Computational Statistics and Data Analysis 103 (2016) 151–169

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Linear mixed models with marginally symmetric nonparametric random effects

Hien D. Nguyen *, Geoffrey J. McLachlan

School of Mathematics and Physics, University of Queensland, Australia

ARTICLE INFO

Article history:
Received 16 April 2015
Received in revised form 10 October 2015
Accepted 6 May 2016
Available online 19 May 2016

Keywords:
Linear mixed models
Nonparametric maximum likelihood
Marginal symmetry
Random effects
Mixture model
ECM algorithm

ABSTRACT

Linear mixed models (LMMs) are used as an important tool in the data analysis of repeated measures and longitudinal studies. The most common form of LMMs utilizes a normal distribution to model the random effects. Such assumptions can often lead to misspecification errors when the random effects are not normal. One approach to remedy the misspecification errors is to utilize a point-mass distribution to model the random effects; this is known as the nonparametric maximum likelihood-fitted (NPML) model. The NPML model is flexible but requires a large number of parameters to characterize the random-effects distribution. It is often natural to assume that the random-effects distribution be at least marginally symmetric. The marginally symmetric NPML (MSNPML) random-effects model is introduced, which assumes a marginally symmetric point-mass distribution for the random effects. Under the symmetry assumption, the MSNPML model utilizes half the number of parameters to characterize the same number of point masses as the NPML model; thus the model confers an advantage in economy and parsimony. An EM-type algorithm is presented for the maximum likelihood (ML) estimation of LMMs with MSNPML random effects; the algorithm is shown to monotonically increase the log-likelihood and is proven to be convergent to a stationary point of the log-likelihood function in the case of convergence. Furthermore, it is shown that the ML estimator is consistent and asymptotically normal under certain conditions, and the estimation of quantities such as the random-effects covariance matrix and individual *a posteriori* expectations is demonstrated. A simulation study is used to illustrate the gains in efficiency of the MSNPML model over the NPML model under the assumption of symmetry. A pair of real data applications are then used to demonstrate the manner in which the MSNPML model can be used to draw useful statistical inference.

1. Introduction

Linear mixed models (LMMs) are used as a fundamental tool for the statistical analysis of longitudinal data and data with repeated measurements; see McCulloch and Searle (2001, Ch. 6), Pinheiro and Bates (2000, Ch. 1), and Verbeke and Molenberghs (2000) for introductions on the topic. In the style of Laird and Ware (1982), an LMM can be characterized as follows.

Let $\boldsymbol{Y}_j = \left(Y_{j1}, \ldots, Y_{jn_j}\right)^T$ be a vector of n_j responses belonging to individual j, for $j = 1, \ldots, n$, where n is the total number of individuals. Further, let each measurement Y_{jk}, for $k = 1, \ldots, n_j$, be dependent upon covariate vectors $\boldsymbol{x}_{jk} \in \mathbb{R}^p$ and

* Corresponding author.
E-mail address: h.nguyen7@uq.edu.au (H.D. Nguyen).

http://dx.doi.org/10.1016/j.csda.2016.05.005

$\boldsymbol{z}_{jk} \in \mathbb{R}^q$, and for each j, let $\boldsymbol{B}_j \in \mathbb{R}^q$ be a vector of individual random effects arising from the *a priori* probability distribution $F_{\boldsymbol{B}}(\boldsymbol{b})$ with density $f_{\boldsymbol{B}}(\boldsymbol{b})$. We say that the data arises from an LMM if for each j and k,

$$Y_{jk} | \left(\boldsymbol{B}_j = \boldsymbol{b}_j\right) = \boldsymbol{x}_{jk}^T \boldsymbol{\beta} + \boldsymbol{z}_{jk}^T \boldsymbol{b}_j + E_{jk}, \tag{1}$$

where $\boldsymbol{\beta} \in \mathbb{R}^p$ is a vector of fixed effects and E_{jk} is a random error with probability density $f_E(e)$. Here, a superscript T indicates matrix transposition.

The main difficulty that arises in the use of LMMs is the evaluation and manipulation of the marginal density of $\boldsymbol{y}_j$, which has the general form

$$f_{\boldsymbol{Y}_j}\left(\boldsymbol{y}_j\right) = \int_{\mathbb{R}^q} \left[\prod_{k=1}^{n_k} f_E\left(y_{jk} - \boldsymbol{x}_{jk}^T\boldsymbol{\beta} - \boldsymbol{z}_{jk}^T\boldsymbol{b}_j\right)\right] f_{\boldsymbol{B}}\left(\boldsymbol{b}_j\right) \mathrm{d}\boldsymbol{b}_j, \tag{2}$$

and can often be quite complex due to the integration form.

The traditional approach in dealing with the complexities of (2) is to set the error density as

$$f_E(e) = \phi\left(e; 0, \sigma^2\right), \tag{3}$$

where

$$\phi\left(e; \mu, \sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(e-\mu)^2}{2\sigma^2}\right]$$

is the normal density function with mean μ and variance σ^2, and to let $f_{\boldsymbol{B}}(\boldsymbol{b})$ be a multivariate normal density. This approach results in $f_{\boldsymbol{Y}_j}\left(\boldsymbol{y}_j\right)$ having the form of a multivariate normal density function, and allows for simple inference by maximum likelihood (ML) estimation via an expectation–maximization (EM) algorithm; see Laird and Ware (1982, Sec. 4) and McLachlan and Krishnan (2008, Sec. 5.9) for details.

It is well known that the estimation of the fixed effects $\boldsymbol{\beta}$ is robust to the specification of the random-effects distribution. However, this robustness does not extend to the characterization of the random effects in the case of misspecification. The robustness as well as the effects of misspecification are explored in Agresti et al. (2004), Butler and Louis (1992), and McCulloch and Neuhaus (2011). For instance, in all three articles, the authors note that the estimates for the fixed effects tended not to be influenced greatly by the choice of the random-effects model. However, Agresti et al. (2004) note that the usual normal model can be highly inefficient when the true random-effects model is polarizing, such as in the case of binary or mixture random-effects distributions. It is further remarked in McCulloch and Neuhaus (2011) that the estimation of random intercept coefficients can be biased by making incorrect assumptions regarding the shape of the underlying random-effects distribution.

Multiple strategies have been considered for remedying random-effects misspecification. For example, Pinheiro et al. (2001) and Song et al. (2007) considered t distributed random-effects and noise models, whereas Arellano-Valle et al. (2005), Lachos et al. (2010), and Ho and Lin (2010) considered the use of skew normal and t distributed random and noise models. Although a rich class, the use of such distributions often do not allow for simplification of the marginal density (2) and do not allow for enough flexibility to model random-effects distributions with multiple modes or deviations from bell-shaped curves.

Based on the nonparametric maximum likelihood (NPML) principle of Laird (1978), Aitkin (1999) and Butler and Louis (1992) suggested using the point-mass density

$$f_{\boldsymbol{B}}(\boldsymbol{b}) = \sum_{i=1}^{g} \pi_i \delta\left(\boldsymbol{b} - \boldsymbol{\lambda}_i\right) \tag{4}$$

for the random effects, where $\delta(\boldsymbol{x})$ is the Dirac delta function, $g \geq 1$ is the number of point masses, and $\boldsymbol{\lambda}_i \in \mathbb{R}^q$ and $\pi_k > 0$ are the point-mass locations and weights, for $i = 1, .., g$, respectively. To ensure that the total probability of the point masses adds up to unity and that the mean of $f_{\boldsymbol{B}}(\boldsymbol{b})$ is zero, we also require the restrictions that $\sum_{i=1}^{g} \pi_i = 1$ (which implies that $\pi_g = 1 - \sum_{i=1}^{g-1} \pi_i$) and $\sum_{i=1}^{g} \pi_i \boldsymbol{\lambda}_i = \boldsymbol{0}$, where $\boldsymbol{0}$ is a zero vector of appropriate size. We shall refer to densities of the form (4) as NPML-fitted (NPML) densities.

In Agresti et al. (2004) it was shown that NPML densities offered improvements in efficiency for estimating the characteristics of the random-effects density such as the covariance and individual *a posteriori* estimates of the random effects (i.e. $\mathbb{E}\left(\boldsymbol{B}_j | \boldsymbol{Y}_j = \boldsymbol{y}_j\right)$) when compared to the use of parametric random-effects densities in situations, where the true random-effect densities deviated from the assumed parametric form. However, this improvement comes at a cost of modeling complexity, since for any given g and q, the number of parameters required for the specification and estimation of the NPML density is $(g-1)(q+1)$. This number can grow quickly if g or q are large.

We note that although unimodality or bell shape cannot be assumed in general, it may still be acceptable to assume symmetry in the random-effects distribution. Thus, we introduce the marginally symmetric NPML (MSMPML)