Statistical Methodology 32 (2016) 203–217

Contents lists available at ScienceDirect

Statistical Methodology

journal homepage: www.elsevier.com/locate/stamet

Systematic deviation in smooth mixed models for multi-level longitudinal data

Viani A. Biatat Djeundje

University of Edinburgh, United Kingdom

A R T I C L E I N F O

Article history:
Received 11 May 2014
Received in revised form
12 May 2016
Accepted 12 May 2016
Available online 27 May 2016

Keywords:
Centring
Constraints
Covariance
Multi-level
Smoothing

A B S T R A C T

The analysis of longitudinal data or repeated measurements is an important and growing area of Statistics. In this context, data come in different formats but typically, they have a hierarchical or multi-level structure including group and subject components, and the main purpose of the analysis is usually to estimate these components from the data. A standard way to perform this estimation is via mixed models. In this paper, we show that the estimated group effects from standard smooth mixed models can deviate systematically from the underlying group mean, leading to wrong conclusions about the data. We then present two ways to avoid such systematic deviations and misinterpretations when fitting flexible mixed models to multi-level data. The first method is a marginal procedure, and the second method is based on the conditional distribution of the subject effects derived from appropriate constraints. Both methods are robust against mis-specification of the covariance structure in the sense that they allow one to resolve the lack of centring found in standard smooth mixed models.

1. Introduction

Longitudinal data arise widely in Medicine, Psychology, Social Sciences, etc. In a longitudinal study, repeated observations are recorded on a number of subjects over time with grouped or nested structures, causing traditional model assumptions of independence and homogeneity to fail.

E-mail address: viani@aims.ac.za.

http://dx.doi.org/10.1016/j.stamet.2016.05.003
1572-3127/

A solution to this problem is found in the use of mixed models, which allow the appropriate investigation of data with more complex, multi-level and hierarchical structures.

The origin of mixed models can be traced back to Fisher [13] who introduced random effects models to study the correlations of trait values between relatives. Subsequently, Henderson et al. [17] brought up the concept of best linear unbiased predictions (BLUE), and, following the work of Harville [15] and Laird and Ware [18], mixed models attracted considerable attention and became a major area of Statistics. Detailed exposure and recent developments can be found in [6,26,27,1], among others. In essence, measurements on each subject are treated as a cluster, and incorporation of multi-level random effects allows one to make inference using a combination of data from all subjects.

Mixed models can be divided broadly into two groups: (i) the parametric framework in which appropriate parametric functions are sufficient to provide a good description of the data [21,6], and (ii) the semiparametric or smooth framework where one needs to allow room for flexibility in order to capture complex or hidden patterns in the data [2,9,27]. In both cases, the model is made up of two components: the fixed (group) effects and the random (subject) effects. A key issue that arises from such a representation is the appropriate identification of these two effects. For example, when the subject effects are meant to quantify departures from a mean or group effect, one would expect an ANOVA-type condition to be fulfilled: that is, the estimated or predicted subject effects are broadly centred at each time point. A violation of this centring requirement can result in estimated group and subject effects that *deviate systematically* from the data.

Under mild conditions (as we shall see in Section 2.1), it can be shown that the BLUP of the subject effects from parametric mixed models fulfil the centring requirement irrespective of the covariance structure. However, it is not the case for smooth mixed models. Indeed, for many datasets, widely used smooth mixed models for multi-level longitudinal data yield (a) estimated group and subject effects that deviate severely from the underlying data and (b) standard errors with undesirable properties, as we shall elaborated in detail in Section 2.2.

These problems surrounding smooth mixed models were considered by Heckman et al. [16]. Whilst the authors advocated the use of a sandwich procedure to improve the estimates of standard errors, they provided no solution to the *systematic deviation* problem pertaining to estimates of the group and subject effects in the first place. Djeundje and Currie [8] raised the same concerns and described a way to tackle the problem via multiple penalties. However, using penalty arguments is equivalent to assuming a particular covariance structure, and it possible to construct covariance structures that deviate from those based on penalty arguments. In other words, a penalty-based covariance structure can become inappropriate, resulting in a model with undesirable properties. In the context of parametric mixed models for instance, Gurka et al. [14] showed how incorrectly assuming a compound symmetric covariance structure can inflate type I error for inference about the fixed effects.

In this paper, we present two methods that can be used to avoid *systematic deviations* when fitting flexible or smooth mixed models to multi-level data. The first method is a marginal procedure, and the second method is a constrained mixed model in which the joint prior distribution of the subject effects is conditional on appropriate constraints. Unlike the covariance driven approach, these two methods are robust against mis-specified covariance structures in the sense that they allow to ensure appropriate identification of the group and subject effects (as we shall see in Section 3).

The paper is structured as follows. Section 2 introduces some notations, explores parametric mixed models, and raises concerns about smooth mixed models. Section 3 presents ways to handle the problems exposed in Section 2. A discussion and some concluding remarks follow in Section 4.

2. Mixed models and identification

Consider a longitudinal study involving a sample of n subjects in which y_{ij} denotes a continuous random variable observed on subject i at time point $t_{ij}, j = 1, \ldots, n_i$. The values of this variable may be driven by several covariates, some being time-dependent and others measured only at the baseline. In general, let $\boldsymbol{x}_{ij}$ designate the vector of covariates values on subject i at time t_{ij}. We will represent the response vector on subject i by $\boldsymbol{y}_i$, the corresponding time vector by $\boldsymbol{t}_i$, and the associated covariate vector or matrix by $\boldsymbol{x}_i$. A well known modelling framework for such data is

$$\boldsymbol{y}_i = \mu(\boldsymbol{t}_i, \boldsymbol{x}_i) + f_i(\boldsymbol{t}_i, \boldsymbol{x}_i) + \boldsymbol{\varepsilon}_i, \quad i = 1, \ldots, n, \tag{2.1}$$