Random effects regression mixtures for analyzing infant habituation

Derek S. Young\textsuperscript{a}\textsuperscript{*} and David R. Hunter\textsuperscript{b}

\textsuperscript{a}Department of Statistics, University of Kentucky, Lexington, KY 40536, USA; \textsuperscript{b}Department of Statistics, The Pennsylvania State University, University Park, PA 16802, USA

(Received 30 August 2013; accepted 16 December 2014)

Random effects regression mixture models are a way to classify longitudinal data (or trajectories) having possibly varying lengths. The mixture structure of the traditional random effects regression mixture model arises through the distribution of the random regression coefficients, which is assumed to be a mixture of multivariate normals. An extension of this standard model is presented that accounts for various levels of heterogeneity among the trajectories, depending on their assumed error structure. A standard likelihood ratio test is presented for testing this error structure assumption. Full details of an expectation-conditional maximization algorithm for maximum likelihood estimation are also presented. This model is used to analyze data from an infant habituation experiment, where it is desirable to assess whether infants comprise different populations in terms of their habituation time.

Keywords: bootstrap; ECM algorithm; likelihood ratio test; switching regressions; trajectory data

AMS Subject Classification: 62J05; 62P15

1. Introduction

Finite mixture models have long been used as a way to model a sample from a population subdivided into some known number of distinct categories that occur in unknown proportions in the case where each individual’s category membership is unobserved. Finite mixture models appear in areas such as economics, machine learning, neural networks, and the social and behavioral sciences. In addition to a vast library of journal articles about mixture models, there are also books dedicated to the subject; for instance, see Titterington \textit{et al.} \cite{40}, Lindsay \cite{21}, McLachlan and Peel \cite{25}, and Mengersen \textit{et al.} \cite{27}.

\textsuperscript{*}Corresponding author. Email: derek.young@uky.edu

© 2015 Taylor & Francis
In a typical multivariate finite mixture model, the \( m \)-dimensional vectors \( Y_1, \ldots, Y_n \) are a simple random sample from a \( k \)-component mixture distribution in which \( Y_i \) has density

\[
f(y_i \mid \Psi) = \sum_{j=1}^{k} \lambda_j g(y_i \mid \theta_j),
\]

where \( k > 1 \) is a fixed integer (and assumed known for now) and the \( \lambda_j > 0 \) are the component mixing proportions which sum to unity. Much of the mixture modeling literature considers the \( g(\cdot \mid \theta_j) \)'s in Equation (1) to be multivariate normal densities. If a vector of covariates, say \( x_i = (x_{i,1}, \ldots, x_{i,p})^T \), is also observed with a (univariate) response \( y_i \) in the mixture setting, then \( g(y_i \mid \theta_j) \) in Equation (1) becomes \( g(y_i \mid x_i, \theta_j) \) and the resulting model is a mixture of regressions. This model was first studied in detail in the econometrics literature by Quandt [33], who introduced it as the switching regressions model. Wedel and DeSarbo [45] and McLachlan and Peel [25] provide extensive literature reviews on many mixture of regression models along with different estimation methods and applications. Here, we assume that each individual observation is actually an entire set of observations that come from the same regression model. Thus, we may identify each observation with a trajectory – a linear function, in this case – that describes the conditional mean of the response variable as a function of the predictors. Assuming that the regression parameters are random variables that are themselves drawn from a mixture distribution, we arrive at a random effects regression mixture model. The specifics of the model are introduced in Section 2.1.

Closely connected to our discussion is latent class analysis for longitudinal data, which has broad applicability in analyzing psychological data. Often called the growth mixture model (GMM), this approach has been applied in numerous psychological studies over the years. For example, GMMs have been used to characterize alcoholism [28,32], analyze behavior in young children [36,42], and applied in studies about depression [13,15]. Another term for this model is the latent class growth model (LCGM), which has been extensively discussed in Nagin [30] and Andruff et al. [2] as a way to identify distinct subgroups of individuals following a similar pattern of change over time for a given variable. For example, Raudenbush [34] identified the need to characterize different subgroups of people regarding depression since the use of a single average trajectory could mask important individual differences and lead to erroneous conclusions. Louvet et al. [22] found three distinct trajectories that characterized how a sample of athletes dealt with coping across competitions. Vaughn et al. [43] used a LCGM-type model when analyzing multiple homicide offenders. The results of this last study showed three groups within these homicide offenders, which may have an underlying connection to their psychosocial background. Computation for GMMs and LCGMs can be performed using the MPlus software of Muthén and Muthén [29], which has been applied considerably in the literature. Besides GMMs and LCGMs, there are many other relevant techniques that amount to classifying trajectory data. Most notably are classification methods for functional data (see [3] for a review).

Any of the methods discussed earlier are viable strategies for analyzing trajectory data in psychological studies. However, mixture models are also a common analytical tool for interpreting heterogeneity from longitudinal studies in psychology [37]. Our goal is to describe a random effects regression mixture model to characterize distinct subgroups of infants that may possess differing abilities based on a habituation experiment. We present an extension of the random effects regression mixture model in Xu and Hedeker [46] and Gaffney and Smyth [12] where we allow the variance term to differ between the trajectories, which may be more appropriate for the different subgroups. Mixture-of-regression models with random effects have been considered from both the frequentist [10,44,46] and Bayesian [12,20] perspectives. Our manuscript gives a general treatment of the frequentist problem of parameter estimation, for which most of the cited references employ an expectation-maximization (EM) algorithm [9]. The extension we present...
requires use of an expectation-conditional maximization (ECM) algorithm [26], which we carefully detail in the appendix. We also discuss various hypothesis tests that arise in this situation (e.g. testing the different levels of heterogeneity when conditioned on component membership) and their scientific interpretations for the infant habituation data. The code developed for this analysis is available in the R contributed package mixtools [4].

2. Random effects regression mixtures

2.1 The model

Consider the following mixed-effects regression model for observation $i$, $i = 1, \ldots, n$:

$$y_i = U_i \alpha + X_i \beta + \epsilon_i,$$  \hfill (2)

where $y_i$ is an $n_i$-dimensional response vector (trajectory) that is linearly related to two quantities: (1) a known $n_i \times q$ matrix of explanatory variables $U_i$ with an associated $q$-dimensional vector of unknown regression parameters $\alpha$ and (2) a known $n_i \times p$ design matrix $X_i$ with an associated $p$-dimensional vector of unknown individual (random) effects $\beta_i \sim N_p(\mu, \Sigma)$. The $\beta_i$ are distributed independently of one another and of the $\epsilon_i$, which are $n_i$-dimensional random error terms that are, themselves, distributed independently as $N_{n_i}(0, \sigma^2 I_{n_i})$, where $I_d$ denotes the $d \times d$ identity matrix.

Now, assume that the $i$th of $n$ subjects generates an $n_i$-dimensional trajectory that can be modeled similar to Equation (2), but the individual comes from one of $k$ groups, say group $j_i$, and the error terms are now $\epsilon_i \sim N_{n_i}(0, \sigma^2_{j_i} I_{n_i})$. The conditional density for this setting is

$$p(y_i | U_i, X_i, \alpha, \beta_i, \sigma^2_{j_i}) = \phi_{n_i}(y_i | U_i \alpha + X_i \beta_i, \sigma^2_{j_i} I_{n_i}),$$

where $\phi_d(\cdot | \mu, \Sigma)$ denotes the $d$-dimensional normal density with mean $\mu$ and covariance $\Sigma$.

We now assume that the $p$-dimensional vector $\beta_i$ is normally distributed with mean $\mu_{j_i}$ and variance $R_{j_i}$. Notice that the mean of $\beta_i$ and the variances of both $\beta_i$ and $\epsilon_i$ depend on the mixture component $j_i$ from which this trajectory is drawn. Because we are most interested in the conditional distribution of $y_i$ given $X_i$, we ignore the distribution of the $X_i$ here.

Letting $j_i$ specify to which of the $k$ groups the $i$th trajectory belongs, assume that $\beta_i | j_i \sim N_p(\mu_{j_i}, R_{j_i})$ and that $\beta_i$ is conditionally independent of $\epsilon_i$. Let $\varsigma^2 = (\sigma^2_1, \ldots, \sigma^2_k)^\top$, $\mu = (\mu_1^\top, \ldots, \mu_k^\top)^\top$, and $R = (\text{vech}(R_1)^\top, \ldots, \text{vech}(R_k)^\top)^\top$, where $\text{vech}$ is the half-vectorization function which gives, in vector form, the lower triangular portion of the symmetric matrix it operates on. Assuming that $j_i$ takes value $j$ with probability $\lambda_j$, the marginal density of $\beta_i$ is the mixture density

$$p(\beta_i | \lambda, \mu, R) = \sum_{j=1}^k \lambda_j \phi_p(\beta_i | \mu_{j_i}, R_{j_i}).$$  \hfill (3)

Here, we let $\lambda^\top = (\lambda_1, \ldots, \lambda_{k-1})$ denote only the first $k - 1$ component weights, since $\lambda_k$ is a linear combination of these first $k - 1$ values. Finally, let

$$\psi^\top = (\varsigma^2, \mu^\top, R^\top, \lambda^\top, \alpha^\top)$$  \hfill (4)

and $\beta^\top = (\beta_1^\top, \ldots, \beta_n^\top)$.

Under the earlier assumptions and given the component membership $j_i$, the joint distribution of the trajectories $y_i$ and the random effects $\beta_i$ is derived as for a standard random effects regression
model. Straightforward calculations show that, conditional on component membership $j_i$, $(y_i, \beta_i)$ has the following joint multivariate normal distribution:

$$
\begin{bmatrix} y_i \\ \beta_i \end{bmatrix} \sim \mathcal{N}_{n_i+p} \left( \begin{bmatrix} U_i \alpha + X_i \mu_{j_i} \\ \mu_{j_i} \end{bmatrix}, \begin{bmatrix} X_i R_{j_i} X_i^T + \sigma_{j_i}^2 I_{n_i} & X_i R_{j_i} \\ R_{j_i}^T X_i & R_{j_i} \end{bmatrix} \right).$

(5)

Maximum likelihood estimation using the aforementioned framework can now be performed using an ECM algorithm [26]. We note that the incorporation of $\sigma_{j_i}^2$ is what necessitates the use of an ECM algorithm over a traditional EM algorithm. If $\sigma_{j_i}^2 \equiv \sigma^2$ for all $j = 1, \ldots, k$, then our model collapses to the traditional model presented in Xu and Hedeker [46] and, thus, an EM algorithm can be employed. The details of constructing an ECM algorithm for our random effects regression mixture model are presented in the appendix.

### 2.2 Equal variances and equal variance–covariance matrices

As we have noted, the difference between the model presented here and the models presented in Verbecke and Lesaffre [44], Xu and Hedeker [46], and Gaffney and Smyth [12] is that we allow for different trajectory variance terms. This means that, given the component membership of the random effects, we allow the trajectories to be heterogeneous between the different components. Here, we use a likelihood ratio test (LRT) for the hypotheses:

$$H_0 : \text{There exists } \sigma^2 \text{ such that } \sigma_{j_i}^2 \equiv \sigma^2 \text{ for all } j = 1, \ldots, k.$$

$$H_1 : \text{Not all } \sigma_{j_i}^2 \text{ are the same.}$$

(6)

Under the null hypothesis in Equation (6), the LRT statistic is asymptotically $\chi^2_{k-1}$. Note that here standard regularity conditions (see, for example, [8]) are satisfied and so the usual asymptotic results apply to the LRT; we do not have to worry about the difficulties sometimes associated with LRT statistics in a finite mixture model context when the test involves the number of components [6, 21, 25].

We also allow for $k$ different variance–covariance matrices in the mixture structure for the random effects. Gaffney and Smyth [12] did incorporate this into their model, but did not perform any testing for this heterogeneity. We again use an LRT for the hypothesis:

$$H_0 : \text{There exists } R \text{ such that } R_{j_i} \equiv R \text{ for all } j = 1, \ldots, k.$$

$$H_1 : \text{Not all } R_{j_i} \text{ are the same.}$$

(7)

Under the null hypothesis in Equation (7), the LRT statistic is approximately $\chi^2$ on $p(p + 1)(k - 1)/2$ degrees of freedom.

### 2.3 Determining the number of components

Some applications of mixture models specify the number of components based on scientific assumptions. For example, the clinical trial data sets analyzed in Xu and Hedeker [46] are analyzed with the random effects regression mixture model for identification of treatment ‘responders’ versus ‘nonresponders’. Hence, the $\beta_i$ are assumed to come from a two-component mixture of multivariate normals. However, when the number of components is not specified a priori, then one must determine an appropriate number of components to use.
From a likelihood perspective, we consider testing
\[ H_0 : k = k_0, \]
\[ H_1 : k = k_0 + 1, \]
for some positive integer \( k_0 \). Letting \( \hat{\psi}_0 \) and \( \hat{\psi}_1 \) denote the MLEs of \( \psi \) calculated under \( H_0 \) and \( H_1 \), respectively, we could consider the LRT statistic
\[ -2 \log \Delta = 2\{\ell(\hat{\psi}_1) - \ell(\hat{\psi}_0)\}. \]

It is well known that standard regularity conditions do not hold in the setting of Equation (8) and thus the asymptotic distribution of Equation (9) is not the usual chi-squared distribution (see [1,21] for a discussion). However, empirical results have shown that model selection techniques (e.g. the Akaike information criterion and Bayesian information criterion) and bootstrapping for testing the number of components both have good empirical results (see [25] for references). We use the latter approach for our analysis, which we now define in greater detail.

The bootstrapping approach for determining the number of components was proposed by McLachlan [23]. The algorithm is an attempt to approximate the null distribution of the LRT statistic values given in Equation (9). An outline of the algorithm is as follows:

1. Fit a mixture model with \( k_0 \) and \( k_0 + 1 \) components to the data, which leads to the estimates \( \hat{\psi}_0 \) and \( \hat{\psi}_1 \), respectively.
2. Calculate the (observed) log-likelihood ratio statistic in Equation (9). Denote this value by \( \Xi_{\text{obs}} \).
3. Simulate a data set of size \( n \) from the null distribution (the model with \( k_0 \) components).
4. Fit a mixture model with \( k_0 \) and \( k_0 + 1 \) components to the simulated data and calculate the corresponding bootstrap log-likelihood ratio statistic. Denote this value by \( \Xi^*_i \).
5. Repeat steps 3 and 4 \( B \) times to generate the bootstrap sampling distribution of the likelihood ratio statistic, \( \Xi^*_1, \Xi^*_2, \ldots, \Xi^*_B \).
6. Compute the bootstrap \( p \)-value as
\[ p_B = \frac{1}{B} \sum_{i=1}^{B} I\{\Xi_{\text{obs}} \geq \Xi^*_i\}. \]

We implement the aforementioned algorithm by first testing 1 versus 2 components. We obtain \( p_B \) for this test and if it is lower than some significance level \( \alpha \) (say, 0.05), we claim statistical significance and proceed to test 2 versus 3 components. If not, we stop and claim that there is not statistically significant evidence for a two-component fit. We proceed in this manner until we fail to reject the null hypothesis.

### 2.4 Some caveats

There are various difficulties and limitations that arise when estimating mixture models. For example, mixture densities having the same parametric families for the components (like the mixture of multivariate normals in Equation (3)) are invariant under the \( k! \) permutations of the component labels in the parameter vector; see McLachlan and Peel [25] for discussion. From an estimation standpoint, this can become a concern if performing a parametric bootstrap to obtain standard error estimates (as in our application) or if conducting Bayesian inference via Markov chain Monte Carlo samplers. Such a permutation of the component labels results in the well-known label switching problem. Fortunately, there are numerous methods in the literature for
dealing with label switching; see, for example, McLachlan and Peel [25] and Jasra et al. [17]. One common approach for handling this issue is to impose identifiability constraints as discussed in Aitkin and Rubin [1], who suggest restricting the mixing proportions to

$$\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_k.$$  \hspace{1cm} (11)

The aforementioned identifiability constraint is what is applied in our analysis in the next section.

The label switching problem tends to become less problematic and easier to correct when the components are well-separated. This was discussed in Hurn et al. [16] when performing Bayesian inference on finite mixtures of regressions models. Moreover, if the components of a mixture model are not well-separated, then this can also impact the ability to accurately test the number of components when not known a priori. McLachlan [23] showed how the power of the test in Equation (8) for $k_0 = 1$ is poor for the mixture of normals setting (with common variances) unless the components are well-separated.

It is also necessary to consider the possibility of multiple local maxima since the likelihood will have multiple roots. This can become an issue when testing the number of components as one risks making a decision based on convergence to a local maximum. When performing maximum likelihood estimation with an optimization algorithm (e.g. with EM algorithms), it is always a good practice to try multiple starting values and assess the reasonableness of the results. Moreover, one also needs to consider the practical interpretation of the results obtained. For example, when testing the number of components, one should consider if there is some additional research or theory that substantiates why the chosen number of components may be appropriate. For our infant habituation analysis, the components of our mixture model are used to model possible subgroups based on visual habituation trends. We will highlight the literature that suggests different subgroups of infants take different strategies when exposed to habituation experiments.

If using an EM algorithm for maximum likelihood estimation, then convergence of the algorithm can be slow. There are various methods that have been proposed for speeding up convergence of EM algorithms, many of which leverage Aitken’s acceleration method; see McLachlan and Krishnan [24] and McLachlan and Peel [25] for discussion of such methods. Unfortunately, the ECM algorithm we present in the appendix is also very slow to converge. We have not yet incorporated any methods for speeding up convergence of our ECM algorithm, but note that this is a direction for future research.

One final issue is that the likelihood function may be unbounded, which is also a concern when implementing numerical algorithms. Focusing on local maxima on the interior of the parameter space helps circumvent this problem because under certain regularity conditions, there exists a strongly consistent sequence of roots to the likelihood equation that is asymptotically efficient; see McLachlan and Peel [25]. Regarding numerical optimization, it is again suggested to use either informative starting values or try multiple starting values as a way to handle this situation.

### 3. Examples

In this section, we first analyze simulated data to highlight the relevant functions in the mixtools package for a mixed-effects regression mixture analysis. We perform a limited power analysis of the equal variances and equal variance–covariance matrices setting. The analysis of the simulated data is intended to highlight the general application of the procedures in Section 2. We then proceed to analyze the infant habituation data using a random effects regression mixture model. In the following analyses, we highlight some of the limitations discussed at the end of Section 2.
3.1 Simulated data

We begin by generating 50 trajectories from a two-component random effects regression mixture model with quadratic regressions and a fixed explanatory variable. The variables for the random effects, \( x_1, \ldots, x_{50} \), were generated independently from \( \text{Unif}(0, 10) \). The vectors were randomly assigned a length \( n_i \in \{5, 6, 7\} \), with each length selected with equal probability. Thus, the design matrix for the random effects of a trajectory is

\[
X_i = \begin{pmatrix} 1_{n_i} & x_i & x_i^2 \end{pmatrix},
\]

where \( x_i^2 \) is a vector consisting of the squared entries in \( x_i \). The explanatory variables for the fixed effect, \( u_1, \ldots, u_{50} \), were generated independently from the \( \text{Bin}(3, 0.5) \) distribution. The trajectories were then defined by

\[
y_i = u_i \alpha + X_i \beta_i + \epsilon_i, \quad (12)
\]

where \( \epsilon_i \sim N_{n_i}(0, \sigma_{ji}^2 I_{n_i}) \). The parameters for Equation (12) are \( \alpha = 6 \), \( \sigma_1 = 1 \), and \( \sigma_2 = 6 \), while the individual effects were generated according to

\[
\beta_i \sim \begin{cases} 
N_3 \begin{pmatrix} -20 \\ 0.1 \\ 6 \end{pmatrix}, & \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \right) 
\end{cases} \quad \text{with probability 0.2,} 
\]

\[
N_3 \begin{pmatrix} 40 \\ 1 \\ -0.1 \end{pmatrix}, & \begin{pmatrix} 10 & 0.2 & -1 \\ 0.2 & 6 & 2 \\ -1 & 2 & 1 \end{pmatrix} \right) \right)
\end{cases} \quad \text{with probability 0.8.} 
\]

We first ran two simulation studies to check the probability of a type I error when the data are generated according to each of the null hypotheses in Section 2.2. Each simulation was ran 500 times and the 0.05 significance level was used for each test. Each of these tests can be performed using the `test.equality.mixed` function. For the testing scenario in Equation (6), the data were generated according to the structure outlined earlier, but under the null hypothesis where \( \sigma_1^2 = \sigma_2^2 = 1 \). The proportion of times the null hypothesis was rejected is 0.0648. For the testing scenario in Equation (7), the data were also generated according to the structure outlined earlier, but this time under the null hypothesis where \( R_1 = R_2 = I_3 \). The proportion of times the null hypothesis was rejected for this simulation is 0.0440.

Next, a data set simulated using Equations (12) and (13) is used to illustrate the method of bootstrapping for the number of components. This bootstrapping procedure is implemented by using the `boot.comp` function. Plots of this data set, as well as a least-squares fit to each trajectory, can be found in Figure 1(a) and 1(b), respectively.

For testing \( k = 1 \) versus \( k = 2 \) for the number of components, the bootstrapped \( p \)-value is \( p_B = 0.000 \), suggesting strongly that \( k = 2 \) component is more appropriate than \( k = 1 \). However, for \( k = 2 \) versus \( k = 3 \), \( p_B = 0.280 \). Thus, the bootstrapping for the number of components procedure correctly identifies the two-component structure of these data. Moreover, the LRT statistics for testing the equality of variances of the two components and the equality of the variance matrices hypothesis are 108.1 (one df, \( p < 0.0001 \)) and 23.944 (six df, \( p = 0.001 \)), respectively. These results are consistent with the performance of the simulation presented earlier in this section.

The ECM algorithm is implemented by using `regmixEM.mixed` and the output can be found in Table 1. The observed log-likelihood for the two-component fit is \( \ell_o(\psi) = -1032.534 \). The standard errors for the ECM algorithm are calculated using \( B = 500 \) bootstrap samples according to a parametric bootstrapping scheme given in Efron and Tibshirani [11]. This is implemented via the `boot.se` function. Label switching did not appear to be present since the identifiability constraint \( \lambda_1 < \lambda_2 \) is always met despite never being enforced. Figure 2
Figure 1. (a) Plot of the simulated data. The different colors and plotting characters denote from which component the data were generated. (b) Plot of the same data with an ordinary least squares (OLS) line denoting each trajectory. The different colors and line style denote from which component the data were generated.

Table 1. The ECM estimates for the simulated data with bootstrap standard errors. This data set was simulated using the structure outlined in Equations (12) and (13).

| Parameter | Value (SE) | Parameter | Value (SE) |
|-----------|------------|-----------|------------|
| $\mu_1$   | -20.599 (0.397) | $\mu_2$ | 43.159 (0.939) |
|           | -0.131 (0.383) |          | 0.386 (0.484)  |
|           | 5.872 (0.323)  |          | 0.153 (0.148)  |
| $\sigma_1$| 1.007 (0.105)  | $\sigma_2$| 5.558 (0.151)  |
| $\lambda_1$| 0.220 (0.052) | $\lambda_2$| 0.780 (0.052)  |
| vech($R_1$)   | 0.153 (0.524) | vech($R_2$) | 21.805 (6.631) |
|           | -0.255 (0.486) |        | -6.231 (2.564)  |
|           | 0.124 (0.368)  |        | -2.634 (0.832)  |
|           | 1.041 (0.631)  |        | 6.616 (1.790)   |
|           | -0.552 (0.417) |        | 2.259 (0.522)   |
|           | 0.985 (0.426)  |        | 0.809 (0.187)   |
| $\alpha$  | 5.569 (0.212)  |          |              |

shows a histogram of $\lambda_2 - \lambda_1$ for each sample and all differences are positive and well above 0.

Figure 3 depicts two trajectories and their $(y_i, X_i)$ values from the simulated data set with both posterior regression lines plotted. The line generated using the most probable regression coefficients is denoted by a solid line. We have also plotted the regression lines that result from the posterior random effect regressors (the $\beta^{(t)}$ values). Since we fit a two-component model to these data, we have two sets of regressors. Furthermore, each set of regressors has a corresponding posterior membership probability and these weights typically indicate which regression line is a better fit for the trajectory.

Figure 4(a) shows all such regression lines for each trajectory and the corresponding regression coefficients are plotted in Figure 4(b). Figure 4(c) and 4(e) denotes the subsets of those regression lines whose posterior random effect regressors had the highest posterior membership probability
Figure 2. Histogram showing that the identifiability constraint $\lambda_1 < \lambda_2$ was satisfied for the simulated data.

Figure 3. Two trajectories from the first simulated data set. (a) The solid red line corresponds to most probable membership to the first component. (b) The solid green line corresponds to most probable membership to the second component. For each trajectory, the posterior membership probability for the chosen component was essentially 1.

for each component and the corresponding regression coefficients are plotted in Figure 4(d) and 4(f). Clearly, Figure 4(c)–(f) indicate a separation of the trajectories into two groups. Thus, the analysis we presented using random effects regression mixtures captured the structure of the simulated data, including the correct number of components, unequal variances, and unequal variance–covariance matrices.

The simulated data discussed thus far is ideal since the components are well-separated. However, any analysis of random (or mixed) effects regression mixture models is still prone to the issues discussed in Section 2.4. To demonstrate the effects on determining the number of components, consider now a simple random effects regression mixture model measured at
Figure 4. Plots for the first generated data set. (a) All regression lines using the posterior random effect regressors. (b) All posterior random effect regressors in $(\beta_1, \beta_2, \beta_3)$ space. The subsets of plot (a) and plot (b) with highest first component membership probabilities (plots (c) and (d)) and highest second component membership probabilities (plots (e) and (f)) are also given.

$x_i = (0, 2, 4, \ldots, 20)\top$ and where $k = 3$. Thus, each 11-dimensional trajectory is defined by

$$y_i = X_i \beta_i + \epsilon_i,$$  \hspace{1cm} (14)

where the design matrix is $X_i = (1, x_i)$ and $\epsilon_i \sim N(0, \sigma_{ji}^2 I_{11})$, for $j_i \in \{1, 2, 3\}$. The parameters for Equation (14) are $\sigma_1 = 1$, $\sigma_2 = 3$, and $\sigma_3 = 2$, while the individual effects are generated.
according to

\[
\beta_i \sim \begin{cases} 
N_2 \left( \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \tau_1^2 I_2 \right) & \text{with probability 0.2,} \\
N_2 \left( \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \tau_2^2 I_2 \right) & \text{with probability 0.3,} \\
N_2 \left( \begin{bmatrix} 3 \\ 0.1 \end{bmatrix}, \tau_3^2 I_2 \right) & \text{with probability 0.5.}
\end{cases}
\]  

(15)

We consider a well-separated setting where \((\tau_1, \tau_2, \tau_3) = (0.20, 0.19, 0.21)\) and a heavily overlapping setting where \((\tau_1, \tau_2, \tau_3) = (2.00, 1.99, 2.01)\). The regression lines and data generated using these parameters are given in Figure 5. Clearly, one can see how estimation routines with the heavily overlapping setting can be challenging. When testing the number of components using Equation (8) in the well-separated setting, we obtain bootstrap \(p\)-values of 0.000, 0.020,
and 0.574 for $k_0 = 1, 2,$ and 3, respectively. Thus, the correct number of components is determined. However, for the same test in the heavily overlapping setting, we obtain a bootstrap $p$-value of 0.960 for the simplest test of one component (i.e. no mixture structure) versus two components. Thus, we cannot detect the difference between components in the second setting. Such a difficulty in detecting the difference between components is consistent with the results of McLachlan [23].

3.2 Infant habituation data set

Psychological data sets can provide interesting examples for mixture modeling when theories of, say, cognition posit that human populations consist of discrete subgroups. For example, one might assume that different children employ qualitatively different strategies when solving a cognitive task, and any such differences may be detected by fitting a finite mixture model to data collected about the task [39]. Here, we consider an experiment concerning habituation among infants; the motivation will be to determine the appropriateness of a two-component random effects regression mixture model, that is, whether there is evidence for two distinct groups of infants that may then be presumed to possess differing abilities or to use differing strategies.

Habituation is a decrease in response times upon repeated stimulus presentations. Visual habituation studies in infants have attempted to predict later cognitive abilities in childhood; for example, see Slater [35], Hood et al. [14], and Colombo and Mitchell [7]. Much of this research asserts that there are common and stable mechanisms that underly both visual fixation behavior observed during infant habituation studies and later cognitive abilities measured in the child. Hence, it is of scientific interest to investigate possible subgroups of infants based on visual habituation results as this could provide insights into their cognitive development.

The data set involves $n = 47$ infants at 4 months of age. The infants were exposed to a checkerboard pattern on a computer screen and the time (in milliseconds) until they turned away was recorded. Since this is a habituation experiment, there should be a progressive decline of behavioral response observed. This was repeated for $l = 1, \ldots, 11$ trials on each infant. Thus, each infant will have an associated trajectory of length 11. Plots of the data are given in Figure 6.

For our model, the time until the infant turns away has been converted to the log scale, which we call $y_i$. The model of interest is Equation (2), where the design matrix for the random

![Figure 6. (a) Plot of the infant data. (b) Plot of the same data with an OLS line fitting each trajectory.](image)
effects is
\[
X_i = \begin{bmatrix}
1 & x_{i,1} & e^{-x_{i,1}^2} \\
1 & x_{i,2} & e^{-x_{i,2}^2} \\
\vdots & \vdots & \vdots \\
1 & x_{i,11} & e^{-x_{i,11}^2}
\end{bmatrix} = \begin{bmatrix}
1 & 1 & 1 \\
1 & 2 & e^{-1} \\
\vdots & \vdots & \vdots \\
1 & 11 & e^{-100}
\end{bmatrix},
\]
but no fixed effects are present. Notice the exponential term in the third column. This quantity incorporates some curvature into the estimation of the trajectories, which is an attempt to better capture the longer response times that occur before the infants habituate to the repeated stimulus. We also shift all exponents by one to reflect a peak at time \( t = 1 \) (rather than zero), where an infant should exhibit its longest response time since the stimulus is novel to the infant at that initial trial.

For determining the number of components, the test of \( k = 1 \) versus \( k = 2 \) yields \( p_B = 0.0000 \), while the test for \( k = 2 \) versus \( k = 3 \) yields \( p_B = 0.2360 \). Thus, we proceed with a two-component fit for the data. We also test for equal variances and equal variance–covariance matrices, which yield \( p \)-values of 0.0868 and 0.0002, respectively. Based on these results, the model we fit is
\[
y_i = X_i \beta_i + \epsilon_i, \quad (16)
\]
where conditional on component membership \( j_i \), the \( \epsilon_{i,k} \sim N(0, \sigma_k^2) \) are independent, \( i = 1, \ldots, 47 \). Furthermore, the random effects coefficients are assumed to be distributed as
\[
\beta_i \sim \begin{cases} 
N(\mu_1, \Sigma_1) & \text{with probability } \lambda, \\
N(\mu_2, \Sigma_2) & \text{with probability } 1 - \lambda.
\end{cases} \quad (17)
\]

The ECM algorithm output can be found in Table 2. The standard errors for the ECM algorithm are calculated using \( B = 500 \) bootstrap samples. Label switching was present in this bootstrap sample. This was diagnosed by first noting that the standard errors appeared to be fairly large. Since \( \sigma_1 \) and \( \sigma_2 \) appeared to be well-separated in the sample, we simply imposed the identifiability constraint of \( \sigma_2 < \sigma_1 \) in order to correct the label switching (see the histogram in Figure 7).

It is sometimes of scientific interest to estimate a shift in the response variable from 0, which can be done by subtracting a shift parameter \( \theta_i \) from the original response times before taking the

| Parameter | Value (SE) | Parameter | Value (SE) |
|-----------|-----------|-----------|-----------|
| \( \mu_1 \) | 10.176 (0.347) | \( \mu_2 \) | 9.863 (0.391) |
| \( \sigma_1 \) | 0.664 (0.075) | \( \sigma_2 \) | 0.987 (0.060) |
| \( \lambda_1 \) | 0.664 (0.075) | \( \lambda_2 \) | 0.987 (0.060) |
| vech(\( R_1 \)) | 0.595 (0.261) | vech(\( R_2 \)) | 0.065 (0.501) |
| \( \sigma_1 \) | 0.664 (0.075) | \( \sigma_2 \) | 0.987 (0.060) |
| \( \lambda_1 \) | 0.664 (0.075) | \( \lambda_2 \) | 0.987 (0.060) |
| vech(\( R_1 \)) | 0.595 (0.261) | vech(\( R_2 \)) | 0.065 (0.501) |
| \( \sigma_1 \) | 0.664 (0.075) | \( \sigma_2 \) | 0.987 (0.060) |
| \( \lambda_1 \) | 0.664 (0.075) | \( \lambda_2 \) | 0.987 (0.060) |
| vech(\( R_1 \)) | 0.595 (0.261) | vech(\( R_2 \)) | 0.065 (0.501) |
Figure 7. Histogram showing the difference $\sigma_1 - \sigma_2$ after imposing the identifiability constraint $\sigma_2 < \sigma_1$.

Figure 8. Plots of two infants’ times along with the most probable regression line (solid line) and least probable regression line (dashed line) according to the posterior membership probabilities of the ECM algorithm.

log. One method of estimating $\theta_i$ involves taking a specified percentage of the minimum time of each trajectory. Another possible method is advocated for lognormal data by Boswell et al. [5], which looks at a function of the 100$\alpha$th lower, 50th and 100$\alpha$th upper percentiles for a chosen $\alpha$. We attempted to estimate $\theta_i$ using these methods, but found no improvement in the fit of our model.

Figure 8 gives plots of regression lines that result from the posterior random effect regressors for two of the infants. The solid lines have the higher posterior membership probability and appear to be the better fit of the two lines. Figure 9(a) shows all such regression lines for each infant’s trajectory and the corresponding regression coefficients are plotted in Figure 9(b). Figure 9(c) and 9(e) shows the subsets of those regression lines whose posterior random effect regressors had the highest posterior membership probability for each component; the corresponding regression coefficients are plotted in Figure 9(d) and 9(f).
Figure 9. Plots for the infant data set. (a) All regression lines using the posterior random effect regressors. (b) All posterior random effect regressors in \((\beta_1, \beta_2, \beta_3)\) space. The subsets of plot (a) and plot (b) with highest first component membership probabilities (plots (c) and (d)) and highest second component membership probabilities (plots (e) and (f)) are also given.

Figure 9(c) and 9(e) gives a key to interpretation of the two groups that emerge from this analysis. The group of Figure 9(c) and 9(d) tends to exhibit a steeper drop in response times at the earlier occasions than the group of Figure 9(e) and 9(f). The second group appears to follow a more linear trend than the first group. The mixture structure on the random effects coefficients
provides an effective way to capture this characteristic and allows us to classify the infants using all of their measured data across time as opposed to, say, a single time point.

The results presented earlier are also consistent with many studies on infant habituation found in the literature. For example, Oakes [31] discusses how habituation experiments provide insights into the cognitive processes of infants. Subgroups with different patterns of habituation as in our analysis are likely because of inherent differences with infants’ abilities (e.g. memory and sensitivity to feature combinations) and possibly other mental processes. These two subgroups could, potentially, highlight overall differences in learning strategies. Moreover, Thomas and Gilmore [38] note that since infants are sufficiently different in their behavior, it is important that any modeling approach reflects these differences, which is captured using our random effects regression mixture model. Finally, while not measured with the data we analyzed, there could be some common features within the two identified subgroups regarding their neural responses; see Turk-Browne et al. [41] for a discussion of habituation in functional neuroimaging. Overall, our analysis provides potential evidence of two groups of infants with different patterns of habituation.

4. Discussion

The random effects regression mixture model allows one to search for latent group structure in data that comprise multiple independent regression observations. The model we present here allows for \( k \) different variance–covariance matrices in the mixture structure for the random effects. We also provide an extension to the models presented in Xu and Hedeker [46] and Gaffney and Smyth [12] by allowing the error variance terms to differ between the trajectories, thus allowing a more nuanced view of the infant data. Moreover, relative to Gaffney and Smyth [12] who consider this model from a Bayesian perspective, our approach allows for formal testing of various hypotheses regarding the level of heterogeneity among the mixture components. These tests are helpful in deciding upon an appropriate model in case studies, such as the infant data set considered here. Maximum likelihood is used for parameter estimation, which we accomplish using the ECM algorithm of Meng and Rubin [26].

As with any mixture model, the components should have a meaningful interpretation in the context of the data problem. Given the literature establishing the connection between habituation in infants and cognitive development as children, analysis using the random effects regression mixture model can provide a way of classifying infants that may already be demonstrating differing strategies or cognitive abilities. Moreover, the consideration of other functional forms for the mixture components (e.g. higher order polynomials) may reveal the need for a different number of components.

There are a couple of extensions to the random effects regression mixture model that could be of further interest for analyzing trajectory data like the infant habituation data. One extension to consider is a correlated data structure, like the models presented in Laird [18]. For example, we might consider \( \epsilon_i \sim N_{n_i}(0, \sigma^2 \Omega_i) \), where \( \Omega_i \) is an appropriately defined correlation matrix. This would be informative given that a correlated structure for habituation data might be a reasonable assumption.

Another extension might allow for each trajectory to have its own variance term. In other words, the \( \sigma^2_{ji} \) for the variance of the error terms in model (2) would be replaced by \( \sigma^2_i \). Given component membership (i.e. \( j_i \in \{1, \ldots, k\} \)), we have the following hypotheses of interest:

\[
H_0 : \text{Var}[y_i | X_i] = \sigma^2, \quad \forall i = 1, \ldots, n,
\]

\[
H_1 : \text{Var}[y_i | X_i] = \tau^2_{ji}, \quad \forall i = 1, \ldots, n,
\]

\[
H_2 : \text{Var}[y_i | X_i] = \sigma^2_i, \quad \forall i = 1, \ldots, n.
\] (18)
Notice that we write $H_1$ using $\tau_{ji}^2$ to avoid an abuse of notation. If we had written $H_1$ using $\sigma_{ji}^2$, then the parameter space of this hypothesis would include $\{\sigma_1^2, \ldots, \sigma_k^2\}$, while the parameter space of $H_2$ would include $\{\sigma_1^2, \ldots, \sigma_n^2\}$. In other words, the first $k$ variances in the parameter space under $H_2$ need not be equal to the $k$ variances in the parameter space under $H_1$.

Acknowledgements

The authors are grateful to two anonymous referees for helpful comments during the preparation of this article. We also wish to thank Hoben Thomas from the Department of Psychology, Pennsylvania State University, Arnold Lohaus from the Department of Psychology, University of Marburg, and the German Research Foundation (DFG) for providing the infant data set.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by National Science Foundation Award [SES-0518772].

Supplemental data

Supplemental data for this article can be accessed http://dx.doi.org/10.1080/02664763.2014.1000272.

References

[1] M. Aitkin and D.B. Rubin, *Estimation and hypothesis testing in finite mixture models*, J. R. Stat. Soc. Ser. B 47(1) (1985), pp. 67–75.
[2] H. Andruff, N. Carraro, A. Thompson, P. Gaudreau and B. Louvet, *Latent class growth modelling: a tutorial*, Tuter. Quant. Methods Psychol. 5(1) (2009), pp. 11–24.
[3] A. Baillo, A. Cuevas and R. Fraiman, *Classification for functional data*, in *Functional Data Analysis*, F. Ferraty and Y. Romain, eds., Oxford University Press, Oxford, 2010, pp. 259–297.
[4] T. Benaglia, D. Chauveau, D.R. Hunter and D.S. Young, *mixtools: an R package for analyzing finite mixture models*, J. Statist. Softw. 32(6) (2009), pp. 1–29. http://www.jstatsoft.org/v32/i06/.
[5] M.T. Boswell, J.K. Ord and G.P. Patil, *Normal and lognormal distributions as models of size*, in *Statistical Distributions in Ecological Work*, J.K. Ord, G.P. Patil, and C. Taillie, eds., International Co-operative Publishing House, Fairland, MD, 1979, pp. 3–157.
[6] H. Chen, J. Chen and J.D. Kalbfleisch, *A modified likelihood ratio test for homogeneity in finite mixture models*, J. R. Stat. Soc. Ser. B 63(1) (2001), pp. 19–29.
[7] J. Colombo and D.W. Mitchell, *Infant visual habituation*, Neurobiol. Learn. Mem. 92(2) (2009), pp. 225–234.
[8] H. Cramér, *Mathematical Methods of Statistics*, Princeton University Press, Princeton, NJ, 1946.
[9] A.P. Dempster, N.M. Laird and D.B. Rubin, *Maximum likelihood from incomplete data via the EM algorithm*, J. R. Stat. Soc. Ser. B 39(1) (1977), pp. 1–38.
[10] W.S. DeSarbo and W.L. Cron, *A maximum likelihood methodology for clusterwise linear regression*, J. Classification 5(2) (1988), pp. 249–282.
[11] B. Efron and R. Tibshirani, *An Introduction to the Bootstrap*, Chapman & Hall, London, 1993.
[12] S.J. Gaffney and P. Smyth, *Curve clustering with random effects regression mixtures*, in *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics*, C.M. Bishop and B.J. Frey, eds., Society for Artificial Intelligence and Statistics, FL, 2003.
[13] R. Gueorguieva, C. Mallinckrodt and J.H. Krystal, *Trajectories of depression severity in clinical trials of duloxetine*, Arch. Gen. Psychiat. 68(12) (2011), pp. 1227–1237.
[14] B.M. Hood, L. Murray, F. King, R. Hooper, J. Atkinson and O. Braddick, *Habituation changes in early infancy: longitudinal measures from birth to 6 months*, J. Reprod. Infant Psychol. 14(3) (1996), pp. 177–185.
[15] A.M. Hunter, B.O. Muthén, I.A. Cook and A.F. Leuchter, Antidepressant response trajectories and quantitative electroencephalography (QEEG) biomarkers in major depressive disorder, J. Psychiat. Res. 44(2) (2010), pp. 90–98.

[16] M. Hurn, A. Justel and C.P. Robert, Estimating mixtures of regressions, J. Comput. Graph. Stat. 12(1) (2003), pp. 55–79.

[17] A. Jasra, C.C. Holmes and D.A. Stephens, Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling, Stat. Sci. 20(1) (2005), pp. 50–67.

[18] R.A. Johnson and D.W. Wichern, Applied Multivariate Statistical Analysis, 6th ed., Prentice Hall, Upper Saddle River, NJ, 2007.

[19] N.M. Laird, Analysis of Longitudinal and Cluster-Correlated Data, NSF-CBMS Regional Conference Series in Probability and Statistics, Institute of Mathematical Statistics and the American Statistical Association, 2004.

[20] P.J. Lenk and W.S. DeSarbo, Bayesian inference for finite mixtures of generalized linear models with random effects, Psychometrika 65(1) (2000), pp. 93–119.

[21] B.G. Lindsay, Mixture Models: Theory, Geometry and Applications, NSF-CBMS Regional Conference Series in Probability and Statistics Vol. 5, Institute of Mathematical Statistics and the American Statistical Association, 1995.

[22] B. Louvet, P. Gaudreau, A. Menaut, J. Genty and P. Deneuve, Longitudinal patterns of stability and change in coping across three competitions: A latent class growth analysis, J. Sport Exercise Psychol. 29(1) (2007), pp. 100–117.

[23] G.J. McLachlan, On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture, J. R. Stat. Soc. Ser. C 36(3) (1987), pp. 318–324.

[24] G.J. McLachlan and T. Krishnan, The EM Algorithm and Extensions, Wiley, New York, 1997.

[25] G.J. McLachlan and D. Peel, Finite Mixture Models, Wiley, New York, 2000.

[26] X.-L. Meng and D.B. Rubin, Maximum likelihood estimation via the ECM algorithm: a general framework, Biometrika 80(2) (1993), pp. 267–278.

[27] K.L. Mengersen, C.P. Robert and D.M. Titterington, Mixtures: Estimation and Applications, Wiley, West Sussex, UK, 2011.

[28] B.O. Muthén and L.K. Muthén, Integrating person-centered and variable-centered analyses: growth mixture modeling with latent trajectory classes, Alcohol. Clin. Exp. Res. 24(6) (2000), pp. 882–891.

[29] L.K. Muthén and B.O. Muthén, MPlus User’s Guide, 6th ed., Muthén and Muthén, Los Angeles, CA, 1998–2011.

[30] D.S. Nagin, Group-Based Modeling of Development, Harvard University Press, Cambridge, 2005.

[31] L.M. Oakes, Using habituation of looking time to assess mental processes in infancy, J. Cognit. Dev. 11(3) (2010), pp. 255–268.

[32] M.L. Oxford, L.D. Gilchrist, D.M. Morrison, M.R. Gillmore, M.J. Lohr and S.M. Lewis, Alcohol use among adolescent mothers: Heterogeneity in growth curves, predictors, and outcomes of alcohol use over time, Prevent. Sci. 4(1) (2003), pp. 15–26.

[33] R.E. Quandt, A new approach to estimating switching regressions, J. Am. Stat. Assoc. 67(338) (1972), pp. 306–310.

[34] S.W. Raudenbush, Comparing personal trajectories and drawing causal inferences from longitudinal data, Annu. Rev. Psychol. 52 (2001), pp. 501–525.

[35] A. Slater, Can measures of infant habituation predict later intellectual ability? Arch. Dis. Child. 77(6) (1997), pp. 474–476.

[36] M. Stoolmiller, Synergistic interaction of child manageability problems and parent-discipline tactics in predicting future growth in externalizing behavior for boys, Dev. Psychol. 37(6) (2001), pp. 814–825.

[37] H. Thomas and M.P. Dahlin, Individual development and latent groups: analytical tools for interpreting heterogeneity, Dev. Rev. 25(2) (2005), pp. 133–154.

[38] H. Thomas and R.O. Gilmore, Habituation assessment in infancy, Psychol. Methods 9(1) (2004), pp. 70–92.

[39] H. Thomas and J.I. Horton, Competency criteria and the class inclusion task: modeling judgments and justifications, Dev. Psychol. 33(6) (1997), pp. 1060–1073.

[40] D.M. Titterington, A.F.M. Smith and U.E. Makov, Statistical Analysis of Finite Mixture Distributions, Wiley, New York, 1985.

[41] N.B. Turk-Browne, B.J. Scholl and M.M. Chun, Babies and brains: habituation in infant visual processing and functional neuroimaging, Front. Hum. Neurosci. 2(16) (2008), pp. 1–11.

[42] P.A.C. van Lier, B.O. Muthén, R.M. van der Sar and A.A.M. Crijnen, Preventing disruptive behavior in elementary schoolchildren: impact of a universal classroom-based intervention, J. Consult. Clin. Psychol. 72(3) (2004), pp. 467–478.

[43] M.G. Vaughn, M. DeLisi, K.M. Beaver and M.O. Howard, Multiple murder and criminal careers: a latent class analysis of multiple homicide offenders, Forensic Sci. Int. 183(1–3) (2009), pp. 67–73.

[44] G. Verbecke and E. Lesaffre, A linear mixed-effects model with heterogeneity in the random-effects population, J. Am. Stat. Assoc. 91(433) (1996), pp. 217–221.

[45] M. Wedel and W.S. DeSarbo, A review of recent developments in latent class regression models, in Advanced Methods of Marketing Research, R. Bagozzi, ed., Blackwell, London, 1994, pp. 352–388.
Appendix. Details of ECM algorithm

From the conditional distribution of $y_i$, given component membership in Equation (5), we obtain the (observed-data) log-likelihood function

$$
\ell_c(\psi) = \log \left( \prod_{i=1}^{n} \prod_{j=1}^{k} \left( \frac{1}{\sqrt{2\pi} \sigma_j} \exp \left( -\frac{1}{2\sigma_j^2} \| y_i - U_i \alpha - X_i \beta_j \|^2 \right) \right) \phi_{\psi}(\beta_j | \mu_j, R_j) \right)^{Z_{ij}}.
$$

As is standard practice in finite mixture model EM algorithms, we let $Z_{ij} = I(y_i = j)$ denote the (unobserved) indicator that the $i$th observation comes from the $j$th component. The unobservable $\beta$ are also treated as missing, so $(Z, \beta)$ constitutes the missing data for this problem.

An EM algorithm consists of two alternating steps, repeated until convergence: the E-step and the M-step. An ECM algorithm operates on only part of the parameter vector at once. In other respects, though, it is just like an EM algorithm: During each E-step, the expectation of the complete-data log-likelihood function is calculated. In the corresponding CM-step, some part of the parameter vector is held fixed (this is the ‘conditioning’) while the expected log-likelihood is maximized in the other parameters. This process repeats with different subsets of the parameters until finally, each parameter has been updated at least once, and the ECM iteration is complete.

In the E-step for our model, we form the conditional expectation, given the observed $y_1, \ldots, y_n$, of the complete-data log-likelihood function

$$
\ell_c(\psi) = \log \left( \prod_{i=1}^{n} \prod_{j=1}^{k} \left( \frac{1}{\sqrt{2\pi} \sigma_j} \exp \left( -\frac{1}{2\sigma_j^2} \| y_i - U_i \alpha - X_i \beta_j \|^2 \right) \right) \phi_{\psi}(\beta_j | \mu_j, R_j) \right)^{Z_{ij}}.
$$

For iteration $t, t = 0, 1, \ldots$, we take the expectation of $\ell_c(\psi)$ with respect to the conditional density

$$
p(Z, \beta | U_i, X_i, y_i, \psi^{(t)}) = p(Z | U_i, X_i, y_i, \psi^{(t)}) p(\beta | Z, U_i, X_i, y_i, \psi^{(t)}),
$$

where $\psi$ is defined in Equation (4). Recalling that the $Z_{ij}$ are marginally Bernoulli distributed with parameter $\lambda_j$, we have

$$
Z_{ij}^{(t)} \overset{\text{def}}{=} P(Z_{ij} = 1 | U_i, X_i, y_i, \psi^{(t)}) = \frac{\lambda_j \phi_{\psi}(y_i | U_i \alpha^{(t)} + X_i \beta_j^{(t)}; R_j \beta_j^{(t)} + \sigma_j^{2(t)} I_n)}{\sum_{j=1}^{k} \lambda_j \phi_{\psi}(y_i | U_i \alpha^{(t)} + X_i \beta_j^{(t)}; R_j \beta_j^{(t)} + \sigma_j^{2(t)} I_n)}.
$$

The conditional density for $\beta_j$ in Equation (A2) may be expressed up to a constant of proportionality not involving $\beta$, as

$$
p(y_i | \beta_j, Z_i, U_i, X_i, \psi^{(t)}) \propto \exp \left\{ -\frac{1}{2\sigma_j^2} \| y_i - U_i \alpha - X_i \beta_j \|^2 \right\} \exp \left\{ -\frac{1}{2} R_j^{(t)} (\beta_j - \mu_j^{(t)}) \| (\beta_j - \mu_j^{(t)}) \right\}.
$$

where, after some algebra, we obtain

$$
V_{ij}^{(t)} = \left( R_j^{(t)} + \frac{1}{\sigma_j^2} X_i^T X_i \right)^{-1}
$$
and

$$\beta_{ij}^{(t)} = \frac{1}{\sigma_{ij}^{2(t)}} V_{ij}^{(t)} \left( R_{ij}^{(t)} - 1 \right) \mu_{ij}^{(t)} + \frac{1}{\sigma_{ij}^{2(t)}} X_{ij}^{(t)} (y_i - U_i \alpha - X_i \mu_{ij}^{(t)}) + \mu_{ij}^{(t)} \quad (A6)$$

as the conditional variance and mean of the multivariate normal conditional distribution of $\beta_{ij}$ given that the $i$th observation comes from the $j$th component. [In Equation (A4), we know the component $j$ because we are conditioning on $Z_i$ in that equation.] Notice that we applied Bayes’ theorem in deriving Equation (A4) and combined the complete data and an estimate of $\psi$ to approximate this density. Hence, $\beta_{ij}^{(t)}$ is simply an empirical Bayes estimate of the random effects. Note that $\beta_{ij}$ will not always resemble the parameters one would obtain from ordinary least squares regression; indeed, these parameters can be quite different, particularly when $\hat{Z}_j$ is a small probability.

Using the joint distribution of $(Z, \beta)$ given $(U, X, y, \psi^{(t)})$ as summarized by Equations (A2)–(A6), we may find the expectation of the terms in the complete-data log-likelihood function (A2), as required in the E-step of our ECM algorithm, using several well-known identities involving multivariate normal distributions. To wit, the only problematic expectations are

$$E \left[ \sum_{i=1}^{n} \sum_{j=1}^{k} \frac{Z_{ij}}{2\sigma_{ij}^2} | y_i - U_i \alpha - X_i \beta_j |^2 | U_i, X_i, y_i \right]$$

$$= E \left[ E \left[ \sum_{i=1}^{n} \sum_{j=1}^{k} \frac{Z_{ij}}{2\sigma_{ij}^2} | y_i - U_i \alpha - X_i \beta_j |^2 | Z_i, U_i, X_i, y_i \right] \right]$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{k} Z_{ij} E \left[ \beta_j^{(t)} | Z_i, U_i, X_i, y_i \right]$$

and, similarly,

$$E \left[ \sum_{i=1}^{n} \sum_{j=1}^{k} Z_{ij} | R_j^{(t)} - 1/2 (\beta_j^{(t)} - \mu_j)^2 | Z_i, U_i, X_i, y_i \right]$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{k} Z_{ij} E \left[ \beta_j^{(t)} | R_j^{(t)} - 1/2 (\beta_j^{(t)} - \mu_j)^2 + \text{tr}(R_j^{(t)} V_{ij}^{(t)}) \right] \quad (A7)$$

and

$$E \left[ \sum_{i=1}^{n} \sum_{j=1}^{k} Z_{ij} | y_i - U_i \alpha - X_i \beta_j |^2 | U_i, X_i, y_i \right]$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{k} Z_{ij} E \left[ \beta_j^{(t)} | y_i - U_i \alpha - X_i \beta_j |^2 \right]$$

$$= \sum_{i=1}^{n} \sum_{j=1}^{k} Z_{ij} E \left[ \beta_j^{(t)} | y_i - U_i \alpha - X_i \beta_j |^2 + \text{tr}(X_i^T X_i V_{ij}^{(t)}) \right] \quad (A8)$$

where we have used the facts that $\text{tr}(X_i^T X_i V_{ij}^{(t)}) = \text{tr}(X_i^T X_i V_{ij}^{(t)})$ and if $W \sim N_p(\nu, \Sigma)$, then $E \|W\|^2 = \sum_{i=1}^{p} \nu_i^2 + \text{tr}(\Sigma)$.

Thus, the expectation of the complete-data log-likelihood, which must be maximized as a function of $\psi$, equals

$$-\frac{N}{2} \log(2\pi) - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{k} \hat{Z}_{ij}^{(t)} n_i \log(\sigma_{ij}^2)$$

$$- \sum_{i=1}^{n} \sum_{j=1}^{k} \hat{Z}_{ij}^{(t)} \left[ | y_i - U_i \alpha - X_i \beta_j^{(t)} |^2 + \text{tr}(X_i^T X_i V_{ij}^{(t)}) \right]$$

$$+ \sum_{i=1}^{n} \sum_{j=1}^{k} Z_{ij} \hat{\nu}_j - \frac{np}{2} \log(2\pi) - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{k} Z_{ij}^{(t)} \log(|R_j|)$$

$$- \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{k} Z_{ij}^{(t)} \left[ | R_j^{(t)} - 1/2 (\beta_j^{(t)} - \mu_j)^2 + \text{tr}(R_j^{(t)} V_{ij}^{(t)}) \right] \quad (A9)$$

The first CM-step of the ECM algorithm maximizes the aforementioned expectation over all parameters except $\alpha$. As with any mixture model EM (or EM-type) algorithm, only a single term of Equation (A9) involves the mixing proportions $\lambda$, and this term is maximized by

$$\lambda_j^{(t+1)} = \frac{1}{n} \sum_{i=1}^{n} \hat{Z}_{ij}^{(t)}.$$

For the means of the random effects, we obtain

$$\mu_j^{(t+1)} = \frac{\sum_{i=1}^{n} \hat{Z}_{ij}^{(t)} \beta_j^{(t)}}{\sum_{i=1}^{n} \hat{Z}_{ij}^{(t)}}.$$
Next, since we are dealing with component-specific error terms as stated in Equation (2), we get
\[
\sigma_j^{2(t+1)} = \frac{\sum_{i=1}^{n} \hat{Z}_{ij}^{(t)} \parallel y_i - U_i \alpha^{(t)} - X_i \beta^{(t)}_j \parallel^2 + \text{tr}(X_i^\top X_i \beta^{(t)}_j)}{\sum_{i=1}^{n} \hat{Z}_{ij}^{(t)} n_i}.
\]

Alternatively, if we assume the same variance for each component, we obtain
\[
\sigma^{2(t+1)} = \frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{n} \hat{Z}_{ij}^{(t)} \parallel y_i - U_i \alpha^{(t)} - X_i \beta^{(t)}_j \parallel^2 + \text{tr}(X_i^\top X_i \beta^{(t)}_j) = \frac{\sum_{j=1}^{k} \left( \sum_{i=1}^{n} \hat{Z}_{ij}^{(t)} n_i \right) \sigma_j^{2(t+1)}}{N}.
\]

Finally, \(R_j^{(t+1)}\) may be found by using the following fact, found for instance in [18]: Given a \(p \times p\) symmetric positive definite matrix \(A\) and a scalar \(a > 0\), it follows that
\[
-a \log |\Sigma| - \frac{1}{2} \text{tr}(\Sigma^{-1} A) \leq -a \log |A| + pa \log(2a) - pa
\]
for all positive definite \(p \times p\) matrices \(\Sigma\), with equality holding only for \(\Sigma = (1/2a)A\). Focusing on the terms in Equation (A9) that depend on \(R_j\), we have
\[
-\frac{1}{2} \sum_{i=1}^{n} \hat{Z}_{ij}^{(t)} \log(|R_j|) - \frac{1}{2} \sum_{i=1}^{n} \hat{Z}_{ij}^{(t)} \parallel \beta_j^{(t)} - \mu^{(t+1)}_j \parallel^2 + \text{tr}(R_j^{-1} V^{(t)}_j)
\]
\[
= -\frac{1}{2} \sum_{i=1}^{n} \hat{Z}_{ij}^{(t)} \log(|R_j|) - \frac{1}{2} \text{tr} \left[ R_j^{-1} \sum_{i=1}^{n} \hat{Z}_{ij}^{(t)} (\beta_j^{(t)} - \mu^{(t+1)}_j)(\beta_j^{(t)} - \mu^{(t+1)}_j) + V^{(t)}_j \right].
\]

We conclude that
\[
R_j^{(t+1)} = \frac{\sum_{i=1}^{n} \hat{Z}_{ij}^{(t)} (\beta_j^{(t)} - \mu^{(t+1)}_j)(\beta_j^{(t)} - \mu^{(t+1)}_j) + V^{(t)}_j}{\sum_{i=1}^{n} \hat{Z}_{ij}^{(t)}}.
\]

or, if we assume the same \(R\) matrix for each component,
\[
R^{(t+1)} = \frac{1}{N} \sum_{j=1}^{k} \sum_{i=1}^{n} \hat{Z}_{ij}^{(t)} (\beta_j^{(t)} - \mu^{(t+1)}_j)(\beta_j^{(t)} - \mu^{(t+1)}_j) + V^{(t)}_j = \sum_{j=1}^{k} \left( \frac{\sum_{i=1}^{n} \hat{Z}_{ij}^{(t)}}{N} \right) R_j^{(t+1)}.
\]

In the second E-step, calculate \(\hat{Z}_{ij}^{(t+1/2)}\), \(V^{(t+1/2)}\), and \(\beta_j^{(t+1/2)}\) as in Equations (A3), (A5), and (A6), respectively, but using the values \(\mu^{(t+1)}_j\), \(R_j^{(t+1)}\), and \(\sigma_j^{2(t+1)}\). For the second CM-step, calculate the update for the fixed effects regression coefficients as
\[
\alpha^{(t+1)} = \left[ \sum_{i=1}^{n} \sum_{j=1}^{k} \hat{Z}_{ij}^{(t+1/2)} U_i U_i^\top \right]^{-1} \left[ \sum_{i=1}^{n} \sum_{j=1}^{k} \hat{Z}_{ij}^{(t+1/2)} U_i (y_i - X_i \beta_j^{(t+1/2)}) \right].
\]

When maximizing Equation (A9) with respect to \(\sigma^2\), the formula involves \(\alpha\), and vice versa. Hence why we need to condition on the updates and appeal to an ECM algorithm. On the other hand, if \(\sigma^2 \equiv \sigma^2\), then the formula for \(\alpha^{(t+1)}\) simplifies and is free of \(\sigma^2\). Hence \(\alpha^{(t+1)}\) would be calculated under the first CM-step with the remaining parameters, thus collapsing to a regular EM algorithm.

The ECM algorithm is iterated until the stopping criterion \(\ell_v(\psi^{(t+1)}) - \ell_v(\psi^{(t)}) < \epsilon\) is attained for some fixed small \(\epsilon > 0\).