Clustering of Longitudinal Data Using Mixture of Extended Linear Mixed-Effect Models

Changjiang Xu* Celia M. T. Greenwood† Vicky Tagalakis‡ Antonio Ciampi§

Abstract

We model longitudinal data from a heterogeneous population as samples from a mixture of extended linear mixed-effects models, and develop an expectation-maximization (EM) algorithm based on Monte Carlo (MC) sampling, called EMMC, to estimate the model parameters. The algorithm EMMC is implemented separately using linear mixed-effect model for each cluster, and thus converges much faster than the standard EM algorithm. We present an evaluation of the approach through simulations. We also applied our algorithm to clustering of International Normalized Ratio (INR) trajectories following warfarin initiation. Four clusters of the INR trajectories were determined using model selection criterion AIC. In contrast, when using BIC for model selection, the cluster consisting of the worst cases, where INR cannot be controlled, was missed.

Key Words: Longitudinal data, mixed-effects, mixture model, EM algorithm, clustering

1. Introdution

Longitudinal study involves repeated measurements on each subject or individual over long periods of time. Such repeated measures data are usually heteroscedastic and correlated within subjects, and measurement times may be unequally spaced within an individual and varying across individuals. Model-based methods or finite mixture models (Titterington et al., 1985; McLachlan and Basford, 1988; McLachlan and Peel, 2000; Fraley and Raftery, 2002) have been proposed for clustering such longitudinal data, for example, see Pauler and Laird (2000), De la Cruz-Mesia et al. (2008), and McNicholas and Murphy (2010).

For model-based clustering, we need to specify a family of statistical models for the components of the mixture models. Two families have previously been proposed. Banfield and Raftery (1993) proposed a general framework for the family of multivariate normal mixtures by parameterizing covariance matrices through eigenvalue decomposition. McNicholas and Murphy (2008) developed a parsimonious Gaussian mixture model (PGMM) family based on a mixture of factor analyzers (McLachlan et al., 2003). The PGMM family of models is well suited to the analysis of high-dimensional data because the number of covariance parameters is linear in the dimensionality of the data under consideration. These two families of models are implemented in the R (R Development Core Team, 2013) software packages mclust (Fraley and Raftery, 1999, 2007) and pgmm (McNicholas et al., 2010), respectively. McNicholas and Murphy (2010) also proposed a Gaussian mixture model with a modified Cholesky decomposed covariance structure (Pourahmadi, 1999) for clustering of longitudinal data. These families of models take account of the heteroscedasticity and correlation within subjects, but ignore random variations from multiple sources or between subjects.

More recently, the family of linear mixed-effects (LME) models has been used in the model-based clustering context (Celeux et al., 2005; De la Cruz-Mesia et al., 2008), to

*McGill University, and Lady Davis Institute for Medical Research, Jewish General Hospital, Canada. Currently Donnelly Centre, University of Toronto, Canada.

†McGill University, and Lady Davis Institute for Medical Research, Jewish General Hospital, Canada.

‡McGill University, and Lady Davis Institute for Medical Research, Jewish General Hospital, Canada.

§Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Canada

provide a more flexible and powerful tool for the analysis of random variations (Pinheiro and Bates, 2000). Although the LME models include random variations, they neglect heteroscedasticity and correlation within subjects by assuming that the correlation matrix of within subject errors is the identity matrix. Extended linear mixed-effects models relax this assumption and allow heteroscedastic and correlated within-subject errors (Pinheiro and Bates, 2000). Hence, the extended LME model offers a viable option for modeling longitudinal data.

In this paper, we consider the family of extended LME models for clustering of longitudinal data. Following previous work (Ciampi et al., 2012), we model the longitudinal data as samples from a mixture of extended LME models. The EM algorithm (Dempster et al., 1977; Titterington et al., 1985) is popularly used to estimate mixture model parameters by maximizing a complete data likelihood. However, the standard EM algorithm is highly nonlinear and can be slow to converge. We propose an EM algorithm with Monte Carlo sampling, called EMMC, that is implemented separately using LME model for each sampling cluster. The EMMC algorithm converges quickly.

The rest of this paper is organized as follows. Section 2 introduces a mixture of extended LME models. The EM algorithms for parameter estimation are presented in Section 3. Sections 4 is devoted to evaluation of the approach through limited simulations. In Section 5, we analyzed clinical data set on initiation of warfarin.

2. A mixture of extended linear mixed-effects models

Consider a data set containing M subjects or individuals assumed to come from K clusters. For subject i in cluster k we model the observed response as a linear mixed model

$$y_i = X_i\beta_k + Z_i b_{ik} + \epsilon_{ik}, \tag{1}$$

where $y_i = (y_{i1}, \ldots, y_{i,n_i})'$, y_{ij} is the response variable of subject i at measurement time point t_{ij}, $j = 1, \ldots, n_i$, and

- β_k is a vector of fixed effects, and X_i is a corresponding design matrix of $n_i \times p$;
- $b_{ik} \sim N(0, \Psi_k)$ corresponding to random effects, and Z_i is a design matrix of $n_i \times q$;
- $\epsilon_{ik} \sim N(0, \Lambda_{ik})$ representing modeling errors, where $\Lambda_{ik} > 0$, positive definite.

The random effects, b, and modeling errors, ϵ, are assumed to be independent. The covariance matrix of y_i, can therefore be written

$$\Sigma_{ik} = Z_i \Psi_k Z_i' + \Lambda_{ik}.$$

Let $\mathcal{C}_k$ denote the set of subjects in the cluster k, and $\alpha_k = \Pr\{\mathcal{C}_k\}$ be the mixture proportion for cluster k, satisfying $\sum_{k=1}^{K} \alpha_k = 1$. Let $\mu_{ik} = X_i\beta_k$. Then the set of observations for subject i follow a multivariate Gaussian mixture distribution with probability density function

$$f(y|\theta) = \sum_{k=1}^{K} \alpha_k \varphi(y|\mu_{ik}, \Sigma_{ik}), \tag{2}$$

where

$$\varphi(y|\mu_{ik}, \Sigma_{ik}) = \frac{1}{\sqrt{(2\pi)^{n_i}|\Sigma_{ik}|}} \exp\{-\frac{1}{2}(y - \mu_{ik})'\Sigma_{ik}^{-1}(y - \mu_{ik})\} \tag{3}$$

is the density of the multivariate normal distribution, and $\theta = \{\alpha, \beta, \Psi, \Lambda\}$ contain all unknown parameters. The covariance matrix of the modeling errors may be redefined as

$\Lambda_{ik} = \sigma_k^2 V_i(\phi_k)$, depending on some additional parameters. Then $\theta = \{\alpha, \beta, \Psi, \phi, \sigma^2\}$. In this article, we use parameter notation without subscripts to represent the set of all corresponding parameters.

3. Model estimation

The likelihood function for the model (1) is

$$l(\theta|y) = \sum_{i=1}^{M} \log f(y_i|\theta) = \sum_{i=1}^{M} \log(\sum_{k=1}^{K} \alpha_k \varphi(y_i|\mu_{ik}, \Sigma_{ik})).$$

Estimating the model parameters by directly maximizing the likelihood function usually is not possible due to its complexity. Instead, the EM algorithm can be used for estimating parameters of statistical models with missing or latent variables. Let $\delta_{ik} = 1\{i \in \mathcal{C}_k\}$ be an indicator function of subject i being in cluster k, and let $\delta_i = (\delta_{i1}, \ldots, \delta_{iK})'$. Both the random effects, b, and cluster indicators, δ, are unobserved, and can be called missing or latent variables. Nevertheless, the joint probability of Y_i, b_i, and δ_i, is

$$f(y_i, b_i, \delta_i|\theta) = \prod_{k=1}^{K} \{\alpha_k \varphi(y_i|X_i\beta_k + Z_i b_{ik}, \Lambda_{ik}) \varphi(b_{ik}|0, \Psi_k)\}^{\delta_{ik}}.$$

We therefore have a complete data log-likelihood function

$$l(\theta|y, b, \delta) = \sum_{i=1}^{M} \sum_{k=1}^{K} \delta_{ik} \{\log \alpha_k + \log \varphi(y_i|X_i\beta_k + Z_i b_{ik}, \Lambda_{ik}) + \log \varphi(b_{ik}|0, \Psi_k)\}. \quad (4)$$

Instead of maximizing $l(\theta|y)$ directly, we can use the EM algorithm to maximize the expectation of the complete data log-likelihood $E_{b,\delta}\{l(\theta|y, b, \delta)\}$. This is achieved by two alternating steps, the E or expectation step and M or maximization step, repeated over multiple iterations.

3.1 The standard EM algorithm (EM)

E-step

Let $\theta^{(s)}$ denote the value of the parameters after iteration s. Then the E-step at iteration $s+1$ involves the computation of a Q-function, $Q(\theta|\theta^{(s)}) = E_{b,\delta}\{l(\theta|y, b, \delta)|y, \theta^{(s)}\}$. Omitting the constant C in (4), we have

$$Q(\theta|\theta^{(s)}) = -\frac{1}{2} \sum_{i=1}^{M} \sum_{k=1}^{K} \tau_{ik} \{-2 \log \alpha_k + [\log |\Psi_k| + tr(\Psi_k^{-1} B_{ik})] + [\log |\Lambda_{ik}| + tr(\Lambda_{ik}^{-1} A_{ik})]\}, \quad (5)$$

where

$$\begin{aligned}
\tau_{ik} &= E\{\delta_{ik}|y, \theta^{(s)}\} = \Pr\{i \in \mathcal{C}_k|y_i, \theta^{(s)}\} = \frac{\alpha_k^{(s)} \varphi(y_i|X_i\beta_k^{(s)}, \Sigma_{ik}^{(s)})}{\sum_{j=1}^{K} \alpha_j^{(s)} \varphi(y_i|X_i\beta_j^{(s)}, \Sigma_{ij}^{(s)})}, \\
A_{ik} &= E\{e_{ik}e_{ik}'|y, \theta^{(s)}\} = (y_i - X_i\beta_k - Z_i\gamma_{ik})(y_i - X_i\beta_k - Z_i\gamma_{ik})' + Z_i\Gamma_{ik}Z_i' \triangleq A_{ik}(\beta_k), \\
B_{ik} &= E\{b_{ik}b_{ik}'|y, \theta^{(s)}\} = \gamma_{ik}\gamma_{ik}' + \Gamma_{ik},
\end{aligned}$$

and

$$\begin{aligned}\Sigma_{ik}^{(s)} &= Z_i\Psi_k^{(s)}Z_i' + \Lambda_{ik}^{(s)},\\ \gamma_{ik} &= E\{b_{ik}|y,\theta^{(s)}\} = \Psi_k^{(s)}Z_i'\Sigma_{ik}^{(s)^{-1}}(y_i - X_i\beta_k^{(s)}),\\ \Gamma_{ik} &= Var\{b_{ik}|y,\theta^{(s)}\} = (\Psi_k^{(s)} - \Psi_k^{(s)}Z_i'\Sigma_{ik}^{(s)^{-1}}Z_i\Psi_k^{(s)}).\end{aligned}$$

All these quantities and then the Q-function, $Q(\theta|\theta^{(s)})$, are straightforward to calculate using the above equations given $\theta^{(s)}$.

M-step

The M-step at iteration $s+1$ needs to update the parameters θ by maximizing the Q-function, $Q(\theta|\theta^{(s)})$. Letting the first partial derivatives of the Q-function be zero and using the constraint $\sum_{k=1}^{K}\alpha_k = 1$ and $\Lambda_{ik} = \sigma_k^2 V_i(\phi_k)$, we have

$$\begin{aligned}\hat{\alpha}_k &= \frac{1}{M}\sum_{i=1}^{M}\tau_{ik}\\ \hat{\beta}_k &= (\sum_{i=1}^{M}\tau_{ik}X_i'V_{ik}^{-1}X_i)^{-1}\sum_{i=1}^{M}\tau_{ik}X_i'V_{ik}^{-1}(y_i - Z_i\gamma_{ik})\\ \hat{\Psi}_k &= \frac{\sum_{i=1}^{M}\tau_{ik}B_{ik}}{\sum_{i=1}^{M}\tau_{ik}}\\ \hat{\sigma}_k^2 &= \frac{\sum_{i=1}^{M}\tau_{ik}tr\{V_{ik}^{-1}A_{ik}(\beta_k)\}}{\sum_{i=1}^{M}\tau_{ik}n_i}\end{aligned}$$

where $\sum_{i=1}^{M}\tau_{ik} \neq 0$ and $V_{ik} = V_i(\phi_k)$. The parameters ϕ_k or equivalently the matrices V_{ik} can be estimated by minimizing the last term in the Q-function (5):

$$\xi(\phi_k,\beta_k,\sigma_k^2) = \sum_{i=1}^{M}\tau_{ik}\{\log(\sigma_k^{2n_i}|V_{ik}|) + \sigma_k^{-2}tr(V_{ik}^{-1}A_{ik}(\beta_k))\}. \quad (6)$$

Substituting $\hat{\beta}_k$ and $\hat{\sigma}_k^2$ into (6), we have $\xi(\phi_k,\hat{\beta}_k,\hat{\sigma}_k^2) = \sum_{i=1}^{M}\tau_{ik}\{n_i\log\hat{\sigma}_k^2 + \log|V_{ik}| + n_i\}$. Then the ϕ_k are estimated by

$$\hat{\phi}_k = \arg\min\{\xi(\phi_k) = \xi(\phi_k,\hat{\beta}_k,\hat{\sigma}_k^2)\}.$$

The E-step and M-step are repeated until convergence. The individual or subject i is finally classified into the cluster $\hat{k}_i = \arg\max_k\{\tau_{ik}\}$.

3.2 EM algorithm using Monte Carlo sampling (EMMC)

The indictor variables for cluster membership, $(\delta_{i1},\ldots,\delta_{iK})$, follow a multinomial distribution with probabilities $\tau_i = (\tau_{i1},\ldots,\tau_{iK})$. Let $(\delta_{i1}^{(h)},\ldots,\delta_{iK}^{(h)})$, $h = 1,\ldots,H$, be H samples of clusters membership taken from the multinomial distribution, $Multinomial(1,\tau_i)$. Then we use $\delta_i^{(h)}$ to replace τ_i in the Q-function (5), and have

$$Q_h(\theta|\theta^{(s)}) = -\frac{1}{2}\sum_{k=1}^{K}Q_{h,k}(\theta|\theta^{(s)}),$$

where

$$Q_{h,k}(\theta|\theta^{(s)}) = \sum_{\delta_{ik}^{(h)}=1} \{-2\log\alpha_k + [\log|\Psi_k| + tr(\Psi_k^{-1}B_{ik})] + [\log|\Lambda_{ik}| + tr(\Lambda_{ik}^{-1}A_{ik})]\},$$

which is a Q-function of a linear mixed-effects (LME) models. Let $\hat{\theta}_h = \arg\max_\theta Q_h(\theta|\theta^{(s)})$. Then $\hat{\theta}_h = (\hat{\theta}_{h,1}, \ldots, \hat{\theta}_{h,K})$ that can be estimated separately in each cluster as a LME model (Pinheiro and Bates, 2000) by

$$\hat{\theta}_{h,k} = \arg\max_{\theta_k} Q_{h,k}(\theta|\theta^{(s)})$$

The mixing parameters are estimated by

$$\hat{\theta} = \frac{1}{H}\sum_{h=1}^{H}\hat{\theta}_h.$$

The parameter estimation for LME models is much faster than for a mixture of extended LME models. Hence the EM algorithm with Monte Carlo sampling converges quickly within each iteration. However, typically, a large Monte Carlo sample size is required to estimate the parameters within an acceptable tolerance. That is, H needs to be large. So the EMMC is computationally demanding overall.

3.3 Initial values and number of clusters

The EM algorithm can be quite sensitive to the choice of starting values. A number of different strategies for choosing starting values have been proposed (McLachlan and Peel, 2000). As Celeux et al. (2005) and Ciampi et al. (2012), we perform k-means clustering of regression parameters obtained from linear regressions on each individual or subject to obtain starting values.

The number of clusters in the finite mixture models may be estimated using Akaike information criterion (AIC) (Akaike, 1974) or Bayesian information criterion (BIC) (Schwarz, 1978). AIC selects more clusters and then has less missing clusters than BIC. So, for practical purpose, we may prefer AIC to BIC.

4. Simulations

We simulated data containing 200 individuals from 4 clusters that mimic 4 different patterns of time evolution: worsening, slowly worsening, slowly improving and improving. The number of individuals in each cluster was chosen to be 30, 43, 57, and 70, respectively. For individual i in cluster k, the responses were generated from the following model:

$$y_{ijk} = \beta_{0,k} + \beta_{1,k}t_{ijk} + b_{0,jk} + b_{1,jk}t_{ijk} + \sigma_k\epsilon_{ijk},$$

where $j = 1, \ldots, n_i$, n_i is the number of measures for individual i, t_{ijk} is the j-th measured time point, $\beta_{s,k}$, $s = 0, 1$, are fixed effects, $b_{s,jk} \sim N(0, \psi_{s,k})$ are random effects leading to cluster-specific and individual-specific patterns and correlation, and ϵ_{ijk} follows AR(1). The true parameters are shown in Table 1. It is important to note that the measurement times are unequally spaced for each individual, and n_i, the number of observations, was allowed to range from 15 to 25. We generated 500 datasets.

Firstly, assume that we know the true number of clusters, $K = 4$. We estimated the model parameters using the two algorithms: EM and EMMC, for each simulated dataset.

For each algorithm, the minimum and maximum iterations are set to be 100 and 300, respectively. The E and M iteration processes will be terminated if a log-likelihood increase is less than 0.001 or the iterations are over 300. The number of Monte Carlo samples for EMMC is 200. The choice of these numbers of iterations here is arbitrary but set large enough to ensure convergence. The mean and standard deviation of the estimated parameters over the 500 simulations for each algorithm are shown in Table 1. The average run time per simulation for this simulation setting is 2.39 hours for EM and 12.48 hours for EMMC. All simulations were run on the Colosse cluster of Compute Canada which has 2.8GHz blades in groups of 8, with a total of 24Gb RAM per 8 blades.

Secondly, we allowed the number of clusters to vary between 1 and 7, i.e., $1 \leq K \leq 7$. Then we estimated the number of clusters using both AIC and BIC model selection criteria. The frequencies of the numbers of clusters selected by AIC and BIC for the EM algorithm are shown in Figure 1. The simulation examples show that BIC selects a more accurate number of clusters than AIC, but sometimes underestimates K leading to potentially missing interesting trajectory patterns.

Table 1: True parameters and average estimates of the parameters over 500 simulations assuming the true number of clusters are known, that is, $K = 4$. The values in parentheses are sample standard deviations of the estimates.

Method	Cluster	α	β_0	β_1	ψ_0	ψ_1	ϕ	σ^2
True parameters								
	1	0.150	8	-0.75	0.50	0.01	0.45	1.00
	2	0.215	6	-0.25	0.50	0.01	0.45	1.41
	3	0.285	4	0.25	0.50	0.01	0.45	1.73
	4	0.350	2	0.75	0.50	0.01	0.45	2.00
Estimated parameters								
EM	1	0.157(0.028)	7.980(0.231)	-0.734(0.051)	0.450(0.221)	0.013(0.011)	0.447(0.05)	1.014(0.098)
	2	0.221(0.024)	5.931(0.306)	-0.223(0.075)	0.372(0.265)	0.015(0.013)	0.451(0.047)	1.438(0.132)
	3	0.283(0.031)	3.915(0.267)	0.272(0.062)	0.317(0.267)	0.015(0.012)	0.452(0.041)	1.761(0.136)
	4	0.339(0.029)	1.957(0.170)	0.754(0.017)	0.398(0.259)	0.010(0.003)	0.451(0.034)	2.013(0.119)
EMMC	1	0.153(0.024)	7.982(0.219)	-0.745(0.039)	0.490(0.221)	0.011(0.006)	0.448(0.049)	1.010(0.095)
	2	0.216(0.016)	5.968(0.279)	-0.243(0.060)	0.501(0.230)	0.010(0.004)	0.447(0.044)	1.417(0.116)
	3	0.284(0.021)	3.983(0.234)	0.256(0.054)	0.479(0.208)	0.010(0.004)	0.445(0.038)	1.732(0.118)
	4	0.347(0.025)	1.994(0.154)	0.752(0.018)	0.485(0.224)	0.010(0.003)	0.449(0.032)	1.999(0.114)

5. Clinical example

Data on response to warfarin therapy was collected on 99 patients with atrial fibrillation presenting to the anticoagulation clinic between March 2001 and June 2012 at a single tertiary-care center and enrolled in a previous study. Response to warfarin is measured with the International Normalized Ratio (INR), and each patient has multiple measurements (6 $\sim$ 33) measured over a time period of up to 240 days (8 months) post initiation of warfarin therapy. Patient responses to warfarin are known to be highly variable and dose needs to be carefully adjusted in some patients to achieve optimal anticoagulant effect. We modeled the data as samples from a mixture of K linear mixed models with autoregressive errors, where K is to be estimated from the data using either AIC or BIC. We found the best fit was 4 clusters using AIC, and 3 clusters using BIC. An uncommon cluster with substantial INR instability was only identified using the AIC criterion. Figure 2 shows typical patterns of INR trajectories.

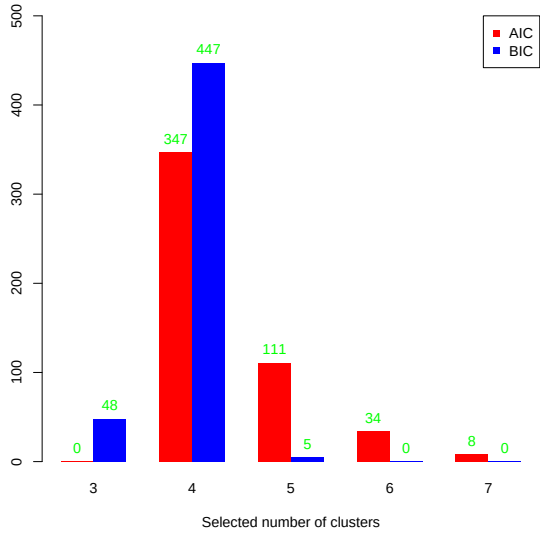

Figure 1: Frequencies of cluster numbers selected by AIC and BIC for EM algorithm.

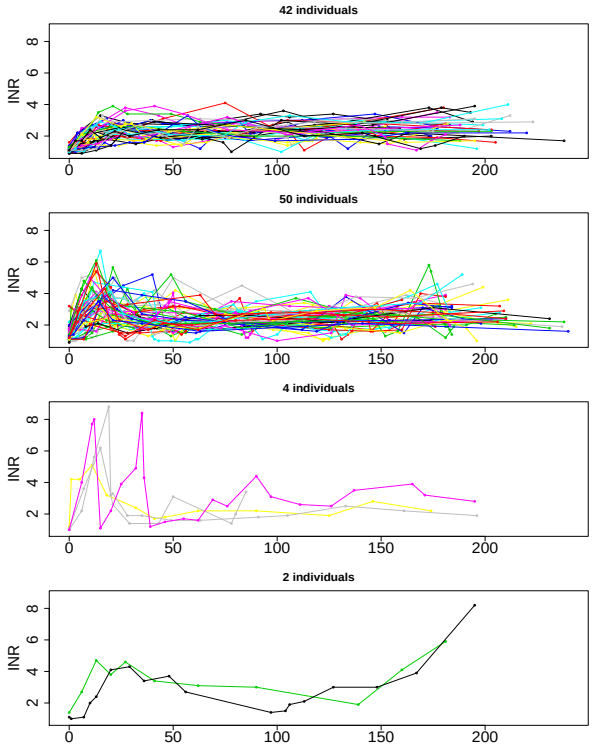

Figure 2: Typical patterns of INR trajectories

6. Conclusion

A mixture of extended LME models has been proposed for clustering of longitudinal data, and two EM-type algorithms: standard EM and Monte Carlo sampling based EM, have been developed to estimate the mixture model.

BIC remains the most prevalent mixture model selection technique within the literature. Our simulation examples also showed that BIC tends to select a more accurate number of clusters than AIC but also can miss some true clusters. In practice, it is recommended to look at the clusters selected by both AIC and BIC since the small clusters found only by AIC may have relevance.

The standard errors of $\hat{\theta}$ may be given by the diagonal elements of the asymptotic covariance matrix of the maximum-likelihood estimates (Ciampi et al., 2012) , which is equal to the inverse of the expected Information matrix and can be approximated by the Louis (1982) decomposition.

The extended LME models are based on the Gaussian or normal distribution. Mixture of different types of distributions, such as mixtures of Gaussian and uniform distributions or t-distributions (McNicholas and Subedi, 2012), has received attention recently and it could be interesting to extend this work to more general distributions.

Although total computation time for EMMC is longer than for EM, the former can be parallelized so that total elapsed time for estimation could be considerably faster.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control 19*(6), 716–723.

Banfield, J. D. and A. E. Raftery (1993). Model-based gaussian and non-gaussian clustering. *Biometrics 49*(3), 803–821.

Celeux, G., O. Martin, and C. Lavergne (2005). Mixture of linear mixed models for clustering gene expression profiles from repeated microarray experiments. *Statistical Modelling 5*, 243–267.

Ciampi, A., H. Campbell, A. Dyachenko, B. Rich, J. McCusker, and M. G. Cole (2012). Model-based clustering of longitudinal data: application to modeling disease course and gene expression trajectories. *Communications in Statistics - Simulation and Computation 41*(7), 992–1005.

De la Cruz-Mesia, R., F. A. Quintana, and G. Marshall (2008). Model-based clustering for longitudinal data. *Computational Statistics and Data Analysis 52*, 1441–1457.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of Royal Statistical Society, Series. B 39*(1), 1–38.

Fraley, C. and A. E. Raftery (1999). Mclust: Software for model-based cluster analysis. *Journal of Classification 16*, 297–306.

Fraley, C. and A. E. Raftery (2002). Model-based clustering, discriminant analysis and density estimation. *Journal of the American Statistical Association 97*, 611–631.

Fraley, C. and A. E. Raftery (2007). Model-based methods of classification: using the mclust software in chemometrics. *Journal of Statistical Software 18*(6).

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of Royal Statistical Society, Series. B 44*, 226–233.

McLachlan, G. and K. Basford (1988). *Mixture Models: Inference and Applications to Clustering*. New York: Marcel Dekker.

McLachlan, G. and D. Peel (2000). *Finite Mixture Models*. New York: Wiley.

McLachlan, G. J., D. Peel, and R. W. Bean (2003). Modelling high-dimensional data by mixtures of factor analyzers. *Computational Statistics & Data Analysis 41*, 379–388.

McNicholas, P. D. and T. B. Murphy (2008). Parsimonious gaussian mixture models. *Statistics and Computing 18*(3), 285–296.

McNicholas, P. D. and T. B. Murphy (2010). Model-based clustering of longitudinal data. *The Canadian Journal of Statistics 38*(1), 153–168.

McNicholas, P. D., T. B. Murphy, A. F. McDaid, and D. Frost (2010). Serial and parallel implementations of model-based clustering via parsimonious gaussian mixture models. *Computational Statistics and Data Analysis 54*(3), 711–723.

McNicholas, P. D. and S. Subedi (2012). Clustering gene expression time course data using mixtures of multivariate t-distributions. *Journal of Statistical Planning and Inference 142*(5), 1114–1127.

Pauler, D. K. and N. M. Laird (2000). A mixture model for longitudinal data with application to assessment of noncompliance. *Biometrics 56*, 464–472.

Pinheiro, J. C. and D. M. Bates (2000). *Mixed-Effects Models in S and S-PLUS*. New York: Springer.

Pourahmadi, M. (1999). Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation. *Biometrika 86*(3), 677–690.

R Development Core Team (2013). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics 6*(2), 461–464.

Titterington, D. M., A. F. M. Smith, and U. E. Makov (1985). *Statistical Analysis of Finite Mixture Distributions*. Chichester: John Wiley & Sons.