Page Proof Instructions and Queries

Journal Title: Statistical Methods in Medical Research
Article Number: 1221201

Thank you for choosing to publish with us. This is your final opportunity to ensure your article will be accurate at publication. Please review your proof carefully and respond to the queries using the circled tools in the image below, which are available in Adobe Reader DC* by clicking **Tools** from the top menu, then clicking **Comment**.

Please use *only* the tools circled in the image, as edits via other tools/methods can be lost during file conversion. For comments, questions, or formatting requests, please use T. Please do *not* use comment bubbles/sticky notes.

*If you do not see these tools, please ensure you have opened this file with **Adobe Reader DC**, available for free at get.adobe.com/reader or by going to Help > Check for Updates within other versions of Reader. For more detailed instructions, please see us.sagepub.com/ReaderXProofs.

No.	Query
GQ1	Please confirm that all author information, including names, affiliations, sequence, and contact details, is correct.
GQ2	Please review the entire document for typographical errors, mathematical errors, and any other necessary corrections; check headings, tables, and figures.
GQ3	Please confirm that the Funding and Conflict of Interest statements are accurate.
GQ4	Please ensure that you have obtained and enclosed all necessary permissions for the reproduction of artistic works, (e.g. illustrations, photographs, charts, maps, other visual material, etc.) not owned by yourself. Please refer to your publishing agreement for further information.
GQ5	Please note that this proof represents your final opportunity to review your article prior to publication, so please do send all of your changes now.
GQ6	Please note, only ORCID iDs validated prior to acceptance will be authorized for publication; we are unable to add or amend ORCID iDs at this stage.

Original Research Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
1–21

DOI: 10.1177/09622802231221201
journals.sagepub.com/home/smm

Sage

Fixed and random effect selections in generalized linear mixed models

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

GQ1

Abstract

GQ2 GQ4 GQ5

Generalized linear mixed models are commonly used to describe relationships between correlated responses and covariates in medical research. In this paper, we propose a simple and easily implementable regularized estimation approach to select both fixed and random effects in generalized linear mixed model. Specifically, we propose to construct and optimize the objective functions using the confidence distributions of model parameters, as opposed to using the observed data likelihood functions, to perform effect selections. Two estimation methods are developed. The first one is to use the joint confidence distribution of model parameters to perform simultaneous fixed and random effect selections. The second method is to use the marginal confidence distributions of model parameters to perform the selections of fixed and random effects separately. 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 patient health outcomes after cancer therapies.

Keywords

Confidence distribution, generalized linear mixed model, variable selection, regularization, adaptive Lasso

1 Introduction

Generalized linear mixed models (GLMMs) are a commonly used class of models to describe the relationship between correlated responses and covariates in biomedical research. Researchers often want to determine the fixed effects and (or) the random effects of covariates for the outcome variables from a pool of covariates using the variable selection approaches. Our study is motivated by two longitudinal cancer studies. The first study longitudinally measures tumor size in lung cancer patients during the cycles of radiation therapy. The second study followed up with breast cancer patients for the incidence of common mammographic sequelae after they received breast-conserving surgery and radiation therapy. To account for the intra-patient correlation among repeatedly measured outcomes, GLMM analyses have been employed for both studies. To gain insight about patients' prognostics and health management, both studies aim to identify important demographic and clinical covariates that may predict the outcomes as fixed effects. Selections of random effects are also considered to evaluate heterogeneous effects.

[1]Rutgers School of Public Health, Piscataway, NJ, USA
[2]Rutgers Cancer Institute of New Jersey, New Brunswick, NJ, USA
[3]Bristol Myers Squibb, Berkeley Height, NJ, USA
[4]New York Institute of Technology, New York, NY, USA
[5]Deloitte, Parsippany, NJ, USA
[6]George Washington University Hospital, Washington, DC, USA

Corresponding author:
Shou-En Lu, Rutgers School of Public Health, Piscataway, NJ, USA.
Email: shouen.lu@rutgers.edu

In the statistical literature, several variable selection approaches for the GLMM have been proposed. For instance, the selection using the information criterion,[1–5] e.g. Akaike information criterion or Bayesian information criterion (BIC), etc, has been commonly used to determine the final model among a number of candidate models. For the popular regularized estimation approach, some methods are proposed to select both fixed and random effects for the linear mixed models (LMMs),[6–9] and some for the GLMM.[10] Some approaches emphasize on the selection of fixed effects only,[11,12] and some focus on the selection of random effects only.[13] In general, these methods are computationally extensive and complicated, primarily due to the complexity of optimizing the objective function constructed from the marginal likelihood function of the observed data. In the typical GLMM estimation process, the marginal likelihood function involves the integration with respect to the distribution of the random effects. Except for the LMM with normal responses and identity link, the marginal likelihood function generally does not have a closed-form solution and is typically approximated using numerical methods.[14–18] With the addition of penalty terms to the marginal likelihood function, optimizing the objective functions can be even more computationally challenging (see the GLMM regularized estimation approaches referenced above as examples). Recently, Hui, Müeller and Welsh[19] proposed a penalized quasi-likelihood (PQL) estimation for GLMM by approximating the marginal likelihood using the quasi-likelihood function, with sparsity inducing penalties on both fixed and random effect coefficients. Their simulation studies demonstrated much improved computational efficiency, compared to some existing methods.

In this paper, we propose a regularized estimation based on the confidence distribution approach.[20] The seed idea of the confidence distribution could be traced back to Bayes[21] and Fisher.[22] However, the concept and its applications have advanced extensively in recent years.[23,25,24,26] The confidence distribution can be viewed as a sample—DIFadd-dependent distribution function, and used to estimate and provide statistical inference for a parameter of interest.[27] Rather than optimizing the objective functions based on the likelihood function using observed data, we propose to construct and optimize the objective functions using the confidence distributions of model parameters, based on the asymptotic distribution of the model parameter estimators.[20] Because the confidence distribution of the model parameters is a multivariate normal distribution, we demonstrate that the objective function using the marginal likelihood function of the observed data can be approximated by the objective function constructed from the joint confidence distribution of the model parameters. Then, based on the joint and marginal confidence distributions of the model parameters, we propose two regularized estimations to perform simultaneous and separate selections of fixed and random effects, respectively. With proper choices for the regularization parameters, we show that the proposed estimators have the properties of estimation consistency, selection consistency, and the oracle property. Because the confidence distributions considered in our paper are based on the asymptotic distributions of the maximum likelihood estimators (MLEs), we consider finite-dimensional variable selections of the fixed and random effects as the asymptotic distributions of these MLEs are typically established for finite dimensions. Our approach may not be applicable for high dimensional variable selections of the fixed and random effects.

To the best of our knowledge, there are only a limited number of tools that perform GLMM regularized variable selections (e.g. the R packages `rpql`[19] for GLMM joint fixed and random effect selection, `glmmLasso`[11] for GLMM fixed effects selection only, and the R code of Bondell's method[6] for LMM fixed and random effect selections, available at https://blogs.unimelb.edu.au/howard-bondell/). Therefore, the availability of tools to perform computationally efficient regularized estimation for GLMM is highly desirable. As demonstrated later, our methods are simple, computationally efficient, and can be easily implemented using existing software packages without the need to develop new algorithms specific to GLMM.

The rest of this paper is organized as follows. In Section 2, we provide a brief review of the statistical inference in GLMM. In Section 3, we delineate 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 apply the proposed methods to the two examples of cancer studies. We conclude this paper with a discussion in Section 7.

2 Generalized linear mixed models

Consider a sample of n independent clusters. Let $\boldsymbol{y}_i = (y_{i1}, y_{i2}, \ldots, y_{im_i})^T$ and y_{ij} denote the jth measurement of the ith cluster, where $i = 1, 2, \ldots, n$, and $j = 1, 2, \ldots, m_i$. Let $\boldsymbol{x}_{ij}$ be a $(p_f + 1)$-variate vector of covariates corresponding to the fixed effects, and $\boldsymbol{z}_{ij}$ be a $(p_r + 1)$-variate vector of covariates corresponding to the random effects. Both $\boldsymbol{x}_{ij}$ and $\boldsymbol{z}_{ij}$ include 1 for the intercept. Typically, $\boldsymbol{z}_{ij}$ is a subset of $\boldsymbol{x}_{ij}$. Conditional on the random effects $\boldsymbol{b}_i$, we assume that the responses $y_{ij}'s$ follow a distribution of the exponential family with conditional mean μ_{ij} 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 $\boldsymbol{\mu}_i = (\mu_{i1}, \mu_{i2}, \ldots, \mu_{im_i})^T$, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_{p_f})^T$ is the fixed effect regression coefficients, $\boldsymbol{b}_i$ is a vector of random effects assumed to follow a multivariate normal distribution $\boldsymbol{N}_{p_r+1}(\boldsymbol{0}, \boldsymbol{I}_{p_r+1})$ with variance-covariance $\boldsymbol{I}_{p_r+1}$ being a $(p_r + 1)\times(p_r+1)$ identity matrix, and $\boldsymbol{\Gamma}$ is a $(p_r+1)\times(p_r+1)$ Cholesky decomposition lower triangular matrix depending on the parameter $\boldsymbol{\gamma}$ such that $\boldsymbol{\Gamma b}_i$ follows $\boldsymbol{N}_{(p_r+1)}(\boldsymbol{0}, \boldsymbol{D})$ and $\boldsymbol{D} = \boldsymbol{\Gamma\Gamma}^T$. Moreover, we assume that $\boldsymbol{\gamma}$ is a vector consisting of the row elements of the lower triangular components of $\boldsymbol{\Gamma}$ such that the length of $\boldsymbol{\gamma}$ is $(p_r + 1)(p_r + 2)/2$. For simplicity, we assume the canonical link such that $g(\boldsymbol{\mu_i}) = \boldsymbol{\eta}_i$. Consider $p_f < \infty$ and $p_r < \infty$ such that both $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are of finite dimensions. The model parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \boldsymbol{\gamma}^T, \phi)^T$ can be estimated by maximizing the marginal likelihood of $\boldsymbol{y}$ by 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 ϕ 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 MLE of $\boldsymbol{\theta}$ by

$$\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\beta}}^T, \hat{\boldsymbol{\gamma}}^T, \hat{\phi})^T = \arg\max_{\boldsymbol{\theta}} \log \mathcal{L}(\boldsymbol{\theta}; \boldsymbol{y})$$

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 $\hat{\boldsymbol{\Sigma}} = \boldsymbol{\mathcal{I}}^{-1}(\hat{\boldsymbol{\theta}})$.[28,29]

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 the penalty functions for model parameters. 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$_\theta$ $Q^o(\boldsymbol{\theta})$, 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 estimate $\boldsymbol{\theta}$, including the methods for the commonly used GLMM estimations,[14,15,18] and the methods for the regularized LMM or GLMM estimations.[6,10,12,19] To alleviate such a computational complexity in the regularized estimation process, we propose to perform the regularized estimation by optimizing the objective function constructed from the confidence distribution based on the MLE $\hat{\boldsymbol{\theta}}$.

Inference based on the confidence distribution has been extensively studied in the statistical literature (see literary works[30,20,31] and the references therein for a comprehensive review). In short, a confidence distribution can be viewed as a sample dependent distribution function that can be used to estimate and provide all aspects of statistical inference for a parameter of interest. This useful feature has been applied in Liu, Liu and Xie[24] for meta-analysis and Tian, Wang and Cai et al.[25] for joint inference about a set of constrained parameters in survival analysis. Then, based on Singh, Xie and Strawderman[20] and Liu et al.,[24] we write the confidence density of the parameter $\boldsymbol{\theta}$ according to the asymptotic distribution of $\hat{\boldsymbol{\theta}}$:

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

where p denotes the length of $\boldsymbol{\theta}$ and $\det(C)$ is the determinant of a matrix C. Note that $h(\boldsymbol{\theta})$ is a multivariate normal density. Taking the logarithm of $h(\boldsymbol{\theta})$, we get

$$-\log[h(\boldsymbol{\theta})] = \frac{1}{2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T\left(n^{-1}\hat{\boldsymbol{\Sigma}}\right)^{-1}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) + c \tag{4}$$

where c is some constant free of $\boldsymbol{\theta}$. Consider the following approximation. It can be seen that

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

since $\partial\log\mathcal{L}(\boldsymbol{\theta};\boldsymbol{y})/\partial\boldsymbol{\theta}^T|_{\theta=\hat{\theta}} = 0$ and constant $c' = n^{-1}\log\mathcal{L}(\hat{\boldsymbol{\theta}};\boldsymbol{y}) - c$ is free of $\boldsymbol{\theta}$. Thus $n^{-1}Q^o(\boldsymbol{\theta}) \approx n^{-1}\log[h(\boldsymbol{\theta})] + \kappa_\rho^o(\boldsymbol{\beta}) + \kappa_\tau^o(\boldsymbol{\gamma}) + c'$. This motivates us to construct the following objective function $Q(\boldsymbol{\theta})$ to approximate $Q^o(\boldsymbol{\theta})$ to perform regularized estimation using the confidence density $-\log[h(\boldsymbol{\theta})]$ in (4). Specifically, let

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

Note that the objective function $Q(\boldsymbol{\theta})$ takes the same form as the objective function of the least squares approximation (LSA) approach proposed by Wang and Leng[32] for the generalized linear models. Define the regularized estimator by

$$\hat{\boldsymbol{\theta}}_{\rho\tau} = \arg\min_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}) \tag{6}$$

It is interesting to notice the connection of our method with some methods that estimate $\boldsymbol{\theta}$ by optimizing $Q^o(\boldsymbol{\theta})$. For instance, Ibrahim et al.[10] estimates $\boldsymbol{\theta}$ by optimizing $Q^o(\boldsymbol{\theta})$ using a Monte Carlo EM algorithm that involves a Markov chain Monte Carlo sampling approach to approximate $\log\mathcal{L}(\boldsymbol{\theta};\boldsymbol{y})$; Hui et al. (2017) uses a quasi-likelihood to approximate $\log\mathcal{L}(\boldsymbol{\theta};\boldsymbol{y})$ to estimate $\boldsymbol{\theta}$. Our method approximates $\log\mathcal{L}(\boldsymbol{\theta};\boldsymbol{y})$ using the log confidence density, $-\log[h(\boldsymbol{\theta})]$, to estimate $\boldsymbol{\theta}$. Note that, in our method, performing the numerical approximation of the integration in $\log\mathcal{L}(\boldsymbol{\theta};\boldsymbol{y})$ is only required to derive $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\Sigma}}$, which can be achieved using existing software packages (e.g. `Proc Glimmix` in SAS, and `lme4` package in R). Once $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\Sigma}}$ are obtained, constructing and optimizing $Q(\boldsymbol{\theta})$ to estimate $\boldsymbol{\theta}$ no longer involves the numerical integration in $\log\mathcal{L}(\boldsymbol{\theta};\boldsymbol{y})$. Thus, the computational burden is greatly alleviated and the computational efficiency is much improved. 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 using simulation studies.

3.2 Statistical properties of the proposed estimator

To facilitate statistical inference (e.g. deriving the confidence intervals (CIs)), one may consider to use the adaptive LASSO[33] or the smoothly clipped absolute deviation (SCAD)[34] penalty functions for $\kappa_\rho(\cdot)$ and $\kappa_\tau(\cdot)$. In this paper, we focus on the adaptive LASSO framework. With little effort, SCAD regularization can be adopted in our proposed procedure. To be specific, we consider the following objective function:

$$Q(\boldsymbol{\theta}) = (\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})^T\left(n^{-1}\hat{\boldsymbol{\Sigma}}\right)^{-1}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}) + n\left(\sum_{f=1}^{p_f}\rho_f|\beta_f| + \sum_{m=2}^{p_r+1}\tau_m||\boldsymbol{\gamma}_m||\right)$$

by choosing the adaptive LASSO with $\kappa_\rho(\boldsymbol{\beta}) = \sum_{f=1}^{p_f}\rho_f|\beta_f|$ for the fixed effect selection, where $\rho_f's$are the adaptive weights that control the penalty with respect to $|\beta_f|$, for $f = 1, 2, \ldots, p_f$. For the random-effect selection, we use the adaptive group LASSO, following the rationale by He et al.[35]: Let $\boldsymbol{\gamma}_m$ denote the mth row of $\boldsymbol{\Gamma}$, then $\boldsymbol{\gamma}_m\boldsymbol{\gamma}_m^T = \boldsymbol{D}_{mm}$ which is the mth variance component of the random effects $\boldsymbol{\Gamma}\boldsymbol{b}_i$, for $m = 1, 2, \ldots, p_r + 1$. 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 $\boldsymbol{\Gamma}\boldsymbol{b}_i$ involving $(\boldsymbol{\Gamma}\boldsymbol{b}_i)_m$ are also 0. As a result, if a row vector $\boldsymbol{\gamma}_m$ is not selected, the random effect $(\boldsymbol{\Gamma}\boldsymbol{b}_i)_m$ and the corresponding component in $\boldsymbol{z}$ are excluded from the model and the positive-definitiveness of $\boldsymbol{D}$ is preserved. Thus, the adaptive group LASSO penalty is chosen as $\kappa_\tau(\boldsymbol{\gamma}) = \sum_{m=2}^{p_r+1}\tau_m||\boldsymbol{\gamma}_m||$ and $\tau_m's$ are the adaptive weights 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. Moreover, the parameter $\boldsymbol{\gamma}$ can be expressed as $\boldsymbol{\gamma} = (\gamma_1, \boldsymbol{\gamma}_2^T, \ldots, \boldsymbol{\gamma}_m^T)^T$.

Without loss of generality, we assume that only the first f_0 fixed effect covariates $\boldsymbol{x}_{ij}$ and the first r_0 random effect covariates $\boldsymbol{z}_{ij}$, both including the intercepts, are informative. Therefore, we write the true value of $\boldsymbol{\theta}$ by $\boldsymbol{\theta}_0 = (\boldsymbol{\beta}_0^T, \boldsymbol{\gamma}_0^T, \phi)^T$, where $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{0a}^T, \boldsymbol{\beta}_{0b}^T = \mathbf{0}^T)$ and $\boldsymbol{\gamma}_0 = (\boldsymbol{\gamma}_{0a}^T, \boldsymbol{\gamma}_{0b}^T = \mathbf{0}^T)$ such that $\boldsymbol{\beta}_{0a} = (\beta_{01}, \beta_{02}, \ldots, \beta_{0,f_0})^T$ with $\beta_{0j} \neq 0$ for $j = 1, 2, \ldots, f_0$, and $\boldsymbol{\gamma}_{0a}$ corresponds to the first r_0 rows in the lower triangle of $\boldsymbol{\Gamma}$ and each of these r_0 row vectors is non-zero. Similarly, $\hat{\boldsymbol{\theta}}_{\rho\tau} = (\hat{\boldsymbol{\beta}}_{\rho\tau}^T, \hat{\boldsymbol{\gamma}}_{\rho\tau}^T, \hat{\phi}_{\rho\tau})^T$, $\hat{\boldsymbol{\beta}}_{\rho\tau} = (\hat{\boldsymbol{\beta}}_{\rho\tau,a}^T, \hat{\boldsymbol{\beta}}_{\rho\tau,b}^T)^T$, and $\hat{\boldsymbol{\gamma}}_{\rho\tau} = (\hat{\boldsymbol{\gamma}}_{\rho\tau,a}^T, \hat{\boldsymbol{\gamma}}_{\rho\tau,b}^T)^T$.

Obviously, $Q(\boldsymbol{\theta})$ is strictly convex in $\boldsymbol{\theta}$. We establish the consistency and oracle properties of $\hat{\boldsymbol{\theta}}_{\rho\tau}$ in Theorem 1.

Theorem 1. *Let $a_{f,n} = \max\{\rho_j, j \leq f_0\}$, $b_{f,n} = \min\{\rho_j, j > f_0\}$, $a_{r,n} = \max\{\tau_j, j \leq r_0\}$, 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 $n^{1/2}b_{r,n} \xrightarrow{P} \infty$, $Pr(\hat{\boldsymbol{\beta}}_{\rho\tau,b} = \mathbf{0}$ and $\hat{\boldsymbol{\gamma}}_{\rho\tau,b} = \mathbf{0}) \to 1$.

(3) (Oracle property) Let $\boldsymbol{\theta}_{0a} = (\boldsymbol{\beta}_{0a}^T, \boldsymbol{\gamma}_{0a}^T, \phi_0)^T$ and $\hat{\boldsymbol{\theta}}_{\rho\tau,a} = (\hat{\boldsymbol{\beta}}_{\rho\tau,a}^T, \hat{\boldsymbol{\gamma}}_{\rho\tau,a}^T, \hat{\phi}_0)^T$. 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 $n^{1/2}b_{r,n} \xrightarrow{P} \infty$, then $n^{1/2}(\hat{\boldsymbol{\theta}}_{\rho\tau a} - \boldsymbol{\theta}_{0a}) \xrightarrow{D} \mathcal{N}(\mathbf{0}, [(\boldsymbol{\Sigma}^{-1})_{\boldsymbol{\theta}_{0a}}]^{-1})$, 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 $[(\boldsymbol{\Sigma}^{-1})_{\boldsymbol{\theta}_{0a}}]^{-1}$ can be consistently estimated by $[(\hat{\boldsymbol{\Sigma}}^{-1})_{\boldsymbol{\theta}_{0a}}]^{-1}$, where $(\hat{\boldsymbol{\Sigma}}^{-1})_{\boldsymbol{\theta}_{0a}}$ is the submatrix of $\hat{\boldsymbol{\Sigma}}^{-1}$ corresponding to $\boldsymbol{\theta}_{0a}$.

A sketch of the proof is provided in the Appendix.

3.3 An alternative estimation

Recall that the objective function $Q(\boldsymbol{\theta})$ is built by the confidence density $h(\boldsymbol{\theta})$ in (3), according to the joint asymptotic distribution of $\hat{\boldsymbol{\theta}}$. Note that $h(\boldsymbol{\theta})$ is a 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 another estimation based on the marginal confidence densities with respective to $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ in (3) to separately estimate $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$. These marginal confidence densities 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 to construct separate objective functions for $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\gamma}}$ in the following:

$$Q_f(\boldsymbol{\beta}) = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T [n^{-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}}]^{-1} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + n \sum_{f=1}^{p_f} \rho_f |\beta_f|$$

$$Q_r(\boldsymbol{\gamma}) = (\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma})^T [n^{-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\gamma}}]^{-1} (\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}) + n \sum_{m=2}^{p_r+1} \tau_m ||\boldsymbol{\gamma}_m||$$

where $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\gamma}}$ are submatrices of $\hat{\boldsymbol{\Sigma}}$ corresponding to the marginal variance-covariance of $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\gamma}}$, respectively. Define the regularized estimators as $\hat{\boldsymbol{\beta}}_{\rho}^s = \arg\min_{\boldsymbol{\beta}} Q_f(\boldsymbol{\beta})$ and $\hat{\boldsymbol{\gamma}}_{\tau}^s = \arg\min_{\boldsymbol{\gamma}} Q_r(\boldsymbol{\gamma})$. The CD-separate estimation allows for the flexibility of performing the selection of fixed effects only or the selection of random effects only, and enables the performance of fixed- and (or) random-effect selections in case only $\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}$,and $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\gamma}}$ are available by convenience.

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 $\hat{\boldsymbol{\gamma}}$ in $h(\boldsymbol{\theta})$, 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, although the CD-joint estimators and CD-separate estimators are different estimators. Recall that $\boldsymbol{\beta}_0$ and $\boldsymbol{\gamma}_0$ are expressed as $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{0a}^T, \boldsymbol{\beta}_{0b}^T = \mathbf{0}^T)^T$ and $\boldsymbol{\gamma}_0 = (\boldsymbol{\gamma}_{0a}^T, \boldsymbol{\gamma}_{0b}^T = \mathbf{0}^T)^T$. Similarly, we write $\hat{\boldsymbol{\beta}}_{\rho}^s = (\hat{\boldsymbol{\beta}}_{\rho,a}^{s\,T}, \hat{\boldsymbol{\beta}}_{\rho,b}^{s\,T})^T$ and $\hat{\boldsymbol{\gamma}}_{\tau}^s = (\hat{\boldsymbol{\gamma}}_{\tau,a}^{s\,T}, \hat{\boldsymbol{\gamma}}_{\tau,b}^{s\,T})^T$. Obviously, both $Q_f(\boldsymbol{\beta})$ and $Q_r(\boldsymbol{\gamma})$ are convex in $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, respectively. We establish the consistency and oracle properties for $\hat{\boldsymbol{\beta}}_{\rho}^s$ and $\hat{\boldsymbol{\gamma}}_{\tau}^s$ as follows.

Theorem 2. *Let $a_{f,n} = \max\{\rho_j, j \leq f_0\}$, and $b_{f,n} = \min\{\rho_j, j > f_0\}$. Then the regularized estimator $\hat{\boldsymbol{\beta}}_{\rho}^s$ 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* $n^{1/2}b_{f,n} \xrightarrow{P} \infty$, $Pr(\hat{\boldsymbol{\beta}}^s_{\rho,b} = \mathbf{0}) \to 1$.

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

Theorem 3. *Let* $a_{r,n} = \max\{\tau_j, j \le r_0\}$, *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* $n^{1/2}b_{r,n} \xrightarrow{P} \infty$, $Pr(\hat{\boldsymbol{\gamma}}^s_{\tau,b} = \mathbf{0}) \to 1$.

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

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

Although the dispersion parameter $\boldsymbol{\phi}$ is often a nuisance parameter and thus omitted in the previously described separate estimation, it can actually be included, for instance, by combining $\boldsymbol{\phi}$ with $\boldsymbol{\gamma}$. Then we modify $Q_r(\boldsymbol{\gamma})$ by $Q^*_r(\boldsymbol{\gamma}, \boldsymbol{\phi})$ given below, based on the (marginal) joint distribution of $\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\phi}}$ in (3):

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

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

4 Optimization and determination of tuning parameters

To obtain $\hat{\boldsymbol{\theta}}_{\rho\tau}$ via optimizing $Q(\boldsymbol{\theta})$, we follow the method of Zhang and Lu[36] and rewrite the objective function $Q(\boldsymbol{\theta})$ as

$$Q(\boldsymbol{\theta}) = (\Lambda\boldsymbol{\theta} - \Psi)^T (\Lambda\boldsymbol{\theta} - \Psi) + n\kappa_\rho(\boldsymbol{\beta}) + n\kappa_\tau(\boldsymbol{\gamma}) \tag{7}$$

where $\Psi = \Lambda\hat{\boldsymbol{\theta}}$, and Λ can be obtained using the singular value decomposition such that $(n^{-1}\hat{\boldsymbol{\Sigma}})^{-1} = \Lambda^T\Lambda$. Then the function in (7) is a typical convex optimization problem and can be solved by standard software packages, for instance, the R packages `glmnet`[37] and `gglasso`.[38] The same approach can also be applied to optimize $Q_f(\boldsymbol{\beta})$, $Q_r(\boldsymbol{\gamma})$, and $Q^*_r(\boldsymbol{\gamma},\boldsymbol{\phi})$. For instance, 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 $\Lambda_{\boldsymbol{\gamma}}$ are obtained using the singular value decomposition such that $[n^{-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}}]^{-1} = \Lambda^T_{\boldsymbol{\beta}}\Lambda_{\boldsymbol{\beta}}$ and $[n^{-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\gamma}}]^{-1} = \Lambda^T_{\boldsymbol{\gamma}}\Lambda_{\boldsymbol{\gamma}}$. As a result, $Q_f(\boldsymbol{\beta}) = (\Lambda_{\boldsymbol{\beta}}\boldsymbol{\beta} - \Psi_{\boldsymbol{\beta}})^T(\Lambda_{\boldsymbol{\beta}}\boldsymbol{\beta} - \Psi_{\boldsymbol{\beta}}) + n\kappa_\rho(\boldsymbol{\beta})$ and $Q_r(\boldsymbol{\gamma}) = (\Lambda_{\boldsymbol{\gamma}}\boldsymbol{\gamma} - \Psi_{\boldsymbol{\gamma}})^T(\Lambda_{\boldsymbol{\gamma}}\boldsymbol{\gamma} - \Psi_{\boldsymbol{\gamma}}) + n\kappa_\tau(\boldsymbol{\gamma})$, respectively. In our simulations and data analysis, we used R package `gglasso` to optimize $Q(\boldsymbol{\theta})$, $Q_r(\boldsymbol{\gamma})$, and $Q^*_r(\boldsymbol{\gamma}, \boldsymbol{\phi})$, and `glmnet` to optimize $Q_f(\boldsymbol{\beta})$.

Typically, the tuning parameters ρ_f's and τ_m's can be chosen using the approaches of cross validation or generalized cross validation. But, these methods can be computationally extensive. With the simple solution suggested by Zou,[33] we consider $\rho_f = \lambda|\hat{\beta}_f|^{-\varphi_f}$ and $\tau_m = \lambda||\hat{\boldsymbol{\gamma}}_m||^{-\varphi_r}$ for the CD-joint estimation method, and $\rho_f = \lambda_f|\hat{\beta}_f|^{-\varphi_f}$ and $\tau_m = \lambda_r||\hat{\boldsymbol{\gamma}}_m||^{-\varphi_r}$ for the CD-separate estimation method, for $f = 1, 2, \ldots, p_f$ and $m = 2, 3, \ldots, p_r + 1$, where $\hat{\beta}_f$ and $\hat{\boldsymbol{\gamma}}_m$ are the maximum likelihood estimates for β_f and γ_m, respectively, and φ_f and φ_r are pre-specified positive numbers. In the CD-joint method, the adaptive LASSO penalties for the fixed effects and random effects are linked by the tuning parameter $\lambda > 0$. The CD-separate estimation allows each of the fixed- and random-effect selections to have its own tuning parameter to control the shrinkage. Because the MLEs $\hat{\beta}_f$'s and $\hat{\boldsymbol{\gamma}}_m$'s are $\sqrt{n}$-consistent, it can be verified that the tuning parameters considered above satisfy the conditions required by Theorems 1, 2, and 3, provided that $n^{1/2}\lambda \to 0$, $n^{(1+\varphi_f)/2}\lambda \to \infty$, and $n^{(1+\varphi_r)/2}\lambda \to \infty$ as well as $n^{1/2}\lambda_f \to 0$, $n^{(1+\varphi_f)/2}\lambda_f \to \infty$, $n^{1/2}\lambda_r \to 0$ and $n^{(1+\varphi_r)/2}\lambda_r \to \infty$. Thus it suffices to

select $\lambda \in R+ = [0, \infty)$, $\lambda_f \in R+ = [0, \infty)$ and $\lambda_r \in R+ = [0, \infty)$. Therefore, to determine λ, λ_f, and λ_r, we consider to minimize BIC, per recommendations by prior research.[32,40] Specifically, for the CD-joint estimation, we 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_ρ is the number of non-zero coefficients in $\hat{\boldsymbol{\beta}}_{\rho\tau}$, and df_τ is the number of groups with non-zero within-group coefficients $\hat{\boldsymbol{\gamma}}_{\rho\tau}$. For the CD-separate estimation, we define $BIC_{f,\rho} = n(\hat{\boldsymbol{\beta}}^s_\rho - \hat{\boldsymbol{\beta}})^T \hat{\boldsymbol{\Sigma}}^{-1}_\beta (\hat{\boldsymbol{\beta}}^s_\rho - \hat{\boldsymbol{\beta}}) + (\log n) df_\rho$ for optimizing $Q_f(\boldsymbol{\beta})$, and $BIC_{r,\tau} = n(\hat{\boldsymbol{\gamma}}^s_\tau - \hat{\boldsymbol{\gamma}})^T \hat{\boldsymbol{\Sigma}}^{-1}_\gamma (\hat{\boldsymbol{\gamma}}^s_\tau - \hat{\boldsymbol{\gamma}}) + (\log n) df_\tau$ 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 optimizing $Q_r(\boldsymbol{\gamma})$ and $Q^*_r(\boldsymbol{\gamma}, \phi)$, respectively.

5 Simulation studies

We conducted simulation studies to examine the performance of the proposed CD methods and compared them with the methods by Hui et al.[19] (`rpql` R package), Bondell et al.[6] (available at https://blogs.unimelb.edu.au/howard-bondell/), and the method of Groll et al.[11] (`glmmLasso` R package). Hui et al.[19] and Bondell et al.[6] used adaptive Lasso penalties. Groll et al.[11] used the Lasso penalty. We applied these methods because either the R packages or the R code are publicly available. Data were simulated under 3 scenarios (i.e. LMM, random effects logistic regression models, and random effects Poisson models) according to model (1), with results summarized in Tables 1 to 5. Moreover, we have extended our methods to nested random effects in GLMM, with detailed descriptions of model specifications and simulation results in the Supplemental Materials.

In Scenario 1, we generated $y'_{ij}s$ from the LMM such that $y_{ij}|\boldsymbol{b}_i$ follows $N(\boldsymbol{x}^T_{ij}\boldsymbol{\beta} + \boldsymbol{z}^T_{ij}\boldsymbol{\Gamma}\boldsymbol{b}_i, \sigma^2)$ 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 for the random effects Poisson model with a log link. In all three scenarios, we chose $p_f = 15$ for fixed effects and $p_r = 3$ for random effects. The true value for $\boldsymbol{\beta}$ was $\boldsymbol{\beta}_0 = (\mathbf{1}_6, \mathbf{0}_{10})$ for LMM and random effects logistic regression models (Scenarios 1 and 2), and $\boldsymbol{\beta}_0 = (1, -\mathbf{1}_5, \mathbf{0}_{10})$ for random effects Poisson models (Scenarios 3). The true 4×4 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, and $vech(\boldsymbol{D}) = (3, 1.2, 0.8, 0; 2, 0.5, 0; 1, 0; 0)$ for Scenarios 2 and 3, i.e. only the first three components of $\boldsymbol{z}_{ij}$, including the random intercept, are informative. As a result, we express the corresponding Cholesky decomposition lower triangular matrix by showing the row elements in the lower triangle as $\boldsymbol{\Gamma} = (3; 1.6, 1.2; 0.2, 0.57, 0.8; \mathbf{0}_4)$ 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 numbers of clusters n and cluster size m. Covariates $\boldsymbol{x}_{ij} = (1, x_{ij,1}, x_{ij,2}, \ldots, x_{ij,p_f})^T$ and $\boldsymbol{z}_{ij} = (1, z_{ij,1}, z_{ij,2}, \ldots, z_{ij,p_r})^T$, for $p_f = 15$ and $p_r = 3$, were generated from a mix of continuous, categorical (binary) variables and interactions of continuous and categorical variables. Specifically, $x_{ij,1}, x_{ij,3}, x_{ij,6} \sim x_{ij,8}$ are generated from the standard normal distribution, $x_{ij,2}, x_{ij,12} \sim x_{ij,14}$ were generated from exponential distribution with mean 1 (exponential(1)), $x_{ij,4}, x_{ij,9} \sim x_{ij,11}$ were generated from Bernoulli (0.4) distribution; $x_{ij,4}$ and $x_{ij,15}$ are interaction terms with $x_{ij,5} = x_{ij,1} * x_{ij,4}$ and $x_{ij,15} = x_{ij,6} * x_{ij,9}$. Moreover, $z_{ij,l} = x_{ij,l}$, for $l = 1, 2, 3$. All continuous covariates were standardized to have mean 0 and variance 1.

For the proposed methods, we refer CD-joint and CD-separate estimations as CD-J and CD-S, respectively. We showed the mean of the estimates, empirical standard error (ESE), percentage of selection (% Sel), and the average computation time (Time (mins)). For the proposed CD methods, coverage probability of 95% CIs (CovP) was calculated based on the oracle properties in Theorems 1 to 3. In addition to the CD methods, we also fitted the GLMM models to obtain the model estimates as initial values to implement the method of Hui et al.[19] (referred to as the rPQL method), following the examples in Hui.[39] When we calculated the computation time, we included the time for fitting the GLMM models, when applicable, as well as the time of the regularization process, including the determination of tuning parameters. The R `lme4` package was used to derive the GLMM estimates. For the CD-J and CD-S, the tuning parameters λ, λ_f, and λ_r were determined from 10^ω, where ω went from -4 to 4 by 0.01 (801 values in total). For the rPQL method, we used the function *lseq*(), provided by the `rpql` package,[39] to determine the tuning parameters from $(0, 100]$ using syntax "$lseq(1e-6, 10^2, length = 200)$" (200 values in total). Several ranges wider than $(0, 100]$ were applied, each for $50 - 100$ simulation runs, and results were similar. For `glmmLasso`, the tuning parameters were chosen from 10^ω, where ω went from -4 to 4 by 0.08 (101 values in total). We applied different ranges and different numbers of tuning parameters for the CD-methods, rPQL and glmmLasso methods, with CD-methods using the most number of tuning parameters. Because the computation time of the rPQL and glmmLasso methods was longer, especially for large n's, we determined to use smaller numbers of tuning parameters for these methods to facilitate the progress of the simulation studies, after trying various ranges of tuning parameters and making sure results were similar.

Moreover, we performed additional simulations to examine the impact of φ_f and φ_r (see results in the Supplemental Materials). Specifically, we considered the values of φ_f and φ_r to be 0.25, 1 and 4 for the proposed CD-J, CD-S, and the

Table 1. Linear mixed model: Fixed effect selection.

(n, m)	Method	β_0	I	I	I	I	I	I	$\mathbf{0}_{10}$
(30, 10)	CD-J	$\hat{\beta}_{\rho\tau}$	1.011	0.996	0.986	0.996	0.997	0.992	0.000
		ESE	0.565	0.388	0.199	0.071	0.146	0.156	0.042
		CovP(%)	93.0	92.4	93.9	93.0	94.1	93.9	-
		% Sel	100.0	99.8	100.0	100.0	100.0	100.0	8.0
	CD-S	$\hat{\beta}^s_\rho$	1.011	0.99	0.98	0.99	0.991	0.987	0.000
		ESE	0.565	0.391	0.201	0.071	0.147	0.157	0.030
		CovP(%)	92.6	92.0	93.5	92.6	93.7	93.5	-
		% Sel	100.0	99.2	100.0	100.0	100.0	100.0	4.0
	rPQL	$\hat{\beta}^{rPQL}$	0.916	0.791	0.87	0.998	0.987	0.976	0.000
		ESE	0.473	0.500	0.318	0.061	0.137	0.166	0.000
		% Sel	100.0	76.1	95.8	100	100	99.8	7.9
	Bondell et al.	$\hat{\beta}^B$	-	0.176	0.681	0.988	0.932	0.937	0.001
		ESE	-	0.249	0.257	0.069	0.158	0.174	0.001
		% Sel	-	45.2	99.2	100.0	100.0	100.0	6.1
(60, 6)	CD-J	$\hat{\beta}_{\rho\tau}$	0.979	0.979	0.997	0.991	0.99	0.994	−0.001
		ESE	0.417	0.286	0.148	0.073	0.142	0.166	0.037
		CovP(%)	93.4	92.8	93.8	91.6	94.5	92.9	-
		% Sel	100.0	100.0	100.0	100.0	100.0	100.0	6.6
	CD-S	$\hat{\beta}^s_\rho$	0.979	0.973	0.992	0.986	0.984	0.989	−0.001
		ESE	0.417	0.288	0.149	0.074	0.143	0.167	0.028
		CovP(%)	93.4	92.7	93.7	90.9	94.5	92.5	-
		% Sel	100.0	99.9	100.0	100.0	100.0	100.0	3.4
	rPQL	$\hat{\beta}^{rPQL}$	0.907	0.797	0.861	0.995	0.972	0.97	0.000
		ESE	0.489	0.488	0.319	0.07	0.16	0.197	0.000
		% Sel	100.0	78.9	95.9	100.0	100.0	99.5	12.5
	Bondell et al.	$\hat{\beta}^B$	-	0.238	0.774	0.996	0.964	0.981	0.001
		ESE	-	0.223	0.174	0.051	0.106	0.117	0.001
		% Sel	-	76.5	100.0	100.0	100.0	100.0	6.5
(120, 6)	CD-J	$\hat{\beta}_{\rho\tau}$	0.993	0.991	0.993	0.994	0.996	0.995	0.000
		ESE	0.291	0.197	0.11	0.05	0.103	0.116	0.023
		CovP(%)	93.8	94.0	93.1	93.9	94.3	93.1	-
		% Sel	100.0	100.0	100.0	100.0	100.0	100.0	3.9
	CD-S	$\hat{\beta}^s_\rho$	0.993	0.988	0.99	0.992	0.993	0.992	0.000
		ESE	0.291	0.198	0.11	0.05	0.104	0.116	0.017
		CovP(%)	93.7	94	93	93	94.1	93.2	-
		% Sel	100.0	100.0	100.0	100.0	100.0	100.0	1.96
	rPQL	$\hat{\beta}^{rPQL}$	1.021	0.936	0.957	0.999	0.988	0.998	−0.001
		ESE	0.347	0.333	0.196	0.055	0.11	0.128	−0.001
		% Sel	100	91.6	99.6	100.0	100.0	100.0	7.7
	Bondell et al.	$\hat{\beta}^B$	-	0.351	0.834	1.008	0.992	0.978	0.002
		ESE	-	0.188	0.132	0.052	0.11	0.118	0.002
		% Sel	-	95.5	100.0	100.0	100.0	100.0	7.3
(500, 6)	CD-J	$\hat{\beta}_{\rho\tau}$	1.001	0.999	0.999	0.998	1.002	1.001	0.000
		ESE	0.142	0.094	0.052	0.024	0.049	0.053	0.007
		CovP(%)	94.2	95.4	95.5	93.8	95	94.7	-
		% Sel	100.0	100.0	100.0	100.0	100.0	100.0	1.5
	CD-S	$\hat{\beta}^s_\rho$	1.001	0.999	0.998	0.997	1.001	1.001	0.000
		ESE	0.142	0.094	0.052	0.024	0.049	0.053	0.005
		CovP(%)	94.2	95.3	95.5	93.6	95.1	94.7	-
		% Sel	100.0	100.0	100.0	100.0	100.0	100.0	0.6
	rPQL	$\hat{\beta}^{rPQL}$	1.023	0.996	0.991	0.996	0.995	1.000	0.000
		ESE	0.141	0.106	0.055	0.025	0.051	0.054	0.000
		% Sel	100.0	100.0	100.0	100.0	100.0	100.0	0.6

PQL: penalized quasi-likelihood; ESE: empirical standard error.

Table 2. Linear mixed model: Random effect selection.

(n, m)		γ_0	3	1.6	1.2	0.2	0.57	0.8	$\mathbf{0}_4$	Time (Mins)
(30, 10)	CD-J	$\hat{\gamma}_{\rho\tau}$	3.066	1.609	1.189	0.194	0.508	0.645	0.000	0.003 ± 0.001
		ESE	0.467	0.346	0.200	0.206	0.198	0.177	0.007	
		CovP(%)	88.0	88.6	86.9	88.6	87.1	68.3	-	
		% Sel	100.0	100.0	100.0	98.7	98.7	98.7	0.2	
	CD-S	$\hat{\gamma}^s_\rho$	3.066	1.623	1.200	0.206	0.558	0.719	0.000	0.003 ± 0.001
		ESE	0.467	0.344	0.200	0.217	0.207	0.163	0.029	
		CovP(%)	88.0	89.4	86.9	87.1	88.6	78.1	-	
		% Sel	100.0	100.0	100.0	100.0	100.0	100.0	0.7	
	rPQL	$\hat{\gamma}^{rPQL}$	2.978	1.587	1.282	0.206	0.465	0.920	0.060	0.332 ± 0.474
		ESE	0.263	0.225	0.151	0.156	0.175	0.123	0.185	
		% Sel	99.5	99.5	99.5	99.5	99.5	99.5	0.0	
	Bondell et al.	$\hat{\gamma}^B$	2.935	1.634	0.259	1.089	0.583	0.652	0.004	3.587 ± 1.119
		ESE	0.400	0.318	0.206	0.227	0.203	0.156	0.005	
		% Sel	100.0	100.0	100.0	100.0	100.0	100.0	1.8	
(60, 6)	CD-J	$\hat{\gamma}_{\rho\tau}$	3.116	1.633	1.207	0.193	0.511	0.699	0.000	0.004 ± 0.001
		ESE	0.353	0.258	0.174	0.152	0.165	0.158	0.000	
		CovP(%)	83.8	90.1	84.6	91.6	87.4	76.6	-	
		% Sel	100.0	100.0	100.0	99.7	99.7	99.7	0.0	
	CD-S	$\hat{\gamma}^s_\rho$	3.116	1.660	1.229	0.214	0.574	0.794	0.027	0.003 ± 0.001
		ESE	0.353	0.257	0.174	0.167	0.171	0.144	0.175	
		CovP(%)	85.1	90.2	85.1	89.5	88.4	87.4	-	
		% Sel	100.0	100.0	100.0	100.0	100.0	100.0	0.0	
	rPQL	$\hat{\gamma}^{rPQL}$	2.978	1.587	1.282	0.206	0.465	0.920	0.060	0.113 ± 0.489
		ESE	0.263	0.225	0.151	0.156	0.175	0.123	0.185	
		% Sel	99.5	99.5	99.5	99.5	99.5	99.5	0.0	
	Bondell et al.	$\hat{\gamma}^B$	3.011	1.651	0.228	1.149	0.564	0.728	0.001	18.452 ± 5.107
		ESE	0.283	0.226	0.140	0.144	0.142	0.099	0.002	
		% Sel	100.0	100.0	100.0	100.0	100.0	100.0	0.4	
(120, 6)	CD-J	$\hat{\gamma}_{\rho\tau}$	3.055	1.609	1.202	0.193	0.545	0.744	0.000	0.009 ± 0.002
		ESE	0.247	0.180	0.114	0.103	0.110	0.102	0.000	
		CovP(%)	86.2	89.9	88.3	94.0	92.1	83.0	-	
		% Sel	100.0	100.0	100.0	100.0	100.0	100.0	0.0	
	CD-S	$\hat{\gamma}^s_\rho$	3.055	1.619	1.209	0.199	0.566	0.772	0.000	0.005 ± 0.001
		ESE	0.247	0.180	0.114	0.107	0.113	0.096	0.014	
		CovP(%)	86.2	89.6	89.2	92.8	91.9	88.4	-	
		% Sel	100.0	100.0	100.0	100.0	100.0	100.0	0.2	
	rPQL	$\hat{\gamma}^{rPQL}$	2.990	1.591	1.272	0.206	0.466	0.926	0.048	0.638 ± 0.347
		ESE	0.204	0.163	0.105	0.108	0.121	0.079	0.140	
		% Sel	99.6	99.6	99.6	99.6	99.6	99.6	0.0	
	Bondell et al.	$\hat{\gamma}^B$	3.062	1.663	0.237	1.159	0.594	0.734	0.003	86.251 ± 19.771
		ESE	0.216	0.172	0.103	0.185	0.123	0.088	0.004	
		% Sel	100.0	100.0	100.0	100.0	100.0	100.0	1.3	
(500, 6)	CD-J	$\hat{\gamma}_{\rho\tau}$	3.023	1.608	1.200	0.202	0.559	0.782	0.000	0.034 ± 0.004
		ESE	0.117	0.092	0.056	0.055	0.055	0.047	0.000	
		CovP(%)	87.3	89.2	90.6	93.6	93.6	87.9	-	
		% Sel	100.0	100.0	100.0	100.0	100.0	100.0	0.0	
	CD-S	$\hat{\gamma}^s_\rho$	3.023	1.611	1.202	0.204	0.566	0.791	0.000	0.020 ± 0.003
		ESE	0.117	0.092	0.056	0.055	0.055	0.047	0.006	
		CovP(%)	87.3	89.6	90.6	93.4	93.1	90.0	-	
		% Sel	100.0	100.0	100.0	100.0	100.0	100.0	0.1	
	rPQL	$\hat{\gamma}^{rPQL}$	2.999	1.576	1.277	0.202	0.467	0.945	0.000	43.226 ± 7.230
		ESE	0.107	0.098	0.051	0.057	0.046	0.040	0.096	
		% Sel	99.1	99.1	99.1	99.1	99.1	99.1	0.0	

PQL: penalized quasi-likelihood; ESE: empirical standard error.

Table 3. Random effect logistic regression model: Fixed effect selection.

(n, m)	Method	β_0	I	I	I	I	I	I	$\mathbf{0}_{10}$	Time (Mins)
(30, 10)	CD-J	$\hat{\beta}_{\rho\tau}$	1.394	1.272	1.260	1.209	1.278	1.318	0.004	0.078 ± 0.036
		ESE	1.297	1.143	1.101	0.855	1.386	1.384	0.361	
		CovP(%)	82.1	83.9	83.5	76.4	83.2	77.8	-	
		% Sel	100.0	96.8	98.9	99.8	92.7	92.4	14.5	
	CD-S	$\hat{\beta}^s_\rho$	1.394	1.261	1.251	1.212	1.273	1.318	0.002	0.078 ± 0.036
		ESE	1.297	1.144	1.101	0.851	1.386	1.383	0.361	
		CovP(%)	87.6	88.7	91.9	79.4	91.1	81.4	-	
		% Sel	100.0	96.1	98.6	99.8	92.2	91.4	12.9	
	rPQL	$\hat{\beta}^{rPQL}$	0.718	0.599	0.570	0.652	0.547	0.601	0.001	0.816 ± 0.625
		ESE	0.381	0.431	0.353	0.327	0.487	0.540	0.001	
		% Sel	100.0	77.4	82.5	88.3	66.0	65.5	7.0	
	glmmLasso	$\hat{\beta}^g$	0.631	0.343	0.292	0.334	0.21	0.351	−0.001	1.203 ± 0.398
		ESE	0.227	0.193	0.197	0.193	0.257	0.252	−0.001	
		% Sel	100.0	93.7	87.2	91.4	59.9	86.1	18.6	
(80, 10)	CD-J	$\hat{\beta}_{\rho\tau}$	1.048	1.021	0.998	1.001	1.027	1.013	0.000	0.163 ± 0.033
		ESE	0.331	0.289	0.236	0.189	0.333	0.373	0.06	
		CovP(%)	88.4	90.2	86.1	89.2	90.6	89.3	-	
		% Sel	100.0	100.0	100.0	100.0	99.7	99.0	5.5	
	CD-S	$\hat{\beta}^s_\rho$	1.048	1.02	0.999	1.006	1.029	1.017	0.000	0.163 ± 0.033
		ESE	0.331	0.289	0.234	0.186	0.332	0.371	0.06	
		CovP(%)	90.8	93.3	91.9	92.5	94.0	93.4	-	
		% Sel	100.0	99.9	100.0	100.0	99.6	98.9	4.5	
	rPQL	$\hat{\beta}^{rPQL}$	0.664	0.604	0.619	0.689	0.642	0.605	0.000	0.339 ± 0.041
		ESE	0.206	0.238	0.175	0.186	0.344	0.365	0.000	
		% Sel	100.0	94.7	98.3	97.3	88.7	81.7	5.4	
	glmmLasso	$\hat{\beta}^g$	0.55	0.409	0.401	0.46	0.375	0.465	0.001	7.407 ± 3.471
		ESE	0.151	0.113	0.111	0.104	0.192	0.171	0.001	
		% Sel	100.0	100.0	100.0	100.0	99.5	100.0	43.4	
(200, 6)	CD-J	$\hat{\beta}_{\rho\tau}$	1.046	1.021	1.01	0.981	1.005	0.989	−0.001	0.282 ± 0.082
		ESE	0.392	0.303	0.293	0.2	0.324	0.37	0.055	
		CovP(%)	82.1	87.7	82.3	85.5	89.7	90.8	-	
		% Sel	100.0	100.0	100.0	100.0	99.9	99.9	4.5	
	CD-S	$\hat{\beta}^s_\rho$	1.046	1.023	1.013	0.987	1.009	0.995	0.000	0.282 ± 0.082
		ESE	0.392	0.303	0.292	0.2	0.324	0.369	0.054	
		CovP(%)	87.9	92.8	91.3	91.4	95.4	94.8	-	
		% Sel	100.0	100.0	100.0	100.0	99.9	99.9	3.2	
	rPQL	$\hat{\beta}^{rPQL}$	0.640	0.524	0.501	0.591	0.532	0.505	0.001	2.361 ± 0.282
		ESE	0.179	0.280	0.254	0.265	0.364	0.377	0.001	
		% Sel	100.0	85.4	87.6	88.6	79.8	73.2	4.6	
(200, 10)	CD-J	$\hat{\beta}_{\rho\tau}$	1.014	1.009	0.999	0.988	0.998	0.983	0.000	0.350 ± 0.057
		ESE	0.206	0.171	0.139	0.112	0.192	0.210	0.036	
		CovP(%)	87.2	90.3	87.2	90.8	91.3	91.5	-	
		% Sel	100.0	100.0	100.0	100.0	100.0	100.0	5.3	
	CD-S	$\hat{\beta}^s_\rho$	1.014	1.005	0.995	0.984	0.994	0.979	0.000	0.350 ± 0.057
		ESE	0.206	0.171	0.137	0.111	0.193	0.211	0.025	
		CovP(%)	89.6	93.7	91.4	93.2	94.6	93.8	-	
		% Sel	100.0	100.0	100.0	100.0	100.0	100.0	1.9	
	rPQL	$\hat{\beta}^{rPQL}$	0.609	0.415	0.350	0.463	0.397	0.359	0.000	0.585 ± 0.493
		ESE	0.076	0.277	0.256	0.310	0.339	0.309	0.000	
		% Sel	100.0	74.8	73.3	73.3	64.4	65.2	2.7	
	glmmLasso	$\hat{\beta}^g$	0.49	0.416	0.439	0.493	0.442	0.502	0.001	61.5 ± 35.362
		ESE	0.094	0.066	0.069	0.066	0.116	0.098	0.001	
		% Sel	100	100	100	100	100	100	54.3	

(continued)

Table 3. Random effect logistic regression model: Fixed effect selection.

(n, m)	Method	β_0	1	1	1	1	1	1	$\mathbf{0}_{10}$	Time (Mins)
(500, 6)	CD-J	$\hat{\beta}_{\rho\tau}$	1.008	1.008	1.001	0.983	0.999	0.986	0.000	0.546 ± 0.091
		ESE	0.126	0.105	0.098	0.121	0.062	0.105	0.020	
		CovP(%)	89.3	91.8	88.5	93.6	88.2	94.0	-	
		% Sel	100.0	100.0	100.0	100.0	100.0	100.0	3.6	
	CD-S	$\hat{\beta}^s_\rho$	1.010	0.995	0.978	0.971	0.979	0.966	0.000	0.546 ± 0.091
		ESE	0.126	0.106	0.098	0.121	0.062	0.105	0.010	
		CovP(%)	90.5	93.6	94.9	94.1	93.7	94.2	-	
		% Sel	100.0	100.0	100.0	100.0	100.0	100.0	1.4	
	rPQL	$\hat{\beta}^{rPQL}$	0.609	0.415	0.350	0.463	0.397	0.359	0.000	33.081 ± 3.608
		ESE	0.076	0.277	0.256	0.310	0.339	0.309	0.000	
		% Sel	100.0	74.8	73.3	73.3	64.4	65.2	2.7	

PQL: penalized quasi-likelihood; ESE: empirical standard error.

rPQL method because the `rpql` package provided the flexibility to specify the penalty weight. In this section, we reported results of $\varphi_f = \varphi_r$=1 for the CD-J and CD-S methods, and $\varphi_f = \varphi_r$=4 for the rPQL method because of the better performance.

Scenario 1. For LMM, we showed the performance of the proposed CD-J and CD-S methods, and compared them with the rPQL method and Bondell's method. For the Bondell's method, we did not show the intercept estimates because they were not readily provided. Also, we did not continue with Bondell's method for $(n, m) = (500, 6)$ because of being unable to complete the estimation process after 48 hours for a single try. For fixed effect selection using the CD-J and CD-S methods (Table 1), results were vey similar. The estimates in both $\hat{\beta}_{\rho\tau}$ and $\hat{\beta}^s_\rho$ were very close to the true values. The types of covariates (e.g. continuous vs. binary, symmetric vs. right-skewed, and interactions) had little impact on the biasedness of the estimates. The ESE, as expected, decreased with n. The coverage probability of 95% CI for the proposed CD-J and CD-S was generally close to the nominal 95% level. For the performance of the variable selection, the selection of true covariates by the CD methods was close to 100%. The noise covariates were selected but at a very low rate, especially for large n. For the rPQL and Bondell's methods, we noted that the first 3 fixed effect estimates in $\hat{\beta}^{rPQL}$ (rPQL) and $\hat{\beta}^B$ (Bondell's method), whose corresponding covariates are associated with random effects, showed some bias from the true values and low selection rate when the number of clusters n is moderately small (e.g. $n = 30, 60$), but the bias and selection rate improved as n increased. For the random effect selections (Table 2), the parameter estimates of the proposed CD methods were generally close to the true values; the ESE decreases with n, and the coverage probability of 95% CI is close to but slightly under the nominal 95% level. The selection rate of true covariates was nearly 100%. The noise covariates were selected but at a very lower rate. The rPQL and Bondell's methods were similar, too. In terms of the computation time, the CD methods generally took less than ($\ll$) 1 minute, while the other two methods can take much longer, especially when n is large (say, $n = 500$). When n is moderate to large, the proposed CD methods can be an attractive and competitive approach.

Scenario 2. For the random effects logistic regression models, we compared the CD methods with rPQL and glmmLasso[11] for the fixed effect selection (Table 3). Note that glmmLasso only performs the fixed effect selection. When we used glmmLasso, we included the correct random-effect covariates in the models. For the CD methods, there was some bias in the fixed effect estimates when n is small (e.g. $n = 30$), but the bias and ESE decreased with n. When n is 80 or more, the bias becomes minimal. The coverage probability was lower than the nominal 95% level for small n (e.g. $n = 30$), largely due to the bias of $\hat{\boldsymbol{\beta}}_{\rho\tau}$ and $\hat{\boldsymbol{\beta}}^s_\rho$, but improved as n increased. The performance of CD-S was slightly better than the CD-J, with a slightly lower false selection rate and better coverage rate of the 95% CI. For both CD methods, the computation time was primarily spent in deriving $\hat{\theta}$ and $\hat{\Sigma}$. The time to obtain $\hat{\beta}_{\rho\tau}$, $\hat{\beta}^s_\rho$, $\hat{\gamma}_{\rho\tau}$, and $\hat{\gamma}^s_\rho$ in the regularized estimation process was minimal ($\ll$ 1 minute). Compared to rPQL and glmmLasso, the bias of the estimates, the selection rate and computation time, the CD methods were obviously better. For the random effect selection (Table 4), results were similar. For the CD methods, there was some bias in the random effect estimates when n is small (e.g. $n = 30$). When n increased to 80 or more, the bias became minimal. The selection rate was lower when n was small which resulted in lower coverage rate of the 95% CIs. We also noticed that the selection rate of the random effect corresponding to $x_{ij,2}$ was low. It might be because $x_{ij,2}$ follows the exponential distribution with mean 1 and is right-skewed. After increasing m (e.g. $(n, m) = (200, 6)$ increased to $(n, m) = (200, 10)$) and/or increasing n, the selection rate and the probability coverage rate of the 95% CIs improved. Compared to rPQL, the bias of the estimates and the selection rate of the CD methods were better.

Table 4. Random effect logistic regression model: Random effect selection.

(n, m)	Method	γ_0	1.73	0.69	1.23	0.46	0.15	0.88	$\mathbf{0}_4$
(30, 10)	CD-J	$\hat{\gamma}_{\rho\tau}$	2.181	0.662	1.139	0.324	0.018	0.339	0.003
		ESE	1.284	1.023	1.657	0.738	0.549	1.472	0.297
		CovP(%)	79.9	81.9	73.1	27.5	23.9	27.9	-
		% Sel	100.0	86.0	86.0	32.7	32.7	32.7	1.6
	CD-S	$\hat{\gamma}^s_\tau$	2.181	0.770	1.375	0.476	0.074	0.493	0.003
		ESE	1.284	1.047	1.629	0.839	0.662	1.507	0.297
		CovP(%)	89.0	88.5	91.2	44.5	40.4	45.4	-
		% Sel	100.0	96.2	96.2	50.5	50.5	49.2	1.6
	rPQL	$\hat{\gamma}^{rPQL}$	1.127	0.421	0.651	0.300	0.126	0.261	0.000
		ESE	0.070	0.105	0.133	0.132	0.194	0.170	0.014
		% Sel	90.0	86.6	87.7	69.4	68.5	70.8	60.1
(80, 10)	CD-J	$\hat{\gamma}_{\rho\tau}$	1.761	0.617	1.037	0.316	0.060	0.455	0.000
		ESE	0.293	0.289	0.299	0.315	0.212	0.403	0.000
		CovP(%)	87.9	91.4	77.4	61.8	61.3	60.5	-
		% Sel	100.0	99.6	99.6	63.6	63.6	63.6	0.0
	CD-S	$\hat{\gamma}^s_\tau$	1.761	0.677	1.145	0.417	0.096	0.632	0.000
		ESE	0.293	0.301	0.288	0.340	0.271	0.419	0.000
		CovP(%)	92.9	94.0	88.6	74.7	73.2	76.8	-
		% Sel	100.0	100.0	100.0	78.1	78.1	78.0	0.0
	rPQL	$\hat{\gamma}^{rPQL}$	1.108	0.416	0.597	0.291	0.091	0.235	0.006
		ESE	0.180	0.211	0.212	0.222	0.281	0.220	0.156
		% Sel	99.3	97.3	97.3	84.3	84.0	84.3	41.4
(200, 6)	CD-J	$\hat{\gamma}_{\rho\tau}$	1.740	0.636	1.031	0.356	0.067	0.507	0.000
		ESE	0.250	0.242	0.287	0.315	0.205	0.407	0.006
		CovP(%)	83.2	89.3	69.5	63.4	66.3	65.5	-
		% Sel	100.0	99.9	99.9	68.7	68.7	68.7	0.1
	CD-S	$\hat{\gamma}^s_\tau$	1.740	0.695	1.127	0.456	0.103	0.666	0.000
		ESE	0.250	0.248	0.278	0.326	0.257	0.417	0.006
		CovP(%)	92.0	93.3	85.6	76.3	75.8	78.4	-
		% Sel	100.0	99.9	99.9	81.0	81.0	81.0	0.1
	rPQL	$\hat{\gamma}^{rPQL}$	1.108	0.416	0.597	0.291	0.091	0.235	0.006
		ESE	0.180	0.211	0.212	0.222	0.281	0.220	0.156
		% Sel	82.0	74.5	76.5	53.5	53.0	53.5	11.4
(200, 10)	CD-J	$\hat{\gamma}_{\rho\tau}$	1.714	0.648	1.117	0.398	0.109	0.672	0.000
		ESE	0.166	0.174	0.173	0.193	0.156	0.235	0.000
		CovP(%)	89.3	92.2	78.5	85.5	91.6	72.9	-
		% Sel	100.0	100.0	100.0	95.2	95.2	95.2	0.0
	CD-S	$\hat{\gamma}^s_\tau$	1.714	0.674	1.167	0.452	0.128	0.779	0.000
		ESE	0.166	0.177	0.169	0.190	0.178	0.201	0.000
		CovP(%)	93.8	93.9	89.2	92.7	92.7	90.9	-
		% Sel	100.0	100.0	100.0	99.4	99.4	99.4	0.0
	rPQL	$\hat{\gamma}^{rPQL}$	1.108	0.416	0.597	0.291	0.091	0.235	0.006
		ESE	0.180	0.211	0.212	0.222	0.281	0.220	0.156
		% Sel	99.9	99.4	99.4	85.4	85.7	85.8	20.0
(500, 6)	CD-J	$\hat{\gamma}_{\rho\tau}$	1.714	0.648	1.117	0.398	0.109	0.672	0.000
		ESE	0.166	0.174	0.173	0.193	0.156	0.235	0.000
		CovP(%)	89.3	92.2	78.5	85.5	91.6	72.9	-
		% Sel	100.0	100.0	100.0	95.2	95.2	95.2	0.0
	CD-S	$\hat{\gamma}^s_\tau$	1.714	0.674	1.167	0.452	0.128	0.779	0.000
		ESE	0.166	0.177	0.169	0.190	0.178	0.201	0.000
		CovP(%)	93.8	93.9	89.2	92.7	92.7	90.9	-
		% Sel	100.0	100.0	100.0	99.4	99.4	99.4	0.0
	rPQL	$\hat{\gamma}^{rPQL}$	0.985	0.384	0.545	0.245	0.056	0.174	0.000
		ESE	0.093	0.134	0.156	0.242	0.230	0.130	0.000
		% Sel	82.4	71.8	72.9	57.6	56.5	57.6	1.2

PQL: penalized quasi-likelihood; ESE: empirical standard error. [a] *CovP*(%) calculated after excluding incorrect non-selections (0's

Table 5. Random effects poisson regression model: Fixed effect and random effect selections.

Fixed effects	Method	β_0	1	−1	−1	−1	−1	−1	$\mathbf{0}_{10}$	Time (mins)
(30, 10)	CD-J	$\hat{\beta}_{\rho\tau}$	0.990	−1.020	−0.989	−0.999	−0.988	−0.986	0.001	0.113 ± 0.090
		ESE	0.313	0.292	0.209	0.028	0.098	0.079	0.012	
		CovP(%)	94.6	90.3	91.3	86.0	71.9	76.3	-	
		% Sel	100.0	100.0	100.0	100.0	100.0	100.0	5.8	
	CD-S	$\hat{\beta}^s_\rho$	0.990	−1.019	−0.988	−0.998	−0.987	−0.985	0.000	0.113 ± 0.090
		ESE	0.313	0.292	0.209	0.028	0.098	0.079	0.008	
		CovP(%)	94.6	90.6	91.3	86.6	90.6	87.0	-	
		% Sel	100.0	100.0	100.0	100.0	100.0	100.0	1.7	
	rPQL	$\hat{\beta}^{rPQL}$	1.107	−0.964	−0.938	−0.996	−0.989	−1.001	0.000	0.062 ± 0.013
		ESE	0.331	0.308	0.271	0.042	0.108	0.119	0.000	
		% Sel	100.0	97.9	99.1	100.0	99.9	100.0	0.9	
(200, 6)	CD-J	$\hat{\beta}_{\rho\tau}$	0.991	−1.008	−1.000	−1.001	−1.004	−1.001	0.000	0.260 ± 0.197
		ESE	0.135	0.114	0.087	0.014	0.044	0.037	0.004	
		CovP(%)	94.3	93.1	93.9	91.0	77.6	87.8	-	
		% Sel	100.0	100.0	100.0	100.0	100.0	100.0	2.1	
	CD-S	$\hat{\beta}^s_\rho$	0.991	−1.008	−1.000	−1.000	−1.003	−1.001	0.000	0.260 ± 0.196
		ESE	0.135	0.114	0.087	0.014	0.044	0.037	0.003	
		CovP(%)	94.7	93.1	94.7	91.8	95.9	93.9	-	
		% Sel	100.0	100.0	100.0	100.0	100.0	100.0	0.8	
	rPQL	$\hat{\beta}^{rPQL}$	1.000	−1.000	−1.006	−0.998	−0.997	−0.999	0.000	2.971 ± 0.716
		ESE	0.135	0.113	0.100	0.013	0.027	0.030	0.000	
		% Sel	100.0	100.0	100.0	100.0	100.0	100.0	0.2	
(500, 6)	CD-J	$\hat{\beta}_{\rho\tau}$	1.000	−1.000	−0.997	−1.000	−1.003	−1.002	0.000	0.565 ± 0.190
		ESE	0.079	0.064	0.056	0.007	0.029	0.022	0.002	
		CovP(%)	94.9	97.7	92.2	95.9	75.1	85.3	-	
		% Sel	100.0	100.0	100.0	100.0	100.0	100.0	0.7	
	CD-S	$\hat{\beta}^s_\rho$	1.000	−0.999	−0.997	−1.000	−1.003	−1.002	0.000	0.565 ± 0.190
		ESE	0.079	0.064	0.056	0.007	0.029	0.022	0.001	
		CovP(%)	94.9	97.7	92.2	96.3	94.9	91.7	-	
		% Sel	100.0	100.0	100.0	100.0	100.0	100.0	0.2	
	rPQL	$\hat{\beta}^{rPQL}$	1.009	−1.003	−0.999	−0.999	−0.999	−0.998	0.000	43.025 ± 7.323
		ESE	0.082	0.067	0.056	0.016	0.018	0.007	0.000	
		% Sel	100.0	100.0	100.0	100.0	100.0	100.0	100.0	
Random effects		γ_0	1.73	0.69	1.23	0.46	0.15	0.88	$\mathbf{0}_4$	
(30, 10)	CD-J	$\hat{\gamma}_{\rho\tau}$	1.703	0.679	1.136	0.416	0.136	0.722	0.000	
		ESE	0.267	0.257	0.163	0.204	0.185	0.135	0.000	
		CovP(%)	89.6	92.3	84.3	91.0	90.6	73.2	0.0	
		% Sel	100.0	100.0	100.0	100.0	100.0	100.0	-	
	CD-S	$\hat{\gamma}^s_\rho$	1.703	0.693	1.165	0.445	0.147	0.800	0.000	
		ESE	0.267	0.260	0.163	0.212	0.198	0.132	0.000	
		CovP(%)	89.6	94.3	88.3	93.6	89.0	86.0	-	
		% Sel	100.0	100.0	100.0	100.0	100.0	100.0	0.0	
	rPQL	$\hat{\gamma}^{rPQL}$	1.573	0.631	1.143	0.387	0.156	0.761	0.001	
		ESE	0.247	0.276	0.176	0.233	0.197	0.168	0.034	
		% Sel	100.0	100.0	100.0	100.0	100.0	100.0	0.0	
(200, 6)	CD-J	$\hat{\gamma}_{\rho\tau}$	1.727	0.684	1.216	0.453	0.138	0.839	0.000	
		ESE	0.101	0.106	0.074	0.085	0.075	0.063	0.001	
		CovP(%)	95.5	93.9	89.8	95.1	94.7	85.3	-	
		% Sel	100.0	100.0	100.0	100.0	100.0	100.0	0.0	
	CD-S	$\hat{\gamma}^s_\rho$	1.727	0.688	1.224	0.461	0.141	0.859	0.000	
		ESE	0.101	0.107	0.074	0.086	0.077	0.061	0.001	
		CovP(%)	95.5	95.1	90.2	96.7	93.5	89.8	-	
		% Sel	100.0	100.0	100.0	100.0	100.0	100.0	0.0	

(continued)

Table 5. Random effects poisson regression model: Fixed effect and random effect selections.

Fixed effects	Method	β_0	1	−1	−1	−1	−1	−1	$\mathbf{0}_{10}$	Time (mins)
	rPQL	$\hat{\gamma}^{rPQL}$	1.626	0.652	1.191	0.426	0.155	0.842	0.000	
		ESE	0.093	0.106	0.072	0.086	0.078	0.060	0.000	
		% Sel	99.8	99.8	99.8	99.8	99.8	99.8	0.0	
(500, 6)	CD-J	$\hat{\gamma}_{\rho\tau}$	1.716	0.681	1.227	0.446	0.145	0.856	0.000	
		ESE	0.067	0.068	0.047	0.060	0.048	0.040	0.000	
		CovP(%)	92.2	94.0	94.0	90.8	94.0	88.0	-	
		% Sel	100.0	100.0	100.0	100.0	100.0	100.0	0.0	
	CD-S	$\hat{\gamma}^{s}_{\rho}$	1.716	0.683	1.231	0.450	0.147	0.865	0.000	
		ESE	0.067	0.068	0.046	0.061	0.049	0.040	0.000	
		CovP(%)	92.2	94.5	95.4	92.6	94.0	90.8	-	
		% Sel	100.0	100.0	100.0	100.0	100.0	100.0	0.0	
	rPQL	$\hat{\gamma}^{rPQL}$	1.629	0.652	1.200	0.428	0.148	0.848	0.000	
		ESE	0.061	0.065	0.040	0.054	0.047	0.035		
		% Sel	99.5	99.5	99.5	99.5	99.5	99.5	0.0	

PQL: penalized quasi-likelihood; ESE: empirical standard error.

Scenario 3. For the random effects Poisson regression models (Table 5), the fixed effect and random effect estimates of CD-J, CD-S and rPQL methods were very close to the true values. The selection rate of true covariates was close to 100%, and nearly no noise covariates was selected by all three methods. Compared to rPQL, the computation time of the CD methods was shorter, especially when n is large.

6 Applications: Data analysis

We applied the proposed CD approach to the two longitudinal cancer studies that motivated our methods. Regression coefficient estimates were reported and the 95% CIs by the CD methods were calculated based on the oracle properties in Theorems 2 and 3. Moreover, we used φ_f=φ_r=1 in the CD methods, and φ_f=φ_r=4 in the rPQL method, based on the results from simulation studies.

Data Example 1. This study longitudinally measured the tumor volume during the cycles of radiation therapy in 111 patients with unresectable, locally advanced, non-small cell lung cancer (NSCLC). Most patients were treated with concurrent chemoradiation therapy (CRT) as it offers much improved survival outcomes, compared to a sequential combination of chemotherapy followed by radiation therapy.[41] To measure the response to CRT, the cone beam computed tomography, as part of the image guided radiation therapy and known to have higher precision, has been used to measure the tumor volumes. The investigators were interested in knowing which demographic and clinical factors are associated with shrinkage of tumor volume over the treatment cycles. We then fitted the linear mixed model and applied the proposed CD methods.

Data with 777 observations from the 111 NSCLC patients were included in the data analysis. The outcome variable, tumor volume, was log-transformed to ensure normality. Potential covariates considered for the fixed effect selection included Weeks (weeks from the start of radiation cycle), Age (age when radiation therapy started), Gender (women vs. men), smoking (yes vs. no), mean lung dose, and Lung V20. Mean lung dose (MLDGy) measured how much radiation dose the normal lung tissues have received, and Lung V20 (LungV20) was the portion of normal lung volume that received 20 Gy of radiation dose. Weeks and MLDGy were considered for the random effect selection. All continuous variables were standardized to have mean 0 and variance 1. In order to evaluate the performance of the proposed methods , we have randomly generated 2 noise variables following standard normal distribution and included them to the fixed effect selection. One of these noise variables was also included in the random effect selection. Results were summarized in Table 6. For the fixed effect selection by the CD-J method, the randomly generated noise variables were not selected. Weeks, Age, Gender, MLDGy, and LungV20 were selected. Specifically, the selected fixed effect estimates suggested that tumor volume decreased with Weeks ($\hat{\beta}_{\rho\tau} = -0.114$, 95% CI: $[-0.126, -0.102]$), and was larger in men than women ($\hat{\beta}_{\rho\tau} = 0.477$, 95% CI: $[0.010, 0.943]$). The mean tumor size also increased with MLDGy ($\hat{\beta}_{\rho\tau} = 0.689$, 95% CI: $[0.004, 1.373]$) and decreased with LungV20 ($\hat{\beta}_{\rho\tau} = -0.499$, 95% CI: $[-1.177, 0.180]$). Random intercept ($\hat{\Gamma}_{11} = 1.190$, 95% CI: $[1.188, 1.192]$) and random slope for Weeks ($\hat{\Gamma}_{21} = 0.000$, 95% CI: $[0.000, 0.000]$, and $\hat{\Gamma}_{22} = 0.062$, 95% CI: $[0.062, 0.063]$) were selected, suggesting a positive within-person correlation and a heterogeneous effect by Weeks. Results from CD-S were similar. We

Table 6. Linear mixed model analysis of lung cancer data.

	CD-J method		CD-S method		rPQL	Bondell et al.
Fixed effects	$\hat{\beta}_{\rho\tau}$	(95% CI)	$\hat{\beta}^{s}_{\rho}$	(95% CI)	$\hat{\beta}^{rPQL}$	$\hat{\beta}^{B}$
Intercept	3.572	(2.742, 4.403)	3.869	(3.537, 4.202)	3.909	-
Weeks	−0.114	(−0.126, −0.102)	−0.111	(−0.123, −0.099)	−0.110	−0.112
Age (in years)	0.125	(−0.106, 0.356)	0.087	(−0.131, 0.304)	0	0
Gender (men vs. women)	0.477	(0.010, 0.943)	0.588	(0.168, 1.008)	0.592	0.430
Smoking (yes vs. no) Yes	0		0		0	3.948
MLDGy	0.689	(0.004, 1.373)	0.860	(0.238, 1.482)	0.146	0.177
LungV20	−0.499	(−1.177, 0.180)	−0.635	(−1.259, −0.010)	−0.092	0
Noise 1	0		0		0	0
Noise 2	0		0		0	0
Random effects						
	$\hat{\gamma}_{\rho\tau}$	(95% CI)	$\hat{\gamma}^{s}_{\tau}$	(95% CI)	$\hat{\gamma}^{rPQL}$	$\hat{\gamma}^{B}$
Intercept (Γ_{11})	1.190	(1.188, 1.192)	1.244	(1.242 1.246)	1.308	1.521
Weeks (Γ_{21})	0.000	(0.000, 0.000)	0.002	(0.001, 0.002)	0.003	0.004
Weeks (Γ_{22})	0.062	(0.062, 0.063)	0.062	(0.061, 0.063)	0.051	0.060
MLDGy ($\Gamma_{31}, \Gamma_{32}, \Gamma_{33}$)	0		0		0	0
Noise 1 ($\Gamma_{41}, \Gamma_{42}, \Gamma_{43}, \Gamma_{44}$)	0		0		0	0

PQL: penalized quasi-likelihood; CI: confidence interval; MLD: mean lung dose.

also applied the rPQL and Bondell's methods to this dataset. Neither of these methods selected the noise variables. The rPQL didn't select Age in the fixed effect. Bondell's method did not select Age but selected Smoking. For the random effect selection, all four methods selected the same random effects.

Data Example 2. Patients with early-stage breast cancer are commonly treated with breast-conserving therapy (BCT), which includes lumpectomy followed by radiation therapy. Prior studies with long-term follow-up have demonstrated equivalent overall survival in those treated with lumpectomy and radiation, compared with those who underwent mastectomy.[42,43] Because mammographic alterations after BCT can mimic or hide tumor recurrence, they become clinically relevant when unnecessary biopsies or delayed diagnoses occur. Then this study longitudinally followed up the mammographic changes in early-stage breast cancer patients after BCT, and is interested in identifying covariates associated with the incidence or changes of common mammographic sequelae.[44]

Data from 89 patients with a total of 605 longitudinally measured observations were included in the data analysis. Among several image parameters, we fitted a random effects logistic regression model with calcification (yes vs. no) as the dependent variable, and applied the proposed CD-S method to select fixed effects from the following covariates: age (years), years from radiation therapy, African Americans (yes vs. no), Her2/neu positive (yes vs. no), adjuvant chemo and/or hormonal therapy (yes vs. no), smoking (yes vs. no), bilateral disease (yes vs. no) and 2 unrelated and randomly generated standard normal noise variables. Potential covariates for the random effect selection included the intercept, age, and the two unrelated noise variables that were also included in the fixed effect selection. All continuous variables were standardized to have mean 0 and variance 1. Results were summarized in Table 7. For the fixed effect selection, the randomly generated noise variables were not selected. Age, years since radiation therapy, African Americans, Her2/neu positive, and Bilateral Disease were selected. For instance, estimates of the selected fixed effects suggested that the risk of calcification increased with age ($\hat{\beta}_{\rho\tau} = 4.035$, 95% CI: [2.579, 5.492]), years since radiation therapy ($\hat{\beta}_{\rho\tau} = 1.270$, 95% CI: [0.824, 1.717]), and was lower in African Americans ($\hat{\beta}_{\rho\tau} = -4.330$, 95% CI: [−7.199, −1.462]). For random effects, age ($\hat{\Gamma}_{21} = 6.002$, 95% CI: [3.095, 8.910], and $\hat{\Gamma}_{22} = 5.049$, 95% CI: [2.657, 7.442]) was selected, in addition to the random intercept ($\hat{\Gamma}_{11} = 3.374$, 95% CI: [2.125, 4.622]), suggesting a positive within-person correlation and a heterogeneous effect by age. Results from the CD-S method were similar. We applied the rPQL and glmmLasso methods to this dataset. Because glmmLasso only performs the fixed effect selection, we only included the intercept and age as random effects in the model and reported the parameter estimates for reference. The rPQL method selected the same fixed effect covariates as the CD-S method, but did not select any random effects. The method of glmmLasso selected more fixed effect covariates. The magnitude of the selected fixed effect coefficient estimates of these two methods were generally smaller than that by the CD methods, consistent with the observations from the simulation studies.

Table 7. Random effects logistic regression model analysis of calcification data.

Fixed effects	CD-J method		CD-S method		rPQL	glmmLasso
	$\hat{\beta}_{\rho\tau}$	(95% CI)	$\hat{\beta}^s_\rho$	(95% CI)	$\hat{\beta}^{rPQL}$	$\hat{\beta}^g$
Intercept	4.614	(2.482, 6.748)	4.615	(3.118, 6.112)	1.251	1.361
Age	4.035	(2.579, 5.492)	4.022	(2.031, 6.013)	0.442	0.954
Years from radiation therapy	1.270	(0.824, 1.717)	1.259	(0.768, 1.749)	0.446	0.916
African Americans	−4.330	(−7.199, −1.462)	−4.315	(−7.884, −0.745)	−2.054	−2.889
Her2/neu positive	−1.969	(−3.418, −0.520)	−1.877	(−3.336, −0.418)	−0.965	−1.265
Bilateral disease	1.915	(−0.494, 4.325)	1.790	(−0.854, 4.434)	0.778	0.814
Smoking	0		0		0	0.457
Adjuvant therapy	0		0		0	0.502
Noise 1	0		0		0	−0.060
Noise 2	0		0		0	0
Random effects	$\hat{\gamma}_{\rho\tau}$	(95% CI)	$\hat{\gamma}^s_\tau$	(95% CI)	$\hat{\gamma}^{rPQL}$	$\hat{\gamma}^g$
Intercept (Γ_{11})	3.374	(2.125, 4.622)	3.184	(1.488, 4.879)	0	2.160
Age (Γ_{21})	6.002	(3.095, 8.910)	5.955	(0.935, 10.975)	0	0.117
Age (Γ_{22})	5.049	(2.657, 7.442)	5.022	(1.914, 8.129)	0	1.385
Noise 1 ($\Gamma_{31}, \Gamma_{32}, \Gamma_{33}$)	0		0		0	0
Noise 2 ($\Gamma_{41}, \Gamma_{42}, \Gamma_{43}, \Gamma_{44}$)	0		0		0	0

PQL: penalized quasi-likelihood; CI: confidence interval.

7 Discussion

In this paper, we propose a regularized estimation approach using the confidence distributions of model parameters to select both fixed and random effects in GLMMs. Specifically, we propose to construct and optimize the objective functions based on the confidence distributions of model parameters, as opposed to the objective functions typically constructed from the likelihood function using the observed data. This can greatly alleviate the computational burden in approximating the integral in the log marginal likelihood function in the regularization step.

We started from showing 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 MLEs. As a result, we propose the CD-joint estimation method by constructing the objective functions using the confidence density, as opposed to the observed data likelihood function. Because the joint confidence density of the fixed effect and random effect parameters is a multivariate normal density, the parameters of the joint distribution and the marginal distribution are the same. Therefore, we propose the CD-separate estimation method by constructing the objective functions based on the marginal confidence density corresponding to the fixed effect and random effect parameters, respectively. In spite that the CD-joint estimators and CD-separate estimators are different estimators, the true values of the underlying parameters of these estimators are the same. With a proper choice of regularization parameters in the adaptive LASSO framework, we show that the proposed estimators have consistency and oracle properties. Based on the asymptotic properties of the proposed estimators and our simulation studies, when the sample size (i.e. the number of independent cluster, n) is large, our methods generally performed well and do not require to refit a GLMM model after the variable selection step. However, when n is small, we noticed that our methods may result in bigger bias and false selection rate (e.g. $n = 30$ and $m = 10$ in the simulation studies of random effect logistic regression analysis). Thus, similar to Hui et al.,[19] we recommend the hybrid estimation approach, e.g. refitting a GLMM model using REML after the variable selection step, to improve the finite sample performance. Moreover, because GLMM typically requires the assumption that the random effects follow a multivariate normal distribution with mean 0, we applied this assumption in our simulation studies to assess the performance of the proposed methods. In case this assumption is violated and the random effects are following a right-skewed distribution with non-zero means, it is likely to cause biased GLMM estimates and affect the asymptotic distributions of the estimators. Because the proposed CD methods are built upon the asymptotic distributions of the GLMM estimators, we expect the performance of our methods will be negatively affected due to the deviation from this assumption.

The proposed CD methods require obtaining the GLMM estimates in order to apply the confidence distribution approach and perform the variable selections, while other methods, e.g. the Bondell's[6] method, may not require this condition, and some may only need good initial estimates to implement their methods, e.g. the rPQL[19] method. To implement our methods in practice, one may construct the proposed objective functions by using existing software packages (e.g. `lme4` package in R and `Proc Glimmix` in SAS) that provide readily available solutions of $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\Sigma}}$. Then optimizing the proposed

objective functions to obtain the regularized estimators no longer involves the numerical approximation of the integral in the log marginal likelihood function, and thus boosts computational efficiency. Moreover, optimizing the proposed objective functions in the regularized estimation process can also use existing software packages without the need to develop new computational algorithms specific to GLMM.

When the data are independent, the proposed CD-J method generally does not work due to the singularity of the joint variance-covariance of the fixed and random effect parameter estimates. For the proposed CD-S method, the fixed effect selection still works well, but the random effect selection does not, due to the same singularity issue of the variance-covariance of the random effect estimates. Therefore, we recommend to apply our methods when the dependent variables are correlated and the GLMM analysis can produce meaningful random effect estimates.

Due to the asymptotic normality property of the MLE, we notice that the proposed CD-based objective functions take the similar forms to the objective function based on the LSA proposed by Wang and Leng[32] for the generalized linear models. Simulation studies demonstrate the consistency, oracle properties and computational efficiency, especially when the number of independent clusters n is large. The coverage probability of the 95% CI seemed to be lower than the nominal 95% level in some cases. Replacing $\hat{\boldsymbol{\Sigma}}$ by other consistent estimators of $\boldsymbol{\Sigma}$, for instance, the robust sandwich variance estimator[45,46] might improve the coverage probability.

For future work, we will extend our method to the framework of generalized estimating equation approach[47] for correlated outcome data. The extension to other likelihood-based approaches for complex modeling, such as the joint analysis of survival and longitudinal data analysis, will also be explored.

Acknowledgment

Drs Lu and Kim contributed equally to this paper. The authors would like to thank Dr Steven Feigenberg, Professor of Radiation Oncology at the Hospital of the University of Pennsylvania, for providing partial data in the lung cancer study example.

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The research of SL, and SJ was partially supported by NIH/NCI CCSG Grant 3P30CA072720. JQC was partially supported by NSF CCF-1909963 and NSF CNS-2120350.

ORCID iD

Shou-En Lu https://orcid.org/0000-0002-0916-1842

Supplemental material

Supplemental material for this article is available online.

References

1. Keselman HJ, Algina J, Kowalchuk RK, et al. A comparison of two approaches for selecting covariance structures in the analysis of repeated measurements. *Commun Stat–Simul Comput* 1998; **27**: 591–604.
2. Vaida F and Blanchard S. Conditional Akaike information for mixed-effects models. *Biometrika* 2005; **92**: 351–370.
3. Gurka MJ. Selecting the best linear mixed model under REML. *Am Stat* 2006; **60**: 19–26.
4. Ibrahim JG, Zhu H and Tang N. Model selection criteria for missing-data problems using the EM algorithm. *J Am Stat Assoc* 2008; **103**: 1648–1658.
5. Liang H, Wu H and Zou G. A note on conditional AIC for linear mixed effects-models. *Biometrika* 2008; **95**: 773–778.
6. Bondell HD, Krishna A and Ghosh SK. Joint variable selection for fixed and random effects in linear mixed-effects models. *Biometrics* 2010; **66**: 1069–1077.
7. Peng H and Lu Y. Model selection in linear mixed effect models. *J Multivar Anal* 2012; **109**: 109–129.
8. Lin B, Pang Z and Jiang J. Fixed and random effects selection by REML and pathwise coordinate optimization. *J Comput Graph Stat* 2013; **22**: 341–355.
9. Pan J and Shang J. Adaptive LASSO for linear mixed model selection via profile log-likelihood. *Commun Stat - Theory Method* 2018; **47**: 1882–1900.
10. Ibrahim JG, Zhu H, Garcia RI, et al. Fixed and random effects selection in mixed effects models. *Biometrics* 2011; **67**: 495–503.
11. Groll A and Tutz G. Variable selection for generalized linear mixed models by L1-Penalized estimation. *Stat Comput* 2014; **24**: 137–154.

12. Schelldorfer J, Meier L and Bühlmann P. Glmmlasso: an algorithm for high-dimensional generalized linear mixed models using L1-Penalization. *J Comput Graph Stat* 2014; **23**: 460–477.
13. Pan J and Huang C. Random effects selection in generalized linear mixed models via ahrinkage penalty function. *Stat Comput* 2014; **24**: 725–738.
14. Wolfinger R. Laplace's approximation for nonlinear mixed models. *Biometrika* 1993; **80**: 791–795.
15. Pinheiro JC and Bates DM. Approximations to the log-likelihood function in the nonlinear mixed-effects model. *J Comput Graph Stat* 1995; **4**: 12–35.
16. Westfall PH. Multiple testing of general contrasts using Llogical constraints and correlations. *J Am Stat Assoc* 1997; **92**: 299–306.
17. Lange K. *Numerical Analysis for Statisticians*. New York: Springer-Verlag, 1999.
18. Pinheiro JC and Chao EC. Efficient Laplacian and adaptive Gaussian quadrature algorithms for multilevel generalized linear mixed models. *J Comput Graph Stat* 2006; **15**: 58–81.
19. Hui FKC, Müeller S and Welsh AH. Joint selection in mixed models using regularized PQL. *J Am Stat Assoc* 2017; **112**: 1323–1333.
20. Singh K, Xie M and Strawderman WE. Confidence distribution (CD) distribution estimator of a parameter. In *Complex datasets and inverse problems*. Institute of Mathematical Statistics, 2007. pp. 132–150.
21. Bayes T. An essay towards solving a problem in the doctrine of chances. *Phil Trans Roy Soc* 1763; **53**: 370–418; Reprinted in *Biometrika*, 45 (1958), 293–315.
22. Fisher RA. On the mathematical foundations of theoretical statistics. *Philos Trans R Soc Lond A* 1922; **222**: 309–368.
23. Xie M, Singh K and Strawderman WE. Confidence distributions and a unifying framework for meta-analysis. *J Am Stat Assoc* 2011; **106**: 320–333.
24. Liu D, Liu RY and Xie M. Multivariate meta-analysis of heterogeneous studies using only summary statistics: efficiency and robustness. *J Am Stat Assoc* 2015; **110**: 326–340.
25. Tian L, Wang R, Cai T, et al. The highest confidence density region and its usage for joint inferences about constrained parameters. *Biometrics* 2011; **67**: 604–610.
26. Wang W, Lu S-E, Cheng JQ, et al. Multivariate survival analysis in big data: a divide-and-combine approach. *Biometrics* 2021; 2021, Apr 13. DOI: 10.1111/biom.13469.
27. Cox DR. Discussion. *Int Stat Rev* 2013; **81**: 40–41.
28. Vonesh EF. Conditional second-order generalized estimating equations for generalized linear and nonlinear mixed-effects models. *Biometrika* 1996; **83**: 447–452.
29. Pan Z and Lin DY. Goodness-of-fit methods for generalized linear mixed models. *Biometrics* 2005; **61**: 1000–1009.
30. Efron B. Bayes and likelihood calculations from confidence intervals. *Biometrika* 1993; **80**: 3–26.
31. Xie M and Singh K. Confidence distribution, the frequentist distribution estimator of a parameter: a review. *Int Stat Rev* 2013; **81**: 3–39.
32. Wang H and Leng C. Unified LASSO estimation by least squares approximation. *J Am Stat Assoc* 2007; **102**: 1039–1048.
33. Zou H. The adaptive lasso and its oracle properties. *J Am Stat Assoc* 2006; **101**: 1418–1429.
34. Fan J and Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *J Am Stat Assoc* 2001; **96**: 1348–1360.
35. He Z, Tu W, Wang S, et al. Simultaneous variable selection for joint models of longitudinal and survival outcomes. *Biometrics* 2015; **71**: 178–187.
36. Zhang HH and Lu W. Adaptive Lasso for Cox's proportional hazards model. *Biometrika* 2007; **94**: 691–703.
37. Friedman J, Hastie T and Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw* 2010; **33**: 1–22.
38. Yang Y and Zou H. A fast unified algorithm for computing Group-Lasso penalized learning problems. *Stat Comput* 2015; **25**: 1129–1141.
39. Hui FKC. *R package version 0.8*. 2020.
40. Wang H and Leng C. A note on adaptive group lasso. *Comput Stat Data Anal* 2008; **52**: 5277–5286.
41. Jabbour SK, Kim S, Haider SA, Xu X, et al. Reduction in tumor volume by Cone Beam computed tomography predicts overall survival in non-small cell lung cancer treated with chemoradiation therapy. *Int J Radiat Oncol Biol Phys* 2015; **92**: 627–633.
42. Fisher B, Anderson S, Bryant J, et al. Twenty-year follow-up of a randomized trial comparing total mastectomy, lumpectomy, and lumpectomy plus irradiation for the treatment of invasive breast cancer. *N Engl J Med* 2002; **347**: 1233–1241.
43. Veronesi U, Cascinelli N, Mariani L, et al. Twenty-year follow-up of a randomized study comparing breast-conserving surgery with radical mastectomy for early breast cancer. *N Engl J Med* 2002; **347**: 1227–1232.
44. Tian S, Paster LF, Kim S, et al. Comparison of mammographic changes across three different fractionation schedules for early-stage breast cancer. *Int J Radiat Oncol Biol Phys* 2016; **95**: 597–604.
45. Freedman DA. On the So-Called "Huber Sandwich Estimator" and "Robust Standard Errors". *Am Stat* 2006; **60**: 299–302.
46. Wang T and Merkle EC. merDeriv: derivative computations for linear mixed effects models with application to robust standard errors. *J Stat Softw* Code Snippets, 2018; **87**: 1–16.
47. Liang KY and Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika* 1986; **73**: 13–22.

Appendix A: Proof of Theorem 1

(1) Since the objective function $Q(\boldsymbol{\theta})$ is strictly convex in $\boldsymbol{\theta}$, a local consistent minimizer is the global consistent minimizer. Therefore, it suffices to show the existence of a local consistent minimizer, then the estimation consistency follows immediately. Following Fan and Li[34] and letting $\mathbf{u} = (u_0, u_1, u_2, \ldots, u_{p_f}, v_1, \mathbf{v}_2^T, \ldots, \mathbf{v}_{p_r+1}^T)^T$, where u_l's and v_1 are scalar, for $l = 0, 1, 2, \ldots, p_f$, and $\mathbf{v}_m^T$'s are vectors of length m, for $m = 2, 3, \ldots, p_r + 1$. Let $p = \text{length}(\boldsymbol{\theta})$. 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

$$\lim_{n\to\infty} P\left\{ \inf_{\mathbf{u}\in\mathbf{R}^p:||\mathbf{u}||=C} Q(\boldsymbol{\theta}_0 + n^{-1/2}\mathbf{u}) > Q(\boldsymbol{\theta}_0) \right\} > 1 - \epsilon \tag{8}$$

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

To show this, consider

$$\begin{aligned}
&Q(\boldsymbol{\theta}_0 + n^{-1/2}\mathbf{u}) - Q(\boldsymbol{\theta}_0) \\
&= \mathbf{u}^T\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{u} + 2\mathbf{u}^T\hat{\boldsymbol{\Sigma}}^{-1}\left\{n^{1/2}(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}})\right\} + n\sum_{f=1}^{p_f}\rho_f(|\beta_{0f} + n^{-1/2}u_f| - |\beta_{0f}|) \\
&\quad + n\sum_{m=2}^{p_r+1}\tau_m(||\boldsymbol{\gamma}_{0m} + n^{-1/2}\mathbf{v}_m|| - ||\boldsymbol{\gamma}_{0m}||) \\
&\geq \mathbf{u}^T\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{u} + 2\mathbf{u}^T\hat{\boldsymbol{\Sigma}}^{-1}\left\{n^{1/2}(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}})\right\} + n\sum_{\{f:\beta_{0f}\neq 0\}}\rho_f(|\beta_{0f} + n^{-1/2}u_f| - |\beta_{0f}|) \\
&\quad + n\sum_{\{m:\boldsymbol{\gamma}_{0m}\neq 0\}}\tau_m(||\boldsymbol{\gamma}_{0m} + n^{-1/2}\mathbf{v}_m|| - ||\boldsymbol{\gamma}_{0m}||) \\
&\geq \mathbf{u}^T\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{u} + 2\mathbf{u}^T\hat{\boldsymbol{\Sigma}}^{-1}\left\{n^{1/2}(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}})\right\} - n\sum_{\{f:\beta_{0f}\neq 0\}}\rho_f|n^{-1/2}u_f| \\
&\quad - n\sum_{\{m:\boldsymbol{\gamma}_{0m}\neq 0\}}\tau_m||n^{-1/2}\mathbf{v}_m|| \\
&\geq \mathbf{u}^T\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{u} + 2\mathbf{u}^T\hat{\boldsymbol{\Sigma}}^{-1}\left\{n^{1/2}(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}})\right\} - (n^{1/2}f_0 a_{f,n} + n^{1/2}r_0 a_{r,n})||\mathbf{u}||
\end{aligned} \tag{9}$$

followed by $\boldsymbol{\beta}_{0b} = \mathbf{0}$, $\boldsymbol{\gamma}_{0b} = \mathbf{0}$, the triangle inequality, $a_{f,n} = max\{\rho_j, j \leq f_0\}$ and $a_{r,n} = max\{\tau_j, j \leq r_0\}$. 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 (9) is $o_p(1)$. Because $\hat{\boldsymbol{\theta}}$ and $\hat{\Sigma}$ are consistent estimators of $\boldsymbol{\theta}$ and Σ, respectively, the second term in (9) 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 $\boldsymbol{\Sigma}$ and its estimate $\hat{\boldsymbol{\Sigma}}$ are positive semidefinite, the first term in (9) is larger than $\mu_{min}(\hat{\boldsymbol{\Sigma}}^{-1})C^2 \xrightarrow{P} \mu_{min}(\boldsymbol{\Sigma}^{-1})C^2$, where $\mu_{min}(.)$ refers to the minimal eigenvalue. It follows that, with probability going to 1, the first term in (9) is larger than $\mu_{min}(\boldsymbol{\Sigma}^{-1})C^2$ 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, (8) holds and the proof of estimation consistency is completed.

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

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

where $\hat{\boldsymbol{\Sigma}}_{(\beta_j)}^{-1}$ represents the row vector of $\hat{\boldsymbol{\Sigma}}^{-1}$ corresponding to the position of β_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 $n^{1/2}\rho_j \geq n^{1/2}b_{f,n} \xrightarrow{P} \infty$. 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 $f_0 < j \leq p_f$. Similarly, suppose $\hat{\boldsymbol{\gamma}}_{\rho\tau,m} \neq 0$ for some $r_0 < m \leq p_r + 1$, then by definition

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

where $\hat{\boldsymbol{\Sigma}}^{-1}_{(\boldsymbol{\gamma}_m)}$ represents the submatrix consisting of of the row vectors of $\hat{\boldsymbol{\Sigma}}^{-1}$ corresponding to the position of $\boldsymbol{\gamma}_m$. It can be shown that the first term on the right hand side of (11) is $O_p(1)$. Based on the condition $n^{1/2}b_{r,n} \xrightarrow{P} \infty$, we have $n^{1/2}\tau_m \geq n^{1/2}b_{r,n} \xrightarrow{P} \infty$. Then to satisfy (11), 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 $r_0 < m \leq p_r + 1$. This completes the proof of selection consistency.

(3) To prove the oracle property, we first ease the notation within the scope of this proof, by re-arranging $\boldsymbol{\theta}_0$, $\boldsymbol{\theta}$, and $\hat{\boldsymbol{\theta}}_{\rho\tau}$, according to the order of $\boldsymbol{\theta}_0 = (\boldsymbol{\theta}_{0a}^T, \boldsymbol{\theta}_{0b}^T = \mathbf{0}^T)^T$, such that $\boldsymbol{\theta} = (\boldsymbol{\theta}_a^T, \boldsymbol{\theta}_b^T)^T$ and $\hat{\boldsymbol{\theta}}_{\rho\tau} = (\hat{\boldsymbol{\theta}}_{\rho\tau a}^T, \hat{\boldsymbol{\theta}}_{\rho\tau b}^T)^T$. Similarly, we use $\boldsymbol{\Sigma}$ and $\hat{\boldsymbol{\Sigma}}$ to represent the re-arranged matrices according to the order of parameters in $\boldsymbol{\theta}_0$. Moreover, we decompose $\boldsymbol{\Sigma}$ and $(\boldsymbol{\Sigma})^{-1}$ into block matrices as:

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

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

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

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

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

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

$$n^{1/2}(\hat{\boldsymbol{\theta}}_{\rho\tau,a} - \boldsymbol{\theta}_{0a}) = n^{1/2}(\hat{\boldsymbol{\theta}}_a - \boldsymbol{\theta}_{0a}) + \left(\hat{\boldsymbol{\Omega}}_{aa}\right)^{-1}\hat{\boldsymbol{\Omega}}_{ab}\hat{\boldsymbol{\theta}}_b - 1/2\left(\hat{\boldsymbol{\Omega}}_{aa}\right)^{-1}\boldsymbol{D}(\hat{\boldsymbol{\theta}}_{\rho\tau,a}) \tag{13}$$

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

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

Given that

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

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

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

Providing the fact that

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

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

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