Linear Algebra and its Applications 506 (2016) 603–613

Contents lists available at ScienceDirect

Linear Algebra and its Applications

www.elsevier.com/locate/laa

Identifiability of covariance parameters in linear mixed effects models

Wei Wang

Department of Surgical Outcomes and Analysis, Kaiser Permanente, San Diego, USA

A R T I C L E   I N F O

Article history:
Received 22 November 2015
Accepted 10 June 2016
Available online 17 June 2016
Submitted by W.B. Wu

MSC:
15A03
15A18
15B05
62J99

Keywords:
Covariance matrix structures
Identifiability
Mixed effects models
Random effects
Variance components

A B S T R A C T

Building a linear mixed model often involves selection of the parametrized covariance matrix structures for the random components of the model. Parameters in the covariance matrix of the response then consist of those from the random effects and from the random residual error. However, some specifications of the structures can result in the parameters not identifiable, even if the model is not over-parametrized. Software output can look normal with no indication of error when fitting non-identifiable models. In our simulation studies, we found no implication of model non-identifiable about half of the times. We derive model identifiability conditions which only rely on properties of the known design matrix associated with the random effects and the specific structures being used. The results can be applied to study identifiability for commonly used covariance matrix structures.

1. Introduction

Data of a clustered structure, or a longitudinal or repeated-measures structure are often encountered in a variety of disciplines such as agriculture, biology, medicine, and

E-mail address: wei.wang@kp.org.

http://dx.doi.org/10.1016/j.laa.2016.06.022

social sciences. A linear mixed effects model (LMM) may be well suited for the analysis of these types of data; see for instance, [1–10]. Let $\mathbf{y}$ be the response vector and let $\mathbf{X}$ and $\mathbf{Z}$ be known, non-random design matrices. Let $\boldsymbol{\beta}$ be the fixed effects, $\mathbf{u}$ be the random effects and $\boldsymbol{\epsilon}$ be the random residual error. A linear mixed model is written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\epsilon},$$
$$\mathbf{u} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}_u),\ \boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}_\epsilon),\ \mathbf{u} \text{ is independent of } \boldsymbol{\epsilon}. \tag{1}$$

The covariance matrices $\boldsymbol{\Sigma}_u$ and $\boldsymbol{\Sigma}_\epsilon$ are usually parametrized and of certain structures. For instance, a multiple of an identity (MI) structured $\boldsymbol{\Sigma}_\epsilon$ has the form $\sigma^2\mathbf{I}$ with unknown parameter $\sigma^2 > 0$. Let $\boldsymbol{\Sigma}_u$ and $\boldsymbol{\Sigma}_\epsilon$ be parametrized by $\boldsymbol{\theta}_u$ and $\boldsymbol{\theta}_\epsilon$ with parameter spaces $\boldsymbol{\Theta}_u$ and $\boldsymbol{\Theta}_\epsilon$ respectively. It follows that the covariance matrix of $\mathbf{y}$ is $\boldsymbol{\Sigma}_y(\boldsymbol{\theta}_u, \boldsymbol{\theta}_\epsilon) = \mathbf{Z}\boldsymbol{\Sigma}_u(\boldsymbol{\theta}_u)\mathbf{Z}' + \boldsymbol{\Sigma}_\epsilon(\boldsymbol{\theta}_\epsilon)$ with unknown parameters $\boldsymbol{\theta}_u$ and $\boldsymbol{\theta}_\epsilon$ to be estimated.

To build an LMM, one often needs to choose covariance structures for $\boldsymbol{\Sigma}_u$ and $\boldsymbol{\Sigma}_\epsilon$ [7, 11–13]. Fitting an LMM with a variety of covariance structures to choose has been provided by statistical software such as R, SAS, SPSS, STATA and HLM [14–17]. In choosing $\boldsymbol{\Sigma}_u(\boldsymbol{\theta}_u)$ and $\boldsymbol{\Sigma}_\epsilon(\boldsymbol{\theta}_\epsilon)$, certain combination may lead to parameters $\boldsymbol{\theta}_u$ and $\boldsymbol{\theta}_\epsilon$ in $\boldsymbol{\Sigma}_y$ not identifiable, even if $\boldsymbol{\Sigma}_y$ is not over-parametrized. That is, there may exist $\boldsymbol{\theta}_u^* \in \boldsymbol{\Theta}_u$ ($\boldsymbol{\theta}_u^* \neq \boldsymbol{\theta}_u$) and $\boldsymbol{\theta}_\epsilon^* \in \boldsymbol{\Theta}_\epsilon$ ($\boldsymbol{\theta}_\epsilon^* \neq \boldsymbol{\theta}_\epsilon$) such that $\boldsymbol{\Sigma}_y(\boldsymbol{\theta}_u, \boldsymbol{\theta}_\epsilon) = \boldsymbol{\Sigma}_y(\boldsymbol{\theta}_u^*, \boldsymbol{\theta}_\epsilon^*)$. Since $\mathbf{u}$ and $\boldsymbol{\epsilon}$ are both normally distributed, so is $\mathbf{y}$. Thus, distinct parameter values produce the same $\boldsymbol{\Sigma}_y$ component of the likelihood of $\mathbf{y}$, and result in model non-identifiability. The best linear unbiased predictor (BLUP) $\hat{\mathbf{u}} = \boldsymbol{\Sigma}_u(\boldsymbol{\theta}_u)\mathbf{Z}'\boldsymbol{\Sigma}_y^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$ and the residual $\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\hat{\mathbf{u}}$ are often used in model checks and diagnostics [14,17]. When the model is not identifiable, the BLUP and the residual can be $\hat{\mathbf{u}}^* = \boldsymbol{\Sigma}_u(\boldsymbol{\theta}_u^*)\mathbf{Z}'\boldsymbol{\Sigma}_y^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$ and $\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\hat{\mathbf{u}}^*$. In principle, model identifiability shall be checked before fitting [12], is viewed as a necessary property for the adequacy of a statistical model [1], and is closely related to parameter estimability [18].

In practice, there are N individuals and each response vector $\mathbf{y}_i$ is modeled as in (1). The $\mathbf{u}_i$'s and $\boldsymbol{\epsilon}_i$'s are mutually independent with associated covariance matrices $\boldsymbol{\Sigma}_u$ and $\boldsymbol{\Sigma}_{\epsilon i}$. Covariance parameters in the joint model of all individuals are then ($\boldsymbol{\theta}_u, \boldsymbol{\theta}_{\epsilon i}$, $i = 1, \ldots, N$). It is shown that the joint model is identifiable if and only if at least one individual model is identifiable [19]. Checking identifiability of the joint model is then reduced to the checking of an individual model (1).

We study identifiability of the covariance parameters ($\boldsymbol{\theta}_u, \boldsymbol{\theta}_\epsilon$) in the LMM (1) when $\boldsymbol{\Sigma}_y$ is not over-parametrized. It is clear if $\mathbf{X}$ is of full column rank, there is no $\boldsymbol{\beta}^* \neq \boldsymbol{\beta}$ producing the same mean vector of $\mathbf{y}$. For simplicity, we do not distinguish identifiability of the covariance parameters and identifiability of the model parameters, and use the two terms interchangeably. [19] studies model identifiability focusing on a few structures or a special 2×2 $\mathbf{Z}$. Our results extend those of [19] and can be applied to study a variety of structures.

Novel results are presented in Section 3, where we first provide a characterization of the matrix $\mathbf{H}_\mathbf{Z} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$ when the model is not identifiable. Based on the character-