Sankhyā : The Indian Journal of Statistics
2018, Volume 80-A, Part 2, pp. 385-410

Efficient Shrinkage for Generalized Linear Mixed Models Under Linear Restrictions

T. Thomson
University of Winnipeg, Winnipeg, Canada
S. Hossain
University of Winnipeg, Winnipeg, Canada

Abstract

In this paper, we consider the pretest, shrinkage, and penalty estimation procedures for generalized linear mixed models when it is conjectured that some of the regression parameters are restricted to a linear subspace. We develop the statistical properties of the pretest and shrinkage estimation methods, which include asymptotic distributional biases and risks. We show that the pretest and shrinkage estimators have a significantly higher relative efficiency than the classical estimator. Furthermore, we consider the penalty estimator LASSO (Least Absolute Shrinkage and Selection Operator), and numerically compare its relative performance with that of the other estimators. A series of Monte Carlo simulation experiments are conducted with different combinations of inactive predictors, and the performance of each estimator is evaluated in terms of the simulated mean squared error. The study shows that the shrinkage and pretest estimators are comparable to the LASSO estimator when the number of inactive predictors in the model is relatively large. The estimators under consideration are applied to a real data set to illustrate the usefulness of the procedures in practice.

AMS (2000) *subject classification.* Primary 62J07; Secondary 62F03, 62F10, 62J12.
Keywords and phrases. Asymptotic distributional bias and risk, Generalized linear mixed models, LASSO, Likelihood ratio test, Monte Carlo simulation, Shrinkage and pretest estimators

1 Introduction

Longitudinal data, also known as panel data, involving binary or count responses is a frequent occurrence in many fields such as biology, economics, and health research. For example, physicians may investigate the presence of a particular drug side-effect which can be modelled as a binary outcome (1 - yes; 0 - no) with factors pertaining to the patient's diet and lifestyle. Typically, the physician would observe the patient's status prior to receiving

the medication, during the treatment period, and shortly after the drug is administered. In this case, analysis with the linear regression model cannot be applied due to the non-normality of the error structure, and the lack of independence among observations.

The generalized linear mixed models (GLMMs) is an extension of the generalized linear models (GLMs) (Nelder and Wedderburn, 1972) that is widely used to model correlated and clustered responses. The basic idea of the GLMMs is that it incorporates the fixed effects formulation of the GLMs, however it extends to include subject-specific random effects to capture the correlation within the data. Specifically, the GLMMs model the conditional distribution of a response variable Y given the $q \times 1$ vector of unobserved random effects $\boldsymbol{u}$, and a $p \times 1$ vector of fixed effect covariates $\boldsymbol{x}$ via the linear combination $\boldsymbol{x}^\top \boldsymbol{\beta} + \boldsymbol{z}^\top \boldsymbol{u}$. Here, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown regression coefficients, and $\boldsymbol{z}$ is a $q \times 1$ vector of random effect covariates. Associated with the random effects is a $\kappa \times 1$ vector $\boldsymbol{\theta}$ of the variance and covariance components. The aim here is to estimate the parameters associated with the fixed and random effects, however we proceed to estimate $\boldsymbol{\beta}$ by using the likelihood procedure, and consider $\boldsymbol{\theta}$ to be a nuisance parameter. The GLMMs require an established relationship between the expected value of Y and $\boldsymbol{x}^\top \boldsymbol{\beta} + \boldsymbol{z}^\top \boldsymbol{u}$. Such a relationship exists when the probability distribution of the response is a member of the exponential family.

In this paper, we propose the shrinkage and pretest estimation approaches to estimate the regression parameters of a GLMM. Basically, we estimate the parameters when there are many potential covariates under investigation. Although we optimally select the active covariates in the model, there are many situations in which over-modelling takes place, and one wishes to reduce the number of active variables in the model. In doing so, emphasis is placed on maximizing the predictive power while minimizing the number of active variables in the model. The James-Stein shrinkage estimation strategy is one such method that allows the researcher to achieve the aforementioned goal because it incorporates information from the inactive covariates when estimating the coefficients of the active covariates. Recently, Thomson et al. (2014) considered shrinkage and penalty estimation strategies in the linear regression model with autoregressive errors. Hossain et al. (2015) introduced shrinkage and penalty estimators in a GLM when there are many active predictors and some of them may not have influence on the response. Thomson et al. (2015) further extended this line of work by using shrinkage estimation for time series following a GLMs. Many extensions of shrinkage estimation and associated theoretical investigations have significantly boosted the

popularityof this approach (Fallahpour et al., 2012; Lian, 2012; Hossain and Ahmed, 2014). It is to the best of the authors' knowledge that this problem remains open.

The aim of this paper is to develop the James-Stein shrinkage estimation, and the LASSO penalized estimation methods, and compare its performance with the maximum likelihood estimate for a GLMM when some of the covariates may be subject to a linear restriction. The literature on penalty estimators has been growing very rapidly in recent years; here we only give a limited number of studies that are the most relevant to the present paper. Tibshirani (1996) introduced the LASSO, which imposes a bound on the L_1 norm of the coefficients. This results in both shrinkage and variable selection due to the nature of the constraint region which often results in several coefficients becoming identically zero. Groll and Tutz (2014) implemented the LASSO estimator to the GLMMs, where they use a gradient ascent algorithm to maximize the penalized log-likelihood. They show that the LASSO estimator can be used in high-dimensional settings with several covariates under consideration; this contrasts with common procedures. For review articles on penalized regression, see Fan and Lv (2010). Recent works involving the LASSO and other forms of penalized estimation with GLMMs include Fahrmeir and Kneib (2011) & Schelldorfer et al. (2014). For recent works involving the LASSO, see Simon et al. (2014), Zhang and Zou (2014), Li and Shao (2015), Lu and Su (2016), & Arnold and Tibshirani (2016).

The outline of this paper is as follows. Section 2 begins with preliminary definitions, and proceeds with estimation strategies. Section 3 examines the asymptotic distributional biases and risks of the various estimators, where the risk comparisons and proofs are presented in the Appendix. In Section 4, we use a Monte Carlo simulation to evaluate the numerical performance of the various estimators with respect to the maximum likelihood estimator. Section 5 follows with an application to a real data example. Section 6 provides a summary with concluding remarks.

2 Generalized Linear Mixed Models and Estimation

In this section, we implement the GLMMs to our proposed estimation strategies. The GLMMs, Fisher information matrix, likelihood ratio test, and various estimators are presented in the following subsections.

For $i = 1, \cdots, N$ and $j = 1, \cdots, n_i$, let y_{ij} denote the response for the ith subject measured at the jth time points, so that $\boldsymbol{y}_i = (y_{i1}, \cdots, y_{in_i})^\top$ denotes the $n_i \times 1$ vector of observations for the ith subject. Corresponding to each $\boldsymbol{y}_i$ is a $p \times 1$ vector, and a $q \times 1$ vector of covariates associated

with fixed and random effects denoted by $\boldsymbol{x}_{ij} = (x_{ij1}, \cdots, x_{ijp})^\top$, and $\boldsymbol{z}_{ij} = (z_{ij1}, \cdots, z_{ijq})^\top$, respectively. Also, let $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)^\top$ be the $p \times 1$ vector of regression coefficients for the fixed effects, and $\boldsymbol{u}_i = (u_{i1}, \cdots, u_{iq})^\top$ be the $q \times 1$ vector of random effects for the ith subject. Furthermore, suppose that conditional on $\boldsymbol{u}_i$, the elements of $\boldsymbol{y}_i$ are independent, and follows a distribution in the exponential family:

$$f_{y_{ij}|\boldsymbol{u}_i}(y_{ij}|\boldsymbol{u}_i, \beta, \phi) = \exp\left(\frac{y_{ij}\psi_{ij} - b(\psi_{ij})}{a_{ij}(\phi)} + c(y_{ij}; \phi)\right), \tag{1}$$

where $a_{ij}(.)$, $b(.)$, and $c(.)$ are known functions with $a_{ij}(\phi) = \frac{\phi}{\omega_{ij}}$, ω_{ij} is known as the *prior weight*, ϕ is a *dispersion* parameter, and ψ_{ij} is the *canonical* parameter of the distribution. We also assume that the vector of unobserved random effects $\boldsymbol{u}_i$ follows a distribution

$$\boldsymbol{u}_i \sim f_{\boldsymbol{u}_i}(\boldsymbol{u}_i|\boldsymbol{\theta}), \tag{2}$$

with parameters $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_\kappa)^\top$, which is not functionally related with the regression parameter $\boldsymbol{\beta}$. In practice, $f_{\boldsymbol{u}_i}(\boldsymbol{u}_i|\boldsymbol{\theta})$ is considered to be multivariate normal with mean $\mathbf{0}$ and variance-covariance matrix comprised of the components in $\boldsymbol{\theta}$ along the main diagonal. However, there are scenarios in which this assumption is violated, and we proceed without any specific distributional assumptions of $\boldsymbol{u}_i$. Furthermore, assume that for some monotone differential link function $g(.)$, the conditional mean and variance of y_{ij} given the random effects is

$$\mu_{ij} = E(y_{ij}|\boldsymbol{u}_i) = b'(\psi_{ij}) = g^{-1}(\eta_{ij}) \quad \text{and} \quad \text{Var}(y_{ij}|\boldsymbol{u}_i) = a_{ij}(\phi)b''(\psi_{ij}), \tag{3}$$

where $\eta_{ij} = \boldsymbol{x}_{ij}^\top\boldsymbol{\beta} + \boldsymbol{z}_{ij}^\top\boldsymbol{u}_i$ is the linear predictor, and $b' = \frac{\partial b}{\partial \psi_{ij}}$.

By setting $g(\mu_{ij}) = \psi_{ij} = (b')^{-1}(\mu_{ij})$, we may find the link function pertaining to the conditional distribution of y_{ij}. A special case of interest is the *canonical link function* $g(\mu_{ij}) = \psi_{ij}(\mu_{ij}) = \eta_{ij}$. This link function includes many popular models for continuous and discrete data. When the distribution of our response in (1) follows a normal distribution, the link function is the identity function so that (3) simplifies to the natural linear mixed model setup. In the following subsequent sections, we will focus our attention on a discrete response. More specifically, when the response is binary or a count of some occurrence.

2.1. Maximum Likelihood Estimation From (1) and (2), the marginal likelihood of the parameters $\boldsymbol{\gamma} = (\boldsymbol{\beta}^\top, \boldsymbol{\theta}^\top, \phi)^\top$, given the vector of responses $\boldsymbol{y} = (\boldsymbol{y}_1^\top, \cdots, \boldsymbol{y}_N^\top)^\top$ is

$$\begin{aligned} L(\boldsymbol{\gamma}|\boldsymbol{y}) &= \prod_{i=1}^{N} \int \prod_{j=1}^{n_i} f_{y_{ij}|\boldsymbol{u}_i}(y_{ij}|\boldsymbol{u}_i, \boldsymbol{\beta}, \phi) f_{\boldsymbol{u}_i}(\boldsymbol{u}_i|\boldsymbol{\theta}) d\boldsymbol{u}_i \\ &= \prod_{i=1}^{N} \int f_{\boldsymbol{y}_i|\boldsymbol{u}_i}(\boldsymbol{y}_i|\boldsymbol{u}_i) f_{\boldsymbol{u}_i}(\boldsymbol{u}_i|\boldsymbol{\theta}) d\boldsymbol{u}_i. \end{aligned} \tag{4}$$

We see that (4) is comprised of the conditional distribution of $\boldsymbol{y}_i$ given $\boldsymbol{u}_i$ and the marginal distribution of $\boldsymbol{u}_i$. This is done to model the joint distribution of $\boldsymbol{y}_i$ and $\boldsymbol{u}_i$, which is unobservable. Although we consider ϕ to be a nuisance parameter, there are many cases such as binary or Poisson regression models where the dispersion parameter is fixed at $\phi = 1$. We therefore proceed as in Davis et al. (2012), and assume that $\phi = 1$, and take $\boldsymbol{\gamma} = (\boldsymbol{\beta}^\top, \boldsymbol{\theta}^\top)^\top$. In cases of counts with over-dispersion, we may use the method of moments estimator (McCulloch et al., 2008, chap. 5).

$$\hat{\phi} = \frac{1}{N^* - p} \sum_{i=1}^{N} \sum_{j=1}^{n_i} \frac{y_{ij} - \mu_{ij}}{b''(\psi_{ij})},$$

where $N^* = \sum_{i=1}^{N} n_i$ is the total number of observations.

To obtain the unrestricted score equations, we make use of the log-likelihood function as

$$\begin{aligned} \ell^*(\boldsymbol{\gamma}|\boldsymbol{y}) &= \log(L(\boldsymbol{\gamma}|\boldsymbol{y})) \\ &= \sum_{i=1}^{N} \log \left(\int f_{\boldsymbol{y}_i|\boldsymbol{u}_i}(\boldsymbol{y}_i|\boldsymbol{u}_i) f_{\boldsymbol{u}_i}(\boldsymbol{u}_i|\boldsymbol{\theta}) d\boldsymbol{u}_i \right). \end{aligned}$$

From this, we solve the corresponding score equations to obtain the unrestricted maximum likelihood estimators of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$. These score equations appear in McCulloch et al. (2008), and are given as

$$S_{\boldsymbol{\beta}}(\boldsymbol{\gamma}) = \sum_{i=1}^{N} E \left(\frac{\partial \log f_{\boldsymbol{y}_i|\boldsymbol{u}_i}(\boldsymbol{y}_i|\boldsymbol{u}_i, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} | \boldsymbol{y}_i \right) = \boldsymbol{0}, \tag{5}$$

$$S_{\boldsymbol{\theta}}(\boldsymbol{\gamma}) = \sum_{i=1}^{N} E \left(\frac{\partial \log f_{\boldsymbol{u}_i}(\boldsymbol{u}_i|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} | \boldsymbol{y}_i \right) = \boldsymbol{0}, \tag{6}$$

where the expectation is taken with respect to the conditional distribution of $\boldsymbol{u}_i$ given $\boldsymbol{y}_i$. These equations cannot be solved explicitly for either $\boldsymbol{\beta}$ or $\boldsymbol{\theta}$, and the integration in (4) often lead to numerically intractable solutions, whenever there are a large number of random effects (i.e. for large q). Therefore, we solve (5) and (6) via numerical optimization to obtain the maximum likelihood estimators. This can be done by using a Gauss-Hermite approximation, or a Laplace approximation (see McCulloch et al., 2008, chap. 14). We denote the unrestricted maximum likelihood estimator (UE) of $\boldsymbol{\gamma}$ as $\hat{\boldsymbol{\gamma}} = (\hat{\boldsymbol{\beta}}_U^\top, \hat{\boldsymbol{\theta}}_U^\top)^\top$. Although we obtained the estimate $\hat{\boldsymbol{\theta}}_U$, we consider it as a nuisance parameter, and primarily focus on $\hat{\boldsymbol{\beta}}_U$.

2.2. Fisher Information Matrix and Restricted Estimator We now proceed to obtain the observed Fisher information matrix as derived by Louis (1982). Let $\ell(\boldsymbol{\gamma}|\boldsymbol{y}) = \sum_{i=1}^N \log(f_{\boldsymbol{y}_i|\boldsymbol{u}_i}(\boldsymbol{y}_i|\boldsymbol{u}_i, \boldsymbol{\beta}, \phi) f_{\boldsymbol{u}_i}(\boldsymbol{u}_i|\boldsymbol{\theta})) = \sum_{i=1}^N \log(f_{\boldsymbol{y}_i|\boldsymbol{u}_i}(\boldsymbol{y}_i|\boldsymbol{u}_i)) + \sum_{i=1}^N \log(f_{\boldsymbol{u}_i}(\boldsymbol{u}_i|\boldsymbol{\theta}))$ be the complete data likelihood. Then the information matrix is

$$\begin{aligned} \boldsymbol{I}(\boldsymbol{\gamma}|\boldsymbol{y}) &= -\sum_{i=1}^N E\left(\frac{\partial t_i(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}^\top}|\boldsymbol{y}_i\right) - \sum_{i=1}^N E\left(t_i(\boldsymbol{\gamma}) t_i(\boldsymbol{\gamma})^\top|\boldsymbol{y}_i\right) \\ &\quad + \sum_{i=1}^N E\left(t_i(\boldsymbol{\gamma})|\boldsymbol{y}_i\right) E\left(t_i(\boldsymbol{\gamma})|\boldsymbol{y}_i\right)^\top, \end{aligned}$$

where $t_i(\boldsymbol{\gamma}) = \frac{\partial \ell(\boldsymbol{\gamma}|\boldsymbol{y})}{\partial \boldsymbol{\gamma}}$, and the expectations are taken with respect to the conditional distribution of $\boldsymbol{u}_i$ given $\boldsymbol{y}_i$. For testing the particular hypothesis on $\boldsymbol{\beta}$, we partition the observed information matrix as

$$\boldsymbol{I}(\boldsymbol{\gamma}|\boldsymbol{y}) = \begin{bmatrix} \boldsymbol{I}(\boldsymbol{\beta}, \boldsymbol{\beta}) & \boldsymbol{I}(\boldsymbol{\beta}, \boldsymbol{\theta}) \\ \boldsymbol{I}(\boldsymbol{\theta}, \boldsymbol{\beta}) & \boldsymbol{I}(\boldsymbol{\theta}, \boldsymbol{\theta}) \end{bmatrix}. \tag{7}$$

We work with such a partition as our inference is centred around $\boldsymbol{\beta}$. Hence we consider the general linear hypothesis

$$H_0 : \boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{c} \quad \text{vs.} \quad H_1 : \boldsymbol{A}\boldsymbol{\beta} \neq \boldsymbol{c}, \tag{8}$$

where $\boldsymbol{A}$ is an $r \times p$ matrix of full row rank, $r \leq p$, and $\boldsymbol{c}$ is an $r \times 1$ vector of known constants. This motivates us to define the restricted parameter space $\Omega = \{(\boldsymbol{\beta}^\top, \boldsymbol{\theta}^\top)^\top | \boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{c}\}$. In order to maximize the log-likelihood function under Ω, we implement a modified version of the gradient projection (GP) algorithm (Jamshidian, 2004) which searches through active constraint sets to determine the optimal solution.

The GLMMs are unlike the case in Jamshidian (2004), so we maximize the marginal likelihood (4) under the equality constraint on the regression parameters. Hence we use the observed information matrix obtained by evaluating the conditional expectations in order to accommodate for the mixed model situation.

Since the parameter vector $\boldsymbol{\gamma}$ is a function of $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$, we take the inverse of the observed information matrix (7)

$$\boldsymbol{I}^{-1}(\boldsymbol{\gamma}|\boldsymbol{y}) = \boldsymbol{W}(\boldsymbol{\gamma}) = \begin{bmatrix} \boldsymbol{W}_{11}(\boldsymbol{\gamma}) & \boldsymbol{W}_{12}(\boldsymbol{\gamma}) \\ \boldsymbol{W}_{21}(\boldsymbol{\gamma}) & \boldsymbol{W}_{22}(\boldsymbol{\gamma}) \end{bmatrix},$$

where

$$\begin{aligned} \boldsymbol{W}_{11}(\boldsymbol{\gamma}) &= \left(\boldsymbol{I}(\boldsymbol{\beta},\boldsymbol{\beta}) - \boldsymbol{I}(\boldsymbol{\beta},\boldsymbol{\theta})\boldsymbol{I}^{-1}(\boldsymbol{\theta},\boldsymbol{\theta})\boldsymbol{I}(\boldsymbol{\theta},\boldsymbol{\beta})\right)^{-1}, \\ \boldsymbol{W}_{22}(\boldsymbol{\gamma}) &= \left(\boldsymbol{I}(\boldsymbol{\theta},\boldsymbol{\theta}) - \boldsymbol{I}(\boldsymbol{\theta},\boldsymbol{\beta})\boldsymbol{I}^{-1}(\boldsymbol{\beta},\boldsymbol{\beta})\boldsymbol{I}(\boldsymbol{\beta},\boldsymbol{\theta})\right)^{-1}, \\ \boldsymbol{W}_{12}(\boldsymbol{\gamma}) &= -\boldsymbol{I}^{-1}(\boldsymbol{\beta},\boldsymbol{\beta})\boldsymbol{I}(\boldsymbol{\beta},\boldsymbol{\theta})\boldsymbol{W}_{22}(\boldsymbol{\gamma}), \end{aligned}$$

and $\boldsymbol{W}_{21}(\boldsymbol{\gamma}) = \boldsymbol{W}_{12}^{\top}(\boldsymbol{\gamma})$. With this, we are able to obtain the generalized score vector

$$S^*(\boldsymbol{\gamma}|\boldsymbol{y}) = \boldsymbol{W}(\boldsymbol{\gamma})\left(S_{\boldsymbol{\beta}}(\boldsymbol{\gamma})^{\top}, S_{\boldsymbol{\theta}}(\boldsymbol{\gamma})^{\top}\right)^{\top} = \left(S_1^*(\boldsymbol{\gamma})^{\top}, S_2^*(\boldsymbol{\gamma})^{\top}\right)^{\top},$$

where $S_1^*(\boldsymbol{\gamma}) = \boldsymbol{W}_{11}(\boldsymbol{\gamma})S_{\boldsymbol{\beta}}(\boldsymbol{\gamma}) + \boldsymbol{W}_{12}(\boldsymbol{\gamma})S_{\boldsymbol{\theta}}(\boldsymbol{\gamma})$, and $S_2^*(\boldsymbol{\gamma}) = \boldsymbol{W}_{21}(\boldsymbol{\gamma})S_{\boldsymbol{\beta}}(\boldsymbol{\gamma}) + \boldsymbol{W}_{22}(\boldsymbol{\gamma})S_{\boldsymbol{\theta}}(\boldsymbol{\gamma})$.

If the UE satisfies the constraint so that $\hat{\boldsymbol{\gamma}} \in \Omega$, then the restricted maximum likelihood estimator (RE) is equivalent to $\hat{\boldsymbol{\gamma}}$. Otherwise, we proceed with the following modified algorithm of Jamshidian (2004) with an initial value $\hat{\boldsymbol{\gamma}}_0$, chosen from Ω. Following the steps below, this algorithm converts the significant constrained set until no change in $\ell^*(\boldsymbol{\gamma}|\boldsymbol{y})$.

1. Compute $\boldsymbol{W}(\hat{\boldsymbol{\gamma}}_0)$ at the initial estimate $\hat{\boldsymbol{\gamma}}_0$ of $\boldsymbol{\gamma}$.

2. Compute $\mathscr{B}(\hat{\boldsymbol{\gamma}}_0) = \boldsymbol{A}^{\top}\left(\boldsymbol{A}\boldsymbol{W}_{11}(\hat{\boldsymbol{\gamma}}_0)\boldsymbol{A}^{\top}\right)^{-1}\boldsymbol{A}$ and $\boldsymbol{d} = (\boldsymbol{d}_1^{\top}, \boldsymbol{d}_2^{\top})^{\top}$, where $\boldsymbol{d}_1 = [I - \boldsymbol{W}_{11}(\hat{\boldsymbol{\gamma}}_0)\mathscr{B}(\hat{\boldsymbol{\gamma}}_0)]S_1^*(\hat{\boldsymbol{\gamma}}_0)$, and $\boldsymbol{d}_2 = -\boldsymbol{W}_{21}(\hat{\boldsymbol{\gamma}}_0)\mathscr{B}(\hat{\boldsymbol{\gamma}}_0)S_1^*(\hat{\boldsymbol{\gamma}}_0) + S_2^*(\hat{\boldsymbol{\gamma}}_0)$.

3. (a) If $\boldsymbol{d} = \boldsymbol{0}$, stop and declare convergence, otherwise

   (b) Go to Step 4.

4. Obtain a new estimate $\hat{\boldsymbol{\gamma}}_1 = \hat{\boldsymbol{\gamma}}_0 + \alpha \boldsymbol{d}$ by choosing $\alpha = \arg\max_\alpha \{\ell^*(\hat{\boldsymbol{\gamma}}_0 + \alpha \boldsymbol{d} | \boldsymbol{y})\}$.

5. Replace $\hat{\boldsymbol{\gamma}}_0$ with $\hat{\boldsymbol{\gamma}}_1$, and go to Step 1.

Jamshidian (2004) observed that the gradient projection is a generalized steepest ascent algorithm, and therefore, standard global convergence theory of steepest ascent algorithms warrants its global convergence (Luenberger and Ye, 2008, chap. 6). We will denote the RE as $\hat{\boldsymbol{\gamma}}_R = (\hat{\boldsymbol{\beta}}_R^\top, \hat{\boldsymbol{\theta}}_R^\top)^\top$.

2.3. Likelihood Ratio Test Based on the estimators $\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\gamma}}_R$, we construct the likelihood ratio test statistic

$$D_N = 2\left(\ell^*(\hat{\boldsymbol{\gamma}}|\boldsymbol{y}) - \ell^*(\hat{\boldsymbol{\gamma}}_R|\boldsymbol{y})\right).$$

It is well known that under H_0 in (8), the statistic D_N asymptotically follows a chi-square distribution with r degrees of freedom, see Davis et al. (2012). Recently, Davis et al. (2012) used the likelihood ratio test statistic to evaluate linear inequality constraints in a GLMM, and assess the asymptotic distribution of this test statistic under such constraints.

Observe that D_N may also be used to test constraints involving the variance components $\boldsymbol{\theta}$, provided that some certain regularity conditions for the observed Fisher information matrix are satisfied. In this case, the hypothesis in (8) can be extended by replacing $\boldsymbol{\beta}$ with $\boldsymbol{\gamma} = (\boldsymbol{\beta}^\top, \boldsymbol{\theta}^\top)^\top$, and by adding κ columns to the matrix $\boldsymbol{A}$.

In the subsequent subsections, the pretest and shrinkage estimators are loosely defined by the likelihood ratio test statistic. The purpose of this statistic is to place weights on the UE and RE in order to maximize the predictive power while under the specified constraint in (8).

2.4. Pretest Estimator The pretest estimator (PT) of $\boldsymbol{\beta}$ based on $\hat{\boldsymbol{\beta}}_U$ and $\hat{\boldsymbol{\beta}}_R$ is defined as

$$\hat{\boldsymbol{\beta}}_P = \hat{\boldsymbol{\beta}}_U - I(D_N \leq \chi^2_{r,\alpha})(\hat{\boldsymbol{\beta}}_U - \hat{\boldsymbol{\beta}}_R), \quad r \geq 1,$$

where $I(A)$ is an indicator function of a set A, and $\chi^2_{r,\alpha}$ is the α-level critical value of the approximate distribution of D_N under H_0. Based on the indicator function, $\hat{\boldsymbol{\beta}}_P$ chooses $\hat{\boldsymbol{\beta}}_R$ or $\hat{\boldsymbol{\beta}}_U$ according to whether H_0 is accepted or rejected, respectively. It is important to note that $\hat{\boldsymbol{\beta}}_P$ is bounded and performs better than $\hat{\boldsymbol{\beta}}_R$ in certain areas of the parameter space. For details, see Judge and Bock (1978) & Ahmed et al. (2006).

2.5. Shrinkage and Positive Shrinkage Estimators The shrinkage estimator (SE), $\hat{\boldsymbol{\beta}}_S$ and the positive shrinkage estimator (PSE), $\hat{\boldsymbol{\beta}}_+$ are members of the class of estimators

$$\hat{\boldsymbol{\pi}}(h) = \hat{\boldsymbol{\beta}}_R + h(D_N)(\hat{\boldsymbol{\beta}}_U - \hat{\boldsymbol{\beta}}_R).$$

By taking $h(x) = 1 - \left(1 - \frac{(r-2)}{x}\right)$, for $x > 0$ and $h(x) = \max\left\{0, \frac{(r-2)}{x}\right\}$, we get $\hat{\boldsymbol{\beta}}_S$ and $\hat{\boldsymbol{\beta}}_+$, respectively. More precisely, we have

$$\begin{aligned}\hat{\boldsymbol{\beta}}_S &= \hat{\boldsymbol{\beta}}_R + (1 - (r-2)D_N^{-1})(\hat{\boldsymbol{\beta}}_U - \hat{\boldsymbol{\beta}}_R)\\ \hat{\boldsymbol{\beta}}_+ &= \hat{\boldsymbol{\beta}}_R + (1 - (r-2)D_N^{-1})^+(\hat{\boldsymbol{\beta}}_U - \hat{\boldsymbol{\beta}}_R),\end{aligned}$$

with $r \geq 3$ and $z^+ = \max(0, z)$. The shrinkage estimator can be viewed as a smoothed version of the pretest estimator which selects the UE, $\hat{\boldsymbol{\beta}}_U$ when the statistic D_N is large and the RE, $\hat{\boldsymbol{\beta}}_R$ otherwise.

2.6. Penalized Estimator: LASSO Penalized regression estimates are obtained by maximizing the likelihood function (4) subject to a specified constraint

$$\arg\max_{\boldsymbol{\beta}}\{\ell^*(\boldsymbol{\gamma}|\boldsymbol{y}) - \boldsymbol{P}_{\lambda_k}(.)\},$$

where $\boldsymbol{P}_{\lambda_k}(.)$ is a penalty function, λ_k are the penalty parameters, and $k = 1, \cdots, p$. The LASSO estimator implements the penalty function $\boldsymbol{P}_{\lambda_k}(.)$ as

$$\boldsymbol{P}_{\lambda_k}(.) = \lambda\|\boldsymbol{\beta}\|_1,$$

where $\|.\|_1$ is the L_1-norm, and $\lambda \geq 0$ is the regularization parameter that have to be determined via information criteria or cross-validation. To obtain the LASSO estimator, we use the full gradient algorithm based on Goeman (2010), in which the algorithm can amend to situations where subsets of the parameter space are not penalized. The algorithm starts with a global intercept model with random effect (i.e. $g(\mu_{ij}) = \boldsymbol{\beta}_0 + \boldsymbol{z}_{ij}^\top \boldsymbol{u}_i$), then iterates the algorithm until convergence. The non-zero estimates are the combination of the gradient decent and the Fisher scoring algorithms. See, Groll and Tutz (2014) for details and implementation of the algorithm.

3 Asymptotic Properties of the Estimators

In this section, we derive the asymptotic joint normality for the UE and RE. First note that under the regularity conditions listed in the Appendix A, $\sqrt{N}(\boldsymbol{A}\hat{\boldsymbol{\beta}}_U - \boldsymbol{c}) \xrightarrow{\mathcal{L}} \mathcal{N}(\boldsymbol{\delta}, \boldsymbol{A}\boldsymbol{B}^{-1}\boldsymbol{A}^\top)$, where $\boldsymbol{B} = \lim_{N\to\infty} \boldsymbol{I}(\boldsymbol{\beta}, \boldsymbol{\beta})/N$ converges in probability to a non-random $p \times p$ positive definite matrix, and $\boldsymbol{\delta}$ is defined below. It is well known that the effective domain of risk dominance of shrinkage estimators over the UE is near the null hypothesis $H_0 : \boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{c}$; and as we increase the sample size N, this domain becomes narrower. To avoid asymptotic degeracy, we consider a sequence of local alternatives

$$K_{(N)} : \boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{c} + \frac{\boldsymbol{\delta}}{\sqrt{N}}, \tag{9}$$

where $\boldsymbol{\delta} = (\delta_1, \delta_2, \cdots, \delta_r)^\top \in \Re^r$ is fixed. Observe that $\boldsymbol{\delta} = \mathbf{0}$ implies that $\boldsymbol{A\beta} = \boldsymbol{c}$ for all N, which is a special case of (9). Under local $K_{(N)}$, the following theorem facilitates the theoretical comparison and numerical computation of the ADB and ADR of the estimators:

Theorem 3.1. *Under some regularity conditions in the Appendix, and the sequence of local alternative (9), we have the joint distributions:*

$$(i). \quad \begin{bmatrix} \boldsymbol{\eta}_1 \\ \boldsymbol{\eta}_2 \end{bmatrix} \sim N_{2p}\left(\begin{bmatrix} \mathbf{0} \\ -\boldsymbol{J}_0\boldsymbol{\delta} \end{bmatrix}, \begin{bmatrix} \boldsymbol{B}^{-1} & \boldsymbol{B}^{-1} - \boldsymbol{B}^* \\ \boldsymbol{B}^{-1} - \boldsymbol{B}^* & \boldsymbol{B}^{-1} - \boldsymbol{B}^* \end{bmatrix} \right)$$

$$(ii). \quad \begin{bmatrix} \boldsymbol{\eta}_2 \\ \boldsymbol{\eta}_3 \end{bmatrix} \sim N_{2p}\left(\begin{bmatrix} \boldsymbol{J}_0\boldsymbol{\delta} \\ -\boldsymbol{J}_0\boldsymbol{\delta} \end{bmatrix}, \begin{bmatrix} \boldsymbol{B}^* & \mathbf{0} \\ \mathbf{0} & \boldsymbol{B}^{-1} - \boldsymbol{B}^* \end{bmatrix} \right),$$

where $\boldsymbol{\eta}_1 = \lim_{n\to\infty} \sqrt{n}(\hat{\boldsymbol{\beta}}_U - \boldsymbol{\beta})$, $\boldsymbol{\eta}_2 = \lim_{n\to\infty} \sqrt{n}(\hat{\boldsymbol{\beta}}_U - \hat{\boldsymbol{\beta}}_R)$, $\boldsymbol{\eta}_3 = \lim_{n\to\infty} \sqrt{n}(\hat{\boldsymbol{\beta}}_R - \boldsymbol{\beta})$, $\boldsymbol{J}_0 = \boldsymbol{B}^{-1}\boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{B}^{-1}\boldsymbol{A}^\top)^{-1}$, *and* $\boldsymbol{B}^* = (\boldsymbol{I} - \boldsymbol{J}_0\boldsymbol{A})\boldsymbol{B}^{-1}$.

Using this theorem, we can obtain the main results of this section. We present the asymptotic distributional bias (ADB) and asymptotic distributional risk (ADR) results. We begin with the bias.

3.1. Asymptotic Distributional Bias Consider a sequence of parameter values $\boldsymbol{\beta}$ and a sequence of estimators $\hat{\boldsymbol{\beta}}_*$. Assume that $\sqrt{n}(\hat{\boldsymbol{\beta}}_* - \boldsymbol{\beta})$ converges in distribution as $n \to \infty$ to some random variable $\boldsymbol{X}$ with distribution $\tilde{G}$. Then the ADB of $\hat{\boldsymbol{\beta}}_*$ is defined by

$$\mathrm{ADB}(\hat{\boldsymbol{\beta}}_*, \boldsymbol{\beta}) = \mathrm{E}\left[\boldsymbol{X}\right] = \int \boldsymbol{x} d\tilde{G}(\boldsymbol{x}).$$

Here we define the notations in order to describe our ADB and ADR results. Let Z_1 and Z_2 be $\chi^2_{r+2}(\Delta)$ and $\chi^2_{r+4}(\Delta)$ random variables, respectively. The distribution function of a non-central χ^2 variable with non-centrality parameter Δ and degrees of freedom g is denoted by $H_g(x, \Delta) = P\left(\chi^2_g(\Delta) \le x\right)$. Let $\chi^2_{r,\alpha}$ be the α-level critical value of central χ^2 distribution. In view of Theorem 3.1, the asymptotic biases of the estimators give us the following Theorem.

Theorem 3.1.1. *If the condition of Theorem 3.1 hold, then:*

$$\begin{aligned}
ADB(\hat{\boldsymbol{\beta}}_U) &= \mathbf{0}; \quad ADB(\hat{\boldsymbol{\beta}}_R) = -\boldsymbol{J}_0\boldsymbol{\delta}; \\
ADB(\hat{\boldsymbol{\beta}}_P) &= -\boldsymbol{J}_0\boldsymbol{\delta} H_{r+2}(\chi^2_{r,\alpha}, \Delta); \quad ADB(\hat{\boldsymbol{\beta}}_S) = -(r-2)\boldsymbol{J}_0\boldsymbol{\delta} E\left(Z_1^{-1}\right); \\
ADB(\hat{\boldsymbol{\beta}}_+) &= -\boldsymbol{J}_0\boldsymbol{\delta}\left((r-2)E\left(Z_1^{-1}(1 - I(Z_1 < r-2))\right) + H_{r+2}(r-2, \Delta)\right).
\end{aligned}$$

PROOF. See Appendix B.

Remark 3.1. *To compare ADBs, let* $\boldsymbol{\tau} = -\boldsymbol{J}_0\boldsymbol{\delta}/\sqrt{\Delta}$ *where* $\Delta = \boldsymbol{\delta}^\mathsf{T}(\boldsymbol{A}\boldsymbol{B}^{-1}\boldsymbol{A}^\mathsf{T})^{-1}\boldsymbol{\delta}$. *All of the ADB expressions in Theorem 3.1.1. is a scaler multiple of* Δ *with the vector* $\boldsymbol{\tau}$. *Thus the value of* $\boldsymbol{\tau}$ *remains the same except for this change of scaler multiplication. Now to compare the ADBs of different estimators it suffices to compare these scalar terms. The scalar term in the ADB of* $\hat{\boldsymbol{\beta}}_R$ *is* $\sqrt{\Delta}$, *which is an unbounded function of* Δ. *On the other hand, the scalar terms in the ADBs of* $\hat{\boldsymbol{\beta}}_P$, $\hat{\boldsymbol{\beta}}_S$, *and* $\hat{\boldsymbol{\beta}}_+$ *are bounded in* Δ. *For example, since* $E(Z_1^{-1})$ *is not an increasing function of* Δ, *the ADB of* $\hat{\boldsymbol{\beta}}_S$ *starts from the origin, increases to a maximum, and then decreases towards 0 as* $\Delta \to \infty$. *The characteristics of* $\hat{\boldsymbol{\beta}}_P$, *and* $\hat{\boldsymbol{\beta}}_+$ *are similar to that of* $\hat{\boldsymbol{\beta}}_S$. *We now turn to the ADRs of the estimators.*

3.2. Asymptotic Distributional Risk To derive expressions for the ADRs of the estimators, we define a quadratic loss function

$$\mathcal{L}(\hat{\boldsymbol{\beta}}_*; \boldsymbol{M}) = \left(\sqrt{N}(\hat{\boldsymbol{\beta}}_* - \boldsymbol{\beta})\right)^\top \boldsymbol{M} \left(\sqrt{N}(\hat{\boldsymbol{\beta}}_* - \boldsymbol{\beta})\right),$$

where $\boldsymbol{M}$ is a nonnegative weight matrix (typically, $\boldsymbol{M} = \boldsymbol{I}_{p\times p}$, which is the usual quadratic loss). Using a general $\boldsymbol{M}$ provides a loss function that weights other $\boldsymbol{\beta}$'s differently. The expected loss function, or the risk function, is defined as $E\left(\lim_{N\to\infty} \mathcal{L}(\hat{\boldsymbol{\beta}}_*; \boldsymbol{M})\right)$. Let $\hat{\boldsymbol{V}}_N$ be the variance-covariance matrix of $\sqrt{N}(\hat{\boldsymbol{\beta}}_* - \boldsymbol{\beta})$, and let $\mathrm{R}_N(\hat{\boldsymbol{\beta}}_*; \boldsymbol{M})$ be the expected value of the loss function, that is, the quadratic risk. We have

$$\mathrm{R}_N(\hat{\boldsymbol{\beta}}_*; \boldsymbol{M}) = \mathrm{trace}(\boldsymbol{M}\hat{\boldsymbol{V}}_N) + \left(\mathrm{ADB}(\hat{\boldsymbol{\beta}}_*, \boldsymbol{\beta})\right)^\top \boldsymbol{M} \left(\mathrm{ADB}(\hat{\boldsymbol{\beta}}_*, \boldsymbol{\beta})\right).$$

If $\lim_{N\to\infty} \mathrm{R}_N(\hat{\boldsymbol{\beta}}_*; \boldsymbol{M})$ exists, it is called *asymptotic risk*. In order to define this quantity, let $\sqrt{N}(\hat{\boldsymbol{\beta}}_* - \boldsymbol{\beta})$ converges in distribution to $\boldsymbol{\Xi}$ as N tends to infinity. The ADR of $\hat{\boldsymbol{\beta}}_*$ is defined as $\mathrm{ADR}(\hat{\boldsymbol{\beta}}_*; \boldsymbol{M}) = E\left(\mathrm{trace}(\boldsymbol{\Xi}^\top \boldsymbol{M}\boldsymbol{\Xi})\right)$. This leads us to the following Theorem:

Theorem 3.2.1. *If the condition of Theorem 3.1 hold, then:*

$$\begin{aligned}
ADR(\hat{\boldsymbol{\beta}}_U; \boldsymbol{M}) &= trace\left(\boldsymbol{M}\boldsymbol{B}^{-1}\right), \\
ADR(\hat{\boldsymbol{\beta}}_R; \boldsymbol{M}) &= ADR(\hat{\boldsymbol{\beta}}_U; \boldsymbol{M}) - trace\left(\boldsymbol{M}\boldsymbol{B}^{-1}\boldsymbol{A}^\top \boldsymbol{J}_0^\top\right) + \boldsymbol{\delta}^\top \boldsymbol{J}_0^\top \boldsymbol{M}\boldsymbol{J}_0\boldsymbol{\delta}, \\
ADR(\hat{\boldsymbol{\beta}}_P; \boldsymbol{M}) &= ADR(\hat{\boldsymbol{\beta}}_U; \boldsymbol{M}) - trace\left(\boldsymbol{M}\boldsymbol{B}^{-1}\boldsymbol{A}^\top \boldsymbol{J}_0^\top\right) H_{r+2}(\chi^2_{r,\alpha}, \Delta) \\
&\quad - \boldsymbol{\delta}^\top \boldsymbol{J}_0^\top \boldsymbol{M}\boldsymbol{J}_0\boldsymbol{\delta}\left(H_{r+4}(\chi^2_{r,\alpha}, \Delta) - 2H_{r+2}(\chi^2_{r,\alpha}, \Delta)\right),
\end{aligned}$$

$$
\begin{aligned}
ADR(\hat{\boldsymbol{\beta}}_S;\boldsymbol{M}) &= ADR(\hat{\boldsymbol{\beta}}_U;\boldsymbol{M}) + (r-2)trace\left(\boldsymbol{M}\boldsymbol{B}^{-1}\boldsymbol{A}^\top\boldsymbol{J}_0^\top\right)\\
&\quad\times\left((r-2)E(Z_1^{-2}) - 2E(Z_1^{-1})\right) + (r-2)\\
&\quad\times\left((r-2)E(Z_2^{-2}) - 2E(Z_2^{-1} - Z_1^{-1})\right)\boldsymbol{\delta}^\top\boldsymbol{J}_0^\top\boldsymbol{M}\boldsymbol{J}_0\boldsymbol{\delta},\\
ADR(\hat{\boldsymbol{\beta}}_+;\boldsymbol{M}) &= ADR(\hat{\boldsymbol{\beta}}_S;\boldsymbol{M}) - trace\left(\boldsymbol{M}\boldsymbol{B}^{-1}\boldsymbol{A}^\top\boldsymbol{J}_0^\top\right)\\
&\quad\times E\left((1-(r-2)Z_1^{-1})^2 I(Z_1 < r-2)\right)\\
&\quad- E\left((1-(r-2)Z_2^{-1})^2 I(Z_2 < r-2)\right)\boldsymbol{\delta}^\top\boldsymbol{J}_0^\top\boldsymbol{M}\boldsymbol{J}_0\boldsymbol{\delta}\\
&\quad+2E\left((1-(r-2)Z_1^{-1})I(Z_1 < r-2)\right)\boldsymbol{\delta}^\top\boldsymbol{J}_0^\top\boldsymbol{M}\boldsymbol{J}_0\boldsymbol{\delta}.
\end{aligned}
$$

Proof. See Appendix B.

Remark 3.2. *From Theorem 3.2.1, we note that subject to some suitable weight matrix* $\boldsymbol{M}$*, the* $ADR(\hat{\boldsymbol{\beta}}_U;\boldsymbol{M})$ *and* $ADR(\hat{\boldsymbol{\beta}}_R;\boldsymbol{M})$ *follow directly from Theorem 3.1; thus, we have* $ADR(\hat{\boldsymbol{\beta}}_U;\boldsymbol{M}) = trace\left(\boldsymbol{M}\boldsymbol{B}^{-1}\right)$ *and* $ADR(\hat{\boldsymbol{\beta}}_R;\boldsymbol{M}) = trace\left(\boldsymbol{M}\boldsymbol{B}^{*}\right) + \boldsymbol{\delta}^\top\boldsymbol{J}_0^\top\boldsymbol{M}\boldsymbol{J}_0\boldsymbol{\delta}$*. By comparing the risk of the estimators, we see that, as* Δ *moves away from 0, the risk of* $\hat{\boldsymbol{\beta}}_R$ *becomes unbounded. Thus,* $\hat{\boldsymbol{\beta}}_R$ *may not behave well when the value of* $\boldsymbol{\delta}$ *is different from the specified* $\mathbf{0}$ *vector. For all* $\Delta \in (0,\infty)$*,* $ADR(\hat{\boldsymbol{\beta}}_U) < trace\left(\boldsymbol{M}\boldsymbol{B}^{-1}\right)$*, hence* $\hat{\boldsymbol{\beta}}_S$ *provides greater estimation accuracy than* $\hat{\boldsymbol{\beta}}_U$*. Indeed, the ADR function of the SE is monotone in* Δ*, where the smallest value is achieved at* $\Delta = 0$ *and the largest is trace* $\left(\boldsymbol{M}\boldsymbol{B}^{-1}\right)$*. Hence,* $\hat{\boldsymbol{\beta}}_S$ *outperforms* $\hat{\boldsymbol{\beta}}_U$ *and is an admissible estimator when compared with* $\hat{\boldsymbol{\beta}}_U$*. Furthermore, the PSE,* $\hat{\boldsymbol{\beta}}_+$ *asymptotically superior to* $\hat{\boldsymbol{\beta}}_S$ *in the entire parameter space induced by* Δ*. Therefore,* $\hat{\boldsymbol{\beta}}_+$ *is also superior to* $\hat{\boldsymbol{\beta}}_U$*.*

We next report on a simulation study, which compares the performance of the estimators of this section and the penalty estimators for finite sample sizes.

4 Simulation Study

In this Section, we assess the performance of the proposed estimators with respect to the UE through a Monte Carlo simulation study. Our simulations are based on the models

$$\mu_{ij} = (1+\exp(-\eta_{ij}))^{-1}, \quad \text{and} \quad \mu_{ij} = \exp(\eta_{ij}),$$

for a binary and count response, respectively, where $\eta_{ij} = \boldsymbol{x}_{ij}^\top\boldsymbol{\beta} + \boldsymbol{z}_{ij}^\top\boldsymbol{u}_i$ is the linear predictor. We consider $n_i = 3$ observations per subject, where we have a total of $N = 75$ and $N = 150$ subjects for the count and binary

response, respectively. In our study, we consider generating $p = 7, 12, 17$ and 21 fixed effect covariates, a random intercept, and a random effect covariate generated from the standard normal distribution.

Each of the p fixed effect covariates $\boldsymbol{x}_{ij} = (x_{ij1}, \cdots, x_{ij,p})^\top$ were generated from a separate n_i-multivariate normal distribution with mean $\mathbf{0}$ and variance-covariance matrix $\sigma^2_{\boldsymbol{x}}\boldsymbol{\rho}_{\boldsymbol{x}}$, where $\sigma^2_{\boldsymbol{x}} = 0.21$, and $\boldsymbol{\rho}_{\boldsymbol{x}}$ is an exchangeable correlation matrix with parameter $\rho = 0.8$. The purpose of selecting a small $\sigma^2_{\boldsymbol{x}}$ value is to generate manageable response values, and the objective of generating correlated fixed effect covariates is to capture the between-subject correlation. The basic idea is to then generate subject-specific random effects in η_{ij} which then acts as noise. This gives us the necessary conditional independence, and establishes the GLMM framework.

To generate the random effects, we begin by specifying $\boldsymbol{\theta} = (0.7, 0.7, 0)$ as the variance parameters, that is, $q = 2$, and $\frac{q(q-1)}{2} = 1$ covariance parameter. The random effects $\boldsymbol{u}_i$ are then generated from a bivariate normal distribution with mean $\mathbf{0}$, and variance-covariance matrix comprised of the non-zero elements of $\boldsymbol{\theta}$.

We consider a special case of the hypothesis $H_0 : \boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{c}$, where we partition the fixed effect regression vector into two sub-vectors $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^\top, \boldsymbol{\beta}_2^\top)^\top$, where $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are assumed to have dimensions $p_1 \times 1$ and $p_2 \times 1$, respectively, such that $p = p_1 + p_2$ (i.e. $p_2 = r$, and $p_1 = p - r$). We are interested in estimating the sub-vector $\boldsymbol{\beta}_1$ by incorporating the information of $\boldsymbol{\beta}_2$ into the estimation procedure, where we consider the null hypothesis to be $H_0 : \boldsymbol{\beta}_2 = \mathbf{0}$. We specify $p_1 = 4$ throughout the study, and set $\boldsymbol{\beta}_1 = (-1.85, 1.2, -1.3, 2.13)^\top$, and $\beta_0 = 0.4$ as the global intercept term. The weight matrix $\boldsymbol{M}$ in the quadratic loss function from the previous Section is set to the $p \times p$ identity matrix.

We define the parameter, $\Delta = \|\boldsymbol{\beta} - \boldsymbol{\beta}^{(0)}\|^2$, where $\|.\|$ is the Euclidean norm, $\boldsymbol{\beta}^{(0)} = (\boldsymbol{\beta}_1^\top, \boldsymbol{m}^\top)^\top$, and $\boldsymbol{m}$ is a zero vector with various dimensions. The samples were generated using $\Delta = (0,\ 0.03,\ 0.05,\ 0.07,\ 0.1,\ 0.3,\ 0.54,\ 1,\ 2)$, and $\Delta = (0,\ 0.03,\ 0.05,\ 0.07,\ 0.1,\ 0.3,\ 0.54,\ 1,\ 2,\ 4)$ for the count and binary response, respectively. We used 1000 replications in the simulation, as a further increase did not change the results significantly.

The objective here is to investigate the behaviour of the shrinkage estimators for $\Delta \geq 0$, and the LASSO estimator for $\Delta = 0$; Ahmed et al. (2012) show that penalty estimators do not correctly estimate the model parameters for $\Delta > 0$. We combine the gradient ascent optimization with the Fisher scoring algorithm similar to Goeman (2010) to find the entire solution path for the LASSO estimator, where the optimal λ value is determined by the BIC criterion.

The criterion for comparing the performance of any estimator $\hat{\boldsymbol{\beta}}_*$ in our study is the relative mean square efficiency (RMSE) of $\hat{\boldsymbol{\beta}}_*$ to $\hat{\boldsymbol{\beta}}_U$ and is defined as $\text{RMSE}(\hat{\boldsymbol{\beta}}_*) = \text{MSE}(\hat{\boldsymbol{\beta}}_*)/\text{MSE}(\hat{\boldsymbol{\beta}}_U)$, where $\hat{\boldsymbol{\beta}}_*$ is the proposed estimator. Thus, a RMSE value less than 1 indicates risk reduction relative to $\hat{\boldsymbol{\beta}}_U$.

The simulation results are presented in Figs. 1 & 2, and Table 1, where the inactive set of parameters is $\boldsymbol{\beta}_2 = (\Delta, \mathbf{0})$, and $\mathbf{0}$ is the null vector of length $p_2 - 1$.

4.1. Binary Response We summarize our findings (with reference to Fig. 1) of the simulation study with a binary response as follows:

(i) With the exception of $p_2 = 3$, the minimum RMSE occurred at a slight departure from the null hypothesis. This is attributed to sampling

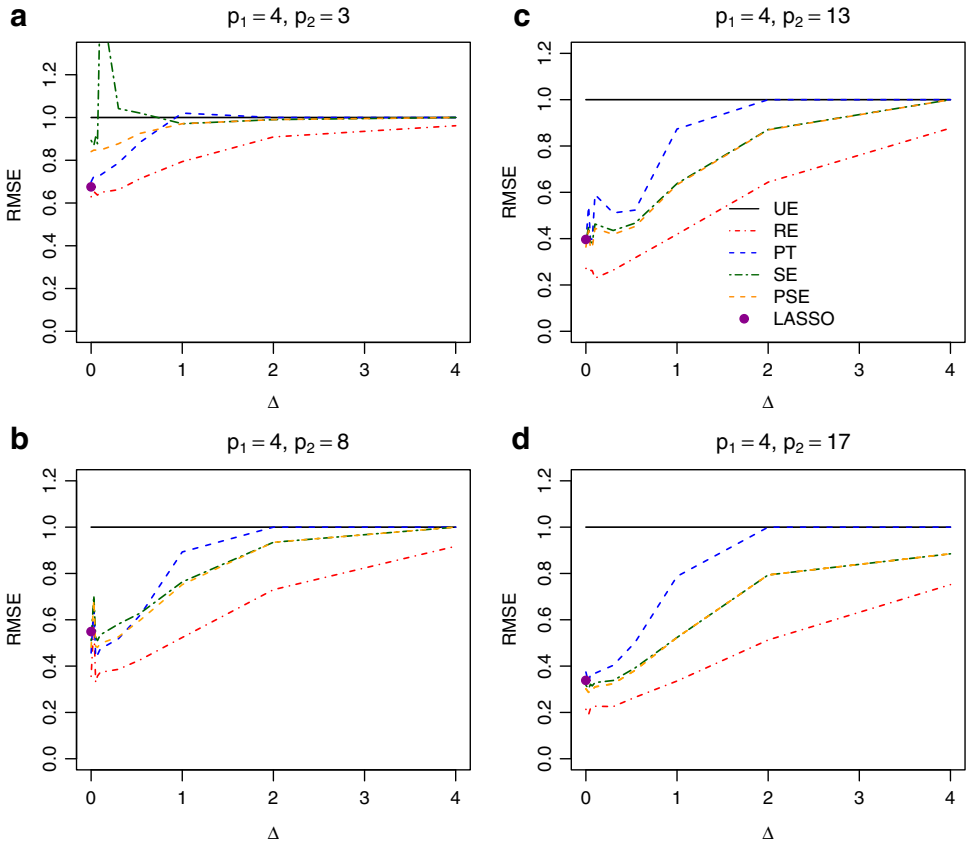

Figure 1: RMSEs of the estimators with respect to the UE when the subspace misspecifies Δ as zero, $N = 150$, and the response is binary. Here, p_2 is the number of inactive predictors

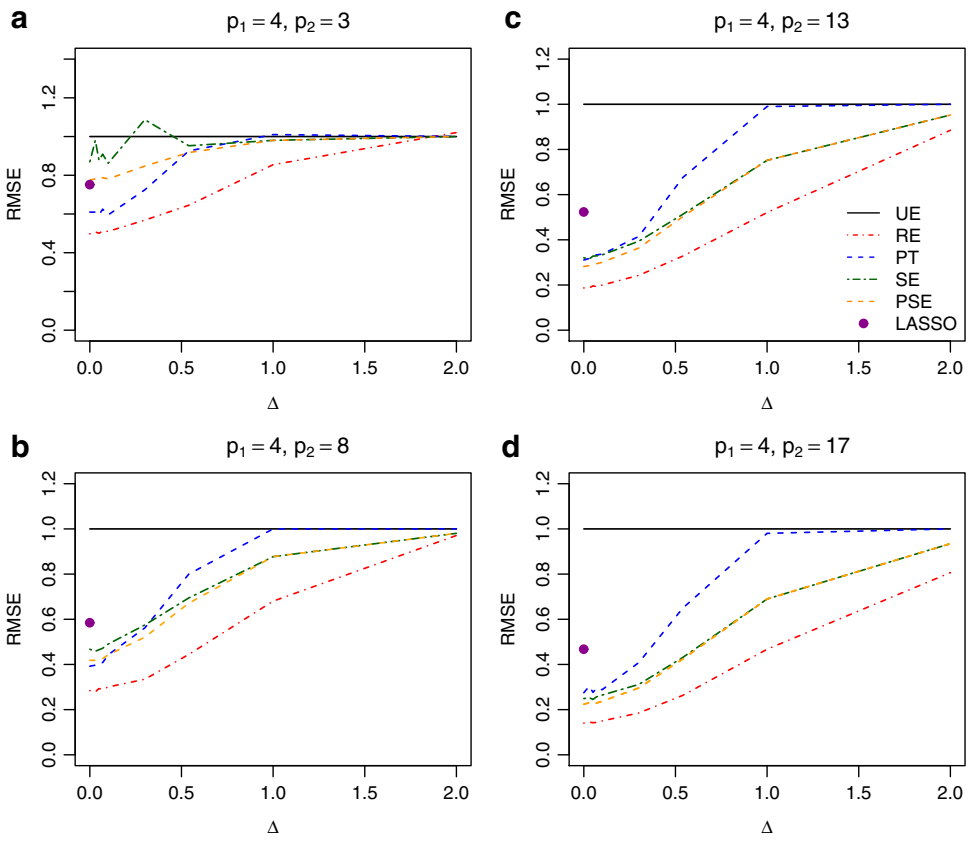

Figure 2: RMSEs of the estimators with respect to the UE when the subspace misspecifies Δ as zero, $N = 75$, and the response is a count. Here, p_2 is the number of inactive predictors

error within the simulation study. In general, we observed that as p_2 increases, the RMSE decreases. Due to the nature of the RE, it is the optimal estimator over the entire parameter space.

(ii) The RMSE of the RE is diverging towards ∞ as $\Delta \to \infty$ at a slow rate. The RMSE of the other estimators are bounded and converge to 1 as $\Delta \to \infty$ with RMSE($\hat{\boldsymbol{\beta}}_P$) being the first to converge.

(iii) The LASSO estimator outperformed the PSE for $p_2 = 3$. The relative performance of the LASSO estimator over the PT cannot be determined.

Table 1: RMSE of the RE, PT, SE, PSE, and LASSO estimator with respect to the UE when the restricted parameter space is correct (i.e. $\Delta = 0$)

Method	$p_2 = 3$	$p_2 = 8$	$p_2 = 13$	$p_2 = 17$
	Binary response ($p_1 = 4$, $N = 150$)			
RE	0.629	0.356	0.271	0.214
PT	0.699	0.457	0.385	0.375
SE	0.893	0.510	0.385	0.317
PSE	0.840	0.481	0.364	0.302
LASSO	0.676	0.549	0.397	0.338
	Count response ($p_1 = 4$, $N = 75$)			
RE	0.498	0.284	0.187	0.140
PT	0.610	0.392	0.311	0.275
SE	0.870	0.467	0.321	0.249
PSE	0.775	0.418	0.282	0.224
LASSO	0.752	0.585	0.524	0.467

(iv) For $\Delta = 0$ and small-to-medium values of p_2, the PT is outperformed by the PSE. However, this superiority of the PT over the PSE diminishes as p_2 increases. For large p_2, the PT underperformed with respect to the PSE. Overall, the risk function decreases as p_2 increases.

(v) The PSE outperformed the SE at $0 \leq \Delta < 2$, and $\hat{\boldsymbol{\beta}}_{S+}$ converges to $\hat{\boldsymbol{\beta}}_S$ for $\Delta \geq 2$. Therefore, the PSE is declared to be superior over the SE for all values of Δ. Overall, the greatest risk reductions occur for parameter values near the restriction $\Delta = 0$.

4.2. Count Response We summarize our findings (with reference to Fig. 2) of the simulation study with a count response as follows:

(i) The minimum RMSE occurred at $\Delta = 0$, see Table 1, with the exception of one case attributed to random error. In general, we have that as p_2 increases, the RMSE decreases. As Δ departs from 0, the RE outperformed the other estimators. The only case that we observed $\text{RMSE}(\hat{\boldsymbol{\beta}}_R) > 1$ is when $p_2 = 3$ at $\Delta = 2$.

(ii) The RMSE of the RE diverges towards ∞ as $\Delta \to \infty$ at a slow rate. The RMSE of the other estimators are bounded and approaches 1 as $\Delta \to \infty$ with $\text{RMSE}(\hat{\boldsymbol{\beta}}_P)$ being the first to converge.

(iii) The PSE outperformed the SE at $0 \leq \Delta < 1$, and $\hat{\boldsymbol{\beta}}_{S+}$ converges to $\hat{\boldsymbol{\beta}}_S$ for $\Delta \geq 1$. Therefore, the PSE outperforms the SE for all values of

Δ. However, the RE outperformed both the SE and PSE for medium and large values of Δ.

(iv) For $\Delta = 0$ and small-to-medium values of p_2, the PT outperformed the PSE. This superiority of the PT over the PSE diminishes as p_2 increases. For large p_2, the PT underperformed with respect to the PSE.

(v) The LASSO estimator outperformed the PSE estimator for $p_2 = 3$, and underperformed for $p_2 > 4$. Furthermore, the LASSO estimator is declared inferior to the PT for all values of p_2.

5 Example: Indonesian Preschool Respiratory Infections

The data was collected by Sommer et al. (1984), where they conducted a study of over 3000 preschool children in the Aceh province of Indonesia to determine the causes and effects of vitamin A deficiency within this demographic. This particular data set can be obtained from Diggle et al. (2002), where 275 children were examined for up to six consecutive quarters, to see whether they suffered from respiratory infection or xerophthalmia, an ocular manifestation of vitamin A deficiency. This is a subset of a cohort studied by Sommer et al. (1984). The purpose of this study is to determine which health-related factors can best determine the presence of a respiratory infection.

These longitudinal data consists of a binary response, y_{ij}, to indicate the presence of a respiratory infection (1 - yes; 0 - no) for the ith child at time-point j, where $i = 1, \cdots, 275$, and $j = 1, 2, \cdots, 6$. The reported covariates pertaining to this data set is the observed height, weight, gender (1 - male; 0 - female), baseline age (in months), xerophthalmia, time, sin(time), and cos(time). Diggle et al. (2002) provided an initial analysis of the data, where they found based on a few fitted models that the prevalence of respiratory infection is a function of age^2 and is sinusoidally based on time. Furthermore, they also found that there is a quadratic relationship between time and the logarithm of the risk of respiratory infection. Therefore, we consider the model

$$\begin{aligned}
\mu_{ij} &= E(y_{ij}|\boldsymbol{u}_i) = (1 + \exp(-\eta_{ij}))^{-1}, \\
\eta_{ij} &= \beta_0 + \beta_1 \text{Gender}_i + \beta_2 \text{Height}_{ij} + \beta_3 \text{Age}_i + \beta_4 \text{Age}_i^2 + \beta_5 \cos j \\
&\quad + \beta_6 \sin j + \beta_7 \text{Xerophthalmia}_{ij} + \beta_8 j + \beta_9 j^2 + \boldsymbol{u}_i,
\end{aligned}$$

where Height_{ij}, and Age_i are standardized.

We propose several candidate models, and use the AIC & BIC as a selection criteria, see Table 2. It shows the AIC and BIC values for the seven models M1-M7. Based on these values, models 5 and 6 have the lowest AIC and BIC values compare to the other models and model M6 is the preferred one to explain the risk factors for respiratory infection. Model M6 shows that Gender_i, Height_{ij}, Age_i, Age_i^2, and $\cos j$ are the significant risk factors and the other four may not be risk factors for this infection. Reminiscent to the simulation study, we obtain the inactive set of parameters $H_0 : \boldsymbol{\beta}_2 = \mathbf{0}$, where $\boldsymbol{\beta}_2 = (\beta_6, \beta_7, \beta_8, \beta_9)^\top$ is a $p_2 \times 1$ vector, and $p_2 = 4$. Observe that xerophthalmia is one of the insignificant covariates, which is due to the limited number of cases within the data.

5.1. Bootstrap A bootstrap sampling scheme (Wu and Chiang, 2000) is conducted to compute the estimates, standard errors, and RMSEs of the proposed estimates. We randomly sample 275 subjects with replacement from the original data set, and let $\{\boldsymbol{Y}_i^*, \boldsymbol{X}_i^*, \boldsymbol{Z}_i^*;\ \ 1 \le i \le 275, 1 \le j \le 6\}$ be the longitudinal bootstrap sample. The entire measurements of some subjects in the original sample may appear multiple times in the new bootstrap sample. We then refit the GLMM model using these data based on the same method that would be applied in Section 5 to obtain bootstrap estimates and conduct 1000 replications iteratively. The point estimates, standard errors, and RMSEs of significant coefficients are reported in Table 3. The

Table 2: Seven different models (say, M1-M7) for the Indonesian pre-school respiratory infections data, and their AIC and BIC values. This shows that model 6 is the optimal model

Covariates	M1	M2	M3	M4	M5	M6	M7
Fixed intercept	Y	Y	Y	Y	Y	Y	Y
Gender	Y	Y	N	N	Y	Y	Y
Height	Y	Y	Y	N	Y	Y	Y
Age	Y	Y	Y	Y	Y	Y	Y
Age^2	Y	Y	Y	Y	Y	Y	Y
$\cos j$	N	N	N	N	Y	Y	Y
$\sin j$	N	N	N	N	Y	N	Y
Xerophthalmia	N	N	N	N	N	N	Y
j	N	Y	Y	Y	N	N	Y
j^2	N	Y	Y	Y	N	N	Y
Random intercept	Y	Y	Y	Y	Y	Y	Y
AIC	681.94	678.19	681.20	684.97	672.09	670.79	674.89
BIC	712.48	718.91	716.83	715.51	712.81	706.42	730.88

Table 3: Estimates (first row) and standard errors (second row) for Gender_i (β_1), Height_{ij} (β_2), Age_i (β_3), Age_i^2 (β_4), and $\cos j$ (β_5)

Estimators	β_1	β_2	β_3	β_4	β_5	RMSE
UE	−0.677	−0.429	−0.998	−0.551	−0.645	1.000
	(0.282)	(0.157)	(0.214)	(0.218)	(0.282)	
RE	−0.689	−0.450	−0.989	−0.545	−0.677	0.550
	(0.278)	(0.146)	(0.210)	(0.216)	(0.222)	
PT	−0.685	−0.442	−0.994	−0.548	−0.660	0.802
	(0.280)	(0.152)	(0.213)	(0.218)	(0.246)	
SE	−0.681	−0.435	−0.996	−0.549	−0.655	0.766
	(0.281)	(0.154)	(0.214)	(0.219)	(0.257)	
PSE	−0.681	−0.435	−0.996	−0.550	−0.655	0.740
	(0.281)	(0.152)	(0.213)	(0.217)	(0.257)	
LASSO	−0.539	−0.246	−0.838	−0.578	−0.088	0.838
	(0.220)	(0.192)	(0.209)	(0.208)	(0.224)	

RMSEs of RE, PT, SE, PSE and LASSO with respect to UE are 0.55, 0.802, 0.766, 0.740 and 0.838, respectively. The findings are consistent with the simulation results and asymptotic findings.

6 Conclusions

In this paper, we compared the relative performance of the RE, PT, SE, PSE, and LASSO estimators with respect to the UE in the context of a GLMM when some of the covariates may be restricted to a linear subspace. In doing so, we have presented a closed form of the bias and risk expressions, and used a Monte Carlo simulation study to explore the bias and risk properties of the estimators under consideration. We conclude that the risk improvement of the RE over all other estimators is substantial at and near the restriction $\boldsymbol{A\beta} = \boldsymbol{c}$. However, the improvement diminishes as we move away from this restriction. As Δ increases, the risk of the PT crosses the risk of the UE, reaching a maximum, and then decreasing monotonically to the risk of the UE. Furthermore, the PSE outperformed the UE in the entire parameter space, and will outperform the restricted estimator for large enough Δ. The risk of the pretest estimator is less than the UE at and near the restriction. Finally, we applied the proposed estimation methods to a real data example to evaluate the relative performance of the proposed estimators. The findings are in agreement with the simulation study, and theoretical results.

References

AHMED, S., HUSSEIN, A. and SEN, P. (2006). Risk comparison of some shrinkage M-estimators in linear models. *Journal of Nonparametric Statistics* **18**, 401–415.

AHMED, S.E., HOSSAIN, S. and DOKSUM, K.A. (2012). LASSO And shrinkage estimation in Weibull censored regression models. *Journal of Statistical Planning and Inference* **142**, 1273–1284.

ARNOLD, T.B. and TIBSHIRANI, R.J. (2016). Efficient implementations of the generalized Lasso dualpath algorithm. *J. Comput. Graph. Stat.* **25**, 1–27.

CHEN, F. and NKURUNZIZA, S. (2015). A class of Stein-rules in multivariate regression model with structural changes. *Scand. J. Stat.* **43**, 83–102.

DAVIS, K.A., PARK, C.G. and SINHA, S.K. (2012). Testing for generalized linear mixed models with cluster correlated data under linear inequality constraints. *Can. J. Stat.* **40**, 243–258.

DIGGLE, P.J., HEAGERTY, P., LIANG, K-Y. and ZEGER, S.L. (2002). *Analysis of Longitudinal Data*, 2nd edn. Oxford University Press, Oxford.

FAHRMEIR, L. and KNEIB, T. (2011). *Bayesian smoothing and regression for longitudinal, spatial and event history data*. Oxford University Press, Oxford.

FALLAHPOUR, S., AHMED, S.E. and DOKSUM, K.A. (2012). ℓ_1 penalty and shrinkage estimation in partially linear models with random coefficient autoregressive errors. *Appl. Stoch. Model. Bus. Ind.* **28**, 236–250.

FAN, J. and LV, J. (2010). A selective overview of variable selection in high dimensional feature space. *Stat. Sin.* **20**, 101–148.

GOEMAN, J.J. (2010). Penalized estimation in the Cox proportional hazards model. *Biom. J.* **52**, 70–84.

GROLL, A. and TUTZ, G. (2014). Variable selection for generalized linear mixed models by l_1 penalized estimation. *Stat. Comput.* **24**, 137–154.

HOSSAIN, S. and AHMED, S.E. (2014). Shrinkage estimation and selection for a logistic regression model. *CRM Proceedings-Contemporary Mathematics* **622**, 159–176.

HOSSAIN, S., AHMED, S.E. and DOKSUM, K.A. (2015). Shrinkage, pretest, and penalty estimators in generalized linear models. *Stat. Methodology* **24**, 52–68.

JAMSHIDIAN, M. (2004). On algorithms for restricted maximum likelihood estimation. *Comput. Stat. Data Anal.* **45**, 137–157.

JUDGE, G.G. and BOCK, M.E. (1978). *The statistical implication of pretest and Stein-Rule estimators in econometrics*. North-Holland, Amsterdam.

LI, Q. and SHAO, J. (2015). Regularizing LASSO: A consistent variable selection method. *Stat. Sin.* **25**, 975–992.

LIAN, H. (2012). Shrinkage estimation for identification of linear components in additive models. *Statistics & Probability Letters* **82**, 225–231.

LOUIS, T.A. (1982). Finding the observed information matrix when using the EM algorithm. *J. R. Stat. Soc. Ser. B Methodol.* **44**, 226–233.

LU, X. and SU, L. (2016). Shrinkage estimation of dynamic panel data models with interactive fixed effects. *J. Econ.* **190**, 148–175.

LUENBERGER, D.G. and YE, Y. (2008). *Linear and nonlinear programming*. Springer Science & Business Media, New York.

MCCULLOCH, C.E., SEARLE, S.R. and NEUHAUS, J.M. (2008). *Generalized, Linear, and Mixed Models*. Wiley, Hoboken.

NELDER, J.A. and WEDDERBURN, R.W.M. (1972). Generalized linear models. *J. R. Stat. Soc. Ser. A* **135**, 370–384.

schelldorfer, j., meier, l. and bühlmann, p. (2014). GLMMLasso: an algorithm for high-dimensional generalized linear mixed models using l_1-penalization. *J. Comput. Graph. Stat.* **23**, 460–477.

simon, n., friedman, j., hastie, t. and tibshirani, r. (2014). A sparse-group LASSO. *J. Comput. Graph. Stat.* **22**, 231–245.

sommer, a., katz, j. and tarwotjo, i. (1984). Increased risk of respiratory infection and diarrhea in children with pre-existing mild vitamin a deficiency. *Am. J. Clin. Nutr.* **40**, 1090–1095.

thomson, t., hossain, s. and ghahramani, m. (2014). Application of shrinkage estimation in linear regression models with autoregressive errors. *J. Stat. Comput. Simul.* **85**, 3335–3351.

thomson, t., hossain, s. and ghahramani, m. (2015). *Efficient estimation for time series following generalized linear models* **58**, 493–513.

tibshirani, r. (1996). Regression shrinkage and selection via the LASSO. *J. R. Stat. Soc. Ser. B* **58**, 267–288.

wu, c.o. and chiang, c. (2000). Kernel smoothing on varying coefficient models with longitudinal dependent variable. *Stat. Sin.* **10**, 433–456.

zhang, t. and zou, h. (2014). Sparse precision matrix estimation via LASSO penalized D-trace loss. *Biometrika* **101**, 103–120.

Appendix A: Regularity conditions

(a) The parameter space $\boldsymbol{\Omega}$ for $\boldsymbol{\gamma} = (\boldsymbol{\beta}^\top, \boldsymbol{\theta}^\top)^\top$ is compact. The score functions $S_{\boldsymbol{\beta}}(\boldsymbol{\gamma})$ and $S_{\boldsymbol{\theta}}(\boldsymbol{\gamma})$ are continuous functions of $\boldsymbol{\gamma}$ for all $\boldsymbol{y}$ and measurable functions of $\boldsymbol{y}$ for each $\boldsymbol{\gamma} \in \boldsymbol{\Omega}$.

(b) There exists unique MLE $\hat{\boldsymbol{\gamma}}$ in $\boldsymbol{\Omega}$ for $\ell^*(\boldsymbol{\gamma})$. The moments of $\partial \ell^*(\boldsymbol{\gamma})/\partial \boldsymbol{\beta}$ exist at least up to the third order.

(c) The design matrices $\boldsymbol{X}$ and $\boldsymbol{Z}$ in model (3) are of full rank and all of their elements are bounded by a single finite real number.

Appendix B: Proof of Theorems 3.1.1 and 3.2.1

The following Lemma is needed for the derivation of bias and risk functions of Theorems 3.1.1 and 3.2.1:

Lemma 1. *Let h be a Borel measurable and real-valued integrable function, and $X \sim \mathcal{N}_p(\boldsymbol{\delta}, \boldsymbol{\Sigma}_p)$, where $\boldsymbol{\Sigma}_p$ is a nonnegative definite matrix with rank $r \leq p$. Also Let $\boldsymbol{R}$ be a $p \times p$ nonnegative definite matrix with rank k such that $\boldsymbol{\Sigma}_p \boldsymbol{R}$ is an idempotent matrix, $\boldsymbol{R}\boldsymbol{\Sigma}_p\boldsymbol{R} = \boldsymbol{R}$; $\boldsymbol{\Sigma}_p\boldsymbol{R}\boldsymbol{\Sigma}_p = \boldsymbol{\Sigma}_p$; and $\boldsymbol{\Sigma}_p\boldsymbol{R}\boldsymbol{\delta} = \boldsymbol{\delta}$, and let $\boldsymbol{W} = \boldsymbol{R}^{1/2}\boldsymbol{W}^*\boldsymbol{R}^{1/2}$, where $\boldsymbol{W}^*$ nonnegative definite matrix, Then,*

1) $E\left(h\left(\boldsymbol{X}^\top \boldsymbol{R}\boldsymbol{X}\right)\boldsymbol{W}\boldsymbol{X}\right) = E\left(h\left(\chi^2_{r+2}(\boldsymbol{\delta}^\top \boldsymbol{R}\boldsymbol{\delta})\right)\right)\boldsymbol{W}\boldsymbol{\delta}$

$$2)\quad E\left(h\left(\boldsymbol{X}^{\top}\boldsymbol{R}\boldsymbol{X}\right)\boldsymbol{X}^{\top}\boldsymbol{W}\boldsymbol{X}\right) = E\left(h\left(\chi^2_{r+2}(\boldsymbol{\delta}^{\top}\boldsymbol{R}\boldsymbol{\delta})\right)\right) trace(\boldsymbol{R}\boldsymbol{\Sigma}_p) + E\left(H\left(\chi^2_{r+4}(\boldsymbol{\delta}^{\top}\boldsymbol{R}\boldsymbol{\delta})\right)\right)\boldsymbol{\delta}^{\top}\boldsymbol{W}\boldsymbol{\delta},$$

The outline of proof of this lemma is given in Chen and Nkurunziza (2015).

B.1: Proof of Theorem 3.1.1:

$$\begin{aligned}
\mathrm{ADB}(\hat{\boldsymbol{\beta}}_U) &= \lim_{N\to\infty} E\left(\sqrt{N}(\hat{\boldsymbol{\beta}}_U - \boldsymbol{\beta})\right) = \mathbf{0}.\\
\mathrm{ADB}(\hat{\boldsymbol{\beta}}_R) &= \lim_{N\to\infty} E\left(\sqrt{N}(\hat{\boldsymbol{\beta}_R} - \boldsymbol{\beta})\right) = \lim_{N\to\infty} E\left(\sqrt{N}(\hat{\boldsymbol{\beta}}_U - \boldsymbol{\beta} + \hat{\boldsymbol{\beta}_R} - \hat{\boldsymbol{\beta}}_U)\right)\\
&= \mathbf{0} - (\boldsymbol{B}^{-1}\boldsymbol{A}^{\mathsf{T}}(\boldsymbol{A}\boldsymbol{B}^{-1}\boldsymbol{A}^{\mathsf{T}})^{-1} \lim_{N\to\infty} E\left(\sqrt{N}(\boldsymbol{A}\hat{\boldsymbol{\beta}}_U - \boldsymbol{c})\right)\\
&= -\boldsymbol{B}^{-1}\boldsymbol{A}^{\mathsf{T}}(\boldsymbol{A}\boldsymbol{B}^{-1}\boldsymbol{A}^{\mathsf{T}})^{-1}\boldsymbol{\delta} = -\boldsymbol{J}_0\boldsymbol{\delta},\\
&\quad \text{where } \boldsymbol{J}_0 = \boldsymbol{B}^{-1}\boldsymbol{A}^{\mathsf{T}}(\boldsymbol{A}\boldsymbol{B}^{-1}\boldsymbol{A}^{\mathsf{T}})^{-1}.\\
\mathrm{ADB}(\hat{\boldsymbol{\beta}}_P) &= \lim_{N\to\infty} E\left(\sqrt{N}(\hat{\boldsymbol{\beta}_P} - \boldsymbol{\beta})\right)\\
&= \lim_{N\to\infty} E\left(\sqrt{N}(\hat{\boldsymbol{\beta}}_U - I(D_N \le \chi^2_{r,\alpha})(\hat{\boldsymbol{\beta}}_U - \hat{\boldsymbol{\beta}}_R) - \boldsymbol{\beta})\right)\\
&= -\boldsymbol{J}_0\boldsymbol{\delta}E\left(I(\chi^2_{r+2,\alpha}(\Delta) \le \chi^2_{r,\alpha})\right), \quad \text{by Lemma 1}\\
&= -\boldsymbol{J}_0\boldsymbol{\delta}H_{r+2}(\chi^2_{r,\alpha}, \Delta).\\
\mathrm{ADB}(\hat{\boldsymbol{\beta}}_S) &= \lim_{N\to\infty} E\left(\sqrt{N}(\hat{\boldsymbol{\beta}_S} - \boldsymbol{\beta})\right)\\
&= \lim_{N\to\infty} E\left(\sqrt{N}(\hat{\boldsymbol{\beta}}_R + (\hat{\boldsymbol{\beta}}_U - \hat{\boldsymbol{\beta}}_R)\right.\\
&\quad \left. -(r-2)D_N^{-1}(\hat{\boldsymbol{\beta}}_U - \hat{\boldsymbol{\beta}}_R) - \boldsymbol{\beta})\right)\\
&= -(r-2)\boldsymbol{J}_0\boldsymbol{\delta}E\left(\chi^{-2}_{r+2,\alpha}(\Delta)\right), \quad \text{by Lemma 1}\\
&= -(r-2)\boldsymbol{J}_0\boldsymbol{\delta}E\left(Z_1^{-1}\right), \quad \text{where } Z_1 = \chi^2_{r+2,\alpha}(\Delta).\\
\mathrm{ADB}(\hat{\boldsymbol{\beta}}_+) &= \lim_{N\to\infty} E\left(\sqrt{N}(\hat{\boldsymbol{\beta}}_+ - \boldsymbol{\beta})\right)\\
&= \lim_{N\to\infty} E\left(\sqrt{N}(\hat{\boldsymbol{\beta}}_S - \boldsymbol{\beta})\right)\\
&\quad - \lim_{N\to\infty} E\left(\sqrt{N}(1-(r-2)D_N^{-1})(\hat{\boldsymbol{\beta}}_U - \hat{\boldsymbol{\beta}}_R)I(D_N < r-2)\right)\\
&= -\boldsymbol{J}_0\boldsymbol{\delta}\{(r-2)E\left(Z_1^{-1}(1 - I(Z_1 < r-2))\right) + H_{r+2}(r-2, \Delta)\},\\
&\quad \text{by Lemma 1.}
\end{aligned}$$

B.2: Proof of Theorem 3.2.1:

The asymptotic MSE of $\hat{\boldsymbol{\beta}}_U$ is

$$\text{MSE}(\hat{\boldsymbol{\beta}}_U) \quad = \lim_{N\to\infty} E\left(N(\hat{\boldsymbol{\beta}}_U - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}}_U - \boldsymbol{\beta})^{\mathsf{T}}\right) = \boldsymbol{B}^{-1}.$$

Hence, it follows that $\text{ADR}(\hat{\boldsymbol{\beta}}_U; \boldsymbol{M}) = \text{trace}\left(\boldsymbol{M}\boldsymbol{B}^{-1}\right)$. The asymptotic MSE of $\hat{\boldsymbol{\beta}}_R$ is

$$\begin{aligned}\text{MSE}(\hat{\boldsymbol{\beta}}_R) &= \lim_{N\to\infty} E\left(N(\hat{\boldsymbol{\beta}}_R - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}}_R - \boldsymbol{\beta})^{\mathsf{T}}\right) \\ &= \lim_{N\to\infty} E\Big(N(\hat{\boldsymbol{\beta}}_U - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}}_U - \boldsymbol{\beta})^{\mathsf{T}} + N\boldsymbol{J}_0(\boldsymbol{A}\hat{\boldsymbol{\beta}}_U - \boldsymbol{c}) \\ &\quad \times (\boldsymbol{A}\hat{\boldsymbol{\beta}}_U - \boldsymbol{c})^{\mathsf{T}}\boldsymbol{J}_0^{\mathsf{T}} - 2N\boldsymbol{J}_0(\boldsymbol{A}\hat{\boldsymbol{\beta}}_U - \boldsymbol{c})(\hat{\boldsymbol{\beta}}_U - \boldsymbol{\beta})^{\mathsf{T}}\Big).\end{aligned}$$

The first term can be written as $\lim_{N\to\infty} E\left(N(\hat{\boldsymbol{\beta}}_U - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}}_U - \boldsymbol{\beta})^{\mathsf{T}}\right) = \boldsymbol{B}^{-1}$, for the second term, let $\boldsymbol{D} = \sqrt{N}(\boldsymbol{A}\hat{\boldsymbol{\beta}} - \boldsymbol{c})$, then

$$\begin{aligned}\lim_{N\to\infty} E\left(N\boldsymbol{J}_0(\boldsymbol{A}\hat{\boldsymbol{\beta}}_U - \boldsymbol{c})(\boldsymbol{A}\hat{\boldsymbol{\beta}}_U - \boldsymbol{c})^{\mathsf{T}}\boldsymbol{J}_0^{\mathsf{T}}\right) &= \lim_{N\to\infty} E\left(\boldsymbol{J}_0\boldsymbol{D}\boldsymbol{D}^{\mathsf{T}}\boldsymbol{J}_0^{\mathsf{T}}\right) \\ &= \boldsymbol{J}_0\boldsymbol{A}\boldsymbol{B}^{-1} + \boldsymbol{J}_0\boldsymbol{\delta}\boldsymbol{\delta}^{\mathsf{T}}\boldsymbol{J}_0^{\mathsf{T}}.\end{aligned}$$

Finally the third term can be written as $-2\lim_{N\to\infty} E\left(N\boldsymbol{J}_0(\boldsymbol{A}\hat{\boldsymbol{\beta}}_U - \boldsymbol{c})(\hat{\boldsymbol{\beta}}_U - \boldsymbol{\beta})^{\mathsf{T}}\right) = -2\boldsymbol{J}_0\boldsymbol{A}\boldsymbol{B}^{-1}$.

By summing the three terms, the ADR of $\hat{\boldsymbol{\beta}}_R$ can be written as

$$\text{ADR}(\hat{\boldsymbol{\beta}}_R; \boldsymbol{M}) \quad = \text{ADR}(\hat{\boldsymbol{\beta}}_U; \boldsymbol{M}) - \text{trace}\left(\boldsymbol{M}\boldsymbol{B}^{-1}\boldsymbol{A}^{\mathsf{T}}\boldsymbol{J}_0^{\mathsf{T}}\right) + \boldsymbol{\delta}^{\mathsf{T}}\boldsymbol{J}_0^{\mathsf{T}}\boldsymbol{M}\boldsymbol{J}_0\boldsymbol{\delta}.$$

The asymptotic MSE of $\hat{\boldsymbol{\beta}}_P$ is

$$\begin{aligned}\text{MSE}(\hat{\boldsymbol{\beta}}_P) &= \lim_{N\to\infty} E\left(N(\hat{\boldsymbol{\beta}}_P - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}}_P - \boldsymbol{\beta})^{\mathsf{T}}\right) \\ &= \lim_{N\to\infty} E\Big(N(\hat{\boldsymbol{\beta}}_U - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}}_U - \boldsymbol{\beta})^{\mathsf{T}} + NI(D_N \le \chi^2_{r,\alpha})(\hat{\boldsymbol{\beta}}_U - \hat{\boldsymbol{\beta}}_R) \\ &\quad \times(\hat{\boldsymbol{\beta}}_U - \hat{\boldsymbol{\beta}}_R)^{\mathsf{T}} - 2NI(D_N \le \chi^2_{r,\alpha})(\hat{\boldsymbol{\beta}}_U - \hat{\boldsymbol{\beta}}_R)(\hat{\boldsymbol{\beta}}_U - \boldsymbol{\beta})^{\mathsf{T}}\Big).\end{aligned}$$

Consider the second term

$$\begin{aligned}
&\lim_{N\to\infty} E\left(NI(D_N \le \chi^2_{r,\alpha})(\hat{\boldsymbol{\beta}}_U - \hat{\boldsymbol{\beta}}_R)(\hat{\boldsymbol{\beta}}_U - \hat{\boldsymbol{\beta}}_R)^\mathsf{T}\right) \\
= \;& \boldsymbol{J}_0 \lim_{N\to\infty} E\left(I(D_N \le \chi^2_{r,\alpha})\boldsymbol{D}\boldsymbol{D}^\mathsf{T}\right)\boldsymbol{J}_0^\mathsf{T} \\
= \;& \boldsymbol{B}^{-1}\boldsymbol{A}^\mathsf{T}\boldsymbol{J}_0^\mathsf{T} H_{r+2}(\chi^2_{r,\alpha},\Delta) + \boldsymbol{J}_0\boldsymbol{\delta}\boldsymbol{\delta}^\mathsf{T}\boldsymbol{J}_0^\mathsf{T} H_{r+4}(\chi^2_{r,\alpha},\Delta), \quad \text{by Lemma 1}
\end{aligned}$$

and the third term

$$\begin{aligned}
&-2\lim_{N\to\infty} E\left(NI(D_N \le \chi^2_{r,\alpha})(\hat{\boldsymbol{\beta}}_U - \hat{\boldsymbol{\beta}}_R)(\hat{\boldsymbol{\beta}}_U - \boldsymbol{\beta})^\mathsf{T}\right) \\
= \;& -2\boldsymbol{J}_0 \lim_{N\to\infty} E\Big(\boldsymbol{D}\sqrt{N}\{(\hat{\boldsymbol{\beta}}_U - \hat{\boldsymbol{\beta}}_R) - \boldsymbol{B}^{-1}\boldsymbol{A}^\mathsf{T}(\boldsymbol{A}\boldsymbol{B}^{-1}\boldsymbol{A}^\mathsf{T})^{-1} \\
& \times (\boldsymbol{A}\boldsymbol{\beta} - \boldsymbol{c})\}^\mathsf{T} I(D_N \le \chi^2_{r,\alpha})\Big) \\
= \;& -2\boldsymbol{J}_0 \lim_{N\to\infty} E\left(\boldsymbol{D}\{\boldsymbol{J}_0\boldsymbol{D} - \boldsymbol{J}_0\sqrt{N}(\boldsymbol{A}\boldsymbol{\beta} - \boldsymbol{c})\}^\mathsf{T} I(D_N \le \chi^2_{r,\alpha})\right) \\
= \;& -2\boldsymbol{B}^{-1}\boldsymbol{A}^\mathsf{T}\boldsymbol{J}_0^\mathsf{T} H_{r+2}(\chi^2_{r,\alpha},\Delta) - 2\boldsymbol{J}_0\boldsymbol{\delta}\boldsymbol{\delta}^\mathsf{T}\boldsymbol{J}_0^\mathsf{T} \\
& \times \left(H_{r+4}(\chi^2_{r,\alpha},\Delta) - H_{r+2}(\chi^2_{r,\alpha},\Delta)\right).
\end{aligned}$$

Hence, The ADR of $\hat{\boldsymbol{\beta}}_P$ is

$$\begin{aligned}
\mathrm{ADR}(\hat{\boldsymbol{\beta}}_P;\boldsymbol{M}) &= \mathrm{trace}\left(\boldsymbol{M}\mathrm{MSE}(\hat{\boldsymbol{\beta}}_P)\right) \\
&= \mathrm{ADR}(\hat{\boldsymbol{\beta}}_U;\boldsymbol{M}) - \mathrm{trace}\left(\boldsymbol{M}\boldsymbol{B}^{-1}\boldsymbol{A}^\mathsf{T}\boldsymbol{J}_0^\mathsf{T}\right) H_{r+2}(\chi^2_{r,\alpha},\Delta) \\
&\quad - \boldsymbol{\delta}^\mathsf{T}\boldsymbol{J}_0^\mathsf{T}\boldsymbol{M}\boldsymbol{J}_0\boldsymbol{\delta}\left(H_{r+4}(\chi^2_{r,\alpha},\Delta) - 2H_{r+2}(\chi^2_{r,\alpha},\Delta)\right).
\end{aligned}$$

The asymptotic MSE of $\hat{\boldsymbol{\beta}}_S$ is

$$\begin{aligned}
\mathrm{MSE}(\hat{\boldsymbol{\beta}}_S) &= \lim_{N\to\infty} E\left(N(\hat{\boldsymbol{\beta}}_S - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}}_S - \boldsymbol{\beta})^\mathsf{T}\right) \\
&= \lim_{N\to\infty} E\Big(N(\hat{\boldsymbol{\beta}}_U - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}}_U - \boldsymbol{\beta})^\mathsf{T} + N(r-2)^2 D_N^{-2}(\hat{\boldsymbol{\beta}}_U - \hat{\boldsymbol{\beta}}_R) \\
&\quad \times(\hat{\boldsymbol{\beta}}_U - \hat{\boldsymbol{\beta}}_R)^\mathsf{T} - 2N(r-2)D_N^{-1}(\hat{\boldsymbol{\beta}}_U - \hat{\boldsymbol{\beta}}_R)(\hat{\boldsymbol{\beta}}_U - \boldsymbol{\beta})^\mathsf{T}\Big).
\end{aligned}$$

Consider the second term

$$\begin{aligned}
&\lim_{N\to\infty} E\left(N(r-2)^2 D_N^{-2}(\hat{\boldsymbol{\beta}}_U - \hat{\boldsymbol{\beta}}_R)(\hat{\boldsymbol{\beta}}_U - \hat{\boldsymbol{\beta}}_R)^\mathsf{T}\right) \\
= \;& (r-2)^2 \lim_{N\to\infty} \boldsymbol{J}_0 E\left(D_N^{-2}\boldsymbol{D}\boldsymbol{D}^\mathsf{T}\right)\boldsymbol{J}_0^\mathsf{T} \\
= \;& (r-2)^2\boldsymbol{B}^{-1}\boldsymbol{A}^\mathsf{T}\boldsymbol{J}_0^\mathsf{T} E(Z_1^{-2}) + (r-2)^2\boldsymbol{J}_0\boldsymbol{\delta}\boldsymbol{\delta}^\mathsf{T}\boldsymbol{J}_0^\mathsf{T} E(Z_2^{-2}), \\
& \text{where } Z_2 = \chi^2_{r+4,\alpha}(\Delta)).
\end{aligned}$$

The third term can be expressed as

$$\begin{aligned}
& -2(r-2)\lim_{N\to\infty} E\left(ND_N^{-1}(\hat{\boldsymbol{\beta}}_U-\hat{\boldsymbol{\beta}}_R)(\hat{\boldsymbol{\beta}}_U-\boldsymbol{\beta})^{\mathsf{T}}\right)\\
= \;& -2(r-2)\boldsymbol{J}_0\lim_{N\to\infty} E\left(D_N^{-1}\boldsymbol{D}\sqrt{N}(\hat{\boldsymbol{\beta}}_U-\boldsymbol{\beta})^{\mathsf{T}}\right)\\
= \;& -2(r-2)\boldsymbol{J}_0\lim_{N\to\infty} E\left(\boldsymbol{D}\boldsymbol{D}^{\mathsf{T}}D_N^{-1}\right)\boldsymbol{J}_0^{\mathsf{T}}+2(r-2)\boldsymbol{J}_0\lim_{N\to\infty} E\left(\boldsymbol{D}D_N^{-1}\right)\\
& \times\sqrt{N}(\boldsymbol{A}\boldsymbol{\beta}-\boldsymbol{c})^{\mathsf{T}}\boldsymbol{J}_0^{\mathsf{T}}\\
= \;& -2(r-2)\boldsymbol{B}^{-1}\boldsymbol{A}^{\mathsf{T}}\boldsymbol{J}_0^{\mathsf{T}}E(Z_1^{-1})-2(r-2)\boldsymbol{J}_0\boldsymbol{\delta}\boldsymbol{\delta}^{\mathsf{T}}\boldsymbol{J}_0^{\mathsf{T}}\left(E(Z_2^{-1}-Z_1^{-1})\right),\\
& \text{by Lemma 1.}
\end{aligned}$$

Hence, the ADR of $\hat{\boldsymbol{\beta}}_S$ is

$$\begin{aligned}
\mathrm{ADR}(\hat{\boldsymbol{\beta}}_S;\boldsymbol{M}) &= \mathrm{trace}\left(\boldsymbol{M}\mathrm{MSE}(\hat{\boldsymbol{\beta}}_S)\right)\\
&= \mathrm{ADR}(\hat{\boldsymbol{\beta}}_U;\boldsymbol{M})+(r-2)\mathrm{trace}\left(\boldsymbol{M}\boldsymbol{B}^{-1}\boldsymbol{A}^{\mathsf{T}}\boldsymbol{J}_0^{\mathsf{T}}\right)\\
&\quad\times\left((r-2)E(Z_1^{-2})-2E(Z_1^{-1})\right)+(r-2)\\
&\quad\times\left((r-2)E(Z_2^{-2})-2E(Z_2^{-1}-Z_1^{-1})\right)\boldsymbol{\delta}^{\mathsf{T}}\boldsymbol{J}_0^{\mathsf{T}}\boldsymbol{M}\boldsymbol{J}_0\boldsymbol{\delta}.
\end{aligned}$$

The asymptotic MSE of $\hat{\boldsymbol{\beta}}_+$ is

$$\begin{aligned}
\mathrm{MSE}(\hat{\boldsymbol{\beta}}_+) &= \lim_{N\to\infty} E\left(N(\hat{\boldsymbol{\beta}}_+-\boldsymbol{\beta})(\hat{\boldsymbol{\beta}}_+-\boldsymbol{\beta})^{\mathsf{T}}\right)\\
&= \mathrm{MSE}(\hat{\boldsymbol{\beta}}_S)+\lim_{N\to\infty} E\left(N(1-(r-2)D_N^{-1})^2(\hat{\boldsymbol{\beta}}_U-\hat{\boldsymbol{\beta}}_R)\right.\\
&\quad\left.\times(\hat{\boldsymbol{\beta}}_U-\hat{\boldsymbol{\beta}}_R)^{\mathsf{T}}I(D_N<r-2)\right)\\
&\quad-2\lim_{N\to\infty} E\left(N(1-(r-2)D_N^{-1})(\hat{\boldsymbol{\beta}}_U-\hat{\boldsymbol{\beta}}_R)\{(\hat{\boldsymbol{\beta}}_R-\boldsymbol{\beta})\right.\\
&\quad\left.+(1-(r-2)D_N^{-1})(\hat{\boldsymbol{\beta}}_U-\hat{\boldsymbol{\beta}}_R)\}^{\mathsf{T}}I(D_N<r-2)\right)\\
&= \mathrm{MSE}(\hat{\boldsymbol{\beta}}_S)-\lim_{N\to\infty} E\left(N(1-(r-2)D_N^{-1})^2(\hat{\boldsymbol{\beta}}_U-\hat{\boldsymbol{\beta}}_R)\right.\\
&\quad\left.\times(\hat{\boldsymbol{\beta}}_U-\hat{\boldsymbol{\beta}}_R)^{\mathsf{T}}I(D_N<r-2)\right)\\
&\quad-2\lim_{N\to\infty} E\left(N(1-(r-2)D_N^{-1})I(D_N<r-2)\right.\\
&\quad\left.\times(\hat{\boldsymbol{\beta}}_U-\hat{\boldsymbol{\beta}}_R)(\hat{\boldsymbol{\beta}}_R-\boldsymbol{\beta})^{\mathsf{T}}\right).
\end{aligned}$$

Consider the second term

$$\begin{aligned} & -\lim_{N\to\infty} E\left(N(1-(r-2)D_N^{-1})^2(\hat{\boldsymbol{\beta}}_U-\hat{\boldsymbol{\beta}}_R)(\hat{\boldsymbol{\beta}}_U-\hat{\boldsymbol{\beta}}_R)^{\mathsf{T}} I(D_N<r-2)\right) \\ = \; & -\boldsymbol{B}^{-1}\boldsymbol{A}^{\mathsf{T}}\boldsymbol{J}_0^{\mathsf{T}} E\left((1-(r-2)Z_1^{-1})^2 I(Z_1<r-2)\right) \\ & -\boldsymbol{J}_0\boldsymbol{\delta}\boldsymbol{\delta}^{\mathsf{T}}\boldsymbol{J}_0^{\mathsf{T}} E\left((1-(r-2)Z_2^{-1})^2 I(Z_2<r-2)\right). \end{aligned}$$

Now consider the third term

$$\begin{aligned} & -2\lim_{N\to\infty} E\left(N(1-(r-2)D_N^{-1})I(D_N<r-2)(\hat{\boldsymbol{\beta}}_U-\hat{\boldsymbol{\beta}}_R)(\hat{\boldsymbol{\beta}}_R-\boldsymbol{\beta})^{\mathsf{T}}\right) \\ = \; & -2\boldsymbol{J}_0\lim_{N\to\infty} E\left((1-(r-2)D_N^{-1})I(D_N<r-2)\boldsymbol{D}\sqrt{N}(\hat{\boldsymbol{\beta}}_R-\boldsymbol{\beta})^{\mathsf{T}}\right) \\ = \; & 2\boldsymbol{J}_0\boldsymbol{\delta}\boldsymbol{\delta}^{\mathsf{T}}\boldsymbol{J}_0^{\mathsf{T}} E\left((1-(r-2)Z_1^{-1})I(Z_1<r-2)\right), \quad \text{by Lemma 1.} \end{aligned}$$

By collecting all of the terms, it then follows that

$$\begin{aligned} \mathrm{ADR}(\hat{\boldsymbol{\beta}}_+;\boldsymbol{M}) &= \mathrm{trace}\left(\boldsymbol{M}\mathrm{MSE}(\hat{\boldsymbol{\beta}}_+)\right) \\ &= \mathrm{ADR}(\hat{\boldsymbol{\beta}}_S;\boldsymbol{M}) - \mathrm{trace}\left(\boldsymbol{M}\boldsymbol{B}^{-1}\boldsymbol{A}^{\mathsf{T}}\boldsymbol{J}_0^{\mathsf{T}}\right) \\ &\quad \times E\{(1-(r-2)Z_1^{-1})^2 I(Z_1<r-2)\} \\ &\quad -E\left((1-(r-2)Z_2^{-1})^2 I(Z_2<r-2)\right)\boldsymbol{\delta}^{\mathsf{T}}\boldsymbol{J}_0^{\mathsf{T}}\boldsymbol{M}\boldsymbol{J}_0\boldsymbol{\delta} \\ &\quad +2E\left((1-(r-2)Z_1^{-1})I(Z_1<r-2)\right)\boldsymbol{\delta}^{\mathsf{T}}\boldsymbol{J}_0^{\mathsf{T}}\boldsymbol{M}\boldsymbol{J}_0\boldsymbol{\delta}. \end{aligned}$$

T. Thomson
University of Winnipeg,
Winnipeg, MB, Canada

S. Hossain
Department of Mathematics and
Statistics, University of Winnipeg,
515 Portage Ave, Winnipeg, MB,
Canada
E-mail: sh.hossain@uwinnipeg.ca

Paper received: 19 July 2016.