

Model Based Clustering for Longitudinal Data

Rolando De la Cruz-Mesía* Fernando A. Quintana, Guillermo Marshall†

April 2, 2007

Abstract

A model-based clustering method is proposed for clustering individuals on the basis of measurements taken over time. Data variability is taken into account through non-linear hierarchical models leading to a mixture of hierarchical models. We study both frequentist and Bayesian estimation procedures. From a classical viewpoint, we discuss maximum likelihood estimation of this family of models through the EM algorithm. From a Bayesian standpoint, we develop appropriate Markov chain Monte Carlo (MCMC) sampling schemes for the exploration of target posterior distribution of parameters. The methods are illustrated with the identification of hormone trajectories that are likely to lead to adverse pregnancy outcomes in a group of pregnant women.

Keywords: EM-algorithm, Cluster analysis, Markov chain Monte Carlo, Mixture model, Non-linear models, Random effects.

1 Introduction

The use of mixture models for clustering is sometimes referred to as model-based probabilistic clustering (Fraley and Raftery, 1998, 2002), since a particular functional form for the component densities must be assumed. Finite mixture models are widely used for clustering data in a variety of applications (see McLachlan and Basford, 1988).

Many standard clustering algorithms are based on the assumption that the vectors to be clustered are realizations of random vectors from some parametric statistical model. These models usually place no restriction on the mean structure via covariates or otherwise. However, in many applications there is potential for parsimonious representation of the mean. For example, medical studies often yield time series-type data where each d -dimensional vector consists of measurements at d different time points. In such cases, it seems natural to model the mean via regression and we will show that

*Departamento de Salud Pública, Facultad de Medicina, Pontificia Universidad Católica de Chile, Marcoleta 434, Santiago, Casilla 114D, CHILE. rolando@med.puc.cl. Partially funded by grant FONDECYT 3060071

†Departamento de Estadística, Facultad de Matemáticas, Pontificia Universidad Católica de Chile, Casilla 306, Correo 22, Santiago, CHILE. {quintana,gm}@mat.puc.cl. Partially funded by grants FONDECYT 1060729 and 1060721

there is a decided advantage in doing so, specially when tempered with the ability to detect clusters that are well defined but deviate from the model.

In longitudinal medical studies, measurements taken over time on individuals usually show a highly unbalanced structure, e.g., measurement times may be unequally spaced within a individual and may differ across individuals. Traditionally, clustering algorithms, such as K -means, have operated on points or on feature vectors of fixed-dimensional size (Hartigan, 1975). In contrast, however, data from longitudinal studies do not frequently come in a convenient fixed-dimensional form. Thus, model-based clustering has some inherent advantages compared to non-probabilistic clustering techniques (see Li, 2006). In addition to providing a generative and predictive model for the data, such methodology can conveniently handle missing and irregularly spaced measurements. Also, one key advantage of using a model-based approach is that, in addition to the clustering itself, we obtain a measure of uncertainty for the assignment of each individual via the posterior probabilities of cluster membership (see (10) and (11) in Section 5). Fraley and Raftery (2002) have shown effectiveness of model-based clustering in a number of practical applications including clustering of medical data, gene expression data, web-logs data, image data, and spatial data.

In this paper, we formulate a class of regression models based on the mixture of hierarchical nonlinear models. This class can be viewed as an extension of finite mixtures of nonlinear models in which cluster specific random effects are included to account for within-cluster variability. Thus, the finite mixture of hierarchical nonlinear models provides formal estimates of individual and population probabilities of membership in each cluster, population level fixed effects for each cluster, and individual-specific random effects for each cluster conditional on membership.

Most parameter estimation methods for mixture models can be classified into two categories. One is the likelihood-based approach and the other is the Bayesian approach. Maximum likelihood estimation is greatly facilitated by the EM algorithm, while the Bayesian approach has benefited from the development of the Gibbs sampler. With the EM algorithm, latent variables (or “missing data”) are introduced, which allows finite mixture models to be fit by iteratively fitting weighted versions of the component models. So, for example, a K component finite mixture of nonlinear models can be fit via maximum likelihood (ML) by fitting K weighted nonlinear models, updating the weights and iterating to convergence. Mixture models with random effects pose an additional challenge to ML estimation as the marginal likelihood involves an integral that cannot be typically evaluated in closed form. This challenge is similar to that found with ordinary (non-mixture) hierarchical nonlinear models. Estimation in a Bayesian framework is now feasible using posterior simulation via MCMC methods. Bayes estimators for mixture models are well defined as long as the prior distributions are proper (Roeder and Wasserman, 1997). Important papers on the Bayesian analysis of mixture following MCMC methods include Diebolt and Robert (1994) and Escobar and West (1995).

In this article, we study both frequentist and Bayesian approaches for parameter estimation. The maximum likelihood estimation is carried out via the Monte Carlo EM algorithm. From a Bayesian viewpoint, we outline a sampling strategy for fitting the models, which consists of a sequence of

Gibbs and Metropolis-Hastings steps, where the latter is required when no closed form is available in some full conditional in the implementation of the Gibbs sampling algorithm.

The rest of this paper is organized as follows. In Section 2 we review related work about mixture models. We formulate the component mixture of non-linear hierarchical models in Section 3. In Section 4, we outline the EM algorithm and the Bayesian framework via MCMC methods to estimate the model. Section 5 illustrates how the methods can be used to approach the problem of clustering trajectories. In Section 6 the problem of selecting a particular mixture of hierarchical models is considered. The model class and estimation methods are illustrated with a real data example in Section 7. Finally, we give a brief discussion in Section 8.

2 Review of Clustering via Mixture Models with Regression Structure

Finite mixture models with regression structure have been extensively studied in the statistical literature and commonly applied to problems in fields such as epidemiology, medicine, genetics, economics, engineering, marketing and in the physical and social sciences. One of the earliest works was that of Quandt (1972) who defined a two-component mixture likelihood for so-called switching regressions. The methodology demonstrated the ability to find underlying group behavior by maximizing the likelihood using a conjugate gradient algorithm. Later, Quandt and Ramsey (1978) developed a procedure using the method of moments to estimate the mixture parameters for switching regressions. Hosmer (1974) also defined a two component mixture likelihood containing regression components but used maximum likelihood to estimate the mixture parameters in an iterative process. Essentially, he developed an EM algorithm for mixtures of regressions coming from two clusters. DeSarbo and Cron (1988) developed the modern EM-based procedure for mixtures of linear regressions with any number of clusters. Jones and McLachlan (1992) extend this work to multivariate data. Hurn et al. (2003) discuss solutions to the label-switching problem for Bayesian inference with regression mixtures, and Viele and Tong (2002) present consistency results of the posterior distribution.

Recently, many researchers have incorporated random effects into a wide variety of regression models to account for correlated response and multiple sources of variation. Linear and nonlinear models with fixed and random effects are two model classes that have attracted an enormous amount of attention in recent years (see Davidian and Giltinan, 1995; Vonesh and Chinchilli, 1997; Pinheiro and Bates, 2000; Verbeke and Molenberghs, 2000; Fitzmaurice et al., 2004). In a mixture model context, Gaffney and Smyth (2003) developed a random effects regression mixture framework and derived a maximum a posteriori based EM algorithm to perform inference. James and Sugar (2003) developed a functional clustering model for sparsely sampled functional data. Celeux et al. (2005) proposed a mixture of linear mixed models. They used the EM algorithm for estimating the parameters. Pfeifer (2004) considered the problem of model-based clustering based on semi-parametric mixed effects models. More recently, Booth et al. (2005) proposed a Bayesian approach to clustering multivariate

data, based on a multi-level linear mixed model.

3 Mixture of Nonlinear Hierarchical Models

In this section we introduce the finite mixture of nonlinear hierarchical models and present the hormone trajectories data as a motivating application.

3.1 Model Specification

The goal of cluster analysis is to partition a collection of individuals into homogeneous subsets. Model-based clustering has recently received a great deal of attention and has provided promising results in various applications (Fraley and Raftery, 2002). In this approach, the data are viewed as coming from a mixture of distributions, each representing a different cluster. Let $\mathbf{y}_i = (y_{i1}, \dots, y_{in_i})$ denote a vector of repeated observations for individual i that is assumed to arise from one of K populations, with densities $\mathbf{f}_k(\mathbf{g}_k(\boldsymbol{\theta}_{ik}, \mathbf{x}_{ik}); \mathbf{W}_{ik})$ indexed by a mean $\mathbf{g}_k(\cdot)$ and covariance matrix \mathbf{W}_{ik} , for $i = 1, \dots, m$ and $k = 1, \dots, K$. The matrix \mathbf{W}_{ik} only depends on i for its dimension, that is, \mathbf{W}_{ik} has dimension $n_i \times n_i$. We assume $\mathbf{g}_k(\cdot)$ to be a nonlinear function of unknown individual-specific parameters, $\boldsymbol{\theta}_{ik}$, and known covariates, \mathbf{x}_{ik} . For each k , the parameter vector $\boldsymbol{\theta}_{ik}$, of dimension p , follows a population distribution $\mathcal{N}_p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$. We will assume that, conditional on $(\boldsymbol{\theta}_{i1}, \dots, \boldsymbol{\theta}_{iK}, \mathbf{W}_{i1}, \dots, \mathbf{W}_{iK}, \pi_1, \dots, \pi_K)$, \mathbf{y}_i follows a mixture model

$$\mathbf{y}_i \sim \sum_{k=1}^K \pi_k \mathbf{f}_k(\mathbf{g}_k(\boldsymbol{\theta}_{ik}, \mathbf{x}_{ik}); \mathbf{W}_{ik}), \quad (1)$$

where π_k ($\pi_k \geq 0$, $\sum_{k=1}^K \pi_k = 1$) is the probability that a individual belongs to the k th cluster. Thus model (1) represents a mixture of nonlinear hierarchical models. We assume in (1) the component densities $\mathbf{f}_k(\mathbf{g}_k(\boldsymbol{\theta}_{ik}, \mathbf{x}_{ik}); \mathbf{W}_{ik})$ to have multivariate normal distribution, i.e., $\mathbf{y}_i \sim \mathcal{N}_{n_i}(\mathbf{g}_k(\boldsymbol{\theta}_{ik}, \mathbf{x}_{ik}), \mathbf{W}_{ik})$, if individual i belongs to cluster k . Depending on the context, various assumptions can be made about the covariance matrix \mathbf{W}_{ik} . Typically \mathbf{W}_{ik} is required to be $\sigma_k^2 \mathbf{I}_{n_i}$, $i = 1, \dots, m$, reflecting the assumption that individuals have exchangeable errors. In longitudinal applications, exchangeable models are common. In other situations, banded or first-order autoregressive forms where \mathbf{W}_{ik} depends on a small number of free parameters, are more common (see De la Cruz-Mesía and Marshall, 2003, 2006, and references therein). For the remainder of this article we assume that $\mathbf{W}_{ik} = \sigma_k^2 \mathbf{I}_{n_i}$ in order to reduce the number of parameters to be estimated.

Each of the component densities in (1) is a individual-specific model, defined in terms of individual random-effects. The component models can be similar in form, varying only in mean specification, or have entirely different functional forms with parameters of different dimensions and meanings across submodels. A special case we use here corresponds to component models that are similar in form, and have exactly the same mean structure, but with different parameter values. Also, we assume that K has a fixed given value; discussion on how to choose K will be given in the next three sections.

Logistic regression has been proposed by (Hosmer and Lemeshow, 2000) to classify individuals when there are known sub-populations. However, their method was not designed to handle longitudinal data. In any case, the sub-populations are not predefined in our application. Regression tree methodology, such as CART (Breiman et al., 1984), is a standard non-parametric method that can handle unknown sub-populations. However, when applying CART to longitudinal data (Segal, 1992), there are difficulties to accommodate unequally spaced and highly unbalanced data. The same problems afflict the multivariate adaptive regression splines approach for longitudinal data (MASAL) of Zhang (1997). Therefore a parametric modeling approach as we propose may provide satisfactory answers while avoiding some of the complexities inherent to more sophisticated alternatives.

A model similar to (1) was considered by Pauler and Laird (2000) who formulate a class of two-component mixtures of hierarchical nonlinear models for longitudinal data. They also outline a sampling strategy for posterior simulation, which consists of a sequence of Gibbs, Metropolis-Hastings, and reversible jump steps, where the later is required for switching between component models of different dimensions.

3.2 Biomarker Example

Motivation for model (1) comes from a study in a private fertilization obstetrics clinic in Santiago, Chile, where the detection of pathological pregnancies was desired. Assisted reproduction treatment entails a risk of ectopic pregnancy and early pregnancy loss. Thus, early prediction of outcome is important in pregnancies following assisted reproduction treatment. In these pregnancies, the incidence of ectopic pregnancies varies from double to nearly 5-fold compared with that in spontaneous pregnancies. In particular, patients with tubal factor infertility are at an increased risk of ectopic pregnancies and should therefore receive special attention to avoid further impairment of fertility. The rate of multiple gestation is also high (20–25%) and early pregnancy loss is common, which causes anxiety in the couples involved.

Markers have been sought to distinguish between normal and abnormal pregnancies before verification of live intrauterine pregnancy by transvaginal sonography is possible. Serum Beta Human Chorionic Gonadotropin (β -HCG) has been found to be predictive of pregnancy outcome. It is well known in obstetrics that, among other hormones, the β -HCG shows dramatic changes in women during pregnancy. It has been established, also, that values of the β -HCG are different in women who have normal pregnancies with terminal deliveries than in women who have spontaneous abortions or other types of adverse pregnancy outcomes.

In a normal pregnancy, the level of this hormone approximately doubles every 1.5 days up to 5 weeks after the last menstrual period, and then every 3.5 days from the 7th week on (Frits and Guo, 1987). After the first trimester, levels should gradually decrease over time and in fact quickly decrease to zero after the pregnancy is ended. However, abnormally large levels of β -HCG may indicate choriocarcinoma (a quick growing form of cancer that occurs in a woman’s uterus after a pregnancy, miscarriage, or abortion), Down syndrome in the fetus, hydatidiform mole (a rare mass or growth that forms inside the uterus at the beginning of a pregnancy), or ovarian cancer. Lower than

normal β -HCG levels may indicate ectopic pregnancy, a miscarriage or spontaneous abortion. In any case, a failure to exhibit normal growth patterns in β -HCG levels should be usually interpreted as a complicated pregnancy.

Using clinical criteria, these patients were grouped into normal and abnormal pregnancies. As reported by Marshall and Barón (2000), the normal group represents women with a normal delivery, whilst the abnormal group represents women who had any complication resulting in a non-terminal delivery and loss of the fetus.

On 173 young women, representing different pregnancies over a period of 2 years, the marker β -HCG was measured during the first 80 days of gestational age and one of the main targets of the study was to evaluate these concentrations at early stages of pregnancy, with the purpose of identifying women with a high risk of loss. Figure 1 presents the time profile in log scale for these 173 women. A non-linear relationship of the log β -HCG by day of gestational age is common for most women. We assumed these women belonged to one of $K = 2$ subpopulations: normal and abnormal pregnancies. The simple compartmental model commonly assumed for such relationship, in the k th subpopulation, leads to an underlying nonlinear model of the form

$$\mathbf{y}_i | z_{ik} = 1 \sim \mathcal{N}_{n_i}(\mathbf{g}(\boldsymbol{\theta}_{ik}, \mathbf{t}_i); \sigma_k^2 \mathbf{I}_{n_i}), \quad (2)$$

where

$$\mathbf{g}(\boldsymbol{\theta}_{ik}, \mathbf{t}_i) = \frac{\theta_{i1k}}{1 + \exp\{-(\mathbf{t}_i - \theta_{i2k})/\theta_{i3k}\}}.$$

Here $\mathbf{y}_i = (y_{i1}, \dots, y_{in_i})'$ denotes the longitudinal measurements on i th patient taken at arbitrary times $\mathbf{t}_i = (t_{i1}, \dots, t_{in_i})'$, $i = 1, \dots, m = 173$, and $z_{ik} = 1$ denotes that the individual i belongs to subpopulation k . We assume $\boldsymbol{\theta}_{ik} = (\theta_{i1k}, \theta_{i2k}, \theta_{i3k})' \sim \mathcal{N}_3(\boldsymbol{\mu}_k, \tau_k^2 \mathbf{I}_3)$. It is easy to see that model (2) along with the population distributions above form a finite mixture of nonlinear hierarchical models (1), i.e., to analyze these data, we define a finite mixture of nonlinear hierarchical models for subpopulations of normal and abnormal pregnant women and use this to estimate individual and population probabilities of normal pregnancy.

Figure 1 also shows heterogeneous profiles. The idea is to find groups of patients with similar profiles, and ultimately link these groups to the pregnancy outcome prediction. Determining the number of groups is thus part of the inferential problem.

4 Parameter Estimation

The likelihood for model (1) is invariant under permutations of the K components. This is not a problem for a deterministic algorithm such as the EM, but it complicates the inference from sampling procedures such as MCMC, because the labels of components may be randomly switched during the iterative process. See the discussion below.

We show first how to proceed with the EM algorithm.

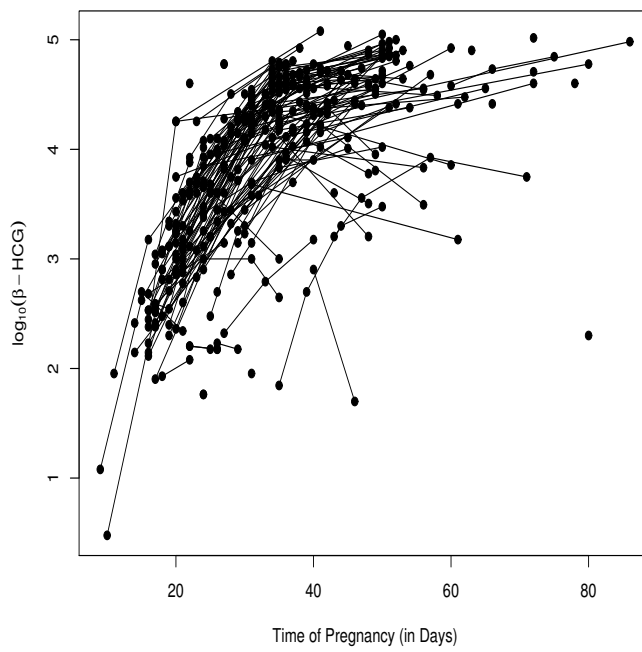


Figure 1: Observed profiles of β -HCG for all 173 women.

4.1 MLE via an EM-type algorithm

Following McLachlan and Basford (1988), inference in mixture models is facilitated by the introduction of latent variables $\mathbf{z}_i = (z_{i1}, \dots, z_{iK})$ with

$$z_{ik} = \begin{cases} 1 & \text{if } \mathbf{y}_i \text{ belongs to cluster } k \\ 0 & \text{otherwise.} \end{cases}$$

We assume that \mathbf{z}_i , $i = 1, \dots, m$, are iid realizations from a multinomial distribution with probabilities (π_1, \dots, π_K) satisfying $\sum_{k=1}^K \pi_k = 1$, and that the density of \mathbf{y}_i given \mathbf{z}_i is

$$\prod_{k=1}^K [f(\mathbf{g}(\boldsymbol{\theta}_{ik}, \mathbf{x}_{ik}); \sigma_k^2 \mathbf{I}_{n_i})]^{z_{ik}}.$$

We make use of an EM-type algorithm methodology that takes into account the incomplete structure of the data. In model (1) we denote the mixture proportions by $\boldsymbol{\pi} = (\pi_1, \dots, \pi_K)$, and the nonlinear model parameters for cluster k by $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_K)$, where $\boldsymbol{\beta}_k = (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \sigma_k^2)$. Here, missing data are of two types: the indicator vectors $\mathbf{z} = (z_i, i = 1, \dots, m)$ and the random effects $\boldsymbol{\theta}_{ik}$ for individual i in the k th cluster.

Then it is easy to derive the complete-data log-likelihood as

$$l(\boldsymbol{\beta}, \boldsymbol{\pi} | \mathbf{y}, \mathbf{z}, \boldsymbol{\theta}) = \sum_{i=1}^m \sum_{k=1}^K z_{ik} \log(\pi_k p(\mathbf{y}_i, \boldsymbol{\theta}_{ik} | \boldsymbol{\beta}_k))$$

where the vector \mathbf{y}_i , of size n_i , contains all the recorded values for individual i , and $\boldsymbol{\theta}_{ik}$ denotes the random-effect vector for individual i in cluster k . Thus we have

$$l(\boldsymbol{\beta}, \boldsymbol{\pi} | \mathbf{y}, \mathbf{z}, \boldsymbol{\theta}) = \sum_{i=1}^m \sum_{k=1}^K z_{ik} \log \pi_k + \sum_{i=1}^m \sum_{k=1}^K z_{ik} h(\boldsymbol{\beta}_k | \mathbf{y}_i, \boldsymbol{\theta}_{ik}),$$

where

$$\begin{aligned} h(\boldsymbol{\beta}_k | \mathbf{y}_i, \boldsymbol{\theta}_{ik}) &= C - \frac{n_i}{2} \log \sigma_k^2 - \frac{1}{2\sigma_k^2} \|\mathbf{y}_i - \mathbf{g}(\boldsymbol{\theta}_{ik}, \mathbf{x}_{ik})\|^2 - \frac{1}{2} \log |\boldsymbol{\Sigma}_k| \\ &\quad - \frac{1}{2} (\boldsymbol{\theta}_{ik} - \boldsymbol{\mu}_k)' \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{\theta}_{ik} - \boldsymbol{\mu}_k). \end{aligned}$$

for some constant C .

4.1.1 E step

At iteration $s > 0$, this step consists of computing the expectation of the complete log-likelihood knowing the observed data and the current value of the parameters $\boldsymbol{\beta}^{(s)}$, $\boldsymbol{\pi}^{(s)}$. In the nonlinear hierarchical model mixture context we get

$$\begin{aligned} \mathcal{Q}(\boldsymbol{\beta}, \boldsymbol{\pi} | \boldsymbol{\beta}^{(s)}, \boldsymbol{\pi}^{(s)}) &= \mathbb{E}(l(\boldsymbol{\beta}, \boldsymbol{\pi} | \mathbf{y}, \mathbf{z}, \boldsymbol{\theta}) | \mathbf{y}, \boldsymbol{\beta}^{(s)}, \boldsymbol{\pi}^{(s)}) \\ &= \sum_{i=1}^m \sum_{k=1}^K \hat{z}_{ik}^{(s)} \log \pi_k \\ &\quad + \sum_{i=1}^m \sum_{k=1}^K \hat{z}_{ik}^{(s)} \mathbb{E} \left[h(\boldsymbol{\beta}_k | \mathbf{y}_i, \boldsymbol{\theta}_{ik}) | \mathbf{y}, \boldsymbol{\beta}^{(s)} \right], \end{aligned} \quad (3)$$

where

$$\hat{z}_{ik}^{(s)} = \Pr(z_{ik} = 1 | \mathbf{y}_i, \boldsymbol{\beta}^{(s)}, \boldsymbol{\pi}^{(s)}) = \frac{\pi_k^{(s)} \mathbf{p}(\mathbf{y}_i | \boldsymbol{\beta}_k^{(s)})}{\sum_{l=1}^K \pi_l^{(s)} \mathbf{p}(\mathbf{y}_i | \boldsymbol{\beta}_l^{(s)})}$$

denotes the conditional probability that \mathbf{y}_i arises from the k th cluster, and $\mathbf{p}(\mathbf{y}_i | \boldsymbol{\beta}_k)$ is the marginal distribution obtained using Monte Carlo integration, i.e, for some large T , we compute

$$\mathbf{p}(\mathbf{y}_i | \boldsymbol{\beta}_k) \approx \frac{1}{T} \sum_{l=1}^T \mathbf{p}(\mathbf{y}_i | \boldsymbol{\theta}_k^{(l)}, \boldsymbol{\beta}_k),$$

with $\boldsymbol{\theta}_k^{(1)}, \dots, \boldsymbol{\theta}_k^{(T)} \stackrel{\text{iid}}{\sim} \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$.

4.1.2 M step

This stage consists of finding the values maximizing $\mathcal{Q}(\boldsymbol{\beta}, \boldsymbol{\pi} | \boldsymbol{\beta}^{(s)}, \boldsymbol{\pi}^{(s)})$. It can now be shown that the value $(\boldsymbol{\beta}^{(s+1)}, \boldsymbol{\pi}^{(s+1)})$ of $(\boldsymbol{\beta}, \boldsymbol{\pi})$ that maximizes $\mathcal{Q}(\boldsymbol{\beta}, \boldsymbol{\pi} | \boldsymbol{\beta}^{(s)}, \boldsymbol{\pi}^{(s)})$ is given by

$$(\boldsymbol{\pi}^{(s+1)}, \boldsymbol{\mu}^{(s+1)}, \boldsymbol{\Sigma}^{(s+1)}, \sigma^{2(s+1)})'$$

where, for $k = 1, \dots, K$

$$\begin{aligned} \pi_k^{(s+1)} &= \frac{1}{m} \sum_{i=1}^m \hat{z}_{ik}^{(s)}, \\ \boldsymbol{\mu}_k^{(s+1)} &= \frac{1}{\sum_{i=1}^m \hat{z}_{ik}^{(s)}} \sum_{i=1}^m \hat{z}_{ik}^{(s)} \mathbb{E}(\boldsymbol{\theta}_{ik} | \mathbf{y}, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \sigma_k^2), \\ \boldsymbol{\Sigma}_k^{(s+1)} &= \frac{1}{\sum_{i=1}^m \hat{z}_{ik}^{(s)}} \sum_{i=1}^m \hat{z}_{ik}^{(s)} \mathbb{E}\{(\boldsymbol{\theta}_{ik} - \boldsymbol{\mu}_k^{(s+1)})(\boldsymbol{\theta}_{ik} - \boldsymbol{\mu}_k^{(s+1)})' | \mathbf{y}, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \sigma_k^2\}, \end{aligned}$$

and

$$\sigma_k^2 = \frac{1}{\sum_{i=1}^m \hat{z}_{ik}^{(s)}} \sum_{i=1}^m \hat{z}_{ik}^{(s)} \mathbb{E}\{\|\mathbf{y}_i - \mathbf{g}(\boldsymbol{\theta}_{ik}, \mathbf{x}_{ik})\|^2 | \mathbf{y}, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \sigma_k^2\}.$$

We introduce now the following notation: let $\bar{\boldsymbol{\theta}}_{ik} = \mathbb{E}(\boldsymbol{\theta}_{ik} | \mathbf{y}, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \sigma_k^2)$, $\boldsymbol{\Theta}_{ik} = \text{Cov}(\boldsymbol{\theta}_{ik} | \mathbf{y}, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \sigma_k^2)$, $\bar{\mathbf{g}}_{ik} = \mathbb{E}(\mathbf{g}(\boldsymbol{\theta}_{ik}) | \mathbf{y}, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \sigma_k^2)$, and $\boldsymbol{\Psi}_{ik} = \text{Cov}(\mathbf{g}(\boldsymbol{\theta}_{ik}) | \mathbf{y}, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \sigma_k^2)$. Then

$$\pi_k^{(s+1)} = \frac{1}{m} \sum_{i=1}^m \hat{z}_{ik}^{(s)}, \tag{4}$$

$$\boldsymbol{\mu}_k^{(s+1)} = \frac{1}{\sum_{i=1}^m \hat{z}_{ik}^{(s)}} \sum_{i=1}^m \hat{z}_{ik}^{(s)} \bar{\boldsymbol{\theta}}_{ik}, \tag{5}$$

$$\boldsymbol{\Sigma}_k^{(s+1)} = \frac{1}{\sum_{i=1}^m \hat{z}_{ik}^{(s)}} \sum_{i=1}^m \hat{z}_{ik}^{(s)} \{(\bar{\boldsymbol{\theta}}_{ik} - \boldsymbol{\mu}_k^{(s+1)})(\bar{\boldsymbol{\theta}}_{ik} - \boldsymbol{\mu}_k^{(s+1)})' + \boldsymbol{\Theta}_{ik}\}, \tag{6}$$

and

$$\sigma_k^2 = \frac{1}{\sum_{i=1}^m \hat{z}_{ik}^{(s)}} \sum_{i=1}^m \hat{z}_{ik}^{(s)} \{ \|\mathbf{y}_i - \bar{\mathbf{g}}_{ik}\|^2 + \text{tr}(\mathbf{\Psi}_{ik}) \}. \quad (7)$$

In the special case where $\mathbf{\Sigma}_k = \tau_k^2 \mathbf{I}_p$, we get

$$\tau_k^{2(s+1)} = \frac{1}{p \sum_{i=1}^m \hat{z}_{ik}^{(s)}} \sum_{i=1}^m \hat{z}_{ik}^{(s)} \{ \|\bar{\boldsymbol{\theta}}_{ik} - \boldsymbol{\mu}_k^{(s+1)}\|^2 + \text{tr}(\mathbf{\Theta}_{ik}) \}. \quad (8)$$

Thus, in order to implement the EM algorithm the quantities $\bar{\boldsymbol{\theta}}_{ik}$, $\mathbf{\Theta}_{ik}$, $\bar{\mathbf{g}}_{ik}$, and $\mathbf{\Psi}_{ik}$ need to be evaluated at each iteration of the algorithm. One practical problem arising from our use of the EM algorithm is that there is no closed form expressions for any of $\bar{\boldsymbol{\theta}}_{ik}$, $\mathbf{\Theta}_{ik}$, $\bar{\mathbf{g}}_{ik}$, or $\mathbf{\Psi}_{ik}$, and hence no closed form expressions for any of (5), (6) (or 8), or (7). We use Monte Carlo integration to calculate these quantities. Details are given below.

From our earlier definition $\bar{\boldsymbol{\theta}}_{ik} = \int \boldsymbol{\theta}_{ik} \mathbf{p}(\boldsymbol{\theta}_{ik} | \mathbf{y}_i, \boldsymbol{\mu}_k, \mathbf{\Sigma}_k, \sigma_k^2) d\boldsymbol{\theta}_{ik}$. However, due the nonlinearity of random effects in the response scale, $\mathbf{p}(\boldsymbol{\theta}_{ik} | \mathbf{y}_i, \boldsymbol{\mu}_k, \mathbf{\Sigma}_k, \sigma_k^2)$ is not available in closed form. Nevertheless, sampling from $\mathbf{p}(\boldsymbol{\theta}_{ik} | \boldsymbol{\mu}_k, \mathbf{\Sigma}_k)$ is straightforward and therefore we switch to the alternative expression

$$\bar{\boldsymbol{\theta}}_{ik} = \frac{\int \boldsymbol{\theta}_k \mathbf{p}(\mathbf{y}_i | \boldsymbol{\theta}_k, \sigma_k^2) \mathbf{p}(\boldsymbol{\theta}_k | \boldsymbol{\mu}_k, \mathbf{\Sigma}_k) d\boldsymbol{\theta}_k}{\int \mathbf{p}(\mathbf{y}_i | \boldsymbol{\theta}_k, \sigma_k^2) \mathbf{p}(\boldsymbol{\theta}_k | \boldsymbol{\mu}_k, \mathbf{\Sigma}_k) d\boldsymbol{\theta}_k}.$$

To implement the Monte Carlo integration, take, for some large T ,

$$\boldsymbol{\theta}_k^{(1)}, \dots, \boldsymbol{\theta}_k^{(T)} \stackrel{\text{iid}}{\sim} \mathcal{N}(\boldsymbol{\mu}_k, \mathbf{\Sigma}_k)$$

and put

$$\bar{\boldsymbol{\theta}}_{ik} = \frac{\sum_{l=1}^T \boldsymbol{\theta}_k^{(l)} \mathbf{p}(\mathbf{y}_i | \boldsymbol{\theta}_k^{(l)}, \sigma_k^2)}{\sum_{l=1}^T \mathbf{p}(\mathbf{y}_i | \boldsymbol{\theta}_k^{(l)}, \sigma_k^2)}.$$

To obtain $\mathbf{\Theta}_{ik}$, first put $\bar{\mathbf{\Theta}}_{ik} = E(\boldsymbol{\theta}_{ik} \boldsymbol{\theta}_{ik}' | \mathbf{y}_i, \boldsymbol{\mu}_k, \mathbf{\Sigma}_k, \sigma_k^2)$, which is given by

$$\bar{\mathbf{\Theta}}_{ik} = \frac{\sum_{l=1}^T \boldsymbol{\theta}_k^{(l)} \boldsymbol{\theta}_k'^{(l)} \mathbf{p}(\mathbf{y}_i | \boldsymbol{\theta}_k^{(l)}, \sigma_k^2)}{\sum_{l=1}^T \mathbf{p}(\mathbf{y}_i | \boldsymbol{\theta}_k^{(l)}, \sigma_k^2)},$$

so that $\mathbf{\Theta}_{ik} = \bar{\mathbf{\Theta}}_{ik} - \bar{\boldsymbol{\theta}}_{ik} \bar{\boldsymbol{\theta}}_{ik}'$. Values for $\bar{\mathbf{g}}_{ik}$ and $\mathbf{\Psi}_{ik}$ are obtained in an identical manner, that is,

$$\bar{\mathbf{g}}_{ik} = \frac{\sum_{l=1}^T \mathbf{g}(\boldsymbol{\theta}_k^{(l)}, \mathbf{x}_{ik}) \mathbf{p}(\mathbf{y}_i | \boldsymbol{\theta}_k^{(l)}, \sigma_k^2)}{\sum_{l=1}^T \mathbf{p}(\mathbf{y}_i | \boldsymbol{\theta}_k^{(l)}, \sigma_k^2)},$$

and $\Psi_{ik} = \bar{\Psi}_{ik} - \bar{g}_{ik}\bar{g}'_{ik}$, where

$$\bar{\Psi}_{ik} = \frac{\sum_{l=1}^T \mathbf{g}(\boldsymbol{\theta}_k^{(l)}, \mathbf{x}_{ik}) \mathbf{g}(\boldsymbol{\theta}_k^{(l)}, \mathbf{x}_{ik})' \mathbf{p}(\mathbf{y}_i | \boldsymbol{\theta}_k^{(l)}, \sigma_k^2)}{\sum_{l=1}^T \mathbf{p}(\mathbf{y}_i | \boldsymbol{\theta}_k^{(l)}, \sigma_k^2)}.$$

Convergence to the true values of $\bar{\boldsymbol{\theta}}_{ik}$, etc., is almost sure, so it only needs to be established which value of T leads to the required accuracy for these values.

Starting with $(\boldsymbol{\beta}^{(0)}, \boldsymbol{\pi}^{(0)})$, at sth iterative step the algorithm moves from a state $(\boldsymbol{\beta}^{(s)}, \boldsymbol{\pi}^{(s)})$ to $(\boldsymbol{\beta}^{(s+1)}, \boldsymbol{\pi}^{(s+1)})$, which is described by steps (4), (5), (6) (or (8)), and (7). Sufficient conditions for convergence are given in Dempster et al. (1977) and Wu (1983). As with all iterative searches for an MLE, a number of starting points should be considered to ensure that a true global maximum has been found.

4.1.3 Standard Errors

A motivation for using the EM algorithm is often that the likelihood based on the observed data is difficult or impossible to evaluate. Using an EM algorithm, maximum likelihood estimates of parameters are readily obtained, but the algorithm does not immediately yield asymptotic standard errors of these estimates.

From Guo and Thompson (1994) the variance-covariance matrix \mathbf{V} of the estimates can be written as

$$\mathbf{V}^{-1} = \mathbf{I}_c - \text{Var} \left(\frac{\partial \log \mathbf{p}(\mathbf{y}, \mathbf{z}, \boldsymbol{\theta} | \boldsymbol{\beta}, \boldsymbol{\pi})}{\partial (\boldsymbol{\beta}, \boldsymbol{\pi})} \bigg| \mathbf{y} \right) \bigg|_{(\boldsymbol{\beta}, \boldsymbol{\pi}) = (\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\pi}})},$$

where \mathbf{I}_c is the complete data expected information matrix given by

$$\mathbf{I}_c = \mathbb{E}(\mathbf{I}_0(\boldsymbol{\beta}, \boldsymbol{\pi} | \mathbf{y}, \mathbf{z}, \boldsymbol{\theta}) | \mathbf{y}, \boldsymbol{\beta}, \boldsymbol{\pi}) |_{(\boldsymbol{\beta}, \boldsymbol{\pi}) = (\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\pi}})},$$

and

$$\mathbf{I}_0(\boldsymbol{\beta}, \boldsymbol{\pi} | \mathbf{y}, \mathbf{z}, \boldsymbol{\theta}) = \frac{-\partial^2 \log p(\mathbf{y}, \mathbf{z}, \boldsymbol{\theta} | \boldsymbol{\beta}, \boldsymbol{\pi})}{\partial (\boldsymbol{\beta}, \boldsymbol{\pi})^2},$$

where the matrix $\mathbf{I}_0(\boldsymbol{\beta}, \boldsymbol{\pi} | \mathbf{y}, \mathbf{z}, \boldsymbol{\theta})$ is given by

$$\mathbf{I}_0(\boldsymbol{\beta}, \boldsymbol{\pi} | \mathbf{y}, \mathbf{z}, \boldsymbol{\theta}) = \begin{pmatrix} \mathbf{A} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{B} & \mathbf{C} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}' & \mathbf{D} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{E} \end{pmatrix}.$$

Matrices \mathbf{A} , \mathbf{B} , \mathbf{C} , \mathbf{D} , and \mathbf{E} are respectively given by

$$-\frac{\partial^2 \log p(\mathbf{y}, \mathbf{z}, \boldsymbol{\theta} | \boldsymbol{\beta}, \boldsymbol{\pi})}{\partial \pi_k^2}, \quad -\frac{\partial^2 \log p(\mathbf{y}, \mathbf{z}, \boldsymbol{\theta} | \boldsymbol{\beta}, \boldsymbol{\pi})}{\partial \mu_k^2}, \quad -\frac{\partial^2 \log p(\mathbf{y}, \mathbf{z}, \boldsymbol{\theta} | \boldsymbol{\beta}, \boldsymbol{\pi})}{\partial \boldsymbol{\mu}_k \boldsymbol{\Sigma}_k},$$

$$-\frac{\partial^2 \log p(\mathbf{y}, \mathbf{z}, \boldsymbol{\theta} | \boldsymbol{\beta}, \pi)}{\partial \boldsymbol{\Sigma}_k^2}, \quad \text{and} \quad -\frac{\partial^2 \log p(\mathbf{y}, \mathbf{z}, \boldsymbol{\theta} | \boldsymbol{\beta}, \pi)}{\partial (\sigma_k^2)^2}.$$

Using formulas given in Jennrich and Schluchter (1986) we get $\mathbf{A} = \text{diag}(a_1, \dots, a_{K-1})$, with $a_k = \sum_{i=1}^m \left\{ \frac{z_{ik}}{\pi_k^2} + \frac{z_{iK}}{\pi_K^2} \right\}$, for $k = 1, \dots, K-1$. Furthermore, $\mathbf{B} = \boldsymbol{\Sigma}_k^{-1} \sum_{i=1}^m z_{ik}$, and

$$C_{qr} = \mathbf{H}'_q \boldsymbol{\Sigma}_k^{-1} \frac{\partial \boldsymbol{\Sigma}_k}{\partial \Sigma_{kr}} \boldsymbol{\Sigma}_k^{-1} \sum_{i=1}^m z_{ik} (\boldsymbol{\theta}_{ik} - \boldsymbol{\mu}_k)$$

for $1 \leq q \leq p$, $1 \leq r \leq p^2$, and \mathbf{H}_q is the q th column of the $p \times p$ identity matrix \mathbf{I}_p . Also

$$D_{qr} = \frac{1}{2} \text{tr} \left(\boldsymbol{\Sigma}_k^{-1} \frac{\partial \boldsymbol{\Sigma}_k}{\partial \Sigma_{kq}} \boldsymbol{\Sigma}_k^{-1} \left\{ \sum_{i=1}^m [2z_{ik} (\boldsymbol{\theta}_{ik} - \boldsymbol{\mu}_k) (\boldsymbol{\theta}_{ik} - \boldsymbol{\mu}_k)' - \boldsymbol{\Sigma}_k] \right\} \boldsymbol{\Sigma}_k^{-1} \frac{\partial \boldsymbol{\Sigma}_k}{\partial \Sigma_{kr}} \right)$$

for $1 \leq q, r \leq p^2$, and

$$E = -\frac{1}{2\sigma_k^4} \sum_{i=1}^m n_i z_{ik} + \frac{1}{\sigma_k^6} \sum_{i=1}^m z_{ik} \|\mathbf{y}_i - \mathbf{g}(\boldsymbol{\theta}_{ik}, \mathbf{x}_{ik})\|^2.$$

In the special case $\boldsymbol{\Sigma}_k = \tau_k^2 \mathbf{I}_p$, we get

$$\begin{aligned} \mathbf{B} = B &= \tau_k^{-2} \sum_{i=1}^m z_{ik} \\ \mathbf{C} = C &= \frac{1}{\tau_k^4} \sum_{i=1}^m z_{ik} (\boldsymbol{\theta}_{ik} - \boldsymbol{\mu}_k) \\ \mathbf{D} = D &= -\frac{p}{2\tau_k^4} \sum_{i=1}^m z_{ik} + \frac{1}{\tau_k^6} \sum_{i=1}^m z_{ik} \|\boldsymbol{\theta}_{ik} - \boldsymbol{\mu}_k\|^2, \end{aligned}$$

where, in this case, B , C and D become scalar quantities.

4.2 Bayesian Estimation via MCMC

We have seen that estimation for nonlinear hierarchical mixture models is straightforward using the EM algorithm. Estimation in a Bayesian framework can be done using posterior simulation via Markov chain Monte Carlo (MCMC) methods. Bayes estimators for mixture models are well defined so long as the prior distributions are proper. Provided that suitable (conjugate) priors are used, the posterior density will be proper, thereby allowing the application of MCMC methods such as the Gibbs sampler to provide an accurate approximation to the Bayes solution (see Roeder and Wasserman, 1997; Stephens, 2000).

Here, it is also useful to employ the latent allocation variables \mathbf{z}_i introduced in Section 4.1. Recall the corresponding complete likelihood is

$$\prod_{i=1}^m \prod_{k=1}^K \{\pi_k p(\mathbf{y}_i, \boldsymbol{\theta}_{ik} | \boldsymbol{\beta}_k)\}^{z_{ik}}.$$

By simplicity, the prior on $(\boldsymbol{\beta}, \boldsymbol{\pi})$ is assumed to be a product of conjugate densities

$$P(\boldsymbol{\beta}, \boldsymbol{\pi}) = P(\boldsymbol{\pi}) \prod_{k=1}^K P(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \sigma_k^2). \quad (9)$$

Now, we are concerned with Bayesian inference about the model parameters $\boldsymbol{\beta}$, $\boldsymbol{\pi}$ and the classification indicators \mathbf{z} . We use the Gibbs sampler for estimating parameters, as explained next.

4.2.1 Prior Specification

We now consider the problem of choosing prior distributions for parameters $\boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}_k$ and σ_k^2 of model (1). We assume prior independence for parameters in (9), i.e., $P(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \sigma_k^2) = P(\boldsymbol{\mu}_k)P(\boldsymbol{\Sigma}_k)P(\sigma_k^2)$. If the expected fraction of individuals belonging to a specific cluster is small, posterior distributions of the population and individual-specific parameters in the corresponding submodel of (1) will be poorly estimated. Therefore, proper subjective priors are necessary for all population parameters in the component densities. If subjective priors are not available, a sensitivity analysis should be performed across a range of sensible priors.

The conjugate prior on $\boldsymbol{\pi}$ will always be taken as symmetric Dirichlet distribution.

$$\boldsymbol{\pi} = (\pi_1, \dots, \pi_K) \sim \mathcal{D}(\delta, \dots, \delta),$$

and the conjugate prior distribution of $\boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}_k^{-1}$, and σ_k^2 are normal, Wishart, and inverse gamma distributions, respectively:

$$\boldsymbol{\mu}_k \sim \mathcal{N}_p(\boldsymbol{\mu}_{k0}, \boldsymbol{\Sigma}_{k0}), \quad \boldsymbol{\Sigma}_k^{-1} \sim \mathcal{W}(r_{k0}, [r_{k0} \mathbf{R}_{k0}]^{-1}), \quad \sigma_k^2 \sim \mathcal{IG}(a_{k0}, b_{k0}).$$

The Wishart prior is parameterized such that its mean is \mathbf{R}_{0k}^{-1} . In the special case $\boldsymbol{\Sigma}_k = \tau_k^2 \mathbf{I}_p$ the prior for τ_k^2 is inverse gamma, i.e.,

$$\tau_k^2 \sim \mathcal{IG}(c_{k0}, d_{k0}).$$

The inverse gamma prior is parameterized as $\pi(x) \propto x^{-(a+1)} \exp(-1/cx)$. In practice the specification of hyperparameters $\boldsymbol{\mu}_{k0}$, $\boldsymbol{\Sigma}_{k0}$, r_{k0} , \mathbf{R}_{k0} (or c_{k0} , d_{k0}), a_{k0} , and b_{k0} may be difficult, so we can take the values of hyperparameters in such a way that we get non-informative priors in the limiting case when no (or minimal) prior information is available. For example, the prior choice for r_{k0} is $r_{k0} = p$, which is most non-informative in the sense that its distribution is flattest. Similarly, \mathbf{R}_{k0} is chosen to be an approximate prior estimate of $\boldsymbol{\Sigma}_k$.

4.2.2 Full Conditionals

The full conditionals for implementing Gibbs sampling are given by

$$\begin{aligned} \boldsymbol{\pi} | \dots &\sim \mathcal{D}(\delta + m_1, \dots, \delta + m_K), \\ \Pr(z_{ik} = 1 | \dots) &= \frac{\pi_k^{(s)} p(\mathbf{y}_i | \boldsymbol{\theta}_{ik}^{(s)}, \sigma_k^{2(s)})}{\sum_{l=1}^K \pi_l^{(s)} p(\mathbf{y}_i | \boldsymbol{\theta}_{il}^{(s)}, \sigma_l^{2(s)})}, \quad k = 1, \dots, K, \quad i = 1, \dots, m, \end{aligned}$$

$$\begin{aligned}
\boldsymbol{\mu}_k | \dots &\sim \mathcal{N}(\mathbf{V}_k(m\boldsymbol{\Sigma}_k^{-1}\bar{\boldsymbol{\theta}}_k + \boldsymbol{\Sigma}_{k0}^{-1}\boldsymbol{\mu}_{k0}), \mathbf{V}_k), \\
\boldsymbol{\Sigma}_k^{-1} | \dots &\sim \mathcal{W}(m + r_{k0}, \tilde{\mathbf{R}}_k), \\
\sigma_k^2 | \dots &\sim \mathcal{IG}\left(\frac{2a_{k0} + \sum_{i=1}^m n_i z_{ik}}{2}, \frac{2b_{k0}}{2 + \sum_{i=1}^m z_{ik} \|\mathbf{y}_i - \mathbf{g}(\boldsymbol{\theta}_{ik}, \mathbf{x}_{ik})\|^2}\right),
\end{aligned}$$

where $\bar{\boldsymbol{\theta}}_k = m^{-1} \sum_{i=1}^m z_{ik} \boldsymbol{\theta}_{ik}$, $\mathbf{V}_k^{-1} = (m_k \boldsymbol{\Sigma}_k^{-1} + \mathbf{R}_{k0}^{-1})$, $\tilde{\mathbf{R}}_k = (r_{k0} \mathbf{R}_{k0} + \sum_{i=1}^m z_{ik} (\boldsymbol{\theta}_{ik} - \boldsymbol{\mu}_k)(\boldsymbol{\theta}_{ik} - \boldsymbol{\mu}_k)')^{-1}$, $m_k = \sum_{i=1}^m z_{ik}$, and where ‘ $|\dots$ ’ denotes conditioning on all other variables. In the special case $\boldsymbol{\Sigma}_k = \tau_k^2 \mathbf{I}_p$, the full conditional for τ_k^2 is

$$\tau_k^2 | \dots \sim \mathcal{IG}\left(\frac{2c_{k0} + pm_k}{2}, \frac{2d_{k0}}{2 + \sum_{i=1}^m z_{ik} (\boldsymbol{\theta}_{ik} - \boldsymbol{\mu}_k)' \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{\theta}_{ik} - \boldsymbol{\mu}_k)}\right).$$

Generating samples from the full conditional distributions for $(\mathbf{z}, \boldsymbol{\pi}, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, (\text{or } \tau_k^2), \sigma_k^2)$ is straightforward since they all have convenient forms.

The full conditional for $\boldsymbol{\theta}_{ik}$ is not available analytically. This suggests to carry out a Metropolis-Hastings step. Denote the proposal distribution by $q(\boldsymbol{\theta}_{ik} | \mathbf{a}_{ik}, \mathbf{v}_{ik})$, which we take as a p -dimensional Normal law with mean \mathbf{a}_{ik} and variance-covariance matrix \mathbf{v}_{ik} . In order to increase the efficiency of the algorithm the matrix \mathbf{v}_{ik} is determined as follows. Firstly, a maximum likelihood estimate $\hat{\mathbf{v}}_{ik}$ of \mathbf{v}_{ik} is obtained. Then a preliminary (random-walk) Metropolis-Hastings run is performed with the posterior distribution of $\boldsymbol{\theta}_{ik}$ as the target distribution using, at this stage, $q(\boldsymbol{\theta}_{ik} | \boldsymbol{\theta}_{ik}^{(s)}, c_i \hat{\mathbf{v}}_{ik})$ as the proposal distribution at the $(s+1)$ th iteration, where c_i is a suitable tuning parameter whose value is assumed known. After running such a chain, one obtains a sample $\{\hat{\boldsymbol{\theta}}_{ik}^{(s)} : s \geq s_0\}$ from the posterior distribution of $\boldsymbol{\theta}_{ik}$ and we set $\mathbf{v}_{ik} = c'_i \tilde{\mathbf{v}}_{ik}$, where $\tilde{\mathbf{v}}_{ik}$ denotes the corresponding sample variance-covariance matrix and c'_i another suitable tuning parameter chosen to get sure that the acceptance rate is satisfactory (typically between 0.2 and 0.5), see Gelman et al. (1996).

Details for implementation of MCMC algorithm are provided in the Appendix.

4.2.3 Label switching

In finite mixture models, an identifiability problem arises from the invariance of the likelihood under permutation of the component labels unless strong prior information is used (Stephens, 2000). Under the Bayesian standpoint, this leads to symmetric and multimodal posterior distributions with up to $K!$ copies of each ‘‘genuine’’ mode, complicating inference on the parameters. This may cause label switching during the MCMC iterations, hence typical averages of MCMC samples of the parameters may yield unreasonable estimates of the mixture parameters. Traditional approaches to this problem impose identifiability constraints on the parameters, for instance $\pi_1 < \dots < \pi_K$. These constraints, however, do not always solve the problem. Therefore, we adopt the relabeling procedure suggested by Celeux (1998) at each iteration of MCMC. General background on solutions that have been previously suggested for this problem can be found in Jasra et al. (2005) who categorize them to artificial identifiability constraints (Diebolt and Robert, 1994; Richardson and Green, 1997), random permutation sampling (Frühwirth-Schnatter, 2001), relabeling algorithms (Stephens, 2000; Celeux, 1998), and label invariant loss functions methods (Celeux et al., 2000; Hurn et al., 2003).

5 Allocation

From a classical viewpoint each individual i can be allocated to cluster k on the basis of the estimated posterior probabilities. Let $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\pi}})$ denote the maximum likelihood estimates of $(\boldsymbol{\beta}, \boldsymbol{\pi})$. The estimated posterior probability that individual \mathbf{y}_i belongs to cluster k is given by

$$\mathbf{p}(z_{ik} = 1 | \mathbf{y}_i, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\pi}}) = \frac{\hat{\pi}_k \mathbf{p}(\mathbf{y}_i | \hat{\boldsymbol{\beta}}_k)}{\sum_{\ell=1}^K \hat{\pi}_\ell \mathbf{p}(\mathbf{y}_i | \hat{\boldsymbol{\beta}}_\ell)}, \quad (10)$$

where $\mathbf{p}(\mathbf{y}_i | \hat{\boldsymbol{\beta}}_k)$ is obtained via Monte Carlo integration, i.e., for some large T , we compute

$$\mathbf{p}(\mathbf{y}_i | \hat{\boldsymbol{\beta}}_k) \approx \frac{1}{T} \sum_{l=1}^T \mathbf{p}(\mathbf{y}_i | \boldsymbol{\theta}_{ik}^{(l)}, \hat{\sigma}^2),$$

with $\boldsymbol{\theta}_{ik}^{(l)} \sim \mathcal{N}_p(\hat{\boldsymbol{\mu}}_k, \hat{\boldsymbol{\Sigma}}_k)$, for $k = 1, \dots, K$. The i th individual is then allocated according to the values of \hat{z}_i :

$$\hat{z}_i = \arg \max_{1 \leq k \leq K} \mathbf{p}(z_{ik} = 1 | \mathbf{y}_i, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\pi}}),$$

i.e., to the cluster maximizing the allocation probabilities $\mathbf{p}(z_{ik} = 1 | \mathbf{y}_i, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\pi}})$.

Once the label switching is taken care of, the MCMC samples can be used to draw posterior inference. Of particular interest are the allocation vector, \mathbf{z} . We compute the marginal posterior probability that individual i is allocated to cluster k as:

$$\begin{aligned} \mathbf{p}(z_{ik} = 1 | \mathbf{y}_i) &= \int \mathbf{p}(z_{ik} = 1 | \mathbf{y}_i, \boldsymbol{\Omega}) \mathbf{p}(\boldsymbol{\Omega} | \mathbf{y}^m) d\boldsymbol{\Omega} \\ &= \int \frac{\pi_k \mathbf{p}(\mathbf{y}_i | \boldsymbol{\Omega}_k)}{\sum_{\ell=1}^K \pi_\ell \mathbf{p}(\mathbf{y}_i | \boldsymbol{\Omega}_\ell)} \mathbf{p}(\boldsymbol{\Omega} | \mathbf{y}^m) d\boldsymbol{\Omega} \\ &\approx \sum_{s=1}^S \frac{\pi_k^{(s)} \mathbf{p}(\mathbf{y}_i | \boldsymbol{\Omega}_k^{(s)})}{\sum_{\ell=1}^K \pi_\ell^{(s)} \mathbf{p}(\mathbf{y}_i | \boldsymbol{\Omega}_\ell^{(s)})}. \end{aligned} \quad (11)$$

where \mathbf{y}^m denote all data and $\boldsymbol{\Omega}$ denotes all the random variables. The posterior allocation of individual i can then be estimated by the mode of its marginal posterior density:

$$\hat{z}_i = \arg \max_{1 \leq k \leq K} \mathbf{p}(z_{ik} = 1 | \mathbf{y}_i).$$

Class prediction

Let us now consider prediction of the class membership for a future individual \mathbf{y}_f . Using maximum likelihood, this is done by computing

$$\mathbf{p}(z_{fk} = 1 | \mathbf{y}_f, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\pi}}) = \frac{\hat{\pi}_k \mathbf{p}(\mathbf{y}_f | \hat{\boldsymbol{\beta}}_k)}{\sum_{\ell=1}^K \hat{\pi}_\ell \mathbf{p}(\mathbf{y}_f | \hat{\boldsymbol{\beta}}_\ell)},$$

with $\mathbf{p}(\mathbf{y}_f|\hat{\boldsymbol{\beta}}_k)$ obtained via Monte Carlo integration. Then the individual is allocated according to \hat{z}_{fk} :

$$\hat{z}_f = \arg \max_{1 \leq k \leq K} \mathbf{p}(z_{fk} = 1 | \mathbf{y}_f, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\pi}}).$$

The MCMC output can also be used to predict the class membership of future individual \mathbf{y}_f . The classification probability that a future individual \mathbf{y}_f belongs to the k th class is

$$\begin{aligned} \mathbf{p}(z_{fk} = 1 | \mathbf{y}_f, \mathbf{y}^m) &= \int \mathbf{p}(z_{fk} = 1 | \mathbf{y}_f, \boldsymbol{\Omega}) \mathbf{p}(\boldsymbol{\Omega} | \mathbf{y}^m) d\boldsymbol{\Omega} \\ &\propto \int \frac{\pi_k \mathbf{p}(\mathbf{y}_f | \boldsymbol{\Omega}_k)}{\sum_{\ell=1}^K \pi_\ell \mathbf{p}(\mathbf{y}_f | \boldsymbol{\Omega}_\ell)} \mathbf{p}(\boldsymbol{\Omega} | \mathbf{y}^m) d\boldsymbol{\Omega} \\ &\approx \sum_{s=1}^S \pi_k^{(s)} \mathbf{p}(\mathbf{y}_f | \boldsymbol{\Omega}_k^{(s)}). \end{aligned}$$

Then the individual is allocated according to \hat{z}_f :

$$\hat{z}_f = \arg \max_{1 \leq k \leq K} \mathbf{p}(z_{fk} = 1 | \mathbf{y}_f, \mathbf{y}^m).$$

6 Selecting the Number of Clusters

In the previous sections, it was implicitly assumed that the number of clusters, K , was fixed. However, one of the questions of scientific interest is assessing the reliability of the output from their clustering analysis. This is equivalent to the question of determining the number of true clusters that exist in the data and for determining the number of the components in a mixture model (Roeder and Wasserman, 1997).

Several measures have been proposed for choosing the clustering model (parametrization and number of clusters); see, e.g., Chapter 6 of McLachlan and Peel (2000). We use the Bayesian Information Criterion (BIC) approximation to the Bayes factor (Schwarz, 1978), which adds a penalty to the log-likelihood based on the number of parameters, and has performed well in a number of applications (see Dasgupta and Raftery, 1998; Fraley and Raftery, 1998, 2002). The BIC has the form

$$\text{BIC} = 2\widehat{\text{loglik}}_{\mathcal{M}} - p_{\mathcal{M}} \log(\# \text{ of observations}), \quad (12)$$

where $\widehat{\text{loglik}}_{\mathcal{M}}$ is the maximized log-likelihood for the model and data and $p_{\mathcal{M}}$ is the number of independent parameters to be estimated in the model \mathcal{M} . The BIC procedure is to choose the model for which the BIC criterion is maximized.

From a Bayesian viewpoint we use the Bayes factor (Kass and Raftery, 1995) as a selection tool. For the computation, we use the Chib's estimator (Chib, 1995) which is based on the Gibbs output and is specially suitable for mixture models. Detailed discussion on how to compute the Bayes factor is given in Chib (1995).

7 Analysis of the Pregnant Women Data

The approaches will be illustrated by considering the clustering of cases on the basis of longitudinal measurements on pregnant women after assisted reproduction. The analysis of the β -HCG concentrations for the 173 women is carried out using model (2). Our goal is to identify clusters of trajectories and to describe any clusters that are predictive of normal or abnormal pregnancy probability.

The outcome that we analyze are the vectors of time-varying β -HCG measurements for the 173 women. Approximately 30 per cent of the 173 women had one β -HCG measurement, 31 per cent had two, 34 per cent had three, and 5 per cent had four or more measurements. The 173 women altogether contribute a total of 375 observations, where the number of samples per woman ranged from 1 through 6, with median of 2. Figure 1 presents the individual-specific \log_{10} β -HCG profiles. In order to obtain initial parameter estimates we fit a single-cluster version of the model using the NLME library of Pinheiro and Bates (2000). We next applied a model-based clustering to the random effects estimated using the MCLUST package (Fraley and Raftery, 1999) with $K = 2, 3, 4, 5$. To facilitate fair comparison of our classical and Bayesian analyses, we selected very vague prior distributions in our analysis. That is, we use proper priors, but with hyperparameter values chosen so that the priors will have minimal impact relative to the data.

For the EM algorithm, the number T of samples for the Monte Carlo integration, was taken to be 10 000. When implementing the Gibbs sampling, we chose starting points in a neighborhood of the MLEs of model parameters. We also used other starting points obtaining similar results. We used 800 000 iterations with 100 000 sweeps as burn-in. Samples were collected at a spacing of 700 iterations, to obtain approximately independent samples. Finally we totaled 1 000 posterior Monte Carlo samples. To avoid the label switching problem, we adopt the relabeling procedure suggested by Celeux (1998) at each iteration of MCMC. To diagnose convergence, we suggest any of the convergence criteria discussed in the literature, for example, those included in the BOA package (Smith, 2004). Because of the high dimensional parameter vector, we prefer to use diagnostics, such as proposed by Geweke (1995), which do not require multiple parallel chains.

The mixture method of clustering requires the specification of the number of underlying number of clusters, K , to be fitted to the model. This decision was based on having the clinical classification of the data into two groups, but using the BIC criterion we found that the number the cluster selected was $K = 2$ (see Figure 2). Also, we found that the Bayes factor criterion favored $K = 2$ (data not shown).

Table 1 shows parameter estimates using both methods. The results are similar. The classification results for both methods are given in Table 2. We can see that the clinical classification and the “statistical diagnosis” are different for 40 and 42 women using Bayesian and classical methods, respectively. Examination of the posterior probabilities showed that 15 of these individuals are decisively assigned to a different group than the one corresponding to the clinical classification. Although the use of Bayesian methods has only slightly improved the outright clustering, it does produce a less extreme probabilistic clustering for the misallocated women. Another comparison

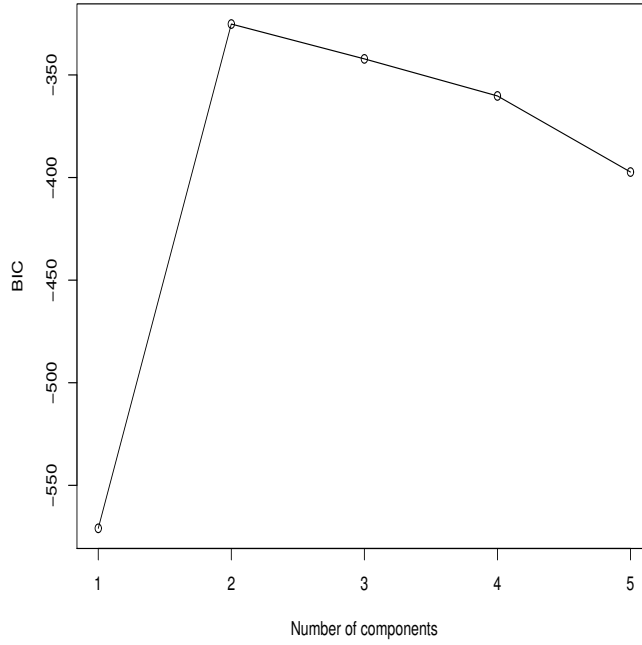


Figure 2: Values of the BIC criterion for the data.

Table 1: Summary of model fitting.

| Parameter | MCMC | | EM | |
|--------------|--------|-------|--------|-------|
| | mean | sd | mean | sd |
| μ_{11} | 4.749 | 0.037 | 4.762 | 0.038 |
| μ_{12} | 15.550 | 0.269 | 15.511 | 0.274 |
| μ_{13} | 6.743 | 0.356 | 6.958 | 0.349 |
| μ_{21} | 4.214 | 0.151 | 4.229 | 0.105 |
| μ_{22} | 13.970 | 1.129 | 13.923 | 1.145 |
| μ_{23} | 8.648 | 1.633 | 8.811 | 1.620 |
| σ_1^2 | 0.025 | 0.007 | 0.022 | 0.007 |
| σ_2^2 | 0.254 | 0.050 | 0.241 | 0.055 |
| τ_1^2 | 0.026 | 0.015 | 0.032 | 0.012 |
| τ_2^2 | 0.393 | 0.120 | 0.411 | 0.109 |
| π_1 | 0.547 | 0.058 | 0.551 | 0.055 |

Table 2: Agreements and differences between the clinical and model classifications using Bayesian and Classical methods.

| classification | Model classification | | | | Groups |
|----------------|----------------------|----------|-----------|----------|--------|
| | Bayesian | | Classical | | |
| | Normal | Abnormal | Normal | Abnormal | |
| Normal | 94 | 30 | 95 | 29 | 124 |
| Abnormal | 10 | 39 | 13 | 36 | 49 |
| | 104 | 69 | 108 | 65 | 173 |

between the clinical classification and the model fit can be obtained by examining the estimated parameters for the model with their counterparts using the clinical classification. Agreement was fairly close.

Figure 3 shows the β -HCG trajectories for the two groups. In general, in one group we observe a steady growth of the trajectories. For the other group, however, profiles tend to have sharp increases at the start and then decrease towards the end of the window, or to have exceptionally high or low values. It is clear that the method has appropriately grouped similar women together. On the basis of maximizing classification probabilities, we see that the model classifies the normal women as being in component 1.

8 Discussion

We studied classical (EM algorithm) and Bayesian (MCMC) estimation of a proposed mixture of hierarchical nonlinear models. The model and methods can be useful in situations where repeated measures over time are available and the profiles show a nonlinear relationship across time. In a cluster analysis context, this is expected to lead to more reliable clustering structures since it allows to take advantage of the powerful hierarchical nonlinear models methodology in the mixtures framework. And in many situations, it could be crucial to distinguish the statistical individuals according to their variability.

An extension of the finite mixture model that would be of particular interest in many applications is the modeling of cluster membership probabilities as a function of covariates. In the context of our example, this could be accomplished using a logistic form for the population proportion of normal pregnancies, $\pi = e^{\alpha\mathbf{x}} / (1 + e^{\alpha\mathbf{x}})$ with \mathbf{x} a vector of covariates for each women and α a vector of regression parameters (Peng et al., 1996). Identifying associations between women covariates, such as age, number of previous pregnancies, and normal pregnancy tendencies can be useful for targeting specific individuals in future analysis. In our example a number of woman had missing covariate values.

Also, in our approach the number of clusters was fixed and we proposed to choose a hierarchical nonlinear model mixture with the BIC criterion. A further step we may consider, from a Bayesian

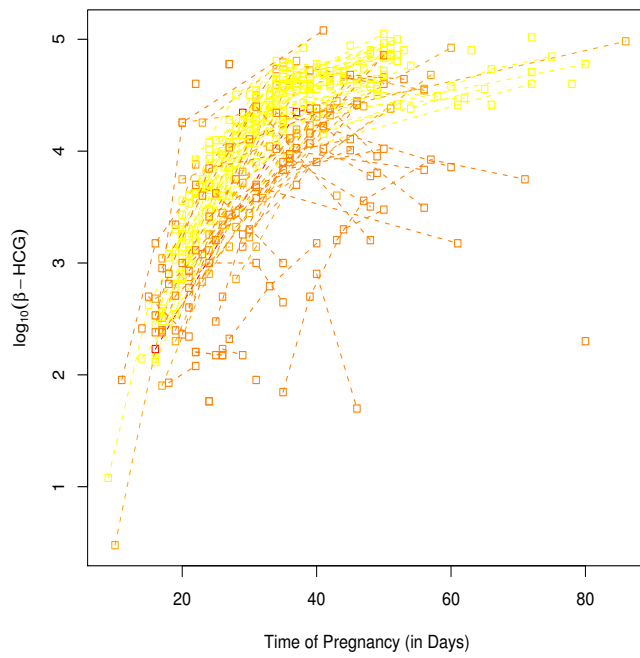


Figure 3: Trajectories with labels.

viewpoint, is to treat K as random, i.e., we could formulate the clustering problem in terms of a hierarchical nonlinear model mixture with an unknown number of components and use the reversible jump Markov chain Monte Carlo technique to define a sampler that moves between different dimensional spaces. Also, in our application, the hormone trajectories are functions of the time. Functional data analysis (Ramsay and Silverman, 1997) is an ideal alternative approach for modeling such relationships. Research along these lines is currently in progress.

Acknowledgments

The authors would like to thank two referees for their valuable comments which have helped to improve this paper.

Appendix

MCMC Algorithm Implementation Details, with superscripts indicating the iteration.

1. $s = 0$. Fix initial values $\mathbf{z}^{(0)}, \boldsymbol{\pi}^{(0)}, \boldsymbol{\mu}^{(0)}, \boldsymbol{\Sigma}^{(0)}$ (or $\tau^{2(0)}, \sigma^{2(0)}, \boldsymbol{\theta}^{(0)}$).

2. Sample $z_i^{(s+1)}$ from

$$p_{ik} = \frac{\pi_k^{(s)} p(\mathbf{y}_i | \boldsymbol{\theta}_{ik}^{(s)}, \sigma_k^{2(s)})}{\sum_{l=1}^K \pi_l^{(s)} p(\mathbf{y}_i | \boldsymbol{\theta}_{il}^{(s)}, \sigma_l^{2(s)}), \quad k = 1, \dots, K, \quad i = 1, \dots, m.$$

3. Sample $\boldsymbol{\pi}^{(s+1)} = (\pi_1^{(s+1)}, \dots, \pi_K^{(s+1)})$ from $\mathcal{D}(\delta + m_1^{(s+1)}, \dots, \delta + m_K^{(s+1)})$ where $m_k^{(s+1)} = \sum_{i=1}^m z_{ik}^{(s+1)}$.

4. Sample $\boldsymbol{\mu}_k^{(s+1)}$ from

$$\mathcal{N}(\mathbf{V}_k (m \boldsymbol{\Sigma}_k^{-1} \bar{\boldsymbol{\theta}}_k + \boldsymbol{\Sigma}_{k0}^{-1} \boldsymbol{\mu}_{k0}), \mathbf{V}_k).$$

5. Sample $\boldsymbol{\Sigma}_k^{(s+1)}$ from

$$\mathcal{W}(m + r_{k0}, \tilde{\mathbf{R}}_k).$$

5.1 In the special case $\boldsymbol{\Sigma}_k = \tau_k^2 \mathbf{I}_p$, sample $\tau_k^{2(s+1)}$ from

$$\mathcal{IG}\left(\frac{2c_{k0} + pm_k}{2}, \frac{2d_{k0}}{2 + \sum_{i=1}^m z_{ik} (\boldsymbol{\theta}_{ik} - \boldsymbol{\mu}_k)' \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{\theta}_{ik} - \boldsymbol{\mu}_k)}\right).$$

6. Sample $\sigma_k^{2(s+1)}$ from

$$\mathcal{IG}\left(\frac{2a_{k0} + \sum_{i=1}^m n_i z_{ik}}{2}, \frac{2b_{k0}}{2 + \sum_{i=1}^m z_{ik} \|\mathbf{y}_i - \mathbf{g}(\boldsymbol{\theta}_{ik}, \mathbf{x}_{ik})\|^2}\right).$$

7. Sample from the full conditional distribution of $\boldsymbol{\theta}_{ik}^{(s+1)}$ given the observations \mathbf{y} and the (vector of) parameters

$$(\mathbf{z}^{(s+1)}, \boldsymbol{\pi}^{(s+1)}, \boldsymbol{\mu}^{(s+1)}, \boldsymbol{\Sigma}^{(s+1)} \text{ (or } \tau^{2(s+1)}), \sigma^{2(s+1)})$$

via a (random walk) Metropolis-Hastings step with proposal distribution

$$q(\boldsymbol{\theta}_{ik}^{(s+1)} | \boldsymbol{\theta}_{ik}^{(s)}, \mathbf{v}_{ik}).$$

8. $s = s + 1$. Go to step 2.

References

- Booth, J. G., Casella, G., Hobert, J. P., 2005. Clustering using objective functions and stochastic search. Submitted.
- Breiman, L., Fridman, J. H., Olshen, R. A., Stone, C. J., 1984. Classification and Regression Trees. Wadsworth International Group, Inc., Belmont.
- Celeux, G., 1998. Bayesian inference for mixtures: The label switching problem. In: Payne, R., Green, P. (Eds.), *In COMPSTAT 98*. Physica-Verlag, pp. 227–232.
- Celeux, G., Hurn, M., Robert, C. P., 2000. Computational and inferential difficulties with mixture posterior distribution. *Journal of the American Statistical Association* 95, 957–970.
- Celeux, G., Lavergne, C., Martin, O., 2005. Mixture of linear mixed models for clustering gene expression profiles from repeated microarray experiments. *Statistical Modelling* 5, 243–267.
- Chib, S., 1995. Marginal likelihood from the Gibbs output. *Journal of American Statistical Association* 90, 1313–1321.
- Dasgupta, A., Raftery, A. E., 1998. Detecting features in spatial point processes with clutter via model-based clustering. *Journal of the American Statistical Association* 93, 294–302.
- Davidian, M., Giltinan, D. M., 1995. *Nonlinear Models for Repeated Measurement Data*. Chapman & Hall, London.
- De la Cruz-Mesía, R., Marshall, G., 2003. A Bayesian approach for nonlinear regression model with continuous errors. *Communications in Statistics: Theory and Methods* 32 (8), 1631–1646.
- De la Cruz-Mesía, R., Marshall, G., 2006. Nonlinear random effects models with continuous time autoregressive errors: A Bayesian approach. *Statistics in Medicine* 25 (9), 1471–1484.
- Dempster, A. E., Laird, N. M., Rubin, D. B., 1977. Maximum likelihood with incomplete data via the E-M algorithm. *Journal of the Royal Statistical Society, Series B* 39, 1–38.
- DeSarbo, W. S., Cron, W. L., 1988. A maximum likelihood methodology for clusterwise linear regression. *Journal of Classification* 5 (1), 249–282.
- Diebolt, J., Robert, C. P., 1994. Estimation of finite mixture distributions through bayesian samplings. *Journal of the Royal Statistical Society, Series B* 56, 363–375.

- Escobar, M. D., West, M., 1995. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association* 90, 577–588.
- Fitzmaurice, G. M., Laird, N. M., Ware, J. H., 2004. *Applied longitudinal analysis*. Wiley.
- Fraley, C., Raftery, A. E., 1998. How many clusters? which clustering method? answers via model-based cluster analysis. *The Computer Journal* 41, 578–588.
- Fraley, C., Raftery, A. E., 1999. MCLUST: Software for model-based cluster analysis. *Journal of Classification* 16, 297–306.
- Fraley, C., Raftery, A. E., 2002. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association* 97, 611–631.
- Frits, M. A., Guo, S. M., 1987. Doubling time of human chorionic gonadotropin (hCG) in early normal pregnancy: relationship to hCG concentration and gestational age. *Fertil Steril* 47, 584–589.
- Frühwirth-Schnatter, S., 2001. Markov chain Monte Carlo estimation of classical and dynamic switching and mixture models. *Journal of the American Statistical Association* 96, 194–209.
- Gaffney, S. J., Smyth, P., 2003. Curve clustering with random effects regression mixtures. In: Bishop, C. M., Frey, B. J. (Eds.), *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics*. Key West, FL.
- Gelman, A., Roberts, G. O., Gilks, W. R., 1996. Efficient Metropolis jumping rules. In: Bernardo, J. M., Berger, J. O., Dawid, A. P., Smith, A. F. M. (Eds.), *Bayesian Statistics 5*. Oxford University Press, Oxford, pp. 599–607.
- Geweke, J., 1995. Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In: Bernardo, J. M., Berger, J. O., Dawid, A. P., Smith, A. F. M. (Eds.), *In Bayesian Statistics 5*. Vol. 4. Oxford University Press, Oxford, pp. 169–194.
- Guo, S. W., Thompson, E. A., 1994. Monte Carlo estimation of mixed models for large complex pedigrees. *Biometrics* 50, 417–432.
- Hartigan, J. A., 1975. *Clustering Algorithms*. Wiley, New York.
- Hosmer, D., 1974. Maximum likelihood estimates of the parameters of a mixture of two regression lines. *Communications in Statistics* 3 (10), 995–1006.
- Hosmer, D. W., Lemeshow, S., 2000. *Applied logistic regression*, 2nd ed. John Wiley & Sons Inc., New York.
- Hurn, M., Justel, A., Robert, C. P., 2003. Estimating mixtures of regressions. *Journal of Computational and Graphical Statistics* 12 (1), 55–79.

- James, G., Sugar, C., 2003. Clustering for sparsely sampled functional data. *Journal of the American Statistical Association* 98, 397–408.
- Jasra, A., Holmes, C. C., Stephens, D. A., 2005. Markov chain Monte Carlo and the label switching problem in Bayesian mixture modelling. *Statistical Science* 20 (1), 50–67.
- Jennrich, R. I., Schluchter, M. D., 1986. Unbalanced repeated-measures models with structured covariance matrices. *Biometrics* 42 (4), 805–820.
- Jones, P. N., McLachlan, G. J., 1992. Fitting finite mixture models in a regression context. *Australian Journal of Statistics* 34 (2), 233–240.
- Kass, R. E., Raftery, A. E., 1995. Bayes factors. *Journal of American Statistical Association* 90, 773–795.
- Li, B., 2006. A new approach to cluster analysis: the clustering-function-based method. *Journal of the Royal Statistical Society, Series B* 68, 457–476.
- Marshall, G., Barón, A. E., 2000. Linear discriminant models for unbalanced longitudinal data. *Statistics in Medicine* 19, 1969–1981.
- McLachlan, G. J., Basford, K. E., 1988. *Mixture Models : Inference and Applications to Clustering*. Marcel Dekker.
- McLachlan, G. J., Peel, D., 2000. *Finite Mixture Models*. Wiley, New York.
- Pauler, D. K., Laird, N. M., 2000. A mixture model for longitudinal data with application to assessment of noncompliance. *Biometrics* 56, 464–472.
- Peng, F., Jacobs, R. A., Tanner, M. A., 1996. Bayesian inference in mixtures-of-experts and hierarchical mixtures-of-experts with an application to speech recognition. *Journal of the American Statistical Association* 91, 953–960.
- Pfeifer, C., 2004. Classification of longitudinal profiles based on semi-parametric regression with mixed effects. *Statistical Modelling* 4, 314–323.
- Pinheiro, J. C., Bates, D. M., 2000. *Mixed-effects models in S and S-PLUS*. Springer, New York.
- Quandt, R. E., 1972. A new approach to estimating switching regressions. *Journal of the American Statistical Association* 57, 306–310.
- Quandt, R. E., Ramsey, J. B., 1978. Estimating mixtures of normal distributions and switching regressions. *Journal of the American Statistical Association* 73, 730–738.
- Ramsay, J. O., Silverman, B. W., 1997. *Functional Data Analysis*. Springer, New York.
- Richardson, S., Green, P. J., 1997. On Bayesian analysis of mixture models with an unknown number of components. *Journal of the Royal Statistical Society, Series B* 59 (4), 731–792.

- Roeder, K., Wasserman, L., 1997. Practical Bayesian density estimation using mixtures of normals. *Journal of the American Statistical Association* 92, 894–902.
- Schwarz, G., 1978. Estimating the dimension of a model. *The Annals of Statistics* 6, 461–464.
- Segal, M. R., 1992. Tree-structured methods for longitudinal data. *Journal of the American Statistical Association* 87, 407–418.
- Smith, B. J., 2004. Bayesian Output Analysis Program (BOA). Version 1.1.2 for S-PLUS and R. Available at <http://www.public-health.uiowa.edu/boa>.
- Stephens, M., 2000. Dealing with label switching in mixture models. *Journal of the Royal Statistical Society, Series B* 62, 795–809.
- Verbeke, G., Molenberghs, G., 2000. *Linear Mixed Models for Longitudinal Data*. Springer, New York.
- Viele, K., Tong, B., 2002. Modeling with mixtures of linear regressions. *The Annals of Statistics* 27, 439–460.
- Vonesh, E. F., Chinchilli, V. M., 1997. *Linear and Nonlinear Models for the Analysis of Repeated Measurements*. Marcel Dekker.
- Wu, C. F. J., 1983. On the convergence properties of the EM algorithm. *The Annals of Statistics* 11 (1), 95–103.
- Zhang, H., 1997. Multivariate adaptive splines for analysis of longitudinal data. *Journal of Computational and Graphical Statistics* 6, 74–91.