114 | (-0.126, -0.101) |
| Age | 0 | | | | |
| Gender | 0.434 | (0.016, 0.852) | Gender | 0.548 | (0.074, 0.852) |
| MLDGy | 0.912 | (0.299, 1.526) | MLDGy | 0.800 | (0.095, 1.506) |
| LungV20 | -0.789 | (-1.399, -0.180) | LungV20 | -0.624 | (-1.328,-0.079) |
| Smoking | 0 | | | | |
| Noise 1 | 0 | | | | |
| Noise 2 | 0 | | | | |
| | 0 | | | | |
| | Random Effects | | | Random Effects | |
| Intercept ($\hat{\Gamma}_{11}$) | 1.233 | (1.231, 1.235) | Intercept ($\hat{\Gamma}_{11}$) | 1.259 | (1.258, 1.260) |
| Weeks ($\hat{\Gamma}_{21}$) | 0.002 | (0.001, 0.002) | Weeks ($\hat{\Gamma}_{21}$) | 0.001 | (0.001, 0.001) |
| Weeks ($\hat{\Gamma}_{22}$) | 0.062 | (0.061, 0.063) | Weeks ($\hat{\Gamma}_{22}$) | 0.062 | (0.062, 0.063) |
| Age ($\hat{\Gamma}_{31}, \hat{\Gamma}_{32}, \hat{\Gamma}_{33}$) | 0 | | | | |
| MLDGy ($\hat{\Gamma}_{41}, \hat{\Gamma}_{42}, \hat{\Gamma}_{43}, \hat{\Gamma}_{44}$) | 0 | | | | |
| Residual ($\hat{\sigma}^2$) | 0.009 | | | Residual ($\hat{\sigma}^2$) | 0.009 |

Table 5: Random Effects Logistic Regression Model Analysis of Calcification Data

| Variable Selection using CD approach | | | Maximum Likelihood Estimates after Refit | | |
|---|---|---|---|---|---|
| | Fixed Effects | | | Fixed Effects | |
| Intercept | 4.298 | (2.449, 6.148) | Intercept | 4.451 | (2.610, 6.292) |
| Age | 3.945 | (1.579, 6.312) | Age | 4.058 | (1.670, 6.445) |
| Years from Radiation Therapy | 1.411 | (1.048, 1.773) | Years from Radation Therapy | 1.468 | (1.107, 1.828) |
| African Americans | -5.878 | (-10.503, -1.253) | African Americans | -6.188 | (-10.767, -1.609) |
| Her2/neu positive | -1.617 | (-2.899, -0.335) | Her2/neu positive | -1.738 | (-3.007, -0.470) |
| Adjuvant Therapy | 0 | | | | |
| Smoking | 0 | | | | |
| Bilateral | 0 | | | | |
| Noise 1 | 0 | | | | |
| Noise 2 | 0 | | | | |
| | Random Effects | | | Random Effects | |
| Intercept ($\hat{\Gamma}_{11}$) | 4.372 | (2.228, 6.518) | Intercept ($\hat{\Gamma}_{11}$) | 4.358 | (1.255 5.390) |
| Age ($\hat{\Gamma}_{21}$) | 7.292 | (0.175,14.410) | Age ($\hat{\Gamma}_{21}$) | 6.825 | (-3.554, 9.501) |
| Age ($\hat{\Gamma}_{22}$) | 6.318 | (2.912, 9.725) | Age ($\hat{\Gamma}_{22}$) | 6.310 | (2.074, 8.679) |
| Noise 1 ($\hat{\Gamma}_{31}$,$\hat{\Gamma}_{32}$, $\hat{\Gamma}_{33}$) | 0 | | | | |