Detecting influential observations in a model-based cluster analysis

Detecting Influential Observations in a Model-Based Cluster Analysis

Liesbeth Bruckers[1] **Geert Molenberghs**[1,2] **Geert Verbeke**[2,1] **Helena Geys**[3]

[1] *I-BioStat, Universiteit Hasselt, B-3500 Hasselt, Belgium*

[2] *I-BioStat, Universiteit Hasselt, B-3500 Hasselt, Belgium*

[3] *Janssen Pharmaceutica, B-2430 Beerse, Belgium*

Abstract

Finite-mixture models have been used to model population heterogeneity and to relax distributional assumptions. These models are also convenient tools for clustering and classification of complex data such as, for example, repeated-measurements data. The performance of model-based clustering algorithms is sensitive to influential and outlying observations. Methods for identifying outliers in a finite-mixture model have been described in the literature. Approaches to identify influential observations are less common. In this paper, we apply local-influence diagnostics to a finite-mixture model with known number of components. The methodology is illustrated on real-life data.

Some Keywords: Local Influence; Model-Based Clustering; Finite-Mixture Model.

1 Introduction

Cluster analyses are used to reveal latent groups in data. Observations are grouped in such a way that observations in the same group or cluster are more similar than observations belonging to different groups. Various methods can be employed to find clusters in data. For multivariate data, hierarchical and nonhierarchical algorithms are among the most popular ones ([1]). However, these algorithms are less appropriate for data exhibiting complex structures, as is the case for repeated-measurements

and spatial data. For these data types, model-based clustering by means of finite-mixture models is a flexible alternative.

In general, the performance of a clustering algorithm, and hence also the performance of model-based clustering, may be impacted by influential observations. Therefore, it is important to carefully examine the effect an observation has on the cluster result. Identification of influential observations has been described for hierarchical and nonhierarchical cluster analysis. For hierarchical clustering algorithms, proposals to measure the influence of a single observation on the clustering process can be found in Jolliffe et al. [2], Kim et al. [3], and Chen and Milligan [4]. Methods for detecting influential observations in nonhierarchical cluster analysis have, for example, been studied by Cerioli [5] and Cuesta-Albertos et al. [6].

Although outlier detection for model-based clustering algorithms has been investigated (see e.g., McLachlan and Peel [7], Wang et al. [8]), identification of influential observations has, to our knowledge, not yet been described.

In this paper, we show how local-influence diagnostics, as introduced by Cook [9], can be used to identify influential observations when clustering repeated-measurements data by a finite-mixture model. The influence on the mixture's mean profiles and also on the posterior probabilities, used for the classification of an individual subject, will be looked into. The paper is organized as follows. Model-based clustering and the finite-mixture model are introduced in Section 2. Section 3 briefly summarizes the outlier detection method proposed by Wang [8] for mixture populations. The concepts of local-influence diagnostics are sketched in Section 4 and applied to real-life data in Section 5. A targeted simulation study is used to investigate the performance of the local-influence diagnostics applied on a two-component mixture model. The results are presented in the Appendix (Section 7.4).

2 Finite-Mixture Models as a Tool for Clustering

Finite-mixture models [7] are latent-variable models that express the distribution of variables as a mixture of a finite number of component distributions. Finite-mixture models have been used to investigate the performance of estimators in non-normal situations and to develop robust estimators. Finite-mixture models also provide a framework for clustering

They have been used for this purpose in a wide range of applications in marketing, social, psychosocial, **medical and related research, where the data could be seen as arising from two or more populations. York et al., for example, use a finite mixture model to investigate drug-response heterogeneity in an asthma pharmacogenetic study ([10]). Finite mixture models are also used in medical image analysis ([11]).**

Finite-mixture modeling addresses the population heterogeneity in the observed data by means of categorical latent classes, that represent homogenous subpopulations. Class membership is latent and thus needs to be inferred from the data. In its general form the finite-mixture model for a r-dimensional response vector $\boldsymbol{Y}_i$ is written as: $f_i(\boldsymbol{y}_i;\boldsymbol{\pi}) = \sum_{k=1}^{K} \pi_k f_{ik}(\boldsymbol{y}_i)$. Here, π_k is the kth mixing proportion or the probability that an observation belongs to the kth subpopulation or component and $f_{ik}(\boldsymbol{y}_i)$ its corresponding density. K represents the total number of subpopulations and $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)'$, with $0 < \pi_k < 1$, for all $k = 1, \ldots, K$ and $\sum_{k=1}^{K} \pi_k = 1$.

When modeling repeated-measurements data, unobserved individual heterogeneity in the evolution of an outcome over time is in general captured by continuous latent variables. In the framework of mixed models, random effects are used to address the correlation between measurements of the same subject and at the same time they allow for subject-specific evolutions of the response. Let $\boldsymbol{Y}_i$ $=(Y_{i1}, \ldots, Y_{in_i})'$ denote the vector of n_i repeated observations for subject i ($i = 1, \ldots, N$), and let $f(\boldsymbol{y}_i|\boldsymbol{b}_i)$ be the corresponding density, conditional on a q-dimensional vector $\boldsymbol{b}_i$ of random effects. The choice of the distribution of Y_{ij} given $\boldsymbol{b}_i$ is driven by the nature of the response, i.e., generally normal for continuous data, binomial for binary data, and Poisson for count data.

The mean of outcome Y_{ij} is modeled as:

$$E(Y_{ij}|\boldsymbol{b}_i, \boldsymbol{\beta}; \boldsymbol{x}_{ij}, \boldsymbol{z}_{ij}) = g(\boldsymbol{x}_{ij}, \boldsymbol{\beta}, \boldsymbol{z}_{ij}, \boldsymbol{b}_i).$$

The function $g(\cdot)$ is an arbitrary function. The p-dimensional vector $\boldsymbol{\beta}$ contains the fixed-effects parameters at population level, the q-dimensional vector $\boldsymbol{b}_i$ are the random effects of the ith subject, and the vectors $\boldsymbol{x}_{ij}$ and $\boldsymbol{z}_{ij}$ contain covariate information for the ith subject at measurement j, corresponding to the fixed and random effects, respectively. The random effects are assumed to follow a distribution, $f(\boldsymbol{b}_i)$, frequently normal. The model assumes that the random effects are

drawn from one homogeneous population of random effects. However, heterogeneity in the random-effects population is to be expected when the study population consists of a number of (unlabelled) subpopulation. Therefore, the mixed model has to be expanded to include categorical as well as continuous latent variables. The categorical latent variable captures heterogeneity between subjects arising from the fact that they belong to different subpopulations. **When different classes constitute the mixture, and no variation across individuals within classes is allowed (except for residual variation), the model is referred to as a latent class growth mixture model [12].** When within-class variation of individuals is allowed for, through continuous latent random effects, the model is termed growth mixture model [13].

We will follow here the approach of Verbeke and Lesaffre [14] and Spiessens et al. [15] to model heterogeneity in repeated-measurements data by means of continuous and categorical latent variables. They specify a mixture for the distribution of the random effects. More specifically, they extend the normality assumption of the random effects $\boldsymbol{b}_i$ to incorporate mixtures of normal components,

$$\boldsymbol{b}_i \sim \sum_{k=1}^{K} \pi_k N(\boldsymbol{\mu}_k, D_k),$$

where, as before, π_k is the proportion of subjects belonging to subpopulation k, described by the multivariate normal distribution $N(\mu_k, D_k)$. For identifiability we require that, $\boldsymbol{\mu}_i \neq \boldsymbol{\mu}_j$, $\pi_1 \geq \pi_2 \geq \ldots \geq \pi_K > 0$, and $E(\boldsymbol{b}_i) = \sum_{k=1}^{K} \pi_k \boldsymbol{\mu}_k = 0$. Under this assumption, the marginal density function of $\boldsymbol{Y}_i$ turns out to be a mixture of densities $f_{ik}(\boldsymbol{y}_i)$ with mixture probabilities $\pi_1, \ldots, \pi_K$:

$$f_i(\boldsymbol{y}_i) = \sum_{k=1}^{K} \pi_k f_{ik}(\boldsymbol{y}_i). \tag{1}$$

In this expression $f_{ik}(\boldsymbol{y}_i)$ is the marginal density function, for the kth subpopulation, corresponding to a mixed model with random effects that are normally distributed with mean $\boldsymbol{\mu}_k$ and covariance matrix D_k.

Estimates for all parameters in the model are obtained by maximising the log-likelihood

$$l(\boldsymbol{\theta}|\boldsymbol{y}) = \sum_{i=1}^{N} \log \left[\sum_{k=1}^{K} \pi_k f_{ik}(\boldsymbol{y}_i|\boldsymbol{\Psi}) \right], \tag{2}$$

by means of the Expectation-Maximisation (EM) algorithm [16]: The vector $\boldsymbol{\theta}$ in (2) contains all parameters in the model $\boldsymbol{\theta} = (\boldsymbol{\Psi}', \boldsymbol{\pi}')' = (\boldsymbol{\beta}', (\boldsymbol{\mu_1}', \ldots, \boldsymbol{\mu_K}'), (\text{vech}(D_1)', \ldots, \text{vech}(D_K)'), \boldsymbol{\pi}')'$ with vech(D_k) containing all upper-triangular elements of D_k stacked on top of each other. The EM algorithm assumes the presence of missing observations, which for a finite-mixture model are the group memberships.

When the goal of the statistical analysis is not only to obtain parameter estimates but also assignment of the subjects to the subpopulation they belong to, we term this model-based clustering. A subject's posterior probabilities, $\pi_k f_{ik}(\boldsymbol{y}_i)/f_i(\boldsymbol{y}_i)$, are used to classify its longitudinal profile into one of the K components. Spiessens, Verbeke, and Komàrek [15] have developed a SAS macro, based on the SAS procedure NLMIXED, that implements the EM algorithm for fitting nonlinear and generalised linear models with finite normal mixtures as random-effects distributions. The macro also classifies the longitudinal profiles into the different components.

3 Outlier Detection for a Finite-Mixture Model

Wang et al. [8] describe a procedure that looks for outliers from a mixture of normal distributions, where at least some ground-truth information (labelling) is available and the number of components in the mixture is known. Sain et al. [17] extended this procedure to the case where no ground-truth information is available and the number of components is unknown. Specifically, the authors assume that the training data of size N is a sample from a mixture distribution of K distributions, $f_i(\boldsymbol{y_i}) = \sum_{k=1}^{K} \pi_k f_{ik}(\boldsymbol{y_i}; \boldsymbol{\mu}_k, \Sigma_k)$, with $f_{ik}(\boldsymbol{y_i}; \boldsymbol{\mu}_k, \Sigma_k)$ a normal distribution. To investigate if a new observation, $\boldsymbol{y}_{N+1}$, is obtained from the mixture population or from an outlier population they use a likelihood ratio test statistic that does not require the distribution of the outlier population. The classical likelihood ratio test statistic is the ratio of the maximized likelihood functions $L_0(\boldsymbol{\theta}_0) = [\prod_{i=1}^{N} f_i(\boldsymbol{y}_i; \boldsymbol{\theta}_0)] f(\boldsymbol{y}_{N+1} \boldsymbol{\theta}_0)$ under H_0, and $L_1(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1) = [\prod_{i=1}^{N} f_i(\boldsymbol{y}_i; \boldsymbol{\theta}_0)] h(\boldsymbol{y}_{N+1}; \boldsymbol{\theta}_1)$ under H_1, with $h(\boldsymbol{y}; \boldsymbol{\theta}_1)$ the density associated with the outlier population. When there is only a single observation from the outlier population, maximizing L_1 is hard. Wang et al. [8] noted that a viable test statistic could be based on the ratio $L_0/\tilde{L}_1$, with $\tilde{L}_1(\boldsymbol{\theta}_0) = \prod_{i=1}^{N} f_i(\boldsymbol{y}_i; \boldsymbol{\theta}_0)$, eliminating the need for $h(\boldsymbol{y}; \boldsymbol{\theta}_1)$. Thus, in essence, the principle of equi-ignorance is employed, and the

distribution of the outlier is not needed. The resulting modified likelihood ratio test statistic

$$W(\boldsymbol{y}_{N+1}; \boldsymbol{y}_1, \ldots, \boldsymbol{y}_N) = \frac{\sup_{\theta_0 \in \Theta} L_0(\boldsymbol{\theta}_0)}{\sup_{\theta_0 \in \Theta} \tilde{L}_1(\boldsymbol{\theta}_0)}$$

will take small values when $\boldsymbol{y}_{N+1}$ departs from f. The null distribution of W is obtained through nonparametric bootstrap [18]. The authors examined the power of the outlier test based on W via simulations.

4 Review of General Theory for Local Influence

Local influence was presented by Cook ([21], [9]) and used by several authors since. The impact of individuals and measurements on the analysis is assessed by comparing standard maximum likelihood estimates with those resulting from slightly perturbing the contribution of an individual or a measurement. The method is to be contrasted with global influence, or case deletion, where impact is assessed by simply deleting an individual or measurement. While local influence comes with a certain amount of technicality, it is easy and fast to calculate in practice, and in many cases leads to interpretable components of influence. Lesaffre and Verbeke [22] introduced an influence assessment paradigm for the linear mixed model. A review of several diagnostic procedures for the linear mixed model is given in Mun and Lindstrom [23]. Verbeke et al. [24] used local influence for longitudinal Gaussian data with dropout, while incomplete binary data were studied by Jansen et al. [25]. Verbeke and Molenberghs [26] and Molenberghs and Verbeke [27] study the method and provide ample references. Ouwens, Tan, and Berger [28] applied local influence to the generalized linear mixed model for count data, i.e., the Poisson-normal model.

Let the log-likelihood for the chosen model take the form

$$\ell(\boldsymbol{\theta}) \; = \; \sum_{i=1}^{\mathsf{N}} \ell_i(\boldsymbol{\theta}), \tag{3}$$

in which $\ell_i(\boldsymbol{\theta})$ is the contribution of the ith individual to the log-likelihood. Let

$$\ell(\boldsymbol{\theta}|\boldsymbol{\omega}) \;=\; \sum_{i=1}^{N} w_i \ell_i(\boldsymbol{\theta}), \tag{4}$$

now denote the perturbed version of $\ell(\boldsymbol{\theta})$, depending on an N-dimensional vector $\boldsymbol{\omega}$ of weights, assumed to belong to an open subset Ω of $\mathbb{R}^N$. The original log-likelihood (3) follows for $\boldsymbol{\omega} = \boldsymbol{\omega}_0 = (1, 1, \ldots, 1)'$. The perturbed log-likelihood gives more or less weight to log-likelihood contributions of single subjects.

Let $\widehat{\boldsymbol{\theta}}$ be the maximum likelihood estimator for $\boldsymbol{\theta}$, obtained by maximizing $\ell(\boldsymbol{\theta})$, and let $\widehat{\boldsymbol{\theta}}_{\boldsymbol{\omega}}$ denote the estimator for $\boldsymbol{\theta}$ under $\ell(\boldsymbol{\theta}|\boldsymbol{\omega})$. Cook [9] proposed to measure the distance between $\widehat{\boldsymbol{\theta}}_{\boldsymbol{\omega}}$ and $\widehat{\boldsymbol{\theta}}$ by the so-called likelihood displacement, defined by

$$\mathrm{LD}(\boldsymbol{\omega}) \;=\; 2\left(\ell(\hat{\boldsymbol{\theta}}) - \ell(\hat{\boldsymbol{\theta}}_{\boldsymbol{\omega}})\right).$$

$\mathrm{LD}(\boldsymbol{\omega})$ will be large if $\ell(\boldsymbol{\theta})$ is strongly curved at $\widehat{\boldsymbol{\theta}}$ (which means that $\boldsymbol{\theta}$ is estimated with high precision) and small otherwise. A graph of $\mathrm{LD}(\boldsymbol{\omega})$ versus $\boldsymbol{\omega}$ brings out information on the influence of case-weight perturbations. The graph is the geometric surface formed by the values of the $(N+1)$-dimensional vector

$$\boldsymbol{\xi}(\boldsymbol{\omega}) \;=\; \begin{pmatrix} \boldsymbol{\omega} \\ \\ \mathrm{LD}(\boldsymbol{\omega}) \end{pmatrix}$$

as $\boldsymbol{\omega}$ varies throughout Ω. Following Cook [9] and Verbeke and Molenberghs [26], we will refer to $\boldsymbol{\xi}(\boldsymbol{\omega})$ as an influence graph. It is unfeasible to evaluate $\mathrm{LD}(\boldsymbol{\omega})$ for all $\boldsymbol{\omega}$. Cook [9] describes the sensitivity of $\ell(\hat{\boldsymbol{\theta}})$ by looking at small perturbations for case weights around $\boldsymbol{\omega}_{\mathbf{0}}$, i.e., the local behaviour of $\mathrm{LD}(\boldsymbol{\omega})$ around $\boldsymbol{\omega}_{\mathbf{0}}$. The normal curvature C_h of $\mathrm{LD}(\boldsymbol{\omega})$ at $\boldsymbol{\omega}_{\mathbf{0}}$, in the direction of a unit vector $\boldsymbol{h}$ in Ω, was used to quantify the local behaviour of $\mathrm{LD}(\boldsymbol{\omega})$ around $\boldsymbol{\omega}_{\mathbf{0}}$. Cook [9] derived a convenient computational scheme. **Let Δ_i be the s-dimensional vector of second-order derivatives of $\ell(\boldsymbol{\theta}|\boldsymbol{\omega})$, with respect to ω_i and all s components of $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_s)'$, and evaluated at $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}$ and at $\boldsymbol{\omega} = \boldsymbol{\omega}_0$. For the log-likelihood in (4), the pth element of Δ_i is**

equal to $\frac{\partial l_i(\boldsymbol{\theta})}{\partial \theta_p}$ **(**$p = 1, \ldots, s$**), evaluated at** $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}$ **and at** $\boldsymbol{\omega} = \boldsymbol{\omega}_0$**.**

Also, write Δ for the $s \times N$ matrix with Δ_i in the ith column. Let $\ddot{L}$ denote the $s \times s$ matrix of second-order derivatives of $\ell(\boldsymbol{\theta})$, evaluated at $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}$. For any unit vector $\boldsymbol{h}$ in Ω, it follows that:

$$C_h \;=\; 2 \Big| \, \boldsymbol{h}' \Delta' \ddot{L}^{-1} \Delta \boldsymbol{h} \, \Big| . \tag{5}$$

Various choices for $\boldsymbol{h}$ have received specific attention. First, one can focus on a single subject i only, by choosing $\boldsymbol{h} = \boldsymbol{h_i}$, the zero vector with a sole value 1 in the ith position. The normal curvature is then called the total local influence and is given by

$$C_i \;\equiv\; C_{h_i} \;=\; 2 \Big| \, \boldsymbol{\Delta}_i' \ddot{L}^{-1} \boldsymbol{\Delta}_i \, \Big| . \tag{6}$$

Large values of C_i are obtained for subjects for which small perturbations in case weight result locally in a large log-likelihood displacement.

Second, $\boldsymbol{h} = \boldsymbol{h}_{\max}$ can be chosen as the direction of maximal normal curvature $C_{\max}$. It was shown that $\boldsymbol{h}_{\max}$ is the eigenvector of $-\Delta' \ddot{L}^{-1} \Delta$ corresponding to the largest eigenvalue ([29] , [30], [26], [31]). $\boldsymbol{h}_{\max}$ permits detection of individuals that are simultaneously influential.

The total local influence of individual i can be expressed in terms of the nonzero eigenvalues, $\lambda_1 \geq \ldots \geq \lambda_s > 0$ and normalized orthogonal eigenvectors $\boldsymbol{h}_{\max} \equiv \boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_s$ of $-\Delta' \ddot{L}^{-1} \Delta$:

$$C_i = 2 \sum_{j=1}^{s} \lambda_j \nu_{ij}^2,$$

with ν_{ij} the ith component of $\boldsymbol{\nu}_j$. $C_{\max}$ is twice the largest eigenvalue, $C_{\max} = 2 \cdot \lambda_1$. This holds a warning: it is possible for C_i to be large without the same holding for the ith component in $\boldsymbol{h}_{\max}$, provided the corresponding components are large for some of the secondary eigenvectors. It is thus recommended to examine both the C_i and $\boldsymbol{h}_{\max}$.

Lesaffre and Verbeke [22] proposed a threshold for C_i above which an individual is defined as "remarkable". They state that the ith subject is influential if C_i is larger than the cutoff value $2 \sum_{i=1}^{\mathsf{N}} C_i / N$.

The methodology still applies when interest is in a subset $\boldsymbol{\theta}_1$ of $\boldsymbol{\theta} = (\boldsymbol{\theta}_1', \boldsymbol{\theta}_2')'$. It follows that (Verbeke

and Molenberghs [26]) the influence on the estimation of the subset $\boldsymbol{\theta}_1$ is given by:

$$C_h(\boldsymbol{\theta}_1) = 2\left|\boldsymbol{h}'\Delta'\left[\ddot{L}^{-1} - \begin{pmatrix} 0 & 0 \\ 0 & \ddot{L}_{22}^{-1} \end{pmatrix}\right]\Delta\boldsymbol{h}\right| \leq C_h, \tag{7}$$

$\ddot{L}_{22}$ is defined by the partition of $\ddot{L} = \begin{pmatrix} \ddot{L}_{11} & \ddot{L}_{12} \\ \ddot{L}_{21} & \ddot{L}_{22} \end{pmatrix}$ according to the dimensions of $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$. Should $\ddot{L}_{12} = 0$, then $C_h = C_h(\boldsymbol{\theta}_1) + C_h(\boldsymbol{\theta}_2)$. For weakly correlated sub-vectors, this decomposition holds approximately.

To obtain local-influence diagnostics for a finite-mixture model with K components (see Section 2), first and second derivatives of $l(\boldsymbol{\theta}|\boldsymbol{\omega})$ with respect to ω_i and all components of $\boldsymbol{\theta}$, have to be obtained. Under the finite-mixture model, $\boldsymbol{\theta}$ contains the fixed- and random-effects parameters describing the profiles for the K components and the mixture probabilities. Further, the contribution $l_i(\boldsymbol{\theta})$ of the ith subject to the log-likelihood is $l_i(\boldsymbol{\theta}) = \log\left[\sum_{k=1}^{K}\pi_k f_{ik}(\boldsymbol{y}_i|\boldsymbol{\theta})\right]$. Expressions for the derivatives involved in C_i, for a finite-mixture model, can be found in the Appendix (Section 7.1).

Often, interest is not only in the stability of the components' mean profiles but also on the influence on the posterior probabilities. The theory of local influence, as described above, allows, in an elegant way, quantification of the influence of subject i on the posterior probability of subject j. To this end, log-likelihood (4) has to be parameterized as a function of the posterior probabilities. Given the relation between the posterior probabilities and the mixture probabilities, $\pi_{jk} = \pi_k f_{jk}(\boldsymbol{y}_j)/\sum_{k=1}^{K}\pi_k f_{jk}(\boldsymbol{y}_j)$, it is possible to express the contribution of the ith individual to the log-likelihood as a function of the posterior probability of the jth individual. The log-likelihood then takes the form $l(\boldsymbol{\Psi}', \boldsymbol{\pi}_j') = \sum_{i=1}^{N} l_i(\boldsymbol{\Psi}', \boldsymbol{\pi}_j')$, with $\boldsymbol{\pi}_j = (\pi_{j1}, \ldots, \pi_{jK})'$ the vector of posterior probabilities for subject j. The likelihood as a function of the posterior probabilities, $l_i(\boldsymbol{\Psi}', \boldsymbol{\pi}_j')$, is presented in the Appendix (Section 7.2). The local influence of subject i, on the posterior probabilities of subject j can be obtained via (7).

5 Data Applications

5.1 Pharmacokinetic Data

Nonlinear mixed models are often used in pharmacokinetics to study how a drug disperses through subjects. To illustrate the local influence approach on a nonlinear model we will use data presented by Pinheiro and Bates [34]. Serum concentrations of the drug theophylline was measured in 12 subjects over a 25-hour period after oral administration. Pinheiro and Bates considered a first-order compartment model, allowing for random variability between subjects. Let Y_{ij} denote the observed concentration of the ith subject at time t_{ij}, D the dose of theophylline, k_{ei} the elimination rate constant for subject i, k_{ai} the absorption rate constant for subject i, Cl_i the clearance for subject i, and ε_{ij} normal errors. The model for the observed concentration is specified as:

$$Y_{ij} = \frac{Dk_{ei}k_{ai}}{Cl_i(k_{ai} - k_{ei})}[\exp(-k_{ei}t_{ij}) - \exp(-k_{ai}t_{ij})] + \varepsilon_{ij}. \tag{8}$$

The clearance, elimination, and absorption rates for subject i were functions of fixed and random effects:

$$Cl_i = \exp(\beta_1 + b_{i1}), \tag{9}$$

$$k_{ai} = \exp(\beta_2 + b_{i2}), \tag{10}$$

$$k_{ei} = \exp(\beta_3). \tag{11}$$

The random effects allow for heterogeneity between subjects. The $\boldsymbol{b}_i = (b_{i1}, b_{i2})'$ are assumed to follow a multivariate normal distribution with mean zero and an unknown covariance matrix.

The expected concentration level in the body as a function of time, for a typical patient (i.e., random effects equal to zero) is displayed in Figure 1. The model fit criteria for this homogeneity model are as follows: log-likehlihood -178.2, AIC 368.5, and BIC 371.4. When carrying out a two-component heterogeneity model the fit criteria are: log-likelihood -162.9, AIC 343.9, and BIC 348.3; indicating a better fit. The mixing probabilities are .53 and .47. Based on their posterior probability, 6 subjects were classified into the first component and 6 in the second. The expected concentration for both

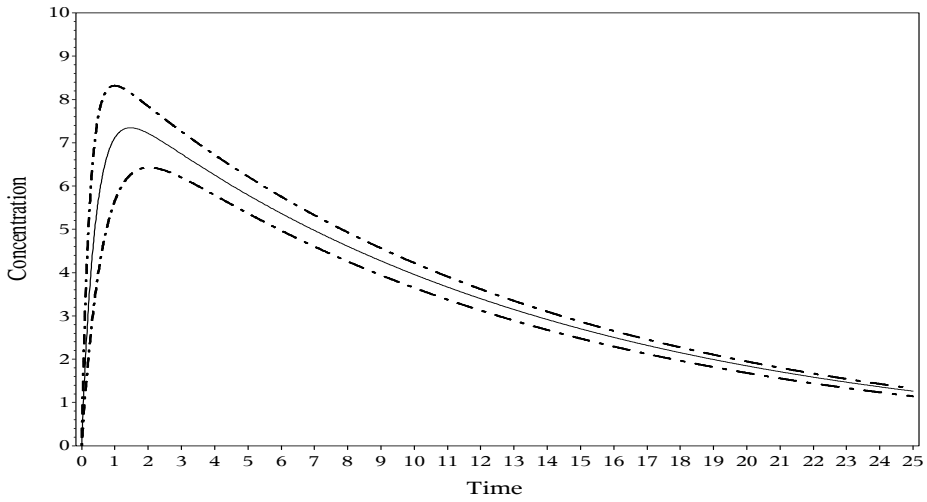

Figure 1: *Evolution of the concentrations - Theophylline data. For each component in the mixture distribution, the evolution for a typical patient is displayed (full line: one-component model, $-.-.$: two-component model).*

components, for a typical subject, can be found in Figure 1. The two components distinguish in terms of the maximum concentration level attained and the time after adminstration that the maximum concentration is reached. The first component reaches its maximum concentration level faster, and the maximum level attained is higher than compared to the maximum level of the second component.

An influence analysis for the two-component heterogeneity model does not reveal cases being locally influential (see Figure 2). All subjects' C_i are below the cut-off value of 2.14. The components of $\boldsymbol{h}_{\max}$ indicate that subjects 1, 2 , 5, 10, and 12 exhibit larger contributions in the direction of maximal curvature; C_{max} equals 12.76. When looking into the plots of the total local influence for a specific fixed or random parameter, these subjects have value of C_i exceeding the cut-off (data not shown).

To evaluate the likelihood function of the nonlinear model for the pharmacokinetic data numerical integration was used. The accuracy of the likelihood evaluation, and as such also of the local influence diagnostics, is a function of the number of quadrature points used in the numerical integration. Increasing the number of quadrature points will result in more accurate diagnostics.

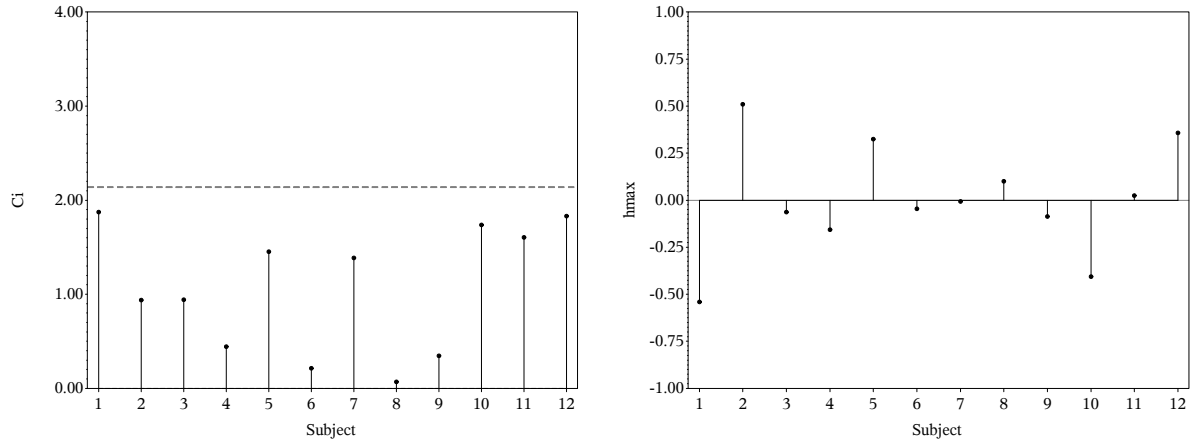

Figure 2: *Local-influence analysis - Theophylline data. The left-hand side figure displays the total local influence versus the patient numbers. The horizontal line represents the cut-off value for C_i. The right-hand side figure shows the indexplot of the components of $\boldsymbol{h}_{max}$.*

5.2 EEG Data

The aim of EEG studies is to characterize the effects of psychotropic drugs on cortical brain activity, on the basis of spectral electro-encephalograms. An EEG study in rats, conducted at Janssen Pharmaceutica (Belgium), is used. Although the brain waves of rats and humans are observed in comparable frequency bands, not all functionalities are the same. There are, however, more similarities than differences, making experiments measuring the electrical brain activity in rats very interesting to study the effect of psychoactive agents on the activity of human brains.

Depending on the frequency measurements range, the brain activity is referred to as delta activity (below 4 Hz per second), theta activity (4–7.5 Hz per second), alpha activity (8-12.5 Hz per second), beta activity (13–30 Hz per second), and gamma activity (above 30 Hz per second). Except for the delta activity, activities are refined in low and high activity (e.g., $\alpha_1, \alpha_2, \ldots$). So the EEG results in a multivariate outcome vector (i.e. 9 outcomes: $\alpha_1, \alpha_2, \beta_1, \beta_2, \theta_1, \theta_2, \gamma_1, \gamma_2, \delta$) to describe brain activity. These outcomes can be measured over time and at different positions in the brain.

The EEG study includes 10 psychoactive compounds at 4 different doses, including a placebo dose. To each compound, 32 rats were randomly assigned, 8 per dose group. The brain signals of the rats in active wake state are monitored over time. For this purpose the 9-

variate response is measured over time. A baseline measurement is taken at adminstration of the drug. A total of 8 follow-up (post drug administration) measurements are taken. The first follow-up measurement is forty-five minutes after administration of the psychoactive agent, whereafter a measurement is obtained every 15 minutes. Furthermore, the brain signal is monitored at six different positions in the brain (left and right frontal, left and right parietal, left and right occipital). For each rat this results in a 9-variate longitudinal response ($\alpha_1, \alpha_2, \ldots$), for each of the 6 different positions in the brains. We will focus on the γ_2 frequencies obtained at the left prefrontal cortex for two psychoactive compounds, PCP and Donepezil, administered at the highest dose. This reduces the data set to 16 rats, 8 rats administered the highest dose of PCP and 8 rats administered the highest dose of Donepezil. Gamma waves are related to strong mental activity like solving problems, fear, and awareness. PCP in low to moderate doses acts as a stimulant, whilst at higher doses it has a sedative effect. Donepezil is a cholinesterase inhibitor and is used to treat moderate to severe dementia of the Alzheimer's type.

To visualize the data, the individual γ_2 longitudinal profiles are given in Figure 3. The response of interest is the percentage change with respect to the measurement at baseline Y_{ib} (administration of the drug): $Y_{ij}^{'} = 100(Y_{ij} - Y_{ib})/Y_{ib}$. At baseline all percentage changes are by definition equal to zero. The graphical display therefore excludes the baseline data. In graphical displays and in the statistical models, time zero refers to the first measurement obtained after administering the drug (i.e., after 45 minutes).

Heterogeneity is seen in the γ_2 waves; some rats have a decrease in the frequency while for others an increase is obtained as an effect of the drug. This heterogeneity is of course caused by administrating 2 different drugs.

The heterogeneity model will be assumed, with a quadratic evolution of the relative γ_2 ratio over time and a random intercept that is a mixture of 2 normal distributions. So for component k ($k = 1, 2$) in the mixture we have:

$$Y_{ij}^k = \beta_0^k + \beta_1^k t_{ij} + \beta_2^k t_{ij}^2 + b_i + \varepsilon_{ij}, \tag{12}$$

where β_0^k, β_1^k, and β_2^k are component-specific fixed parameters describing the mean γ_2 profiles, b_i

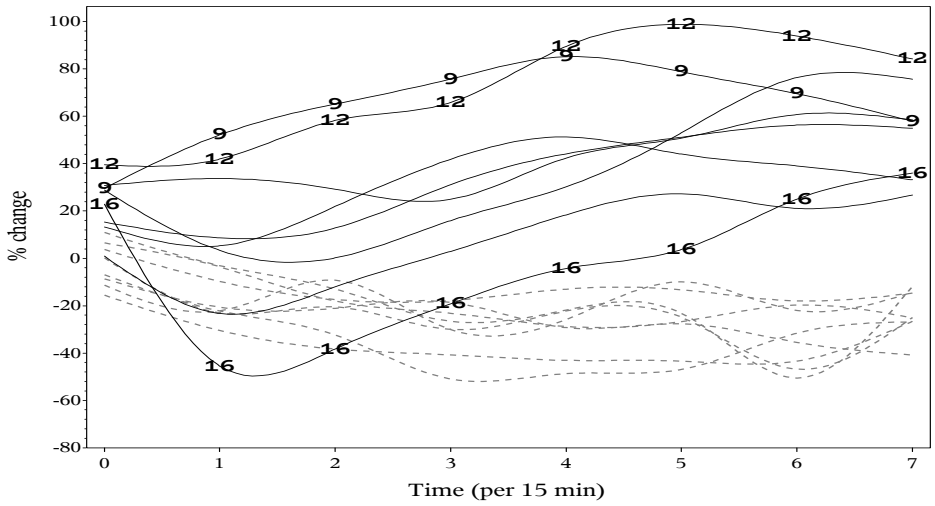

Figure 3: *Smoothed observed %change for γ_2 profiles (full lines: PCP, dotted lines: Donepezil) - EEG data. The origin of the time axis is at the first measurement after adminstration of the drug.*

are rat specific intercepts sampled from a 2-component model, and $\varepsilon_{ij} \sim N(0, \sigma_\varepsilon)$.

When applying the cluster algorithm, information about the drug a rat was given, was not taken into account. The log-likelihoods (BIC) for the one and two-component model were respectively -571.1 (1156.1) and -531.7 (1088.4). The model hypothesizing a mixture is outperforming the one-component model. Classification of the rats into the two components, based on their posterior probabilities, perfectly coincides with the two drug groups included in the analysis. The 8 rats on PCP (ids 9–16) are classified together into cluster 1, and the 8 rats on Donepezil (ids 1–8) into cluster 2.

The results of a local-influence analysis for the two-component model are displayed in Figure 4.

Three rats (ids 9, 12, and 16) of the PCP group are locally influential, based on their C_i value. The observed profiles of these rats are highlighted in Figure 3. These rats also have a large component in the direction of maximal curvature $\boldsymbol{h}_{\max}$. The maximal curvature equals 10.63.

The contribution to $\boldsymbol{h}_{\max}$ is largest for rats 16, 14 and 9. Rats 16 and 14 were also characterized by a high C_i. However, for rat 14 the C_i was not rated as exceptionally high.

To study the influence on subsets of parameters of model (12), expression (7) was used. The local-influence diagnostics were obtained for the cluster specific average profiles, the random components,

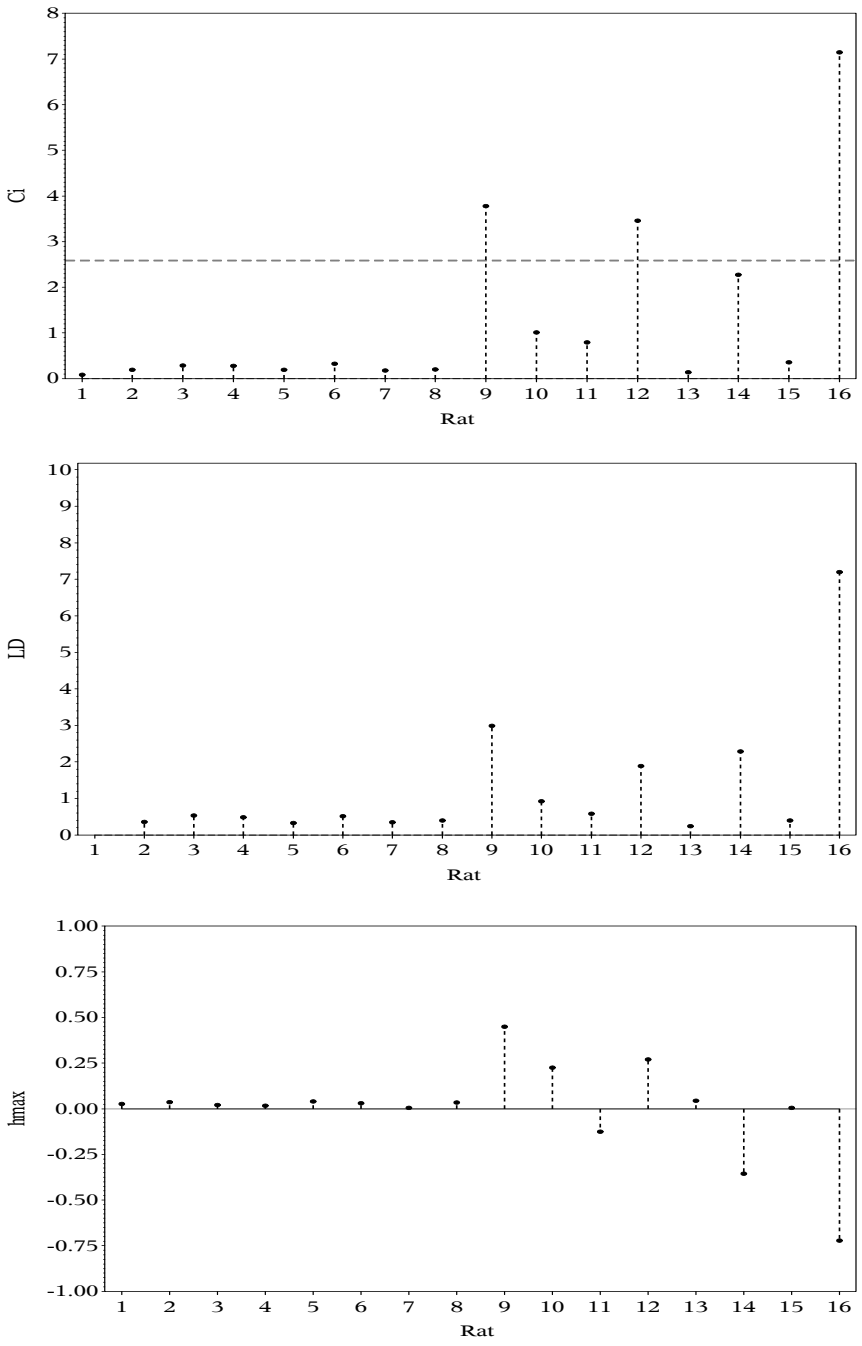

Figure 4: *Total local influence, likelihood displacement, and direction of maximal curvature versus rat identification numbers - EEG data. The horizontal line in the total local influence graph represents the cut-off value for C_i.*

and the mixture probability. The results are presented in Figure 5.

The influence of rats 9, 12, and 16 is visible in the set of fixed parameters characterizing the average

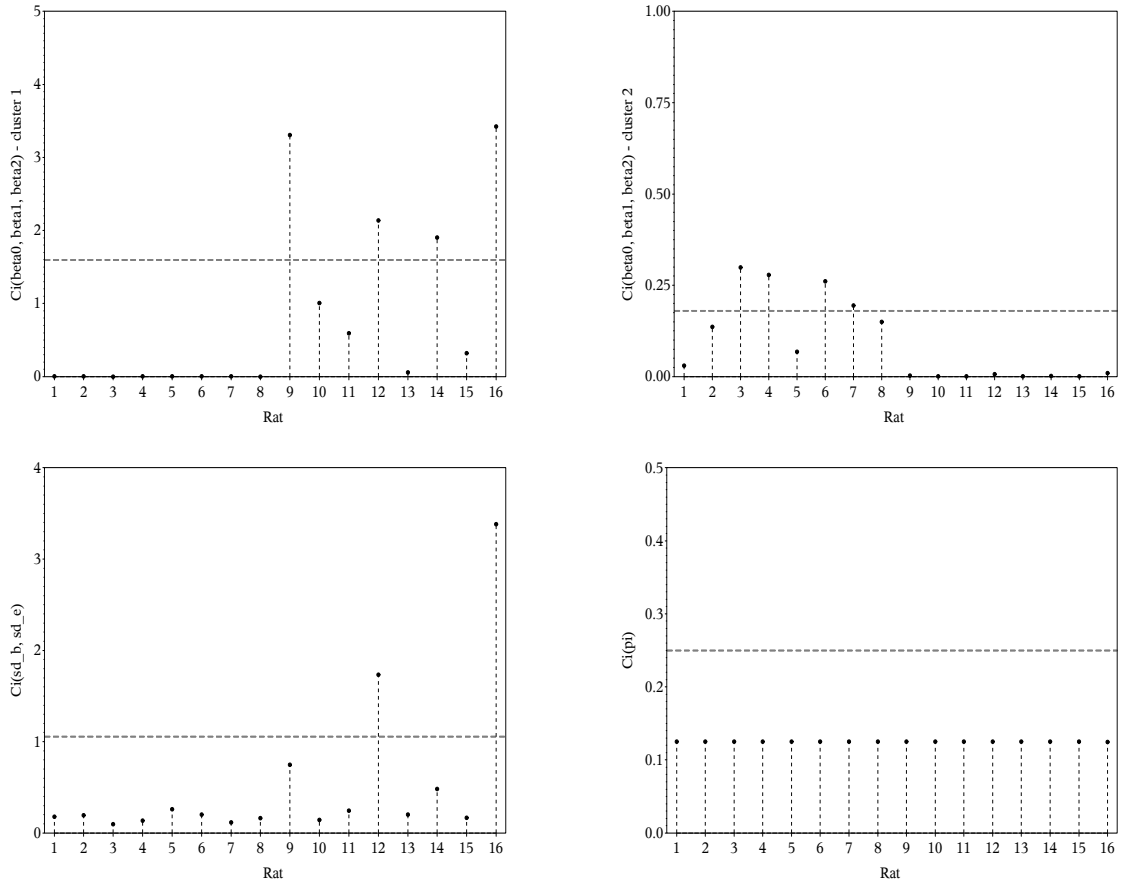

Figure 5: *Plot of local-influence diagnostics for the cluster specific average profiles, the random components and the mixture probability - EEG data. The horizontal lines represent the cut-off values for the displayed influences.*

evolution of the cluster they belong too (cluster 1), their influence on the average profile of cluster 2 is negligible. Rats 12 and 16 are also influential for the random effects (σ_b) and residual error (σ_ϵ), their diagnostics exceed the cutoff value. The mixture probability is not subject to small perturbations in the case-weights. This is not a surprise, the influence on cluster size can not be extremely different for different observations, with only a small set of observations to be clustered.

Local-influence diagnostics for the posterior probabilities are presented in a heatmap (Figure 6), summarizing the local influence that rat i has on the posterior probability of rat j, in the crossing of ith column with the jth row. The values in the graph are standardized, a value of one corresponds to a $C_i(\pi_{j1})$ equal to the cutoff value above which rat i is considered to be influential for the posterior probability of rat j to belong to component 1. The influence on the posterior probabilities

of rats 7, 12, and 14 could not be investigated. **The hessian of the log-likelihood is singular when parameterizing it as a function of the posterior probabilities of these rats. While it is hard to list all situations in which a singular hessian can occur, it is well understood that a non continuously differentiable likelihood, e.g. as a result of quasi complete separation ([35], [36]), or reduced sample size may result in a singular hessian. For rats 7, 12 and 14, the posterior probability to belong to the first cluster is either extremely close to 1 or 0, indicating separation.**

It can be seen that a perturbation in the case-weight of rat 16 influences the posterior probabilities of the other rats. Rat 9 influences the posterior probability of rats 11 and 16, rat 12 influences the posterior probabilities of rats 9, 10, 11, 13 and 15. The other influence diagnostics did not exceed the cutoff values.

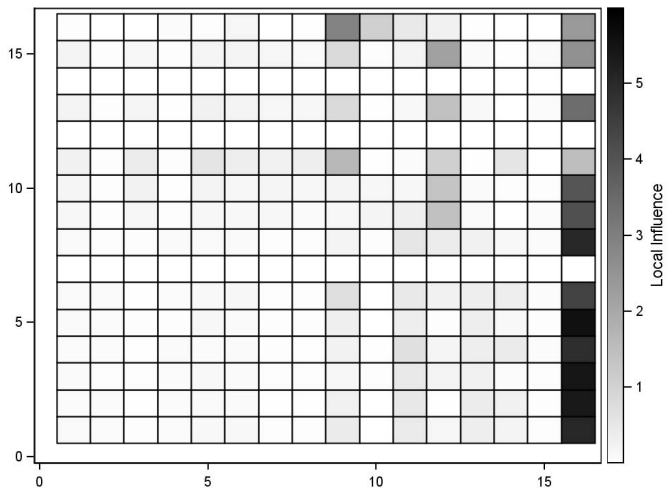

Figure 6: *Local-influence diagnostics for the posterior probabilities to belong to the first component of the mixture - EEG data. The crossing of ith column with the jth row displays the local influence that rat i has on the posterior probability of rat j to belong to the first component of the mixture. Values above 1 are considered to be influential.*

Classical diagnostics are generally based on case deletion. The likelihood displacements obtained by deletion of one rat at a time from the analysis is given in Figure 4. **This likelihood displacement is corresponds to the vector $\boldsymbol{\omega}$ with $\omega_i = 0$ and $\omega_j = 1$ for all $j \neq i$.** The largest likelihood displacements are seen for rats 9, 12, 14 and 16. It is reassuring that these are also the rats that stand out in the local-influence analysis. The influence measures do however not agree for the ranking of

rats 12 and 14.

To investigate if rat 16 is to be considered an outlier, the detection procedure described by Sain et al. (1999) was employed. The profiles of the first 15 rats are assumed to be sampled from a two-component mixture population, and the modified likelihood ratio test is used to see whether the profile of rat 16 belongs to an outlier population or not. The value of the modified likelihood ratio test statistic W equals 0.46. Applying this procedures for rats 9 and 12, the value of the modified likelihood ratio statistic equals 0.74 and 0.84, respectively. The null distribution of the test statistic was obtained via 999 nonparametric bootstrap samples. The 1st (5th) percentile of the distribution equals 0.52 (0.68). Thus, the profile of rat 16 is an outlying observation at the 1% level of significance. On the other hand, the profiles of rat 9 and 12, also flagged in the local-influence analysis, are considered to belong to the two-component mixture population, according to this approach.

The solution of a three component mixture model for the EEG data and the results of a local influence analysis for this model are presented in the Appendix (Section 7.3).

6 Discussion

This article elucidates the usefulness of local influence in model-based cluster analysis.

Local influence quantifies the impact of observations on the analysis. This can, for instance, be done by introducing case-weights in the log-likelihood, such that the contribution of an individual is slightly perturbed. Focus can be put on the effect of individual i only, by choosing the vector of case-weights to be the zero vector with one value of 1 in the ith position. The total local influence is then defined as the normal curvature of the likelihood displacement in the direction of the ith individual.

In this paper, we demonstrated the usefulness of local-influence diagnostics when clustering longitudinal profiles by means of a finite-mixture model, with a priori given number of components. The total local influence measures an individual's influence on the vector of all parameters in the model. Generally, this parameter vector contains (1) a number of fixed-effect parameters to describe

the average evolution of each component in the mixture, (2) random-effect parameters reflecting heterogeneity in the population, and (3) the mixture probabilities. The influence on a subset of the vector of all parameters – for example the influence on the average profile of a specific cluster, or on the mixture probabilities – can also be obtained.

When interest is not only in the stability of the parameters describing the components in the population, but also in the stability of an individual's classification the influence on the posterior probabilities is to be investigated. Local influence is an elegant approach for this. The stability of the posterior probabilities of individual j, can easily be inspected by re-parameterizing the log-likelihood in terms the fixed effects, random effects and the posterior probability of individual j. For the two-component mixtures carried out in this article, the $i \times j$ influence measures were displayed in a heatmap. Local-influence diagnostics were obtained for **two** real-life datasets subjected to a finite mixture model. For the EEG data, the results were compared with an outlier detection procedure for finite-mixture models and a method quantifying the impact of individual data points on the cluster partition when the correct classification is available. Local-influence diagnostics highlighted influential observations, that were not revealed by the traditional case-deletion methods.

7 Appendices : Derivatives

7.1 Derivates of the log-likelihood for a finite-mixture model

This section presents the first and second-order derivatives of the log-likelihood for a finite-mixture model with K components. Let $f_i(\boldsymbol{y}_i|\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k f_{ik}(\boldsymbol{y}_i|\boldsymbol{\Psi})$, with $\boldsymbol{\theta}, \boldsymbol{\pi}$, and $\boldsymbol{\Psi}$ as defined in (2). Individual i (i= 1 ..., N) contributes $l_i(\boldsymbol{\theta}) = \log\left[\sum_{k=1}^{K} \pi_k f_{ik}(\boldsymbol{y}_i|\boldsymbol{\Psi})\right]$ to the log-likelihood. First and second derivatives with respect to the components of $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)'$ and $\boldsymbol{\Psi} = (\Psi_1, \ldots, \Psi_P)'$ are ($l, m = 1, \ldots, K$; $p, q = 1 \ldots, P$):

$$
\begin{aligned}
\frac{\partial l_i}{\partial \pi_l} &= \frac{1}{f_i(\boldsymbol{y}_i)}\left[f_{il}(\boldsymbol{y}_i) - f_{iK}(\boldsymbol{y}_i)\right], \\
\frac{\partial l_i}{\partial \Psi_p} &= \frac{1}{f_i(\boldsymbol{y}_i)}\left[\sum_{k=1}^{K}\pi_k \frac{\partial f_{ik}(\boldsymbol{y}_i)}{\partial \Psi_p}\right],
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial^2 l_i}{\partial \pi_m \partial \pi_l} &= \frac{-1}{f_i^2(\boldsymbol{y}_i)}\left[(f_{im}(\boldsymbol{y}_i) - f_{iK}(\boldsymbol{y}_i))(f_{il}(\boldsymbol{y}_i) - f_{iK}(\boldsymbol{y}_i))\right], \\
\frac{\partial^2 l_i}{\partial \Psi_p \partial \Psi_q} &= \frac{-1}{f_i^2(\boldsymbol{y}_i)}\left[\sum_{k=1}^{K}\pi_k \frac{\partial f_{ik}(\boldsymbol{y}_i)}{\partial \Psi_p}\right]\left[\sum_{k=1}^{K}\pi_k \frac{\partial f_{ik}(\boldsymbol{y}_i)}{\partial \Psi_q}\right] + \frac{1}{f_i(\boldsymbol{y}_i)}\left[\sum_{k=1}^{K}\pi_k \frac{\partial^2 f_{ik}(\boldsymbol{y}_i)}{\partial \Psi_p \partial \Psi_q}\right], \\
\frac{\partial^2 l_i}{\partial \pi_l \partial \Psi_p} &= \frac{-1}{f_i^2(\boldsymbol{y}_i)}\left[\sum_{k=1}^{K}\pi_k \frac{\partial f_{ik}(\boldsymbol{y}_i)}{\partial \Psi_p}\right][f_{il} - f_{iK}] + \frac{1}{f_i(\boldsymbol{y}_i)}\left[\frac{\partial f_{il}(\boldsymbol{y}_i)}{\partial \Psi_p} - \frac{\partial f_{iK}(\boldsymbol{y}_i)}{\partial \Psi_p}\right],
\end{aligned}
$$

7.2 Likelihood for a finite-mixture model as a function of the posterior probabilities

This section sketches the calculation of the likelihood for a finite-mixture model with K components as a function of the posterior probabilities.

We will use the following shortened notation $f_{ik} = f_{ik}(\boldsymbol{y}_i|\boldsymbol{\Psi})$ $(k = 1, \ldots, K)$ and $f_i = f_i(\boldsymbol{y}_i|\boldsymbol{\Psi})$. Let $\boldsymbol{\pi}_i = (\pi_{i1}, \ldots, \pi_{iK})'$ be the vector of posterior probabilities for observation i. The mixture probabilities $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)'$ and the contribution $f_j(\boldsymbol{y}_j)$ of observation j to the likelihood for the finite-mixture model are obtained as follows.

First express the mixture probabilities as a function of the posterior probabilities and the component specific density functions. Given the relation between the posterior probabilities and the mixture probabilities, $\pi_{ik} = \frac{\pi_k f_{ik}}{\sum_{l=1}^{K-1}\pi_l f_{il} + (1-\sum_{l=1}^{K-1}\pi_l) f_{iK}}$, one obtains that $\pi_k = \frac{\pi_{ik}\left[\sum_{l=1, l\neq k}^{K-1}\pi_l(f_{il}-f_{iK}) + f_{iK}\right]}{f_{ik} - \pi_{ik}(f_{ik}-f_{iK})}$ (1). This expression gives the following equations, $\frac{\pi_k \gamma_{ik}}{\pi_{ik}} - \pi_s(f_{is} - f_{iK}) = \sum_{l\neq k,s}^{K-1}\pi_l(f_{il} - f_{iK}) + f_{iK}$ and $\frac{\pi_s \gamma_{is}}{\pi_{is}} - \pi_k(f_{ik} - f_{iK}) = \sum_{l\neq k,s}^{K-1}\pi_l(f_{il} - f_{iK}) + f_{iK}$ with $\gamma_{ik} = f_{ik} - \pi_{ik}(f_{ik} - f_{iK})$. Equating them results in the following expression for the mixture probability, $\pi_s = \frac{\pi_k\left(\frac{\gamma_{ik}}{\pi_{ik}} + (f_{ik} - f_{iK})\right)}{\frac{\gamma_{is}}{\pi_{is}} + (f_{is} - f_{iK})}$, which

simplifies to $\pi_l = \frac{\pi_k \pi_{il} f_{ik}}{\pi_{ik} f_{il}}$. Substituting this expression for the mixture probability into (1) results in

$$\begin{aligned}
\pi_k &= \frac{\pi_{ik} f_{iK}}{f_{ik} - \sum_{l=1}^{K} \frac{f_{ik}}{f_{il}} (f_{il} - f_{iK})} \\
&= \frac{\pi_{ik} f_{iK} \prod_{l=1}^{K-1} f_{il}}{f_{ik} \left[\prod_{l=1}^{K-1} f_{il} - \sum_{l=1}^{K-1} \left(\pi_{il} (f_{il} - f_{iK}) \prod_{m \neq l}^{K-1} f_{im} \right) \right]} \\
&= \frac{\pi_{ik} \prod_{l \neq k}^{K} f_{il}}{\left[\prod_{l=1}^{K-1} f_{il} - \sum_{l=1}^{K-1} \left(\pi_{il} (f_{il} - f_{iK}) \prod_{m \neq l}^{K-1} f_{im} \right) \right]}.
\end{aligned}$$

The kth mixture probability is thus specified as a function of the posterior probabilities and component specific density functions of observation i.

The density function of observation j, $f_j = \sum_{k=1}^{K-1} \pi_k f_{jk} + (1 - \sum_{k=1}^{K-1} \pi_k) f_{jK}$, can now be attained as a function of the posterior probabilities of observation i. Given that $1 - \sum_{k=1}^{K-1} \pi_k$ simplifies to $\frac{\prod_{k=1}^{K-1} f_{il} \pi_{iK}}{\mathsf{D}}$, with $\mathsf{D} = \prod_{l=1}^{K-1} f_{il} - \sum_{l=1}^{K-1} \left(\pi_{il} (f_{il} - f_{iK}) \prod_{m \neq l}^{K-1} f_{im} \right)$ one can show that

$$f_j = \frac{\sum_{k=1}^{K} (\pi_{ik} \prod_{l \neq k}^{K} f_{il}) f_{jk}}{\prod_{l=1}^{K-1} f_{il} - \sum_{l=1}^{K-1} \left(\pi_{il} (f_{il} - f_{iK}) \prod_{m \neq l}^{K-1} f_{im} \right)}.$$

The local influence analysis presented in Section 5.2 of the manuscript considered a two-component mixture model. For K=2, the expression for the prior probability to belong to cluster 1 and the density function are given by:

$$\begin{aligned}
\pi_1 &= \frac{\pi_{i1} f_{2i}}{f_{i1} - \pi_{i1} (f_{1i} - f_{2i})} \\
f_j &= \frac{\pi_{i1} f_{i2} f_{j1} + (1 - \pi_{i1}) f_{i1} f_{j2}}{f_{i1} - \pi_{i1} (f_{1i} - f_{2i})}
\end{aligned}$$

Thus, the contribution of observation j to the likelihood $l_j = \log(f_j)$ is now expressed in terms of the parameters $\boldsymbol{\pi}_{i1}$ and $\boldsymbol{\Psi} = (\Psi_1, \ldots, \Psi_P)'$ $(p, q : 1 \ldots, P)$. The local-influence diagnostics need the first and second-order derivatives with respect to these parameters. These derivatives can be obtained by means of the chain rule or direct derivation of the likelihood. Using the following

notation, $A = \pi_{i1}(f_{i1} - f_{j2}) - f_{i1}f_{i2}$ and $B = \pi_{i1}(f_{i1} - f_{i2} - f_{i1})$ the following expressions are obtained for the first and second derivatives of l_j:

$$\begin{aligned}
\frac{\partial l_j}{\partial \pi_{i1}} &= \frac{\frac{\partial A}{\partial \pi_{i1}}B - A\frac{\partial B}{\partial \pi_{i1}}}{AB} \\
\frac{\partial l_j}{\partial \Psi_p} &= \frac{\frac{\partial A}{\partial \Psi_p}B - A\frac{\partial B}{\partial \Psi_p}}{AB}
\end{aligned}$$

$$\begin{aligned}
\frac{\partial^2 l_j}{\partial^2 \pi_{i1}} &= \frac{(A\frac{\partial B}{\partial \pi_{i1}})^2 - (B\frac{\partial A}{\partial \pi_{i1}})^2}{(AB)^2} \\
\frac{\partial^2 l_j}{\partial \Psi_q \partial \pi_{i1}} &= \frac{\left[\frac{\partial^2 A}{\partial \Psi_p \partial \pi_{i1}}B + \frac{\partial A}{\partial \Psi_p}\frac{\partial B}{\partial \pi_{i1}} - \frac{\partial A}{\partial \pi_{i1}}\frac{\partial B}{\partial \Psi_p} - A\frac{\partial^2 B}{\partial \Psi_p \partial \pi_{i1}}\right](AB) - (\frac{\partial A}{\partial \Psi_p}B - A\frac{\partial B}{\partial \Psi_p})(\frac{\partial A}{\partial \pi_{i1}}B + A\frac{\partial B}{\partial \pi_{i1}})}{(AB)^2} \\
\frac{\partial^2 l_j}{\partial \Psi_q \partial \Psi_p} &= \frac{\left[\frac{\partial^2 A}{\partial \Psi_p \partial \Psi_q}B + \frac{\partial A}{\partial \Psi_p}\frac{\partial B}{\partial \Psi_q} - \frac{\partial A}{\partial \Psi_q}\frac{\partial B}{\partial \Psi_p} - A\frac{\partial^2 B}{\partial \Psi_p \partial \Psi_q}\right](AB) - (\frac{\partial A}{\partial \Psi_p}B - A\frac{\partial B}{\partial \Psi_p})(\frac{\partial A}{\partial \Psi_q}B + A\frac{\partial B}{\partial \Psi_q})}{(AB)^2}
\end{aligned}$$

with

$$\begin{aligned}
\frac{\partial A}{\partial \pi_{i1}} &= f_{i1}f_{j2} - f_{j1}f_{i2} \\
\frac{\partial B}{\partial \pi_{i1}} &= f_{i1} - f_{i2} \\
\frac{\partial A}{\partial \Psi_p} &= \pi_{i1}(\frac{\partial f_{i1}}{\partial \Psi_p}f_{j2} + f_{i1}\frac{\partial f_{j2}}{\partial \Psi_p} - \frac{\partial f_{i2}}{\partial \Psi_p}f_{j1} - f_{i2}\frac{\partial f_{j1}}{\partial \Psi_p}) - \frac{\partial f_{i1}}{\partial \Psi_p}f_{i2} - f_{i1}\frac{\partial f_{i2}}{\partial \Psi_p} \\
\frac{\partial B}{\partial \Psi_p} &= \pi_{i1}(\frac{\partial f_{i1}}{\partial \Psi_p} - \frac{\partial f_{i2}}{\partial \Psi_p}) - \frac{\partial f_{i1}}{\partial \Psi_p} \\
\frac{\partial^2 A}{\partial \Psi_p \partial \pi_{i1}} &= \frac{\partial f_{i1}}{\partial \Psi_p}f_{j2} + f_{i1}\frac{\partial f_{j2}}{\partial \Psi_p} - \frac{\partial f_{i2}}{\partial \Psi_p}f_{j1} - f_{i2}\frac{\partial f_{j1}}{\partial \Psi_p} \\
\frac{\partial^2 A}{\partial \Psi_p \partial \Psi_q} &= \pi_{i1}\left[\frac{\partial^2 f_{i1}}{\partial \Psi_p \partial \Psi_q}f_{j2} + \frac{\partial f_{i1}}{\partial \Psi_p}\frac{\partial f_{j2}}{\partial \Psi_q} + \frac{\partial f_{i1}}{\partial \Psi_q}\frac{\partial f_{j2}}{\partial \Psi_p} + f_{i1}\frac{\partial^2 f_{j2}}{\partial \Psi_p \partial \Psi_q} - \frac{\partial^2 f_{i2}}{\partial \Psi_p \partial \Psi_q}f_{j1} - \frac{\partial f_{i2}}{\partial \Psi_p}\frac{\partial f_{j1}}{\partial \Psi_q} - +\frac{\partial f_{i2}}{\partial \Psi_p}\frac{\partial f_{j1}}{\partial \Psi_q} - f_{i2}\frac{\partial^2 f_{j1}}{\partial \Psi_p \partial \Psi_q}\right] \\
&\quad -\frac{\partial f_{i1}}{\partial \Psi_p}\frac{\partial f_{i2}}{\partial \Psi_q} - \frac{\partial f_{i1}}{\partial \Psi_q}\frac{\partial f_{i2}}{\partial \Psi_p} - f_{i1}\frac{\partial^2 f_{i2}}{\partial \Psi_p \partial \Psi_q} \\
\frac{\partial^2 B}{\partial \Psi_p \partial \Psi_q} &= \pi_{i1}(\frac{\partial^2 f_{i1}}{\partial \Psi_p \partial \Psi_q} - \frac{\partial^2 f_{i2}}{\partial \Psi_p \partial \Psi_q}) - \frac{\partial^2 f_{i1}}{\partial \Psi_p \partial \Psi_q} \\
\frac{\partial^2 B}{\partial \Psi_p \partial \pi_{i1}} &= \frac{\partial f_{i1}}{\partial \Psi_p} - \frac{\partial f_{i2}}{\partial \Psi_p}
\end{aligned}$$

7.3 Three component mixture model for the EEG Data

This section presents the local-influence diagnostics for a three-component mixture model applied to the EEG data. When applying the cluster algorithm, information about the drug a rat was given, was

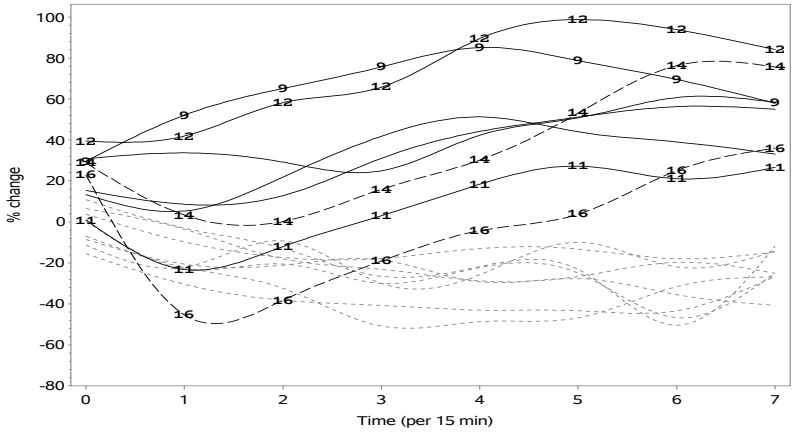

Figure 7: *Smoothed observed %change for γ_2 profiles (full lines: PCP, dotted lines: Donepezil) - EEG data. The origin of the time axis is at the first measurement after adminstration of the drug.*

not taken into account. The log-likelihood (BIC) for the three-component model was -516.2 (1068.5). The model hypothesizing a mixture of three components is outperforming the two-component model; but it results in a small cluster of size two. The 8 rats on Donepezil (ids 1–8) are classified together in cluster 2. The 8 rats on PCP are distributed over two clusters; cluster 3 (ids 14 and 16) and cluster 1 (ids 9–13 and 15). So the three-component mixture splits the PCP cluster revealed in the two-component solution into two clusters. The rat turning out to be most influential (i.e., rat 16), according to local-influence analysis for the two component solution, is a member of the cluster of size two in the three component solution.

The results of a local-influence analysis for the three-component model are displayed in Figure 8.

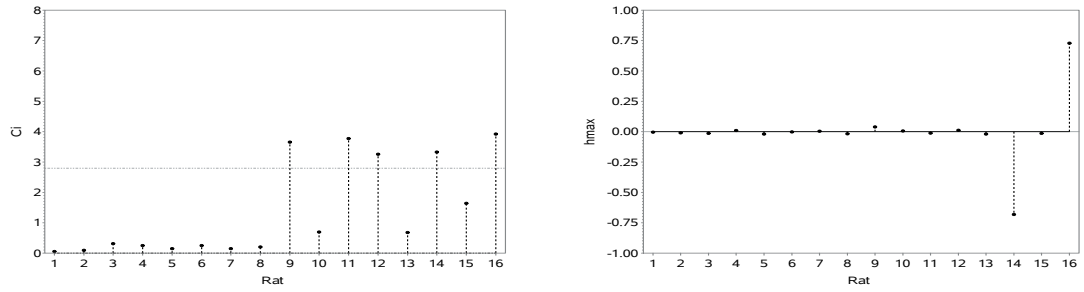

Figure 8: *Total local influence and direction of maximal curvature versus rat identification numbers - EEG data. The horizontal line in the total local influence graph represents the cut-off value for C_i.*

Three rats (ids 9, 11, and 12) of cluster 2 and both rats (ids 14 and 16) of cluster 3 are locally

influential, based on their C_i value. The observed profiles of these rats are highlighted in Figure 7. It is not a surprise that both rats of cluster 3 are highly influential. The influence that an individual rat can effectuate on the average evolution of the cluster it belongs too, becomes more substantial for smaller clusters.

The influence of rats 14 and 16 is visible in the set of fixed parameters characterizing the average evolution of the cluster they belong too (cluster 3), their influence on the average profile of the other clusters is negligible. Rats 14 and 16 are also influential for the mixture probabilities.

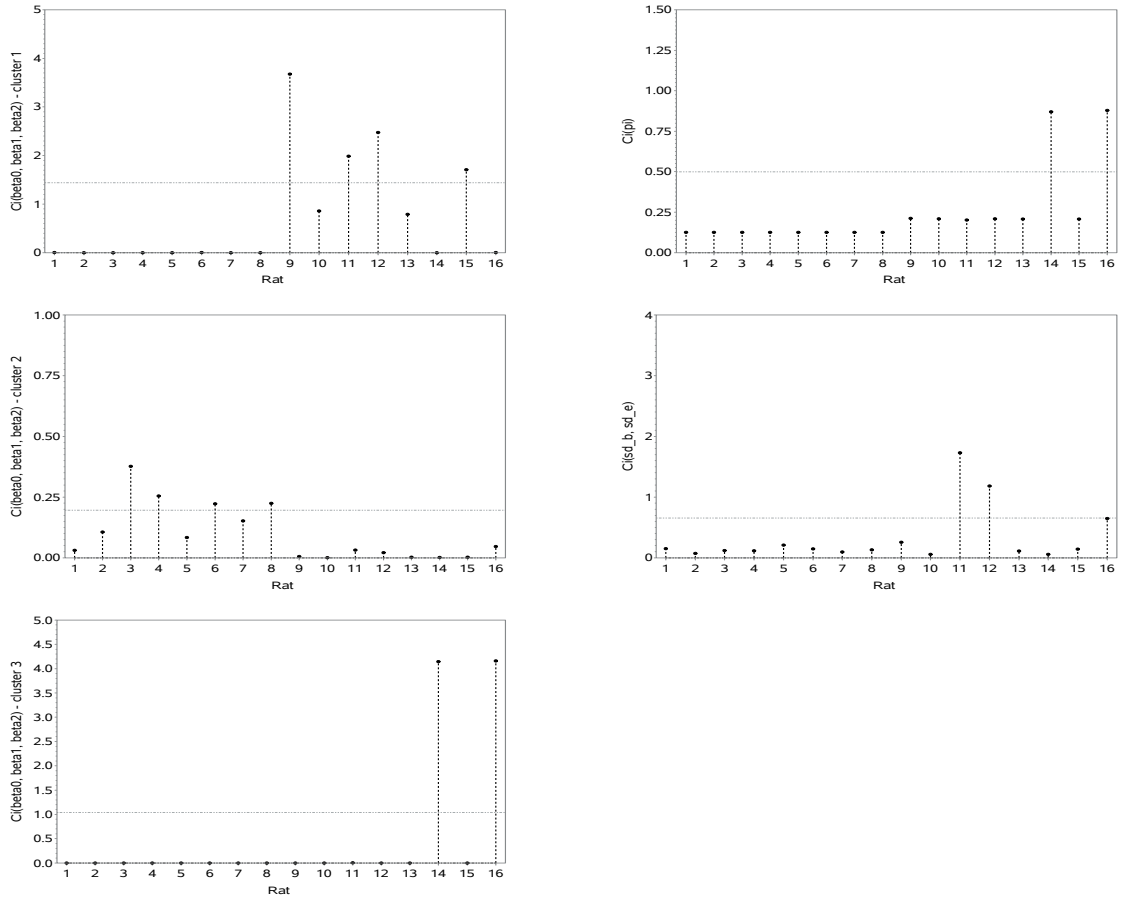

Figure 9: *Plot of local-influence diagnostics for the cluster specific average profiles and the random components - EEG data. The horizontal lines represent the cut-off values for the displayed influences.*

7.4 Simulation Study

To investigate the performance of the local-influence diagnostics in the mixture context a simulation study was set up. The simulation study is inspired on the finite-mixture model obtained for the EEG data, and by no means attempts to be comprehensive.

Data were generated for different settings, each consisting of a mixture with two components and with equal class probability. Repeated measurements at eight time points were generated assuming quadratic individual profiles. Residual variances were normally distributed, homoscedastic, uncorrelated and mixture component invariant. Also the residual variances of the random intercept and random slope were specified to be normally distributed, invariant across the mixture components, and with a zero covariance between them. The values specified for the fixed parameters, the variances of the random intercept and slope, and the residual variance are presented in Table 1. The values are inspired on the average profiles seen in the EEG data, for a two-component mixture model. The sample size per component, for each simulated data set, was fixed to be either 10, 20 or 30. For each setting, 1000 data sets were generated.

In order to evaluate the performance of a local-influence analyses, one individual profile deviating in terms of it's intercept, slope and residual error was subjoined to the generated data sets. The deviation was specified as a deviation from the average profile of the first component (Table 2). Two degrees of deviation were considered, referred to as scheme 1 and scheme 2 with the degree of deviation being largest for scheme 2. The effect of the number of deviating observations was only investigated for the setting with a deviating intercept. Hereto, five deviating observations were added to the generated data.

The results of a local-influence analysis, obtained when applying a two-component model to the simulated data, are presented in Tables 3 and 4.

Each observation's C_i is compared to the cutoff equal to $2 \times \sum C_i/N$. The proportion of deviating observations that was identified as influential and the proportion of observations generated according to component 1 or 2 that are identified as not influential are given in Table 3 and 4. The influence was also studied on subsets of the parameters of the model, i.e. on the cluster specific average

Table 1: *Setting for the Simulation Study: Two-component Mixture Model: Fixed and Random-effects Factors.*

	Component 1	Component 2
Intercept	13	-4.1
Slope	40.8	5.9
Quadratic term	-80.9	-60.6
Random intercept variability	15	15
Random error variability	11	11

Table 2: *Setting for the Simulation Study: Profiles for the Deviating Observation. The deviation is specified as a deviation from the average profile of the first component (Table 1). Two degrees of deviation were considered, referred to as scheme 1 and scheme 2 with the degree of deviation being largest for scheme 2.*

	Deviating in terms of:					
	Intercept		Slope		Random Error	
	Scheme 1	Scheme 2	Scheme 1	Scheme 2	Scheme 1	Scheme 2
Intercept	43	58	13	13	13	13
Slope	40.8	40.8	81.6	122.4	40.8	40.8
Quadratic term	-80.9	-80.9	-80.9	-80.9	-80.9	-80.9
Random intercept variability	15	15	15	15	15	15
Random error variability	11	11	11	11	22	33

profiles, the random component, residual variability and the mixture probability.

The probability that the deviating observation is identified as being influential increases with the extent in which the case deviates from cluster 1 and with the cluster size. Comparing the results for cluster sizes equal to 10 and 30, for scheme 1 for the intercept, shows that the probability to identify the deviating subject as an influential case increases from 0.56 to 0.69. For a cluster size of 20, the probability of identifying an influential case increases from 0.67 for scheme 1 to a probability of 1 for scheme 2. It is seen that the influence of a subject becomes smaller when more observations are deviating. When, for example, looking at scheme 1 and cluster sizes equal to 20, the probability that the deviating case will be identified as influential is 0.67 if there is only one deviating case and this

Table 3: Simulation Study: Probability that an observation is identified as influential, for the cluster components and for the deviating observation(s) in terms of the intercept. The deviating subjects are specified as a deviation from the average profile of the first component (Table 1). Two degrees of deviation were considered, referred to as scheme 1 and scheme 2 with the degree of deviation being largest for scheme 2 (Table 2).

Deviating Scheme	Cluster Size	**Deviating intercept: 1 observation**		**Deviating intercept: 5 observations**	
		observations clusters 1+2	one subject deviating	observations cluster 1+2	five subjects deviating
1	10	0.10	0.56	0.12	0.08
	20	0.10	0.67	0.10	0.21
	30	0.10	0.69	0.10	0.32
2	10	0.06	0.96	0.10	0.19
	20	0.07	1.00	0.07	0.73
	30	0.08	1.00	0.06	0.92

probability decreases to 0.21 when there are five deviating observations.

About 10% of the observations generated according to cluster 1 or cluster 2, are also identified as being influential. This proportion stays more or less constant as a function of the cluster size and only slightly decreases with increasing distance between cluster 1 and the deviating observation.

When the deviating observation differs in terms of the intercept, an influence can be seen on the average profile of cluster 1 (most strongly in the intercept), on the variance of the random intercept and on the residual variance. A deviation in terms of the slope is reflected in an influence on the average profile of cluster 1 and on the residual variance. When the extra observation deviates in terms of its residual error, this is seen in the influence measures for the average profile of cluster 1 and on the residual variance (data not shown).

Table 4 contains the results when the extra observation deviates in terms of the slope or the residual error of the profile. Only the situation where one deviating observation was added is considered.

Although this simulation study should not be over-generalized, the results indicate that observations that were generated as deviating from the average profile of one of the cluster components, are more likely to be identified as influential, as compared to observations generated to belong to the cluster components.

Table 4: Simulation Study: Probability that an observation is identified as influential, for the cluster components and for the deviating observation(s) in terms of the slope or residual error. The deviating subjects are specified as a deviation from the average profile of the first component (Table 1). Two degrees of deviation were considered, referred to as scheme 1 and scheme 2 with the degree of deviation being largest for scheme 2 (Table 2).

Deviating Scheme	Cluster Size	**Deviating Slope**		**Deviating random error**	
		observations clusters 1+2	one subjects deviating	observiations cluster 1+2	one subjects deviating
1	10	0.12	0.13	0.10	0.64
	20	0.11	0.14	0.10	0.68
	30	0.10	0.15	0.10	0.68
2	10	0.11	0.23	0.07	0.91
	20	0.11	0.25	0.07	0.93
	30	0.10	0.25	0.07	0.96

Acknowledgements

Financial support from the IAP research Network P7/06 of the Belgian government (Belgian Science Policy) is gratefully acknowledged. We are grateful to Janssen Pharmaceutica for kind permission to use the EEG data.

References

[1] Johnson RA and Wichern DW. Applied Multivariate Statistical Analysis. *Pearson Prentice Hall* 2007.

[2] Jolliffe IT, Jones B and Morgan BJT. Identifying Influential Observations in Hierarchical Cluster Analysis. *J*ournal of Applied Statistics 1995; 22: 61–80.

[3] Kim S, Kwon S and Cook D. Interactive visualization of hierarchical clusters using MDS and MST. *M*etrika 2000; 51: 39–51.

[4] Cheng R and Milligan GW. Measuring the influence of individual data points in a cluster analysis. *J*ournal of Classification 1996; 13: 315–335.

[5] Cerioli A. A New Method for Detecting Influential Observations in Nonhierarchical Cluster Analysis. Advances in Data Science and Classification. In: Rizzi A , Vichi M, and Bock H (eds) *S*tudies in Classification, Data Analysis, and Knowledge Organization. 1st ed. Heidelberg: Springer-Verlag, 1998, pp. 15-20.

[6] Cuesta-Albertos JA, Gordaliza A and Matran C. Trimmed kMeans: an Attempt to Robustify Quantizers. *T*he Annals of Statistics 1997; 25: 553–576.

[7] McLachlan G and Peel D. *F*inite mixture models . 1st ed. New York: Wiley, 2000.

[8] Wang S, Woodward WA, Gray HL, Wiechecki S and Sain SR. A new test for outlier detection from a multivariate mixture distribution. *J*ournal of Computational and Graphical Statistics 1971; 6: 285–299.

[9] Cook R. Assessment of Local Influence. *J*ournal of the Royal Statistical Society. Series B 1986; 48: 133–169.

[10] York T, Vargas-Irwin C, Anderson W and van den Oord E. Asthma pharmacogenetic study using finite mixture models to handle drug-response heterogeneity. *P*harmacogenomics 2009; 10: 753–767.

[11] Wang Y and Lei T. A new look at finite mixture models in medical image analysis. *S*peech, Image Processing and Neural Networks Proceedings of ICSIPNN 1994. International Conference on Speech, Image Processing and Neural Networks .

[12] Nagin DS. Analyzing developmental trajectories: a semiparametric, group-bases approach. *P*sychological Methods 1999; 4: 139–157.

[13] Muthén B. 2004 Latent variable analysis: Growth mixture modeling and related techniques for longitudinal data. *I*n D. Kaplan (ed.), Handbook of quantitative methodology for the social sciences. Newbury Park, CA: Sage Publications.

[14] Verbeke G and Lesaffre E. A linear mixed effects model with heterogeneity in the random effects population. *J*. Am. Stat. Assoc. 1996; 91: 217–221.

[15] Spiessens B, Verbeke G and Komàrek A. (2002). *A* SAS-macro for the classification of longitudinal profiles using mixtures of normal distributions in nonlinear and generalized linear models. http://www.med.kuleuven.ac.be/biostat/research/software.htm.

[16] Dempster AP, Laird NM and Rubin DB. Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J*ournal of the Royal Statistical Society, Series B 1977; 39: 1–38.

[17] Sain SR, Gray HL, Woodward WA and Fisk MD. Outlier Detection from a Mixture Distribution When Training Data Are Unlabeled. *B*ull. Seismol. Soc. Am. 1999; 89: 294–304.

[18] Efron B and Tibshirani RJ. . *A*n Introduction to the Bootstrap. 1st ed. New York: Chapman & Hall, 1993.

[19] Rand WM. Objective criteria for the evaluation of clustering methods. *J.* Am. Stat. Assoc. 1971; 66: 846–850.

[20] Goodman LA and Kruskal WH. Measures of association for cross classifications. *J.* Am. Stat. Assoc 1954; 49: 732–764.

[21] Cook R and Weisberg S. *R*esiduals and influence in regression. 1st ed. New York -London: Chapman & Hall, 1982.

[22] Lesaffre, E. and Verbeke, G. Local influence in linear mixed models. *B*iometrics 1998; 54: 570–582.

[23] Mun J and Lindstrom MJ. Diagnostics for repeated measurements in linear mixed effects models. *S*tatistics in Medicine 2013; 32: 1361–1375.

[24] Verbeke G, Molenberghs G, Thijs H, Lesaffre E and Kenward MG. Sensitivity analysis for nonrandom dropout: A local influence approach. *B*iometrics 2001; 57: 7–14.

[25] Jansen I, Molenberghs G, Aerts M, Thijs H and Van Steen K. A Local influence approach applied to binary data from a psychiatric study. *B*iometrics 2003; 59: 410–419.

[26] Verbeke G and Molenberghs G. *L*inear Mixed Models for Longitudinal Data. New York: Springer, 2000.

[27] Molenberghs G and Verbeke G. *M*odels for Discrete Longitudinal Data. 1st ed. New York: Springer, 2005.

[28] Ouwens MJNM, Tan FES and Berger MPF. Local influence to detect influential data structures for generalized linear mixed models. *B*iometrics 2001; 57: 1166–1172.

[29] Beckman RJ, Nachtsheim CJ and Cook RD. Diagnostics for mixed model analysis of variance. *T*echnometrics 1987; 29: 413-426.

[30] Verbeke G and Lesaffre E. The linear mixed model. A critical investigation in the context of longitudinal data, In: Gregoire T (ed), *P*roceedings of the Nantucket conference on Modelling Longitudinal and Spatially Correlated Data: Methods, Applications, and Future Directions, Lecture Notes in Statistics. New York: Springer, 1997,pp. 89-99.

[31] Seber GAF. *M*ultivariate Observations. New York: Wiley, 1984.

[32] Draper NR and Smith H. *A*pplied Regression Analysis. 2nd ed. New York: John Wiley & Sons, 1981.

[33] Lindstrom MJ and Bates DM. Nonlinear Mixed Effects Models for Repeated Measures Data. *B*iometrics 1990; 46: 673–687.

[34] Pinheiro JC and Bates D. Approximations to the Log-likelihood Function in the Nonlinear Mixed-effects Model. *J*ournal of Computational and Graphical Statistics 1995; 4: 12–35.

[35] Hosmer D and Lemeshow S. *A*pplied Logistic Regression. 2nd ed. New York: John Wiley & Sons, 2000.

[36] Agresti A. *C*ategorical Data Analysis. 2nd ed. New York: John Wiley & Sons, 2002.