1-1-2008

Fitting finite mixtures of linear mixed models with the EM algorithm

Bettina Grun
University of Wollongong, bettina@uow.edu.au

Fitting finite mixtures of linear mixed models with the EM algorithm

Abstract

Finite mixtures of linear mixed models are increasily applied in differentareas of application. They conveniently allow to account for correlations betweenobservations from the same individual and to model unobserved heterogeneity betweenindividuals at the same time. Different variants of the EM algorithm arepossible for maximum likelihood (ML) estimation. In this paper two different versionsfor fitting this model class are presented. One variant of the EM algorithmrequires weighted ML estimation. As this fitting method might not be readily availablein standard software sufficient conditions which allow to transform a weightedinto an unweighted ML estimation problem are derived.

Keywords

Fitting, finite, mixtures, linear, mixed, models, algorithm

Disciplines

Business | Social and Behavioral Sciences

Publication Details

Grun, B. (2008). Fitting finite mixtures of linear mixed models with the EM algorithm. In P. Brito (Eds.), Compstat 2008 - International Conference on Computational Statistics (pp. 165-173). Germany: Springer.

Fitting Finite Mixtures of Linear Mixed Models with the EM Algorithm

Bettina Grün

Department für Statistik und Mathematik, Wirtschaftsuniversität Wien
Augasse 2-6, 1090 Wien, Austria, *Bettina.Gruen@wu-wien.ac.at*

Abstract. Finite mixtures of linear mixed models are increasily applied in different areas of application. They conveniently allow to account for correlations between observations from the same individual and to model unobserved heterogeneity between individuals at the same time. Different variants of the EM algorithm are possible for maximum likelihood (ML) estimation. In this paper two different versions for fitting this model class are presented. One variant of the EM algorithm requires weighted ML estimation. As this fitting method might not be readily available in standard software sufficient conditions which allow to transform a weighted into an unweighted ML estimation problem are derived.

Keywords: EM algorithm, finite mixture, linear mixed model, unobserved heterogeneity

1 Introduction

Finite mixture models are a popular method for modelling unobserved heterogeneity. In the last decades the original model of finite mixtures of distributions has been extended in several ways and nearly arbitrary component specific models are nowadays used in applications. This development has been faciliated by estimation techniques which constitute a common framework for fitting arbitrary mixture models and which require only to modify the component specific model estimation for different mixture models. This holds for the Expectation-Maximization (EM) algorithm (Dempster, Laird and Rubin (1977)) for maximum likelihood (ML) estimation.

Finite mixtures of mixed effects models allow to account for different kinds of heterogeneity between individuals (Frühwirth-Schnatter 2006). The components of the mixture represent different groups with distinct parameterizations while the random effects allow for individual differences which cluster around a common mean value. These models are applied in several different areas such as marketing (Lenk and DeSarbo (2000)), medicine (Xu and Hedeker (2001)) and bioinformatics (Luan and Li (2004)).

This paper is organized as follows: Section 2 introduces the model. Section 3 outlines two variants of the EM algorithm for ML estimation of this model class and derives sufficient conditions for allowing the use of implementations of fitting algorithms for unweighted mixed-effects models. A short sketch of a possible implementation in R is provided.

2 Model specification

In the following finite mixtures of mixed effects models are considered where the mixed effects are needed to account for correlations between observations from the same individual and the finite mixture models the unobserved heterogeneity between the individuals. This implies that the component memberships of the individuals are fixed.

Assume observations from N individuals are given and for each individual i the data (Y_i, X_i, Z_i, w_i) is given which consists of n_i observations on the dependent variables $Y_i = (y_{ij})_{j=1,\ldots,n_i}$, the covariates for the fixed effects $X_i = (x_{ij})_{j=1,\ldots,n_i}$ and the covariates for the random effects $Z_i = (z_{ij})_{j=1,\ldots,n_i}$. w_i denote the individual specific concomitant variables.

The finite mixture density of mixed effects models with K components is given for the observations of individual i by

$$h(Y_i|X_i, Z_i, w_i, \Theta) = \sum_{k=1}^{K} \pi_k(w_i) \int \prod_{j=1}^{n_i} \phi_1(y_{ij}; x_{ij}\beta_k + z_{ij}b_i^k, \sigma_k^2)\phi_q(b_i^k; 0, \Psi_k)db_i^k$$
$$= \sum_{k=1}^{K} \pi_k(w_i)\phi_{n_i}(Y_i; X_i\beta_k, Z_i\Psi_k Z_i^T + \sigma_k^2 I_{n_i}).$$

$\phi_d(\cdot; \mu, \Sigma)$ denotes the d-dimensional normal distribution with mean μ and variance-covariance matrix Σ. The fixed effects are given by β_k and the random effects by b_i^k. The random effects are assumed to have mean zero which implies that any constant influence is already captured by the fixed effects. This can be ensured by constraining that the covariates Z_i span a subset of the space spanned by X_i over all individuals $i = 1, \ldots, N$.

The variance-covariance matrix of Y_i for component k is given by

$$\Sigma_i^k = \sigma_k^2 \Sigma_{i0}^k = Z_i \Psi_k Z_i^T + \sigma_k^2 I_{n_i}.$$

It is assumed that $\Psi_k = \sigma_k^2 \theta_k$ and $\Sigma_i^k = \sigma_k^2 (Z_i \theta_k Z_i^T + I_{n_i})$.

The component weights $\pi_k(w_i)$ are assumed to fulfill the following conditions for all i:

$$\pi_k(w_i) > 0 \quad \forall k \qquad \text{and} \qquad \sum_{k=1}^{K} \pi_k(w_i) = 1.$$

The most common concomitant variable model for w_i is the multinomial logit model (Dayton and Macready (1988)).

This model specification implies that there exist no common parameters which are constant over the components and hence, each of the components can be separately estimated given the component memberships of the individuals. As the component memberships are fixed for all observations $j = 1, \ldots, n_i$ of individual i, it is also assumed that the concomitant variables

w_i are constant for each individual. A different model specification where the concomitant variables for individual i are given by $W_i = (w_{ij})_{j=1,\ldots,n_i}$ and the component membership π_k is not fixed for each individual is for example given in Yau et al. (2003) and Hall and Wang (2005).

3 Estimation with the EM algorithm

The EM algorithm is in general applied in a missing data context. It is an iterative procedure which alternates between an E(xpectation)-step and a M(aximization)-step. The EM algorithm works on the complete likelihood derived by also including the missing data and exploits the fact that the complete likelihood is in general easier to maximize than the original likelihood. The missing data is integrated out in the E-step by determining the expectation of the complete likelihood given the available data and the current parameter estimates. The expected complete likelihood is then maximized in the M-step.

The EM algorithm has been shown to increase the likelihood in each step and hence to converge for bounded likelihoods. The implementation of the EM algorithm can often be simplified by introducing more variables as missing data. However, the disadvantage is that the convergence of the EM algorithm depends on the amount of missing data and hence, more iterations are needed if the amount of missing data is increased.

For finite mixtures of linear mixed effects models different variants for ML estimation with the EM algorithm have been proposed. In the following two different versions are discussed in detail which differ with respect to the variables they use as missing data.

3.1 Random effects and component memberships as missing data

The most popular variant of the EM algorithm for fitting finite mixtures of linear mixed effects models is where the component memberships as well as the random effects are treated as missing data and imputed in the E-step (see for example Xu and Hedeker (2001) or Celeux et al. (2005)).

For this variant the E-step consists of determining

1. the a posteriori probabilities that an individual i is from component k:

$$\tau_{ik} = \frac{\pi_k(w_i)\phi_{n_i}(X_i\beta_k, Z_i\Psi_k Z_i^T + \sigma_k^2 I_{n_i})}{\sum_{l=1}^{K} \pi_l(w_i)\phi_{n_i}(X_i\beta_l, Z_i\Psi_l Z_i^T + \sigma_l^2 I_{n_i})}$$

   and
2. the mean and the variance of the random effects b_i conditional on the current parameter estimates Θ, the observations Y_i, the covariates X_i and Z_i and the component k. These are calculated using that b_i and Y_i

follow a joint multivariate normal distribution conditional on Θ, X_i, Z_i and k:

$$\begin{aligned}\mu_{b_i,k} &= \mathbb{E}[b_i|Y_i, X_i, Z_i, \Theta, k] \\ &= [\frac{1}{\sigma_k^2} Z_i^T Z_i + \Psi_k^{-1}]^{-1} \frac{1}{\sigma_k^2} Z_i^T (Y_i - X_i\beta_k) \\ \Sigma_{b_i,k} &= \mathbb{V}[b_i|Y_i, X_i, Z_i, \Theta, k] = \left[\frac{1}{\sigma_k^2} Z_i^T Z_i + \Psi_k^{-1}\right]^{-1}.\end{aligned}$$

The expected complete likelihood is given by

$$\sum_{k=1}^{K}\sum_{i=1}^{N} \tau_{ik} \Bigg[\log \pi_k(w_i) - \frac{1}{2}\Bigg((n_i + q)\log(2\pi) + n_i \log \sigma_k^2 + \log|\Psi_k| + \sum_{j=1}^{n_i} \frac{(y_{ij} - z_{ij}\mu_{b_i,k} - x_{ij}\beta_k)^2 + z_{ij}\Sigma_{b_i,k} z_{ij}^T}{\sigma_k^2} + + \mathrm{tr}(\Psi_k^{-1}(\Sigma_{b_i,k} + \mu_{b_i,k}\mu_{b_i,k}^T))\Bigg)\Bigg].$$

$\mathrm{tr}(\cdot)$ denotes the trace of a matrix.

For the M-step the parameters of the concomitant variable model and the component specific model can be separately determined. For the concomitant variable model a weighted multinomial logit model has to be estimated if the component weights are determined through a multinomial logit model. This estimation method is often already available in standard statistical software. For the component specific model the parameters can be determined in closed form by solving the equations derived by determining the derivatives of the expected complete likelihood and setting them to zero:

$$\begin{aligned}\hat{\beta}_k &= \frac{1}{\sum_{i=1}^N \tau_{ik}} \Big(\sum_{i=1}^{N} \tau_{ik} \sum_{j=1}^{n_i} x_{ij}^T x_{ij}\Big)^{-1} \left[\sum_{i=1}^{N} \tau_{ik} \sum_{j=1}^{n_i} x_{ij}^T (y_{ij} - z_{ij}\mu_{b_i,k})\right] \\ \hat{\sigma}_k^2 &= \frac{1}{\sum_{i=1}^N \tau_{ik} n_i} \sum_{i=1}^{N} \tau_{ik} \sum_{j=1}^{n_i} (y_{ij} - z_{ij}\mu_{b_i,k} - x_{ij}\hat{\beta}_k)^2 + z_{ij}\Sigma_{b_i,k} z_{ij}^T \\ \hat{\Psi}_k &= \frac{1}{\sum_{i=1}^N \tau_{ik}} \sum_{i=1}^{N} \tau_{ik} (\Sigma_{b_i,k} + \mu_{b_i,k}\mu_{b_i,k}^T).\end{aligned}$$

3.2 Component memberships as missing data

An alternative implementation would be the straightforward application of the EM algorithm as in general used for finite mixtures, i.e., only the component membership is treated as missing data. This implementation requires

the weighted ML estimation of the linear mixed model for the M-step and the determination of the posterior probabilities in the E-step. Standard software for fitting linear mixed effects models often does not allow for weighted ML estimation or does only account for different variance-covariance matrices for the error term. Under certain conditions an unweighted ML estimation can be used for weighted ML estimation where the observations are suitably transformed. The following corollary gives sufficient conditions.

Corollary 1 (Weighted ML estimation). *The weigthed ML estimate of θ of a linear mixed model with observations (Y_i, X_i, Z_i) and weights τ_i for $i = 1, \ldots, N$ is equivalent to the ML estimate of θ of a linear mixed model with transformed variables $\tilde{X}_i = \sqrt{\tau_i} X_i$ and $\tilde{Y}_i = \sqrt{\tau_i} Y_i$ and the same Z_i if*

$$Z_i \equiv Z \quad \forall i = 1, \ldots, N.$$

Proof. The weighted deviance which is equivalent to -2 log-likelihood is given by

$$\operatorname{dev}(\beta, \theta, \sigma^2) = \sum_{i=1}^{N} \tau_i n_i \log(2\pi\sigma^2) + \tau_i \log|\Sigma_{i0}| + \\ + \frac{\tau_i}{\sigma^2}(Y_i - X_i\beta)^T \Sigma_{i0}^{-1}(Y_i - X_i\beta).$$

The ML estimates of the coefficients $\hat{\beta}$ and the variance $\hat{\sigma}^2$ depend on the weighted residual sum of squares $r_{\tau i}^2$ and for determining the profile deviance they are all functions of θ:

$$r_{\tau i}^2(\theta) = \tau_i (Y_i - X_i\hat{\beta}(\theta))^T \Sigma_{i0}^{-1}(Y_i - X_i\hat{\beta}(\theta)) \tag{1}$$

$$\hat{\sigma}^2(\theta) = \frac{\sum_{i=1}^{N} r_{\tau i}^2(\theta)}{\sum_{i=1}^{N} \tau_i n_i} \tag{2}$$

Given Equation (1) $\hat{\beta}(\theta)$ is given by the generalized least squares estimate for the variance-covariance matrix Σ_{i0}.

The profile deviance is then given by

$$\operatorname{dev}(\theta) = \sum_{i=1}^{N} \tau_i n_i \log\left(2\pi \frac{\sum_{i=1}^{N} r_{\tau i}^2}{\sum_{i=1}^{N} \tau_i n_i}\right) + \tau_i \log|\Sigma_{i0}| + \tau_i n_i. \tag{3}$$

If $Z_i \equiv Z$ for all i and hence also $n_i \equiv n$, this gives

$$\operatorname{dev}(\theta) = \tilde{\tau}\left[n\left(1 + \log(\frac{2\pi}{\tilde{\tau} n}) + \log(\sum_{i=1}^{N} r_{\tau i}^2)\right) + \log|Z\theta Z^T + I_n|\right]$$

where $\tilde{\tau} = \sum_{i=1}^{N} \tau_i$.

The profile deviance for $\tilde{X}_i = \sqrt{w_i} X_i$ and $\tilde{Y}_i = \sqrt{w_i} Y_i$ is given by

$$\text{dev}(\theta) = N \left[n \left(1 + \log(\frac{2\pi}{Nn}) + \log(\sum_{i=1}^{N} r_{\tau i}^2) \right) + \log |Z\theta Z^T + I_n| \right].$$

As the profile deviances are equivalent up to an additive constant and a constant factor they are maximized for the same θ.

The estimates for β and σ^2 are then determined using Equations (1) and (2). The estimate of β is the same for the weighted and the unweighted but transformed fitting problem, because the residual sum of squares term is identical up to a constant factor. Only the estimate of σ^2 has to be modified if the estimates of the unweighted but transformed fitting problem are used. As can be seen in Equation (2) the denominator of the weighted estimation problem is $\sum_{i=1}^{N} \tau_i n_i$ while it is $\sum_{i=1}^{N} n_i$ for the unweighted but transformed problem.

From the weighted profile deviance (Equation 3) it can be seen which changes are necessary to allow for weighted ML estimation. It is not sufficient to only change the residual sum of squares but the weights also influence the sum over the logarithm of the determinant of the individual variance-covariance matrices. Accounting for the weights in the estimation might then not be easily possible if for example the following simplification is used by the software for the determining the determinant

$$|\tilde{Z}\tilde{Z}^T + I_{\sum_{i=1}^{N} n_i}| = |\tilde{Z}^T\tilde{Z} + I_q|.$$

The sufficient conditions indicate that standard software can easily be used in the case where a balanced design is given, i.e., the same observations are available for each individual. Without missing data this occurs for example in bioinformatics where gene expression data is observed over time at a priori specified time points. The conditions might also be more likely applicable in the case where only a random intercept is fitted.

If the entire data set does not fulfill the sufficient conditions, only the subsample fulfilling the conditions might be used in a first step to pre-analyse the data. The entire data set can then be fitted using the EM algorithm where also the random effects are used as missing information but which is initialized in the previously found solution.

If the transformation of the weighted into an unweighted ML estimation problem is not possible, the Classification EM algorithm (CEM; Celeux and Govaert (1992)) can be used instead of the classical EM algorithm. The CEM algorithm allows to use unweigthed ML estimation methods. However, it does not maximize the likelihood but the classification likelihood. The advantage of the CEM algorithm is that it converges in general faster than the EM algorithm, i.e., it needs less iterations. It has been therefore proposed to use the CEM algorithm with different random initializations to find a good

starting point for the ordinary EM algorithm which in the case of finite mixtures of mixed effects models might be the variant where the random effects are also included in the missing data.

3.3 Implementation in R

Both variants of the EM algorithm can easily be implemented in R, an environment for statistical computing and graphics (R Development Core Team (2007)). Package **flexmix** (Leisch (2004)) for example implements the EM algorithm for ML estimation of finite mixture models. It provids the E-step and all data handling and arbitrary mixture models can be fitted by modifying the M-step. The implementation of the package aims at easy extensibility and tries to enable rapid prototyping. In general only a model driver for the component specific model needs to be written which specifies the fitting function. In addition the package also allows fitting of finite mixture models with the CEM algorithm.

The recommended package in R for fitting linear mixed effects models is **nlme** (Pinheiro and Bates (2000)). Function `lme()` allows to specify a weights argument, which can be used to describe the within-group heteroscedasticity structure. An alternative implementation is provided by the package **lme4** (Bates (2007)). The weights argument of function `lmer()` specifies that a weighted residual sum of squares is minimized. Hence, the recommended functions in R do not allow for weighted ML estimation of linear mixed models. The sufficient conditions can be used to determine when it is possible to estimate the transformed problem using this functionality in combination with package **flexmix**.

4 Conclusion and future work

The most common way of fitting finite mixtures of mixed effects models with the EM algorithm is by introducing the component memberships and the random effects as missing data. However, this signifies that this EM algorithm is different from the general application of the EM algorithm for finite mixture models where only the component memberships are used as missing data and the M-step consists of weighted ML estimation of the component specific models.

As the reason for the preference of this variant might be that weighted ML estimation of linear mixed models is not readily available in standard statistical software, this paper investigates which conditions need to be fulfilled that the weighted ML problem is equivalent to an unweighted but transformed ML problem. The results indicate that this is possible in applications where a balanced design is used to collect the data. In addition it is likely to be at least applicable for a subset of the data in random intercept models.

In the future the performance of the two EM algorithms should be compared. The variant where the component memberhips as well as the random effects are used as missing data can be expected to need more iterations while each iteration will take less time as the M-step is given in closed form. The advantage of the other variant is that if the fitting function of the linear mixed model is improved this can be exploited in the M-step. In addition it might be useful to investigate how the fitting function of the linear mixed models has to be modified to allow for weighted ML estimation.

Acknowledgments

This piece of research was supported by the Austrian Science Foundation (FWF) under grant T351.

References

BATES, D. (2007): *lme4: Linear mixed-effects models using S4 classes.* R package version 0.99875-8.

CELEUX, G. and GOVAERT, G. (1992): A Classification EM algorithm for clustering and two stochastic versions. *Computational Statistics & Data Analysis, 14(3), 315–332.*

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

DAYTON, C.M. and MACREADY, G.B. (1988): Concomitant-variable latent-class models. *Journal of the American Statistical Association, 83(401), 173–178.*

DEMPSTER, A.P., LAIRD, N.M. and RUBIN, D. B. (1977): Maximum likelihood from incomplete data via the EM-algorithm. *Journal of the Royal Statistical Society B, 39, 1–38.*

FRÜHWIRTH-SCHNATTER, S. (2006): *Finite Mixture and Markov Switching Models.* Springer.

HALL, D.B. and WANG, L. (2005): Two-component mixtures of generalized linear mixed effects models for cluster correlated data. *Statistical Modelling, 5, 21–37.*

LEISCH, F. (2004): FlexMix: A general framework for finite mixture models and latent class regression in R. *Journal of Statistical Software, 11(8),* 1–18.

LENK, P.J. and DESARBO, W.S. (2000): Bayesian inference for finite mixtures of generalized linear models with random effects. *Psychometrika, 65(1), 93–119.*

LUAN, Y. and LI, H. (2004): Model-based methods for identifying periodically expressed genes based on time course microarray gene expression data. *Bioinformatics, 20(3), 332–339.*

PINHEIRO, J.C. and BATES, D.M. (2000): *Mixed-Effects Models in S and S-Plus.* Springer.

R DEVELOPMENT CORE TEAM (2007): *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria.

XU, W. and HEDEKER, D. (2001): A random-effects mixture model for classifying treatment response in longitudinal clinical trials. *Journal of Biopharmaceutical Statistics, 11(4), 253–273.*

YAU, K.K., LEE, A.H. and NG, A.S. (2003): Finite mixture regression model with random effects: application to neonatal hospital length of stay. *Computational Statistics & Data Analysis, 41, 359–366.*