Latent Class Linear Mixed Models -
A General Approach Implemented via SAS® Macro
with a Tutorial for Clinical Researchers

Steven W. Enck

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Public Health in the Department of Biostatistics.

Chapel Hill
2009

Approved by:

Paul Stewart, Advisor

Lloyd Edwards, Reader

Gerardo Heiss, Reader

Gary Koch, Reader

John Preisser, Reader

Russ Wolfinger, Reader

Abstract

STEVEN W. ENCK: Latent Class Linear Mixed Models -
A General Approach Implemented via SAS® Macro
with a Tutorial for Clinical Researchers
(Under the direction of Paul Stewart)

Linear mixed models provide a flexible, intuitive method for analyzing repeated-measures data when the population being studied can be thought of as either a single population or as a set of known subpopulations. However, in many cases, the underlying subpopulations are not known. Furthermore, the factors that determine the subpopulations can be extremely complex or unmeasurable. In such cases, a different approach is required in order to more accurately analyze the data. The Latent Class Linear Mixed Model (LCLMM) combines the features of the linear mixed model (LMM) with an additional component, which partitions the population into subpopulations or latent classes. This model has usually been specified with relatively simple, restrictive assumptions.

In this dissertation, the methods related to the LCLMM are extended to provide a more general model specification. Fixed-effects may be specified as a combination of class-specific effects and across-class effects. Variances may be specified as being class-specific or equal across classes, a general correlation structure for the random effects is permitted, and multiple residual error variances may be fit. The bound proposed by Hathaway [1985] on the variances to ensure consistency is examined in the context of mixtures of linear mixed models. Class membership probabilities may be specified in one of two ways - via a logistic regression model or using our proposed method in which class membership is estimated based on the relative fit of the underlying linear mixed models. These methods are implemented in a new SAS® macro which offers several options for estimation. In addition to an EM algorithm, gradient-based methods, including quasi-Newton, as well as Hessian-based methods, such as Newton-Raphson, are available to the user. Parameter standard errors are estimated, and predictions of the random effects are derived and calculated. Practical issues, including choosing the number of latent classes and estimation method, are discussed and guidelines are provided based on simulation studies. The stability and advantage of the proposed methods are also examined via simulation study. Finally, our methods are applied to several simple simulated datasets as well as to three real-world

applications to illustrate their usefulness for practical applications.

Table of Contents

List of Tables

List of Figures

Chapter 1

Latent Class Linear Mixed Models - A General Approach

1.1 Introduction

Linear mixed models provide a flexible, intuitive method for analyzing repeated-measures data when the population being studied can be thought of as either a single population or as a set of known subpopulations. However, in many cases, the underlying subpopulations are not known. Furthermore, the factors that determine the subpopulations can be extremely complex or unmeasurable. In such cases, a different approach is required in order to more accurately analyze the data. The Latent Class Linear Mixed Model (LCLMM) combines the features of the linear mixed model (LMM) with an additional component which partitions the population into subpopulations or latent classes. In the existing literature, this model has usually been specified with relatively simple, restrictive assumptions. A single residual error variance and either a diagonal or unstructured variance-covariance matrix of the random effects is typical, with variances usually assumed to be equal across latent classes. In addition, the population is typically assumed to comprise a simple mixture of latent classes. Some models have allowed a structure for class membership represented by a logistic regression model. Subject-specific predictions, a valuable feature of linear mixed models, have not been examined in the context of the LCLMM. Finally, computations have for the most part focused on the use of an EM algorithm, which is known to be relatively slow to converge, and software is not currently available for statisticians to apply the methods.

In this chapter, the methods related to the LCLMM are extended to provide a more general model specification. Fixed-effects may be specified as a combination of class-specific effects and across-class effects. Variances may be specified as being class-specific or equal across classes, a general correlation structure for the random effects is permitted, and multiple residual error variances may be fit. The bound proposed by Hathaway [1985] on the variances to ensure consistency is examined in the context of mixtures of linear mixed models. Class membership probabilities may be specified in one of two ways - via a logistic regression model or using our proposed method in which class membership is estimated based on the relative fit of the underlying linear mixed models. First- and second-derivatives are presented for use in gradient and Hessian-based algorithms. Predictions of the random effects for the LCLMM are derived and calculated. Finally, these methods are applied to lipid data from the Atherosclerosis Risk in Communities (ARIC) study (see ARIC Investigators [1989]).

1.2 Defining the Latent Class Linear Mixed Model

Before examining the literature, it is helpful to first define the model which will be examined throughout this chapter. As noted earlier, the LMM allows for a logical specification of both the means and variances. It is a LMM that provides the basis for the LCLMM. Searle et al [1992] provides a detailed presentation of linear mixed models. A brief review is provided here. For subject i, we assume

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \mathbf{e}_i \tag{1.1}$$

where:

$\mathbf{y}_i$ is an $n_i \times 1$ vector of observations

$\mathbf{X}_i$ is an $n_i \times p_1$ design matrix for the fixed effects

$\boldsymbol{\beta}$ is a $p_1 \times 1$ unknown vector of fixed effects

$\mathbf{Z}_i$ is an $n_i \times q$ design matrix for the random effects

$\mathbf{b}_i$ is a $q \times 1$ unknown vector of random effects

$\mathbf{e}_i$ is an $n_i \times 1$ unknown vector of random error terms

Further, $\mathbf{e}_i$ and $\mathbf{b}_i$ are assumed to be mutually independent of one another and to have the following properties:

$$\begin{aligned}
\mathrm{E}\left(\mathbf{b}_i\right) &= \mathbf{0} \\
\mathrm{Var}\left(\mathbf{b}_i\right) &= \mathbf{D} \\
\mathrm{Cov}\left(\mathbf{b}_i, \mathbf{b}_h'\right) &= \mathbf{0} \text{ for } i \neq h \\
\mathrm{E}\left(\mathbf{e}_i\right) &= \mathbf{0} \\
\mathrm{Var}\left(\mathbf{e}_i\right) &= \mathbf{R}_i \\
\mathrm{Cov}\left(\mathbf{b}_i, \mathbf{e}_i\right) &= \mathbf{0}
\end{aligned} \tag{1.2}$$

And finally, the following distributional properties are assumed to apply to the random effects, $\mathbf{b}_i$, and residuals, $\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}$:

$$\begin{bmatrix} \mathbf{b}_i \\ \mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} \end{bmatrix} \sim N\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{D} & \mathbf{D}\mathbf{Z}_i' \\ \mathbf{Z}_i\mathbf{D} & \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i' + \mathbf{R}_i \end{bmatrix}\right) \tag{1.3}$$

The latent class model adds a further dimension in that each subject's data is modeled as a mixture of K LMMs, one for each of the K latent classes. Note that the value K is specified by the statistician a priori. Details related to selecting K are discussed in Section 2.8. Following the notation in Lin et al [2002a], define for subject i:

$$c_{ik} = \begin{cases} 1 & \text{if subject } i \text{ is a member of class } k \\ 0 & \text{if subject } i \text{ is NOT a member of class } k \end{cases}$$
$$c_{i1}, \ldots, c_{iK} \sim \text{ Multinomial } \left(1, \pi_{i1}, \ldots, \pi_{iK}\right)$$

For simplicity in later derivations, also define $\mathbf{c}_i = \begin{bmatrix} c_{i1} & c_{i2} & \cdots & c_{iK} \end{bmatrix}'$.

The π_{ik}, the multinomial probabilities of being in each latent class, are modeled via a logit model as follows:

$$\pi_{ik} = P\left(c_{ik} = 1 \mid \mathbf{t}_i\right) = \frac{\exp\left(\mathbf{t}_i'\boldsymbol{\alpha}_k\right)}{\sum_{j=1}^{K} \exp\left(\mathbf{t}_i'\boldsymbol{\alpha}_j\right)} \tag{1.4}$$

where:

$\mathbf{t}_i$ is the design vector related to class membership for subject i

$\boldsymbol{\alpha}_k$ is an unknown vector of class-membership parameters for class k with $\boldsymbol{\alpha}_1 = \mathbf{0}$

Further, given that subject i is in class k, define the LMM for subject i as follows:

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{W}_i\boldsymbol{\lambda}_k + \mathbf{Z}_i\mathbf{b}_i + \mathbf{e}_i \tag{1.5}$$

where the following additional definitions are provided:

$\mathbf{W}_i$ is an $n_i \times p_2$ design matrix for the class-specific fixed effects

$\boldsymbol{\lambda}_k$ is a $p_2 \times 1$ unknown vector of class-specific fixed effects for class k

It is useful to note that the parameters in $\boldsymbol{\beta}$ will apply to all subjects through the values of the corresponding column in $\mathbf{X}_i$. The class-specific parameters, $\boldsymbol{\lambda}_k$, however, are different for each latent class.

The error terms and random effects are assumed to have the same properties as in Equation 1.2. Finally, similar to the LMM, the following distributional properties are assumed to apply to the random effects, $\mathbf{b}_i$, and residuals, $\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{W}_i\boldsymbol{\lambda}_k$:

$$\begin{bmatrix} \mathbf{b}_i \\ \mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{W}_i\boldsymbol{\lambda}_k \end{bmatrix} \sim N\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{D} & \mathbf{D}\mathbf{Z}_i' \\ \mathbf{Z}_i\mathbf{D} & \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i' + \mathbf{R}_i \end{bmatrix} \right) \tag{1.6}$$

Lin et al [2002a] creates a further construct which defines a matrix $\mathbf{M}$ as

$$\mathbf{M} = \begin{bmatrix} \vdots & \vdots & & \vdots \\ \boldsymbol{\lambda}_1 & \boldsymbol{\lambda}_2 & \cdots & \boldsymbol{\lambda}_K \\ \vdots & \vdots & & \vdots \end{bmatrix}$$

Recalling the definition of $\mathbf{c}_i$, the $\mathbf{y}_i$ could be written succinctly as:

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{W}_i\left(\mathbf{M}\mathbf{c}_i\right) + \mathbf{Z}_i\mathbf{b}_i + \mathbf{e}_i$$

where the subject's class membership and corresponding $\boldsymbol{\lambda}_k$ are accounted for in the term $(\mathbf{M}\mathbf{c}_i)$.

It is assumed, without loss of generality, that the model is full-rank. This assumption requires that both of the following are true:

$$\begin{bmatrix} \mathbf{t}_1 & \mathbf{t}_2 & \cdots & \mathbf{t}_n \end{bmatrix} \text{ is full rank, and}$$

$$\left[\begin{array}{c|cccc|cccc} \mathbf{X}_1 & \mathbf{W}_1 & . & . & . & \mathbf{Z}_1 & . & . & . \\ \mathbf{X}_2 & . & \mathbf{W}_2 & . & . & . & \mathbf{Z}_2 & . & . \\ \vdots & . & . & \ddots & . & . & . & \ddots & . \\ \mathbf{X}_n & . & . & . & \mathbf{W}_n & . & . & . & \mathbf{Z}_n \end{array}\right] \text{ is full rank.}$$

For notational convenience, denote $\mathbf{\Theta}$ as the combined parameter vector comprised of $\boldsymbol{\alpha}_2 \ldots \boldsymbol{\alpha}_K$, $\boldsymbol{\beta}$, $\boldsymbol{\lambda}_1 \ldots \boldsymbol{\lambda}_K$, and $\boldsymbol{\theta}$, where $\boldsymbol{\theta}$ contains the unique variance components which determine $\mathbf{D}$ and $\mathbf{R}_i$.

It is useful to compare the observed-data likelihoods of the usual LMM with the LCLMM. These likelihoods are conditional on having observed the values $\mathbf{X}_i$, $\mathbf{W}_i$, $\mathbf{Z}_i$, and $\mathbf{t}_i$. In short,

$$\log L\left(\mathbf{\Theta}\right)_{LMM} = \sum_{i=1}^{n} \log f\left(\mathbf{y}_i\right) \tag{1.7}$$

$$\log L\left(\mathbf{\Theta}\right)_{LCLMM} = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \pi_{ik} f\left(\mathbf{y}_i \mid c_{ik} = 1\right) \tag{1.8}$$

with $f\left(\cdot\right)$ being the density defined by

$$\begin{aligned} (\mathbf{y}_i) &\sim N(\mathbf{X}_i\boldsymbol{\beta}, \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i' + \mathbf{R}_i) \\ (\mathbf{y}_i \mid c_{ik} = 1) &\sim N(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{W}_i\boldsymbol{\lambda}_k, \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i' + \mathbf{R}_i) \end{aligned}$$

Several points are of note. First, note the difference between Equations 1.7 and 1.8 - the likelihood for the LCLMM is a weighted average of K LMMs. Second, with K=1, and therefore $\pi_{i1} = 1$, the LCLMM reduces to the usual LMM. Third, if the groups are so well separated that each subject has one $\pi_{ik} = 1$ and the others equal to 0, then the likelihood reduces to that of a LMM with subjects assigned to these groups. Finally, note that the log likelihood for the LCLMM contains the log of the sum over latent classes, making the computations of the first and second derivatives of the log likelihood more complicated.

1.3 Literature Review

With an understanding of the LCLMM in hand, attention turns toward previous research related to these models as well as comparable models in the statistical literature.

For a detailed historical account of maximum likelihood estimation methods used for mixture models, the reader is referred to Redner and Walker [1984]. This article gives particular attention to the EM algorithm and its application to mixtures of densities from exponential families. In addition, Bohning and Seidel [2003] provided a review of more recent developments in mixture models.

A variety of articles have been published which examine latent class models in the context of the LMM. Most have utilized an EM algorithm in calculations, and many have also included a second endpoint, either categorical or time-to-event, which was modeled simultaneously with the repeated measures. Lin et al [2000] proposed a latent class model to jointly model longitudinal data and a categorical event outcome. The longitudinal submodel was a LMM which included random effects with an unstructured variance-covariance matrix. In addition, a single residual error variance was fit. Calculations were performed using a generalized EM algorithm. This model is somewhat similar to that presented in Lin et al [2002a]. However, the second endpoint in the latter article was a survival endpoint which was modeled using a frailty model. In both articles, the mixed model had variances which were modeled as being equal across classes, and class membership was modeled via a logit model, with subjects having unique class-membership probabilities based on their covariates. Lin et al [2004] replaced the survival submodel in the previous article with a multiplicative intensity model to describe a visit process. Lin et al [2002b] proposed a joint analysis of time-to-event data with multiple longitudinal variables via a mixture model. The longitudinal submodel was based on the LMM and allowed for different residual error variances for each longitudinal variable, as well as the inclusion of random effects with an unstructured variance-covariance matrix. However, variances were assumed to be equal across classes. Finally, McCulloch [2003] proposed to jointly model three endpoints simultaneously using a latent class model - a survival endpoint, longitudinal endpoint, and a repeated multivariate binary endpoint. The longitudinal portion of the model included random effects, with one random effect variance and one residual error variance computed across latent classes. The EM algorithm was once again utilized to obtain parameter estimates.

While the articles mentioned above included a second or third endpoint, the next set of articles focused exclusively on generalizing the LMM to simultaneously estimate group membership. Roy

[2003] used a pattern mixture model to model longitudinal data with nonignorable dropouts. Class membership was modeled via an ordinal logistic regression model with length of time in the study as the only covariate. Estimates were calculated using a modified Newton-Raphson algorithm with a multiplier applied to the change in estimates at each iteration. Second derivatives were calculated numerically rather than analytically, and class-specific variances were not considered. The model did include random effects. Celeux et al. [2005] proposed to use a mixture of LMMs to model the clustering of gene expression data. Three simple variance-covariance structures were specified, and each model was described as having either

- variances equal across classes,
- only the random effect variance as class-specific, or
- both random effect and residual error variances class-specific.

The EM algorithm was used to fit the structurally simple models, and the variances were not bounded to ensure a finite likelihood. Class membership was fit as being the same for all subjects.

Research has also focused on models which generalize the random effects distribution of the usual LMM to be a mixture of normal distributions with different means, but with variances equal across classes. This is in contrast to the methods proposed in this dissertation where the random effects distribution is modeled as a mixture of normal distributions with zero mean and class-specific variances. Verbeke and Lesaffre [1996] investigated the impact of heterogeneity in the random effects distribution. The authors proposed a mixture of LMMs in which the random effects distribution was modeled as described above. The authors showed that if normality was assumed in situations where the distribution of the random effects was a finite mixture of normal distributions, then the random effects may be badly estimated. Verbeke and Molenberghs [2000] provided further details related to this model and referred to the model as the 'Heterogeneity Model'. In the proposed model, class-specific fixed-effects were not fit, nor was detail presented regarding possible variance-covariance structures. Spiessens et al [2002] implemented the version of the Heterogeneity Model described in Verbeke and Lesaffre [1996] in a SAS® macro using PROC NLMIXED. The fitted model assumed that the weights for each latent class were the same for all subjects. Xu and Hedeker [2001] proposed a model similar to Verbeke and Molenberghs [2000] and Spiessens et al [2002], in which the authors provided details related to both EM and Fisher scoring algorithms. Proust and Jacqmin-Gadda [2005] proposed a slightly more general model than Verbeke and Lesaffre [1996]. In their revised model, some random effects were modeled by a single multivariate Normal distribution, while others were modeled by a mixture of multivariate Normal distributions, each with a class-specific mean but the same variance-covariance matrix. The group membership probabilities were once again modeled as being the same for all subjects. The authors used a modified Marquardt optimization algorithm, similar to Newton-Raphson, to solve for the parameters. Gusnanto et al [2005] applied a similar model to an application in gene expression data.

Related models have been proposed which have somewhat different model specifications but which share many of the same features of the LCLMM discussed in this research. Muthen et al [2002] proposed a finite mixture model with random effects. While many of the components are similar to the model proposed in this dissertation, the model specification is somewhat different and acts through a continuous latent variable, $\boldsymbol{\eta}_i$. The model is written as follows:

$$\begin{aligned} \mathbf{y}_i &= \boldsymbol{\Lambda}_k \boldsymbol{\eta}_i + \boldsymbol{\epsilon}_i \\ \boldsymbol{\eta}_i &= \boldsymbol{\alpha}_k + \boldsymbol{\Gamma}_k \mathbf{x}_i + \boldsymbol{\varsigma}_i \end{aligned}$$

where: $\mathbf{y}_i$ denotes a vector of continuous outcome variables

$\mathbf{x}_i$ denotes a vector of covariates

$\boldsymbol{\eta}_i$ denotes a vector of continuous variables (latent)

$\boldsymbol{\Lambda}_k$ denotes the parameters related to $\boldsymbol{\eta}_i$ for the k^{th} class

$\boldsymbol{\alpha}_k$ denotes the intercept parameters for the k^{th} class

$\boldsymbol{\Gamma}_k$ denotes the parameters associated with $\mathbf{x}_i$ for the k^{th} class

$\boldsymbol{\epsilon}_i \sim N(0, \boldsymbol{\Theta}_k)$

$\boldsymbol{\varsigma}_i \sim N(0, \boldsymbol{\Psi}_k)$

In words, the continuous latent variable $\boldsymbol{\eta}_i$ is modeled as a function of covariates $\mathbf{x}_i$ with class-specific parameters $\boldsymbol{\Gamma}_k$ and random effects $\boldsymbol{\varsigma}_i$. This latent variable is then modeled as a covariate with class-specific parameters $\boldsymbol{\Lambda}_k$ and error terms $\boldsymbol{\epsilon}_i$. This model can more readily be understood by examining the resulting multivariate Normal distributions, given class membership ($\mathbf{c}_i$), which are shown below. For comparison purposes, the corresponding values for the LCLMM are also provided.

$$\mathbf{y}_i \mid \mathbf{c}_i, \mathbf{x}_i \sim N\left(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)$$

with:

Muthen et al [2002]:

$$\begin{aligned} \boldsymbol{\mu}_k &= \boldsymbol{\Lambda}_k\left(\boldsymbol{\alpha}_k + \boldsymbol{\Gamma}_k \mathbf{x}_i\right) \\ \boldsymbol{\Sigma}_k &= \boldsymbol{\Lambda}_k \boldsymbol{\Phi}_k \boldsymbol{\Lambda}_k' + \boldsymbol{\Theta}_k \end{aligned}$$

LCLMM:

$$\begin{aligned} \boldsymbol{\mu}_k &= \mathbf{X}_i \boldsymbol{\beta} + \mathbf{W}_i \boldsymbol{\lambda}_k \\ \boldsymbol{\Sigma}_k &= \mathbf{Z}_i \mathbf{D} \mathbf{Z}_i' + \mathbf{R}_i \end{aligned}$$

Note the differences in the models. Most notable is the fact that while the variance covariance matrix of the random effects is pre- and post-mulitplied by $\mathbf{Z}_i$ in the case of the LCLMM and $\boldsymbol{\Lambda}_k$ in the case of the Muthen model, the $\mathbf{Z}_i$ matrix does not appear in the mean vector for the LCLMM.

Roy and Lin [2000] proposed a latent variable model in which multiple continuous outcomes were observed over time. At each timepoint, various covariates were assumed to be related to an underlying continuous latent variable. The latent variables for each subject were then assumed to be correlated by including random effects in the model. Estimates were obtained via the EM algorithm. Note that this model is somewhat different than the model examined in this dissertation since there is only one class. The continuous latent variables are instead used to provide a summary measure of all of the covariates at each timepoint. Muthen [2002] provided an overview of methods of statistical analysis which make use of latent variables, presenting four modeling frameworks. Framework C is most similar to the models being considered in this research, with finite mixture modeling included as a particular case. A brief example was provided but detail regarding variance-covariance structures, class-specific variances, etc. was not provided. Yau et al [2003] proposed a two-component mixture regression model in which random effects were included in both a logistic model of the mixture probabilities and the underlying linear models. Computations were performed via an EM algorithm. A single variance was fit for the random effect in the mixture probability portion of the model, and a variance for each group was fit in the linear model portion of the model.

Since the proposed LCLMM combines the traditional LMM with classification into groups, research in the area of model-based cluster analysis provides an additional perspective on these models. These models do not typically include random effects in their model specification and therefore are not as straightforward to use in model specifications where repeated measures are involved. Banfield and Raftery [1993] proposed a clustering algorithm which parameterizes the variance-covariance matrix

$\mathbf{\Sigma}_k$ in terms of its eigenvalue decomposition $\mathbf{\Sigma}_k = \mathbf{D}_k \mathbf{\Lambda}_k \mathbf{D}'_k = \mathbf{D}_k \lambda_k \mathbf{A}_k \mathbf{D}'_k$, where $\mathbf{D}_k$ is the matrix of eigenvectors and $\mathbf{\Lambda}_k$ is a diagonal matrix with the eigenvalues of $\mathbf{\Sigma}_k$ on the diagonal. Note that $\mathbf{\Lambda}_k = \lambda_k \mathbf{A}_k$. Effectively, in their notation, $\mathbf{D}_k$ determines the orientation of the k^{th} cluster, λ_k its size, and $\mathbf{A}_k$ its shape. The authors also discussed the different models that result when different sets of components are allowed to vary by cluster. Details were not provided related to how the models were fit, and necessary conditions to ensure a finite likelihood were not discussed. Fraley and Raftery [2002] provided a review of several clustering methods. One of the two main models which were reviewed maximized the multivariate normal likelihood using an EM algorithm, although few details were provided. The second was described as model-based agglomerative hierarchical clustering, which operates by successively merging pairs of clusters that result in the greatest increase in the classification likelihood. Fraley et al. [2003] proposed a model-based clustering algorithm which incrementally adds clusters as needed until the model fit is no longer improved. In this algorithm, a preliminary mixture model is fit with fewer classes than would be expected. Then the set of observations which are fit worst are reclassified as being in a new cluster, and the model is re-fit. The algorithm continues until adding a cluster results in a decrease to the BIC. Peel and McLachlan [2000] proposed a mixture model based on the multivariate t-distribution. In the model, variances were assumed to be equal across classes, and the same class membership probabilities were assumed to apply to all subjects. The authors used an EM/ECM algorithm to solve for estimates. The variance-covariance matrix was unstructured in this model. Arcidiacono and Jones [2003] applied the ECM algorithm described in Meng and Rubin [1993] to the problem of estimation in finite mixture models. Particular focus was given to likelihoods which can be factored into two quantities, one containing a function related to one set of parameters, and a second which includes potentially all of the parameters. Li [2005] proposed using a multilayer mixture model for applications in clustering. Each individual cluster was modeled by a mixture of normal distributions. In this model, clusters do not share component distributions - each cluster is a mixture of its own set of normal distributions. The proposed model was simple in structure, consisting of means and variances. Random effects, application to repeated-measures data, and the specification of a structure for class membership were not included as part of the proposed model. Celeux and Govaert [1992] proposed a classification EM algorithm to estimate parameters in mixture models. In the approach, the usual E-step was followed, in which an estimate of the probability of being in each class was updated. In a classification step, each subject was assigned to be in the class with the highest probability, and the usual M-step was calculated. A further proposal called for stochastically determining the class assignment based on the probabilities of being in each latent class.

And finally, no literature review would be complete without a review of the capabilities of existing software packages. Several packages are available which provide estimates of the parameters in finite mixture models. Most notably of these are MIXMOD, Mplus, MCLUST, and FlexMix. While all provide for the computation of parameters in specific versions of these models or related models, none focuses on the mixture of LMMs and allows for the general model specification which is the focus of this dissertation. Biernacki et al [2005] provided a review of the features of the software package MIXMOD. This package fits mixture models of multivariate Gaussian distributions. The variance-covariance structure is modeled according to Banfield and Raftery [1993] and various quantities are allowed to be estimated across classes or to be class-specific. A more detailed review of the underlying model described in Banfield and Raftery [1993] was provided earlier in this section. Random effects are not included in this model. Fraley and Raftery [2006] described the models fit in the software package MCLUST. Gaussian mixture models are fit with class-specific means and a variety of variance-covariance structures. The variance-covariance structures ranged from constant variance across latent classes to unconstrained variances to structured variances where some components are allowed to vary across latent classes. The variance-covariance matrix is also structured as in Banfield and Raftery [1993]. The models do not use a random effects specification of the variance-covariance structure, and the mixture probabilities apply to all subjects rather than allowing for different underlying distributions for each subject. Grun and Leisch [2007] described the capabilities of the software package FlexMix. The authors indicate that the package provides the E-step for the EM algorithm for mixture models and the user can select a driver to run the M-step for mixtures of multivariate Gaussian distributions. However, only diagonal or unstructured variance-covariance matrices can be fit with the package. In addition, it is not obvious how mixture probabilities are modeled - i.e. via a logit model, overall, or for each subject, etc. The software includes an option to force each class to have a minimum number of subjects. The Mplus Technical Appendices [1998-2004] provided a review of the technical details of the software package Mplus. The model described earlier in Muthen and Shedden [1999] is one of the models fit in the package. This model shares many of the features discussed in this dissertation, but has a somewhat different model specification.

1.4 New Methods

1.4.1 LCLMM: Structured Variance Components, Equal for All Latent Classes

Motivation and Known Results

The LMM provides a straightforward, intuitive way for statisticians to specify a statistical model, and the LCLMM provides the additional component of determining class membership. However, the literature thus far has not addressed how to fit these models when the variance components are structured. In many applications, it is reasonable to expect that several different repeated measurements may be modeled simultaneously - for example systolic blood pressure, diastolic blood pressure, and heart rate or test scores on math, reading, science, and social studies exams. Each of these measurements would potentially require a different residual error variance, and the random intercepts and slopes for these measurements may be modeled best by fitting a particular correlation structure. The statistical literature has not yet addressed how to allow for such a general model specification. However, these scenarios are addressed in the model proposed in this section.

In Lin et al [2002a], the authors used the EM algorithm to solve a joint latent-class survival/mixed model. The results were maximum likelihood estimates for the parameters with subjects optimally divided among a prespecified number of latent classes. In this model, an unstructured variance-covariance matrix of the random effects was specified, a single residual error variance, σ^2, was fit, and the variances were assumed to be equal across classes. Lin et al [2002b] proposed a joint analysis of time-to-event data with multiple longitudinal variables via a mixture model. The longitudinal submodel allowed for different variances for each longitudinal variable, as well as the inclusion of random effects with an unstructured variance-covariance matrix. Variances were assumed to be equal across classes. The goal in this section is to expand the longitudinal portion of the Lin et al [2002a] and Lin et al [2002b] models to allow for structured variance-covariance matrices, while continuing to model class membership via the logistic regression model.

Two articles provide the basis for the changes proposed in this section. In work by Jennrich and Schluchter [1986], the authors proposed a general method for computing variance components in a maximum likelihood algorithm. In this method, the statistician can specify a general variance structure, and computations proceed through the Newton-Raphson or Fisher scoring algorithms. Rai and Matthews [1993] proposed a modification to the M-step of the EM algorithm. In cases where the

M-step cannot be calculated in closed form, the authors proposed replacing the M-step with a single iteration of a Newton-Raphson maximization of the complete-data likelihood. The authors showed that by using a variable step length, a step can be specified which either increases the likelihood or at worst allows it to remain the same, thus satisfying the criteria for being a Generalized EM algorithm.

The following sections expand upon previous research in several ways. First, building on Rai and Matthews [1993] and Jennrich and Schluchter [1986], the M-step in the EM algorithm is revised to consist of a single iteration of Newton-Raphson based on the expected complete-data likelihood. Second, the first and second derivatives of the unconditional log likelihood of the observed data are derived, and gradient- and Hessian-based algorithms are proposed and implemented. Both of these methods allow the statistician to specify a structure for the variance components that determine $\mathbf{D}$ and $\mathbf{R}$.

Background

The first question that arises relates to determining which likelihood should be maximized in the proposed problem. If we knew each subject's class membership, it would be possible to maximize the likelihood of the observed data given the class membership $(\mathbf{y}_i \mid \mathbf{c}_i)$. However, since the class membership of each subject is an unobservable value, this is not possible. Other choices of likelihoods include the unconditional likelihood of the observed data $(\mathbf{y}_i)$, as well as the joint likelihoods $(\mathbf{y}_i, \mathbf{c}_i, \mathbf{b}_i)$ and $(\mathbf{y}_i, \mathbf{c}_i)$. The unconditional likelihood of the observed data is considered below with respect to the gradient and Hessian-based algorithms. The joint likelihood including the random effects is quite complicated - therefore the focus for the EM algorithm will be to maximize the joint likelihood of the responses and group membership $(\mathbf{y}_i, \mathbf{c}_i)$, with the group membership of each subject taken to be missing data.

Two separate approaches are considered in this section. First, since the LCLMM represents an instance where the data can be considered to be incomplete (i.e., class membership) and the complete-data model is straightforward, the EM algorithm of Dempster et al [1977] represents a logical choice, and is used most often in practice. In the case of complete data, the likelihood corresponding to the joint distribution of the observed data and group classifications $(\mathbf{y}_i, \mathbf{c}_i)$ can be written as:

$$\log L_C(\boldsymbol{\Theta}) = \sum_{i=1}^{n} \sum_{k=1}^{K} c_{ik} \log \left[\pi_{ik} f(\mathbf{y}_i \mid c_{ik} = 1) \right] \tag{1.9}$$

with $f\left(\mathbf{y}_i \mid c_{ik}=1\right)$ being the density defined by

$$\left(\mathbf{y}_i \mid c_{ik}=1\right) \sim N(\mathbf{X}_i\boldsymbol{\beta}+\mathbf{W}_i\boldsymbol{\lambda}_k, \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i'+\mathbf{R}_i)$$

However, since the class membership of each subject is unobservable, the complete data likelihood is replaced during each E-step with its conditional expectation given the observed data $\mathbf{y}_i$, using the current estimates of the parameters, $\boldsymbol{\Theta}^*$. In Dempster et al [1977], the authors refer to this as

$$Q\left(\boldsymbol{\Theta};\boldsymbol{\Theta}^*\right)=\mathrm{E}_{\boldsymbol{\Theta}^*}\left\{\log\, L_C\left(\boldsymbol{\Theta}\right) \mid \mathbf{y}\right\}$$

The goal of each M-step is then to maximize $Q\left(\boldsymbol{\Theta};\boldsymbol{\Theta}^*\right)$. Since the c_{ik} are treated as missing data in the above likelihood specification, the only change that is required in going from $\log\, L_C\left(\boldsymbol{\Theta}\right)$ to $Q\left(\boldsymbol{\Theta};\boldsymbol{\Theta}^*\right)$ is that the value of c_{ik} is replaced by its expected value $\widetilde{c}_{ik}$. It follows that:

$$Q(\boldsymbol{\Theta};\boldsymbol{\Theta^*})=\sum_{i=1}^{n}\sum_{k=1}^{K}\widetilde{c}_{ik}\,\log\,\left[\pi_{ik}f\left(\mathbf{y}_i \mid c_{ik}=1\right)\right] \tag{1.10}$$

In full,

$$Q(\boldsymbol{\Theta};\boldsymbol{\Theta^*})=\sum_{i=1}^{n}\sum_{k=1}^{K}\widetilde{c}_{ik}\left[\begin{array}{c}\log\,\pi_{ik}-\frac{n_i}{2}\,\log 2\pi-\frac{1}{2}\,\log\,\left|\mathbf{Z}_i\mathbf{D}\mathbf{Z}_i'+\mathbf{R}_i\right| \\ -\frac{1}{2}\left(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\beta}-\mathbf{W}_i\boldsymbol{\lambda}_k\right)'\left(\mathbf{Z}_i\mathbf{D}\mathbf{Z}_i'+\mathbf{R}_i\right)^{-1}\left(\mathbf{y}_i-\mathbf{X}_i\boldsymbol{\beta}-\mathbf{W}_i\boldsymbol{\lambda}_k\right)\end{array}\right] \tag{1.11}$$

In the second approach, which requires computation of one or both of the gradient and Hessian, the goal will be to maximize the unconditional likelihood of the observed data, which can be written as follows:

$$\log\, L\left(\boldsymbol{\Theta}\right)=\sum_{i=1}^{n}\log\,\sum_{k=1}^{K}\pi_{ik}f\left(\mathbf{y}_i \mid c_{ik}=1\right) \tag{1.12}$$

with $f\left(\mathbf{y}_i \mid c_{ik}=1\right)$ being the density defined by

$$\left(\mathbf{y}_i \mid c_{ik}=1\right) \sim N(\mathbf{X}_i\boldsymbol{\beta}+\mathbf{W}_i\boldsymbol{\lambda}_k, \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i'+\mathbf{R}_i)$$

Although straightforward in that only the first and second derivatives of the likelihood are needed, finding the maximum of $\log\, L\left(\boldsymbol{\Theta}\right)$ is not an easy task with each subject's contribution to the likelihood involving the log of the sum of the distribution under each of the K models. However, given the quadratic convergence properties of the Newton-Raphson algorithm and possible computational advantages of gradient-based algorithms such as quasi-Newton, these methods may offer the most efficient method of finding solutions.

EM Approach

In order to allow for structured variance components which are equal across classes, I propose building on the Lin et al [2002a] and Lin et al [2002b] articles by incorporating a single iteration of Newton-Raphson within each M-step to compute the updated estimates of the variance components. Since the estimates of the parameters $\boldsymbol{\alpha}_k$ that determine the class probabilities π_{ik} also are not available in closed form, this technique will be used to update these parameters as well.

As mentioned previously, in each E-step of the EM algorothm, the expected complete-data likelihood, $Q\left(\boldsymbol{\Theta};\boldsymbol{\Theta}^*\right)$, is calculated as in Equation 1.11, where the expected value of $\widetilde{c}_{ik}$ can be written as follows:

$$\begin{aligned}
\mathrm{E}\left(c_{ik} \mid \mathbf{y}_i\right) &= \widetilde{c}_{ik} = P\left(c_{ik}=1 \mid \mathbf{y}_i\right) \\
&= \frac{P\left(c_{ik}=1\right) \times P\left(\mathbf{y}_i \mid c_{ik}=1\right)}{\sum_{j=1}^{K}\left[P\left(c_{ij}=1\right) \times P\left(\mathbf{y}_i \mid c_{ij}=1\right)\right]} \\
&= \frac{\pi_{ik} f\left(\mathbf{y}_i \mid c_{ik}=1\right)}{\sum_{j=1}^{K}\left\{\pi_{ij} f\left(\mathbf{y}_i \mid c_{ij}=1\right)\right\}}
\end{aligned}$$

$$\text{with } \left(\mathbf{y}_i \mid c_{ik}=1\right) \sim N(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{W}_i\boldsymbol{\lambda}_k, \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i' + \mathbf{R}_i)$$

The goal of each M-step is to then maximize $Q\left(\boldsymbol{\Theta};\boldsymbol{\Theta}^*\right)$ with respect to the parameters.

In order to maximize $Q\left(\boldsymbol{\Theta};\boldsymbol{\Theta}^*\right)$ with respect to $\boldsymbol{\beta}$, first calculate:

$$\frac{\partial Q(\boldsymbol{\Theta};\boldsymbol{\Theta}^*)}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n}\sum_{k=1}^{K} \widetilde{c}_{ik}\left(\mathbf{X}_i'\boldsymbol{\Sigma}_i^{-1}\mathbf{y}_i - \mathbf{X}_i'\boldsymbol{\Sigma}_i^{-1}\mathbf{X}_i\boldsymbol{\beta} - \mathbf{X}_i'\boldsymbol{\Sigma}_i^{-1}\mathbf{W}_i\boldsymbol{\lambda}_k\right) \tag{1.13}$$

$$\text{with } \boldsymbol{\Sigma}_i = \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i' + \mathbf{R}_i$$

Upon setting this to zero and solving for $\boldsymbol{\beta}$, it follows that:

$$\widehat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{n} \mathbf{X}_i' \boldsymbol{\Sigma}_i^{-1} \mathbf{X}_i\right)^{-1} \sum_{i=1}^{n} \mathbf{X}_i' \boldsymbol{\Sigma}_i^{-1} \left(\mathbf{y}_i - \sum_{k=1}^{K} \widetilde{c}_{ik} \mathbf{W}_i \boldsymbol{\lambda}_k\right) \tag{1.14}$$

Similarly, for $\boldsymbol{\lambda}_k$, where $k = 1, \ldots, K$:

$$\tfrac{\partial Q(\boldsymbol{\Theta};\boldsymbol{\Theta}^*)}{\partial \boldsymbol{\lambda}_k} = \sum_{i=1}^{n} \widetilde{c}_{ik} \left(\mathbf{W}_i' \boldsymbol{\Sigma}_i^{-1} \mathbf{y}_i - \mathbf{W}_i' \boldsymbol{\Sigma}_i^{-1} \mathbf{X}_i \boldsymbol{\beta} - \mathbf{W}_i' \boldsymbol{\Sigma}_i^{-1} \mathbf{W}_i \boldsymbol{\lambda}_k\right) \tag{1.15}$$

$$\widehat{\boldsymbol{\lambda}_k} = \left(\sum_{i=1}^{n} \widetilde{c}_{ik} \mathbf{W}_i' \boldsymbol{\Sigma}_i^{-1} \mathbf{W}_i\right)^{-1} \sum_{i=1}^{n} \widetilde{c}_{ik} \mathbf{W}_i' \boldsymbol{\Sigma}_i^{-1} \left(\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\beta}\right) \tag{1.16}$$

The expected complete-data likelihood $Q(\boldsymbol{\Theta};\boldsymbol{\Theta}^*)$, evaluated at the revised set of parameters which reflect the updated $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\lambda}_k}$, can then be used in a single Newton-Raphson step to calculate structured variance components which are the same for all latent classes. Following the notation in Jennrich and Schluchter [1986], the r^{th} element of the score vector and $r - s^{th}$ element of the Hessian can be computed as follows:

$$\left[\mathbf{s}_{\boldsymbol{\theta}}\right]_r = \tfrac{1}{2} \sum_{i=1}^{n} \sum_{k=1}^{K} \widetilde{c}_{ik} tr \left[\boldsymbol{\Sigma}_i^{-1} \left(\mathbf{e}_{ik} \mathbf{e}_{ik}' - \boldsymbol{\Sigma}_i\right) \boldsymbol{\Sigma}_i^{-1} \dot{\boldsymbol{\Sigma}}_{ir}\right] \tag{1.17}$$

$$\begin{aligned} \left[\mathbf{H}_{\boldsymbol{\theta}\boldsymbol{\theta}}\right]_{rs} &= -\tfrac{1}{2} \sum_{i=1}^{n} \sum_{k=1}^{K} \widetilde{c}_{ik} tr \left[\boldsymbol{\Sigma}_i^{-1} \dot{\boldsymbol{\Sigma}}_{ir} \boldsymbol{\Sigma}_i^{-1} \left(2\mathbf{e}_{ik} \mathbf{e}_{ik}' - \boldsymbol{\Sigma}_i\right) \boldsymbol{\Sigma}_i^{-1} \dot{\boldsymbol{\Sigma}}_{is}\right] \\ &\quad + \tfrac{1}{2} \sum_{i=1}^{n} \sum_{k=1}^{K} \widetilde{c}_{ik} tr \left[\boldsymbol{\Sigma}_i^{-1} \left(\mathbf{e}_{ik} \mathbf{e}_{ik}' - \boldsymbol{\Sigma}_i\right) \boldsymbol{\Sigma}_i^{-1} \ddot{\boldsymbol{\Sigma}}_{i,rs}\right] \end{aligned} \tag{1.18}$$

where:

$$\begin{aligned} \mathbf{e}_{ik} &= \mathbf{y}_i - \mathbf{X}_i \boldsymbol{\beta} - \mathbf{W}_i \boldsymbol{\lambda}_k \\ \dot{\boldsymbol{\Sigma}}_{ir} &= \partial \boldsymbol{\Sigma}_i / \partial \boldsymbol{\theta}_r \\ \ddot{\boldsymbol{\Sigma}}_{i,rs} &= \partial^2 \boldsymbol{\Sigma}_i / \partial \boldsymbol{\theta}_r \partial \boldsymbol{\theta}_s \end{aligned}$$

Similarly, since closed-form solutions for the $\boldsymbol{\alpha}_k$ are not available, a scoring procedure is also used for calculating estimates of the parameters related to class membership. The score vector and Hessian can be computed as follows:

$$\mathbf{s}_{\boldsymbol{\alpha_k}} = \sum_{i=1}^{n} \sum_{j=1}^{K} \widetilde{c}_{ij} \mathbf{t}_i \times \begin{cases} (1 - \pi_{ik}) & \text{if } j = k \\ (-\pi_{ik}) & \text{if } j \neq k \end{cases}$$

$$\mathbf{H}_{\boldsymbol{\alpha_k}\boldsymbol{\alpha_m}} = \sum_{i=1}^{n}\sum_{j=1}^{K}\widetilde{c}_{ij}\mathbf{t}_i\mathbf{t}_i' \times \begin{cases} -\pi_{ik}\left(1-\pi_{im}\right) & \text{if } k=m \\ \pi_{ik}\pi_{im} & \text{if } k \neq m \end{cases}$$

The updated estimates of the variance components and parameters related to class membership can be computed by the following steps:

$$\boldsymbol{\theta} = \boldsymbol{\theta} - j \times \mathbf{H}_{\boldsymbol{\theta\theta}}^{-1}\mathbf{s}_{\boldsymbol{\theta}} \tag{1.19}$$

$$\begin{bmatrix} \boldsymbol{\alpha}_2 \\ \vdots \\ \boldsymbol{\alpha}_K \end{bmatrix} = \begin{bmatrix} \boldsymbol{\alpha}_2 \\ \vdots \\ \boldsymbol{\alpha}_K \end{bmatrix} - j \times \begin{bmatrix} \mathbf{H}_{\boldsymbol{\alpha}_2\boldsymbol{\alpha}_2} & \cdots & \mathbf{H}_{\boldsymbol{\alpha}_2\boldsymbol{\alpha}_K} \\ \vdots & \ddots & \vdots \\ \mathbf{H}_{\boldsymbol{\alpha}_K\boldsymbol{\alpha}_2} & \cdots & \mathbf{H}_{\boldsymbol{\alpha}_K\boldsymbol{\alpha}_K} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{s}_{\boldsymbol{\alpha}_2} \\ \vdots \\ \mathbf{s}_{\boldsymbol{\alpha}_K} \end{bmatrix} \tag{1.20}$$

where j is chosen separately for each set of parameters such that the expected complete data likelihood is increased. By ensuring that the expected complete data likelihood increases at each step, the algorithm represents a generalized EM algorithm, as described in Dempster et al [1977]. Although Rai and Matthews [1993] note that a single iteration of a Newton-Raphson maximization of the complete-data likelihood can be substituted in the M-step in cases where the M-step cannot be calculated in closed form, one modification is necessary in practice for these models. In running test cases of the algorithms, I found it necessary to run the update step for the class-membership parameters several times *within* each M-step in order to minimize the overall number of iterations needed. Since each update of the class-membership parameters is much less time-consuming than a full iteration, the result was typically a faster overall runtime.

We repeat the iterations until the maximum percent change in parameter estimates from one iteration to the next is less than a prespecified tolerance.

Gradient- and Hessian-Based Approaches

Gradient-based algorithms rely on the calculation of both the unconditional log likelihood of the observed data and the first derivatives of this likelihood with respect to the unknown parameters. Hessian-based methods require the additional calculation of the second derivatives of this likelihood.

As noted earlier, the unconditional log likelihood of the observed data is as follows:

$$\log L(\mathbf{\Theta}) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \pi_{ik} f(\mathbf{y}_i \mid c_{ik} = 1) \tag{1.21}$$

with $f(\mathbf{y}_i \mid c_{ik} = 1)$ being the density defined by

$(\mathbf{y}_i \mid c_{ik} = 1) \sim N(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{W}_i\boldsymbol{\lambda}_k, \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i' + \mathbf{R}_i)$

In full,

$$\log L(\mathbf{\Theta}) = \sum_{i=1}^{n} \log \underbrace{\sum_{k=1}^{K} \left\{ \underbrace{\left[\frac{\exp(\mathbf{t}_i'\boldsymbol{\alpha}_k)}{\sum_{j=1}^{K} \exp(\mathbf{t}_i'\boldsymbol{\alpha}_j)} \right]}_{Q_{2ik}} \underbrace{\left[\frac{\exp\left[-\frac{1}{2}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{W}_i\boldsymbol{\lambda}_k)' \mathbf{\Sigma}_i^{-1} (\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{W}_i\boldsymbol{\lambda}_k)\right]}{(2\pi)^{n_i/2} |\mathbf{\Sigma}_i|^{1/2}} \right]}_{Q_{3ik}} \right\}}_{Q_{1i}} \tag{1.22}$$

with $\mathbf{\Sigma}_i = \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i' + \mathbf{R}_i$

The score vector $\mathbf{s}$, containing the first derivatives of $\log L(\mathbf{\Theta})$ with respect to the parameters, and Hessian matrix $\mathbf{H}$, containing the second derivatives of $\log L(\mathbf{\Theta})$ with respect to the parameters, are written as follows:

$$\mathbf{s} = \begin{bmatrix} \mathbf{s}_{\boldsymbol{\beta}} \\ \mathbf{s}_{\boldsymbol{\lambda}_1} \\ \vdots \\ \mathbf{s}_{\boldsymbol{\lambda}_K} \\ \mathbf{s}_{\boldsymbol{\theta}} \\ \mathbf{s}_{\boldsymbol{\alpha}_2} \\ \vdots \\ \mathbf{s}_{\boldsymbol{\alpha}_K} \end{bmatrix} \tag{1.23}$$

and

$$\mathbf{H} = \begin{bmatrix} \mathbf{H}_{\boldsymbol{\beta}\boldsymbol{\beta}} & \mathbf{H}_{\boldsymbol{\beta}\boldsymbol{\lambda}_1} & \cdots & \mathbf{H}_{\boldsymbol{\beta}\boldsymbol{\lambda}_K} & \mathbf{H}_{\boldsymbol{\beta}\boldsymbol{\theta}} & \mathbf{H}_{\boldsymbol{\beta}\boldsymbol{\alpha}_2} & \cdots & \mathbf{H}_{\boldsymbol{\beta}\boldsymbol{\alpha}_K} \\ \mathbf{H}_{\boldsymbol{\lambda}_1\boldsymbol{\beta}} & \mathbf{H}_{\boldsymbol{\lambda}_1\boldsymbol{\lambda}_1} & \cdots & \mathbf{H}_{\boldsymbol{\lambda}_1\boldsymbol{\lambda}_K} & \mathbf{H}_{\boldsymbol{\lambda}_1\boldsymbol{\theta}} & \mathbf{H}_{\boldsymbol{\lambda}_1\boldsymbol{\alpha}_2} & \cdots & \mathbf{H}_{\boldsymbol{\lambda}_1\boldsymbol{\alpha}_K} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{H}_{\boldsymbol{\lambda}_K\boldsymbol{\beta}} & \mathbf{H}_{\boldsymbol{\lambda}_K\boldsymbol{\lambda}_1} & \cdots & \mathbf{H}_{\boldsymbol{\lambda}_K\boldsymbol{\lambda}_K} & \mathbf{H}_{\boldsymbol{\lambda}_K\boldsymbol{\theta}} & \mathbf{H}_{\boldsymbol{\lambda}_K\boldsymbol{\alpha}_2} & \cdots & \mathbf{H}_{\boldsymbol{\lambda}_K\boldsymbol{\alpha}_K} \\ \mathbf{H}_{\boldsymbol{\theta}\boldsymbol{\beta}} & \mathbf{H}_{\boldsymbol{\theta}\boldsymbol{\lambda}_1} & \cdots & \mathbf{H}_{\boldsymbol{\theta}\boldsymbol{\lambda}_K} & \mathbf{H}_{\boldsymbol{\theta}\boldsymbol{\theta}} & \mathbf{H}_{\boldsymbol{\theta}\boldsymbol{\alpha}_2} & \cdots & \mathbf{H}_{\boldsymbol{\theta}\boldsymbol{\alpha}_K} \\ \mathbf{H}_{\boldsymbol{\alpha}_2\boldsymbol{\beta}} & \mathbf{H}_{\boldsymbol{\alpha}_2\boldsymbol{\lambda}_1} & \cdots & \mathbf{H}_{\boldsymbol{\alpha}_2\boldsymbol{\lambda}_K} & \mathbf{H}_{\boldsymbol{\alpha}_2\boldsymbol{\theta}} & \mathbf{H}_{\boldsymbol{\alpha}_2\boldsymbol{\alpha}_2} & \cdots & \mathbf{H}_{\boldsymbol{\alpha}_2\boldsymbol{\alpha}_K} \\ \vdots & \vdots & \cdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{H}_{\boldsymbol{\alpha}_K\boldsymbol{\beta}} & \mathbf{H}_{\boldsymbol{\alpha}_K\boldsymbol{\lambda}_1} & \cdots & \mathbf{H}_{\boldsymbol{\alpha}_K\boldsymbol{\lambda}_K} & \mathbf{H}_{\boldsymbol{\alpha}_K\boldsymbol{\theta}} & \mathbf{H}_{\boldsymbol{\alpha}_K\boldsymbol{\alpha}_2} & \cdots & \mathbf{H}_{\boldsymbol{\alpha}_K\boldsymbol{\alpha}_K} \end{bmatrix} \tag{1.24}$$

where:

$$\mathbf{s}_{\mathbf{x}} = \tfrac{\partial L}{\partial \mathbf{x}} \quad \text{and} \quad \mathbf{H}_{xy} = \tfrac{\partial^2 L}{\partial \mathbf{x} \partial \mathbf{y}} \tag{1.25}$$

Expressions for the elements of **s** and **H** are:

$$s_{\boldsymbol{\beta}} = \sum_{i=1}^{n} \frac{1}{Q_{1i}} \sum_{k=1}^{K} \pi_{ik} \frac{\partial Q_{3ik}}{\partial \boldsymbol{\beta}} \tag{1.26}$$

$$s_{\boldsymbol{\lambda}_k} = \sum_{i=1}^{n} \frac{1}{Q_{1i}} \pi_{ik} \frac{\partial Q_{3ik}}{\partial \boldsymbol{\lambda}_k} \tag{1.27}$$

$$s_{\boldsymbol{\alpha}_k} = \sum_{i=1}^{n} \frac{1}{Q_{1i}} \sum_{j=1}^{K} Q_{3ij} \frac{\partial Q_{2ij}}{\partial \boldsymbol{\alpha}_k} \tag{1.28}$$

$$s_{\boldsymbol{\theta}} = \sum_{i=1}^{n} \frac{1}{Q_{1i}} \sum_{k=1}^{K} \pi_{ik} \frac{\partial Q_{3ik}}{\partial \boldsymbol{\theta}} \tag{1.29}$$

$$H_{\boldsymbol{\beta}\boldsymbol{\beta}} = \sum_{i=1}^{n} \left\{ \left[\frac{1}{Q_{1i}} \sum_{k=1}^{K} \pi_{ik} \frac{\partial^2 Q_{3ik}}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}} \right] + \left[\left(-\frac{1}{Q_{1i}^2} \frac{\partial Q_{1i}}{\partial \boldsymbol{\beta}} \right) \left(\sum_{k=1}^{K} \pi_{ik} \frac{\partial Q_{3ik}}{\partial \boldsymbol{\beta}} \right) \right] \right\} \tag{1.30}$$

$$H_{\boldsymbol{\lambda}_k\boldsymbol{\lambda}_j} = \sum_{i=1}^{n} \left\{ \left[\frac{1}{Q_{1i}} \pi_{ik} \frac{\partial^2 Q_{3ik}}{\partial \boldsymbol{\lambda}_k \partial \boldsymbol{\lambda}_j} \right] + \left[\left(-\frac{1}{Q_{1i}^2} \frac{\partial Q_{1i}}{\partial \boldsymbol{\lambda}_j} \right) \left(\pi_{ik} \frac{\partial Q_{3ik}}{\partial \boldsymbol{\lambda}_k} \right) \right] \right\} \tag{1.31}$$

$$H_{\boldsymbol{\lambda}_k\boldsymbol{\beta}} = \sum_{i=1}^{n} \left\{ \left[\frac{1}{Q_{1i}} \pi_{ik} \frac{\partial^2 Q_{3ik}}{\partial \boldsymbol{\lambda}_k \partial \boldsymbol{\beta}} \right] + \left[\left(-\frac{1}{Q_{1i}^2} \frac{\partial Q_{1i}}{\partial \boldsymbol{\beta}} \right) \left(\pi_{ik} \frac{\partial Q_{3ik}}{\partial \boldsymbol{\lambda}_k} \right) \right] \right\} \tag{1.32}$$

$$H_{\boldsymbol{\alpha}_k\boldsymbol{\alpha}_m} = \sum_{i=1}^{n} \left\{ \left[\frac{1}{Q_{1i}} \sum_{j=1}^{K} Q_{3ij} \frac{\partial^2 Q_{2ij}}{\partial \boldsymbol{\alpha}_k \partial \boldsymbol{\alpha}_m} \right] + \left[\left(-\frac{1}{Q_{1i}^2} \frac{\partial Q_{1i}}{\partial \boldsymbol{\alpha}_m} \right) \left(\sum_{j=1}^{K} Q_{3ij} \frac{\partial Q_{2ij}}{\partial \boldsymbol{\alpha}_k} \right) \right] \right\} \tag{1.33}$$

$$H_{\boldsymbol{\alpha}_k\boldsymbol{\beta}} = \sum_{i=1}^{n} \left\{ \left[\frac{1}{Q_{1i}} \sum_{j=1}^{K} \frac{\partial Q_{2ij}}{\partial \boldsymbol{\alpha}_k} \frac{\partial Q_{3ij}}{\partial \boldsymbol{\beta}} \right] + \left[\left(-\frac{1}{Q_{1i}^2} \frac{\partial Q_{1i}}{\partial \boldsymbol{\alpha}_k} \right) \left(\sum_{j=1}^{K} Q_{2ij} \frac{\partial Q_{3ij}}{\partial \boldsymbol{\beta}} \right) \right] \right\} \tag{1.34}$$

$$H_{\boldsymbol{\alpha}_m\boldsymbol{\lambda}_k} = \sum_{i=1}^{n} \left\{ \left[\frac{1}{Q_{1i}} \frac{\partial Q_{2ik}}{\partial \boldsymbol{\alpha}_m} \frac{\partial Q_{3ik}}{\partial \boldsymbol{\lambda}_k} \right] + \left[-\frac{1}{Q_{1i}^2} \frac{\partial Q_{1i}}{\partial \boldsymbol{\alpha}_m} \left(Q_{2ik} \frac{\partial Q_{3ik}}{\partial \boldsymbol{\lambda}_k} \right) \right] \right\} \tag{1.35}$$

$$H_{\boldsymbol{\theta}\boldsymbol{\theta}} = \sum_{i=1}^{n} \left\{ \left[\frac{1}{Q_{1i}} \sum_{k=1}^{K} \pi_{ik} \frac{\partial^2 Q_{3ik}}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}} \right] + \left[\left(-\frac{1}{Q_{1i}^2} \frac{\partial Q_{1i}}{\partial \boldsymbol{\theta}} \right) \left(\sum_{k=1}^{K} \pi_{ik} \frac{\partial Q_{3ik}}{\partial \boldsymbol{\theta}} \right) \right] \right\} \tag{1.36}$$

$$H_{\boldsymbol{\beta}\boldsymbol{\theta}} = \sum_{i=1}^{n} \left\{ \left[\frac{1}{Q_{1i}} \sum_{k=1}^{K} \pi_{ik} \frac{\partial^2 Q_{3ik}}{\partial \boldsymbol{\beta} \partial \boldsymbol{\theta}} \right] + \left[\left(-\frac{1}{Q_{1i}^2} \frac{\partial Q_{1i}}{\partial \boldsymbol{\theta}} \right) \left(\sum_{k=1}^{K} \pi_{ik} \frac{\partial Q_{3ik}}{\partial \boldsymbol{\beta}} \right) \right] \right\} \tag{1.37}$$

$$H_{\boldsymbol{\lambda}_k\boldsymbol{\theta}} = \sum_{i=1}^{n} \left\{ \left[\frac{1}{Q_{1i}} \pi_{ik} \frac{\partial^2 Q_{3ik}}{\partial \boldsymbol{\lambda}_k \partial \boldsymbol{\theta}} \right] + \left[\left(-\frac{1}{Q_{1i}^2} \frac{\partial Q_{1i}}{\partial \boldsymbol{\theta}} \right) \left(\pi_{ik} \frac{\partial Q_{3ik}}{\partial \boldsymbol{\lambda}_k} \right) \right] \right\} \tag{1.38}$$

$$H_{\boldsymbol{\alpha}_k\boldsymbol{\theta}} = \sum_{i=1}^{n} \left\{ \left[\frac{1}{Q_{1i}} \sum_{j=1}^{K} \frac{\partial Q_{3ij}}{\partial \boldsymbol{\theta}} \frac{\partial Q_{2ij}}{\partial \boldsymbol{\alpha}_k} \right] + \left[\left(-\frac{1}{Q_{1i}^2} \frac{\partial Q_{1i}}{\partial \boldsymbol{\theta}} \right) \left(\sum_{j=1}^{K} Q_{3ij} \frac{\partial Q_{2ij}}{\partial \boldsymbol{\alpha}_k} \right) \right] \right\} \tag{1.39}$$

Note that further details regarding the first and second derivatives are presented in the appendix. For a Newton-Raphson step, the updated estimates of the parameters are computed as follows:

$$\begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{\lambda}_1 \\ \vdots \\ \boldsymbol{\lambda}_k \\ \boldsymbol{\theta} \\ \boldsymbol{\alpha}_2 \\ \vdots \\ \boldsymbol{\alpha}_k \end{bmatrix} = \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{\lambda}_1 \\ \vdots \\ \boldsymbol{\lambda}_k \\ \boldsymbol{\theta} \\ \boldsymbol{\alpha}_2 \\ \vdots \\ \boldsymbol{\alpha}_k \end{bmatrix} - j \times \begin{bmatrix} \mathbf{H}_{\boldsymbol{\beta}\boldsymbol{\beta}} & \mathbf{H}_{\boldsymbol{\beta}\boldsymbol{\lambda}_1} & \cdots & \mathbf{H}_{\boldsymbol{\beta}\boldsymbol{\lambda}_k} & \mathbf{H}_{\boldsymbol{\beta}\boldsymbol{\theta}} & \mathbf{H}_{\boldsymbol{\beta}\boldsymbol{\alpha}_2} & \cdots & \mathbf{H}_{\boldsymbol{\beta}\boldsymbol{\alpha}_k} \\ \mathbf{H}_{\boldsymbol{\lambda}_1\boldsymbol{\beta}} & \mathbf{H}_{\boldsymbol{\lambda}_1\boldsymbol{\lambda}_1} & \cdots & \mathbf{H}_{\boldsymbol{\lambda}_1\boldsymbol{\lambda}_k} & \mathbf{H}_{\boldsymbol{\lambda}_1\boldsymbol{\theta}} & \mathbf{H}_{\boldsymbol{\lambda}_1\boldsymbol{\alpha}_2} & \cdots & \mathbf{H}_{\boldsymbol{\lambda}_1\boldsymbol{\alpha}_k} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{H}_{\boldsymbol{\lambda}_k\boldsymbol{\beta}} & \mathbf{H}_{\boldsymbol{\lambda}_k\boldsymbol{\lambda}_1} & \cdots & \mathbf{H}_{\boldsymbol{\lambda}_k\boldsymbol{\lambda}_k} & \mathbf{H}_{\boldsymbol{\lambda}_k\boldsymbol{\theta}} & \mathbf{H}_{\boldsymbol{\lambda}_k\boldsymbol{\alpha}_2} & \cdots & \mathbf{H}_{\boldsymbol{\lambda}_k\boldsymbol{\alpha}_k} \\ \mathbf{H}_{\boldsymbol{\theta}\boldsymbol{\beta}} & \mathbf{H}_{\boldsymbol{\theta}\boldsymbol{\lambda}_1} & \cdots & \mathbf{H}_{\boldsymbol{\theta}\boldsymbol{\lambda}_k} & \mathbf{H}_{\boldsymbol{\theta}\boldsymbol{\theta}} & \mathbf{H}_{\boldsymbol{\theta}\boldsymbol{\alpha}_2} & \cdots & \mathbf{H}_{\boldsymbol{\theta}\boldsymbol{\alpha}_k} \\ \mathbf{H}_{\boldsymbol{\alpha}_2\boldsymbol{\beta}} & \mathbf{H}_{\boldsymbol{\alpha}_2\boldsymbol{\lambda}_1} & \cdots & \mathbf{H}_{\boldsymbol{\alpha}_2\boldsymbol{\lambda}_k} & \mathbf{H}_{\boldsymbol{\alpha}_2\boldsymbol{\theta}} & \mathbf{H}_{\boldsymbol{\alpha}_2\boldsymbol{\alpha}_2} & \cdots & \mathbf{H}_{\boldsymbol{\alpha}_2\boldsymbol{\alpha}_k} \\ \vdots & \vdots & \cdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{H}_{\boldsymbol{\alpha}_k\boldsymbol{\beta}} & \mathbf{H}_{\boldsymbol{\alpha}_k\boldsymbol{\lambda}_1} & \cdots & \mathbf{H}_{\boldsymbol{\alpha}_k\boldsymbol{\lambda}_k} & \mathbf{H}_{\boldsymbol{\alpha}_k\boldsymbol{\theta}} & \mathbf{H}_{\boldsymbol{\alpha}_k\boldsymbol{\alpha}_2} & \cdots & \mathbf{H}_{\boldsymbol{\alpha}_k\boldsymbol{\alpha}_k} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{s}_{\boldsymbol{\beta}} \\ \mathbf{s}_{\boldsymbol{\lambda}_1} \\ \vdots \\ \mathbf{s}_{\boldsymbol{\lambda}_k} \\ \mathbf{s}_{\boldsymbol{\theta}} \\ \mathbf{s}_{\boldsymbol{\alpha}_2} \\ \vdots \\ \mathbf{s}_{\boldsymbol{\alpha}_k} \end{bmatrix} \tag{1.40}$$

where j is chosen such that the unconditional likelihood of the observed data is increased.

We repeat the iterations until the maximum percent change in parameter estimates from one iteration to the next is less than a prespecified tolerance.

The above derivations describe a step of the Newton-Raphson algorithm. With the analytical gradient and Hessian available for computations, various approaches are possible. Chapter 2 discusses a new SAS® macro which fits LCLMMs using the EM and several gradient- and Hessian-based algorithms.

1.4.2 LCLMM: Structured Class-Specific Variance Components, Diagonal D and R

Motivation and Known Results

In many applications, it is unreasonable to expect that the variances would be equal across classes. For example, in studies of diabetic patients, a great deal of information regarding patient health may be contained in the variability of blood sugar measurements. Current methods typically fit the variances as being equal across classes. In Basford and McLachlan [1985], the authors presented an example to illustrate that fitting a normal mixture model with homoscedastic variances in a situation where the classes are in fact heteroscedastic can have a significant influence on the resulting estimates. In particular, they found that the mixing proportions can be greatly affected, resulting in potentially different results in multivariate clustering analyses. As such, the goal of this section is to expand upon the method proposed in the previous section to allow for variances which differ by latent class. This section examines the case where the matrices **D** and **R** are defined to be diagonal - the next section will further generalize this approach to allow for a correlation structure for the **D** matrix. Similar to the previous section, several different residual error variances may be fit for multiple longitudinal variables.

Key questions arise when generalizing to allow for class-specific variances. For example, if the variances are allowed to vary by latent class, do the estimators still possess statistical properties such as identifiability and consistency? In Yakowitz and Spragins [1968], the authors showed that finite mixtures of multivariate normal distributions with variable mean vectors and variance-covariance matrices are identifiable. However, since it is possible in this model specification to provide a scenario in which a particular class would have a single member with the class-specific variance-component going to zero, the likelihood function is unbounded and therefore the maximum likelihood estimate does not exist. For mixtures of univariate normal models, Hathaway [1985] proposed applying simple constraints in which the smallest variance could become no smaller than a prespecified multiple of the corresponding variances from other classes, with the result shown to be a strongly consistent estimator. He described that the corresponding condition for mixtures of multivariate normal models is to constrain all characteristic roots of $\boldsymbol{\Sigma}_j\boldsymbol{\Sigma}_k^{-1}$ to be greater than or equal to some minimum value $c > 0$, where $\boldsymbol{\Sigma}_j$ represents the variance-covariance matrix for class j and $\boldsymbol{\Sigma}_k$ represents the variance-covariance matrix for class k. Assuming diagonal **D** and **R**, the above constraint for multivariate normal data reduces to a set of comparisons similar to the univariate case, with constraints applied to the variances in both $\mathbf{D}$ and $\mathbf{R}$. These constraints are discussed in detail later in this section. The appendices include a proof which shows that Hathaway's condition is met using the proposed method of bounding both the random effect variances and residual error variances for linear variance-covariance structures.

Ingrassia [2004] provided two algorithms related to the multivariate requirement proposed by Hathaway [1985]. As part of the algorithms, the statistician was required to specify a range of possible eigenvalues of $\boldsymbol{\Sigma}_j \boldsymbol{\Sigma}_k^{-1}$. The first of the proposed algorithms either multiplied one of the two variance-covariance matrices by a fraction < 1 or added a multiple of the identity matrix to ensure that the eigenvalues of $\boldsymbol{\Sigma}_j \boldsymbol{\Sigma}_k^{-1}$ fell in the specified range. The other proposed algorithm used the spectral decomposition of the variance-covariance matrix and adjusted the individual eigenvalues such that the eigenvalues fell in the specified range. There are several drawbacks to these methods in practice. First, in both methods, the statistician must specify an acceptable range of eigenvalues which must both include the true value to be estimated and be sufficiently narrow not to allow one variance to stray too far from the others. It is not apparent how to specify these bounds in practice. Second, in the algorithm which adjusts the entire variance-covariance matrix, all variances are adjusted in order to force some parameters to satisfy the constraint. Therefore some of the variances will likely be inaccurate. Third, while the second proposed algorithm allows for each eigenvalue to be adjusted independently, this will result in a variance-covariance matrix which does not have the same correlation structure as specified prior to constraining the eigenvalues. For example, if the variance-covariance structure is diagonal, the result of constraining one eigenvalue but not another causes the resulting correlation matrix to have non-zero elements on the off-diagonal.

I will again make use of the result from Rai and Matthews [1993] to allow for a revised M-step in which a single step of Newton-Raphson is performed during each iteration. As in the previous section, this will be used to obtain updated estimates of the structured variance-covariance matrices as well as the parameters related to class membership. In addition, the results from Jennrich and Schluchter [1986] will be utilized for the Newton-Raphson calculations in the M-step of the EM algorithm. Gradient- and Hessian-based approaches will similarly be generalized to allow for class-specific variances.

EM Approach

The previous section assumed that the variances were equal across classes. However, if the variances are allowed to differ for each latent class, then the proposed method requires several modifications.

First, most derivations must be revised to account for the fact that different classes have different underlying variances. Second, without a restriction on the class-specific variances, the likelihood for the mixture of multivariate normal distributions is unbounded. Therefore, the class-specific variances must be constrained. Finally, when the residual error variances are fit as being class-specific, the calculation of $\widehat{\boldsymbol{\beta}}$ itself requires maximization over a mixture of multivariate normal distributions. For completeness, I will present the entire method and note how the method differs from the equal across classes variance case.

In the case of complete data, the likelihood corresponding to the joint distribution of the observed data and group classifications $(\mathbf{y}_i, \mathbf{c}_i)$, rewritten to reflect class-specific variances, is as follows:

$$\log L_C(\boldsymbol{\Theta}) = \sum_{i=1}^{n} \sum_{k=1}^{K} c_{ik} \log \left[\pi_{ik} f(\mathbf{y}_i \mid c_{ik}=1)\right] \tag{1.41}$$

with $f(\mathbf{y}_i \mid c_{ik}=1)$ being the density defined by

$$(\mathbf{y}_i \mid c_{ik}=1) \sim N(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{W}_i\boldsymbol{\lambda}_k, \mathbf{Z}_i\mathbf{D}_k\mathbf{Z}_i' + \mathbf{R}_{ik})$$

Note that in the revised specification, $\mathbf{D}$ is replaced with $\mathbf{D}_k$, and $\mathbf{R}_i$ is replaced by $\mathbf{R}_{ik}$ to reflect the fact that the variances are class-specific.

Once again, since the class membership of each subject is unobservable, the complete data likelihood is replaced during each E-step with its conditional expectation given the observed data $\mathbf{y}_i$, using the current estimates of the parameters $\boldsymbol{\Theta}^*$. As noted earlier, this function is referred to in Dempster et al [1977] as $Q(\boldsymbol{\Theta}; \boldsymbol{\Theta}^*) = \mathrm{E}_{\boldsymbol{\Theta}^*}\{\log L_C(\boldsymbol{\Theta}) \mid \mathbf{y}\}$. The only change that is required in going from $\log L_C(\boldsymbol{\Theta})$ to $Q(\boldsymbol{\Theta}; \boldsymbol{\Theta}^*)$ is that the value of c_{ik} is replaced by its expected value $\widetilde{c}_{ik}$. It follows that:

$$Q(\boldsymbol{\Theta}; \boldsymbol{\Theta}^*) = \sum_{i=1}^{n} \sum_{k=1}^{K} \widetilde{c}_{ik} \left[\begin{array}{l} \log \pi_{ik} - \frac{n_i}{2} \log 2\pi - \frac{1}{2} \log \left|\mathbf{Z}_i\mathbf{D}_k\mathbf{Z}_i' + \mathbf{R}_{ik}\right| \\ -\frac{1}{2}(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{W}_i\boldsymbol{\lambda}_k)' (\mathbf{Z}_i\mathbf{D}_k\mathbf{Z}_i' + \mathbf{R}_{ik})^{-1} (\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{W}_i\boldsymbol{\lambda}_k) \end{array} \right] \tag{1.42}$$

where the expected value of $\widetilde{c}_{ik}$ can be written as follows:

$$\mathrm{E}(c_{ik} \mid \mathbf{y}_i) \;=\; \widetilde{c}_{ik} = P(c_{ik} = 1 \mid \mathbf{y}_i)$$

$$
\begin{aligned}
&= \frac{P\left(c_{ik}=1\right) \times P\left(\mathbf{y}_i \mid c_{ik}=1\right)}{\sum_{j=1}^{K}\left[P\left(c_{ij}=1\right) \times P\left(\mathbf{y}_i \mid c_{ij}=1\right)\right]} \\
&= \frac{\pi_{ik} f\left(\mathbf{y}_i \mid c_{ik}=1\right)}{\sum_{j=1}^{K}\left\{\pi_{ij} f\left(\mathbf{y}_i \mid c_{ij}=1\right)\right\}}
\end{aligned}
$$

with $f\left(\mathbf{y}_i \mid c_{ij}=1\right)$ being the density defined by

$$\left(\mathbf{y}_i \mid c_{ij}=1\right) \sim N(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{W}_i\boldsymbol{\lambda}_k, \mathbf{Z}_i\mathbf{D}_k\mathbf{Z}_i' + \mathbf{R}_{ik})$$

The goal of each M step is then to maximize $Q\left(\boldsymbol{\Theta};\boldsymbol{\Theta}^*\right)$ with respect to the parameters.

Since the variances are different for each latent class, maximization of $Q\left(\boldsymbol{\Theta};\boldsymbol{\Theta}^*\right)$ with respect to $\boldsymbol{\beta}$ requires maximization over a mixture of multivariate normal distributions, each with different variances. Therefore, it is not possible to find a closed-form solution for $\widehat{\boldsymbol{\beta}}$. Fortunately, this is a relatively straightforward problem using SAS® PROC NLP. The inputs for this SAS® procedure are $Q\left(\boldsymbol{\Theta};\boldsymbol{\Theta}^*\right)$, as well as its first and second derivatives with respect to $\boldsymbol{\beta}$.

$$\frac{\partial Q(\boldsymbol{\Theta};\boldsymbol{\Theta}^*)}{\partial\boldsymbol{\beta}} = \sum_{i=1}^{n}\sum_{k=1}^{K}\widetilde{c}_{ik}\left(\mathbf{X}_i'\boldsymbol{\Sigma}_{ik}^{-1}\mathbf{y}_i - \mathbf{X}_i'\boldsymbol{\Sigma}_{ik}^{-1}\mathbf{X}_i\boldsymbol{\beta} - \mathbf{X}_i'\boldsymbol{\Sigma}_{ik}^{-1}\mathbf{W}_i\boldsymbol{\lambda}_k\right) \tag{1.43}$$

$$\frac{\partial^2 Q(\boldsymbol{\Theta};\boldsymbol{\Theta}^*)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}} = -\sum_{i=1}^{n}\sum_{k=1}^{K}\widetilde{c}_{ik}\left(\mathbf{X}_i'\boldsymbol{\Sigma}_{ik}^{-1}\mathbf{X}_i\right) \tag{1.44}$$

with $\boldsymbol{\Sigma}_{ik} = \mathbf{Z}_i\mathbf{D}_k\mathbf{Z}_i' + \mathbf{R}_{ik}$

Since calculations for $\widehat{\boldsymbol{\lambda}_k}$ utilize class-specific variance-covariance matrices, closed-form solutions exist:

$$\frac{\partial Q(\boldsymbol{\Theta};\boldsymbol{\Theta}^*)}{\partial\boldsymbol{\lambda}_k} = \sum_{i=1}^{n}\widetilde{c}_{ik}\left(\mathbf{W}_i'\boldsymbol{\Sigma}_{ik}^{-1}\mathbf{y}_i - \mathbf{W}_i'\boldsymbol{\Sigma}_{ik}^{-1}\mathbf{X}_i\boldsymbol{\beta} - \mathbf{W}_i'\boldsymbol{\Sigma}_{ik}^{-1}\mathbf{W}_i\boldsymbol{\lambda}_k\right) \tag{1.45}$$

$$\widehat{\boldsymbol{\lambda}_k} = \left(\sum_{i=1}^{n}\widetilde{c}_{ik}\mathbf{W}_i'\boldsymbol{\Sigma}_{ik}^{-1}\mathbf{W}_i\right)^{-1}\sum_{i=1}^{n}\widetilde{c}_{ik}\mathbf{W}_i'\boldsymbol{\Sigma}_{ik}^{-1}\left(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}\right) \tag{1.46}$$

The expected complete-data likelihood $Q(\boldsymbol{\Theta};\boldsymbol{\Theta}^*)$, evaluated at the revised $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\lambda}_k}$, can then be used in a single Newton-Raphson step to calculate structured, class-specific variances. Following the notation in Jennrich and Schluchter [1986], the r^{th} element of the score vector and $r-s^{th}$ element of the Hessian can be computed as follows:

$$\left[\mathbf{s}_{\boldsymbol{\theta}}\right]_r = \tfrac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{K}\widetilde{c}_{ik} tr\left[\boldsymbol{\Sigma}_{ik}^{-1}\left(\mathbf{e}_{ik}\mathbf{e}_{ik}' - \boldsymbol{\Sigma}_{ik}\right)\boldsymbol{\Sigma}_{ik}^{-1}\dot{\boldsymbol{\Sigma}}_{ik,r}\right] \tag{1.47}$$

$$\begin{aligned}\left[\mathbf{H}_{\boldsymbol{\theta\theta}}\right]_{rs} &= -\tfrac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{K}\widetilde{c}_{ik} tr\left[\boldsymbol{\Sigma}_{ik}^{-1}\dot{\boldsymbol{\Sigma}}_{ik,r}\boldsymbol{\Sigma}_{ik}^{-1}\left(2\mathbf{e}_{ik}\mathbf{e}_{ik}' - \boldsymbol{\Sigma}_{ik}\right)\boldsymbol{\Sigma}_{ik}^{-1}\dot{\boldsymbol{\Sigma}}_{ik,s}\right] \\ &\quad +\tfrac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{K}\widetilde{c}_{ik} tr\left[\boldsymbol{\Sigma}_{ik}^{-1}\left(\mathbf{e}_{ik}\mathbf{e}_{ik}' - \boldsymbol{\Sigma}_{ik}\right)\boldsymbol{\Sigma}_{ik}^{-1}\ddot{\boldsymbol{\Sigma}}_{ik,rs}\right]\end{aligned} \tag{1.48}$$

where:

$$\begin{aligned}\mathbf{e}_{ik} &= \mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{W}_i\boldsymbol{\lambda}_k \\ \dot{\boldsymbol{\Sigma}}_{ik,r} &= \partial\boldsymbol{\Sigma}_{ik}/\partial\boldsymbol{\theta}_r \\ \ddot{\boldsymbol{\Sigma}}_{ik,rs} &= \partial^2\boldsymbol{\Sigma}_{ik}/\partial\boldsymbol{\theta}_r\partial\boldsymbol{\theta}_s\end{aligned}$$

Note that $\boldsymbol{\theta}$ now contains separate variances for each latent class. Since $\dot{\boldsymbol{\Sigma}}_{ik,r}$ and $\ddot{\boldsymbol{\Sigma}}_{ik,rs}$ are non-zero only for components of class k, the summation over k is not necessary. However, since it may be desirable to consider the variances in *either* **D** or **R** as being equal across classes, I have chosen to keep the double-summation to allow for this possibility.

As in the previous section, since closed-form solutions for the $\boldsymbol{\alpha}_k$ are not available, a scoring procedure is also used for calculating estimates of the parameters related to class membership. The same procedure used in the previous section can be used here as well. The updated estimates of the variance components and parameters related to class membership can once again be computed using Equations 1.19 and 1.20.

Recall that Hathaway [1985] set forth a condition for the existence of a consistent maximum likelihood estimator for mixtures of multivariate normal distributions. The condition was that the eigenvalues of $\left[\boldsymbol{\Sigma}_j\boldsymbol{\Sigma}_k^{-1}\right]$ must be $> c\ \forall_{j,k}$, where $\boldsymbol{\Sigma}_j$ represents the variance-covariance matrix for class j and $\boldsymbol{\Sigma}_k$ represents the variance-covariance matrix for class k. In this algorithm, c is prespecified as a small number such as $\frac{1}{10}$ in order to restrict any variance in one class from becoming much smaller than the corresponding variance in another class. In this section, only diagonal specifications of the $\mathbf{D}_k$ and $\mathbf{R}_{ik}$ matrices are considered. Since the eigenvalues of a diagonal matrix are the elements on the diagonal, this constraint on the variances amounts to the following procedure.

For the **D** matrices:

- Examine element (1,1) of each of the matrices $\mathbf{D}_k$
- Define the maximum of these values to be VMax
- If any of the (1,1) elements is less than $c \times$ VMax then set that element equal to $c \times$ VMax
- Repeat this procedure for each of the diagonal elements of $\mathbf{D}_k$

For the $\mathbf{R}$ matrices:

- Examine the value of the first unique variance in the $\mathbf{R}_{ik}$ matrices for each of the k classes
- Define the maximum of these values to be VMax
- If the estimate of this variance in any of the latent classes is less than $c \times$ VMax then set that element equal to $c \times$ VMax
- Repeat this procedure for each of the unique variance in the $\mathbf{R}_{ik}$

We repeat the iterations until the maximum percent change in parameter estimates from one iteration to the next is less than a prespecified tolerance. A proof is presented in the appendices which shows that applying Hathaway's constraint separately to both the variances in $\mathbf{D}$ and $\mathbf{R}$ will result in variance-covariance matrices that satisfy Hathaway's constraint, leading the resulting estimates to be strongly consistent.

Gradient- and Hessian-Based Approaches

Gradient-based algorithms rely on the calculation of both the unconditional log likelihood of the observed data and the first derivatives of this likelihood with respect to the unknown parameters. Hessian-based methods require the additional calculation of the second derivatives of this likelihood. The unconditional log likelihood of the observed data, assuming class-specific diagonal $\mathbf{D}$ and $\mathbf{R}$, is as follows:

$$\log L(\boldsymbol{\Theta}) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \left[\pi_{ik} f(\mathbf{y}_i \mid c_{ik} = 1)\right] \tag{1.49}$$

with $f(\mathbf{y}_i \mid c_{ik} = 1)$ being the density defined by

$$(\mathbf{y}_i \mid c_{ik} = 1) \sim N(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{W}_i\boldsymbol{\lambda}_k, \mathbf{Z}_i\mathbf{D}_k\mathbf{Z}_i' + \mathbf{R}_{ik})$$

In full,

$$\log L\left(\boldsymbol{\Theta}\right) =$$
$$\sum_{i=1}^{n} \log \underbrace{\sum_{k=1}^{K} \left\{ \underbrace{\left[\frac{\exp\left(\mathbf{t}_i' \boldsymbol{\alpha}_k\right)}{\sum_{j=1}^{K} \exp\left(\mathbf{t}_i' \boldsymbol{\alpha}_j\right)} \right]}_{Q_{2ik}} \underbrace{\left[\frac{\exp\left[-\frac{1}{2}\left(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{W}_i\boldsymbol{\lambda}_k\right)' \boldsymbol{\Sigma}_{ik}^{-1} \left(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{W}_i\boldsymbol{\lambda}_k\right)\right]}{\left(2\pi\right)^{n_i/2} \left|\boldsymbol{\Sigma}_{ik}\right|^{1/2}} \right]}_{Q_{3ik}} \right\}}_{Q_{1i}} \tag{1.50}$$

$$\text{with } \boldsymbol{\Sigma}_{ik} = \mathbf{Z}_i \mathbf{D}_k \mathbf{Z}_i' + \mathbf{R}_{ik}$$

The score vector $\mathbf{s}$, containing the first derivatives of log $L\left(\boldsymbol{\Theta}\right)$ with respect to the parameters, and Hessian matrix $\mathbf{H}$, containing the second derivatives of log $L\left(\boldsymbol{\Theta}\right)$ with respect to the parameters, have the same form as in the previous section. However, note that now $\boldsymbol{\theta}$ consists of variances which may be class-specific. Further details regarding the first and second derivatives are presented in the appendix. Once again, for a Newton-Raphson step, the updated estimates of the parameters are computed according to Equation 1.40.

As in the EM approach, a constraint is applied to each of the variances in $\mathbf{D}$ and $\mathbf{R}$ to ensure that for each variance, the estimated value for one class is at least a pre-specified multiple of that variance in other classes.

1.4.3 LCLMM: Structured Class-Specific Variance Components, Structured D and Diagonal R

Motivation and Known Results

While the diagonal specification of the $\mathbf{D}_k$ should allow for a fairly general model specification, it would be advantageous to allow correlation among the random effects and allow the statistician to specify the applicable correlation structure. However, allowing both the correlations and variances in $\mathbf{D}_k$ to vary by latent class results in instances where the variance components could be adjusted in many possible ways to satisfy the constraints proposed by Hathaway [1985]. Therefore, I propose to parameterize the variance-covariance matrices of the random effects, $\mathbf{D}_k$, as follows. First, the correlation structure for the random effects is assumed to be the same across all latent classes. Second, class-specific variances are permitted. Given this structure, I propose to follow a similar approach to that used to constrain the diagonal variance-covariance matrices in the previous chapter. Each individual variance estimate is compared from class to class to ensure that no variance estimate becomes too small when compared to the corresponding variance estimates in other classes.

Required Modifications

Only one change is required to fit the revised model with a structured, non-diagonal, class-specific **D**. The parameters that determine the variance-covariance matrices, contained in $\boldsymbol{\theta}$, now consist of both correlations, which are calculated across latent classes, and variances, which are calculated as being specific to each latent class. As such, the computations for the scoring vector **s** and Hessian **H** will include a summation over classes for the correlations, but will require summations only for the relevant class for the variances.

Note Related to More General Correlation Structures

The previous sections generalized existing methods by first allowing for the specification of a structured variance-covariance matrix with variances equal across classes. Variances were then permitted to vary by latent class but correlation was not permitted between random effects. And finally, variances were permitted to vary by latent class, and a correlation matrix for the random effects was fit across classes.

While these models provide for a fairly general specification of the underlying LMMs, they can easily be modified to allow for particular forms of the variance-covariance structure. For example, the random effects portion could be dropped, and the variance-covariance matrix of the error terms could be specified according to another linear specification. Another alternative would be to keep the structured variance-covariance matrix of the random effects and allow correlation in the variance-covariance matrix of the error terms. Only a few of the various possibilities have been examined in this research, but many more could be easily implemented in future research.

1.4.4 LCLMM: The Relative-Fit Class Membership Model

In previous sections, the LCLMM has been generalized to allow for class-specific, structured, variance-covariance matrices. In these descriptions, class membership has been assumed to be determined based on an underlying logistic regression model. In many cases, the statistician may not know exactly which factors should be included in the model describing class membership, these factors may not be measurable, or the statistician may prefer to simply identify the best-fitting LCLMM with K classes.

However, this model, in which each subject can have its own set of mixture probabilities, would result in many parameters related to class membership - specifically, for n subjects and K classes, the model would require $n \times (K-1)$ parameters. Therefore, I propose in this section a new model which builds upon the generalizations already implemented but replaces the logistic regression model with a model in which class membership is determined by the relative fit of the underlying LMMs. This is much less costly in terms of computational time than fitting a separate set of class membership parameters for each subject, but offers a logical approach with a similar goal in mind.

Recall that the unconditional log likelihood of the observed data, assuming class-specific $\mathbf{D}$ and $\mathbf{R}$, is as follows:

$$\log L(\mathbf{\Theta}) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \left[\pi_{ik} f\left(\mathbf{y}_i \mid c_{ik}=1\right)\right] \tag{1.51}$$

with $f\left(\mathbf{y}_i \mid c_{ik}=1\right)$ being the density defined by

$$(\mathbf{y}_i \mid c_{ik}=1) \sim N(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{W}_i\boldsymbol{\lambda}_k, \mathbf{Z}_i\mathbf{D}_k\mathbf{Z}_i' + \mathbf{R}_{ik})$$

In the revised model, the π_{ik} are determined solely by the relative fit of the underlying LMMs according to the following equation:

$$\hat{\pi}_{ik} = \frac{f\left(\mathbf{y}_i \mid c_{ik}=1; \hat{\mathbf{\Theta}}\right)}{\sum_{j=1}^{K} f\left(\mathbf{y}_i \mid c_{ij}=1; \hat{\mathbf{\Theta}}\right)} \tag{1.52}$$

For example, if two latent classes are fit and the LMM for Class 1 results in a likelihood for the subject's data equal to the likelihood of that subject's data under the model for Class 2, then the subject would be included in each class 50%/50%. If the likelihood for Class 1 was nine times higher than for Class 2, then the subject would be included in Class 1 90 percent and Class 2 only 10 percent.

The gradient for this likelihood has a similar, slightly more complicated, form to that described in the previous sections:

$$s_{\boldsymbol{\beta}} = \sum_{i=1}^{n} \frac{1}{\sum_{k=1}^{K} Q_{3ik}^2} \times \sum_{k=1}^{K} 2 \times Q_{3ik} \times \frac{\partial Q_{3ik}}{\partial \boldsymbol{\beta}} - \sum_{i=1}^{n} \frac{1}{\sum_{k=1}^{K} Q_{3ik}} \times \sum_{k=1}^{K} \frac{\partial Q_{3ik}}{\partial \boldsymbol{\beta}} \tag{1.53}$$

$$s_{\boldsymbol{\lambda}_k} = \sum_{i=1}^{n} \frac{1}{\sum_{k=1}^{K} Q_{3ik}^2} \times 2 \times Q_{3ik} \times \frac{\partial Q_{3ik}}{\partial \boldsymbol{\lambda}_k} - \sum_{i=1}^{n} \frac{1}{\sum_{k=1}^{K} Q_{3ik}} \times \frac{\partial Q_{3ik}}{\partial \boldsymbol{\lambda}_k} \tag{1.54}$$

$$s_{\boldsymbol{\theta}} = \sum_{i=1}^{n} \frac{1}{\sum_{k=1}^{K} Q_{3ik}^2} \times \sum_{k=1}^{K} 2 \times Q_{3ik} \times \frac{\partial Q_{3ik}}{\partial \boldsymbol{\theta}} - \sum_{i=1}^{n} \frac{1}{\sum_{k=1}^{K} Q_{3ik}} \times \sum_{k=1}^{K} \frac{\partial Q_{3ik}}{\partial \boldsymbol{\theta}} \tag{1.55}$$

The Hessian for this likelihood also has a similar, though somewhat more complicated, form to that described in the previous sections. The equation for the second derivative with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ is provided as an example. The other second derivatives have a similar form.

$$\begin{aligned} H_{\boldsymbol{\beta}\boldsymbol{\theta}} &= \sum_{i=1}^{n} 2 \times \left(\sum_{k=1}^{K} Q_{3ik}^2\right)^{-1} \left(\sum_{k=1}^{K} Q_{3ik} \times \frac{\partial^2 Q_{3ik}}{\partial \boldsymbol{\beta} \partial \boldsymbol{\theta}} + \frac{\partial Q_{3ik}}{\partial \boldsymbol{\theta}} \frac{\partial Q_{3ik}}{\partial \boldsymbol{\beta}}\right) \\ &+ \sum_{i=1}^{n} (-4) \times \left(\sum_{k=1}^{K} Q_{3ik}^2\right)^{-2} \left(\sum_{k=1}^{K} Q_{3ik} \times \frac{\partial Q_{3ik}}{\partial \boldsymbol{\theta}}\right) \left(\sum_{k=1}^{K} Q_{3ik} \times \frac{\partial Q_{3ik}}{\partial \boldsymbol{\beta}}\right) \\ &+ \sum_{i=1}^{n} (-1) \times \left(\sum_{k=1}^{K} Q_{3ik}\right)^{-1} \left(\sum_{k=1}^{K} \frac{\partial^2 Q_{3ik}}{\partial \boldsymbol{\beta} \partial \boldsymbol{\theta}}\right) \\ &+ \sum_{i=1}^{n} \left(\sum_{k=1}^{K} Q_{3ik}\right)^{-2} \left(\sum_{k=1}^{K} \frac{\partial Q_{3ik}}{\partial \boldsymbol{\theta}}\right) \left(\sum_{k=1}^{K} \frac{\partial Q_{3ik}}{\partial \boldsymbol{\beta}}\right) \end{aligned}$$

1.4.5 Prediction of the Random Effects

One of the many advantages to using LMMs is the logical specification of the model, and a big part of this specification lies in the random effects which are thought of as being sampled from an underlying distribution. For the LMM, the underlying distribution is the multivariate normal distribution. The LCLMM, in essence, divides the population into subpopulations and then fits the best LMMs to the data in each subpopulation. Therefore, it is important that the LCLMM approach be able to compute predictions of the random effects. In fact, upon closer examination, the random effects for the LCLMM are actually a weighted average of the random effects from each underlying LMM.

From the derivation of the LCLMM, recall that given a subject is a member of class k, the following distributional properties are assumed to apply to the random effects, $\mathbf{b}_i$, and residuals, $\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{W}_i\boldsymbol{\lambda}_k$.

$$\left[\begin{array}{c} \mathbf{b}_i \\ \mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{W}_i\boldsymbol{\lambda}_k \end{array}\right] \sim N\left(\left[\begin{array}{c} \mathbf{0} \\ \mathbf{0} \end{array}\right], \left[\begin{array}{cc} \mathbf{D}_k & \mathbf{D}_k\mathbf{Z}_i' \\ \mathbf{Z}_i\mathbf{D}_k & \mathbf{Z}_i\mathbf{D}_k\mathbf{Z}_i' + \mathbf{R}_i \end{array}\right]\right) \Bigg| \quad c_{ik} = 1 \tag{1.56}$$

The computation of the LCLMM random effects requires maximizing the joint density of the data $\mathbf{y}$

and the random effects $\mathbf{b}_i$ with respect to the terms $\mathbf{b}_i$. The likelihood can be written as follows:

$$\log L\left(\mathbf{y}, \mathbf{b}\right) =$$
$$\sum_{i=1}^{n} \log \underbrace{\sum_{k=1}^{K} \pi_{ik} \underbrace{\left[\frac{\exp\left[-\frac{1}{2}\left\{\left(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{W}_i\boldsymbol{\lambda}_k - \mathbf{Z}_i\mathbf{b}_i\right)' \mathbf{R}_{ik}^{-1}\left(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{W}_i\boldsymbol{\lambda}_k - \mathbf{Z}_i\mathbf{b}_i\right) + \mathbf{b}_i'\mathbf{D}_k^{-1}\mathbf{b}_i\right\}\right]}{(2\pi)^{(n_i+q)/2}\left|\mathbf{R}_{ik}\right|^{1/2}\left|\mathbf{D}_k\right|^{1/2}} \right]}_{P_{3ik}}}_{P_{1i}}$$

However, given that a subject is in class k implies that the best set of predictions of the random effects are the predictions based on the k^{th} LMM. Therefore, the random effects for a subject can be computed as the weighted average of the random effects computed as if that subject were in each latent class:

$$\widetilde{\mathbf{b}}_i \;=\; \text{eBLUP}\left[\mathbf{b}_i\right] \;=\; \text{E}\left[\mathbf{b}_i \mid \hat{\boldsymbol{\Theta}}, \mathbf{y}_i\right] \;=\; \sum_{i=1}^{K} \pi_{ik} \times \widetilde{\mathbf{b}}_{i|k}$$

The resulting random effects are available in the SAS® macro discussed in Chapter 2.

For purposes of calculating standard errors, the second derivative of the above likelihood was derived using the following equations. Note that since the solutions for the random effects themselves can be calculated in closed-form, the gradients used in the second derivative are zero and therefore the form of the Hessian is relatively simple. First and second derivatives are provided in the appendix.

$$s_{\mathbf{b}_i} \;=\; \sum_{i=1}^{n} \frac{1}{P_{1i}} \sum_{k=1}^{K} \pi_{ik} \frac{\partial P_{3ik}}{\partial \mathbf{b}_i} \tag{1.57}$$

$$H_{\mathbf{b}_i\mathbf{b}_i} \;=\; \sum_{i=1}^{n} \left\{ \left[\frac{1}{P_{1i}} \sum_{k=1}^{K} \pi_{ik} \frac{\partial^2 P_{3ik}}{\partial \mathbf{b}_i \partial \mathbf{b}_i} \right] \right\} \tag{1.58}$$

Written in a slightly different way, note that the Hessian is effectively a weighted average of the second derivatives for each underlying LMM, with the weights based on a combination of the fitted class probabilities and the relative fit of the underlying LMMs.

$$H_{\mathbf{b}_i\mathbf{b}_i} = \sum_{i=1}^{n} \sum_{k=1}^{K} \frac{\pi_{ik} P_{3ik}}{\sum_{j=1}^{K} \pi_{ij} P_{3ij}} \left\{ \begin{array}{l} \left[-\mathbf{Z}_i'\mathbf{R}_{ik}^{-1}\mathbf{Z}_i - \mathbf{D}_k^{-1}\right] + \\ \left[\mathbf{Z}_i'\mathbf{R}_{ik}^{-1}\left(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{W}_i\boldsymbol{\lambda}_k - \mathbf{Z}_i\mathbf{b}_i\right) - \mathbf{D}_k^{-1}\mathbf{b}_i\right] \\ \left[\mathbf{Z}_i'\mathbf{R}_{ik}^{-1}\left(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{W}_i\boldsymbol{\lambda}_k - \mathbf{Z}_i\mathbf{b}_i\right) - \mathbf{D}_k^{-1}\mathbf{b}_i\right]' \end{array} \right\}$$

It is interesting to note that the calculations for the Hessian and therefore the standard errors of the random effects will include information related to likely class membership, even in situations where all subjects are fit with the same mixture. Random effect standard errors based on the computed Hessian are calculated in the SAS® macro discussed in Chapter 2.

1.4.6 Computational Methods

The EM and Newton-Raphson approaches have been examined in detail in previous sections. As noted earlier, since the analytical first and second derivatives are available, several computational algorithms can easily be applied in the estimation process. In the SAS® macro presented in Chapter 2, IML modules have been programmed which calculate the first and second derivatives of the likelihood. The engines in SAS®' PROC NLP are then used to provide for a varied set of methods. In the sections below, a brief literature review is presented related to computational methods which have been used in previous research. Details related to SAS®PROC NLP are available in Chapter 4 of the SAS®/OR 9.2 User's Guide [2008].

Literature Review

Most previous research related to LCLMMs has utilized the EM algorithm described in Dempster et al [1977] to arrive at parameter estimates. Since the LCLMM represents an instance where the data can be considered to be incomplete (i.e., class membership) and the complete-data model is straightforward, the EM algorithm is an obvious candidate. As noted in Section 1.3, Lin et al [2000], Lin et al [2002a], Lin et al [2004], Lin et al [2002b], McCulloch [2003], and Celeux et al. [2005] all used the EM algorithm for calculations. Roy [2003] computed estimates via a modified Newton-Raphson algorithm in which second derivatives were calculated numerically. Spiessens et al [2002] implemented the version of the Heterogeneity Model described in Verbeke and Lesaffre [1996] in a SAS® macro using PROC NLMIXED, which uses a quasi-Newton algorithm as the default. Xu and Hedeker [2001] proposed a model similar to Verbeke and Molenberghs [2000] and Spiessens et al [2002], in which the authors provided details related to both EM and Fisher scoring algorithms. Proust and Jacqmin-Gadda [2005] used a modified Marquardt optimization algorithm, similar to Newton-Raphson, to solve for the parameters in a slightly more general model than Verbeke and Lesaffre [1996]. Note that those articles which discuss estimation algorithms other than EM are simpler than the model proposed in this research. The first set of articles which utilize the EM algorithm are reasonably similar in complexity but do not offer the general model specification presented in previous sections.

Since the EM algorithm has been observed to converge slowly in practice, many researchers have explored methods to accelerate convergence. These methods have been applied to a variety of problems, not necessarily related to mixture models. The methods range from simple methods which multiply changes in parameter estimates to quasi-Newton methods which update an estimate of the second derivatives of the likelihood function, thereby gradually resembling a Newton-Raphson algorithm. Laird et al. [1987] used univariate and multivariate Aitken acceleration to improve the speed of convergence of the EM algorithm. This method multiplies the changes in estimates at each EM iteration by a constant (univariate approach) or a matrix (multivariate approach). The authors constructed the matrices for the multivariate approach from the successive changes in the parameters over preceding iterations. Jamshidian and Jennrich [1993] proposed a method to accelerate the EM algorithm based on conjugate gradients. The authors showed how the change in estimates from one EM iteration to the next can be viewed as a generalized gradient of log $L(\mathbf{\Theta})$, therefore lending itself to the application of generalized conjugate gradient methods. Jamshidian and Jennrich [1997] proposed several methods for accelerating the EM algorithm, divided into three groups: pure EM accelerators, EM-type accelerators, and hybrid accelerators. Method QN1 examines the changes in parameter estimates from iteration to iteration which would result from running the next EM step. These are used at each step to update an approximate Jacobian matrix, with the resulting matrix applied to the EM-step in a simple quasi-Newton method. Method QN2 is similar to QN1. However, it also has as its inputs the gradient and Hessian of the complete-data likelihood. Lange [1995b] proposed to replace each M-step with a single iteration of Newton-Raphson, using the first and second derivatives of $Q(\mathbf{\Theta};\mathbf{\Theta}^*)$. This algorithm is typically referred to as the EM gradient algorithm. Lange [1995a] proposed to use Lange [1995b] as the basis for a quasi-Newton method. In addition to working with the first and second derivatives of $Q(\mathbf{\Theta};\mathbf{\Theta}^*)$, the algorithm updates an estimate of the Hessian at each iteration.

Everitt [1984] compared several different computational algorithms for estimation of the parameters in a mixture of two univariate normal distributions. The author found that Newton's method using exact values of the first and second derivatives converged most rapidly and was generally very stable. The EM algorithm was also very stable, but had occasions where convergence was very slow. Aitkin and Aitkin [1996] proposed a hybrid EM/Gauss-Newton algorithm to compute maximum likelihood estimates for mixture models. The proposed algorithm began with five EM iterations and then switched to Gauss-Newton until either convergence was achieved or the log likelihood decreased. If the log likelihood decreased, then the step size was halved up to five times to try to find a step which would increase the likelihood. If the likelihood still decreased after five step-halves, then the EM algorithm was run for five additional iterations from the most recent estimates. A switch to EM was also used if the Hessian was not positive definite in any Gauss-Newton iteration. The authors found 30-40 percent faster time to convergence in the hybrid algorithm. They also found that the initial Gauss-Newton steps almost always decreased the log-likelihood. In some instances, many steps were needed before a Gauss-Newton step increased the likelihood.

Gradient- and Hessian-Based Algorithms Available in SAS® PROC NLP

In Chapter 2, a SAS® macro is presented which implements the EM algorithm presented earlier as well as several gradient and Hessian-based methods. These gradient- and Hessian-based methods use the analytical gradient and/or Hessian along with the engines in SAS®' PROC NLP to obtain estimates. The available methods include conjugate gradient, quasi-Newton, and Newton-Raphson. Additional details and references can be found in Chapter 4 of the SAS®/OR 9.2 User's Guide [2008].

1.4.7 Calculation of Parameter Standard Errors

One advantage to using the Newton-Raphson algorithm for likelihood maximization is that standard errors of the estimates are produced as part of the estimation procedure. Several techniques are available to obtain estimates of parameter standard errors when the EM or gradient-based algorithms are used, although some have had mixed reviews in practice. Meng and Rubin [1991] presented a method for calculating standard errors called the Supplemented EM Algorithm. This method operates by numerically differentiating the EM operator $M(\mathbf{\Theta})$. However, several authors who have applied the Supplemented EM algorithm have run into difficulties in practice. For instance, in McCulloch [1998], the author pointed out that 'for many problems, the Meng and Rubin [1991] method of obtaining standard errors can be numerically unstable.' Jamshidian and Jennrich [2000] examined three methods of obtaining parameter standard errors, one which numerically differentiates the score function, and two which numerically differentiate the M-function, as in Meng and Rubin [1991]. The authors concluded that the method which operates by numerically differentiating the score function was preferred. The quasi-Newton methods described in the previous section also can produce estimates of parameter standard errors as a part of the estimation procedure, similar to Newton-Raphson. In the QN2 method proposed in Jamshidian and Jennrich [1997], an estimate of the Hessian was updated during each iteration. At convergence, the authors proposed to use this approximate Hessian to estimate the standard errors of the parameters. Similarly, a quasi-Newton method proposed in Lange [1995a], which is based on Lange [1995b], also updated an approximate Hessian as part of each iteration.

Building on the methods presented in earlier sections, two practical approaches to estimating parameter standard errors exist which are not too computationally demanding. Since the first and second derivatives of the log likelihood have now been programmed into a publicly available SAS® macro (see Chapter 2), the Hessian is available for use in obtaining estimates of the parameter standard errors. If the computational demands are too great to compute the Hessian, a second option is available. Based on Jamshidian and Jennrich [2000], the Hessian approximated by numerically differentiating the gradient vector appears to provide a viable alternative. Since calculating the gradient is relatively cheap in terms of computational time, this should work for most, if not all, scenarios. Note that both of these methods are available in the SAS® macro discussed in Chapter 2. Since the approximate Hessian updated in the quasi-Newton method is not currently available using SAS®' IML link to PROC NLP, this method was not investigated.

Note that the estimated variance-covariance matrix of the parameters, $\mathbf{V}$, is equal to $-\mathbf{H}^{-1}$, where $\mathbf{H}$ is the Hessian, or the second derivative of the log likelihood.

1.5 Application: Atherosclerosis Risk in Communities (ARIC)

1.5.1 Background

The Atherosclerosis Risk in Communities Study (ARIC) is a prospective epidemiologic study designed to investigate the etiology and natural history of atherosclerosis, the etiology of clinical atherosclerotic diseases, and the variation in cardiovascular risk factors, medical care and disease by race, gender, location, and date. The study was conducted in four communities in the United States - Forsyth County, NC, Jackson, MS, the northwestern suburbs of Minneapolis, MN, and Washington County, MD. Each ARIC field center randomly selected and recruited a cohort sample of approximately 4,000 individuals aged 45-64 from a defined population in their community. A total of 15,792 individuals participated. Study participants received an extensive examination, including medical, social, and demographic data. These participants were reexamined every three years with the first screen (baseline) occurring in 1987-89, the second in 1990-92, the third in 1993-95, and the fourth and final exam in 1996-98. Follow-up occurs yearly by telephone to maintain contact with participants and to assess the health status of the cohort.

A subset of 2,066 members of the ARIC study cohort participated in the Carotid MRI Substudy in 2004-2005. The goal was to recruit 1,200 participants with high values of maximum carotid artery wall thickness at their last ultrasound examination, and 800 individuals randomly sampled from the remainder of the carotid artery wall thickness distribution. Participants had a maximum of the four ARIC cohort examinations, plus one Carotid MRI Substudy examination. Measurement protocols were identical at all five visits.

Fasting blood samples were collected at each examination and assayed for total cholesterol, HDL cholesterol, and triglycerides. LDL cholesterol was calculated according to the Friedewald formula. LDL cholesterol, HDL cholesterol, and triglycerides are commonly considered risk factors for coronary artery disease (CAD) and other related diseases or events and are also routinely measured at annual physicals in the general population. Ballantyne [1998] notes that "clinical trials with statins and other lipid-regulating therapies have conclusively shown that lowering LDL cholesterol decreases both morbidity and mortality from CAD and other vascular diseases." Kwiterovich [1998] notes that "the Framingham Heart Study produced compelling epidemiologic evidence indicating that a low level of HDL cholesterol was an independent predictor of coronary artery disease." And Gotto [1998] points out that "the current evidence argues compellingly for including triglycerides in the evaluation of patient risk for CAD. ... The revived attention to hypertriglyceridemia with respect to increased CAD risk represents an important step in assessing a patient's global risk for developing CAD."

For purposes of analysis, lipid data for the 2,066 individuals in the ARIC Carotid MRI substudy, as well as information related to medication use for controlling cholesterol levels were utilized in the models discussed below. Since the distribution of triglycerides is non-normal, a log transformation for this lab value was used in the analyses.

1.5.2 Methods

The goal of this application is to demonstrate the usefulness of the LCLMM for modeling the data for LDL cholesterol, HDL cholesterol, and triglycerides simultaneously. Since this data represents an example in which each individual has repeated measurements on each lab parameter, the LMM represents one possible method which could be used for analysis. However, this model assumes that the population being studied does not actually consist of several subpopulations, or that those subpopulations are known a priori. However, this assumption may not be a reasonable one. Factors such as cultural background, dietary habits, exercise, etc. all contribute to changes in these lab parameters, and there are likely many additional factors. If the statistician attempted to include all of the associated variables in the model, the resulting model would become too complex for practical use. In addition, questions would arise as to whether these additional variables truly describe the population's behavior or whether they are simply acting as markers or surrogates, attempting to identify a finite number of underlying subpopulations that exist in the data. The LCLMM will be used to simultaneously divide the population into a series of subpopulations while also fitting the best LMM to the data for each of the underlying subpopulations. The resulting estimates will be compared with those from the LMM, and the underlying latent classes will be examined.

The LMM was described in Section 1.2. In this example, the LMM is fit with an intercept and linear and quadratic terms for age for each of the three lipid parameters. Two indicator variables - one for LDL cholesterol and one for HDL cholesterol - are also included to account for the effect of cholesterol medication on these measures. Random intercepts are included for all three lipid parameters, and a random slope is fit for LDL (the variances for the random slopes for HDL and triglycerides were close to 0). The random effects are fit as being correlated via an unstructured variance-covariance matrix, and the error terms are assumed to be uncorrelated, with separate variances for each lipid parameter. Note that age is centered at 59 for purposes of the analysis for both the LMM and LCLMM.

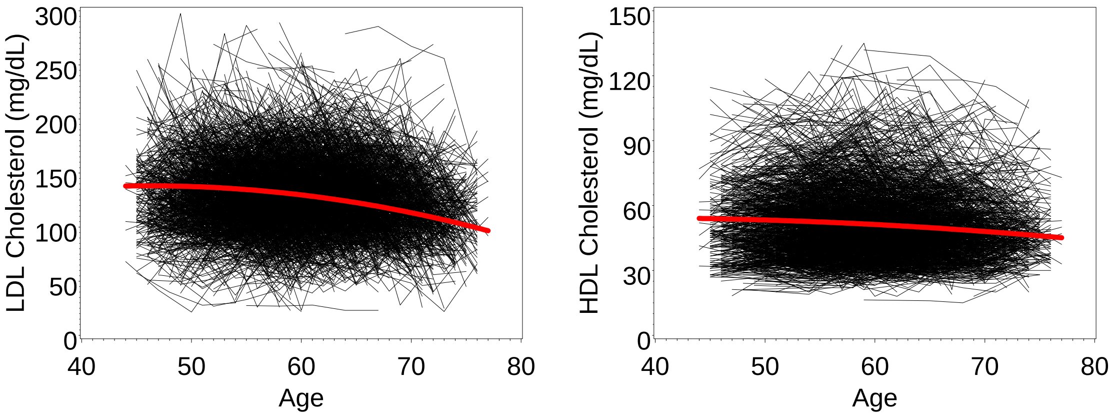
The LCLMM was described in Section 1.2 and expanded in Section 1.4. For the LCLMM, the intercept and linear and quadratic age terms for HDL, LDL, and triglycerides are allowed to vary for each latent class or subpopulation. However, the effect of the cholesterol-lowering medication on HDL and LDL is fit as being the same for each latent class. Random intercepts for all three lipids and a random slope variable for LDL cholesterol are fit in the LCLMM as in the LMM, with correlation permitted between random effects. As noted earlier, the LCLMM requires that the user specify the number of latent classes to fit. Therefore, the LMM (1 class) as well as the LCLMM with 2-10 classes were fit. For the LCLMM runs, two separate runs were actually made - one which required that all underlying latent classes have equal variances and another which allowed the variances to differ by latent class. Note that while the variances are allowed to differ by latent class, the correlation parameters were fit as being the same for each latent class.

1.5.3 Results

Class Trajectories and Class Membership

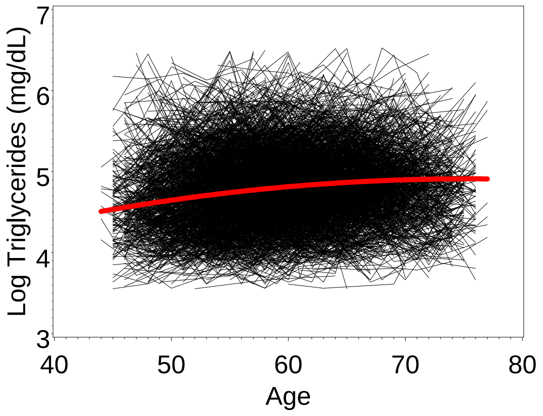
The fitted class trajectories for LDL cholesterol, HDL cholesterol, and triglycerides for the LMM as well as the 2-10 class LCLMMs assuming class-specific variances are displayed in Figures 1.1-1.24. For these plots, the percent of subjects in each class who received cholesterol-lowering medications was used in place of the medication use indicator variables in order to provide an average picture for each latent class. For the 2-5 class models, an additional set of plots is presented in order to provide a more detailed look at the makeup of each latent class. These plots display the observed trajectories for subjects identified in each latent class, along with the fitted class trajectory. For purposes of these plots, a subject is assumed to be in a given class if its highest class probability is at least two times its second-highest class probability.

The LMM, as expected, fits curves through the middle of the data for each lab parameter. The 2-class LCLMM fits two very different patterns. One group, labeled as ’Optimal’, has much higher HDL cholesterol, lower triglycerides, and lower LDL cholesterol than the second group, which is labeled as ’At-Risk’. Notice that LDL cholesterol appears to be decreasing in the ’At-Risk’ group - this is associated with the fact that while both groups tended to be on cholesterol-lowering meds at baseline in equal percentages (20-25 percent), a greater percentage of the ’At-Risk’ group tended to be on cholesterol-lowering meds by visit 5 (56 percent versus 36 percent). The 3-class model appears to add a class which is between the ’Optimal’ and ’At-Risk’ classes, labeled here as ’Average’. Similar to the other models, the 4-class model appears to define an ’Optimal’ group and an ’At-Risk’ group, while fitting two intermediate groups. The two intermediate groups have very similar HDL cholesterol which is roughly midway between the ’Optimal’ and ’At-Risk’ groups. These two groups have different patterns for the other two lab parameters. One of the intermediate groups, labeled ’AvgHDL-HighOther’, has noticeably higher LDL and triglycerides than the group labeled ’AvgHDL-LowOther’. The 5-class model has an ’Optimal’ class, two low HDL classes, which are distinguished by their triglyceride values, and two average HDL classes, distinguished by their LDL values. The 6-class through 10-class models are not described in detail here, but the plots are included with latent classes labeled. As more classes are added, note that there tends to be more overlap between the models.



(a) LDL Cholesterol

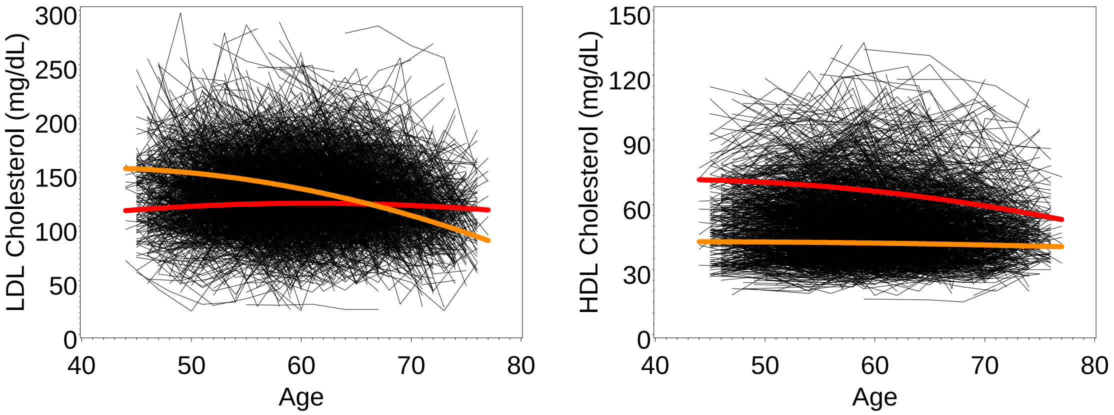
(b) HDL Cholesterol



(c) Triglycerides

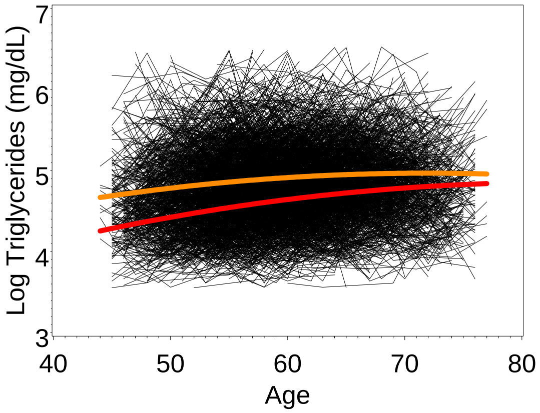
Figure 1.1: ARIC Application: Fitted Lipid Trajectories - Linear Mixed Model

Red = All Subjects



(a) LDL Cholesterol

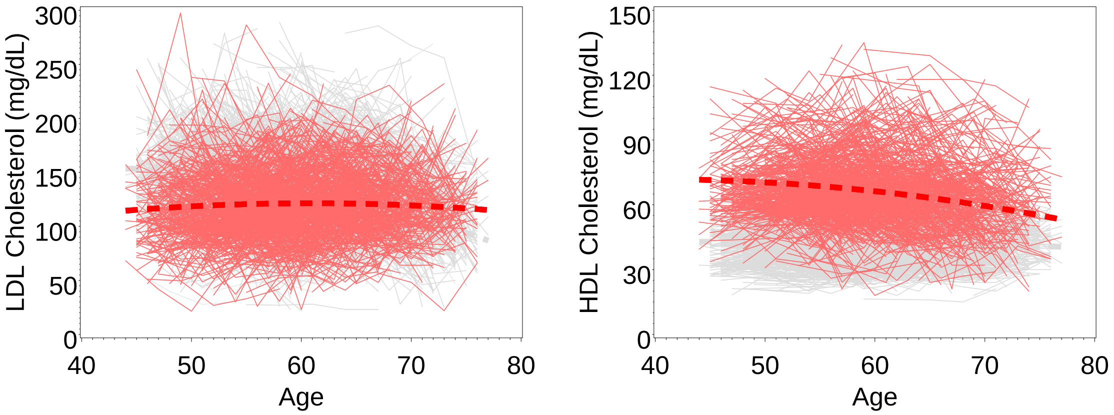
(b) HDL Cholesterol



(c) Triglycerides

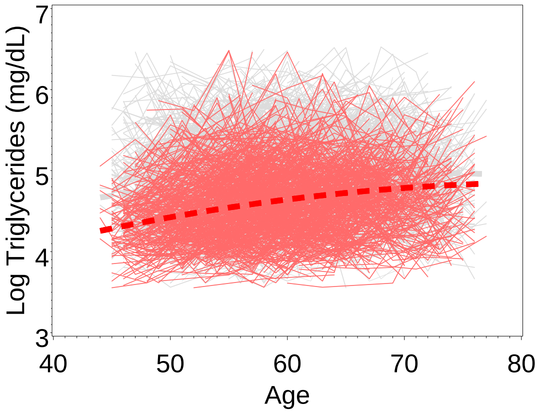
Figure 1.2: ARIC Application: Fitted Lipid Trajectories - LCLMM (2 Classes)

Red = Optimal
Orange = At-Risk



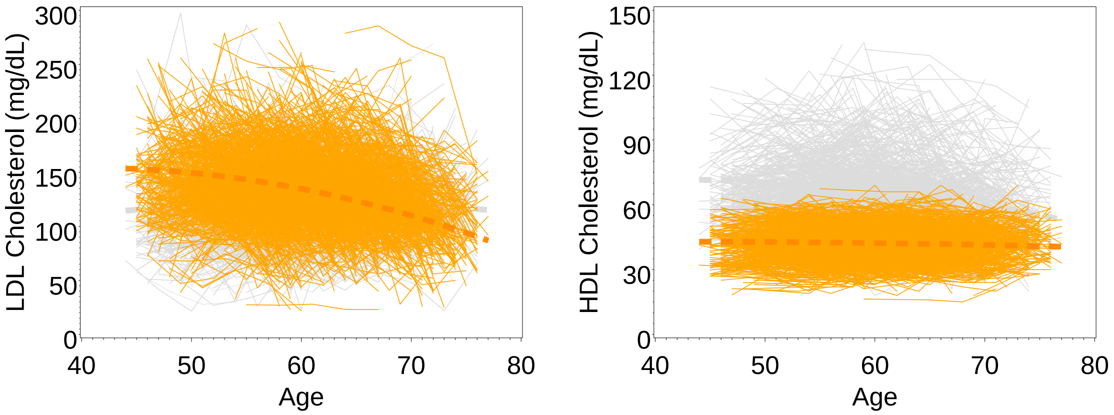
(a) LDL Cholesterol

(b) HDL Cholesterol



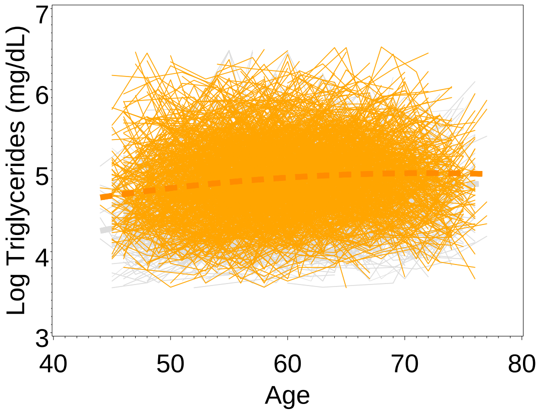
(c) Triglycerides

Figure 1.3: ARIC Application: Fitted Lipid Trajectories - LCLMM (2 Classes) - Optimal Class



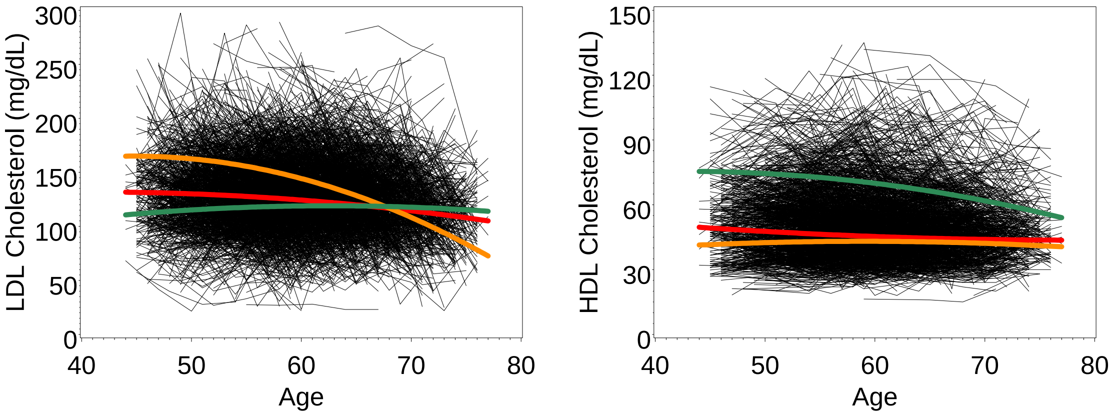
(a) LDL Cholesterol

(b) HDL Cholesterol



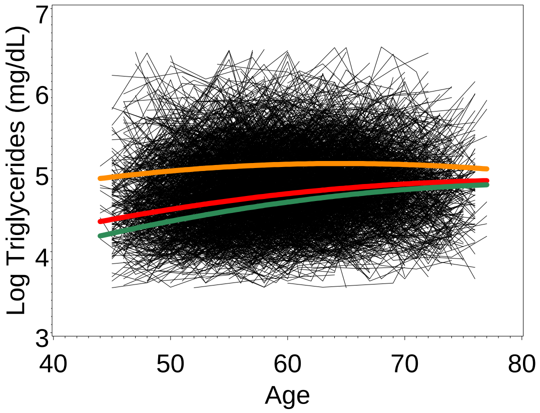
(c) Triglycerides

Figure 1.4: ARIC Application: Fitted Lipid Trajectories - LCLMM (2 Classes) - At-Risk Class



(a) LDL Cholesterol

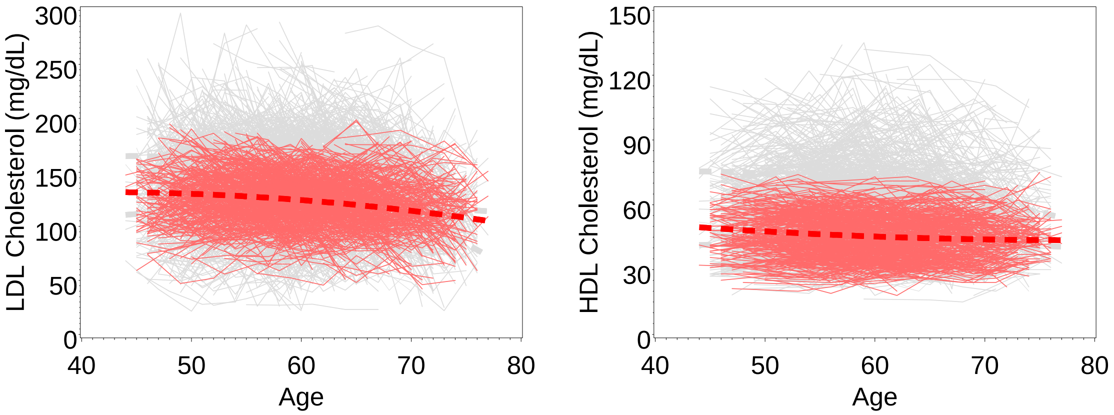
(b) HDL Cholesterol



(c) Triglycerides

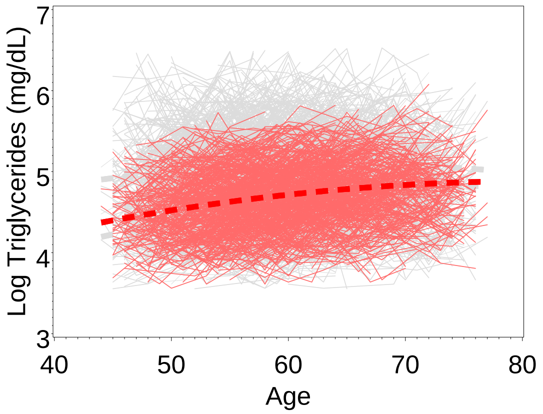
Figure 1.5: ARIC Application: Fitted Lipid Trajectories - LCLMM (3 Classes)

Red = Average
Orange = At-Risk
Green = Optimal



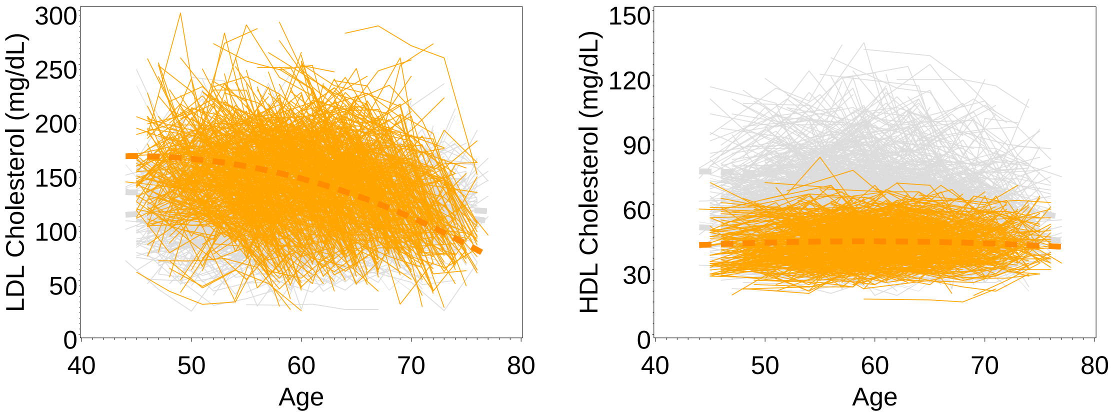
(a) LDL Cholesterol

(b) HDL Cholesterol



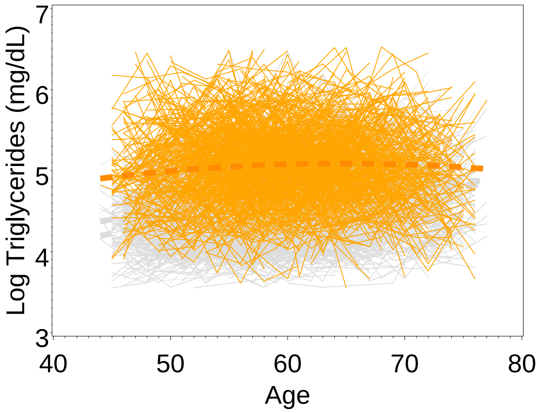
(c) Triglycerides

Figure 1.6: ARIC Application: Fitted Lipid Trajectories - LCLMM (3 Classes) - Average Class



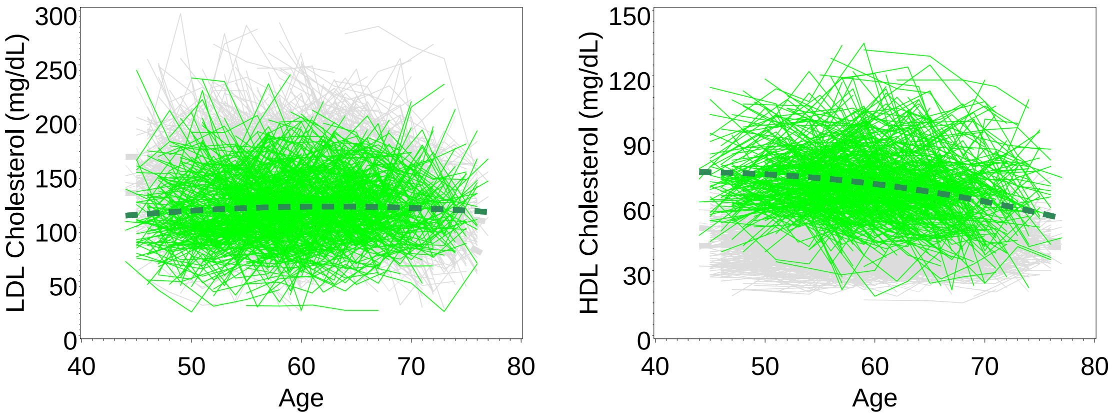
(a) LDL Cholesterol

(b) HDL Cholesterol



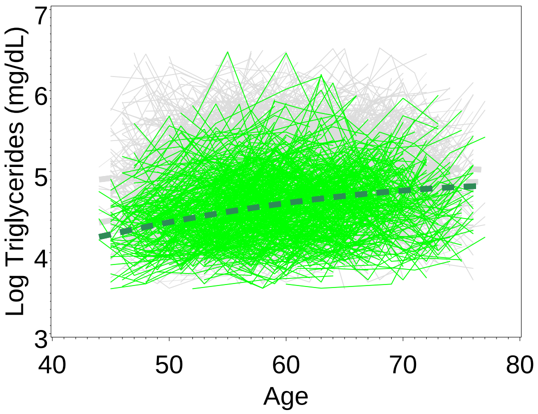
(c) Triglycerides

Figure 1.7: ARIC Application: Fitted Lipid Trajectories - LCLMM (3 Classes) - At-Risk Class



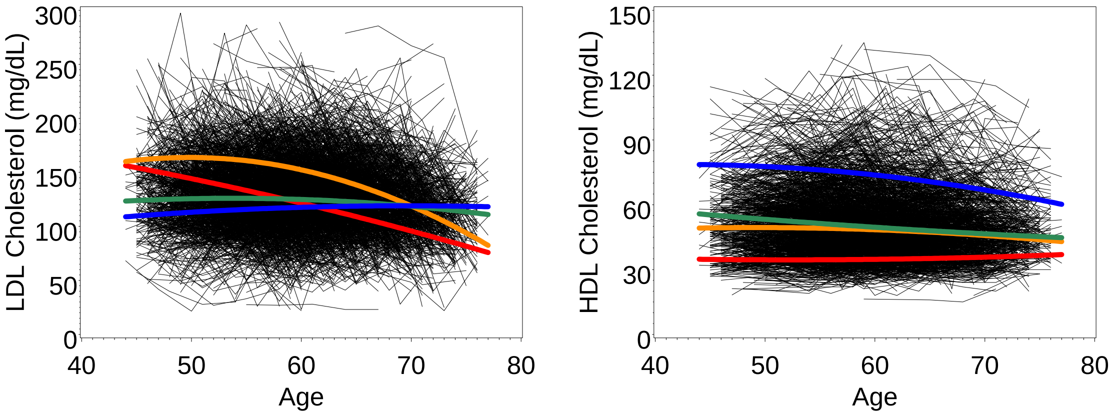
(a) LDL Cholesterol

(b) HDL Cholesterol

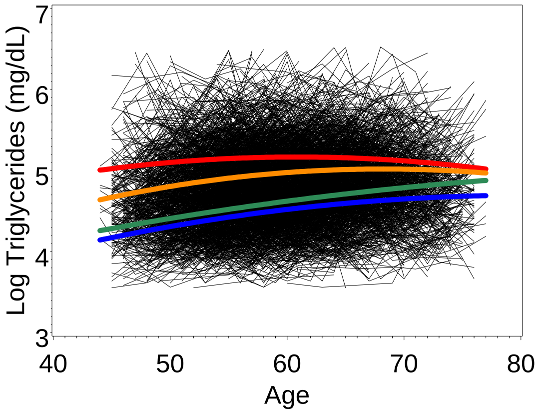


(c) Triglycerides

Figure 1.8: ARIC Application: Fitted Lipid Trajectories - LCLMM (3 Classes) - Optimal Class



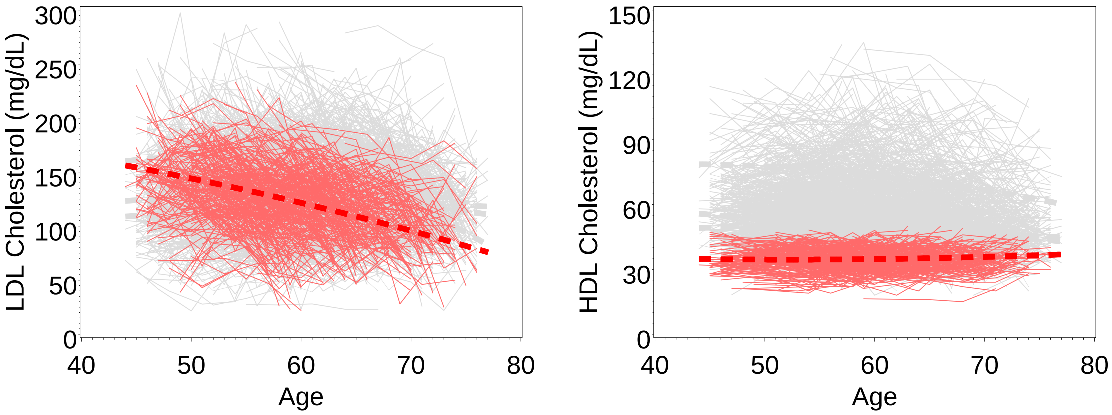
(a) LDL Cholesterol (b) HDL Cholesterol



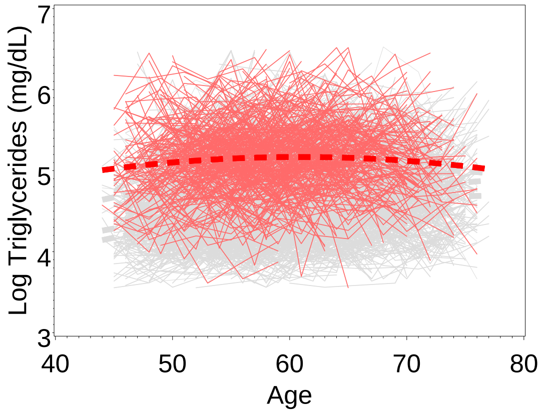
(c) Triglycerides

Figure 1.9: ARIC Application: Fitted Lipid Trajectories - LCLMM (4 Classes)

Red = At-Risk
Orange = AvgHDL-HighOther
Green = AvgHDL-LowOther
Blue = Optimal

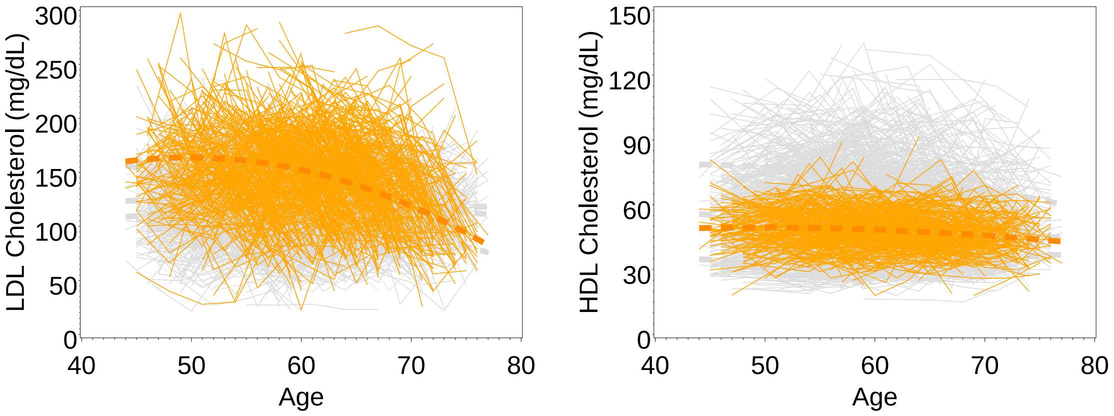


(a) LDL Cholesterol (b) HDL Cholesterol

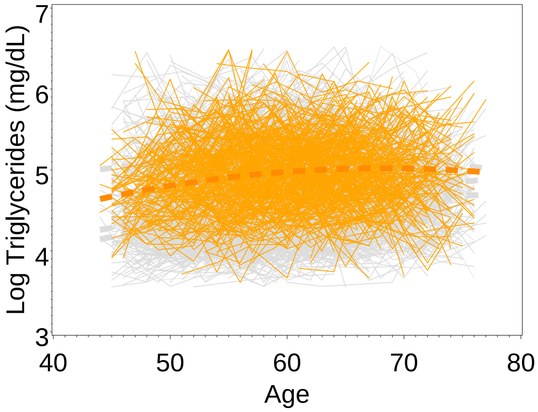


(c) Triglycerides

Figure 1.10: ARIC Application: Fitted Lipid Trajectories - LCLMM (4 Classes) - At-Risk Class

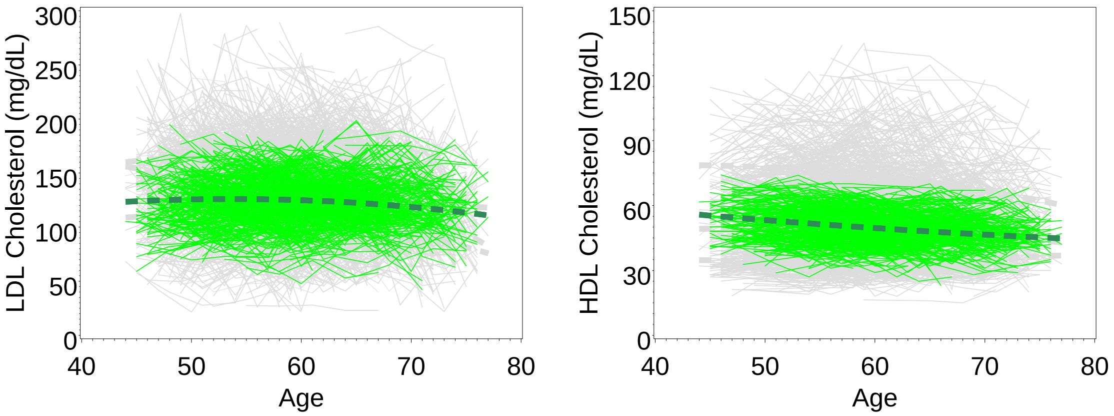


(a) LDL Cholesterol (b) HDL Cholesterol



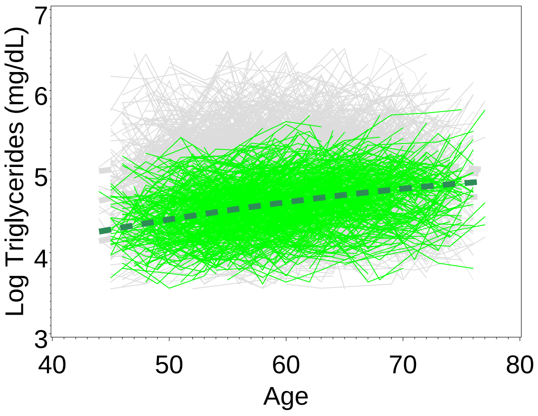
(c) Triglycerides

Figure 1.11: ARIC Application: Fitted Lipid Trajectories - LCLMM (4 Classes) - AvgHDL-HighOther Class



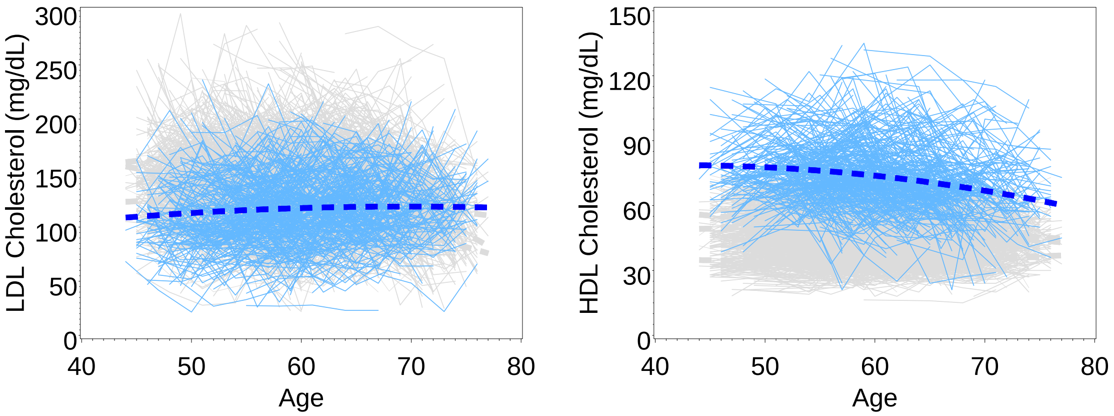
(a) LDL Cholesterol

(b) HDL Cholesterol



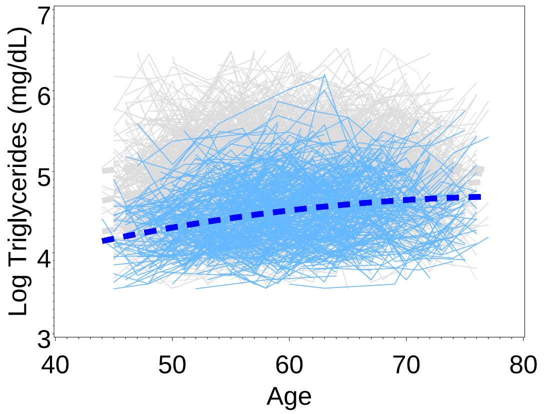
(c) Triglycerides

Figure 1.12: ARIC Application: Fitted Lipid Trajectories - LCLMM (4 Classes) - AvgHDL-LowOther Class



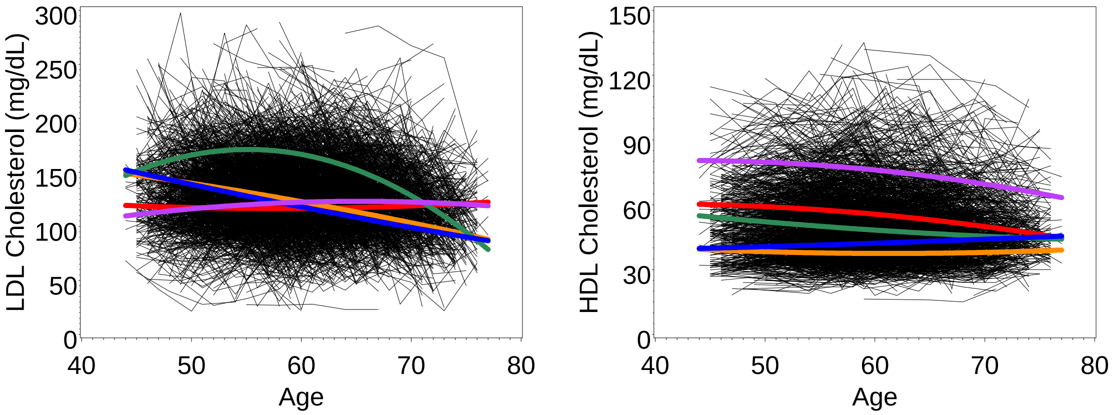
(a) LDL Cholesterol

(b) HDL Cholesterol



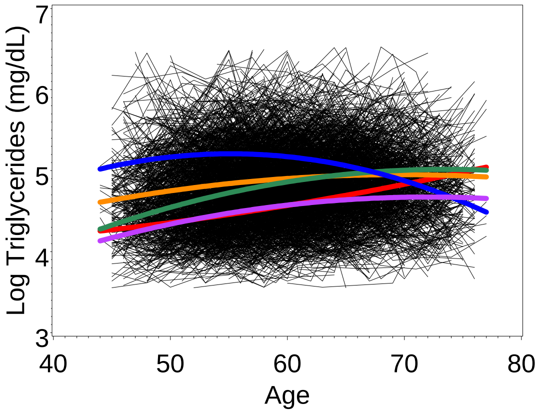
(c) Triglycerides

Figure 1.13: ARIC Application: Fitted Lipid Trajectories - LCLMM (4 Classes) - Optimal Class



(a) LDL Cholesterol

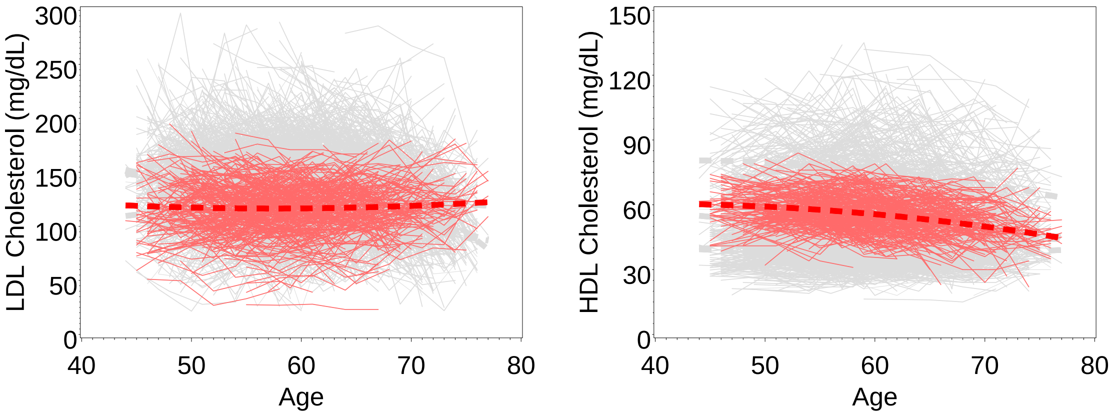
(b) HDL Cholesterol



(c) Triglycerides

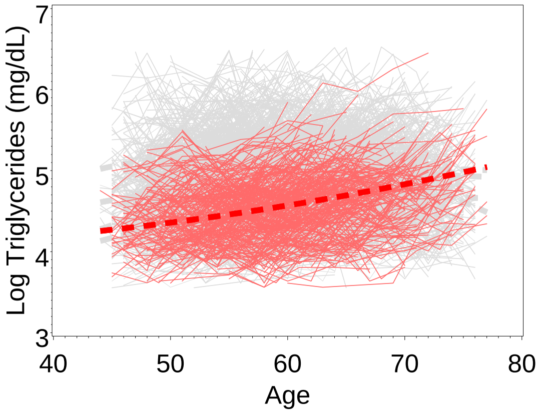
Figure 1.14: ARIC Application: Fitted Lipid Trajectories - LCLMM (5 Classes)

Red = Avg HDL - Lower LDL
Orange = Low HDL - Lower Triglycerides
Green = Avg HDL - Higher LDL
Blue = Low HDL - Higher Triglycerides
Purple = Optimal



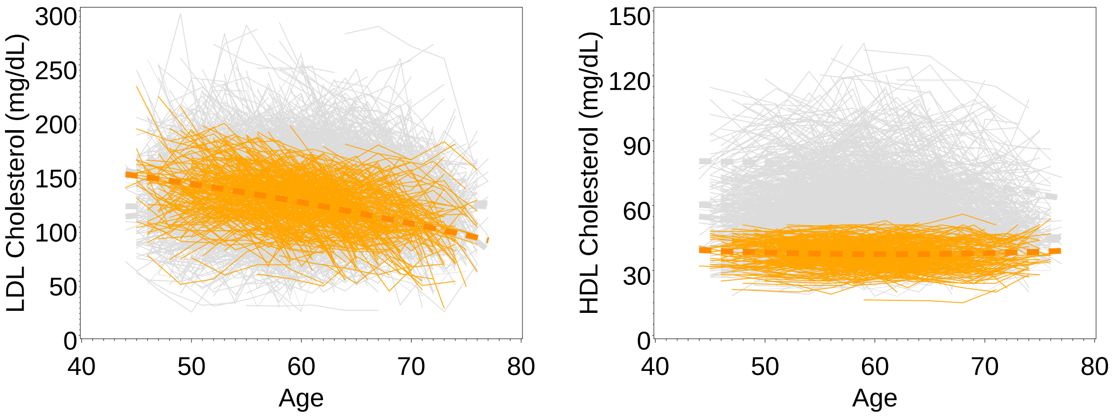
(a) LDL Cholesterol

(b) HDL Cholesterol



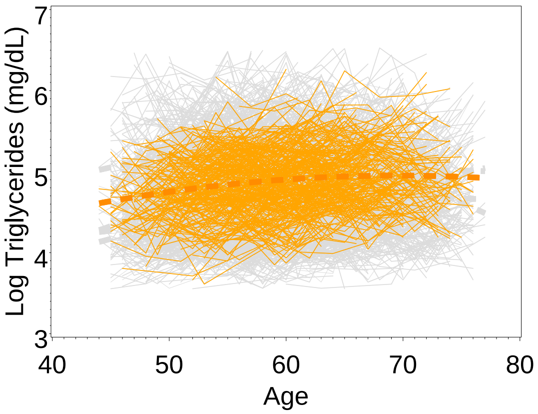
(c) Triglycerides

Figure 1.15: ARIC Application: Fitted Lipid Trajectories - LCLMM (5 Classes) - Avg HDL - Lower LDL Class



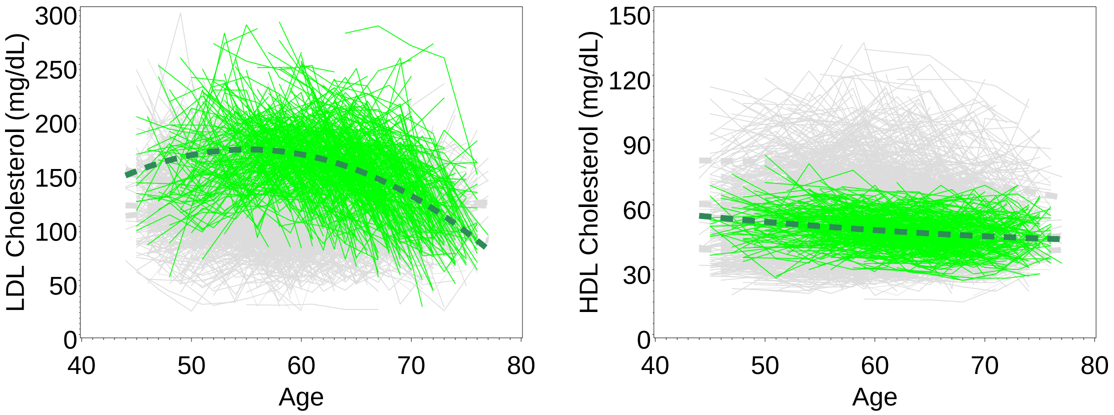
(a) LDL Cholesterol

(b) HDL Cholesterol



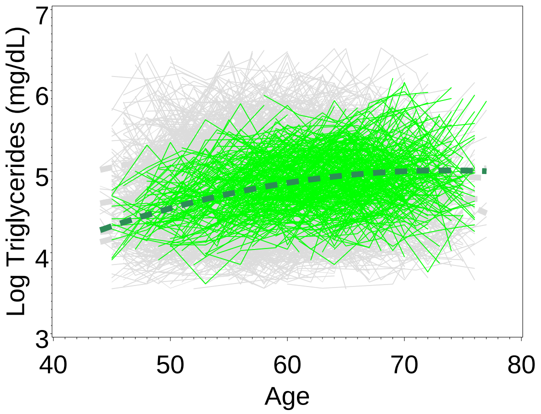
(c) Triglycerides

Figure 1.16: ARIC Application: Fitted Lipid Trajectories - LCLMM (5 Classes) - Low HDL - Lower Triglycerides Class



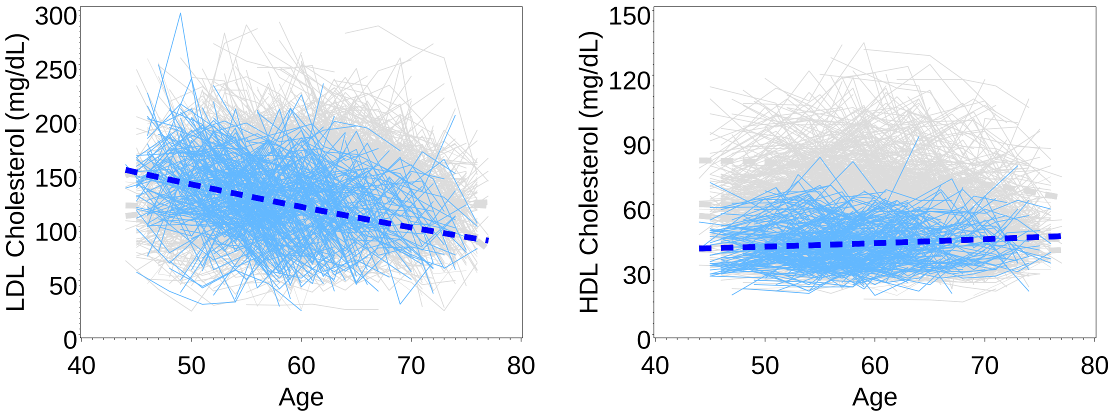
(a) LDL Cholesterol

(b) HDL Cholesterol



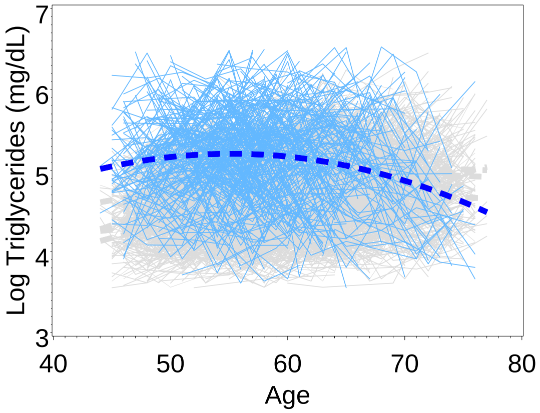
(c) Triglycerides

Figure 1.17: ARIC Application: Fitted Lipid Trajectories - LCLMM (5 Classes) - Avg HDL - Higher LDL Class



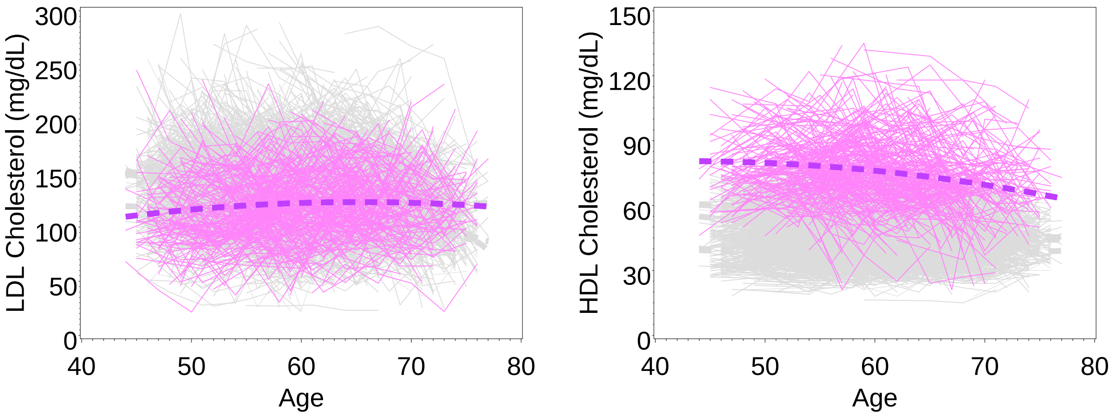
(a) LDL Cholesterol

(b) HDL Cholesterol



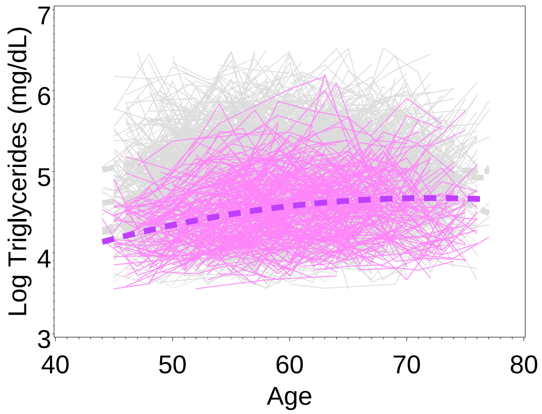
(c) Triglycerides

Figure 1.18: ARIC Application: Fitted Lipid Trajectories - LCLMM (5 Classes) - Low HDL - Higher Triglycerides Class



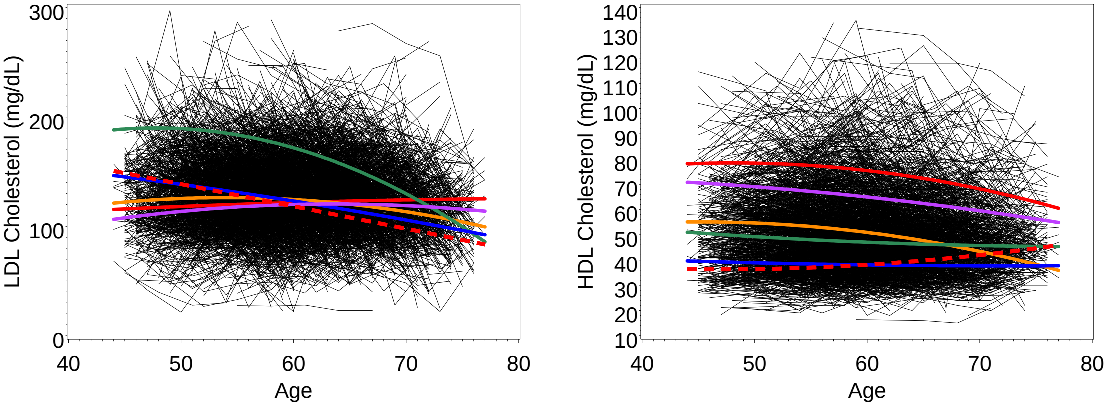
(a) LDL Cholesterol

(b) HDL Cholesterol



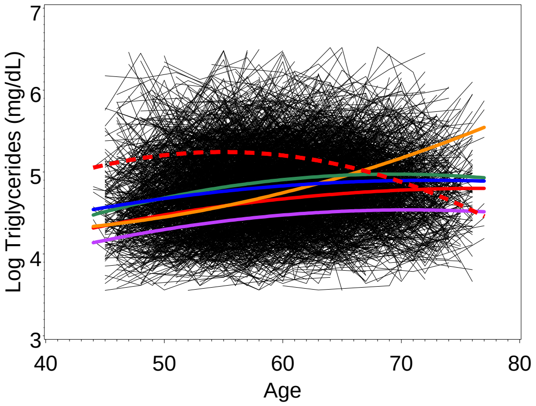
(c) Triglycerides

Figure 1.19: ARIC Application: Fitted Lipid Trajectories - LCLMM (5 Classes) - Optimal Class



(a) LDL Cholesterol

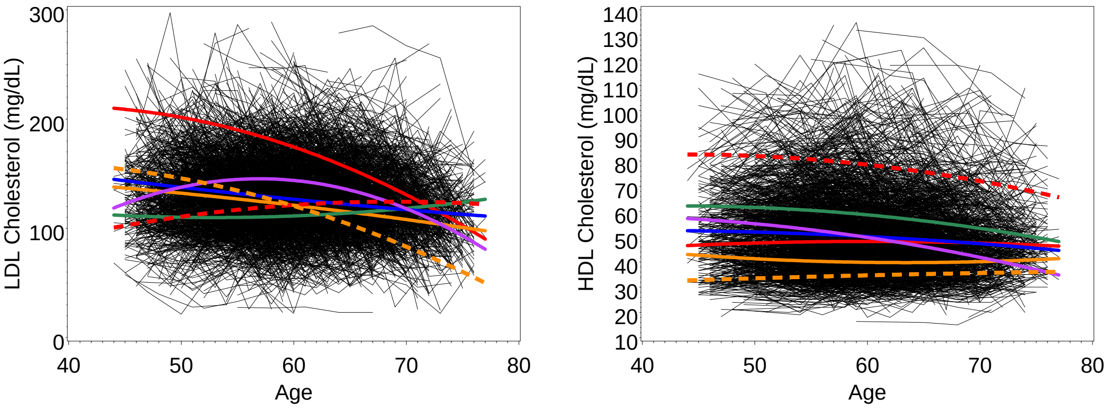
(b) HDL Cholesterol



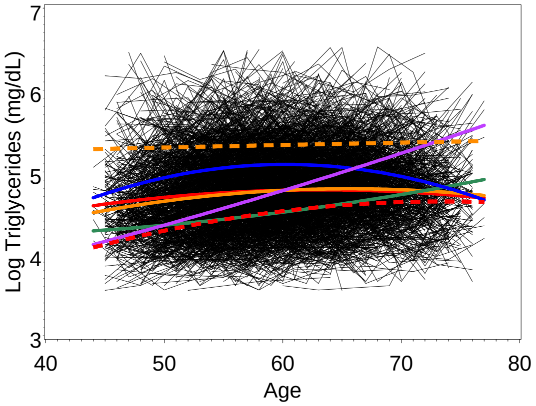
(c) Triglycerides

Figure 1.20: ARIC Application: Fitted Lipid Trajectories - LCLMM (6 Classes)

Red (Solid) = Optimal - Best HDL and Good Triglycerides
Orange = Bad Trends
Green = Highest LDL
Blue = At-Risk
Purple = Optimal - Best Triglycerides and Good HDL
Red (Dotted) = At-Risk - But Signs of Improvement



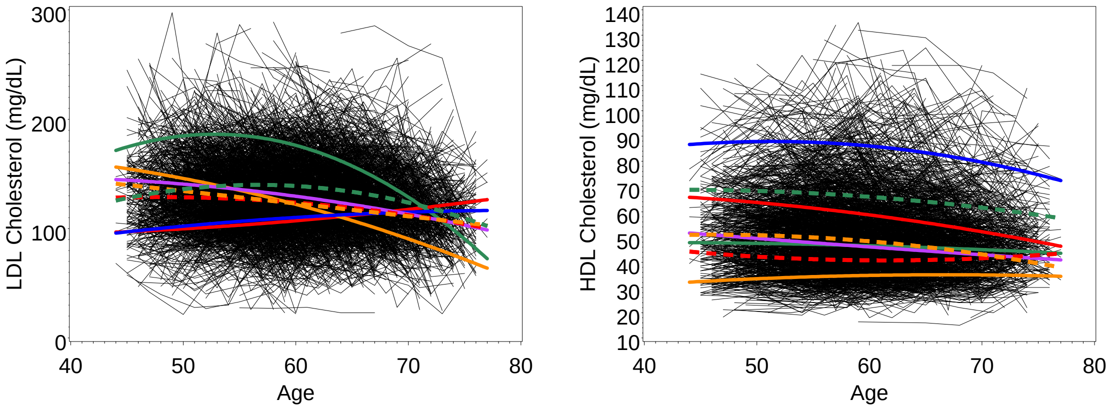
(a) LDL Cholesterol (b) HDL Cholesterol



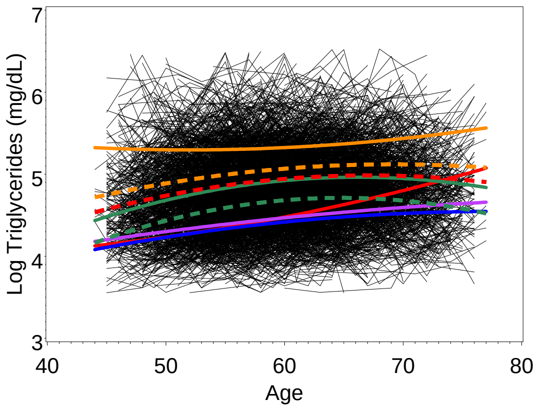
(c) Triglycerides

Figure 1.21: ARIC Application: Fitted Lipid Trajectories - LCLMM (7 Classes)

Red (Solid) = High LDL Only
Orange (Solid) = Low HDL Only
Green = Optimal-Good HDL
Blue = Average
Purple = Bad Trends
Red (Dotted) = Optimal-Best HDL
Orange (Dotted) = At-Risk



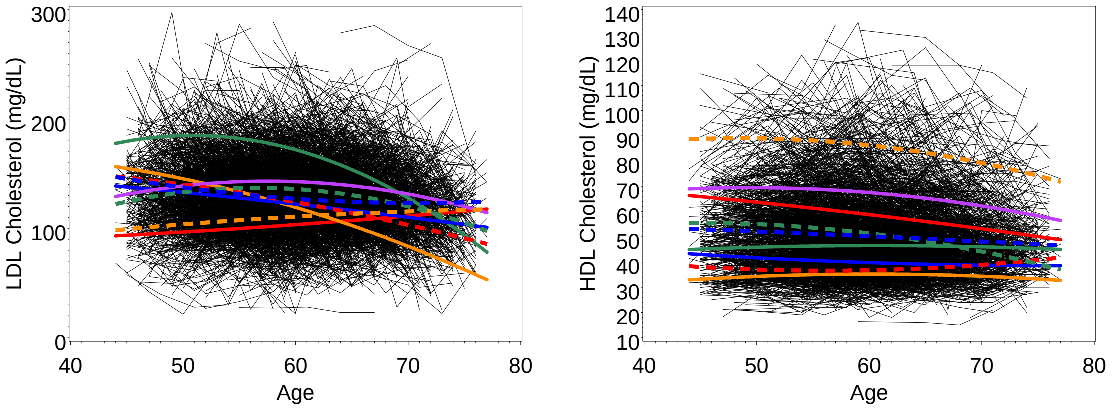
(a) LDL Cholesterol (b) HDL Cholesterol



(c) Triglycerides

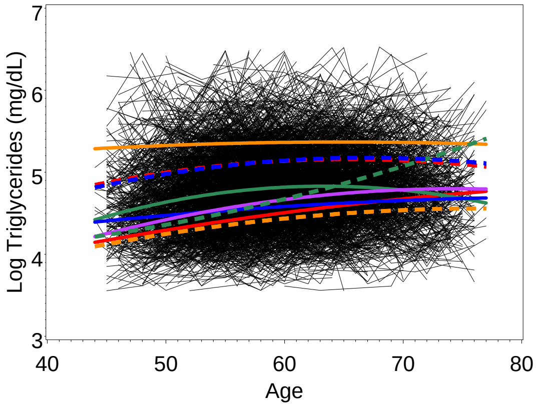
Figure 1.22: ARIC Application: Fitted Lipid Trajectories - LCLMM (8 Classes)

Red (Solid) = Bad Trends
Orange (Solid) = At-Risk
Green (Solid) = High LDL Only
Blue = Optimal
Purple = Low Triglycerides Only
Red (Dotted) = Average
Orange (Dotted) = Elevated Triglycerides Only
Green (Dotted) = Near-Optimal



(a) LDL Cholesterol

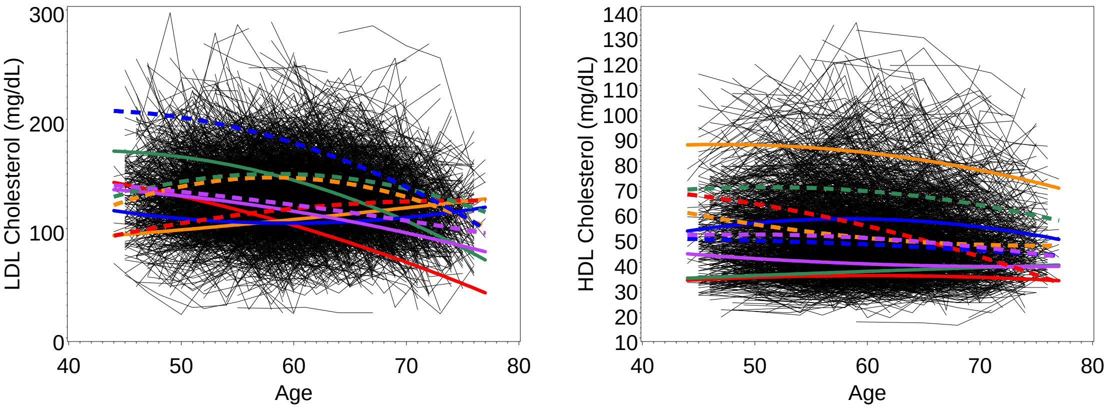
(b) HDL Cholesterol



(c) Triglycerides

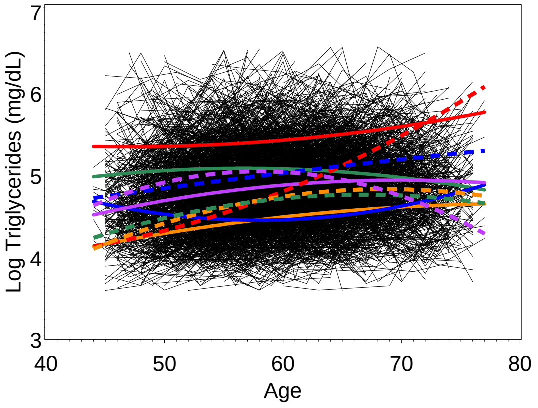
Figure 1.23: ARIC Application: Fitted Lipid Trajectories - LCLMM (9 Classes)

Red (Solid) = Near-Optimal - Good HDL
Orange (Solid) = At-Risk
Green (Solid) = High LDL Only
Blue (Solid) = Low HDL Only
Purple = High HDL Only
Red (Dotted) = Near-At-Risk
Orange (Dotted) = Optimal
Green (Dotted) = Bad Trends
Blue (Dotted) = High Triglycerides Only



(a) LDL Cholesterol

(b) HDL Cholesterol



(c) Triglycerides

Figure 1.24: ARIC Application: Fitted Lipid Trajectories - LCLMM (10 Classes)

Red (Solid) = At-Risk
Orange (Solid) = Optimal
Green (Solid) = Near-At-Risk
Blue (Solid) = Low LDL/Triglycerides
Purple (Solid) = Low HDL
Red (Dotted) = Bad Trends
Orange (Dotted) = Average-I
Green (Dotted) = High HDL
Blue (Dotted) = High LDL
Purple (Dotted) = Average-II

The numbers of subjects in each of the classes described above is presented in Tables 1.1 and 1.2. Since each individual is actually given a probability of being in each latent class as part of the model fit, as opposed to a definitive group, the numbers in these tables classify a subject into a group if that subject has at least twice the probability of being in its most likely class as being in its second-most likely class. It is interesting to note a few things regarding these counts. First, as mentioned earlier, the ARIC Carotid MRI substudy enrolled about 1,200 subjects because they had high values of maximum carotid artery wall thickness at their last ultrasound examination and 800 subjects from the remainder of the distribution. For the 2-class model, the Optimal class was identified as having 715 individuals, while the At-Risk class was identified as having 1,202 subjects. Second, the number of unknown subjects is relatively small - even in models in which many groups are fit, more than 85 percent of subjects are still fairly certain regarding class assignment. This percent of unknown subjects tends to increase slightly as classes are added.

Choosing a Model

Given that the number of latent classes is unknown a priori and that there are two possible models to fit - one with variances equal across class and one with class-specific variances - which model should be selected? This is a topic of great consideration and is examined more generally in a simulation study in Chapter 2. However, for purposes of evaluating the various models fit, selected information criteria are presented in Table 1.3 for the LMM and the LCLMM for 2-10 classes. These measures are based on work summarized or presented in Biernacki and Govaert [1999], Biernacki et al [2000], and Bozdogan [1987], and are discussed in more detail in Chapter 2. Briefly, the criteria are as follows.

$$\begin{aligned}
\text{AIC} &= -2\,\log\,\text{L}\ +2\text{v} && \bullet\text{Akaike Information Criterion}\\
\text{CAIC} &= -2\,\log\,\text{L}\ +\text{v}\ (\log\,\text{n}+1) && \bullet\text{Consistent Akaike Information Criterion}\\
\text{BIC} &= -2\,\log\,\text{L}\ +\text{v}\,\log\,\text{n} && \bullet\text{Bayesian Information Criterion}\\
\text{C} &= -2\,\log\,\text{L}\ +2\text{E} && \bullet\text{Fuzzy Classification Likelihood}\\
\text{ICL} &= -2\,\log\,\text{L}\ +\text{v}\,\log\,\text{n}+2\text{E} && \bullet\text{Integrated Completed Likelihood}
\end{aligned}$$

where:

$$\text{E (Entropy)} = -\sum_{k=1}^{K}\sum_{i=1}^{n}\widetilde{c}_{ik}\,\log\,\widetilde{c}_{ik} \geq 0$$

$$\widetilde{c}_{ik} = \frac{\pi_{ik} f\left(\mathbf{y}_i \mid c_{ik}=1\right)}{\sum_{j=1}^{K}\left\{\pi_{ij} f\left(\mathbf{y}_i \mid c_{ij}=1\right)\right\}}$$

Table 1.1: Percent of Subjects Fit in Each Latent Class (1-6 Classes)

Model	Class	N	Percent
2-Class	Optimal	715	34.6
	At-Risk	1202	58.2
	Unknown	149	7.2
3-Class	Average	678	32.8
	At-Risk	643	31.1
	Optimal	511	24.7
	Unknown	234	11.3
4-Class	At-Risk	399	19.3
	AvgHDL-HighOther	482	23.3
	AvgHDL-LowOther	532	25.8
	Optimal	414	20.0
	Unknown	239	11.6
5-Class	Avg HDL - Lower LDL	386	18.7
	Low HDL - Lower Triglycerides	391	18.9
	Avg HDL - Higher LDL	365	17.7
	Low HDL - Higher Triglycerides	307	14.9
	Optimal	340	16.5
	Unknown	277	13.4
6-Class	Optimal - Best HDL and Good Triglycerides	291	14.1
	Bad Trends	336	16.3
	Highest LDL	284	13.8
	At-Risk	336	16.3
	Optimal - Best Triglycerides and Good HDL	237	11.5
	At-Risk - But Signs of Improvement	282	13.7
	Unknown	300	14.5

Table 1.2: Percent of Subjects Fit in Each Latent Class (7-10 Classes)

Model	Class	N	Percent
7-Class	High LDL Only	221	10.7
	Low HDL Only	304	14.7
	Optimal-Good HDL	281	13.6
	Average	211	10.2
	Bad Trends	210	10.2
	Optimal-Best HDL	271	13.1
	At-Risk	259	12.5
	Unknown	309	15.0
8-Class	Bad Trends	217	10.5
	At-Risk	259	12.5
	High LDL Only	234	11.3
	Optimal	160	7.7
	Low Triglycerides Only	234	11.3
	Average	238	11.5
	Elevated Triglycerides Only	231	11.2
	Near-Optimal	192	9.3
	Unknown	301	14.6
9-Class	Near Optimal - Good HDL	223	10.8
	At-Risk	201	9.7
	High LDL Only	238	11.5
	Low HDL Only	205	9.9
	High HDL Only	163	7.9
	Near-At-Risk	183	8.9
	Optimal	160	7.7
	Bad Trends	216	10.5
	High Triglycerides Only	190	9.2
	Unknown	287	13.9
10-Class	At-Risk	139	6.7
	Optimal	179	8.7
	Near-At-Risk	166	8.0
	Low LDL/Triglycerides	196	9.5
	Low HDL	222	10.8
	Bad Trends	149	7.2
	Average-I	206	10.0
	High HDL	126	6.1
	High LDL	198	9.6
	Average-II	179	8.7
	Unknown	306	14.8

Note that v represents the number of parameters fit in the model, n represents the number of subjects, and E is an entropy measure. For consistency with the other measures, C has been multiplied by -2.

Notice that all of the traditional information-based criteria choose the most complicated 10-class model. This is consistent with results from the simulation study presented in Section 2.8, which found that these criteria tend to overestimate the number of latent classes.

Based on my work with the models, I propose a series of additional information criteria along the lines of BIC. The BIC penalizes the likelihood for each additional parameter fit in the models. However, the model that is the main focus of this research is a model in which class membership is determined by the relative likelihood of an individual's data under each of the underlying LMMs - therefore, no additional parameters are fit for class membership. If the latent classes are so obviously separated that the classes can be easily determined, then this seems appropriate. However, if class membership for an individual is not obvious, it seems as if a penalty needs to be assessed for that individual to account for fitting that individual's mixing proportion. If the individual is fit such that their class membership probability is split between two classes, then it seems the penalty should be the equivalent of fitting one additional parameter in the model. If the individual's probability is split between three classes, then the penalty should be two additional parameters, and so on. Since the point at which an individual can be attributed to one class with certainty or near-certainty is not obvious, several criteria were proposed. The first, BICMod23, looks at the highest probability for each subject - if the second-highest probability is more than half of the highest, then a single penalty parameter is charged. If the third-highest probability is more than half of the highest, then a second penalty parameter is charged and so on. The label '23' is meant to call attention to the fact that a subject with probabilities $\frac{2}{3}$ and $\frac{1}{3}$ would be at the border of having a penalty assessed. A second criteria, BICMod34, would have the border at $\frac{3}{4}$ / $\frac{1}{4}$. BICMod45 has its border at 80 percent / 20 percent and BICMod910 has its border at 90 percent / 10 percent. These criteria are presented for the LMM and LCLMM with 2-10 classes in Table 1.4.

The newly proposed criteria also tend to choose the 10-class model, although the BICMod45 criteria chooses the 7-class model for the equal across classes variance model. The BICMod910 criteria chooses the LMM as the best model, as the penalty is so great that even the 2-class model does not show an improvement. In many respects, these criteria also do not offer a great deal of help in terms of choosing a particular model to use - most do not penalize enough to result in a parsimonious model. Therefore, the next focus in the search for the 'best' model leads to an analysis of the residuals to determine which of the models tends to fit better.

Table 1.3: Linear Mixed Model and LCLMM - Traditional Information Criteria

Variances	# of Classes	ObsLik	AIC	CAIC	BIC	E	C	ICL	v
Equal-Across-Classes	1 (LMM)	-82231.8	164511.5	164567.1	164646.8	0.0	-82231.8	164646.8	24
	2	-81151.2	162368.3	162444.7	162554.2	**231.4**	-81382.6	163017.0	33
	3	-80352.9	160789.8	160887.0	161026.4	268.2	-80621.1	161562.7	42
	4	-79842.3	159786.5	159904.6	160073.8	418.7	-80261.0	160911.3	51
	5	-79407.1	158934.3	159073.2	159272.3	509.7	-79916.9	160291.7	60
	6	-79081.0	158300.0	158459.7	158688.7	509.6	-79590.6	159707.8	69
	7	-78801.8	157759.5	157940.1	158198.9	440.3	-79242.1	159079.6	78
	8	-78451.1	157076.2	157277.6	157566.3	537.9	-78989.0	158642.2	87
	9	-78357.4	156906.9	157129.1	157447.7	543.7	-78901.1	158535.0	96
	10	**-78050.3**	**156310.6**	**156553.7**	**156902.1**	530.4	**-78580.7**	**157962.9**	105
Class-Specific	1 (LMM)	-82231.8	164511.5	164567.1	164646.8	0.0	-82231.8	164646.8	24
	2	-80371.4	160822.8	160915.4	161048.1	**200.7**	-80572.1	161449.5	40
	3	-79187.7	158487.3	158617.0	158802.8	318.7	-79506.4	159440.3	56
	4	-78565.9	157275.9	157442.6	157681.5	355.0	-78920.9	158391.4	72
	5	-78082.1	156340.2	156543.9	156835.9	414.2	-78496.3	157664.3	88
	6	-77700.4	155608.7	155849.5	156194.6	424.3	-78124.6	157043.1	104
	7	-77336.9	154913.9	155191.7	155589.9	449.4	-77786.3	156488.7	120
	8	-77127.8	154527.6	154842.5	155293.8	447.1	-77574.9	156187.9	136
	9	-76786.4	153876.9	154228.8	154733.2	441.8	-77228.3	155616.8	152
	10	**-76555.2**	**153446.5**	**153835.4**	**154392.9**	460.3	**-77015.5**	**155313.4**	168

Table 1.4: Linear Mixed Model and LCLMM - Proposed Modified Bayes Criteria

Variances	# of Classes	BICMod23	BICMod34	BICMod45	BICMod910
Equal-Across-Classes	1 (LMM)	164646.8	164646.8	164646.8	**164646.8**
	2	163882.4	164912.9	165569.4	167485.4
	3	162759.1	163812.5	164469.0	166537.7
	4	162699.7	164371.4	165646.2	169470.5
	5	162539.3	164478.2	166035.4	171271.9
	6	161749.7	163986.3	165635.1	170780.0
	7	161030.9	162679.7	**164069.0**	168213.9
	8	160917.4	163169.2	165039.4	170482.0
	9	160600.3	162859.8	164905.5	170851.9
	10	**159879.2**	**162268.4**	164077.5	169909.4
Class-Specific	1 (LMM)	164646.8	164646.8	164646.8	**164646.8**
	2	162185.5	163101.5	163620.6	165269.4
	3	160634.8	162077.5	163062.2	166031.6
	4	159658.5	161284.4	162414.2	165895.0
	5	159133.5	160896.9	162370.1	166713.5
	6	158706.0	160492.2	161850.9	166186.7
	7	158223.4	160246.2	161566.8	166444.5
	8	157873.9	159667.7	161224.9	166095.0
	9	157206.4	159091.8	160458.2	165786.3
	10	**157026.4**	**159011.1**	**160354.5**	165705.5

Summary measures of the within-subject and total residuals were examined to see if any of the models tended to have a smaller MSE of the residuals (see Table 1.5). These quantities are defined as follows for subject i:

$$\begin{aligned}
\mathbf{r}_{i,total} &= \mathbf{y}_i - \sum_{k=1}^{K} \pi_{ik} \left(\mathbf{X}_i \hat{\boldsymbol{\beta}} + \mathbf{W}_i \hat{\boldsymbol{\lambda}}_k \right) \\
\mathbf{r}_{i,within} &= \mathbf{y}_i - \left[\sum_{k=1}^{K} \pi_{ik} \left(\mathbf{X}_i \hat{\boldsymbol{\beta}} + \mathbf{W}_i \hat{\boldsymbol{\lambda}}_k \right) \right] - \mathbf{Z}_i \tilde{\mathbf{b}}_i \\
\text{MSE} &= \frac{\sum_{i=1}^{n} \mathbf{r}_i' \mathbf{r}_i}{\text{Total \# of Observations}}
\end{aligned}$$

Since the model fits each subject using a mixture determined by the relative fit of the underlying models, it is expected that the residuals should show a noticeable improvement as classes are added. However, keep in mind that if a subject has a great deal of data consistent with one class, but then has a parameter which is inconsistent with that class, then the LCLMM may actually have a worse fit for that parameter.

Notice that all of the models are an improvement over the LMM in terms of decreasing the mean square error of the total residuals. The biggest decrease was about 49 percent for LDL cholesterol, 77 percent for HDL cholesterol, and 53 percent for triglycerides. However, there is not a single model which fits best for all of the lab parameters. Note that while the likelihood is always much better for the class-specific variance model, the models which assume variances are equal across classes have comparable residual measures. Using the within-subject residuals as a guide of overall model fit, I chose to focus on one model with class-specific variances and another with variances equal-across-classes. Note that in the simulation study presented in Section 2.8, the within-subject MSE is found to be the most effective measure in terms of choosing from competing models. For equal-across-classes variances, the 4-class model has the best fit for LDL and is reasonably close to the best fit for HDL and triglycerides. For class-specific variances, the 5-class model also has the best fit for LDL and is reasonably close to the best fit for HDL and triglycerides. This leads to the next question... should the model be fit assuming class-specific variances or variances equal-across-classes?

Variances - Class-Specific or Equal-Across-Classes?

Many previously proposed models have fit the variances as being equal across classes. As noted earlier, Basford and McLachlan [1985] presented an example to illustrate that fitting a normal mixture model

Table 1.5: Summary of MSE of the Residuals

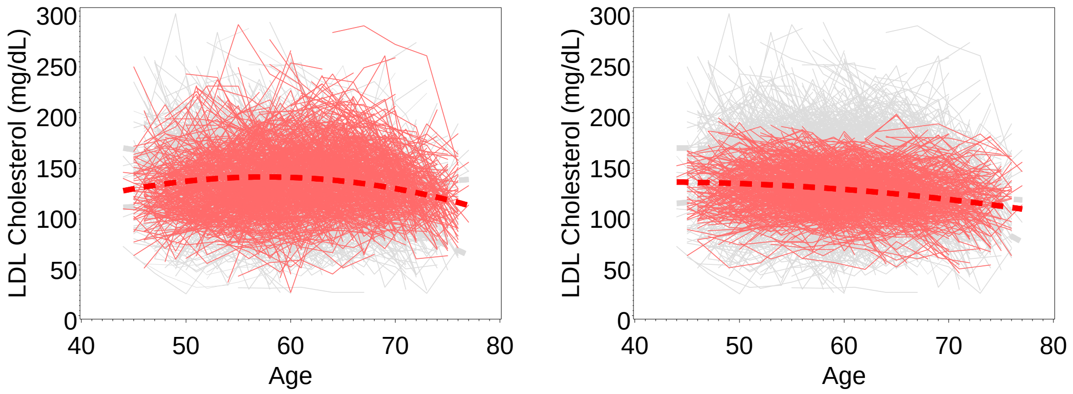
		LDL		HDL		Triglycerides	
Variances	# of Classes	MSE (Total)	MSE (Within)	MSE (Total)	MSE (Within)	MSE (Total)	MSE (Within)
Equal-Across-Classes	1 (LMM)	1266.52	367.488	269.658	48.1396	0.24103	0.073401
	2	1159.65	372.812	135.068	46.0290	0.19761	0.071278
	3	1154.42	381.394	89.588	48.8697	0.19239	0.071675
	4	826.44	**356.578**	122.817	45.3699	0.17877	0.064241
	5	840.50	362.126	106.927	45.5037	0.11690	0.064965
	6	850.50	364.889	89.565	**43.6063**	0.14175	**0.061654**
	7	748.60	380.971	118.886	48.0689	0.13913	0.068594
	8	791.97	357.112	**60.865**	45.3502	0.14527	0.065535
	9	**634.83**	362.785	67.462	46.8258	0.12953	0.065634
	10	687.77	358.773	62.201	46.8550	**0.11322**	0.069984
Class-Specific	1 (LMM)	1266.52	367.488	269.658	48.1396	0.24103	0.073401
	2	1196.82	371.536	131.903	47.9926	0.22063	0.072992
	3	1111.90	372.126	133.078	**47.6210**	0.19571	0.072999
	4	1052.62	368.817	95.099	48.6080	0.17255	0.072665
	5	901.35	**358.826**	96.904	48.4419	0.18417	0.069794
	6	891.88	370.346	94.723	48.5125	0.18557	0.068556
	7	846.97	372.306	85.579	48.9976	0.15179	0.069851
	8	824.82	363.689	77.449	51.3758	**0.13626**	0.073734
	9	828.50	367.474	**71.745**	51.9758	0.14085	0.072628
	10	**642.90**	392.803	84.443	49.5148	0.16001	**0.067432**

with homoscedastic variances in a situation where the classes are in fact heteroscedastic can have a significant influence on the resulting estimates. In particular, they found that the mixing proportions can be greatly affected, resulting in potentially different results in multivariate clustering analyses. Logically, incorporating the different degrees of variation present in each latent class should allow for an improved and more realistic model. For illustration, the 3-class models with class-specific variances and variances equal-across-classes are examined using the ARIC data.

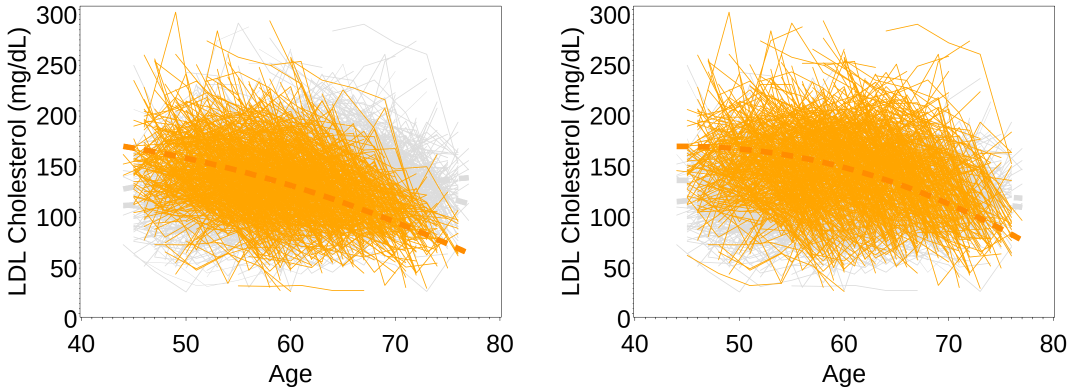
First, spaghetti-plots for subjects identified in each of the three classes are presented in Figures 1.25, 1.26, and 1.27 for both the class-specific variance model and equal-across-classes variance model. It is apparent that the groups have remarkably different degrees of variation, both for the model which accounts for that variation and the model which assumes that the variation is equal for all classes. Note the increased variation present for HDL in the Optimal group in both the equal-across-classes variance model and class-specific variance model. The fitted variances for each class based on the class-specific variance model are presented in Table 1.6 and confirm the sometimes dramatic differences. Note that the residual error variances for HDL range from 22 for the Average class to 138 for the Optimal class. Similar degrees of difference between classes are found for the other lipids as well.

The question arises, "Can these differences in variances affect the choice of groups and/or the underlying models?" Of 2,066 subjects, about 460 were not classified into a particular class (did not have their most likely class twice as likely as their second most likely class) in one or both models. Of the remaining 1,600, about 1,025 were selected to be in the same class in both models. That leaves about 575 who were fit in different classes depending on whether variances were fit as being class-specific or equal-across-classes. Of these, 245 were considered to be in the 'At-Risk' class in the equal-across-classes variance model but not in the class-specific variance model and 186 subjects were considered to be in the 'At-Risk' class for the class-specific variance model but not the equal-across-classes variance model. The remainder were at the border of 'Average' and 'Optimal'. In order to investigate the differences, the random intercepts for each subject for LDL, HDL, and triglycerides from the usual LMM were categorized into quartiles for each of the three lab parameters. Then, the percent of subjects in each of the quartiles who were classified into the 'At-Risk' group in the 3-class LCLMMs was calculated and is presented in Figures 1.28, 1.29, and 1.30. Each figure compares the class membership resulting from the class-specific variance model with the class membership resulting from the equal-across-classes variance model.

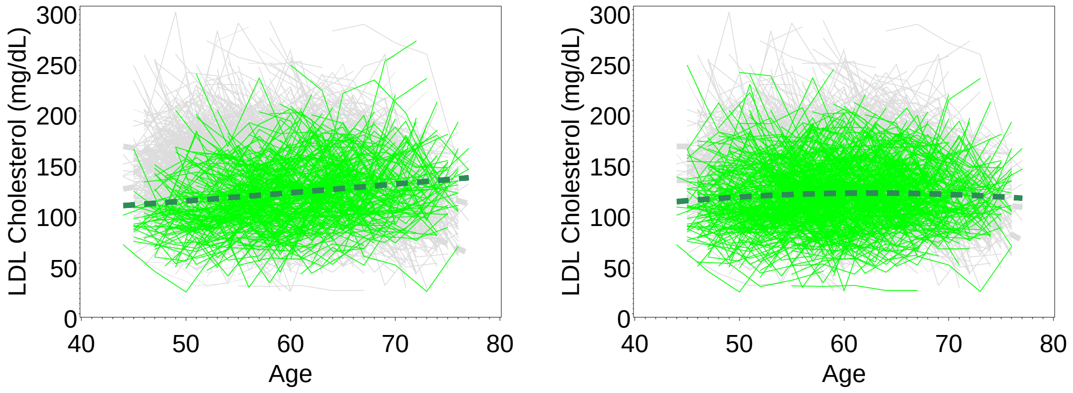
When looking at the LDL/HDL quartiles, the equal-across-classes variance model appears to rely primarily on HDL to make class determinations, as the levels of the bars are very similar across LDL



(a) Equal-Across-Classes Variances - 'Average' Class (b) Class-Specific Variances - 'Average' Class

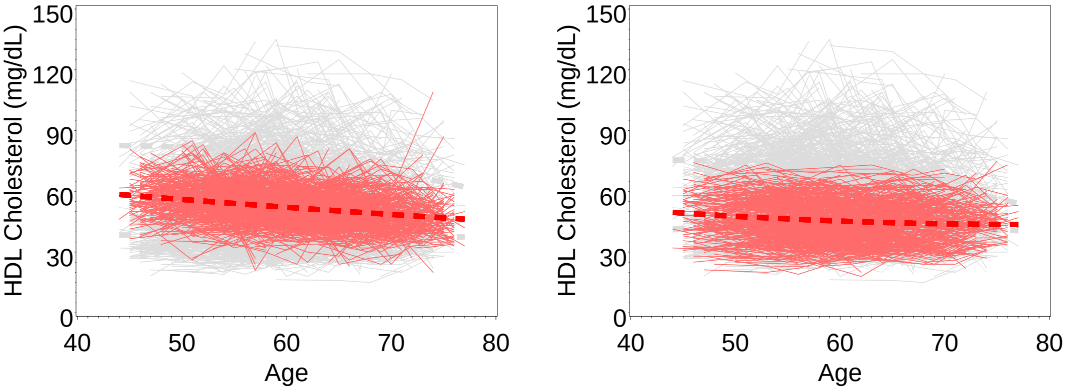


(c) Equal-Across-Classes Variances - 'At-Risk' Class (d) Class-Specific Variances - 'At-Risk' Class



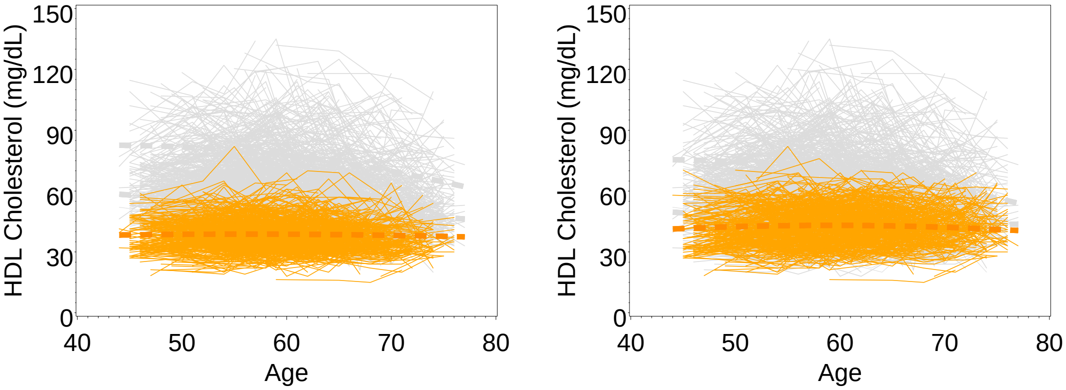
(e) Equal-Across-Classes Variances - 'Optimal' Class (f) Class-Specific Variances - 'Optimal' Class

Figure 1.25: Plot of LDL Cholesterol for Subjects Identified in Each Latent Class



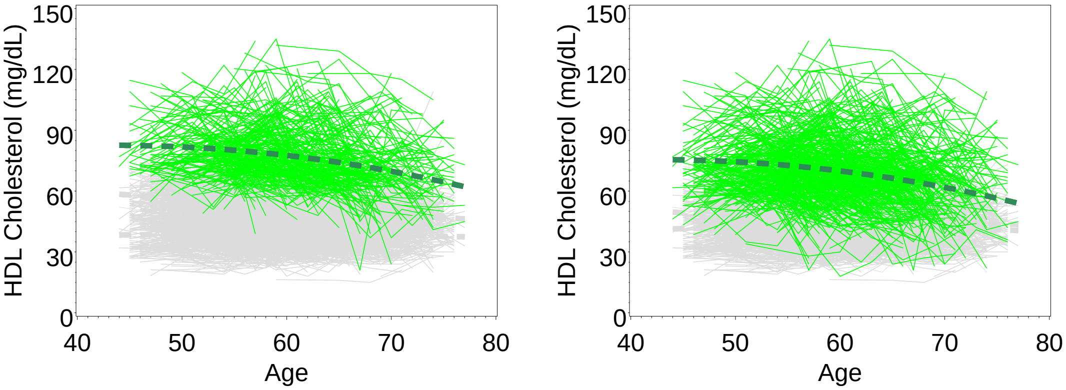
(a) Equal-Across-Classes Variances - 'Average' Class

(b) Class-Specific Variances - 'Average' Class



(c) Equal-Across-Classes Variances - 'At-Risk' Class

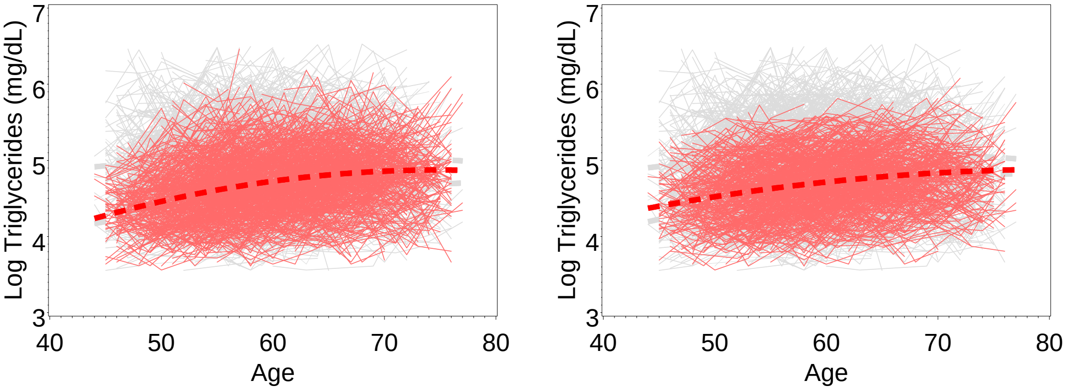
(d) Class-Specific Variances - 'At-Risk' Class



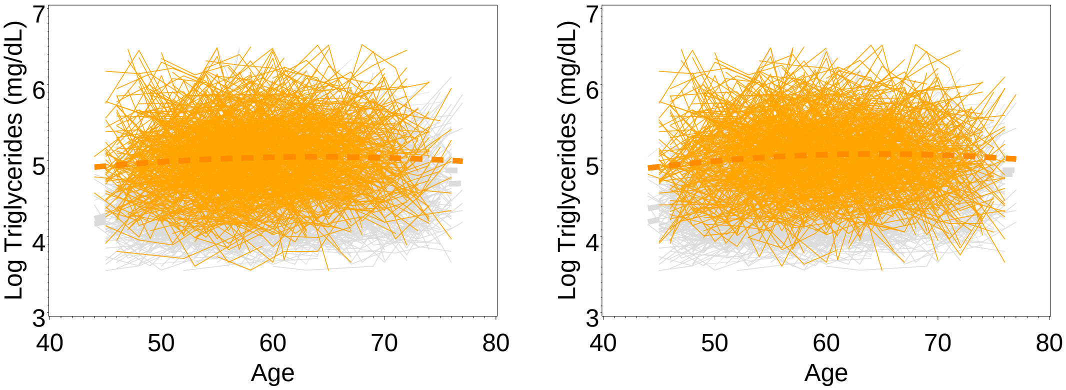
(e) Equal-Across-Classes Variances - 'Optimal' Class

(f) Class-Specific Variances - 'Optimal' Class

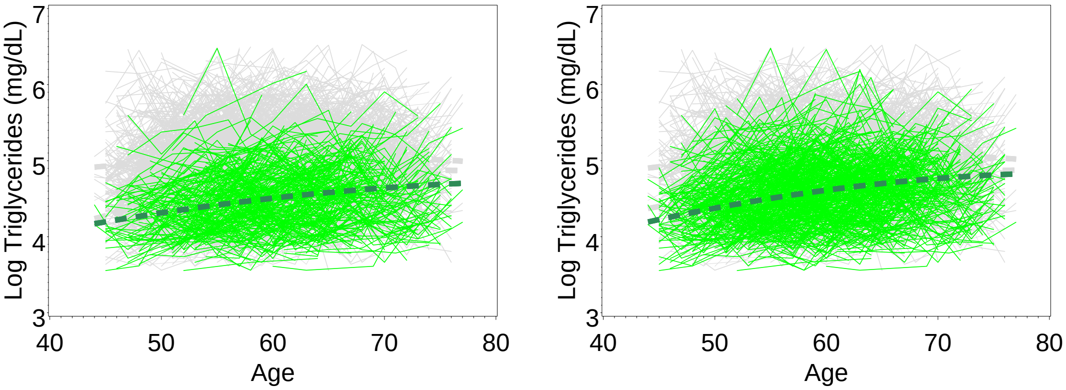
Figure 1.26: Plot of HDL Cholesterol for Subjects Identified in Each Latent Class



(a) Equal-Across-Classes Variances - 'Average' Class (b) Class-Specific Variances - 'Average' Class



(c) Equal-Across-Classes Variances - 'At-Risk' Class (d) Class-Specific Variances - 'At-Risk' Class

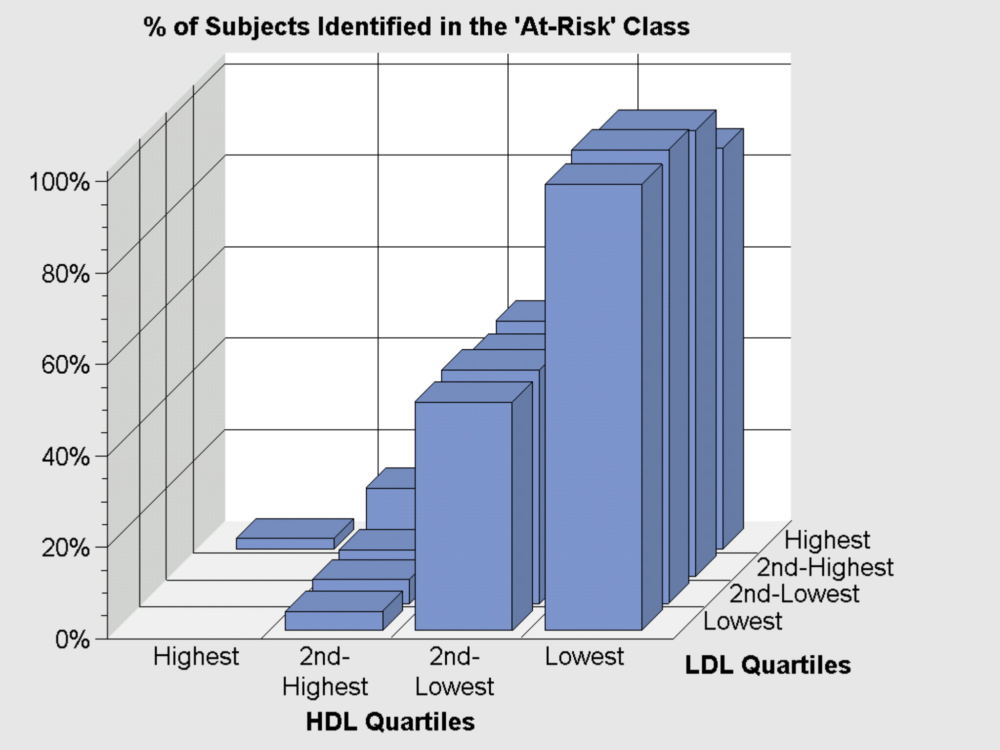


(e) Equal-Across-Classes Variances - 'Optimal' Class (f) Class-Specific Variances - 'Optimal' Class

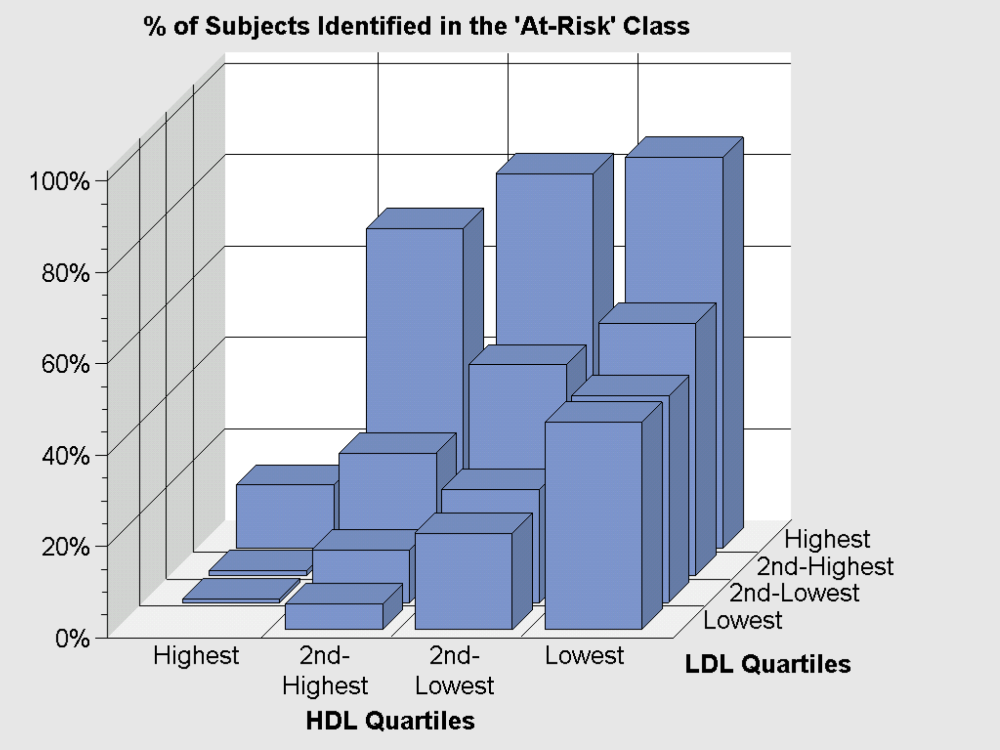
Figure 1.27: Plot of Triglycerides for Subjects Identified in Each Latent Class

Table 1.6: Fitted Variances - LCLMM Class-Specific Variance Model (3-Class)

Variance	Class	Estimate	SE
LDL - Random Intercept	Average	355.66	28.79
	At-Risk	906.42	70.27
	Optimal	643.48	48.95
LDL - Random Slope	Average	1.70	0.24
	At-Risk	1.47	0.65
	Optimal	1.18	0.39
LDL - Residual Error	Average	191.25	7.61
	At-Risk	809.41	28.41
	Optimal	453.72	18.57
HDL - Random Intercept	Average	78.60	5.19
	At-Risk	48.95	3.39
	Optimal	161.58	13.98
HDL - Residual Error	Average	22.38	0.87
	At-Risk	41.51	2.06
	Optimal	138.42	5.46
Triglycerides - Random Intercept	Average	0.10	0.01
	At-Risk	0.12	0.01
	Optimal	0.12	0.01
Triglycerides - Residual Error	Average	0.06	0.002
	At-Risk	0.13	0.005
	Optimal	0.08	0.004

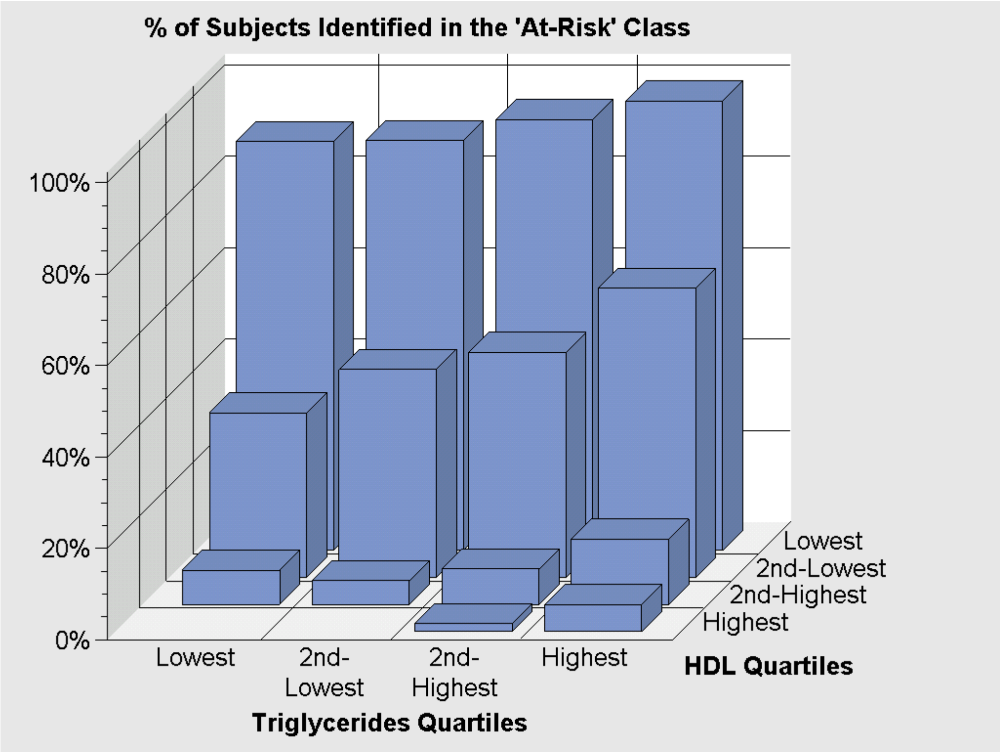


(a) Equal-Across-Classes Variances

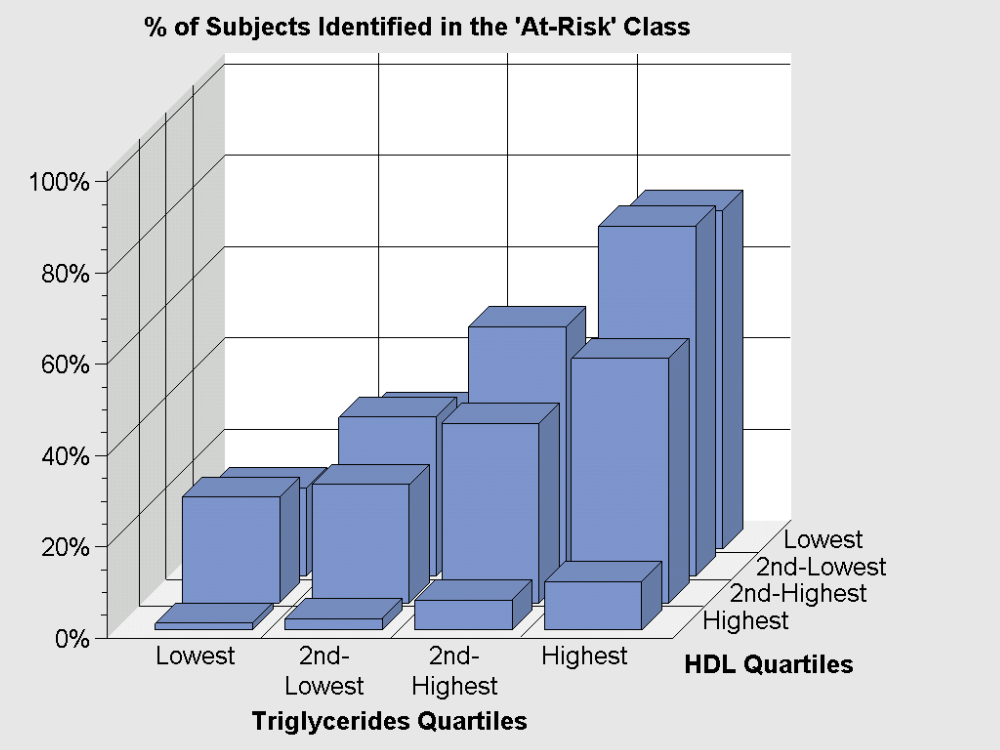


(b) Class-Specific Variances

Figure 1.28: Comparison of Latent Class Membership by LDL/HDL Quartiles

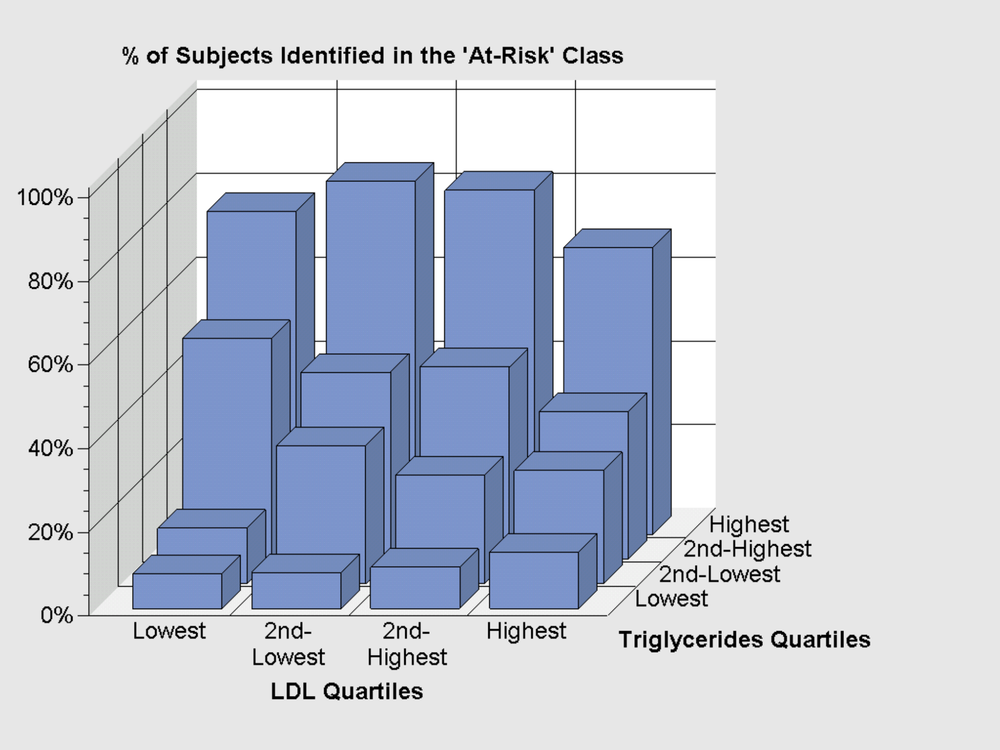


(a) Equal-Across-Classes Variances

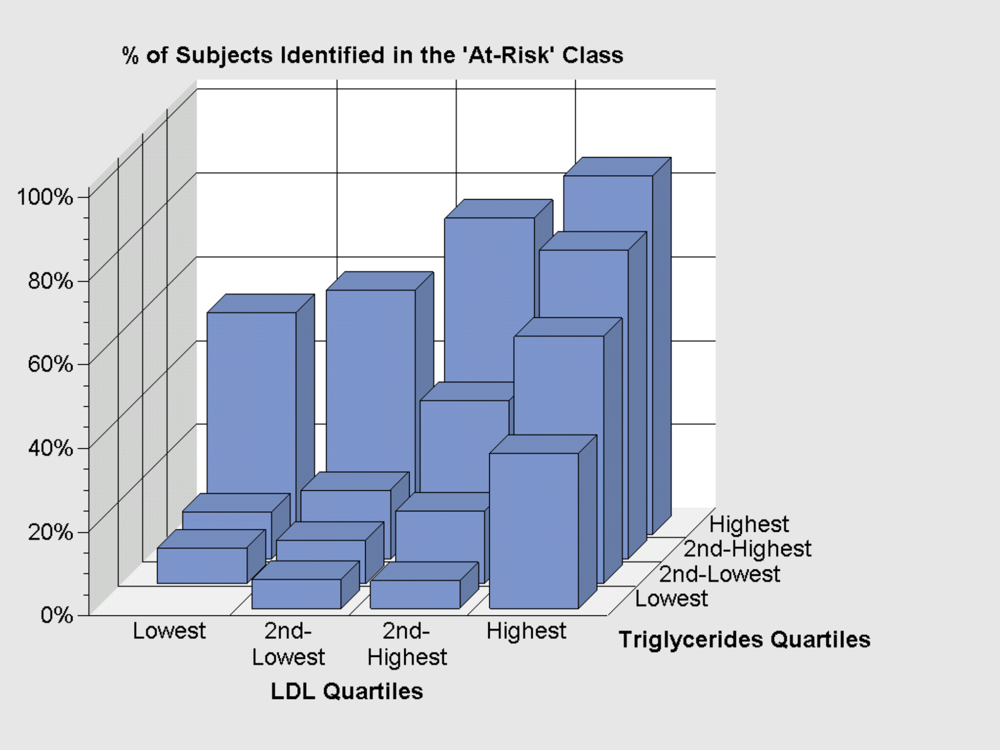


(b) Class-Specific Variances

Figure 1.29: Comparison of Latent Class Membership by HDL/Triglycerides Quartiles



(a) Equal-Across-Classes Variances



(b) Class-Specific Variances

Figure 1.30: Comparison of Latent Class Membership by LDL/Triglycerides Quartiles

levels. As HDL decreases, subjects for the most part become more likely to be classified into the 'At-Risk' class. In the model with class-specific variances, the pattern appears to be more of a step function where an increase in either LDL or HDL will tend to move subjects to the 'At-Risk' group. The chart for HDL vs. triglycerides is very similar in that the class-specific variance model seems to have a more noticeable step function. The LDL vs. triglycerides chart shows the same step pattern for the class-specific variance model, but in the equal-across-classes variance model, it appears that triglycerides is used primarily to determine latent classes, while LDL does not have a noticeable impact on class assignments. Given the dramatic differences between results using the equal-across-classes and class-specific variance models, it seems that if possible, it would be advantageous to utilize information with respect to variation in fitting the model. For completeness, therefore, the parameters for the final 5-class class-specific variance model are presented in Table 1.7. Note that the effect of taking cholesterol medications is a decrease of 15.06 for LDL, but that HDL stays almost unchanged. Also, note that the correlations for the random effects are relatively small after the latent classes have been accounted for. Two correlations are higher than the others. The correlation of the random intercept and slope for LDL have a correlation of -0.37 (SE = 0.049), which indicates that subjects with a higher random intercept tend to have a lower random slope. The correlation of the random intercepts for HDL and triglycerides have a correlation of -0.49 (SE = 0.030), which indicates that subjects with a high random intercept for HDL tend to have a low random intercept for HDL. Interestingly, the correlations between the other random intercepts are not nearly as high.

1.5.4 Identifying the Carotid Artery Thickness Groups

Previous research has examined the relationship between LDL / HDL / triglycerides and carotid artery thickness. Grebe et al [2007] showed that there was a statistically significant negative correlation between HDL and the intima-media thickness (IMT) of the carotid arteries. In addition, the authors found that there was a statistically significant positive correlation between both LDL and IMT and triglycerides and IMT. They found correlations of -0.187 for HDL (p-value=0.022), 0.271 for LDL (p-value=0.001), and 0.185 for triglycerides (p-value=0.023). Therefore, a comparison seemed appropriate in terms of comparing the actual IMT values in each of the identified latent classes.

First, a comparative histogram is presented in Figure 1.31 which shows the distribution of the IMT values in each latent class. Note that the IMT values for the Low HDL-Higher Triglycerides, Low HDL-Lower Triglycerides, and Average HDL-Higher LDL classes tend to be higher than for the

Table 1.7: Final Parameters - LCLMM Class-Specific Variance Model (5-Class)

	Avg HDL-Lower LDL		Low HDL-Lower Trig		Avg HDL-Higher LDL		Low HDL-Higher Trig		Optimal	
Parameter	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE
Effect of Chol. Meds on LDL	-15.06	0.648	⟶		⟶		⟶		⟶	
Effect of Chol. Meds on HDL	-0.13	0.200	⟶		⟶		⟶		⟶	
LDL - Intercept	120.02	1.289	129.20	1.246	175.03	1.708	126.29	1.820	126.32	1.667
LDL - Linear (Age)	0.05	0.095	-1.80	0.124	-1.49	0.246	-2.01	0.215	0.38	0.143
LDL - Quadratic (Age)	0.02	0.008	-0.01	0.011	-0.19	0.020	0.01	0.019	-0.03	0.013
HDL - Intercept	56.10	0.504	37.54	0.381	48.67	0.539	42.10	0.695	76.63	0.798
HDL - Linear (Age)	-0.44	0.040	-0.03	0.024	-0.34	0.052	0.17	0.052	-0.48	0.069
HDL - Quadratic (Age)	-0.01	0.004	0.01	0.002	0.01	0.004	0.00	0.005	-0.02	0.007
Trig - Intercept	4.55	0.022	4.89	0.023	4.84	0.023	5.19	0.035	4.56	0.021
Trig - Linear (Age)	0.02	0.002	0.01	0.001	0.02	0.002	-0.01	0.003	0.02	0.002
Trig - Quadratic (Age)	0.00	0.000	-0.00	0.000	-0.00	0.000	-0.00	0.000	-0.00	0.000
Variance (Random Int - LDL)	415.74	34.887	336.27	39.521	376.49	55.914	397.39	59.758	552.93	52.937
Variance (Random Slope - LDL)	1.00	0.217	2.07	0.345	4.07	0.841	2.32	0.756	1.12	0.501
Variance (Random Int - HDL)	38.66	4.278	26.20	2.649	34.20	5.297	57.00	6.030	104.05	11.441
Variance (Random Int - Trig)	0.11	0.010	0.10	0.009	0.07	0.008	0.15	0.018	0.10	0.009
Variance (Error - LDL)	164.92	9.423	236.53	11.939	754.79	40.088	725.99	39.030	516.68	24.640
Variance (Error - HDL)	41.66	2.121	14.06	0.766	42.29	2.075	54.10	2.900	173.18	8.212
Variance (Error - Trig)	0.06	0.003	0.06	0.003	0.07	0.003	0.18	0.008	0.08	0.003
Corr (Rnd Int LDL/Slope LDL)	-0.37	0.049	⟶		⟶		⟶		⟶	
Corr (Rnd Int LDL/Int HDL)	-0.05	0.039	⟶		⟶		⟶		⟶	
Corr (Rnd Slope LDL/Int HDL)	0.12	0.053	⟶		⟶		⟶		⟶	
Corr (Rnd Int LDL/Int Trig)	0.10	0.035	⟶		⟶		⟶		⟶	
Corr (Rnd Slope LDL/Int Trig)	-0.19	0.045	⟶		⟶		⟶		⟶	
Corr (Rnd Int HDL/Int Trig)	-0.49	0.030	⟶		⟶		⟶		⟶	

Optimal and Average HDL-Lower LDL classes. The mean IMT values for the former set of classes are above 0.9, while the mean values for the latter are around 0.82.

In a second examination of the data, the 2,066 individuals were classified into quartiles based on their IMT values. Figure 1.32 then displays the percent of subjects in each quartile who were classified in the Low HDL-Higher Triglycerides, Low HDL-Lower Triglycerides, and Average HDL-Higher LDL classes. This percentage increases from about 39 percent to 62 percent as you move from the lowest IMT quartile to the highest.

This analysis appears to support the results of Grebe et al [2007]. However, like that study, in which the correlations between the lipid parameters and IMT were between 0.18 and 0.28, the relationships seen in the latent class analysis still leave questions as to other important factors which act upon IMT.

1.5.5 Differential Effect of Obesity

As a further examination of the flexibility of the LCLMM, the effect of obesity was examined as it pertains to the three lipid parameters. Previous sections have already discussed the relationship between the lab parameters and IMT of the carotid arteries. A further question worth examining is 'Does obesity have an effect on an individual's lipid profile after taking into account latent class membership?' In other words, if a subject is identified as being in an at-risk lipid class, does that subject tend to have an even worse profile if they are obese, or does the latent class already reflect this?

To examine this question, the final 5-class LCLMM presented in Table 1.7 was run with one modification. An indicator for the effect of obesity (Yes/No) on LDL cholesterol was included in the model and was allowed to vary by latent class. Similar indictors were included for HDL cholesterol and triglycerides. Obesity in the model is defined as a body mass index of 30 or greater. These parameters would be interpreted as the effect of being obese above and beyond being classified in the given latent class. The estimates and standard errors for these parameters are presented in Table 1.8. In addition, p-values based on a z-test are provided to test whether the effect of obesity is significantly different from zero.

It appears that even after the latent classes are accounted for, there is still an effect of obesity on HDL cholesterol and triglycerides. In all groups, subjects classified as obese had a lower average HDL cholesterol value and a higher average triglyceride value than those not classified as obese, with

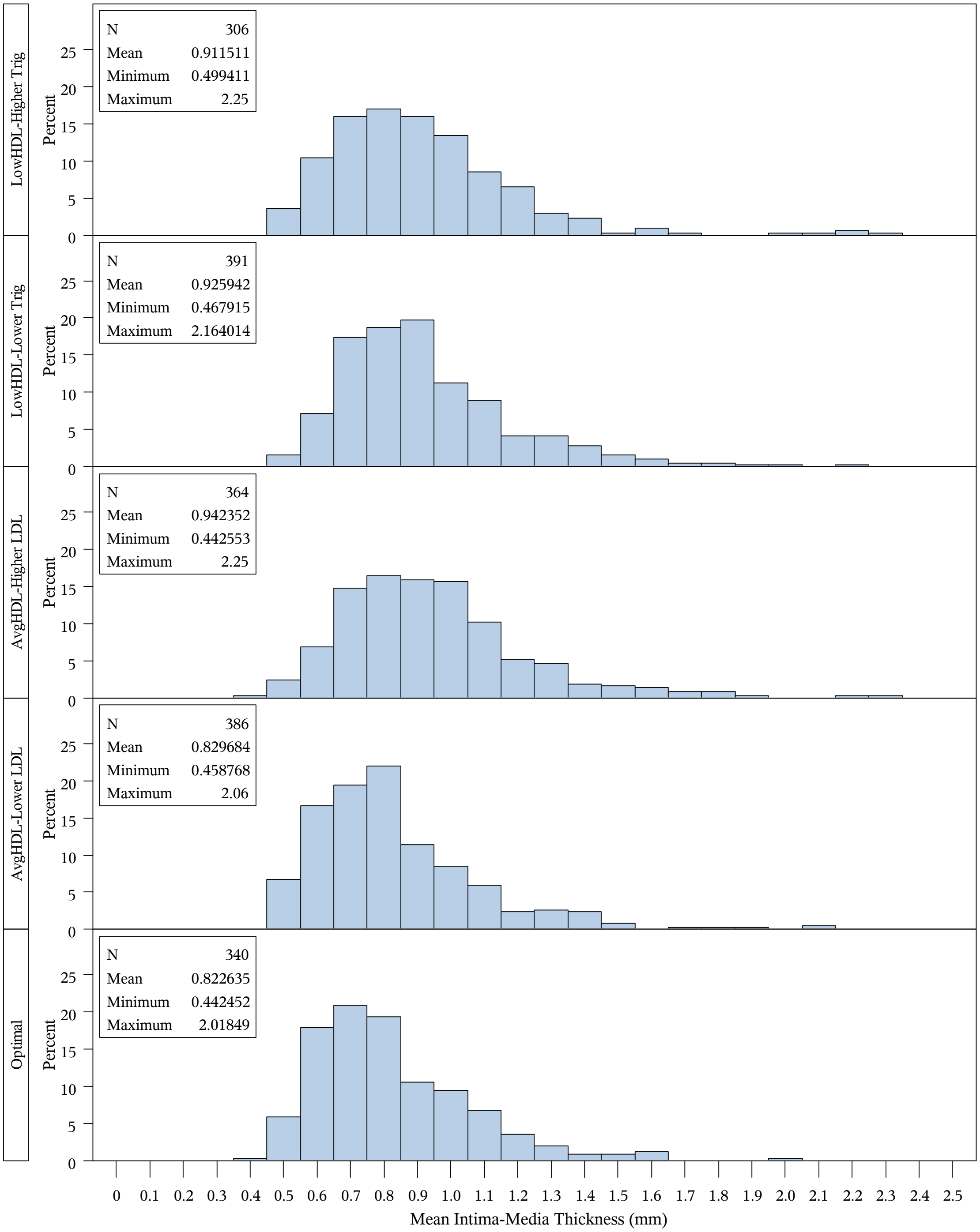


Figure 1.31: Distribution of Mean Intima-Media Thickness by Latent Class

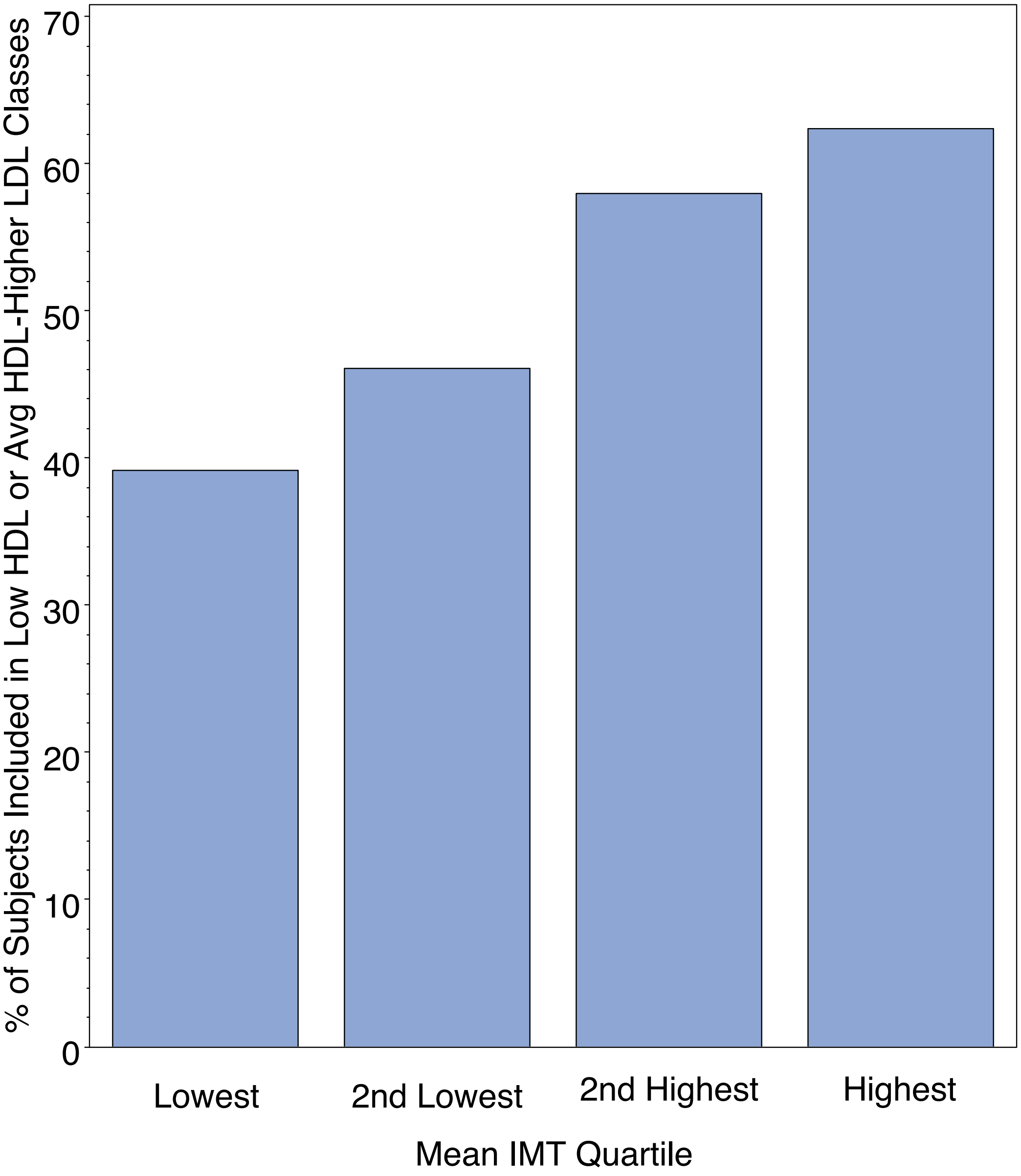


Figure 1.32: Percent of Subjects Classified in the Low HDL or Average HDL-Higher LDL Latent Classes by Mean IMT Quartile

Table 1.8: Effect of Obesity Within Identified Latent Classes

Parameter	Class	Effect of Obesity On Lab Measure	Std Error	P-Value
LDL Cholesterol	Low HDL - Higher Triglycerides	-5.84	3.00	0.0516
	Low HDL - Lower Triglycerides	-0.85	1.60	0.5952
	Avg HDL - Higher LDL	6.68	2.47	0.0068
	Avg HDL - Lower LDL	2.99	1.73	0.0839
	Optimal	3.71	2.77	0.1805
HDL Cholesterol	Low HDL - Higher Triglycerides	-2.34	0.83	0.0048
	Low HDL - Lower Triglycerides	-1.30	0.38	0.0006
	Avg HDL - Higher LDL	-1.28	0.72	0.0754
	Avg HDL - Lower LDL	-5.03	0.77	<0.0001
	Optimal	-6.36	1.32	<0.0001
Triglycerides	Low HDL - Higher Triglycerides	0.1419	0.0342	<0.0001
	Low HDL - Lower Triglycerides	0.1115	0.0237	<0.0001
	Avg HDL - Higher LDL	0.1525	0.0249	<0.0001
	Avg HDL - Lower LDL	0.2042	0.0289	<0.0001
	Optimal	0.1075	0.0316	0.0007

most values being statistically significant. While obesity is associated with a decreased HDL value of 1.28-2.34 in the less optimal classes, it is associated with a much larger 5.03-6.36 decrease in the Optimal and Avg HDL-Lower LDL classes. In other words, although individuals in the ’Optimal’ class tend to have healthy lab measures - low LDL, low triglycerides and high HDL - obese individuals in the ’Optimal’ class will likely have noticeably lower HDL than individuals in the same class who are not obese. Indeed, this is consistent with what you might expect. HDL tends to increase with aerobic exercise, weight loss, and cutting trans-fats from the diet. These are all activities that are more likely a part of the standard routine of someone who is not considered obese. And, while HDL is not likely to show large increases solely due to medication use, LDL can often be lowered by taking appropriate medications. This may help to explain why the association of obesity and LDL cholesterol is not as clear.

Finally, it is worth noting that including the obesity indicators in the LCLMM has the potential to result in the creation of different latent classes since now an additional parameter has been included in the model. However, approximately 93% of the subjects remained in the same latent class in the new model as compared with the old and the fitted class trajectories are very similar.

1.6 Conclusions and Further Research

The Latent Class Linear Mixed Model represents an important tool which researchers can use to understand complex longitudinal data. By solving simultaneously for subpopulation membership as well as the linear mixed models for those subpopulations, the statistician is given a powerful method for understanding the data. This chapter has provided advances in several main areas. First, methods have been extended to allow for structured variance components which can differ by latent class. Second, a new relative-fit model has been proposed which allows the statistician to fit each subject in the class or classes that are most appropriate. And finally, subject-specific predictions have been expanded such that each subject can receive a prediction which reflects the attributes of the class or classes that most represent that subject.

While a great deal has been done to make this model practical and general enough for the complexity of everyday problems, there are several topics which should be examined in future research. First and foremost, it would be advantageous to have a likelihood-based measure which would be useful in choosing the ideal model. There are practical considerations which argue for smaller, more parsimonious models, and other considerations which would argue for more complex models. Balancing these concerns is of great importance. With regard to variance-covariance structures, this chapter has focused on a linear structure. In order to continue to expand the options for statisticians, other correlation structures should be examined, and the bound recommended by Hathaway revisited for these other structures. Finally, this model provides a powerful way to explore complex data and there are likely many applications which could benefit from the application of these models.

Chapter 2

SAS® Macro for Latent Class Linear Mixed Models

2.1 Introduction

Chapter 1 discussed a flexible modeling approach for Latent Class Linear Mixed Models (LCLMM), in which the statistician can tailor the model in several ways to more accurately describe the data being analyzed. Among these choices are selecting fixed and random effects, which fixed effects are fit across-classes and which are fit separately for each class, whether the variances are allowed to vary by class, and how the latent classes themselves are determined. While the LCLMM has been applied in several research settings, it has not been adopted extensively for several reasons. Until now, two of the main drawbacks of the LCLMM have been slow computational speed and the lack of availability of software to run the models. Both of these are addressed in this chapter via the presentation of a new SAS® macro. This macro will allow researchers a way to easily specify models for the means, variances, and class membership, and in turn obtain parameter estimates for the underlying linear mixed models (LMMs), likely class assignments, standard errors of the parameters, and predictions of the random effects. Several computational algorithms can be chosen for estimation, including EM, gradient-based methods such as quasi-Newton, and Hessian-based methods such as Newton-Raphson. Short examples using small simulated datasets are presented to illustrate possible uses of the new macro as well as how to specify a model. In addition, simulation studies are presented to demonstrate the stability and advantage of the LCLMM.

2.2 The Model

A detailed review of the LCLMMs which can be fit in the SAS® macro is provided in Chapter 1. For convenience, a brief review is presented here.

2.2.1 Defining the Latent Class Linear Mixed Model

The LCLMM is effectively a mixture model where the underlying mixture distributions are specified as LMMs. Therefore, a brief introduction of the LMM will lead directly into the specification of the LCLMM. Note that Searle et al [1992] provides a more detailed presentation of the LMM. For subject i, we assume

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \mathbf{e}_i \tag{2.1}$$

where:

$\mathbf{y}_i$ is an $n_i \times 1$ vector of observations

$\mathbf{X}_i$ is an $n_i \times p_1$ design matrix for the fixed effects

$\boldsymbol{\beta}$ is a $p_1 \times 1$ unknown vector of fixed effects

$\mathbf{Z}_i$ is an $n_i \times q$ design matrix for the random effects

$\mathbf{b}_i$ is a $q \times 1$ unknown vector of random effects

$\mathbf{e}_i$ is an $n_i \times 1$ unknown vector of random error terms

Further, $\mathbf{e}_i$ and $\mathbf{b}_i$ are assumed to be mutually independent of one another and to have the following properties:

$$\begin{aligned}
\mathrm{E}\left(\mathbf{b}_i\right) &= \mathbf{0} \\
\mathrm{Var}\left(\mathbf{b}_i\right) &= \mathbf{D} \\
\mathrm{Cov}\left(\mathbf{b}_i, \mathbf{b}_h'\right) &= \mathbf{0} \text{ for } i \neq h \\
\mathrm{E}\left(\mathbf{e}_i\right) &= \mathbf{0} \\
\mathrm{Var}\left(\mathbf{e}_i\right) &= \mathbf{R}_i \\
\mathrm{Cov}\left(\mathbf{b}_i, \mathbf{e}_i\right) &= \mathbf{0}
\end{aligned} \tag{2.2}$$

And finally, the following distributional properties are assumed to apply to the random effects, $\mathbf{b}_i$, and residuals, $\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}$:

$$\begin{bmatrix} \mathbf{b}_i \\ \mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} \end{bmatrix} \sim N\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{D} & \mathbf{D}\mathbf{Z}_i' \\ \mathbf{Z}_i\mathbf{D} & \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i' + \mathbf{R}_i \end{bmatrix} \right) \tag{2.3}$$

The latent class model adds a further dimension in that each subject's data is modeled as a mixture of K LMMs, one for each of the K latent classes. Note that the value K is specified by the statistician a priori. Details related to selecting K are discussed later in this chapter. Following the notation in Lin et al [2002a], define for subject i:

$$c_{ik} = \begin{cases} 1 & \text{if subject } i \text{ is a member of class } k \\ 0 & \text{if subject } i \text{ is NOT a member of class } k \end{cases}$$

$$c_{i1}, \ldots, c_{iK} \sim \text{ Multinomial } (1, \pi_{i1}, \ldots, \pi_{iK})$$

The π_{ik}, the multinomial probabilities of being in each latent class, are modeled via a logit model as follows:

$$\pi_{ik} = P\left(c_{ik} = 1 \mid \mathbf{t}_i\right) = \frac{\exp\left(\mathbf{t}_i'\boldsymbol{\alpha}_k\right)}{\sum_{j=1}^{K} \exp\left(\mathbf{t}_i'\boldsymbol{\alpha}_j\right)} \tag{2.4}$$

where:

$\mathbf{t}_i$ is the design vector related to class membership for subject i

$\boldsymbol{\alpha}_k$ is an unknown vector of class-membership parameters for class k with $\boldsymbol{\alpha}_1 = \mathbf{0}$

Further, given that subject i is in class k, define the LMM for subject i as follows:

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{W}_i\boldsymbol{\lambda}_k + \mathbf{Z}_i\mathbf{b}_i + \mathbf{e}_i \tag{2.5}$$

where the following additional definitions are provided:

$\mathbf{W}_i$ is an $n_i \times p_2$ design matrix for the class-specific fixed effects

$\boldsymbol{\lambda}_k$ is a $p_2 \times 1$ unknown vector of class-specific fixed effects for class k

It is useful to note that the parameters in $\boldsymbol{\beta}$ will apply to all subjects through the values of the corresponding column in $\mathbf{X}_i$. The class-specific parameters, $\boldsymbol{\lambda}_k$, however, are different for each latent class.

The error terms and random effects are assumed to have the same properties as in Equation 2.2. Finally, similar to the LMM, the following distributional properties are assumed to apply to the random effects, $\mathbf{b}_i$, and residuals, $\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{W}_i\boldsymbol{\lambda}_k$:

$$\begin{bmatrix} \mathbf{b}_i \\ \mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{W}_i\boldsymbol{\lambda}_k \end{bmatrix} \sim N\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{D} & \mathbf{D}\mathbf{Z}_i' \\ \mathbf{Z}_i\mathbf{D} & \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i' + \mathbf{R}_i \end{bmatrix} \right) \tag{2.6}$$

It is assumed, without loss of generality, that the model is full-rank. This assumption requires that both of the following are true:

$$\begin{bmatrix} \mathbf{t}_1 & \mathbf{t}_2 & \cdots & \mathbf{t}_n \end{bmatrix} \text{ is full rank, and}$$

$$\left[\begin{array}{c|cccc|cccc} \mathbf{X}_1 & \mathbf{W}_1 & . & . & . & \mathbf{Z}_1 & . & . & . \\ \mathbf{X}_2 & . & \mathbf{W}_2 & . & . & . & \mathbf{Z}_2 & . & . \\ \vdots & . & . & \ddots & . & . & . & \ddots & . \\ \mathbf{X}_n & . & . & . & \mathbf{W}_n & . & . & . & \mathbf{Z}_n \end{array}\right] \text{ is full rank.}$$

For notational convenience, denote $\boldsymbol{\Theta}$ as the combined parameter vector comprised of $\boldsymbol{\alpha}_2 \ldots \boldsymbol{\alpha}_K$, $\boldsymbol{\beta}$, $\boldsymbol{\lambda}_1 \ldots \boldsymbol{\lambda}_k$, and $\boldsymbol{\theta}$, where $\boldsymbol{\theta}$ contains the unique variance components which determine $\mathbf{D}$ and $\mathbf{R}_i$.

It is useful to compare the observed-data likelihoods of the usual LMM with the LCLMM. These likelihoods are conditional on having observed the values $\mathbf{X}_i$, $\mathbf{W}_i$, $\mathbf{Z}_i$, and $\mathbf{t}_i$. In short,

$$\log L\left(\boldsymbol{\Theta}\right)_{LMM} = \sum_{i=1}^{n} \log f\left(\mathbf{y}_i\right) \tag{2.7}$$

$$\log L\left(\boldsymbol{\Theta}\right)_{LCLMM} = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \pi_{ik} f\left(\mathbf{y}_i \mid c_{ik} = 1\right) \tag{2.8}$$

with $f\left(\cdot\right)$ being the density defined by

$$\begin{aligned} (\mathbf{y}_i) &\sim N(\mathbf{X}_i\boldsymbol{\beta}, \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i' + \mathbf{R}_i) \\ (\mathbf{y}_i \mid c_{ik} = 1) &\sim N(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{W}_i\boldsymbol{\lambda}_k, \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i' + \mathbf{R}_i) \end{aligned}$$

Several points are of note. First, note the difference between Equations 2.7 and 2.8 - the likelihood for the LCLMM is a weighted average of K LMMs. Second, with K=1, and therefore $\pi_{i1} = 1$, the LCLMM reduces to the usual LMM. Third, if the groups are so well separated that each subject has one $\pi_{ik} = 1$ and the others equal to 0, then the likelihood reduces to that of a LMM with subjects assigned to these groups. Finally, note that the log likelihood for the LCLMM contains the log of the sum over latent classes, making the computations of the first and second derivatives of the log likelihood more complicated.

2.2.2 LCLMM: The Relative-Fit Class Membership Model

In the model described in the previous section, class membership was assumed to be determined based on an underlying logistic regression model. However, in many cases, the statistician may not know exactly which factors should be included in the model describing class membership, these factors may not be measurable, or the statistician may prefer to simply identify the best-fitting LCLMM with K classes. However, this model, in which each subject can have its own set of mixture probabilities, would result in many parameters related to class membership - specifically, for n subjects and K classes, the model would require $n \times (K-1)$ parameters. Therefore, a second approach was proposed in Chapter 1 which replaces the logistic regression model with a model in which class membership is determined by the relative fit of the underlying LMMs. This is much less costly in terms of computational time than fitting a separate set of class membership parameters for each subject, but offers a logical approach with a similar goal in mind.

Recall that the unconditional log likelihood of the observed data, assuming class-specific **D** and **R**, is as follows:

$$\log L(\boldsymbol{\Theta}) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \left[\pi_{ik} f\left(\mathbf{y}_i \mid c_{ik} = 1\right)\right] \tag{2.9}$$

with $f\left(\mathbf{y}_i \mid c_{ik} = 1\right)$ being the density defined by

$$\left(\mathbf{y}_i \mid c_{ik} = 1\right) \sim N(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{W}_i\boldsymbol{\lambda}_k, \mathbf{Z}_i\mathbf{D}_k\mathbf{Z}_i' + \mathbf{R}_{ik})$$

In the revised model, the π_{ik} are determined solely by the relative fit of the underlying LMMs according to the following equation:

$$\hat{\pi}_{ik} = \frac{f\left(\mathbf{y}_i \mid c_{ik} = 1; \hat{\boldsymbol{\Theta}}\right)}{\sum_{j=1}^{K} f\left(\mathbf{y}_i \mid c_{ij} = 1; \hat{\boldsymbol{\Theta}}\right)} \tag{2.10}$$

For example, if two latent classes are fit and the LMM for Class 1 results in a likelihood for the subject's data equal to the likelihood of that subject's data under the model for Class 2, then the subject would be included in each class 50%/50%. If the likelihood for Class 1 was nine times higher than for Class 2, then the subject would be included in Class 1 90 percent and Class 2 only 10 percent. Further details are presented in Section 1.4.4.

2.2.3 Prediction of the Random Effects

One of the many advantages to using LMMs is the logical specification of the model, and a big part of this specification lies in the random effects which are thought of as being sampled from an underlying distribution. For the LMM, the underlying distribution is the multivariate normal distribution. The LCLMM, in essence, divides the population into subpopulations and then fits the best LMMs to the data in each subpopulation. Therefore, it is important that the LCLMM approach be able to compute predictions of the random effects. In fact, upon closer examination, the random effects for the LCLMM are actually a weighted average of the random effects from each underlying LMM.

From the derivation of the LCLMM, recall that given a subject is a member of class k, the following distributional properties are assumed to apply to the random effects, $\mathbf{b}_i$, and residuals, $\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{W}_i\boldsymbol{\lambda}_k$.

$$\left.\begin{bmatrix} \mathbf{b}_i \\ \mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{W}_i\boldsymbol{\lambda}_k \end{bmatrix} \sim N\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{D}_k & \mathbf{D}_k\mathbf{Z}_i' \\ \mathbf{Z}_i\mathbf{D}_k & \mathbf{Z}_i\mathbf{D}_k\mathbf{Z}_i' + \mathbf{R}_i \end{bmatrix} \right) \;\right|\; c_{ik} = 1 \tag{2.11}$$

The computation of the LCLMM random effects requires maximizing the joint density of the data $\mathbf{y}$ and the random effects $\mathbf{b}_i$, with respect to the terms $\mathbf{b}_i$. The likelihood can be written as follows:

$$\log L(\mathbf{y}, \mathbf{b}) = \sum_{i=1}^{n} \log \underbrace{\sum_{k=1}^{K} \pi_{ik} \underbrace{\left[\frac{\exp\left[-\frac{1}{2}\left\{(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{W}_i\boldsymbol{\lambda}_k - \mathbf{Z}_i\mathbf{b}_i)' \mathbf{R}_{ik}^{-1} (\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{W}_i\boldsymbol{\lambda}_k - \mathbf{Z}_i\mathbf{b}_i) + \mathbf{b}_i'\mathbf{D}_k^{-1}\mathbf{b}_i\right\}\right]}{(2\pi)^{(n_i+q)/2} \left|\mathbf{R}_{ik}\right|^{1/2} \left|\mathbf{D}_k\right|^{1/2}} \right]}_{P_{3ik}}}_{P_{1i}}$$

However, given that a subject is in class k implies that the best set of predictions of the random effects are the predictions based on the k^{th} LMM. Therefore, the random effects for a subject can be computed as the weighted average of the random effects computed as if that subject were in each latent class:

$$\widetilde{\mathbf{b}}_i \;=\; \text{eBLUP}\left[\mathbf{b}_i\right] \;=\; \text{E}\left[\mathbf{b}_i \mid \hat{\boldsymbol{\Theta}}, \mathbf{y}_i\right] \;=\; \sum_{i=1}^{K} \pi_{ik} \times \widetilde{\mathbf{b}}_{i|k}$$

The resulting random effects are available in the SAS® macro discussed in the next section. Additional details related to predictions of the random effects are presented in Section 1.4.5.

2.3 A New SAS® Macro

A new SAS® macro has been developed which allows the user to fit LCLMMs. This macro is publicly available via the following link: http://www.mysitehere.com. This section reviews the input files required to run the SAS® macro, presents some specifics related to how the macro works, and describes the resulting output files.

2.3.1 Directory Structure

In the macro's calling program, a BasePath is specified by the user - for example c:\LatentClass. The directory named as BasePath should contain the program and macros. Two subdirectories should be created underneath the BasePath - these should be named 'in' and 'out' - for example c:\LatentClass\in and c:\LatentClass\out. Input files, described in the next section, will be stored in the 'in' subdirectory, and the macro will output the results to the 'out' subdirectory.

2.3.2 Input Files

Overview

For purposes of describing how to specify the files required by the macro, the below sections assume the researcher is starting with a SAS® dataset which contains one record for each response, with the many responses for a subject stacked so they are on separate records - this is similar to how one would prepare the data to run in SAS® PROC MIXED. As with the LMM, it is not required that all subjects have complete data. Instead, each *record* being modeled must be complete. Therefore, it is necessary to first delete *records* for which there are missing responses or predictors. In terms of preparing to run the data in the SAS® macro, rename/create variables in your SAS® dataset (stacked as described above) so that the following variables are present. Note that all variables except 'subject' must be numeric. In addition, indicator variables must be specified as 1/0, and indicators must be created for each possible value for class variables. In other words, the matrices **X**, **W**, and **Z** must be specified in full.

subject – a unique subject identifier

y – the continuous responses

x1 · · · x? – across-class fixed effect numbers 1-?
– the parameters related to these WILL NOT vary by latent class

w1 · · · w? – class-specific fixed effect numbers 1-?
– separate parameters for these variables will be estimated for each latent class

z1 · · · z? – random effect numbers 1-?

All data for a particular subject should be stored consecutively. In many cases, the z-variables will be a subset of the x-variables, since it is common to have random effects which are also fit as fixed effects. After you've set your data up in this way, the remaining steps are relatively simple.

Continuous Responses

Create a dataset named 'y' which will contain only the variable 'y' from the dataset described above. Create a second dataset named 'info' which will contain only the variable 'subject' from the dataset described above. These datasets will provide information related to the responses and which responses came from which subjects.

Fixed Effects

First, create the dataset of fixed effects which will apply to all latent classes. Create a dataset named 'x' which will contain only the variables x1-x? from the dataset described above. It is important that the variables be present in the dataset in the order X1, X2, X3, etc. since they will be read in from the

’x’ dataset in the same column order as they exist in the dataset, regardless of the variable naming. If you would like all fixed effects to vary by latent class, then do NOT create an ’x’ dataset at all - this will indicate to the macro that all fixed effects vary by class.

Second, create the dataset of class-specific fixed effects. Create a dataset named ’w’ which will contain only the variables w1-w? from the dataset described above. Once again, it is important that the variables be present in the dataset in the order W1, W2, W3, etc. since they will be read in from the ’w’ dataset in the same column order as they exist in the dataset. If you do not want to fit any class-specific fixed effects (i.e. the usual LMM with a mixture distribution for the error or random-effects distribution) - then do not create a ’w’ dataset at all - this will indicate to the macro that all fixed effects are across-class.

Random Effects

Create a dataset named ’z’ which will contain only the variables z1-z? from the dataset described above. Once again, it is important that the variables be present in the dataset in the order Z1, Z2, Z3, etc. since they will be read in from the ’z’ dataset in the same column order as they exist in the dataset.

Variance-Covariance Structure

In this SAS® macro, the variance-covariance structure is determined by the specification of random effects, which are assumed to have a linear correlation structure, as well as a diagonal structure for the error terms, with different residual error variances permitted for different measures.

First, with regard to the residual error variance, create a dataset named ’r_structure’ with a single variable named ’r_structure’ and with the same number of records as ’y’. The variable ’r_structure’ will contain the number of the residual error variance which applies to each record in ’y’. For example, if all records had the same residual error variance, the variable would always contain the value 1. If each subject had five records with one variance followed by five records with a different variance, the variable would contain five 1’s followed by five 2’s for each subject. Note that if this file does not exist, the macro assumes that all records have the same residual error variance.

Second, with regard to the variances of the random effects, create a dataset named ’d_structure’ with a single variable named ’d_structure’ and with the number of records equal to the number of random effects (number of z* variables). The variable will contain the number of the variance which applies. For example, if all random effects had the same variance, the variable would always contain the value 1. If each random effect had a different variance and there were 4 random effects, the file would have 4 records with values for 'd_structure' of 1, 2, 3, and 4. Note that if this file does not exist, the macro assumes that all random effects have a different variance.

Finally, with regard to the correlation structure for the random effects, recall that the correlation between random effects is fit across classes, even if the variances are allowed to differ between classes. Create a dataset named 'd_corr' with the number of records and number of variables equal to the number of random effects (number of z* variables). Name the variables d_corr_1 to d_corr_?. It is easiest to think of this file as the matrix of correlations of the random effects - then it is possible to discuss cell(x,y) of the file. The cells of this file will contain the number of the correlation which applies. For example, if there are four random effects and only the first two are correlated, then the (1,2) cell and (2,1) cell of the file should contain the value 1 and all other cells should be set to 0. If there are 4 random effects and the first two random effects are correlated and the last two random effects are correlated but the first/second are not correlated with the third/fourth, then the (1,2) cell and (2,1) cell of the file should contain the value 1, the (3,4) cell and (4,3) cell of the file should contain the value 2, and all other cells should be set to 0. Note that if this file is not created and the macro is run with the option specified that there should be correlation between the random effects, then the macro assumes that the variance-covariance matrix of the random effects, **D**, is unstructured.

Class Membership

The remaining file to be specified is the file of variables which determine class membership. These are used in the logistic model related to class membership. Create the file 'v' with variables named v1, v2, ... v? in that order. The file should have one record per subject. If all subjects are to have the same mixture, which can be useful when running the usual LMM with mixtures for the error and/or random effect distributions, then a single variable would be present, v1, which would contain 1 for all subjects. If several variables are thought to impact class membership, say age and baseline/change from baseline for some measure, each subject would have 4 variables, named v1 v2 v3 v4, which for each subject would contain the intercept (1) and these three measures. If each subject were permitted to be fit by that's subject's best-fitting model, then there would be a v* variable for each subject - effectively, the 'v' file would be an identity matrix.

Note that this last scenario, in which each subject can be fit in that subject's best-fitting class, results in many parameters and relatively long runtimes. Therefore, the relative-fit model was proposed earlier. If the dataset 'v' is not created and a gradient-based or Hessian-based computational method is selected, then the π_{ik} for each subject is estimated based on the relative fit of the underlying LMMs for that subject's data. For example, if there are two classes and the LMM for Class 1 results in a likelihood for the subject's data equal to the likelihood of that subject's data under the model for Class 2, then the subject would be included in each class 50/50. If the likelihood for Class 1 was nine times higher than for Class 2, then the subject would be included in Class 1 90 percent and Class 2 only 10 percent. This model runs much more quickly and will produce similar results when the classes are somewhat distinct. If the EM method is selected and 'v' is not specified, then a 'v' file is created automatically which is an identity matrix and computations proceed. The relative-fit model is not fit in this case because the unknown-data portion of the EM algorithm, which has been class membership, is now a deterministic function of the likelihoods for each underlying LMM.

2.3.3 Additional Inputs

In addition to the input files, the user must specify additional information prior to running the macro. This includes the following choices: number of latent classes, computational method, whether or not to fit correlation between the random effects, whether variances in $\mathbf{D}$ and/or $\mathbf{R}$ are permitted to vary by class, how to model class membership, and whether to calculate standard errors (and which method to use) or predictions of the random effects. Comments related to these choices are provided in the following sections.

Number of Latent Classes

For each run of the SAS® macro, the user is required to specify the number of latent classes to fit - this is entered in the macro variable 'NumClasses'. In some cases, the user may have a pretty good idea of how many unique classes they are interested in. In other cases, the user may not be sure how many classes exist and would be served best by running the macro several times, each with a different number of classes. Along with the final estimates, the macro computes several information criteria for each model. By comparing information criteria for a range of classes, it may be apparent which number of classes appears to fit best. Section 2.8 presents the results of a simulation study which compares the performance of various information- and residual-based criteria with respect to determining the appropriate number of latent classes. As noted in that study, the user may also wish to compare a more practical measure such as mean square error for each parameter fit. Since this comparison must be made for each unique parameter being modeled, this measure is not automatically output.

Computational Methods

In previous research, LCLMMs have almost exclusively been fit using the EM algorithm. However, with its slow convergence properties and the complexity of the LCLMMs, computation can be too slow for practical use. The new SAS® macro presented here allows the user to select from several gradient-based methods and Hessian-based methods, in addition to running the EM algorithm. A simulation study is presented in Section 2.9 which provides the basis for recommendations as to which algorithm works best for various size models.

The user must choose from the following options - note that the computational method is specified in the macro variable ’Method’:

EM – EM Algorithm
CG – Conjugate Gradient Algorithm
QN – Quasi-Newton Algorithm
NRA – Newton-Raphson Algorithm (without ridging)
NRR – Newton-Raphson Algorithm (with ridging)

The EM and Newton-Raphson algorithms are discussed in detail in Chapter 1. Additional details related to the other algorithms can be found in Chapter 4 of the SAS®/OR 9.2 User’s Guide [2008]. Note that when fitting only one latent class (generalizing the LMM), the user cannot specify ’Method’=’EM’.

Is there non-zero correlation between random effects?

As in the LMM, the user must determine whether to model a correlation structure for the random effects. If this is desired, then the associated macro variable in the latent class program, ’D_HasCorr_YN’, must be set to ’Y’. The structure of the file containing the linear correlation structure was discussed earlier.

Should variances be permitted to vary by latent class?

In many situations, the variances for one class may be very different than another class. If there is reason to believe this may be true, then the model should be run allowing variances to differ by latent class. This is controlled by setting the macro variables ’D_VarDiffByClass_YN’ and ’R_VarDiffByClass_YN’ to either ’Y’ or ’N’. In Chapter 1, it was observed that such a decision can have a noticeable effect on the determination of latent classes and the underlying LMMs.

Class Membership - Structured on Unstructured?

Class membership may be specified in several ways. In some cases, the statistician may simply want to generalize the distributional assumptions of the usual LMM to allow the underlying normal distributions to instead be mixtures of normal distributions. This would be specified by defining the file 'v' to be a single column of 1's and setting the macro variable 'PieMethod' equal to 'STRUCTURED'. In more general situations where interest focuses on discovering unique groups in the data and fitting appropriate LMMs to these groups, class membership may be specified in two very different ways. When the statistician is interested in simply fitting the best K LMMs or is not sure of which factors may be most important in determining class membership, running the analysis with class membership determined solely by the relative fit of the underlying LMMs would be recommended - specify 'PieMethod'='UNSTRUCTURED'. Alternatively, the statistician may already have a set of risk factors which are of interest, and one of the goals of the analysis might be to determine the relationship between the risk factors and latent class membership. In this situation, the file 'v' would be specified with factors relevant to class membership, and 'PieMethod' would be set to 'STRUCTURED'. Note that when fitting only one class (LMM), the user must specify 'PieMethod'='UNSTRUCTURED'.

Standard Errors and Predictions

The main algorithm does not, by default, compute standard errors for the parameters. If you would like to obtain parameter standard errors, set 'CalcParmStdErr'='Y' and specify a computational method for standard errors in the variable 'SEMethod'. For Hessian-based methods, the most efficient method is to compute the standard errors by making use of the computed Hessian - specify 'SEMethod'='HES'. This is probably a good choice even when other methods are used for the main algorithm. The standard errors can also be computed by taking finite differences of the gradient (choose 'SEMethod'='GRD'), or by taking finite differences of the likelihood itself (choose 'SEMethod'='LIK'). Note that this last choice is extremely slow and not recommended.

Miscellaneous

Two additional features are worth note in terms of user options. First, in order to increase the chances of arriving at a global likelihood maximum rather than a local maximum, the user is provided the option of running several preliminary runs from different starting points and using the best of these starting values to initiate the model. This number of base runs is specified in a macro variable 'HowManyBaseRuns', and the seeds used to randomly assign latent classes to initiate those base runs are stored in the macro variables 'SeedForClasses_x'. These preliminary runs iterate for ten iterations with the random effect variances set to zero.

Second, since the derivations for the first and second derivatives are complex, a self-check has been programmed which allows the user to confirm that the method's computations are working properly. If the macro variable 'PerformSelfCheck' is set to 'Y', then at several points in the algorithm, the algorithm will compute quantities using finite differences and print these as well as the values computed using the derived first and second derivatives. This will add to the runtime but can be helpful in that it can confirm that the model is working as it is supposed to.

2.3.4 How the Macro Works

Computation of Starting Values

Several authors have explored methods of choosing the best starting values. While some have examined models which are mixtures of multivariate Normal distributions, none have examined this in the context of the LCLMM. Coleman and Woodruff [2000] proposed to choose a random sample of the data and classify these subjects into groups. This initial clustering would then provide the basis for the starting point of the full algorithm. Karlis and Xekalaki [2003] concluded that the best method is to start from several different initial values, making a small number of iterations from each. The authors then recommended choosing the set of starting values with the largest likelihood after the initial iterations. The authors pointed out that this approach helps to reduce the amount of time spent in areas of a flat likelihood away from the global maximum. Biernacki et al [2003] also examined the issue of choosing starting values for the EM algorithm in multivariate Gaussian mixture models. The authors found that, like Karlis and Xekalaki [2003], running short runs of EM and choosing the one with the highest likelihood tended to work well. The authors further recommended compounding this method by running batches of short runs, choosing the best run from each batch, and then using the best runs as the starting points for another search. However, this method may require a great deal of computational effort and time.

In order to ensure a reasonable set of starting values for the full model fit, a preliminary set of steps is followed at the beginning of the SAS® macro. These methods appear to provide a stable algorithm which will work in most scenarios. The stability of this method and the computations as a whole are examined in Sections 2.5 and 2.6. The steps for computing starting values are as follows:

- Each subject is randomly assigned to a latent class
- Ordinary Least Squares estimates of the fixed effects are calculated using these random class assignments
- 10 preliminary iterations are run without random effects using only the structured **R** matrix. These iterations do not permit variances to differ by latent class. As noted earlier, the user can specify to repeat this process as many times as they would like and choose the best likelihood from all of the various runs - the default is 5.
- The random effect variances are set to 1e-5. Correlations between random effects are set to 1e-10. Preliminary estimates of the variances of the random effects and revised residual error variances are calculated by running one step of Fisher scoring (using the actual second derivative via Newton-Raphson appeared to be less stable here when compared to using its expected value). Variances are still not allowed to differ by latent class.
- 5 EM iterations are run. Variances are still not allowed to differ by latent class, and variances continue to be updated via Fisher scoring.
- If the user chose to allow variances to differ by latent class, then 5 additional EM iterations are run. In these, variances are finally allowed to differ by latent class. Once again, variances are updated using Fisher scoring.
- One additonal EM iteration is run. Variances are updated using Newton-Raphson in this step as well as any future EM steps. Note that if there are no class-specific fixed effects, it was assumed for all prior iterations that the across-class effects were actually class-specific. This allows for some logical separation of the groups. These class-specific effects are dropped in this iteration such that the only parameters that differ by class are the variances.
- The resulting estimates are used as starting values for the main algorithm.

Description of Model Fit

Once starting values are obtained, one of three different routines is run, depending on the choice of algorithm.

The first is the EM algorithm which was also used as part of the process for generating stable starting values. The steps required in the EM algorithm are detailed in Chapter 1. The main change with regard to computation is that variances for the starting values were updated at each step using Fisher scoring while the main algorithm uses Newton-Raphson.

The second routine uses gradient-based methods. An IML module has been programmed to compute the gradient of the likelihood for various scenarios (i.e. fixed effects across-class/class-specific, variances across-class/class-specific, class membership structured/unstructured, etc.). With this in place, the macro utilizes SAS®' built-in engines to compute estimates using the selected algorithm. Methods include conjugate gradient and quasi-Newton.

The third routine uses Hessian-based methods. IML modules have been programmed to compute the gradient and Hessian of the likelihood for various scenarios. Once again, the macro utilizes SAS®' built-in engines to compute estimates using the selected algorithm. Methods include Newton Raphson, with or without ridging.

Screenshot During Model Fit

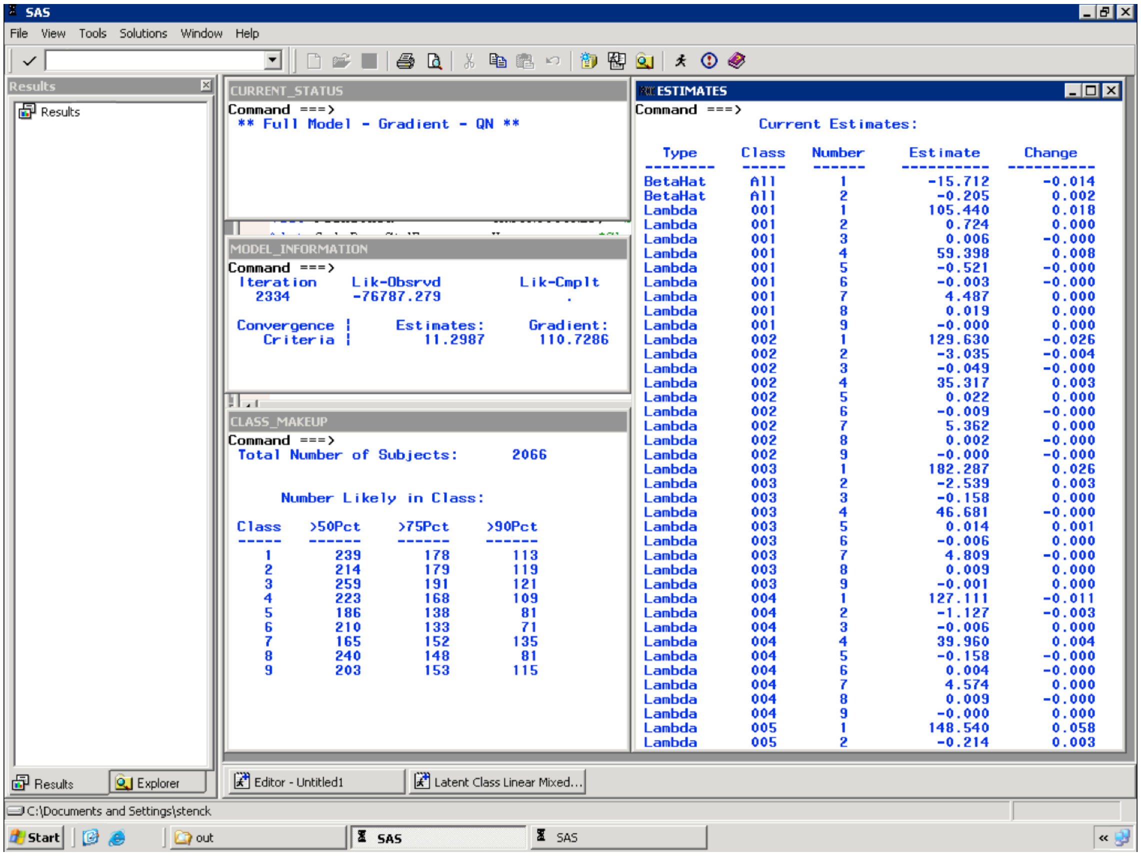
Since a fair amount of information is available at each iteration regarding progress, estimates, class membership, etc., the macro updates information on the screen after each iteration. A sample screenshot is displayed in Figure 2.1.

The top left window displays the current status. This window displays whether the algorithm is obtaining initial estimates or running the full model, and it updates with additional details (i.e. 'calculating betahat') as it moves through the various calculations.

The middle left window displays the observed likelihood (also complete-data likelihood for EM iterations) as well as the convergence criteria. The convergence criteria for the estimates is simply the maximum percent change in estimates from one iteration to the next. The convergence criteria for the gradient is the maximum absolute value of the elements of the gradient at that iteration.

The bottom left window displays the number of subjects being modeled as well as a summary of class membership. For each class, the number of subjects who have greater than a 50 percent probability of being in the class, greater than 75 percent chance, and greater than 90 percent chance are displayed. This allows the user to see how much the classes are changing, how sure the algorithm is about its class assignments, whether a particular class appears to contain just one subject, etc.

Finally, the right window displays the current parameter estimates as well as the changes in those estimates since the last iteration.



CURRENT_STATUS

```
Command ===>
 ** Full Model - Gradient - QN **
```

MODEL_INFORMATION

```
Command ===>
 Iteration     Lik-Obsrvd           Lik-Cmplt
   2334        -76787.279               .

 Convergence |      Estimates:     Gradient:
    Criteria |        11.2987       110.7286
```

CLASS_MAKEUP

```
Command ===>
 Total Number of Subjects:      2066


      Number Likely in Class:

 Class    >50Pct     >75Pct     >90Pct
 -----    ------     ------     ------
    1        239        178        113
    2        214        179        119
    3        259        191        121
    4        223        168        109
    5        186        138         81
    6        210        133         71
    7        165        152        135
    8        240        148         81
    9        203        153        115
```

ESTIMATES

```
Command ===>
                Current Estimates:

   Type      Class   Number     Estimate      Change
 ---------   -----   ------   -----------  -----------
 BetaHat      All       1         -15.712       -0.014
 BetaHat      All       2          -0.205        0.002
 Lambda       001       1         105.440        0.018
 Lambda       001       2           0.724        0.000
 Lambda       001       3           0.006       -0.000
 Lambda       001       4          59.398        0.008
 Lambda       001       5          -0.521       -0.000
 Lambda       001       6          -0.003       -0.000
 Lambda       001       7           4.487        0.000
 Lambda       001       8           0.019        0.000
 Lambda       001       9          -0.000        0.000
 Lambda       002       1         129.630       -0.026
 Lambda       002       2          -3.035       -0.004
 Lambda       002       3          -0.049       -0.000
 Lambda       002       4          35.317        0.003
 Lambda       002       5           0.022        0.000
 Lambda       002       6          -0.009       -0.000
 Lambda       002       7           5.362        0.000
 Lambda       002       8           0.002       -0.000
 Lambda       002       9          -0.000       -0.000
 Lambda       003       1         182.287        0.026
 Lambda       003       2          -2.539        0.003
 Lambda       003       3          -0.158        0.000
 Lambda       003       4          46.681       -0.000
 Lambda       003       5           0.014        0.001
 Lambda       003       6          -0.006        0.000
 Lambda       003       7           4.809       -0.000
 Lambda       003       8           0.009        0.000
 Lambda       003       9          -0.001        0.000
 Lambda       004       1         127.111       -0.011
 Lambda       004       2          -1.127       -0.003
 Lambda       004       3          -0.006        0.000
 Lambda       004       4          39.960        0.004
 Lambda       004       5          -0.158       -0.000
 Lambda       004       6           0.004       -0.000
 Lambda       004       7           4.574        0.000
 Lambda       004       8           0.009       -0.000
 Lambda       004       9          -0.000        0.000
 Lambda       005       1         148.540        0.058
 Lambda       005       2          -0.214        0.003
```

Figure 2.1: Screenshot of SAS® Macro While Running

2.3.5 Output Files

The following output datasets are produced in the estimation process. These are stored in the 'out' subdirectory.

Final_Estimates – Final model information, estimates, and standard errors
Final_Pie – Final probability of being in each latent class (logistic)
Final_Pie_RelFit – Final probability of being in each latent class (relative fit)
Final_Random_Effects – Final random effect estimates and standard errors
Fitted_Values – Fitted values, including and excluding random effects

ParmCorr – Correlation matrix of parameter estimates
AllParms – All parameter estimates by iteration
Gradient – Gradient values for each parameter by iteration
Lik – Likelihood (complete-data and observed) by iteration
Conv_Criteria – Convergence criteria by iteration
RunTime – Runtime in seconds by iteration

BetaHat – Estimates of the across-class fixed-effects by iteration
Lambda – Estimates of the class-specific fixed-effects by iteration
Vars – Estimates of the variances/correlations by iteration

C – (EM only) Bayes class probabilities by iteration
Pie – Probability of being in each latent class by iteration (logistic)
PieParms – Estimates of class-membership parameters by iteration (logistic)

In additon, the estimates at each iteration are saved in a PDF document. This file is named IterationSummary.pdf and is stored in the 'out' subdirectory.

2.4 Description/Specification of Sample Applications

LCLMMs are particularly useful for several modeling scenarios. In order to illustrate the various uses of the model as well as provide examples for the user regarding how to specify the models to run in the new SAS® macro, a small simulated data example is presented for three modeling approaches in the following sections.

2.4.1 Linear Mixed Model with Random Effect Distribution as a Mixture

In the LMM, the error terms and random effects are assumed to be normally distributed. However, in many applications, this may not be accurate. A mixture of two or more normal distributions may more accurately describe the distributions of one or both of these terms. By specifying a mean model in which none of the fixed effects differ by latent class, that all subjects be fit with the same mixture, and that the random effect and/or error variances be permitted to vary by latent class, the LCLMM allows this extension of the LMM to be fit efficiently.

As an example, consider a model in which there are 1,000 subjects. Each subject is measured for some continuous trait ten times. A possible LMM to describe this is as follows:

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

where:

y_{ij} is the value of the trait for subject i at observation j

μ is the mean value for the trait

α_i is the random intercept for subject i

ϵ_{ij} is the error term associated with subject i at observation j

$\alpha_i \sim$ Normal $\left(0, \sigma_r^2\right)$

$\epsilon_{ij} \sim$ Normal $\left(0, \sigma_e^2\right)$

In words, a subject's mean value for the trait is random about the global mean, μ. These random intercepts are distributed according to a normal distribution with mean 0 and variance σ_r^2. The error terms are distributed according to a normal distribution with mean 0 and variance σ_e^2. Now suppose that the distribution of the random intercepts is instead a mixture of two normal distributions.

$$\alpha_i \sim \text{ 50 percent Normal}\left(0, \sigma^2_{rn}\right) \text{ + 50 percent Normal}\left(0, \sigma^2_{ro}\right)$$

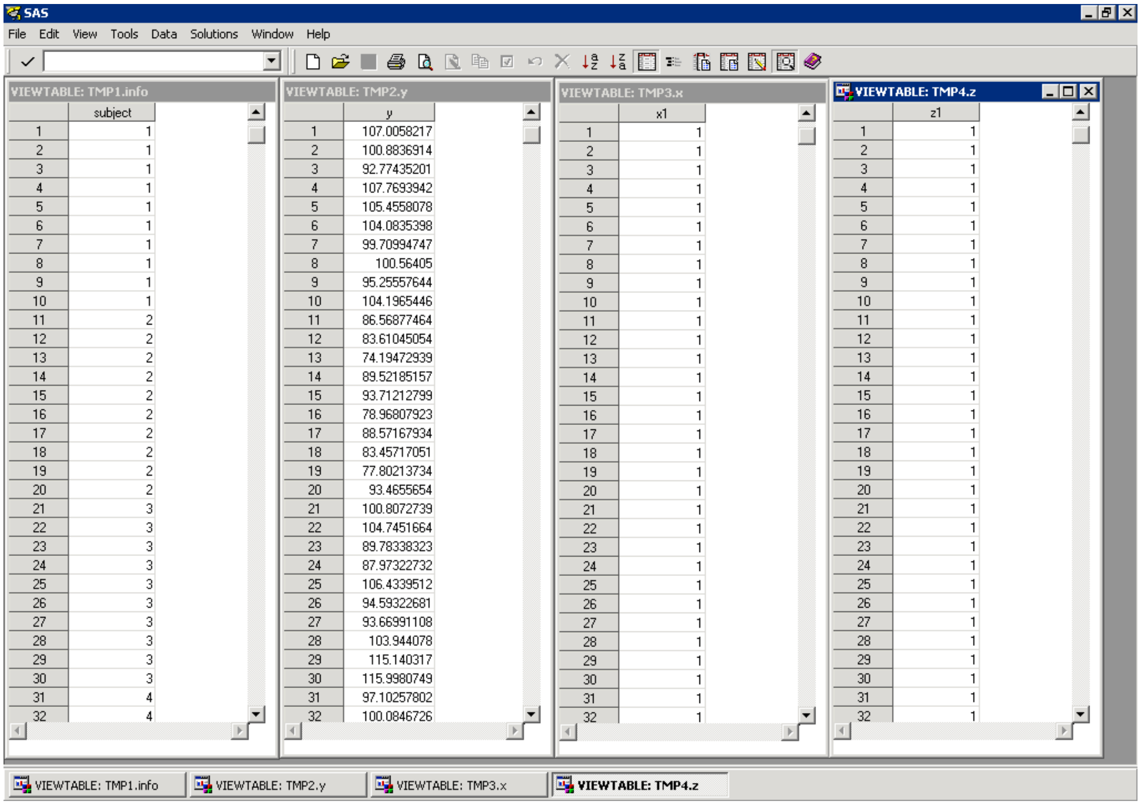
For purposes of illustration, the following values were used to generate data:

$$\begin{aligned} \mu &= 100 \\ \sigma^2_{rn} &= 20 \\ \sigma^2_{ro} &= 200 \\ \sigma^2_{e} &= 40 \end{aligned}$$

In order to run the model through the SAS® macro, the user would specify the files in the format presented in Figures 2.2 and 2.3. Note that it is not necessary to specify D_Corr, D_Structure, and R_Structure since there is only one random effect (no correlation and default for D_Structure is to have a separate variance for each random effect) and since all observations are assumed to have the same residual error variance (also the default). Also, since class membership is determined based on the logistic model, having the same intercept-only specification for each subject in the class-membership model assures that all subjects will be fit with the same mixture.

The usual LMM was run as well as the LCLMM assuming that all subjects have the same mixture. Figure 2.4 displays the options that need to be set to run each of the models. Note that when only one class is run (LMM), the macro variable 'PieMethod' must be set to 'UNSTRUCTURED' and 'Method' cannot be set to 'EM'. The parameter estimates from each model are shown in Table 2.1. Since the LMM effectively fits a combined variance to reflect σ^2_{rn} and σ^2_{ro}, the variable σ^2_{ro} is populated in the table only for the LCLMM. It is interesting to note that the random effect variance estimate from the LMM is very close to the weighted average of the random effect variance estimates from the two latent classes in the structured model.

As seen in Figure 2.5, the predictions of the random effects from the LCLMM are very similar to those from the LMM. However, note that the distribution of these random effects is NOT in fact normal. The histograms of the actual random effects generated in creating the data as well as the predicted random effects from each of the models are presented in Figure 2.6. Notice that, while the



VIEWTABLE: TMP1.info

	subject
1	1
2	1
3	1
4	1
5	1
6	1
7	1
8	1
9	1
10	1
11	2
12	2
13	2
14	2
15	2
16	2
17	2
18	2
19	2
20	2
21	3
22	3
23	3
24	3
25	3
26	3
27	3
28	3
29	3
30	3
31	4
32	4

VIEWTABLE: TMP2.y

	y
1	107.0058217
2	100.8836914
3	92.77435201
4	107.7693942
5	105.4558078
6	104.0835398
7	99.70994747
8	100.56405
9	95.25557644
10	104.1965446
11	86.56877464
12	83.61045054
13	74.19472939
14	89.52185157
15	93.71212799
16	78.96807923
17	88.57167934
18	83.45717051
19	77.80213734
20	93.4655654
21	100.8072739
22	104.7451664
23	89.78338323
24	87.97322732
25	106.4339512
26	94.59322681
27	93.66991108
28	103.944078
29	115.140317
30	115.9980749
31	97.10257802
32	100.0846726

VIEWTABLE: TMP3.x

	x1
1	1
2	1
3	1
4	1
5	1
6	1
7	1
8	1
9	1
10	1
11	1
12	1
13	1
14	1
15	1
16	1
17	1
18	1
19	1
20	1
21	1
22	1
23	1
24	1
25	1
26	1
27	1
28	1
29	1
30	1
31	1
32	1

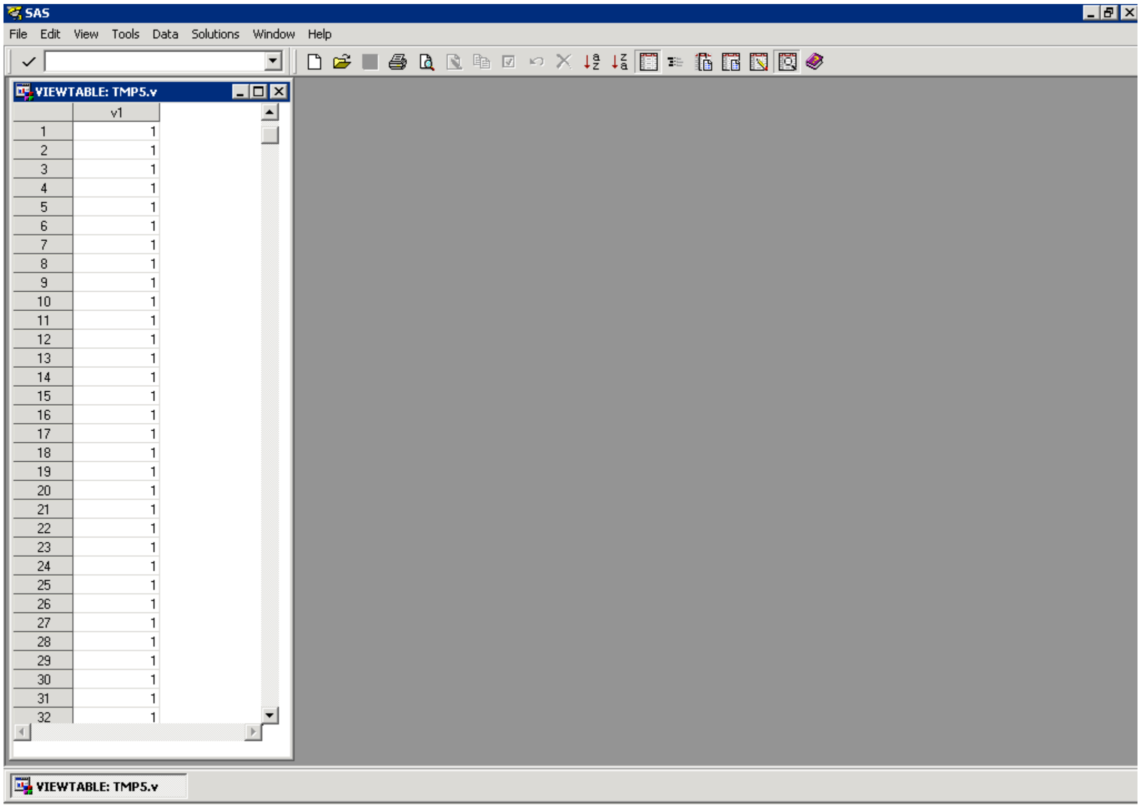
VIEWTABLE: TMP4.z

	z1
1	1
2	1
3	1
4	1
5	1
6	1
7	1
8	1
9	1
10	1
11	1
12	1
13	1
14	1
15	1
16	1
17	1
18	1
19	1
20	1
21	1
22	1
23	1
24	1
25	1
26	1
27	1
28	1
29	1
30	1
31	1
32	1

Figure 2.2: Screenshot of SAS® Datasets Needed to Run Example 1

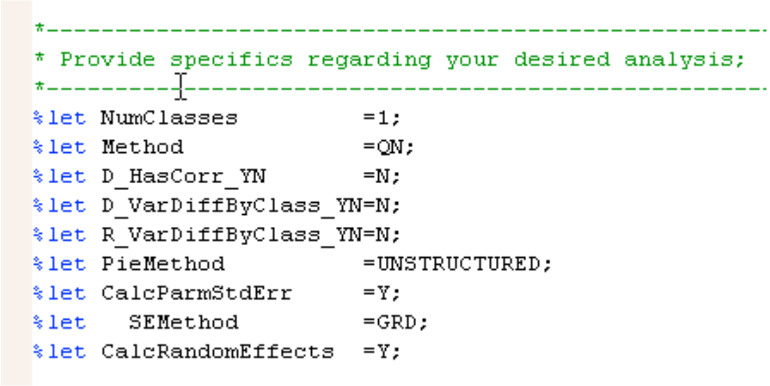
Table 2.1: Random Effect Distribution as a Mixture - Fitted Values

Parameter	Actual Value	LMM Estimate	LCLMM Estimate	Class Probabilities
μ	100	100.12	99.86	
σ^2_{rn}	20	114.83	20.26	52.6%
σ^2_{ro}	200	n/a	219.95	47.4%
σ^2_e	40	40.15	40.15	



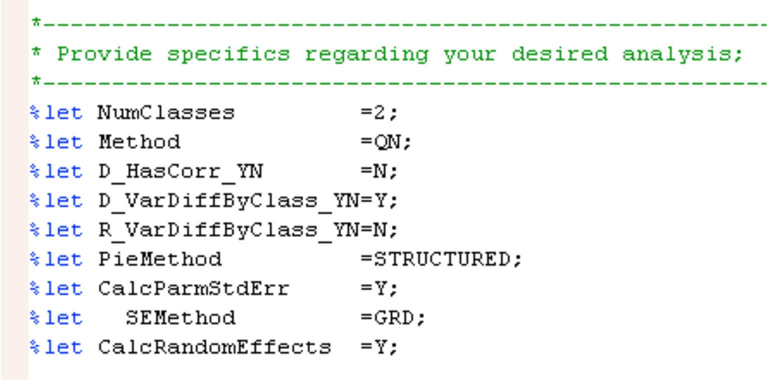
	v1
1	1
2	1
3	1
4	1
5	1
6	1
7	1
8	1
9	1
10	1
11	1
12	1
13	1
14	1
15	1
16	1
17	1
18	1
19	1
20	1
21	1
22	1
23	1
24	1
25	1
26	1
27	1
28	1
29	1
30	1
31	1
32	1

Figure 2.3: Screenshot of Class-Membership SAS® Dataset for Example 1



```
*-----------------------------------------------------
* Provide specifics regarding your desired analysis;
*-----------------------------------------------------
%let NumClasses         =1;
%let Method             =QN;
%let D_HasCorr_YN       =N;
%let D_VarDiffByClass_YN=N;
%let R_VarDiffByClass_YN=N;
%let PieMethod          =UNSTRUCTURED;
%let CalcParmStdErr     =Y;
%let   SEMethod         =GRD;
%let CalcRandomEffects  =Y;
```

(a) Linear Mixed Model



```
*-----------------------------------------------------
* Provide specifics regarding your desired analysis;
*-----------------------------------------------------
%let NumClasses         =2;
%let Method             =QN;
%let D_HasCorr_YN       =N;
%let D_VarDiffByClass_YN=Y;
%let R_VarDiffByClass_YN=N;
%let PieMethod          =STRUCTURED;
%let CalcParmStdErr     =Y;
%let   SEMethod         =GRD;
%let CalcRandomEffects  =Y;
```

(b) Same Mixture LCLMM

Figure 2.4: Options for Running Select Models - Example 1

LMM assumes that the random effects are distributed normally, the predictions of the random effects need not be normally distributed. Each of the models has a similar distribution of random effects and all have a greater percentage of area in the center and heavier tails than a normal distribution.

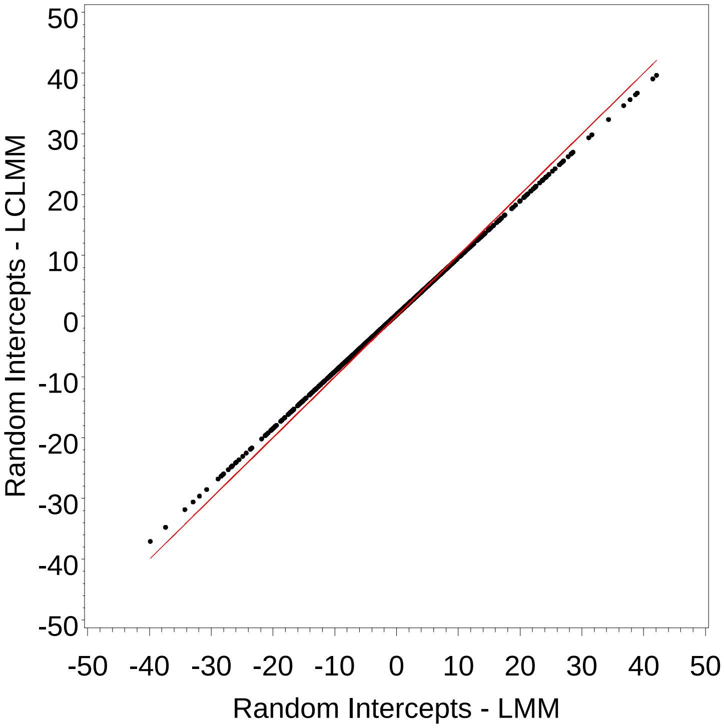
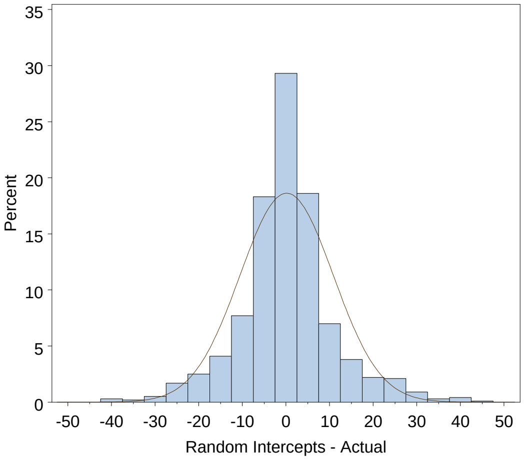
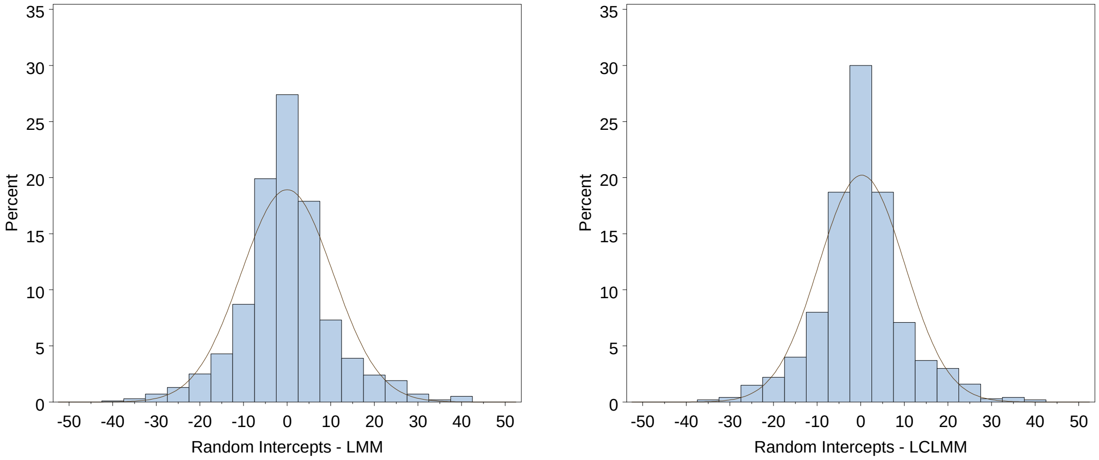


Figure 2.5: Comparison of Random Intercepts - LCLMM vs. Linear Mixed Model

Given the above, where do the models differ? The answer relates mainly to standard errors. The standard errors of the random effects for the LMM are determined based on the assumption that the random effects are normally distributed. In this example, the random intercepts from the LMM had a variance of 3.86. As noted in Section 1.4.5, the standard errors of the random effects for the LCLMM reflect information about each subject's structured class probabilities as well as the relative fit of the underlying LMMs. For the LCLMM, the standard errors of the random effects for each subject are graphed in Figure 2.7 versus the bayesian subject probabilities of being in each latent class (defined in the vector **c** in Section 1.4.1). A horizontal line is included at 3.86 to indicate the random effect standard error computed by the LMM. Note that since the standard error calculation for the LCLMM incorporates information as to which of the underlying distributions is most appropriate, each subject's



(a) Actual



(b) Linear Mixed Model

(c) LCLMM

Figure 2.6: Histograms of Actual and Fitted Random Effects

random intercept could potentially have a different standard error. In Figure 2.7, subjects at the left side (with low probability of being in the higher-standard-error class) are those that are in the very center of the random intercept distribution, while those at the right side are in the tails. Since there is a much greater percentage of random intercepts clumped in the middle of the mixture distribution, the standard errors of the random intercepts for these subjects would be expected to be smaller than under the LMM. However, since the tails in the mixture distribution are more spread out than under the LMM, the standard errors of the random intercepts out in the tails have a higher standard error than under the LMM.

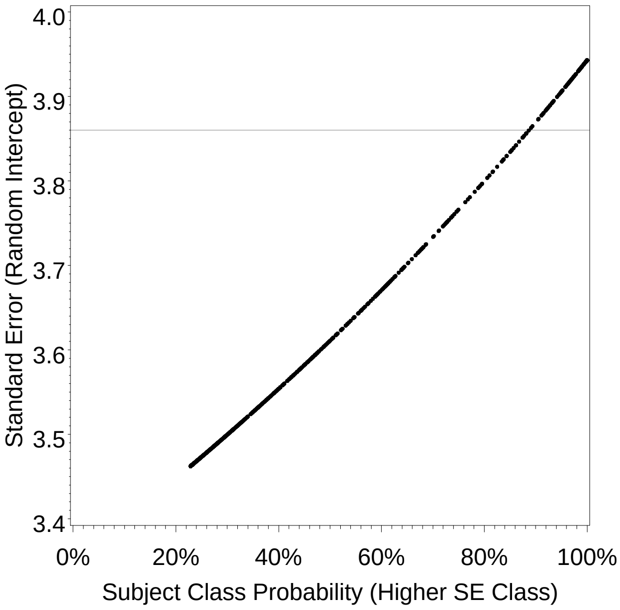


Figure 2.7: Plot of Random Effect Standard Errors from the LCLMM

2.4.2 Relative-Fit Model

LMMs are useful for modeling data when the data being analyzed is made up of a single population or when the subpopulations are known. However, in many cases, the data being analyzed may consist of several subpopulations which are not known a priori. The LCLMM can be very useful in this case. As an example, consider a model in which there are 300 subjects, 50 from each of six groups. Each subject is measured for some trait at ten consecutive time periods, ranging from -4.5 to 4.5. A possible LMM to describe this is as follows:

$$y_{it} = \beta_1 + \beta_2 \times \text{time } + \beta_3 \times \text{time}^2 + \beta_4 \times \text{time}^3 + \gamma_{1i} + \gamma_{2i} \times \text{time } + \epsilon_{it}$$

where:

y_{it} is the observed value of the trait for subject i at time t

β_1 is the fitted intercept

β_2 is the fitted slope for time

β_3 is the fitted quadratic term for time

β_4 is the fitted cubic term for time

γ_{1i} is the random intercept for subject i

γ_{2i} is the random slope for time for subject i

ϵ_{it} is the error term associated with subject i at time t

$\gamma_{1i} \sim \text{ Normal} \left(0, \sigma^2_{int}\right)$

$\gamma_{2i} \sim \text{ Normal} \left(0, \sigma^2_{slope}\right)$

$\epsilon_{it} \sim \text{ Normal} \left(0, \sigma^2_{error}\right)$

$\rho = \text{ corr } (\gamma_{1i}, \gamma_{2i})$

In words, the fitted mean trajectory for all subjects is a cubic curve with respect to time. Each individual is fit with a subject-specific intercept and slope, which are assumed to be distributed about the mean value according to Normal $\left(0, \sigma^2_{int}\right)$ and Normal $\left(0, \sigma^2_{slope}\right)$. Further, the random intercepts and slopes are correlated with correlation coefficient ρ. The error terms are distributed according to Normal $\left(0, \sigma^2_{error}\right)$. The LCLMM adds an additional component in that the population is divided into subpopulations, each with its own cubic curve. In this example, class membership probabilities are estimated based on the relative fit of a subject's data under each of the possible underlying models (see Section 2.2.2).

The following variance and correlation values were used to illustrate the LCLMM. Data for six latent classes were generated based on the fixed-effects parameters shown in Table 2.2.

$$
\begin{aligned}
\sigma^2_{int} &= 40 \\
\sigma^2_{slope} &= 15 \\
\sigma^2_{error} &= 125 \\
\rho &= 0.25
\end{aligned}
$$

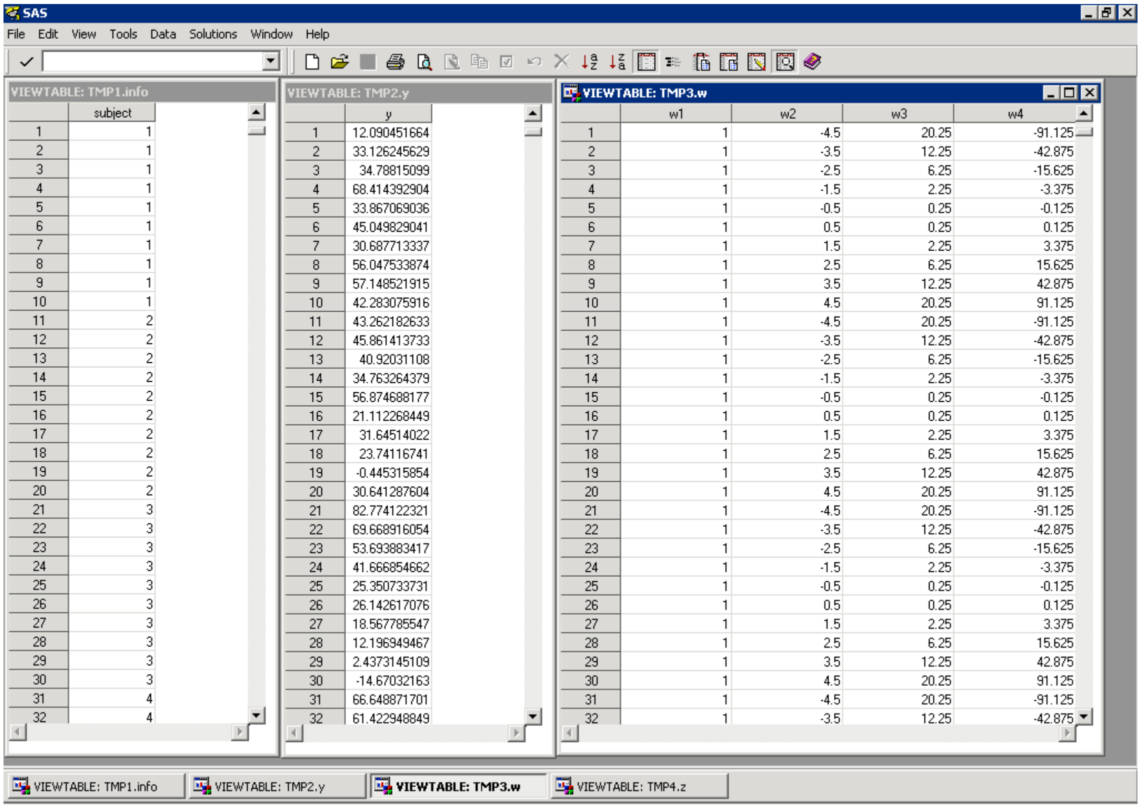
Table 2.2: Relative-Fit Model - Fixed Effects Parameters Used to Generate Data

Parameter	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6
β_1	40	190	125	80	170	130
β_2	-2	6	20	18	18	0
β_3	0	0	0	2.5	-2	0
β_4	0	0	0	0	0	1

In order to run the model through the SAS® macro, the user would specify the files in the format presented in Figures 2.8 and 2.9. Note that it is not necessary to specify D_Corr, D_Structure, and R_Structure since all records have the same residual error variance and the random effects are fit with an unstructured **D** matrix. However, these files are also included in order to illustrate how they would be specified (see Figure 2.10). Also, since class membership is determined based on the relative fit of each underlying LMM, no class-membership file is required for this example. Similarly, since all fixed effects are being allowed to differ by latent class, the file 'x' does not need to be specified.

This model was fit using the usual LMM (1 class) as well as the LCLMM using 2-8 latent classes. Figure 2.11 displays the options that need to be set to run the LMM as well as the LCLMM. The parameter estimates for the 6-class LCLMM are shown in Table 2.3. Note that the LCLMM estimates are very close to the values used to generate the data.

While it is encouraging to see that the model is able to reproduce the parameter values used to generate the data, it is interesting to examine the models fit with other numbers of latent classes. Figures 2.12 and 2.13 present the fitted models for each number of latent classes. Note that a fitted line for each latent class in included as a thick dotted line, while each individual's actual values are plotted using a narrow line. Subjects that have their highest class probability at least twice that of



VIEWTABLE: TMP1.info

	subject
1	1
2	1
3	1
4	1
5	1
6	1
7	1
8	1
9	1
10	1
11	2
12	2
13	2
14	2
15	2
16	2
17	2
18	2
19	2
20	2
21	3
22	3
23	3
24	3
25	3
26	3
27	3
28	3
29	3
30	3
31	4
32	4

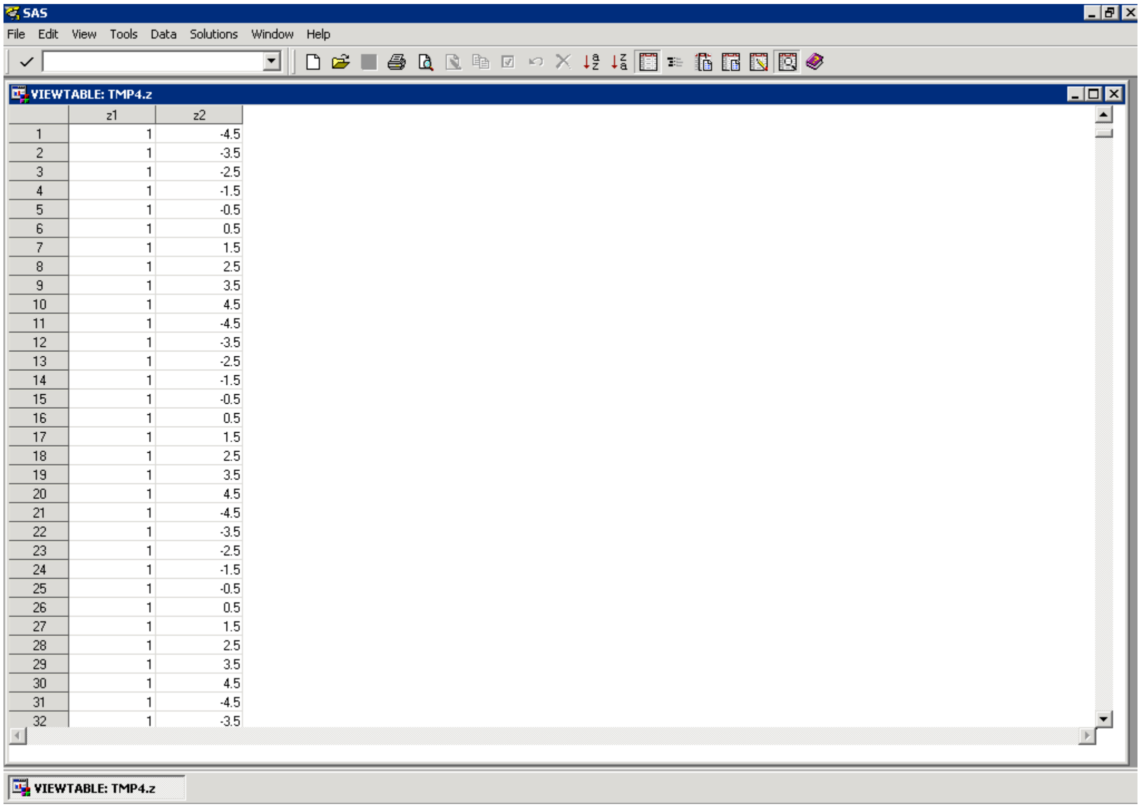
VIEWTABLE: TMP2.y

	y
1	12.090451664
2	33.126245629
3	34.78815099
4	68.414392904
5	33.867069036
6	45.049829041
7	30.687713337
8	56.047533874
9	57.148521915
10	42.283075916
11	43.262182633
12	45.861413733
13	40.92031108
14	34.763264379
15	56.874688177
16	21.112268449
17	31.64514022
18	23.74116741
19	-0.445315854
20	30.641287604
21	82.774122321
22	69.668916054
23	53.693883417
24	41.666854662
25	25.350733731
26	26.142617076
27	18.567785547
28	12.196949467
29	2.4373145109
30	-14.67032163
31	66.648871701
32	61.422948849

VIEWTABLE: TMP3.w

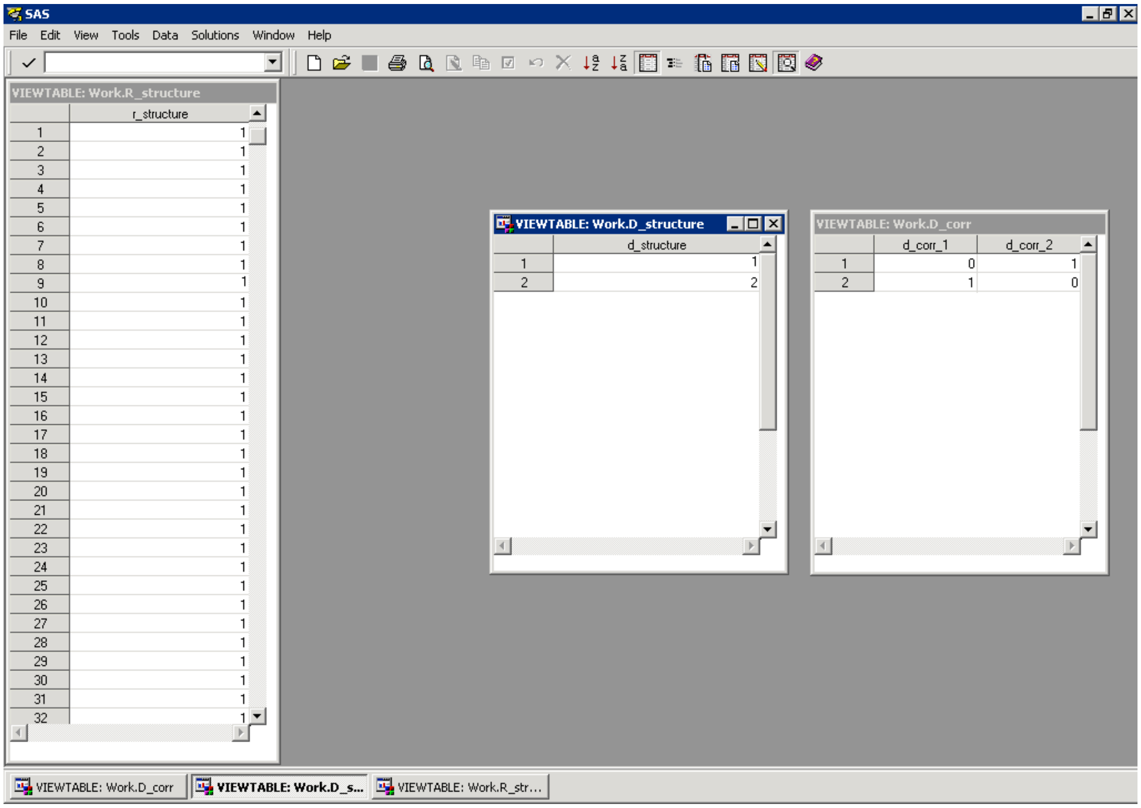
	w1	w2	w3	w4
1	1	-4.5	20.25	-91.125
2	1	-3.5	12.25	-42.875
3	1	-2.5	6.25	-15.625
4	1	-1.5	2.25	-3.375
5	1	-0.5	0.25	-0.125
6	1	0.5	0.25	0.125
7	1	1.5	2.25	3.375
8	1	2.5	6.25	15.625
9	1	3.5	12.25	42.875
10	1	4.5	20.25	91.125
11	1	-4.5	20.25	-91.125
12	1	-3.5	12.25	-42.875
13	1	-2.5	6.25	-15.625
14	1	-1.5	2.25	-3.375
15	1	-0.5	0.25	-0.125
16	1	0.5	0.25	0.125
17	1	1.5	2.25	3.375
18	1	2.5	6.25	15.625
19	1	3.5	12.25	42.875
20	1	4.5	20.25	91.125
21	1	-4.5	20.25	-91.125
22	1	-3.5	12.25	-42.875
23	1	-2.5	6.25	-15.625
24	1	-1.5	2.25	-3.375
25	1	-0.5	0.25	-0.125
26	1	0.5	0.25	0.125
27	1	1.5	2.25	3.375
28	1	2.5	6.25	15.625
29	1	3.5	12.25	42.875
30	1	4.5	20.25	91.125
31	1	-4.5	20.25	-91.125
32	1	-3.5	12.25	-42.875

Figure 2.8: Screenshot of SAS® Datasets Needed to Run Example 2



	z1	z2
1	1	-4.5
2	1	-3.5
3	1	-2.5
4	1	-1.5
5	1	-0.5
6	1	0.5
7	1	1.5
8	1	2.5
9	1	3.5
10	1	4.5
11	1	-4.5
12	1	-3.5
13	1	-2.5
14	1	-1.5
15	1	-0.5
16	1	0.5
17	1	1.5
18	1	2.5
19	1	3.5
20	1	4.5
21	1	-4.5
22	1	-3.5
23	1	-2.5
24	1	-1.5
25	1	-0.5
26	1	0.5
27	1	1.5
28	1	2.5
29	1	3.5
30	1	4.5
31	1	-4.5
32	1	-3.5

Figure 2.9: Screenshot of SAS® Datasets Needed to Run Example 2 (Continued)



VIEWTABLE: Work.R_structure

	r_structure
1	1
2	1
3	1
4	1
5	1
6	1
7	1
8	1
9	1
10	1
11	1
12	1
13	1
14	1
15	1
16	1
17	1
18	1
19	1
20	1
21	1
22	1
23	1
24	1
25	1
26	1
27	1
28	1
29	1
30	1
31	1
32	1

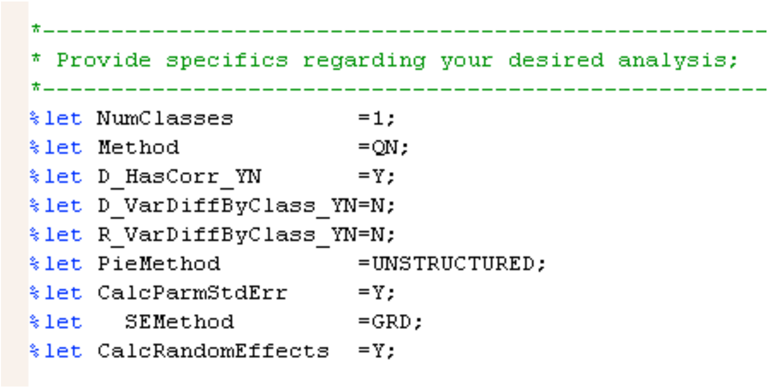
VIEWTABLE: Work.D_structure

	d_structure
1	1
2	2

VIEWTABLE: Work.D_corr

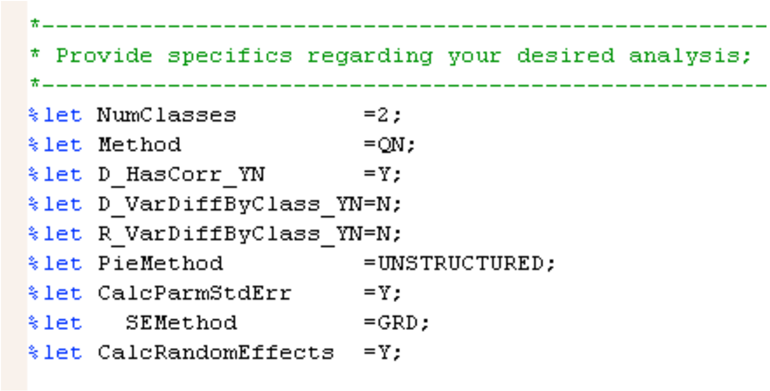
	d_corr_1	d_corr_2
1	0	1
2	1	0

Figure 2.10: Screenshot of SAS® Datasets Needed to Run Example 2 (Additional Files)



```
*----------------------------------------------------
* Provide specifics regarding your desired analysis;
*----------------------------------------------------
%let NumClasses          =1;
%let Method              =QN;
%let D_HasCorr_YN        =Y;
%let D_VarDiffByClass_YN=N;
%let R_VarDiffByClass_YN=N;
%let PieMethod           =UNSTRUCTURED;
%let CalcParmStdErr      =Y;
%let   SEMethod          =GRD;
%let CalcRandomEffects   =Y;
```

(a) Linear Mixed Model



```
*----------------------------------------------------
* Provide specifics regarding your desired analysis;
*----------------------------------------------------
%let NumClasses          =2;
%let Method              =QN;
%let D_HasCorr_YN        =Y;
%let D_VarDiffByClass_YN=N;
%let R_VarDiffByClass_YN=N;
%let PieMethod           =UNSTRUCTURED;
%let CalcParmStdErr      =Y;
%let   SEMethod          =GRD;
%let CalcRandomEffects   =Y;
```

(b) LCLMM (2 Classes)

Figure 2.11: Options for Running Select Models - Example 2

Table 2.3: Relative-Fit 6-Class LCLMM - Actual vs. Fitted

			LCLMM - 6 Classes	
Class	Parameter	Actual Value	Estimate	SE
1	β_1	40.00	39.6	1.15
	β_2	-2.00	-1.59	0.70
	β_3	0.00	-0.03	0.07
	β_4	0.00	-0.04	0.03
2	β_1	190.00	188.78	1.15
	β_2	6.00	6.35	0.70
	β_3	0.00	0.04	0.07
	β_4	0.00	0.02	0.03
3	β_1	125.00	123.27	1.15
	β_2	20.00	20.82	0.70
	β_3	0.00	0.02	0.07
	β_4	0.00	-0.02	0.03
4	β_1	80.00	78.49	1.15
	β_2	18.00	17.76	0.70
	β_3	2.50	2.56	0.07
	β_4	0.00	-0.03	0.03
5	β_1	170.00	169.10	1.15
	β_2	18.00	17.66	0.70
	β_3	-2.00	-2.00	0.07
	β_4	0.00	-0.03	0.03
6	β_1	130.00	131.07	1.15
	β_2	0.00	0.30	0.70
	β_3	0.00	0.05	0.07
	β_4	1.00	1.00	0.03
	σ^2_{int}	40.00	38.71	4.15
	σ^2_{slope}	15.00	14.99	1.34
	σ^2_{error}	125.00	119.71	3.45
	ρ	0.25	0.1648	0.0676

their second-highest class probability are plotted in the color of their selected latent class. From the model fits, several models appear that they could be justified in addition to the 6-class model. The 4-class and 5-class models appear to offer reasonable model fits. It is interesting to note that the 7-class and 8-class models add classes that are almost immediately adjacent to classes that existed in the 6-class model - this is one indication that the 7- and 8-class models may not be the most appropriate.

One way to evaluate model fit is to compare various likelihood measures for the models. Several likelihood measures are provided as part of the macro's output, including AIC, BIC, and ICL. These measures are described in more detail in Section 2.8. For this example, these measures are presented in Table 2.4. Note that the number of estimated parameters for each model, v, is also presented in this table. Each of these measures chooses either seven or eight classes, with the exception of entropy, which chooses either two or six classes.

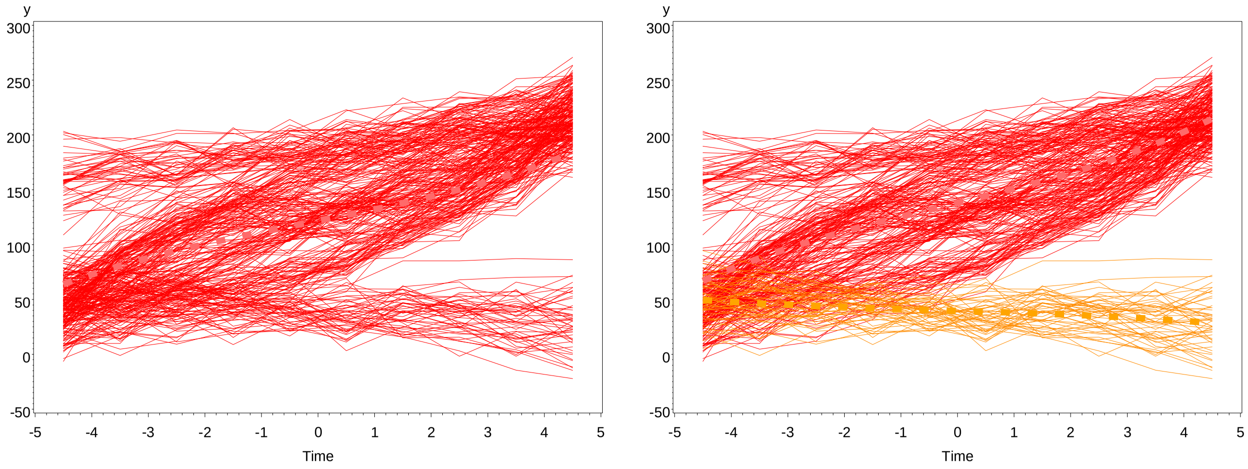
Table 2.4: Relative-Fit Model - Likelihood Measures

	Likelihood Measures							
# of Classes	ObsLik	AIC	CAIC	BIC	E	C	ICL	v
LMM (1)	-13854.8	27725.7	27737.5	27755.3	n/a	27709.7	27755.3	8
2	-13543.0	27110.0	27127.7	27154.4	**0.0**	27085.9	27154.4	12
3	-13118.9	26269.9	26293.5	26329.1	7.0	26251.9	26343.1	16
4	-12761.7	25563.4	25592.9	25637.5	1.1	25525.6	25639.7	20
5	-12344.8	24737.6	24773.1	24826.5	0.3	24690.3	24827.2	24
6	-12012.6	24081.2	24122.5	24184.9	**0.1**	24025.4	24185.1	28
7	-11989.8	24043.5	24090.8	24162.0	6.8	23993.2	**24175.7**	32
8	**-11973.8**	**24019.7**	**24072.8**	**24153.0**	16.7	**23981.1**	24186.4	36

Another way to evaluate the models is to compare a practical measure, such as the mean square error of the residuals. This measure is presented in Table 2.5 for both the total residuals and within-subject residuals. The MSE, as well as the residual measures for subject i, are defined as follows:

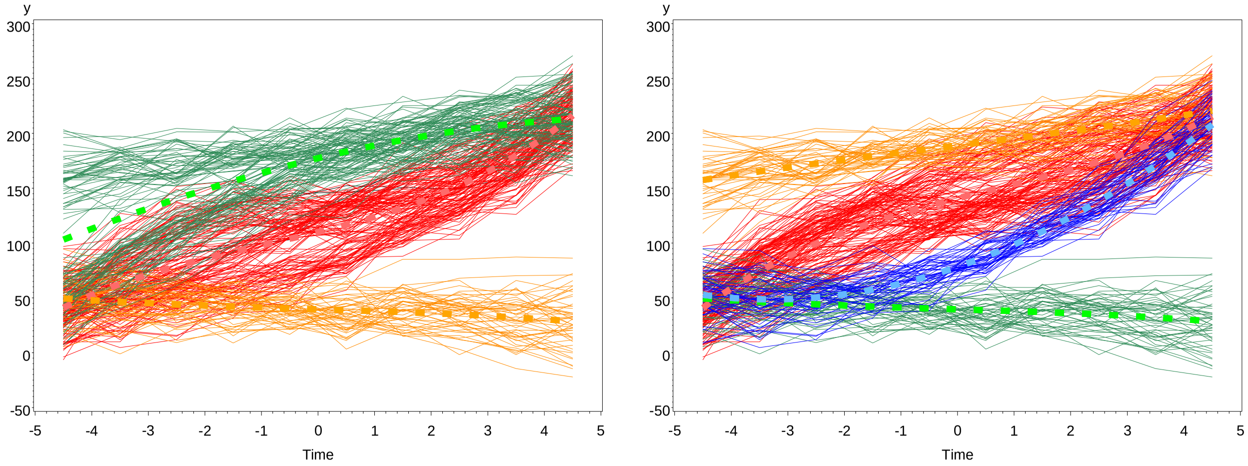
$$\mathbf{r}_{i,total} = \mathbf{y}_i - \sum_{k=1}^{K} \pi_{ik} \left(\mathbf{X}_i \hat{\boldsymbol{\beta}} + \mathbf{W}_i \hat{\boldsymbol{\lambda}}_k \right) \tag{2.12}$$

$$\mathbf{r}_{i,within} = \mathbf{y}_i - \left[\sum_{k=1}^{K} \pi_{ik} \left(\mathbf{X}_i \hat{\boldsymbol{\beta}} + \mathbf{W}_i \hat{\boldsymbol{\lambda}}_k \right) \right] - \mathbf{Z}_i \tilde{\mathbf{b}}_i \tag{2.13}$$



(a) Linear Mixed Model

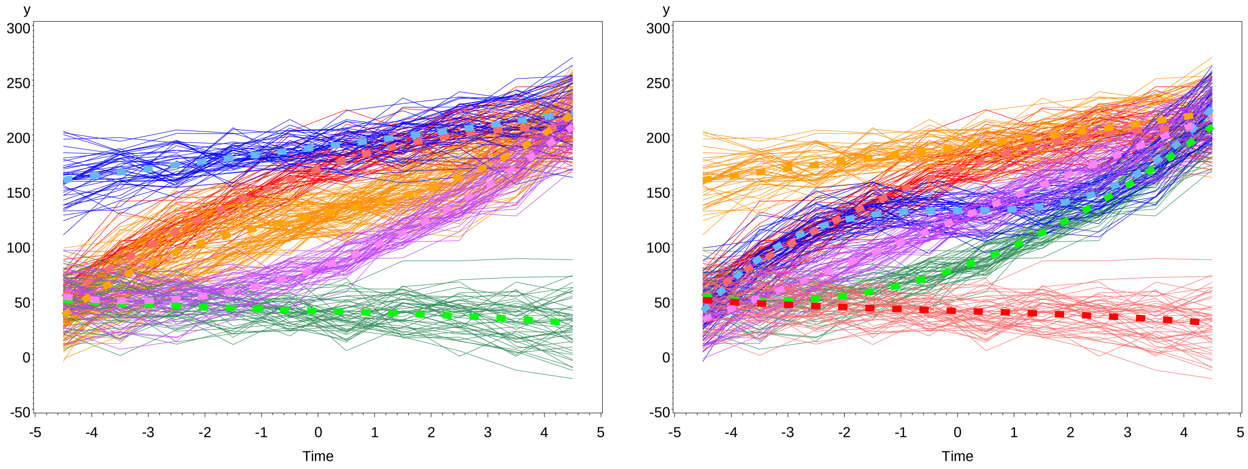
(b) LCLMM - 2 Classes



(c) LCLMM - 3 Classes

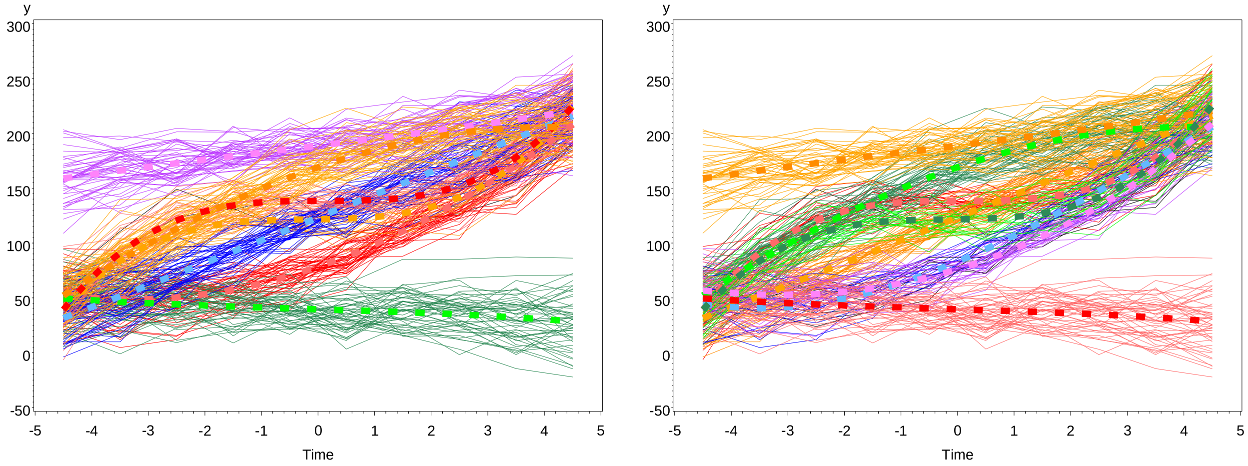
(d) LCLMM - 4 Classes

Figure 2.12: Relative-Fit 6-Class LCLMM - Fitted Models



(a) LCLMM - 5 Classes

(b) LCLMM - 6 Classes



(c) LCLMM - 7 Classes

(d) LCLMM - 8 Classes

Figure 2.13: Relative-Fit 6-Class LCLMM - Fitted Models - Continued

$$\text{MSE} \quad = \quad \frac{\sum_{i=1}^{n} \mathbf{r}_i'\mathbf{r}_i}{\text{Total \# of Observations}} \tag{2.14}$$

For this example, these measures decrease quickly until the 6-class model and then continue to decrease by a very small amount for the 7- and 8-class models. Choosing a parsimonious model would lead one to choose the 6-class model, or potentially the 4- or 5-class models, which also fit reasonably good.

Table 2.5: Relative-Fit Model - Residual Measures

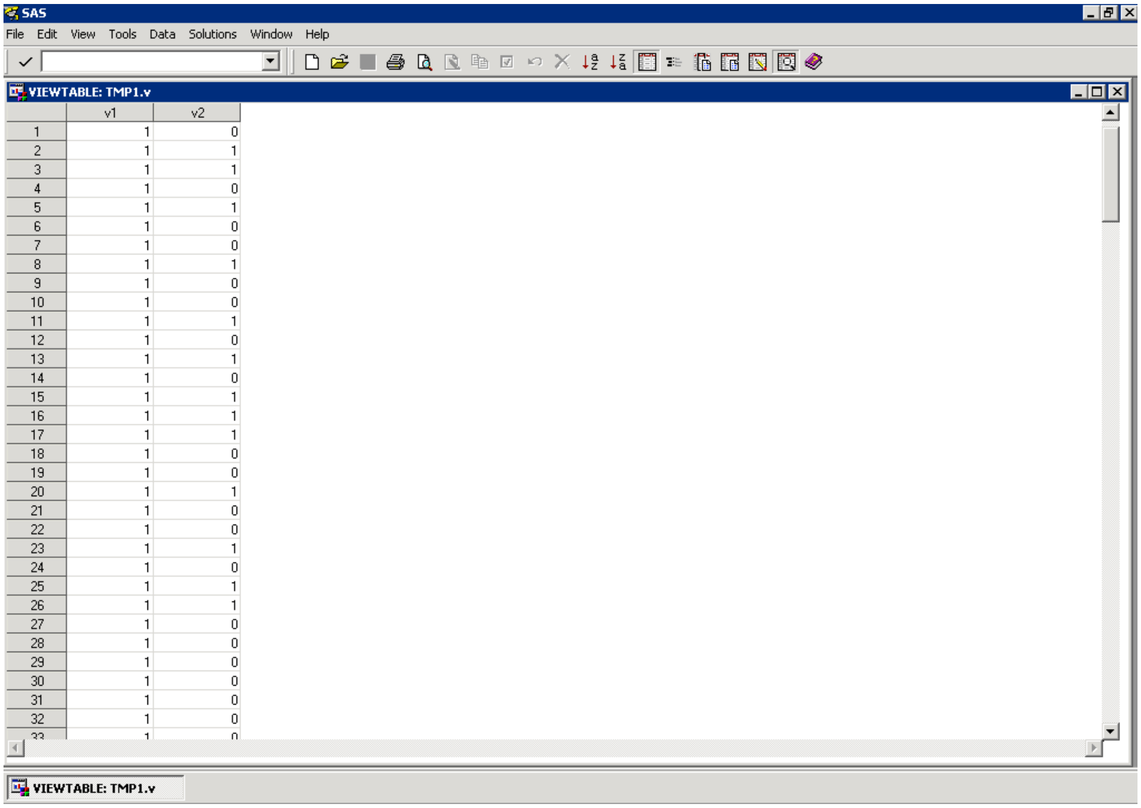
# of Classes	MSE (Total)	MSE (Within-Subject)
LMM (1)	3064.88	234.51
2	1325.38	235.64
3	676.17	185.65
4	428.16	161.50
5	333.90	127.75
6	281.56	99.65
7	275.88	98.33
8	**272.24**	**96.39**

2.4.3 Model with Known Class-Membership Factors

The first example discussed the situation where the user either is modeling a single population or knows the subpopulations and simply wants to generalize the distributional assumptions of the usual LMM. The second example is applicable when the user is unsure which variables may act to determine the underlying latent classes and is looking for the best way of simultaneously dividing the population into groups and fitting a LMM to each group. However, in some cases, the user may have interest in particular factors as they relate to class membership. For example, what is the best way to model the longitudinal data given that a particular set of factors can act to divide the population into groups? Or, does a particular risk factor act to drive an individual toward one class or another? The LCLMM allows the user to specify a logistic regression model for class membership in which the user can enter factors as appropriate.

As an example, consider two of the latent classes from the dataset from the previous example. Classes 4 and 5, the light red and green thick dotted lines in Figure 2.13b, do not overlap except for at the very beginning and end of the plots. This is intentional since the focus of this example is to examine the effect of a variable on class membership via the logistic class membership model. For illustration, $\frac{1}{3}$ of the subjects in Class 4 have a particular characteristic, while $\frac{2}{3}$ of the subjects in Class 5 have that characteristic. The goal here is to fit the best set of LMMs, while simultaneously examining how the trait impacts class membership. Note that while only 50 subjects were fit for each class in the previous example, 100 were fit for each class in this example.

The only change to the model from Section 2.4.2 is the addition of a class-membership file, a portion of which is shown in Figure 2.14. Note that this file has an intercept (v1) and an indicator variable associated with the trait on interest (v2). Roughly $\frac{1}{3}$ of the subjects in Class 4 have a 1 for this variable, while roughly $\frac{2}{3}$ of the subjects in Class 5 have a 1 for this variable.



	v1	v2
1	1	0
2	1	1
3	1	1
4	1	0
5	1	1
6	1	0
7	1	0
8	1	1
9	1	0
10	1	0
11	1	1
12	1	0
13	1	1
14	1	0
15	1	1
16	1	1
17	1	1
18	1	0
19	1	0
20	1	1
21	1	0
22	1	0
23	1	1
24	1	0
25	1	1
26	1	1
27	1	0
28	1	0
29	1	0
30	1	0
31	1	0
32	1	0

Figure 2.14: Screenshot of Class-Membership File Needed to Run Example 3

This model was fit using the LCLMM with two classes to determine whether the model would be able to identify the classes and whether it could determine that having the trait of interest would result in an individual having roughly twice the probability of being classified in Class 5. The parameter estimates are shown in Table 2.6.

Table 2.6: Class Membership Model - Actual vs. Fitted

			LCLMM - 2 Classes	
Class	Parameter	Actual Value	Estimate	SE
4	β_1	80.00	79.19	0.79
	β_2	18.00	18.38	0.51
	β_3	2.50	2.51	0.05
	β_4	0.00	-0.01	0.02
5	β_1	170.00	168.37	0.79
	β_2	18.00	18.30	0.51
	β_3	-2.00	-1.96	0.05
	β_4	0.00	-0.02	0.02
	σ^2_{int}	40.00	35.66	4.76
	σ^2_{slope}	15.00	15.95	1.74
	σ^2_{error}	125.00	117.75	4.16
	ρ	0.25	0.2043	0.0822
	$\boldsymbol{\alpha}_2$–intercept	n/a	-0.5965	0.2020
	$\boldsymbol{\alpha}_2$–trait	n/a	1.2897	0.2987

Note that these estimates are acceptably close to the actual values used to generate the data. Plugging the $\boldsymbol{\alpha}_2$ variables into Equation 2.4, and recalling that $\boldsymbol{\alpha}_1 = \mathbf{0}$, the fitted probability of being in Class 4 if an individual had the specified trait is 33.3%, while the fitted probability of being in Class 5 if an individual had the specified trait is 66.7%. This closely matches the probabilities one would expect based on the creation of the data. The important thing to note here is that since the classes are separated with very little overlap, each subject is fit with nearly 100% probability of being in the appropriate latent class - therefore, the logistic regression portion of the LCLMM effectively reduces to a simple logistic regression model where subjects are yes/no for a particular class. However, the logistic regression portion of the latent class model does NOT require that subjects be 100% in a particular class. A similar class-membership result would have occurred if there was greater overlap between classes but subjects with the trait tended to have twice the probability of being fit in the Class 5 model as in the Class 4 model. In other words, in this example, two out of every three individuals who had the trait were in Class 5 and therefore the logistic model found that there was twice the probability of being in Class 5 if you have the trait. However, if all individuals with the trait had a 66% / 33% mixture and all individuals without the trait had a 33% / 66% mixture, a similar result would have been observed.

2.5 Simulation Study - Stability - Comparison of Quasi-Newton and Newton-Raphson

In order to establish that the proposed algorithm is stable and results in the same final estimates for one estimation method versus another, a simulation study was conducted. The fitted parameters for triglycerides and HDL cholesterol for the Optimal and At-Risk classes from the 4-class class-specific variance model in the ARIC application of Chapter 1 were used to generate data. Recall that this application fit class membership based on the relative fit of each underlying LMM. This approach for fitting class membership was also used in all simulation studies.

Simulations were run to explore the stability of the methods with respect to the following choices:

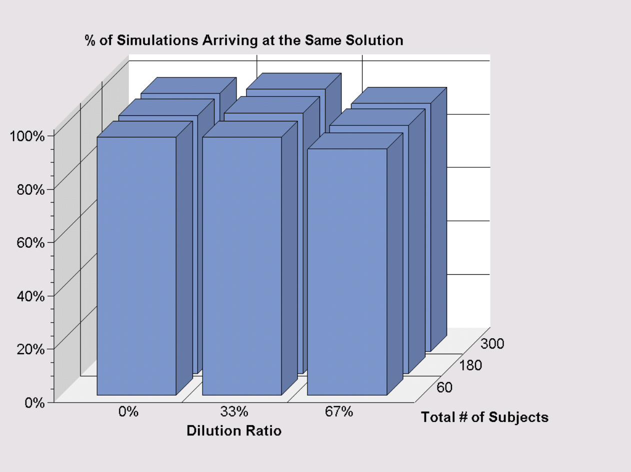
- 3 different size models were run - 60 subjects, 180 subjects, or 300 subjects - this allowed an examination as to whether the stability of the model is dependent on sample size.
- The number of subjects in each of the two classes was split in three different possible ways - 50/50 between the two classes, 75/25, or 90/10 - this allowed an examinations as to whether

the model is as stable when identifying a smaller percentage of outlier subjects as it is when identifying two similarly sized subpopulations.

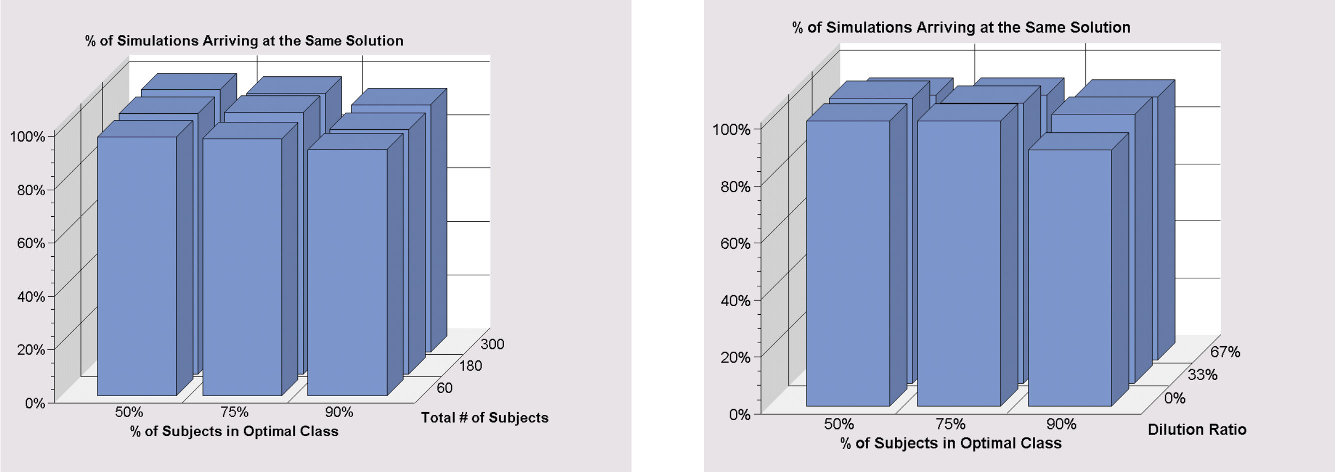
- The groups were separated in varying degrees - 0% dilution (uses parameter estimates for each class based on the ARIC application), 33% dilution, and 67% dilution. For the dilution models, the parameters used to generate the data were a weighted average of the original model and the average of the parameters from the two classes. This was examined to see whether the algorithm is stable when the groups are not as easily identifiable.

For each of the 27 combinations above (3 x 3 x 3), the LCLMM was run 100 times and was set to look for two latent classes. For each run, both the QN and NRA methods were run from the same starting point (a total of 5,400 simulations). The goal of these runs was to determine whether the quasi-Newton and Newton-Raphson methods would arrive at the same final solution if initiated from the same starting values. For purposes of the simulations, the quasi-Newton and Newton-Raphson runs were assumed to arrive at the same solution if the resulting likelihoods were within 0.01% of one another. The results are shown in Figure 2.15. Note that while there are slight drops in stability when 90% of the subjects are in the Optimal class and when the dilution percentage is 67%, these are very small. In general the model is able to find the same solution, regardless of whether Newton-Raphson or quasi-Newton are run. Further, the total number of subjects does not appear to affect the stability of the algorithm. The actual percentages associated with the barcharts are shown in Tables 2.7 to 2.9. It is interesting to note that of the 117 simulations where the Newton-Raphson and quasi-Newton methods arrived at different solutions, the quasi-Newton result had a better likelihood in 88 of the cases, while the Newton-Raphson method had the better result in only 29 of the runs.

In order to test that the stability of the model is not dependent on the size of the model, separation of the data, and percentage of subjects in one class, a logistic regression model was fit. The indicator for the 2,700 simulations as to whether the models found the same solution was modeled in a saturated model including each of the 3-level choices (dilution, size, percent in one class) as well as all 2-way interactions and the 3-way interaction. A backward-selection method was then used to pare down the model to only the necessary components. The final model is shown in Table 2.10 and the odds ratios are presented in Table 2.11. Consistent with the observation above, sample size was removed from the model as not being related to stability. From the odds ratios, it is apparent that only tests of the extremes are found to be statistically significant - tests of the dilution ratio 33% vs 0% and % Optimal 50% vs 75% are not found to be significant. The interaction of the dilution ratio and the



(a) Dilution by Number of Subjects



(b) % in Optimal Class by Number of Subjects

(c) Dilution by % in Optimal Class

Figure 2.15: Stability of Model Fit - Same Starting Values

Table 2.7: Stability of Model Fit - Same Starting Values - Dilution Ratio by Total # of Subjects

Dilution Ratio	Total # of Subjects	% of Simulations Finding Same Solution
0%	60	96.67%
	180	96.67%
	300	96.67%
33%	60	96.64%
	180	97.67%
	300	98.33%
67%	60	92.28%
	180	93.00%
	300	93.00%

Note: The SE for the percentage ranged from 0.74% for 33%/300 to 1.5% for 67%/60.

Table 2.8: Stability of Model Fit - Same Starting Values - Total # of Subjects by % in Optimal Class

Total # of Subjects	% in Optimal Class	% of Simulations Finding Same Solution
60	50%	96.99%
	75%	96.32%
	90%	92.28%
180	50%	97.67%
	75%	98.00%
	90%	91.67%
300	50%	98.33%
	75%	97.00%
	90%	92.67%

Note: The SE for the percentage ranged from 0.74% for 300/50% to 1.6% for 180/90%.

Table 2.9: Stability of Model Fit - Same Starting Values - Dilution Ratio by % in Optimal Class

Dilution Ratio	% in Optimal Class	% of Simulations Finding Same Solution
0%	50%	100.00%
	75%	100.00%
	90%	90.00%
33%	50%	100.00%
	75%	98.33%
	90%	94.30%
67%	50%	92.98%
	75%	92.98%
	90%	92.33%

Note: The SE for the percentage ranged from 0% for several categories to 1.7% for 0%/90%.

% in the Optimal class was also found not to be statistically significant. It can be concluded from the final model that the methods are slightly less stable for datasets which have very little separation between classes and for models which have a relatively small number of subjects in a particular latent class. It would be recommended, therefore, to run several runs from different starting values in these situations and choose the result with the highest likelihood. However, it must be noted that for all of the combinations of dilution/size/%Optimal, the Newton-Raphson and quasi-Newton models found the same solution in 90+% of simulations.

Table 2.10: Stability of Model Fit - Same Starting Values - Logistic Regression Model

Parameter	Level	DF	Estimate	Standard Error	Wald Chi-Square	P-Value
Intercept		1	3.3499	0.1163	828.9934	<0.0001
Dilution Ratio	33%	1	0.4888	0.1633	8.9642	0.0028
Dilution Ratio	67%	1	-0.6593	0.1291	26.0938	<0.0001
% in Optimal	75%	1	0.2778	0.1579	3.0957	0.0785
% in Optimal	90%	1	-0.7764	0.1300	35.6526	<0.0001

Table 2.11: Stability of Model Fit - Same Starting Values - Odds Ratios

Parameter	Levels	Point Estimate	95% Wald Confidence Limits
Dilution Ratio	33% vs 0%	1.375	0.785 - 2.409
Dilution Ratio	67% vs 0%	0.436	0.279 - 0.682
% in Optimal	75% vs 50%	0.802	0.447 - 1.439
% in Optimal	90% vs 50%	0.279	0.170 - 0.460

2.6 Simulation Study - Stability - Repeated Runs From Different Starting Values

The previous simulation study showed that, for most scenarios, the quasi-Newton and Newton-Raphson algorithms consistently find the same final solution when started from the same starting values. The question that follows is, do the methods find the same solution, regardless of the starting values used? As before, the fitted parameters for triglycerides and HDL cholesterol for the Optimal and At-Risk classes from the 4-class class-specific variance model in the ARIC application of Chapter 1 were used to generate data. Class membership, one again, was modeled based on the relative fit of each underlying LMM. The same 27 combinations of choices (3x3x3) described in the previous section were used for this set of simulations. For each of these combinations, the LCLMM was run ten times (five using quasi-Newton and five using Newton-Raphson), with each of the ten runs initiated at different starting values. The LCLMM was again set to look for two latent classes. An ideal result would be that all ten of the runs from different starting values result in the same final solution. The number out of these ten runs which resulted in the solution with the best likelihood was recorded. Similarly, the number for each method was also recorded (max 5). This was repeated with 20 unique datasets for each of the 27 combinations and the mean number of runs which achieved the best likelihood were recorded. For purposes of summarizing the simulations, a run is assumed to arrive at the best solution if the likelihood is within 0.01% of the run with the highest likelihood. As in the first simulation study, this results in a total of 5,400 runs - 27 combinations x 20 different datasets x 10 runs from different starting values. The results are shown in Figure 2.16. Note that, once again, while there are slight drops in the stability of the solutions when 90% of the subjects are in the Optimal class and when the dilution percentage is 67%, these are very small. In general the model is able to repeatedly find the 'best' solution, even when initiated from different starting values. The actual means associated with the barcharts are shown in Tables 2.12 to 2.14.

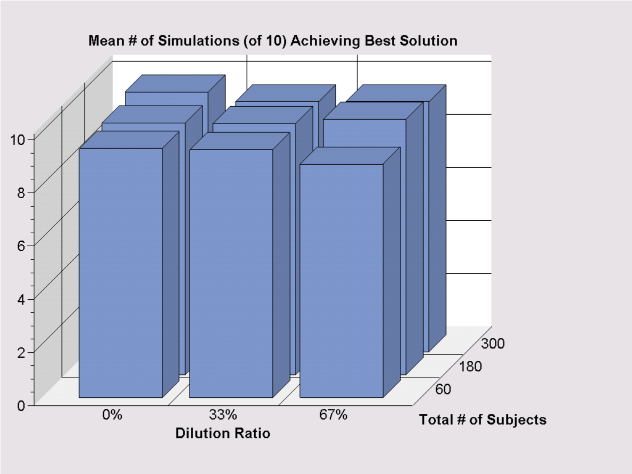
Table 2.12: Stability of Model Fit - Different Starting Values - Dilution Ratio by Total # of Subjects

		Mean Number of Simulations Achieving Best Solution		
Dilution Ratio	Total # of Subjects	Quasi-Newton (of 5)	Newton-Raphson (of 5)	Total (of 10)
0%	60	4.80	4.58	9.38
	180	4.70	4.78	9.48
	300	4.90	4.88	9.78
33%	60	4.66	4.68	9.34
	180	4.75	4.70	9.45
	300	4.80	4.62	9.42
67%	60	4.47	4.32	8.80
	180	4.92	4.70	9.62
	300	4.85	4.57	9.42

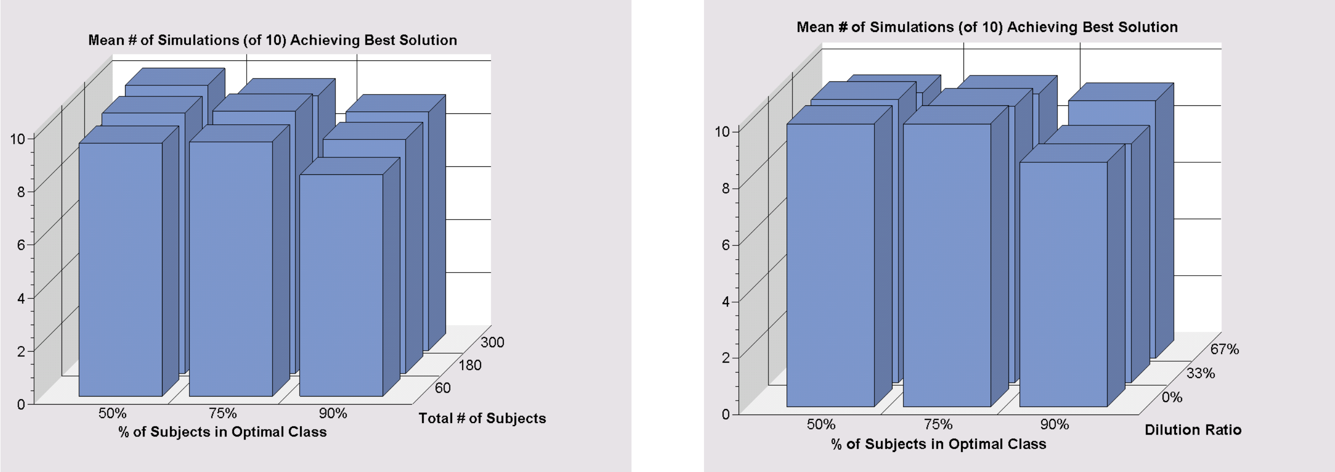
Note: The SE for the Total column ranged from 0.16 for 0%/300 subjects to 0.34 for 67%/60 subjects.

It is interesting to note that as you look at Tables 2.12 to 2.14 and compare the mean number of simulations that achieve the 'best' likelihood between quasi-Newton and Newton-Raphson, the quasi-Newton mean is almost always higher than the Newton-Raphson mean. This is consistent with a result from the previous section, where the quasi-Newton method tended to find the better solution when the methods chose different results.

In order to test that the stability of the model is not dependent on the size of the model, separation of the data, percentage of subjects in one class, and estimation method, a logistic regression model was fit. The focus of this model is to examine which combinations of these factors lead to situations in which less than five runs (the maximum) achieve the 'best' likelihood. For each of the 27 combinations x 20 datasets x 2 methods, an indicator as to whether all five of the runs achieved the 'best' value was defined. This indicator was modeled in a saturated logistic regression model including each of the 3-level choices (dilution, size, percent in one class), estimation method, as well as all 2-way, 3-way, and



(a) Dilution by Number of Subjects



(b) % in Optimal Class by Number of Subjects

(c) Dilution by % in Optimal Class

Figure 2.16: Stability of Model Fit - Different Starting Values

Table 2.13: Stability of Model Fit - Different Starting Values - Total # of Subjects by % in Optimal Class

		Mean Number of Simulations Achieving Best Solution		
Total # of Subjects	% in Optimal Class	Quasi-Newton (of 5)	Newton-Raphson (of 5)	Total (of 10)
60	50%	4.80	4.76	9.56
	75%	4.85	4.76	9.61
	90%	4.30	4.07	8.37
180	50%	4.93	4.88	9.82
	75%	5.00	4.90	9.90
	90%	4.43	4.40	8.83
300	50%	5.00	5.00	10.00
	75%	4.95	4.66	9.62
	90%	4.60	4.40	9.00

Note: The SE for the Total column ranged from 0.00 for 300/50% to 0.36 for 60/90%.

Table 2.14: Stability of Model Fit - Different Starting Values - Dilution Ratio by % in Optimal Class

		Mean Number of Simulations Achieving Best Solution		
Dilution Ratio	% in Optimal Class	Quasi-Newton (of 5)	Newton-Raphson (of 5)	Total (of 10)
0%	50%	5.00	5.00	10.00
	75%	5.00	5.00	10.00
	90%	4.40	4.25	8.65
33%	50%	5.00	5.00	10.00
	75%	4.90	4.87	9.77
	90%	4.32	4.13	8.45
67%	50%	4.73	4.65	9.38
	75%	4.90	4.46	9.36
	90%	4.62	4.48	9.10

Note: The SE for the Total column ranged from 0.00 for several categories to 0.38 for 33/90%.

the 4-way interactions. A backward-selection method was then used to pare down the model to only the necessary components. The final model is shown in Table 2.15 and the odds ratios are presented in Table 2.16. It is interesting to note that dilution ratio and estimation method were removed from the model as not being related to stability. From the odds ratios, it is apparent that only tests of the extremes for the other parameters are found to be statistically significant - tests of the number of subjects 300 vs 180 and % Optimal 50% vs 75% are not found to be significant. The interaction of the total number of subjects and the % in the Optimal class was also found not to be statistically significant. It can be concluded from the final model that the methods are slightly less stable for datasets which have a relatively small number of subjects in a particular latent class and for very small sample sizes. It would be recommended, therefore, to make repeated runs from different starting values, especially in situations where there are small counts or the potential for a class to have a small number of subjects. However, it must be noted once again that for all of the combinations of dilution, size, %Optimal, and estimation method, the mean number of runs which found the 'best' solution was at least 4 out of 5 and in many cases was 4.50-5.00.

Table 2.15: Stability of Model Fit - Different Starting Values - Logistic Regression Model

Parameter	Level	DF	Estimate	Standard Error	Wald Chi-Square	P-Value
Intercept		1	2.7813	0.1516	336.7364	<0.0001
% in Optimal	75%	1	0.4840	0.2162	5.0097	0.0252
% in Optimal	90%	1	-1.3178	0.1672	62.1090	<0.0001
# of Subjects	60	1	-0.5240	0.1516	11.9411	0.0005
# of Subjects	180	1	0.2620	0.1716	2.3324	0.1267

Table 2.16: Stability of Model Fit - Different Starting Values - Odds Ratios

Parameter	Levels	Point Estimate	95% Wald Confidence Limits
% in Optimal	75% vs 50%	0.705	0.308 - 1.612
% in Optimal	90% vs 50%	0.116	0.059 - 0.230
# of Subjects	60 vs 300	0.456	0.266 - 0.781
# of Subjects	180 vs 300	1.000	0.547 - 1.828

2.7 Simulation Study - Accuracy and Advantage of the LCLMM

The previous simulation studies have shown that the method is reasonably stable. The questions that arise next are

- Is the algorithm able to find the values used to generate the data?
- Does the model fit as good or better than the usual LMM?

To address these questions, an additional simulation study was run. The fitted parameters for triglycerides, LDL cholesterol, and HDL cholesterol from the Chapter 1 example for the 4-class class-specific variance model were used to generate data. For each run, 260 subjects were generated, 50 from each of the four latent classes and ten that were intermediate between the six pairs of two classes. These last 60 subjects were included to provide a more typical scenario in which there are some subjects for whom two models fit equally well. Recall from the form of the likelihood that if the underlying latent classes were separated such that they do not overlap (each subject is fit noticeably better in one particular class), then the LCLMM will effectively be maximizing the likelihood of a LMM with the classes known. The proposed set of simulations should provide a more practical example of how close the model could come to identifying the underlying classes and models in a real-world scenario. Class membership, once again, was modeled based on the relative fit of each underlying LMM. Having demonstrated in earlier simulations that the quasi-Newton and Newton-Raphson methods produce nearly identical parameter estimates, only the quasi-Newton method was run. The LCLMM was run for four classes, and the LMM was also run for comparison purposes. 250 unique datasets were run using each model for a total of 500 runs.

Of the 250 simulations, only six (2.4%) resulted in latent class membership which did not obviously match that of the generated data. This left 244 simulations left to examine. The mean fitted parameter estimates for the 244 runs of the LCLMM (4 class) model are presented alongside the actual parameter values used to generate the data in Tables 2.17-2.19. Note that even though 60 of the 260 subjects included in the model were created as being intermediate between the various classes, the fitted parameters are remarkably close to the actual parameter values. P-values which test that fitted parameter means match the actual values used to generate the data are included for each parameter based on a z-test. While several of the parameters have a significant p-value, the percentage difference between the actual values and fitted means is typically small.

In addition, out of 200 subjects from one of the four classes, note that an average of 174 subjects were classified into the group from which the subject was created. The interquartile range for this measure is extremely small at 171-179 and 80 percent of the simulations had between 168 and 182 subjects classified correctly. The minimum and maximum for the number of subjects classified correctly were 114 and 190, respectively.

Table 2.17: Simulation Study Results - Accuracy of the Parameter Estimates - Part 1 of 3

			Parameter Estimates			
Class	Parameter	Actual	Mean	StdErr	PValue	%Diff
AtRisk	LDL Intercept	130.069	130.367	0.297	0.3150	0.23
	LDL Linear Term	-2.391	-2.330	0.017	0.0003	2.55
	LDL Quadratic Term	-0.016	-0.020	0.001	0.0003	25.00
	HDL Intercept	34.615	35.195	0.055	<0.0001	1.68
	HDL Linear Term	0.053	0.038	0.002	<0.0001	28.30
	HDL Quadratic Term	0.004	0.004	0.000	0.1279	0.00
	Trig Intercept	5.167	5.150	0.004	<0.0001	0.33
	Trig Linear Term	0.002	0.003	0.000	<0.0001	50.00
	Trig Quadratic Term	-0.001	-0.001	0.000	0.0609	0.00
	Variance (LDL Random Int)	552.841	525.436	8.865	0.0020	4.96
	Variance (LDL Random Slope)	1.283	1.261	0.032	0.4972	1.71
	Variance (HDL Random Int)	13.408	14.246	0.320	0.0089	6.25
	Variance (Trig Random Int)	0.105	0.102	0.002	0.0338	2.86
	Variance (LDL Error)	403.483	416.861	3.453	0.0001	3.32
	Variance (HDL Error)	17.461	18.657	0.188	<0.0001	6.85
	Variance (Trig Error)	0.106	0.104	0.001	0.0018	1.89
AvgHDL-HighOther	LDL Intercept	161.536	160.895	0.458	0.1614	0.40
	LDL Linear Term	-2.023	-2.026	0.025	0.9213	0.15
	LDL Quadratic Term	-0.110	-0.104	0.002	0.0002	5.45
	HDL Intercept	48.543	49.049	0.105	<0.0001	1.04
	HDL Linear Term	-0.165	-0.179	0.004	0.0002	8.48
	HDL Quadratic Term	-0.008	-0.008	0.000	0.9198	0.00
	Trig Intercept	4.965	4.966	0.004	0.8718	0.02
	Trig Linear Term	0.012	0.012	0.000	0.0051	0.00
	Trig Quadratic Term	-0.001	-0.001	0.000	0.2980	0.00
	Variance (LDL Random Int)	778.488	670.036	14.775	<0.0001	13.93
	Variance (LDL Random Slope)	2.825	2.305	0.065	<0.0001	18.41
	Variance (HDL Random Int)	27.108	25.454	0.790	0.0362	6.10
	Variance (Trig Random Int)	0.093	0.086	0.002	0.0001	7.53
	Variance (LDL Error)	877.219	838.221	6.977	<0.0001	4.45
	Variance (HDL Error)	62.208	63.693	0.557	0.0076	2.39
	Variance (Trig Error)	0.124	0.121	0.001	0.0119	2.42

Table 2.18: Simulation Study Results - Accuracy of the Parameter Estimates - Part 2 of 3

			Parameter Estimates			
Class	Parameter	Actual	Mean	StdErr	PValue	%Diff
AvgHDL-LowOther	LDL Intercept	129.278	128.865	0.246	0.0931	0.32
	LDL Linear Term	-0.294	-0.343	0.015	0.0010	16.67
	LDL Quadratic Term	-0.028	-0.030	0.001	0.0211	7.14
	HDL Intercept	49.879	49.862	0.107	0.8757	0.03
	HDL Linear Term	-0.343	-0.326	0.003	<0.0001	4.96
	HDL Quadratic Term	0.003	0.003	0.000	0.9224	0.00
	Trig Intercept	4.604	4.619	0.004	<0.0001	0.33
	Trig Linear Term	0.020	0.019	0.000	<0.0001	5.00
	Trig Quadratic Term	-0.000	-0.000	0.000	0.0073	-
	Variance (LDL Random Int)	336.792	348.362	6.491	0.0746	3.44
	Variance (LDL Random Slope)	1.151	1.123	0.031	0.3759	2.43
	Variance (HDL Random Int)	44.312	37.554	0.811	<0.0001	15.25
	Variance (Trig Random Int)	0.071	0.072	0.002	0.5918	1.41
	Variance (LDL Error)	201.024	225.971	3.084	<0.0001	12.41
	Variance (HDL Error)	28.009	30.582	0.455	<0.0001	9.19
	Variance (Trig Error)	0.054	0.056	0.001	0.0042	3.70
Optimal	LDL Intercept	121.532	122.380	0.311	0.0065	0.70
	LDL Linear Term	0.330	0.220	0.015	<0.0001	33.33
	LDL Quadratic Term	-0.016	-0.018	0.001	0.0087	12.50
	HDL Intercept	74.254	72.995	0.155	<0.0001	1.70
	HDL Linear Term	-0.512	-0.490	0.005	<0.0001	4.30
	HDL Quadratic Term	-0.015	-0.015	0.000	0.7526	0.00
	Trig Intercept	4.504	4.519	0.004	<0.0001	0.33
	Trig Linear Term	0.018	0.017	0.000	0.0007	5.56
	Trig Quadratic Term	-0.000	-0.000	0.000	0.7205	-
	Variance (LDL Random Int)	654.825	646.843	10.016	0.4255	1.22
	Variance (LDL Random Slope)	1.016	1.121	0.031	0.0006	10.33
	Variance (HDL Random Int)	114.058	93.650	2.017	<0.0001	17.89
	Variance (Trig Random Int)	0.101	0.096	0.001	0.0018	4.95
	Variance (LDL Error)	436.181	440.892	3.634	0.1949	1.08
	Variance (HDL Error)	151.026	143.163	0.918	<0.0001	5.21
	Variance (Trig Error)	0.072	0.073	0.000	0.0183	1.39

Table 2.19: Simulation Study Results - Accuracy of the Parameter Estimates - Part 3 of 3

			Parameter Estimates			
Class	Parameter	Actual	Mean	StdErr	PValue	%Diff
Overall	Correlation (LDL Int/Slope)	-0.342	-0.347	0.007	0.4777	1.46
	Correlation (HDL/LDL Int)	0.058	0.062	0.009	0.6621	6.90
	Correlation (HDL Int/LDL Slope)	0.090	0.068	0.006	0.0007	24.44
	Correlation (Trig/LDL Int)	-0.017	-0.030	0.010	0.2157	76.47
	Correlation (Trig Int/LDL Slope)	-0.022	0.008	0.009	0.0008	136.36
	Correlation (Trig/HDL Int)	-0.269	-0.253	0.008	0.0487	5.95

Based on the above, it appears that the LCLMM was able to accurately recreate the data used to generate the simulations. Nearly 90% of subjects were classified appropriately on average, and the parameter estimates are remarkably close to the estimates used to create the data. This is despite the fact that an additional 60 subjects were included in the model which were intermediate between the classes in order to provide a more practical situation. The question that follows is 'does the LCLMM offer a better model fit than the LMM?'

In order to examine this question, the 250 unique datasets which were run using both the LCLMM and LMM were examined with respect to model fit, and selected summary information is presented in Table 2.20. The mean and mean square error of the within-subject residuals were summarized across all observations. In addition, in order to examine whether certain latent classes tended to show greater improvement than others, these measures were also summarized according to the actual latent classes used to generate the data. The 60 unknown subjects are included in this summary as the 'Unknown' class. For each measure, the means based on both models, the difference between those means, and the standard error of the differences are displayed, as well as the p-value based on a paired z-test of the hypothesis that the results from each model are equal.

Overall, the LCLMM and LMM have reasonably similar average MSEs. The LCLMM tends to have a slightly improved model fit, but not dramatically so. In the At-Risk and Optimal classes, the LCLMM has a somewhat smaller MSE for HDL. Turning attention to the means of the within-subject residuals, note that the overall mean for each lab parameter for the LCLMM is very close to zero. The means for the LMM are zero by definition. When the means are examined for each latent class, another advantage of the LCLMM is apparent. Namely, each subject-specific prediction

Table 2.20: Simulation Study Results - Comparison of Means and MSEs of the Within-Subject Residuals

Class	Statistic	Parameter	Mean (LCLMM)	Mean (LMM)	Difference	StdErr	PValue
Overall	MSE	HDL Cholesterol	54.1492	56.9008	-2.7517	0.0951	<0.0001
		LDL Cholesterol	349.1519	349.5864	-0.4345	0.7830	0.5789
		Triglycerides	0.0738	0.0781	-0.0043	0.0001	<0.0001
	Mean	HDL Cholesterol	-0.0628	0.0000	0.0628	0.0018	<0.0001
		LDL Cholesterol	-0.0557	0.0000	0.0557	0.0034	<0.0001
		Triglycerides	-0.0004	-0.0000	0.0004	0.0000	<0.0001
AtRisk	MSE	HDL Cholesterol	14.7230	24.0205	-9.2975	0.1293	<0.0001
		LDL Cholesterol	294.5671	295.2093	-0.6422	1.3614	0.6372
		Triglycerides	0.0877	0.0993	-0.0116	0.0003	<0.0001
	Mean	HDL Cholesterol	-0.2935	-0.7739	-0.4804	0.0087	<0.0001
		LDL Cholesterol	-0.1902	-1.3433	-1.1531	0.0238	<0.0001
		Triglycerides	0.0029	0.0165	-0.0136	0.0004	<0.0001
AvgHDL-HighOther	MSE	HDL Cholesterol	52.7573	51.0448	1.7125	0.1363	<0.0001
		LDL Cholesterol	640.8770	644.8648	-3.9878	2.8868	0.1672
		Triglycerides	0.1034	0.1024	0.0009	0.0002	<0.0001
	Mean	HDL Cholesterol	-0.1632	0.0100	0.1532	0.0152	<0.0001
		LDL Cholesterol	0.7133	2.0592	-1.3460	0.0333	<0.0001
		Triglycerides	0.0038	0.0081	-0.0042	0.0004	<0.0001
AvgHDL-LowOther	MSE	HDL Cholesterol	23.0832	24.5630	-1.4798	0.0852	<0.0001
		LDL Cholesterol	146.1023	158.3765	-12.2742	0.7670	<0.0001
		Triglycerides	0.0441	0.0502	-0.0060	0.0002	<0.0001
	Mean	HDL Cholesterol	-0.1566	-0.5476	-0.3911	0.0104	<0.0001
		LDL Cholesterol	-0.0340	0.0766	-0.0426	0.0235	0.0696
		Triglycerides	-0.0068	-0.0279	-0.0211	0.0004	<0.0001
Optimal	MSE	HDL Cholesterol	126.5151	131.1604	-4.6453	0.3062	<0.0001
		LDL Cholesterol	317.2034	312.3669	4.8365	1.3054	0.0002
		Triglycerides	0.0595	0.0630	-0.0036	0.0001	<0.0001
	Mean	HDL Cholesterol	0.6265	1.3244	-0.6978	0.0113	<0.0001
		LDL Cholesterol	-0.1728	-0.7917	-0.6188	0.0186	<0.0001
		Triglycerides	-0.0023	0.0036	-0.0014	0.0003	<0.0001
Unknown	MSE	HDL Cholesterol	53.7475	54.2464	-0.4989	0.1317	0.0002
		LDL Cholesterol	347.3665	339.1931	8.1734	0.8841	<0.0001
		Triglycerides	0.0742	0.0760	-0.0018	0.0001	<0.0001
	Mean	HDL Cholesterol	-0.2834	-0.0107	0.2727	0.0116	<0.0001
		LDL Cholesterol	-0.5047	-0.0007	0.5041	0.0176	<0.0001
		Triglycerides	0.0002	-0.0002	-0.0000	0.0002	0.9778

is based on that subject's most appropriate class or combination of classes. In other words, the totality of a subject's information provides valuable input as to which class or classes are most relevant and then predictions reflect the characteristics of those classes. This is in contrast to the LMM, where subject-specific predictions will be adjusted from a population trajectory. In most classes/parameters, the mean of the within-subject residuals is closer to 0 for the LCLMM than for the LMM. The exception appears to be for the 'Unknown' class, although the differences observed for this class are still reasonable. In conclusion, the LCLMM appears to fit slightly better than the LMM overall, with the At-Risk and Optimal classes showing the most improvement. In addition, the means of the residuals for each latent class tend to be closer to 0 under the LCLMM than under the LMM.

2.8 Simulation Study - Comparison of Likelihood Measures

2.8.1 Overview

As noted earlier, the LCLMM requires that the statistician prespecify the number of latent classes. How is this done? What information goes into determining how many classes to fit? These questions have been the subject of a great deal of research, of which just a brief review is provided here. Fraley and Raftery [2002] provided a review of several clustering methods. One of the methods, described as model-based agglomerative hierarchical clustering, operates by successively merging pairs of clusters that result in the greatest increase in the classification likelihood. Similarly, many clustering algorithms use a nearest-neighbor approach to determine which records are similar to others, gradually fine tuning the numbers of clusters. Fraley et al. [2003] proposed a model-based clustering algorithm which incrementally *adds* clusters as needed until the model fit is no longer improved. In this algorithm, a preliminary mixture model is fit with fewer classes than would be expected. Then the set of observations which are fit worst are reclassified as being in a new cluster, and the model is re-fit. The algorithm continues until adding a cluster results in a decrease to the BIC.

It is not immediately obvious why these methods would necessarily result in the 'correct' number of clusters, nor would they offer the most appropriate model fit once the clusters have been determined. Those methods that operate ad-hoc by grouping similar records may produce different results based on different arrangements of the data. And in cases where longitudinal data is present which may have been recorded at different time points, these methods do not appear to be able to fully account for the available information.

In situations where the statistician wishes to determine the most informative clustering of the data, and the underlying data would be best fit by a LMM (i.e. longitudinal, continuous data), the LCLMM seems to provide the most sound approach. In the final model, the records for each latent class would be appropriately modeled via a class-specific LMM. The question, however, still remains - how many classes should be fit? The answer rests in which criteria best account for the tradeoff between the added parameters needed to fit another class and the improvement in the resulting model fit.

2.8.2 Information Criteria

Biernacki and Govaert [1999] examined the performance of several information and classification criteria to determine which were more useful in terms of selecting the best model. Among these were the following measures. Note that v represents the number of parameters fit in the model, n represents the number of subjects, and E/EC represent two different entropy measures, one of which examines all class probabilities while the other examines only the highest class probability.

$$\begin{aligned}
\text{AIC} &= -2\,\log\,\text{L}\ +2\text{v} && \bullet\text{Akaike Information Criterion}\\
\text{AIC3} &= -2\,\log\,\text{L}\ +3\text{v} && \bullet\text{Modified AIC Criterion}\\
\text{BIC} &= -2\,\log\,\text{L}\ +\text{v}\,\log\,\text{n} && \bullet\text{Bayesian Information Criterion}\\
\text{C} &= -2\,\log\,\text{L}\ +2\text{E} && \bullet\text{Fuzzy Classification Likelihood}\\
\text{CLM} &= -2\,\log\,\text{L}\ +2\text{EC} && \bullet\text{Classification Likelihood}
\end{aligned}$$

where:

$$\text{E} = -\sum_{k=1}^{K}\sum_{i=1}^{n} \widetilde{c}_{ik}\,\log\,\widetilde{c}_{ik} \geq 0$$

$$\text{EC} = -\sum_{k=1}^{K}\sum_{i=1}^{n} z_{ik}\,\log\,\widetilde{c}_{ik} \geq 0$$

$$\widetilde{c}_{ik} = \frac{\pi_{ik} f\left(\mathbf{y}_i \mid c_{ik}=1\right)}{\sum_{j=1}^{K}\left\{\pi_{ij} f\left(\mathbf{y}_i \mid c_{ij}=1\right)\right\}}$$

$$z_{ik} \text{ is 1 if the } k^{th} \text{ class has the highest value of } \widetilde{c}_{ik},\ \text{ or 0 otherwise}$$

For consistency with the other measures, C and CLM have been multiplied by -2. Note that for repeated measures studies, the choice of n is not obvious. While the number of subjects is typically used, the extremes are the total number of observations and the total number of subjects. The authors found that AIC and BIC outperformed other criteria in a clustering context.

Biernacki et al [2000] proposed an additional criteria, referred to as the Integrated Classification Likelihood, which is defined as follows:

$$\mathrm{ICL} = -2\ \log\ \mathrm{L}\ \ + \mathrm{v}\ \log\ \mathrm{n} + 2\mathrm{E}$$

The authors found that ICL performs well in determining the relevant number of clusters.

And finally, the last criteria to be considered is the Consistent Akaike Information Criterion (CAIC) proposed by Bozdogan [1987]. This criteria provides a greater penalty than the usual AIC, and is defined as follows:

$$\mathrm{CAIC} = -2\ \log\ \mathrm{L}\ \ + \mathrm{v}\ (\log\ \mathrm{n}\ +\ 1)$$

Based on my work with the models, I propose a series of additional information criteria along the lines of BIC. The BIC penalizes the likelihood for each additional parameter fit in the models. However, the model that is the main focus of this research is a model in which class membership is determined by the relative likelihood of an individual's data under each of the underlying LMMs - therefore, no additional parameters are fit for class membership. If the latent classes are so obviously separated that the classes can be easily determined, then this seems appropriate. However, if class membership for an individual is not obvious, it seems as if a penalty needs to be assessed for that individual to account for fitting that individual's mixing proportion. If the individual is fit such that their class membership probability is split between two classes, then it seems the penalty should be the equivalent of fitting one additional parameter in the model. If the individual's probability is split between three classes, then the penalty should be two additional parameters, and so on. Since the point at which an individual can be attributed to one class with certainty or near-certainty is not obvious, four criteria were proposed. The first, BICMod23, looks at the highest probability for each subject and compares it to the other class probabilities - if the second-highest probability is more than half of the highest, then a single penalty parameter is charged. If the third-highest probability is more than half of the highest, then a second penalty parameter is charged and so on. The label '23' is meant to call attention to the fact that a subject with probabilities 2/3 and 1/3 would be at the border of having a penalty assessed. A second criteria, BICMod34, would have the border at 75 percent/25 percent. BICMod45 has its border at 80 percent/20 percent and BICMod910 has its border at 90 percent/10 percent.

2.8.3 Simulation Study

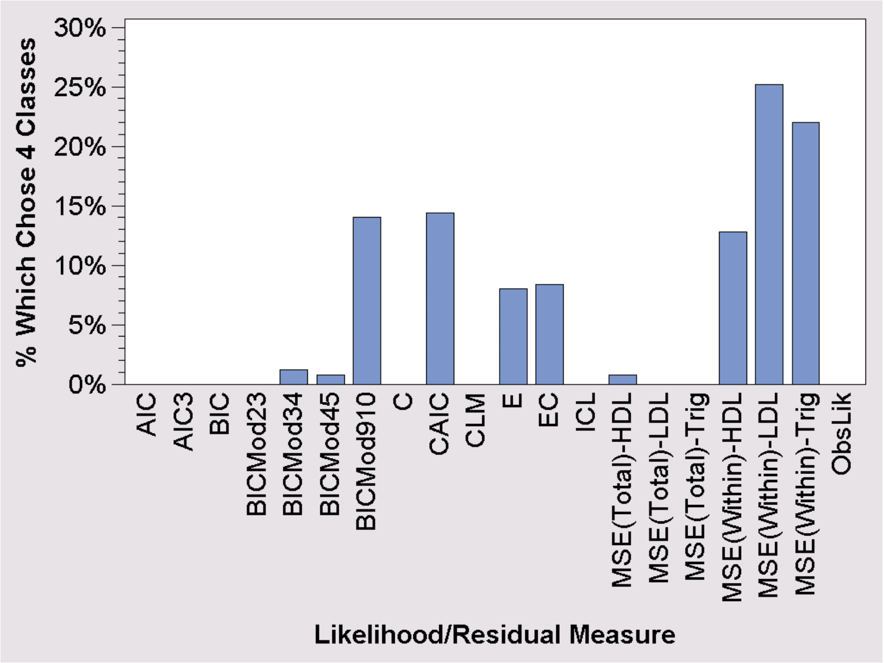
To investigate the usefulness of these information criteria for choosing the most appropriate LCLMM, a simulation study was run. The fitted parameters for triglycerides, LDL cholesterol, and HDL cholesterol from the Chapter 1 example for the 4-class class-specific variance model were used to generate data. For each run, 260 subjects were generated, 50 from each of the four latent classes and ten that were intermediate between the six pairs of two classes. These last 60 subjects were included to provide a more typical scenario in which there are some subjects for whom two models fit equally well. Having demonstrated that the quasi-Newton and Newton-Raphson methods are reasonably consistent, only the quasi-Newton method was run. The LMM and LCLMM for 2-8 classes were run. 250 unique datasets were run for a total of 2,000 runs. Class membership, once again, was modeled based on the relative fit of the underlying LMMs.

In this simulation study, the goal is to determine whether the existing and newly proposed information criteria can effectively determine the appropriate number of latent classes for the model being fit - in this case four classes. In addition to the information criteria, mean square error measures for both the within and total residuals were also examined. These residual measures were defined in Equations 2.12 and 2.13. Based on the eight runs for each unique dataset (for 1-8 classes), a 'best' model was selected based on each likelihood measure. The number of runs (out of 250) which selected each possible number of classes is shown in Table 2.21. Figure 2.17 shows charts of the percentage of runs for which each measure chose the correct number of classes (4) or were within one of this number.

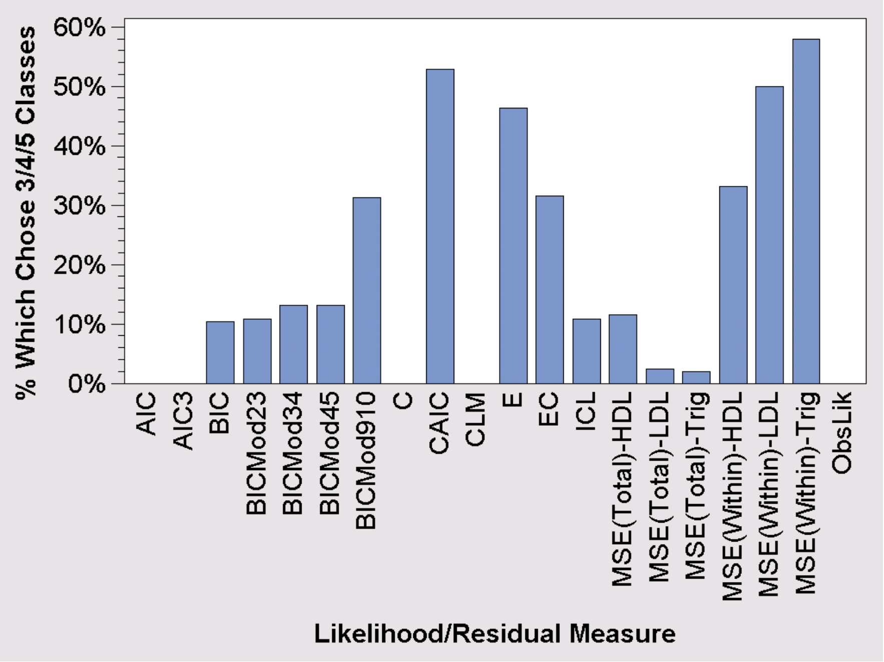
The traditional likelihood measures tended to choose more than four classes. The AIC/C/CLM measures chose the maximum number of classes (8) for nearly all of the 250 unique datasets. The BIC/ICL measures also tended to choose a higher number of classes. Many runs chose seven or eight classes, while a fair number of runs chose five or six classes. The measures which performed best were the entropy (E and EC), MSE (Within), and CAIC measures, with just under 10% choosing four classes for the entropy measures, about 15% choosing four classes for CAIC, and roughly 15-25% choosing four classes for the MSE measures. As in this example, since multiple parameters can be run through the model, it may be possible that some parameters may be fit fine with just two classes, while others are only fit well when additional classes are fit. From the counts for MSE (Within) in Table 2.21, it appears that HDL could be fit reasonably well by just two classes, while LDL appears to be fit best by either one or four classes and triglycerides by four or five classes. This provides a general indication that four or five classes may be appropriate.

Table 2.21: Simulation Study Results - Choosing the # of Latent Classes

	# of Runs Choosing This Number of Latent Classes							
Likelihood/Residual Measure	1	2	3	4	5	6	7	8
AIC	.	.	.	.	.	.	13	237
AIC3	.	.	.	.	.	.	24	226
CAIC	.	.	.	36	96	65	39	14
BIC	.	.	.	.	26	47	92	85
BIC-Mod23	.	.	.	.	27	34	79	110
BIC-Mod34	.	.	.	3	30	29	79	109
BIC-Mod45	.	.	.	2	31	33	70	114
BIC-Mod910	.	.	1	35	42	36	52	84
E	.	53	79	20	17	14	19	48
EC	.	12	35	21	23	18	45	96
C	.	.	.	.	.	.	3	247
CLM	.	.	.	.	.	.	3	247
ICL	.	.	.	.	27	40	93	90
MSE (Total Residuals) - HDL	.	.	2	2	25	35	67	119
MSE (Total Residuals) - LDL	.	.	.	.	6	16	52	176
MSE (Total Residuals) - Triglycerides	.	.	1	.	4	26	63	156
MSE (Within-Subject Residuals) - HDL	.	121	30	32	21	17	16	13
MSE (Within-Subject Residuals) - LDL	73	9	25	63	37	23	15	5
MSE (Within-Subject Residuals) - Triglycerides	.	6	30	55	60	33	42	24



(a) Chose Correct # of Classes



(b) Within 1 Class

Figure 2.17: Simulation Study Results - Choosing the # of Latent Classes

In conclusion, the usual likelihood-based measures tend to choose too many latent classes. The within-subject MSE tended to choose the correct number of latent classes more often than the other measures and should be considered for use until better measures can be proposed and tested.

2.9 Simulation Study - Model Speed

Finally, with the stability and accuracy of the algorithm established, and the advantage of the proposed methods apparent, which method should the user choose for various size models? While there are many variables which can affect runtime, a few are examined in the following simulation study. First, how is runtime impacted when additional subjects are added? Second, what is the effect of fitting additional latent classes? And third, how much time does it add to fit different variances for each latent class?

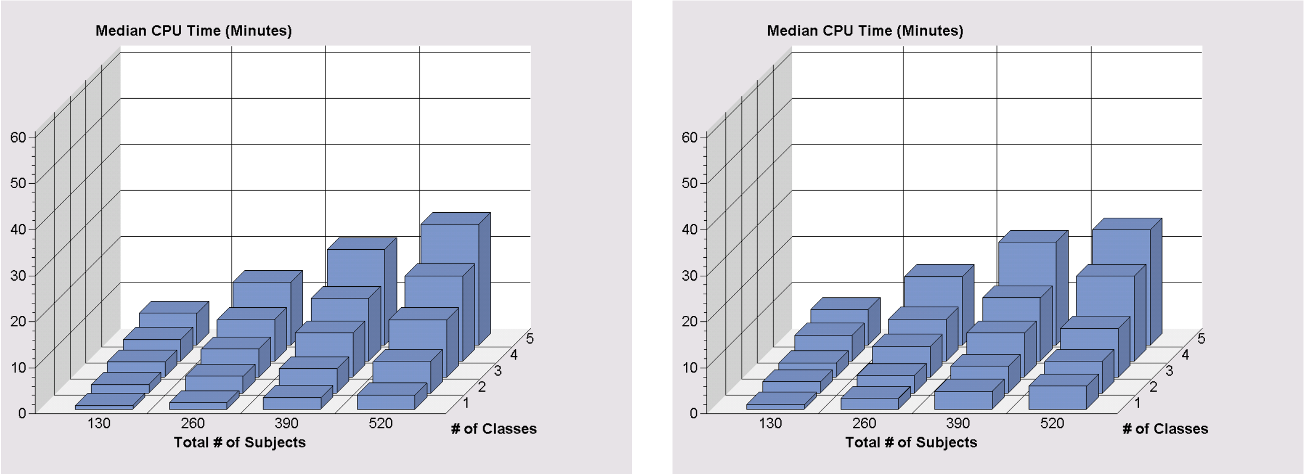
These questions were examined by running simulations based on subjects created from the same 4-class model that has been used in the previous simulation studies. Thirty simulations were run for each possible combination of the below choices, for a total of 2,400 runs.

- A total of 130 subjects, 260 subjects, 390, and 520 subjects were generated
- The LCLMM was run for 1-5 classes
- The model was run in two different ways - once assuming class-specific variances and once assuming variances equal across classes
- Both the QN and NRA methods methods were run

The median CPU runtimes (minutes) are presented in Table 2.22 and in Figure 2.18. The quasi-newton and Newton-Raphson methods have similar runtimes for the equal-across-classes variance models. However, for the class-specific variance approach, in which there can be a much larger number of parameters, the Newton-Raphson method tends to take a noticeably longer period of time to run. For example, for the 5-class class-specific variance model with 520 subjects, the Newton-Raphson method took a median time of 332.6 minutes to run versus only 57.6 minutes for the quasi-Newton method. It is also worth noting that the runtime for the quasi-Newton method does not appear to increase dramatically with increases in sample size or latent classes and has only a modest increase when fitting class-specific variances. These results lead to the conclusion that the quasi-Newton method should be used in general for these rather complex models.

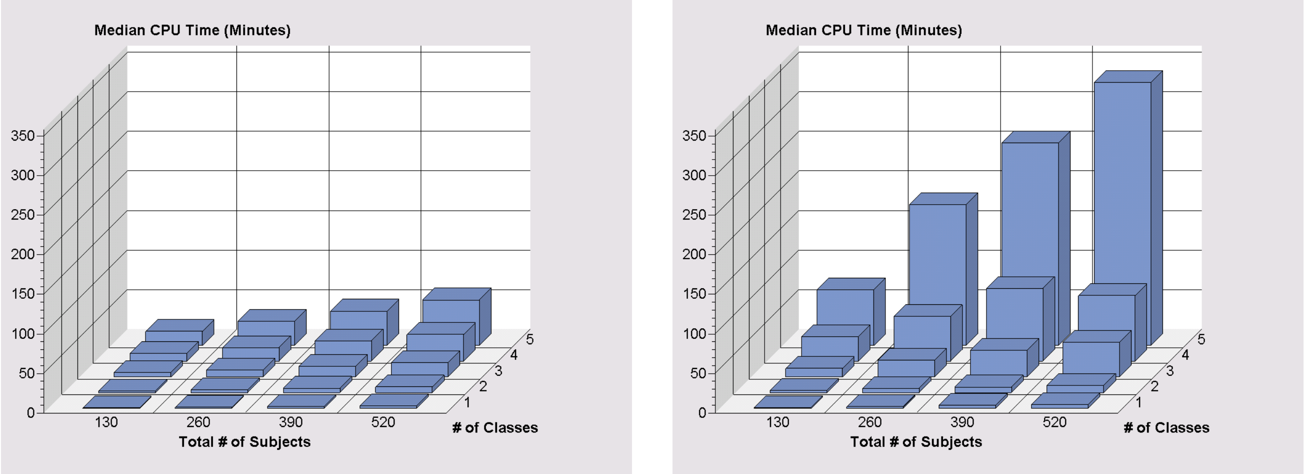
Table 2.22: Simulation Study Results - Median RunTimes (Minutes)

		Equal-Across-Classes		Class-Specific	
# of Subjects	# of Classes	QN	NRA	QN	NRA
130	1	0.8	1.1	0.8	1.1
130	2	1.9	2.5	2.1	3.1
130	3	3.3	3.1	5.9	10.8
130	4	4.7	5.6	10.1	30.9
130	5	6.9	7.8	18.0	70.7
260	1	1.4	2.3	1.4	2.3
260	2	3.7	3.8	3.8	5.5
260	3	6.1	6.8	9.1	21.4
260	4	9.0	9.2	17.2	56.4
260	5	13.7	14.9	30.9	178.1
390	1	2.5	3.9	2.5	3.9
390	2	5.3	5.9	5.7	7.2
390	3	9.6	9.6	13.7	33.7
390	4	13.7	13.8	25.6	91.9
390	5	20.8	22.4	43.5	256.0
520	1	3.1	5.0	3.1	5.0
520	2	6.9	6.9	7.4	9.5
520	3	12.5	10.6	18.5	43.6
520	4	18.5	18.5	34.0	83.0
520	5	26.3	25.1	57.6	332.6



(a) Equal-Across-Classes Variances - QN

(b) Equal-Across-Classes Variances - NRA



(c) Class-Specific Variances - QN

(d) Class-Specific Variances - NRA

Figure 2.18: Simulation Study Results - Median RunTimes (Minutes)

2.10 Conclusions and Further Research

This chapter has presented a new macro, which will allow users to fit mixtures of linear mixed models for a wide variety of problems. In addition to being able to easily generalize the distributional assumptions of the usual linear mixed model, users can explore complex multi-parameter datasets to determine underlying subpopulations and the appropriate linear mixed model for each. Class membership can further be explored by specifying variables to be used in a logistic model for class probabilities. These methods were examined in a series of simulation studies and found to be stable and to provide a real benefit in terms of providing a better model fit than the usual linear mixed model.

Future research should focus first and foremost on developing a dependable measure for selecting the appropriate number of latent classes. While the within-subject MSE appears to perform reasonably well, it would be advantageous to have a better method for making this selection. Ideally, such an approach would allow some sort of tuning parameter to allow the user to specify their sensitivity to the addition of latent classes. For example, in some cases, the user may wish to have a more broad look at class membership, while in others, small clusters of individuals may be of interest. An additional focus of research should examine a hybrid of the relative-fit and structured class membership models. Since there may be instances in which the class membership model fits a subject as being in one class while the subject may be better-fit in another, a model which fits each subject's class probabilities based on a combination of the two would offer an intermediate and perhaps more reliable result.

Chapter 3

Tutorial in Latent Class Linear Mixed Models for Clinical Researchers

3.1 Introduction

Chapter 1 presented an expanded methodology related to Latent Class Linear Mixed Models (LCLMM), including an example using lipid data from the ARIC study. Chapter 2 discussed a new SAS® macro which will allow statisticians and researchers to apply these models easily in practice. In the latter, examples were presented to illustrate how to specify models in the new macro. However, given the complexity of the models and the many questions that arise in specifying a model, it is important to provide statisticians and researchers alike with a more detailed look at the decisions that need to be made when fitting the LCLMM. This is accomplished in this chapter in three sections. The first section presents a brief review of the LCLMM methodology. The second section discusses the decisions that a statistician or researcher must make when fitting a LCLMM. The third section walks through three separate real-world modeling examples step-by-step to illustrate how the models can be used in practice. These should provide a reference as users explore the models more fully.

3.2 The Model

A detailed review of the LCLMM is provided in Chapter 1. For convenience, a brief review is presented here.

3.2.1 Defining the Latent Class Linear Mixed Model

The LCLMM is effectively a mixture model where the underlying mixture distributions are specified as linear mixed models (LMMs). Therefore, a brief introduction of the LMM will lead directly into the specification of the LCLMM. Note that Searle et al [1992] provides a more detailed presentation of the LMM. For subject i, we assume

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \mathbf{e}_i \tag{3.1}$$

where:

$\mathbf{y}_i$ is an $n_i \times 1$ vector of observations

$\mathbf{X}_i$ is an $n_i \times p_1$ design matrix for the fixed effects

$\boldsymbol{\beta}$ is a $p_1 \times 1$ unknown vector of fixed effects

$\mathbf{Z}_i$ is an $n_i \times q$ design matrix for the random effects

$\mathbf{b}_i$ is a $q \times 1$ unknown vector of random effects

$\mathbf{e}_i$ is an $n_i \times 1$ unknown vector of random error terms

Further, $\mathbf{e}_i$ and $\mathbf{b}_i$ are assumed to be mutually independent of one another and to have the following properties:

$$\begin{aligned}
\mathrm{E}\left(\mathbf{b}_i\right) &= \mathbf{0} \\
\mathrm{Var}\left(\mathbf{b}_i\right) &= \mathbf{D} \\
\mathrm{Cov}\left(\mathbf{b}_i, \mathbf{b}'_h\right) &= \mathbf{0} \text{ for } i \neq h \\
\mathrm{E}\left(\mathbf{e}_i\right) &= \mathbf{0} \\
\mathrm{Var}\left(\mathbf{e}_i\right) &= \mathbf{R}_i \\
\mathrm{Cov}\left(\mathbf{b}_i, \mathbf{e}_i\right) &= \mathbf{0}
\end{aligned} \tag{3.2}$$

And finally, the following distributional properties are assumed to apply to the random effects, $\mathbf{b}_i$, and residuals, $\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}$:

$$\begin{bmatrix} \mathbf{b}_i \\ \mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} \end{bmatrix} \sim N\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{D} & \mathbf{D}\mathbf{Z}_i' \\ \mathbf{Z}_i\mathbf{D} & \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i' + \mathbf{R}_i \end{bmatrix}\right) \tag{3.3}$$

The latent class model adds a further dimension in that each subject's data is modeled as a mixture of K LMMs, one for each of the K latent classes. Note that the value K is specified by the statistician a priori. Guidelines related to selecting K are discussed later in this chapter. In addition, a simulation study is presented in Section 2.8, in which the usefulness of various likelihood-based and residual-based criteria is examined with regard to making this selection. Following the notation in Lin et al [2002a], define for subject i:

$$c_{ik} = \begin{cases} 1 & \text{if subject } i \text{ is a member of class } k \\ 0 & \text{if subject } i \text{ is NOT a member of class } k \end{cases}$$

$$c_{i1}, \ldots, c_{iK} \sim \text{ Multinomial } (1, \pi_{i1}, \ldots, \pi_{iK})$$

The π_{ik}, the multinomial probabilities of being in each latent class, are modeled via a logit model as follows:

$$\pi_{ik} = P\left(c_{ik} = 1 \mid \mathbf{t}_i\right) = \frac{\exp\left(\mathbf{t}_i'\boldsymbol{\alpha}_k\right)}{\sum_{j=1}^{K} \exp\left(\mathbf{t}_i'\boldsymbol{\alpha}_j\right)} \tag{3.4}$$

where:

$\mathbf{t}_i$ is the design vector related to class membership for subject i

$\boldsymbol{\alpha}_k$ is an unknown vector of class-membership parameters for class k with $\boldsymbol{\alpha}_1 = \mathbf{0}$

Further, given that subject i is in class k, define the LMM for subject i as follows:

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{W}_i\boldsymbol{\lambda}_k + \mathbf{Z}_i\mathbf{b}_i + \mathbf{e}_i \tag{3.5}$$

where the following additional definitions are provided:

$\mathbf{W}_i$ is an $n_i \times p_2$ design matrix for the class-specific fixed effects

$\boldsymbol{\lambda}_k$ is a $p_2 \times 1$ unknown vector of class-specific fixed effects for class k

It is useful to note that the parameters in $\boldsymbol{\beta}$ will apply to all subjects through the values of the corresponding column in $\mathbf{X}_i$. The class-specific parameters, $\boldsymbol{\lambda}_k$, however, are different for each latent class.

The error terms and random effects are assumed to have the same properties as in Equation 3.2. Finally, similar to the LMM, the following distributional properties are assumed to apply to the random effects, $\mathbf{b}_i$, and residuals, $\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{W}_i\boldsymbol{\lambda}_k$:

$$\begin{bmatrix} \mathbf{b}_i \\ \mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{W}_i\boldsymbol{\lambda}_k \end{bmatrix} \sim N\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{D} & \mathbf{D}\mathbf{Z}_i' \\ \mathbf{Z}_i\mathbf{D} & \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i' + \mathbf{R}_i \end{bmatrix} \right) \tag{3.6}$$

It is assumed, without loss of generality, that the model is full-rank. This assumption requires that both of the following are true:

$$\begin{bmatrix} \mathbf{t}_1 & \mathbf{t}_2 & \cdots & \mathbf{t}_n \end{bmatrix} \text{ is full rank, and}$$

$$\left[\begin{array}{c|cccc|cccc} \mathbf{X}_1 & \mathbf{W}_1 & . & . & . & \mathbf{Z}_1 & . & . & . \\ \mathbf{X}_2 & . & \mathbf{W}_2 & . & . & . & \mathbf{Z}_2 & . & . \\ \vdots & . & . & \ddots & . & . & . & \ddots & . \\ \mathbf{X}_n & . & . & . & \mathbf{W}_n & . & . & . & \mathbf{Z}_n \end{array}\right] \text{ is full rank.}$$

For notational convenience, denote $\boldsymbol{\Theta}$ as the combined parameter vector comprised of $\boldsymbol{\alpha}_2 \ldots \boldsymbol{\alpha}_K$, $\boldsymbol{\beta}$, $\boldsymbol{\lambda}_1 \ldots \boldsymbol{\lambda}_k$, and $\boldsymbol{\theta}$, where $\boldsymbol{\theta}$ contains the unique variance components which determine $\mathbf{D}$ and $\mathbf{R}_i$.

It is useful to compare the observed-data likelihoods of the usual LMM with the LCLMM. These likelihoods are conditional on having observed the values $\mathbf{X}_i$, $\mathbf{W}_i$, $\mathbf{Z}_i$, and $\mathbf{t}_i$. In short,

$$\log L(\boldsymbol{\Theta})_{LMM} = \sum_{i=1}^{n} \log f(\mathbf{y}_i) \tag{3.7}$$

$$\log L(\boldsymbol{\Theta})_{LCLMM} = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \pi_{ik} f(\mathbf{y}_i \mid c_{ik} = 1) \tag{3.8}$$

with $f(\cdot)$ being the density defined by

$$\begin{aligned} (\mathbf{y}_i) &\sim N(\mathbf{X}_i\boldsymbol{\beta}, \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i' + \mathbf{R}_i) \\ (\mathbf{y}_i \mid c_{ik} = 1) &\sim N(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{W}_i\boldsymbol{\lambda}_k, \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i' + \mathbf{R}_i) \end{aligned}$$

Several points are of note. First, note the difference between Equations 3.7 and 3.8 - the likelihood for the LCLMM is a weighted average of K LMMs. Second, with K=1, and therefore $\pi_{i1} = 1$, the LCLMM reduces to the usual LMM. Third, if the groups are so well separated that each subject has one $\pi_{ik} = 1$ and the others equal to 0, then the likelihood reduces to that of a LMM with subjects assigned to these groups. Finally, note that the log likelihood for the LCLMM contains the log of the sum over latent classes, making the computations of the first and second derivatives of the log likelihood more complicated.

3.2.2 The Relative-Fit Class Membership Model

In the model described in the previous section, class membership was assumed to be determined based on an underlying logistic regression model. However, in many cases, the statistician may not know exactly which factors should be included in the model describing class membership, these factors may not be measurable, or the statistician may prefer to simply identify the best-fitting LCLMM with K classes. However, this model, in which each subject can have its own mixture, would result in many parameters related to class membership - specifically, for n subjects and K classes, the model would require $n \times (K-1)$ parameters. Therefore, a second approach was proposed in Chapter 1 which replaces the logistic regression model with a model in which class membership is determined by the relative fit of the underlying LMMs. This is much less costly in terms of computational time than fitting a separate set of class membership parameters for each subject, but offers a logical approach with a similar goal in mind.

Recall that the unconditional log likelihood of the observed data, assuming class-specific **D** and **R**, is as follows:

$$\log L(\mathbf{\Theta}) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \left[\pi_{ik} f\left(\mathbf{y}_i \mid c_{ik} = 1\right)\right] \tag{3.9}$$

with $f\left(\mathbf{y}_i \mid c_{ik} = 1\right)$ being the density defined by

$$\left(\mathbf{y}_i \mid c_{ik} = 1\right) \sim N(\mathbf{X}_i\boldsymbol{\beta} + \mathbf{W}_i\boldsymbol{\lambda}_k, \mathbf{Z}_i\mathbf{D}_k\mathbf{Z}_i' + \mathbf{R}_{ik})$$

In the revised model, the π_{ik} are determined solely by the relative fit of the underlying LMMs according to the following equation:

$$\hat{\pi}_{ik} = \frac{f\left(\mathbf{y}_i \mid c_{ik} = 1; \hat{\boldsymbol{\Theta}}\right)}{\sum_{j=1}^{K} f\left(\mathbf{y}_i \mid c_{ij} = 1; \hat{\boldsymbol{\Theta}}\right)} \tag{3.10}$$

For example, if two latent classes are fit and the LMM for Class 1 results in a likelihood for the subject's data equal to the likelihood of that subject's data under the model for Class 2, then the subject would be included in each class 50/50. If the likelihood for Class 1 was nine times higher than for Class 2, then the subject would be included in Class 1 90 percent and Class 2 only 10 percent. Further details are presented in Section 1.4.4.

3.2.3 Prediction of the Random Effects

One of the many advantages to using LMMs is the logical specification of the model, and a big part of this specification lies in the random effects which are thought of as being sampled from an underlying distribution. For the LMM, the underlying distribution is the multivariate normal distribution. The LCLMM, in essence, divides the population into subpopulations and then fits the best LMMs to the data in each subpopulation. Therefore, it is important that the LCLMM approach be able to compute predictions of the random effects. In fact, upon closer examination, the random effects for the LCLMM are actually a weighted average of the random effects from each underlying LMM.

From the derivation of the LCLMM, recall that given a subject is a member of class k, the following distributional properties are assumed to apply to the random effects, $\mathbf{b}_i$, and residuals, $\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{W}_i\boldsymbol{\lambda}_k$.

$$\left.\begin{bmatrix} \mathbf{b}_i \\ \mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{W}_i\boldsymbol{\lambda}_k \end{bmatrix} \sim N\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{D}_k & \mathbf{D}_k\mathbf{Z}_i' \\ \mathbf{Z}_i\mathbf{D}_k & \mathbf{Z}_i\mathbf{D}_k\mathbf{Z}_i' + \mathbf{R}_i \end{bmatrix}\right) \right| \quad c_{ik} = 1 \tag{3.11}$$

The computation of the LCLMM random effects requires maximizing the joint density of the data $\mathbf{y}$ and the random effects $\mathbf{b}_i$, with respect to the terms $\mathbf{b}_i$. The likelihood can be written as follows:

$$\log L(\mathbf{y}, \mathbf{b}) =$$

$$\sum_{i=1}^{n} \log \underbrace{\sum_{k=1}^{K} \pi_{ik} \underbrace{\left[\frac{\exp\left[-\frac{1}{2}\left\{(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{W}_i\boldsymbol{\lambda}_k - \mathbf{Z}_i\mathbf{b}_i)' \mathbf{R}_{ik}^{-1} (\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{W}_i\boldsymbol{\lambda}_k - \mathbf{Z}_i\mathbf{b}_i) + \mathbf{b}_i'\mathbf{D}_k^{-1}\mathbf{b}_i\right\}\right]}{(2\pi)^{(n_i+q)/2} \left|\mathbf{R}_{ik}\right|^{1/2} \left|\mathbf{D}_k\right|^{1/2}} \right]}_{P_{3ik}}}_{P_{1i}}$$

However, given that a subject is in class k implies that the best set of predictions of the random effects are the predictions based on the k^{th} LMM. Therefore, the random effects for a subject can be computed as the weighted average of the random effects computed as if that subject were in each latent class:

$$\widetilde{\mathbf{b}}_i \;=\; \text{eBLUP}\left[\mathbf{b}_i\right] \;=\; \text{E}\left[\mathbf{b}_i \mid \hat{\boldsymbol{\Theta}}, \mathbf{y}_i\right] \;=\; \sum_{i=1}^{K} \pi_{ik} \times \widetilde{\mathbf{b}}_{i|k}$$

Further details are presented in Section 1.4.5.

3.3 Statistical Questions Related to Specifying a Model

In the LMM, several decisions need to be made by the statistician in order to specify a parsimonious model which addresses the research question. These decisions include deciding which factors are fixed and which are random, as well as deciding upon the structure of the underlying variance-covariance matrices of the random effects and error terms. The LCLMM adds to this list the decision of how many classes to fit, which of the fixed effects, if any, are allowed to vary by latent class, how class membership should be structured, and whether each of the underlying latent classes should have its own estimated variances. Below are commentaries related to each decision.

3.3.1 Fixed Effects - Class-Specific or Equal Across Classes?

In the LMM, the fixed effects are specified such that they are applicable to all subjects or to known subpopulations. In the LCLMM, the subpopulations are not known a priori. As such, it may not be obvious which fixed effects should be permitted to differ by latent class. The main point to keep in mind in making this decision is this... classes will be differentiated based solely on parameters which are allowed to differ by class. Consider the example from Chapter 1 using the ARIC study data. The primary interest in that example was on examining the trajectories of HDL, LDL and triglycerides through time. However, some subjects were taking cholesterol-lowering medications, which would be expected to result in a decrease to LDL cholesterol. This decrease is associated with the individuals taking the medication and the expectation is that the medication should result in a comparable decrease in LDL regardless of the underlying latent class. Therefore, the effect of medication use on cholesterol was fit across-classes. However, the intercept, linear, and quadratic terms for HDL, LDL, and triglycerides were allowed to differ by latent class. In some cases, there may not be prior information to guide this selection. In such cases, the statistician could fit the factor as being class-specific and examine the differences in the estimate by latent class. If the factor is reasonably similar in all classes, then the model could be refit with the factor fit as being equal across classes.

In addition, there may be instances where the statistician would prefer that all fixed effects be the same for each class and allow only the variances to differ by class. The resulting model would effectively be the usual LMM, with the error distribution and/or random effects distribution viewed as being a mixture of multivariate normal distributions.

3.3.2 Variances - Class-Specific or Equal-Across Classes?

There are two main situations which should be considered separately for purposes of this discussion. First, as noted above, the user may wish to fit the usual LMM, but allow the distributions of the random effects and/or error terms to be modeled by a mixture of normal distributions. In this scenario, the LCLMM can be specified such that all fixed effects are fit as being equal across classes, while the variances of the error terms and/or random effects are allowed to differ by latent class. Since the classes are distinguished only by variances and individuals at the center of the distribution may reasonably have come from any of the underlying distributions, it makes sense in this model to restrict each subject to have the same underlying mixture. In the univariate normal case, this would be equivalent to fitting the error terms as being distributed according to a mixture of normals - for example 90 percent distributed Normal(0,1) and 10 percent distributed Normal(0,10). However, the appropriate mixture percentages as well as the variances themselves would be estimated as part of the model. This model specification provides an easy way to generalize the LMM to better fit the data. Note that since the correlation between random effects is held constant across classes, the correlation in the resulting random effect mixture distribution is maintained, while allowing the distributions of the random effects to be represented by a more general distribution.

In models which allow the fixed effects to differ by latent class, the LCLMM attempts to fit the best LMMs to the subjects in each latent class, while determining the latent classes as part of the estimation process. In situations where there are distinctly different variances for each latent class, fitting the variances as class-specific can result in remarkably different class determinations (and therefore models) when compared to results obtained when the variances are estimated as being equal across classes. This is because the information contained in the variation of the data is utilized to distinguish the classes in the class-specific variance case, but not in the equal-across-classes case. In Chapter 1, it was shown using ARIC study data that the determination of class membership was dramatically different based on this important choice. If the statistician is unsure how the variances should be fit, it would be recommended, if sample sizes allow, to fit the variances as being class-specific first to determine the extent of the differences.

3.3.3 How Many Latent Classes Should Be Fit?

Once again, this choice should be considered in two separate contexts. First, in models in which all fixed effects are fit as being equal across classes (generalizing the usual LMM), two classes would be recommended in most cases since this should allow for a varied set of distributions but is still reasonably straightforward.

In models in which the fixed effects are allowed to vary by latent class, the statistician should fit a range for the number of latent classes and examine how the model fit changes as classes are added. As was discussed in Section 2.8, the usual likelihood-based measures tend to select more classes than is necessary. However, the statistician is encouraged to examine practical measures such as mean square error and mean absolute deviation based on the within-subject residuals to compare model fit. In the above-mentioned simulation study, MSE was found to be the most useful of the measures examined, and mean absolute deviation would be expected to have similar results but be less impacted by outlying observations. As is typical, the most parsimonious model should be sought. Therefore, while the model fit may improve slightly as classes are added, a smaller model may be preferred for its ease of interpretation.

3.3.4 Class Membership - Structured or Unstructured?

Class membership may be specified in several ways. As noted previously, in some cases, the statistician may simply want to generalize the distributional assumptions of the usual LMM to allow the underlying normal distributions to instead be mixtures of normal distributions. This is easily specified in the LCLMM and provides a straightforward way to generalize one of the basic assumptions of the LMM. As noted earlier, since individuals at the center of the distribution may reasonably have come from any of the underlying distributions, it makes sense in this model to restrict each subject to have the same underlying mixture. This situation was discussed in the example in Section 2.4.1.

In more general situations where interest focuses on discovering unique groups in the data and fitting appropriate LMMs to these groups, class membership may be specified in two very different ways. In situations where the statistician is interested in simply fitting the best K LMMs or is not sure of which factors may be most important in determining class membership, running the analysis with class membership determined solely by the relative fit of the underlying mixed models would be recommended - this is referred to throughout this document as relative fit class membership or unstructured class membership. The statistician could then examine what is different between the classes and may, as a result, have a better idea which factors are acting to differentiate subjects. For example, does Class 1 tend to have subjects with a particular disease or risk factor? Does Class 2 tend to be older or younger than most of the population?

Alternatively, the statistician may already have a set of risk factors which are of interest, and one of the goals of the analysis might therefore be to determine how the risk factors act to differentiate the groups in terms of their longitudinal measures. By including appropriate risk factors in the class membership model, it may be possible to understand both the risk factors which distinguish the subpopulations as well as how those subpopulations differ in terms of trajectories for the data being studied.

While the structured class membership model can be quite useful in certain situations, the statistician should be cautioned regarding one potential pitfall. In situations where class membership is modeled using too few predictors or predictors which simply do not adequately distinguish the underlying groups, it is likely that some subjects may have fitted class probabilities (based on the parameters in the class membership model) that put an individual in one class, while those subjects would be better fit by the underlying LMM from another class. This would result in a poor fit for selected individuals. In light of this, it is recommended that the statistician compare the fitted class membership probabilities, π_{ik}, with the unstructured probabilities based on the fitted models. The latter set of probabilities could be calculated after model fit according to Equation 3.10 and are indeed provided as part of the output from the macro presented in Chapter 2. In situations where the probabilities are not in agreement, the statistician should focus discussion primarily on the subpopulation trajectories rather than individual subject trajectories.

3.3.5 A Note Related to the Question of Fixed vs. Random?

The choice of fixed effect or random effect has been discussed in the literature (see Searle et al [1992]). However, it should be noted here that sometimes the variation accounted for by the random effects can primarily reflect variation associated with being in one of the underlying latent classes. In other words, what was considered subject-to-subject variation in the LMM may actually be due to the presence of a set of subpopulations. Therefore, as additional latent classes are fit, it is not uncommon to see the variances associated with some of the random effects decrease, in some cases going to zero.

3.4 Example 1 - Cystic Fibrosis - Pulmonary Function Data

Cystic fibrosis (CF) is one of the most common genetic diseases, with about 1 in 3,300 caucasian children born with it in 1997. The disease results in the body producing an unusually thick mucus instead of the typical slippery, watery substance. As a result, the mucus can clog the lungs, block airways, and lead to repeated infections of the respiratory system. In addition, the sticky mucus can block the pancreas, causing the body to be unable to fully absorb nutrients from digested food. In Edwards [2000], the author examines the percent predicted forced expired volume in 1 second (FEV1) from 47 adult CF patients seen at the University of North Carolina pulmonary clinic. For each individual, as few as 4 or as many as 97 measurements were documented, with the median number of measurements being 27. Percent predicted FEV1 was modeled as a linear function of age, and results based on using the LMM were compared with a cross-sectional approach, in which the information contained in having repeated measurements on individuals was effectively ignored. The author demonstrated the value of the LMM, as the cross-sectional model showed very little association between FEV1 and age, while the LMM had a statistically significant age effect (p-value<0.0001). The model from the referenced article can be written as follows:

$$\%\text{PredFEV1}_{it} = \beta_1 + \beta_2 \times \text{AGE}_{it} + \gamma_{1i} + \gamma_{2i} \times \text{AGE}_{it} + \epsilon_{it}$$

where:

$\%\text{PredFEV1}_{it}$ is the % Predicted FEV1 for subject i at age t

β_1 is the overall population intercept (with age centered at age 20)

β_2 is the overall population slope

γ_{1i} is the random intercept for subject i

γ_{2i} is the random slope for subject i

ϵ_{it} is the error term associated with subject i at age t

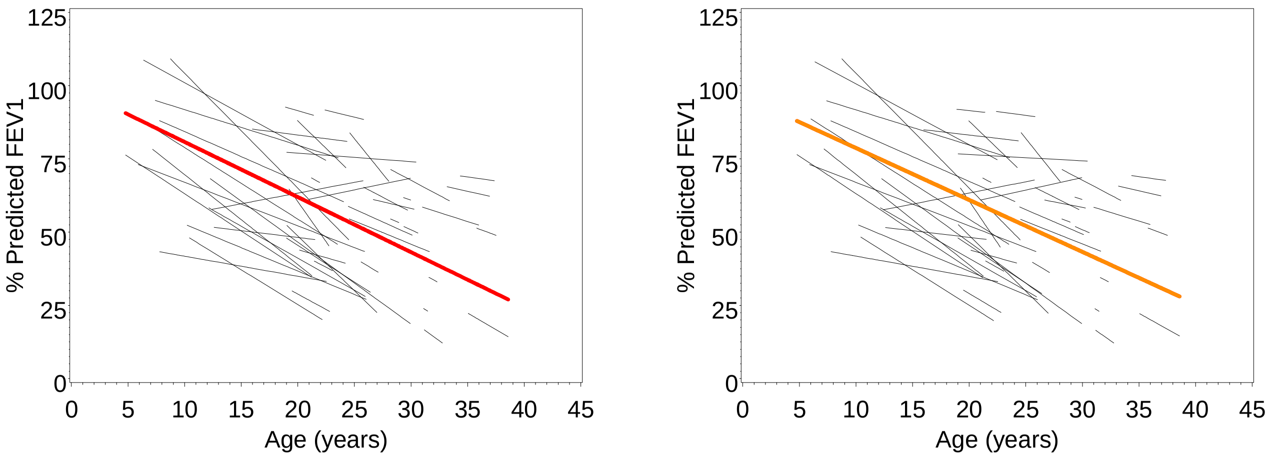
$$\gamma_{1i} \sim \text{ Normal}\left(0, \sigma^2_{int}\right)$$
$$\gamma_{2i} \sim \text{ Normal}\left(0, \sigma^2_{slope}\right)$$
$$\epsilon_{it} \sim \text{ Normal}\left(0, \sigma^2_{error}\right)$$
$$\rho = \text{ corr } \left(\gamma_{1i}, \gamma_{2i}\right)$$

In words, percent predicted FEV1 is modeled as a function of an intercept and a linear term for age. Individuals are modeled as having random intercepts and slopes about the population mean intercept and slope, and the random intercepts and slopes are correlated. The error terms are assumed to be independent with a single variance applicable to all observations.

The plot of the fitted population and subject-specific trajectories for percent predicted FEV1 based on the LMM analysis is presented in Figure 3.1a. While the fitted trajectories do not on first glance appear to indicate that there may be several underlying groups in the data, the LCLMM was applied to the data to examine whether it offers a better model fit. In looking at the fitted individual trajectories, one possible hypothesis is that the subpopulation with higher FEV1 values may be more variable in terms of trajectories. Another possibility is that the random effects have a distribution which would be better classified as a mixture of normals. For comparison purposes, the LCLMM was fit three separate ways. First, the LCLMM was fit with two latent classes where only the variances were allowed to differ by class. This effectively means that the LMM is generalized such that the error and random effect distributions are fit as mixtures of two normal distributions. The mixture percentages for this model were fit as being the same for each individual. This model is referred to below as the LCLMM-Mixture model. A second LCLMM was fit with two latent classes, where each class was permitted to have its own intercept and linear slope, but the variances were assumed to be equal for each latent class. In this second model, each individual was fit as a mixture of the underlying models based on how well the underlying models fit - see Section 3.2.2 for further details. For example, if a subject is fit equally well by both underlying LMMs, then that subject would be modeled 50/50 by those two models. If a subject was fit much better by one model than another, then that subject would have class probabilities closer to 90/10 or 95/5. This model is referred to below as the LCLMM-Relative Fit Equal-Across-Classes Variance model. Finally, the latter model was refit, allowing both the fixed effects and variances to differ by latent class and is referred to below as the LCLMM-Relative Fit Class-Specific Variance model.

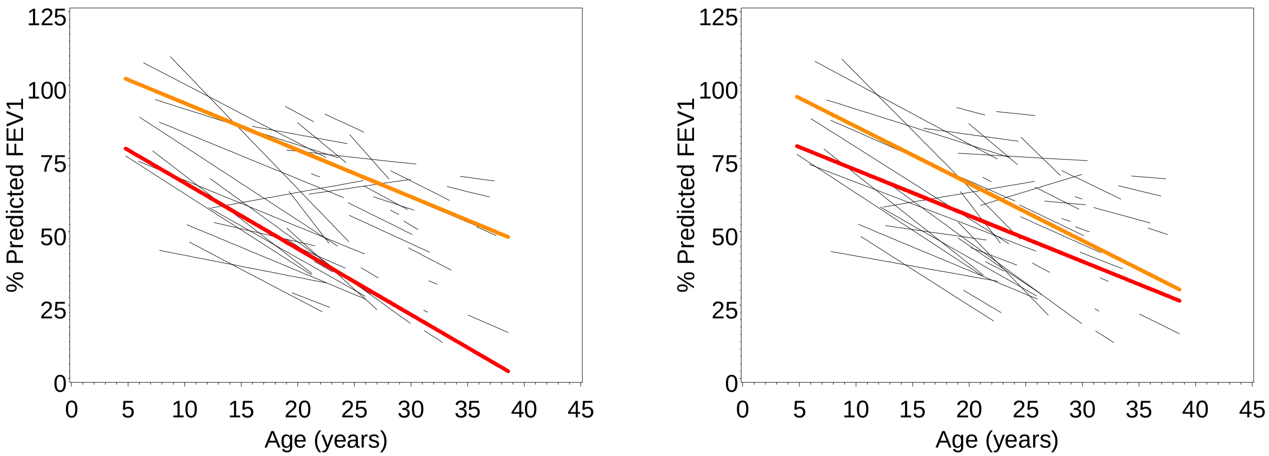
The population and subject-specific trajectories for the LMM and the three versions of the LCLMM are shown in Figure 3.1. Note that all four models produce very similar subject-specific trajectories. This is an indication that the simpler LMM may adequately describe the data, but this question will be addressed later. The parameter estimates for each model are shown in Table 3.1. The LMM and LCLMM-Mixture models produce similar population intercepts and slopes. Note that the LCLMM-Mixture model finds that the best-fitting error mixture is a 52% / 48% mixture of normal distributions with variances of 41 and 155, while the LMM fits a single residual error variance of 109. For illustration, Figure 3.2 displays a normal distribution with mean 0 and variance 109 next to a 52%/48% mixture of two normals with mean 0, one with variance 41 and the other with variance 155. Note the more compact error distribution which is fit in the LCLMM-Mixture model. The random effect mixture is more comparable to the LMM. As you might expect, the LCLMM-Relative Fit models identify higher and lower FEV1 groups. However, from the plots, it does not appear that these are necessarily unique subpopulations - just the best way of dividing the population into two subpopulations. With regard to the initial hypothesis that the subjects with higher FEV1 tend to be more variable, this is confirmed in the LCLMM-Relative Fit Class-Specific Variance model. In this model, the class with higher FEV1 had a random intercept variance of 387 versus just 277 for the lower FEV1 group, a random slope variance of 2.45 versus just 1.67 for the lower FEV1 group, and a residual error variance of 156 versus just 41 for the lower FEV1 group.

Finally, in order to determine whether any of the LCLMMs actually provide a better model fit than the LMM, Table 3.2 presents a variety of likelihood-based criteria as well as the MSE of the within-subject residuals for each of the proposed models. Further details related to these measures are available in Section 2.8. Note that all likelihood-based criteria chose the most complicated of the models, the 2-class LCLMM with class-specific variances, while the within-subject MSE found the model fit to be best under the LMM and LCLMM-Mixture models. This result is consistent with results found in the simulation study in Section 2.8. In that study, the usual likelihood-based criteria tended to choose more classes than were necessary, while the within-subject MSE was most likely to choose the appropriate number of classes. In this case, the models all provide a comparable model fit, and it is encouraging to find that the MSE does not recommend a more complex model. Generalizing the random effect and error distributions does improve the fit slightly, as seen in the slight decrease in the MSE for the LCLMM-Mixture model. However, the overall conclusion here is that the LMM appears to fit the data reasonably well and the more complex LCLMM is not necessary.



(a) Linear Mixed Model

(b) LCLMM - Mixture



(c) LCLMM - Relative Fit Equal-Across-Classes Variance

(d) LCLMM - Relative Fit Class-Specific Variance

Figure 3.1: Population and Subject-Specific Trajectories for Selected Models

Table 3.1: Example 1 - Parameter Estimates

Model	Parameter	Class	Estimate	SE
Linear Mixed Model	β_1	n/a	61.97	3.35
	β_2	n/a	-1.88	0.30
	σ^2_{int}	n/a	397.78	102.31
	σ^2_{slope}	n/a	2.39	0.91
	σ^2_{error}	n/a	109.58	4.26
	ρ	n/a	-0.15	0.22
LCLMM-Mixture - Classes 1 and 2 weighted 52% / 48%	β_1	n/a	60.98	3.28
	β_2	n/a	-1.77	0.28
	σ^2_{int}	1	315.91	124.76
	σ^2_{int}	2	401.83	142.39
	σ^2_{slope}	1	1.70	0.75
	σ^2_{slope}	2	2.41	1.32
	σ^2_{error}	1	41.38	2.96
	σ^2_{error}	2	155.52	8.15
	ρ	n/a	-0.07	0.33
LCLMM-Relative Fit Equal-Across-Classes Variances	β_1	Higher FEV1	77.90	3.14
	β_1	Lower FEV1	44.20	2.70
	β_2	Higher FEV1	-1.60	0.38
	β_2	Lower FEV1	-2.25	0.34
	σ^2_{int}	n/a	123.46	37.00
	σ^2_{slope}	n/a	1.65	0.63
	σ^2_{error}	n/a	110.20	4.28
	ρ	n/a	-0.68	0.12
LCLMM-Relative Fit Class-Specific Variances	β_1	Higher FEV1	66.45	*
	β_1	Lower FEV1	55.48	*
	β_2	Higher FEV1	-1.94	*
	β_2	Lower FEV1	-1.56	*
	σ^2_{int}	Higher FEV1	387.36	*
	σ^2_{int}	Lower FEV1	277.23	*
	σ^2_{slope}	Higher FEV1	2.45	*
	σ^2_{slope}	Lower FEV1	1.67	*
	σ^2_{error}	Higher FEV1	156.59	*
	σ^2_{error}	Lower FEV1	41.08	*
	ρ	n/a	0.04	*

* = SE calculations were unsuccessful for this model

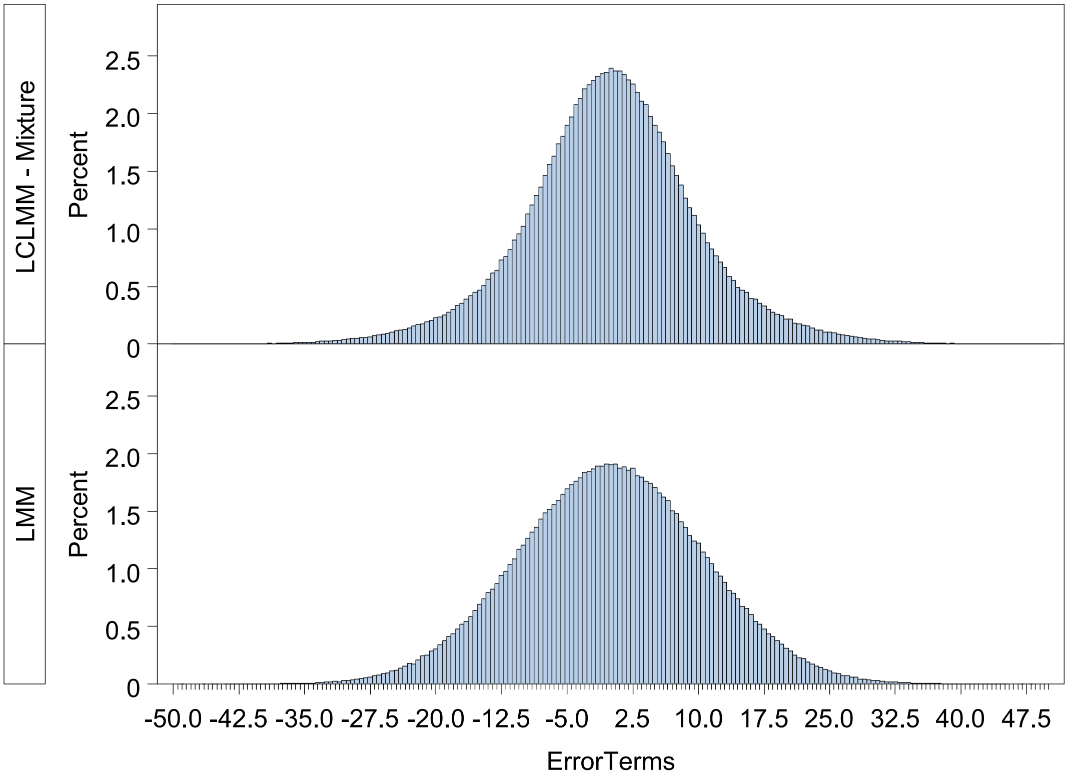


Figure 3.2: Comparison of Error Distributions - LMM vs. LCLMM Mixture

Table 3.2: Example 1 - Information Criteria and MSEs

Criteria	Linear Mixed Model	LCLMM-Mixture	LCLMM Relative-Fit Equal	LCLMM Relative-Fit Class-Specific
Observed Likelihood	-5510.82	-5414.35	-5476.30	**-5385.26**
AIC	11033.63	10848.71	10968.60	**10792.52**
CAIC	11037.66	10855.43	10973.98	**10799.91**
BIC	11044.73	10867.21	10983.40	**10812.87**
Entropy	0.00	6.20	**1.32**	3.29
C	11021.64	10841.10	10955.24	**10777.10**
ICL	11044.73	10879.61	10986.04	**10819.45**
MSE (Within-Subject)	104.28	**104.08**	105.26	104.71
Number of Parameters (v)	6	10	8	11

3.5 Example 2 - Premature Infant Sleep-Wake Data

Holditch-Davis [2004] examined the development of sleeping and waking states during the pre-term and early post-term periods for 134 preterm infants. These infants were at high risk for developmental problems due to low birthweight (less than 1500 g) or mechanical ventilation, and were recruited from three tertiary hospitals (75 from a southeastern perinatal center, five from a northeastern perinatal center, and 54 from a midwestern children's hospital). Weekly behavioral observations were conducted during a 2-hour inter-feeding time in the daytime from enrollment until 44 weeks post-conceptual age or discharge. From one to eight measurements were conducted on each infant (mean=3.3). During the weekly observation, the authors took measurements on several aspects of sleep-wake state - among these were active and quiet sleep, active and quiet waking, as well as a period defined as sleep-wake transition - the definitions of each of these states is provided in the article. The authors recorded the percentage of time each infant spent in these five sleep-wake states during each 2-hour observation period. For purposes of analysis, only subjects who had at least two separate measurements were used. In addition, if a subject had only two measurements, but these measurements were in adjacent weeks, then the data for the subject was dropped. These rules resulted in a total of 103 subjects being analyzed.

In Holditch-Davis [2004], the longitudinal measurements for percentage time in each of the five sleep-wake states were analyzed using the LMM. For each sleep-wake state, a LMM was fit in which the percent of time in that sleep-wake state was modeled as a function of the following covariates: intercept, linear and quadratic terms for postconceptional age, hospital, birthweight, gender, race, the presence of methylxanthine treatment, length of mechanical ventilation, handling due to performing an EEG, and interactions of these variables with postconceptional age. A random intercept and slope for postconceptional age were also included in each model. Note that for modeling purposes, postconceptional age was centered at 29 weeks. In addition, since one of the hospitals recruited only five infants, the indicator for hospital treated the two smaller hospitals as a single entity. A backwards elimination process resulted in models for each sleep-wake state which included the covariates presented in Table 3.3.

The authors plotted the fitted trajectories for each sleep-wake state for the population, as well as the individual fitted trajectories for each infant and found the following general patterns in the data. Active sleep tended to decrease with age, while quiet waking, active waking, and quiet sleep increased with age. Sleep-wake transition tended to increase until 40 weeks and then decreased after 43 weeks.

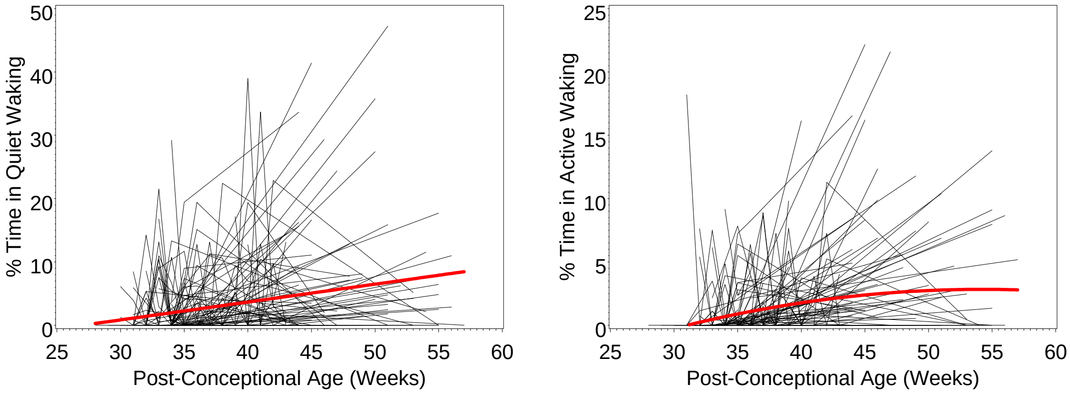
Table 3.3: List of Covariates Used in the Mixed Model Analysis in Holditch-Davis [2004]

Sleep-Wake State	Variables Included in Model
Quiet-Waking - Random intercept	Intercept and Linear Term for PC-Age
Active-Waking - Random intercept	Intercept, Linear/Quadratic terms for PC-Age, Hospital, EEG, and PC-Age x EEG
Sleep-Wake Transition - Random intercept	Intercept, Linear/Quadratic terms for PC-Age, and EEG
Active-Sleep - Random intercept and slope	Intercept, Linear/Quadratic terms for PC-Age, and Mechanical ventilation
Quiet Sleep - Random intercept and slope	Intercept and Linear term for PC-Age

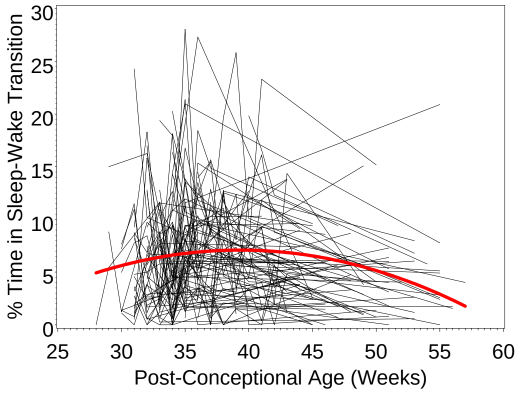
The LMM was first run to confirm that the results from the article could be recreated. The fitted model is presented in Figure 3.3 and shows close agreement with the referenced article.

Since this is an instance in which there are several related parameters, each of which has longitudinal measurements, the LCLMM can potentially shed light on underlying subpopulations that may exist in the premature infant data, as well as fit the best LMM to the data for each of those subpopulations. The models proposed in Holditch-Davis [2004] for each sleep-wake were used as the basis for the LCLMM. However, the intercept and linear/quadratic terms for age were permitted to vary by latent class. Note that other covariates, such as hospital, EEG, and mechanical ventilation, were fit across classes, so their effects are the same for each latent class. The LCLMM was fit for 2-6 classes. Class membership probabilities were modeled based on the relative fit of each of the underlying LMMs. Since the sample size was relatively small, the variances were fit as being equal across classes. As noted in Section 3.3.5, as latent classes are added to the model, it is possible that certain random effect variances may go to zero. This was the case in this analysis, and random effects were removed from the model if the associated variance went to zero. The fitted models for the various numbers of classes are presented in Figures 3.3-3.8. Note that the predicted sleep-wake percentages for each infant are not constrained to sum to 100%.

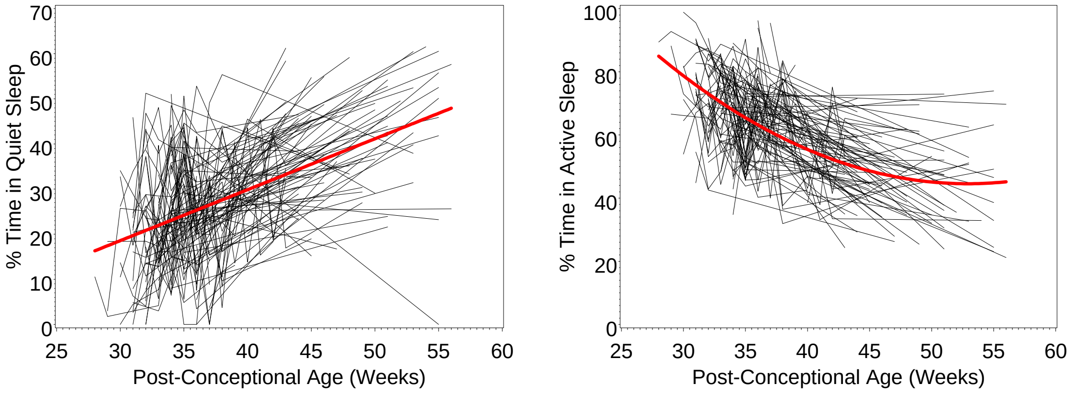
Each of the models presented is the best-fitting model for that number of classes. However, which



(a) Quiet Waking (b) Active Waking

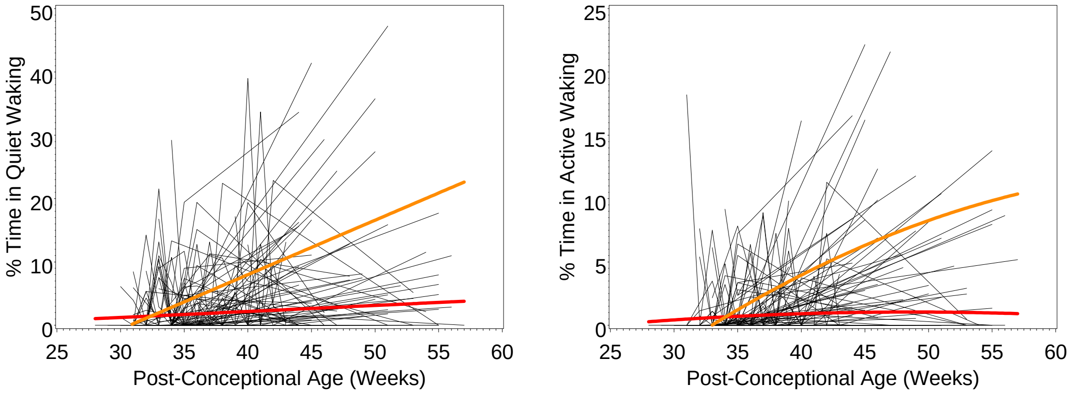


(c) Sleep-Wake Transition

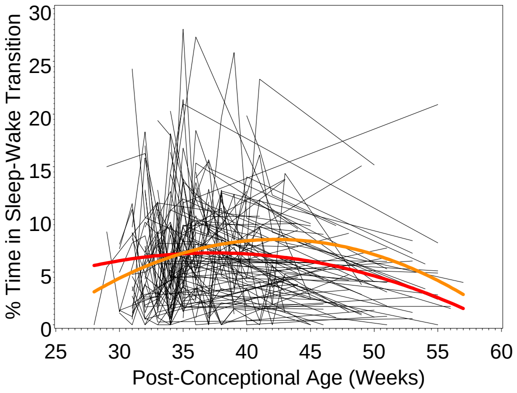


(d) Quiet Sleep (e) Active Sleep

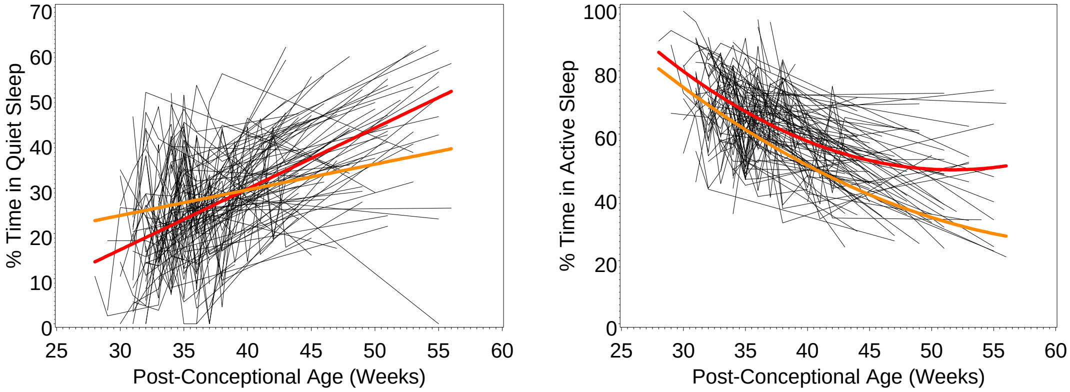
Figure 3.3: Fitted Sleep-Wake State Trajectories Based on the Linear Mixed Model



(a) Quiet Waking (b) Active Waking

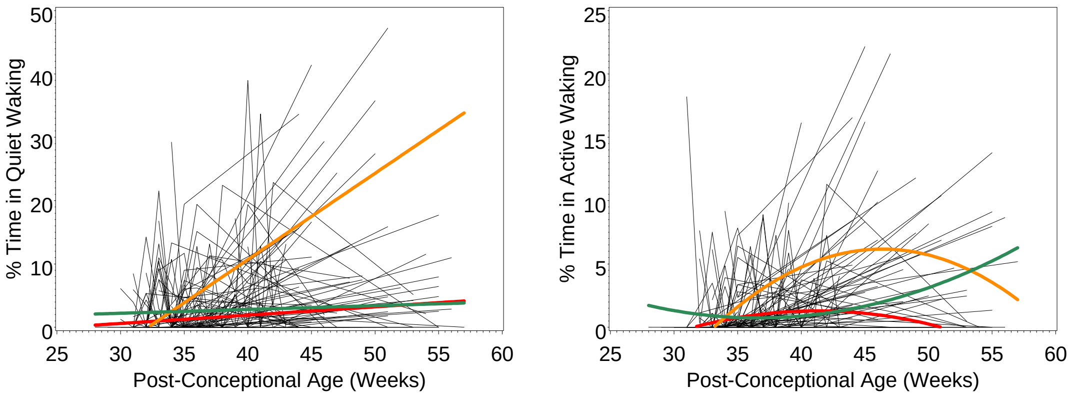


(c) Sleep-Wake Transition



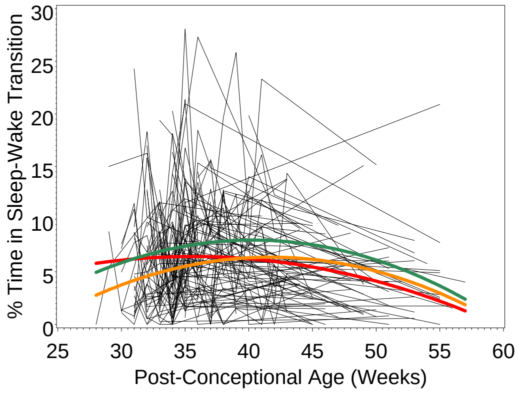
(d) Quiet Sleep (e) Active Sleep

Figure 3.4: Fitted Sleep-Wake State Trajectories Based on the LCLMM - 2 Classes

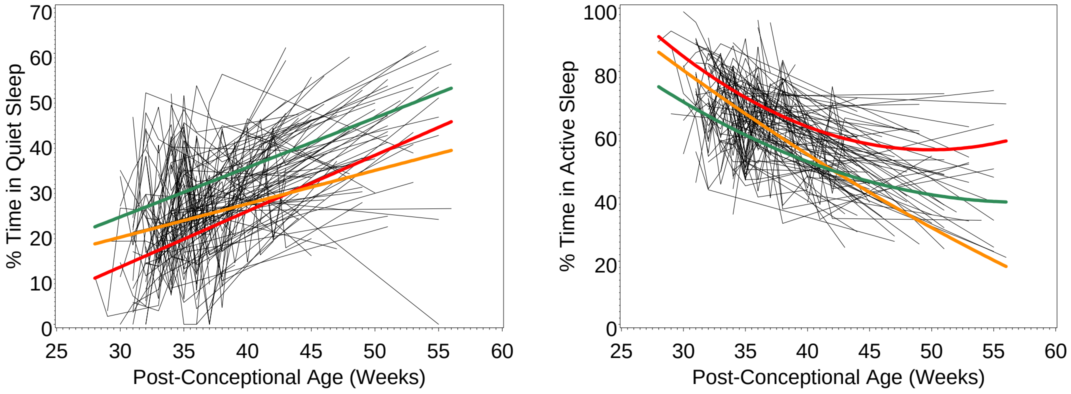


(a) Quiet Waking

(b) Active Waking



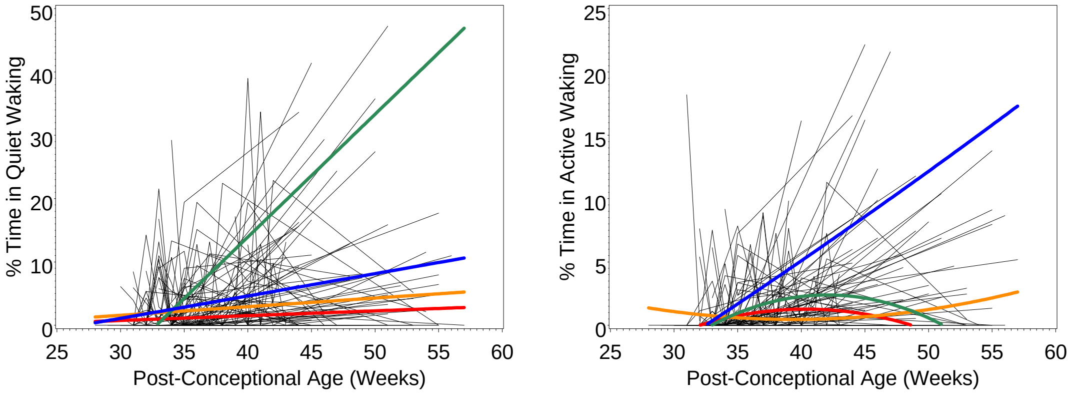
(c) Sleep-Wake Transition



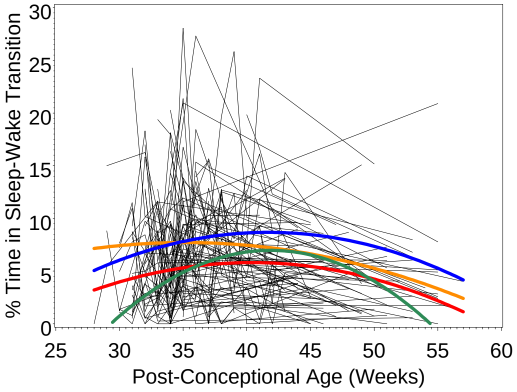
(d) Quiet Sleep

(e) Active Sleep

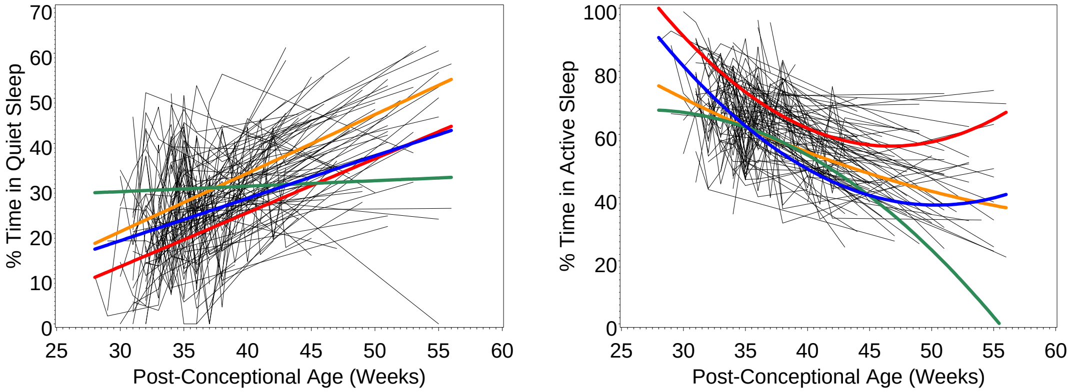
Figure 3.5: Fitted Sleep-Wake State Trajectories Based on the LCLMM - 3 Classes



(a) Quiet Waking (b) Active Waking

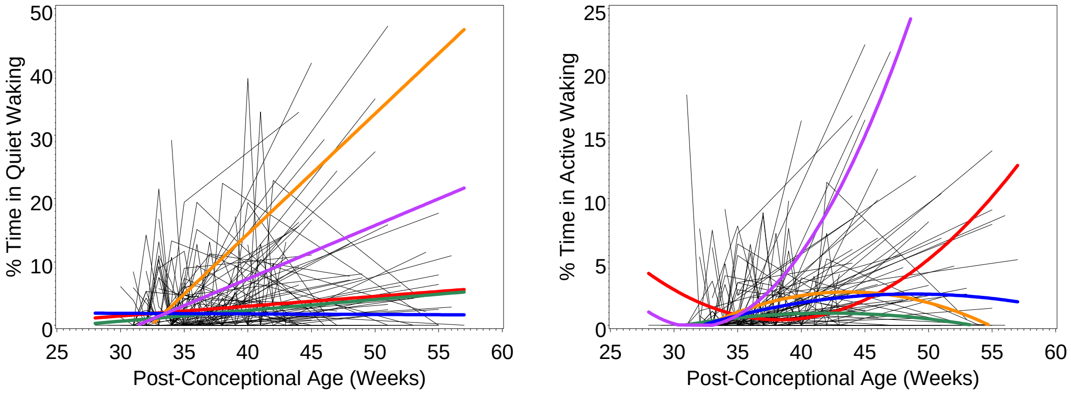


(c) Sleep-Wake Transition

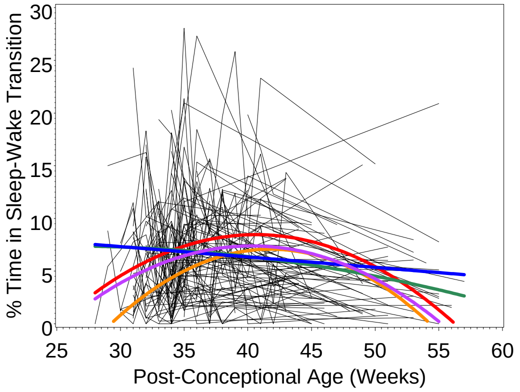


(d) Quiet Sleep (e) Active Sleep

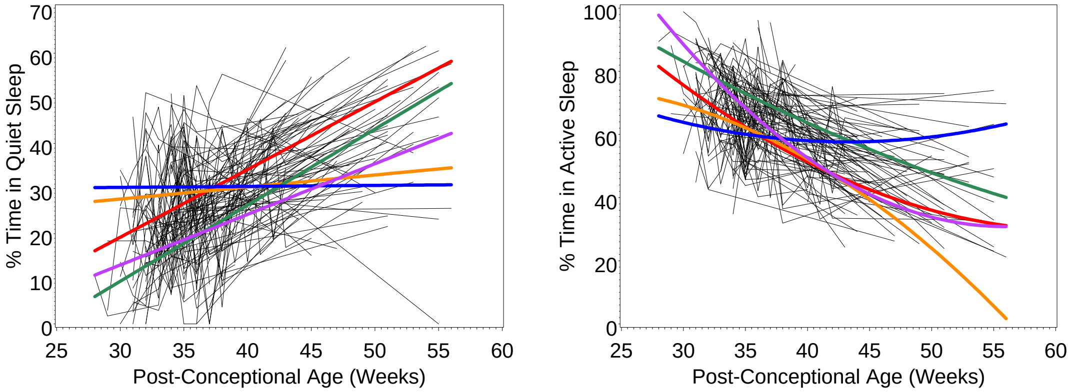
Figure 3.6: Fitted Sleep-Wake State Trajectories Based on the LCLMM - 4 Classes



(a) Quiet Waking (b) Active Waking

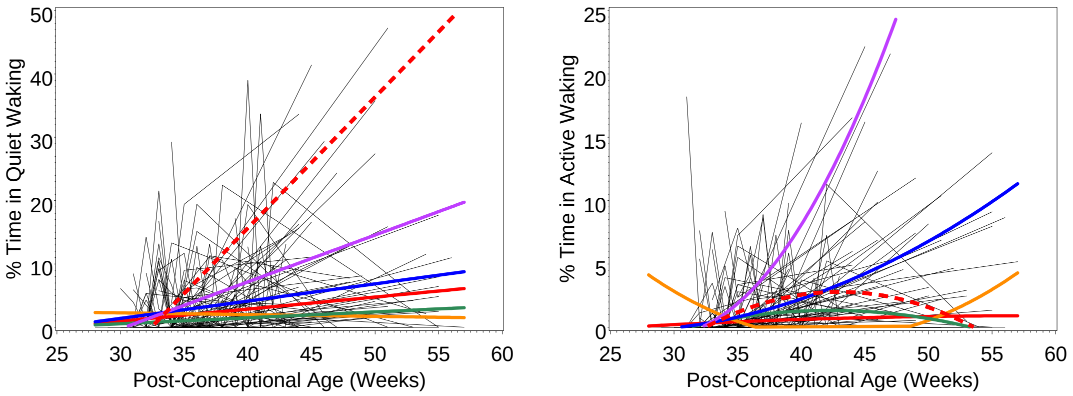


(c) Sleep-Wake Transition

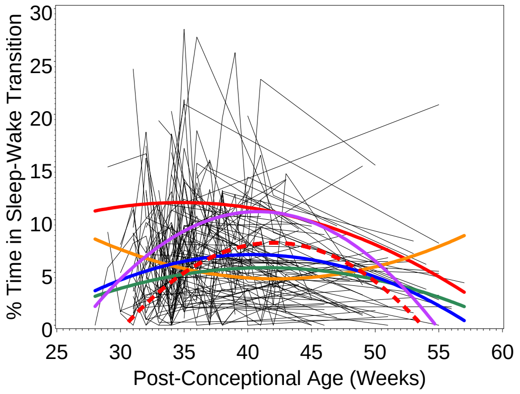


(d) Quiet Sleep (e) Active Sleep

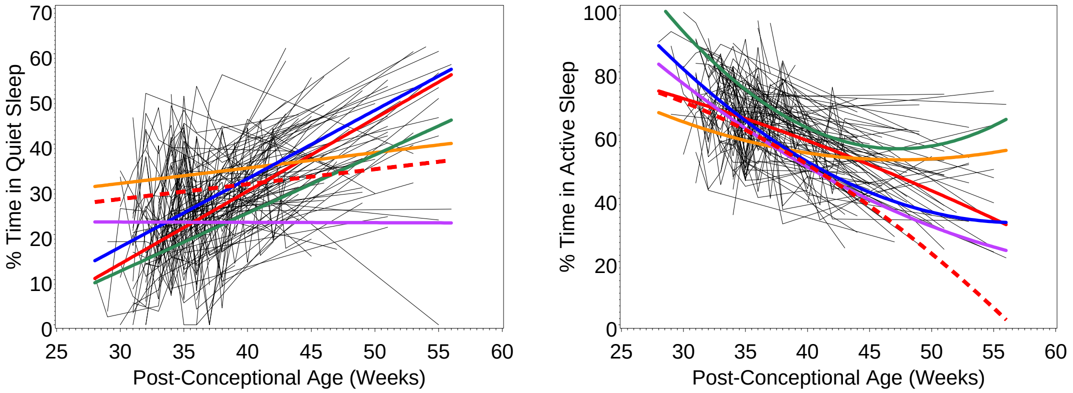
Figure 3.7: Fitted Sleep-Wake State Trajectories Based on the LCLMM - 5 Classes



(a) Quiet Waking (b) Active Waking



(c) Sleep-Wake Transition



(d) Quiet Sleep (e) Active Sleep

Figure 3.8: Fitted Sleep-Wake State Trajectories Based on the LCLMM - 6 Classes

model is best? To examine this, several likelihood-based criteria as well as the MSE of the within-subject residuals are presented in Table 3.4. A simulation study in Section 2.8 showed that the within-subject MSE provided a useful measure by which to choose from competing LCLMMs, while the likelihood-based measures tended to choose too many latent classes. In this case, while the likelihood-based criteria all choose the most complex 6-class model, the within-subject MSE selects the 5-class LCLMM. Based on the within-subject MSE, the 5-class model provides the best fit for several of the sleep-wake states and is among the best-fitting models for the other parameters. In fact, when comparing the MSEs for the 5- and 6-class models, the 5-class model has the same value or better for all of the parameters.

Table 3.4: Example 2 - Information Criteria and MSEs

	Number of Classes Fit					
Measure	1	2	3	4	5	6
Observed Likelihood	-7417.10	-7305.56	-7227.69	-7160.14	-7091.17	**-7043.26**
AIC	14898.20	14701.13	14559.37	14450.27	14338.34	**14268.52**
CAIC	14930.61	14746.71	14612.04	14516.11	14417.34	**14360.69**
BIC	14982.52	14819.69	14696.38	14621.53	14543.85	**14508.28**
Entropy	0.00	**7.95**	10.34	13.47	15.23	12.42
C	14834.2	14627.02	14476.06	14347.22	14212.8	**14111.36**
ICL	14982.52	14835.59	14717.06	14648.47	14574.31	**14533.12**
MSE (Within-Subject)						
- Quiet Waking	37.17	31.56	27.01	26.13	25.07	**25.01**
- Active Waking	7.84	5.83	6.29	4.78	4.30	**4.26**
- Sleep-Wake Transition	16.04	**15.89**	16.03	16.32	16.02	16.92
- Quiet Sleep	100.35	**100.04**	114.15	117.46	102.16	108.06
- Active Sleep	96.55	98.62	109.52	106.60	**96.42**	97.43
Number of Parameters (v)	32	45	52	65	78	91

The final parameter estimates for the 5-class model are shown in Tables 3.5 and 3.6. Note that for ease of review, each class has been identified by a unique feature in its trajectories - i.e. 'Level Quiet Sleep, Level Active Sleep'. More extensive descriptions related to the trajectories exhibited by each class are presented below. For each class, an approximate percent of the population is presented based on the sum of the class probabilities attributed to each latent class.

Table 3.5: Example 2 - Parameter Estimates for the 5-Class Model - Part 1 of 2

		Typical-More Active Sleep		Typical-More Quiet Sleep		Level Quiet/Active Sleep		High Quiet Waking		Increased Quiet/Active Waking	
Sleep-Wake State	Parameter	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE
ACROSS-CLASS:											
Active Waking	Hospital	0.53	0.266								
	EEG	0.47	0.414								
	Age x EEG	0.01	0.038	⟶		⟶		⟶		⟶	
Transition	EEG	-1.24	0.496								
Active Sleep	Mech. Vent.	1.46	0.944								
CLASS-SPECIFIC:											
Quiet Waking	Intercept	0.466	0.759	1.31	0.839	1.890	0.849	-6.47	1.518	-1.957	1.278
	Age	0.171	0.069	0.15	0.078	-0.009	0.080	1.90	0.143	0.843	0.160
Active Waking	Intercept	-0.783	0.662	2.95	0.850	-1.508	0.788	-2.47	1.259	-0.062	0.906
	Age	0.193	0.119	-0.65	0.166	0.354	0.159	0.62	0.279	-0.466	0.234
	Age^2	-0.007	0.004	0.03	0.006	-0.009	0.006	-0.02	0.012	0.085	0.013
Transition	Intercept	7.722	1.355	4.32	1.551	7.836	1.649	0.21	2.779	3.546	1.899
	Age	-0.032	0.251	0.80	0.320	-0.098	0.317	1.15	0.601	0.751	0.477
	Age^2	-0.005	0.009	-0.03	0.012	-0.000	0.011	-0.05	0.026	-0.033	0.026
Quiet Sleep	Intercept	7.761	1.585	17.72	1.673	30.245	1.741	27.41	2.878	11.943	2.800
	Age	1.687	0.144	1.50	0.156	0.022	0.169	0.267	0.288	1.122	0.334
Active Sleep	Intercept	84.104	2.888	77.79	3.378	63.954	3.417	69.86	5.669	91.961	4.321
	Age	-2.158	0.583	-2.97	0.679	-1.002	0.708	-1.04	1.346	-4.626	1.122
	Age^2	0.018	0.022	0.04	0.026	0.035	0.026	-0.06	0.060	0.086	0.061

Table 3.6: Example 2 - Parameter Estimates for the 5-Class Model - Part 2 of 2

		Typical-More Active Sleep		Typical-More Quiet Sleep		Level Quiet/Active Sleep		High Quiet Waking		Increased Quiet/Active Waking	
Sleep-Wake State	Parameter	Est.	SE	Est.	SE	Est.	SE	Est.	SE	Est.	SE
Quiet Waking	Var (Error)	25.352	1.684								
Active Waking	Var (Rand Int)	0.365	0.207	⟶		⟶		⟶		⟶	
	Var (Error)	4.651	0.346								
Transition	Var (Rand Int)	4.775	1.305								
	Var (Error)	18.129	1.361								
Quiet Sleep	Var (Error)	102.455	6.827	⟶		⟶		⟶		⟶	
Active Sleep	Var (Rand Int)	4.520	4.382								
	Var (Error)	99.991	7.618								

Green - ”Typical Patterns - More Active Sleep and Less Quiet Sleep”

- This class has fairly typical trajectories - slightly increasing quiet waking, slightly increasing active waking (although the tips of the curve are impacted by data at the edges), increasing quiet sleep, and decreasing active sleep.
- This class has among the lowest amounts of quiet sleep and among the highest amounts of active sleep.
- This class represents about 32 percent of the population.

Red - ”Typical Patterns - More Quiet Sleep and Less Active Sleep”

- The class also has fairly typical trajectories - slightly increasing quiet waking, slightly increasing active waking, increasing-then-decreasing transition, increasing quiet sleep, and decreasing active sleep - but has a higher amount of quiet sleep than other classes and one of the lower amounts of active sleep.
- This class represents about 28 percent of the population.

Blue - ”Level Quiet Sleep, Level Active Sleep”

- Unlike most other classes, this class did not have an increase in quiet sleep and did not have a decrease in active sleep.
- This class represents about 23 percent of the population.

Orange - ”High Quiet Waking”

- This class had dramatic increases in quiet waking, much more than any other class.
- This class represents about 11 percent of the population.

Purple - ”Increased Quiet Waking and Active Waking”

- This class had a dramatic increase in active waking and one of the largest increases in quiet waking.
- This class represents about 10 percent of the population.

It is useful to divide the population into groups based on their sleep-wake patterns because these patterns have the potential to tell us something about the development of the infants. Previous studies conducted by Halpern et al [1995] and Thoman [1982] found that sleeping and waking reflect underlying functioning of the brain. In addition, Borghese et al [1995], Gertner et al. [2002], Scher et al. [1994], and Whitney and Thoman [1993] found that sleep-wake patterns during the preterm period are related to developmental outcomes. Therefore, the developmental measures from subsequent testing of the premature infants are summarized in Table 3.7 for each class identified in the final model. Note that since data for the developmental measures was not available for all subjects, two N's are presented for each class - the Class N is the number of subjects fit in the class, while N is the number who also had the developmental measure being sumarized. As background, PLS 3, which is short for Preschool Language Scale, is a test which evaluates language development. It can be used to test receptive and expressive language skills with children from two weeks through six years of age. The Fagan test is a test which evaluates infant intelligence by measuring the amount of time the infant spends looking at a new object compared with the time spent looking at a familiar object. The Bayley Scales of Infant Development is an individually administered psychometric assessment of developmental functioning consisting of two scales - the Mental Development Index (MDI) and the Psychomotor Development Index (PDI). Summaries for other relevant information about the mother and infant are presented in Tables 3.8-3.9. Graphical summaries for two developmental measures, as well as birthweight and gestational age are presented in Figures 3.9 and 3.10.

It is important to recognize that since each subject is not assigned 100 percent to be in a particular latent class, but instead is given a probability based on the relative fit of the underlying LMMs, the above-mentioned summaries are not simple statistics. The boxplot and associated table typically display the min, max, 25th and 75th percentiles, as well as the mean and median. However, since every subject is in every class with at least some small percentage, the min and max cannot be calculated as they usually are - otherwise all classes would have the same min and max. For purposes of calculation of the min and max, each subject was classified into a most likely class based on their class probabilities. If a subject was at least twice as likely to be in their most likely class versus their second-most likely class, then that subject was classified as being 'in' their most likely class. Then, the min and max based on subjects in the respective classes were used as each class' min and max for the boxplot. The median and 25th and 75th percentiles have a similar concern. For these, the easiest way to illustrate calculation is by example. To calculate the median and 25th and 75th percentiles for birthweight for Class 1, for example, the birthweight values for all subjects were lined up from smallest to highest, as would normally done. The difference here is that alongside each subject's birthweight, the probability of being in Class 1 is also recorded. These class probabilities are summed and a CDF value is defined for each subject as the cumulative sum of the Class 1 probabilities divided by the total of all of the Class 1 probabilities. Then the birthweight value at the CDF values of 25, 50 and 75 percent were used as the 25th percentile, median, and 75th percentile, respectively. For categorical yes/no responses, a weighted percentage of 'yes' responses is calculated based on each subject's class probability.

Below are some observations related to how the classes compare both in their developmental measures and with respect to other miscellaneous covariates.

Green - "Typical Patterns - More Active Sleep and Less Quiet Sleep"

- This class had the lowest average birthweight (1096.3 grams) - a fair bit lower than the next closest class (1133.7 grams).
- This class also had the lowest gestational age (27.9 weeks).
- This class seemed to score in the middle on most of the developmental scales.

Red - "Typical Patterns - More Quiet Sleep and Less Active Sleep"

- This class had the second-highest mean birth weight (1197 grams) and second-highest gestational age (28.7 weeks).
- This class had the second-highest APGAR scores (5.7 at 1 minute and 7.3 at 5 minutes).
- This class had the highest mean of any class on the PLS-3 (101.2) and the highest or one of the highest means for the other developmental measures.

Blue - "Level Quiet Sleep, Level Active Sleep"

- This class was average in terms of birthweight (1149.9 grams) and gestational age (28.3 weeks).
- This class had the highest mean age of the mother (29.3 years).
- This class had among the lowest developmental scores by many measures.

Orange - "High Quiet Waking"

Table 3.7: Example 2 - Summary of Developmental Outcomes

Measure	Class	Class N	N	Mean	Min	Q1	Median	Q3	Max
MDI (27 Month)	Typical-More Quiet Sleep	27.74	10.88	96.61	62.00	82.38	100.58	106.21	119.00
	High Quiet Waking	10.51	2.17	94.96	69.66	69.66	101.94	111.77	115.00
	Typical-More Active Sleep	31.76	10.25	90.81	52.00	80.03	86.29	97.90	121.00
	Level Quiet/Active Sleep	23.25	6.81	80.78	52.00	64.83	77.03	94.76	102.00
	Incr Quiet/Active Waking	9.74	0.90	80.54	75.46	75.46	81.21	83.86	83.86
PDI (27 Month)	Typical-More Quiet Sleep	27.74	9.84	87.88	73.00	78.35	85.95	93.16	107.00
	High Quiet Waking	10.51	1.28	88.12	87.20	87.20	88.35	88.59	88.65
	Typical-More Active Sleep	31.76	8.37	82.99	74.57	74.57	77.16	84.59	103.00
	Level Quiet/Active Sleep	23.25	7.65	80.91	61.00	75.77	83.13	88.41	95.00
	Incr Quiet/Active Waking	9.74	0.86	92.34	79.79	79.79	97.14	98.15	98.15
PLS 3	Typical-More Quiet Sleep	27.74	15.08	101.24	68.00	85.34	100.14	114.92	132.00
	High Quiet Waking	10.51	4.06	98.51	84.92	84.92	90.41	121.55	127.00
	Typical-More Active Sleep	31.76	18.50	94.25	64.00	84.36	90.30	102.54	123.00
	Level Quiet/Active Sleep	23.25	13.05	93.04	64.00	77.05	92.96	108.23	122.00
	Incr Quiet/Active Waking	9.74	3.31	82.35	62.00	62.00	88.60	90.48	90.83
Mean Fagan	Typical-More Quiet Sleep	27.74	19.11	57.77	48.82	55.42	57.07	58.95	71.72
	High Quiet Waking	10.51	7.11	57.10	52.90	53.92	55.66	57.80	64.22
	Typical-More Active Sleep	31.76	22.14	56.71	48.13	52.42	56.22	58.85	66.34
	Level Quiet/Active Sleep	23.25	14.55	56.16	41.67	53.15	57.20	58.59	63.48
	Incr Quiet/Active Waking	9.74	5.09	55.98	50.52	51.73	55.69	58.86	58.88

Table 3.8: Example 2 - Summary of Infant/Mother Characteristics - Continuous

Measure	Class	N	Mean	Min	Q1	Median	Q3	Max
Birth Weight (grams)	Typical-More Quiet Sleep	27.74	1197.06	582	850.02	1166.27	1397.59	2189
	High Quiet Waking	10.51	1354.52	1035	1195.31	1327.22	1534.78	1980
	Typical-More Active Sleep	31.76	1096.29	562	747.04	1040.75	1333.62	2280
	Level Quiet/Active Sleep	23.25	1149.90	475	896.81	1093.90	1338.66	1798
	Incr Quiet/Active Waking	9.74	1133.67	665	973.31	1131.87	1314.63	1423
Gestational Age (weeks)	Typical-More Quiet Sleep	27.74	28.66	25	26.35	28.15	29.86	33
	High Quiet Waking	10.51	30.33	27	28.47	29.85	31.42	34
	Typical-More Active Sleep	31.76	27.94	23	25.45	27.51	29.52	32
	Level Quiet/Active Sleep	23.25	28.33	24	25.95	27.94	29.46	32
	Incr Quiet/Active Waking	9.74	28.23	26	26.27	27.24	29.10	31
APGAR 1	Typical-More Quiet Sleep	27.74	5.72	1	4.06	5.62	6.71	9
	High Quiet Waking	10.51	5.14	0	2.51	5.03	6.80	9
	Typical-More Active Sleep	31.76	5.36	1	2.97	5.91	6.74	9
	Level Quiet/Active Sleep	23.25	5.19	1	3.13	4.65	6.62	9
	Incr Quiet/Active Waking	9.74	6.10	1	5.24	6.32	6.90	9
APGAR 5	Typical-More Quiet Sleep	27.74	7.31	5	6.13	6.79	7.65	9
	High Quiet Waking	10.51	6.95	1	5.70	6.87	8.12	9
	Typical-More Active Sleep	31.76	7.15	2	6.11	6.99	7.83	9
	Level Quiet/Active Sleep	23.25	7.20	1	6.08	6.78	7.72	9
	Incr Quiet/Active Waking	9.74	7.36	6	6.29	6.80	7.52	9
Mother Age (years)	Typical-More Quiet Sleep	27.74	28.39	19	21.75	26.68	33.56	40
	High Quiet Waking	10.51	27.10	19	20.64	27.48	32.12	38
	Typical-More Active Sleep	31.76	29.22	19	24.20	28.59	33.37	42
	Level Quiet/Active Sleep	23.25	29.27	18	21.94	29.22	32.98	48
	Incr Quiet/Active Waking	9.74	25.58	22	22.29	23.41	27.25	32

Table 3.9: Example 2 - Summary of Infant/Mother Characteristics - Categorical

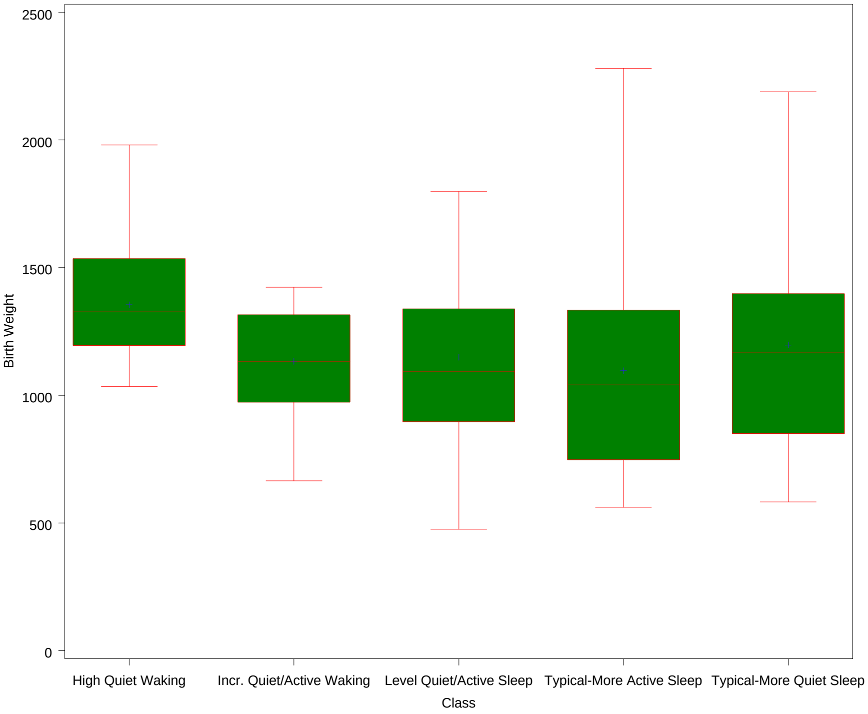
	Percent of Infants with Associated Yes Response				
Covariate	Typical-More Quiet Sleep	High Quiet Waking	Typical-More Active Sleep	Level Quiet/Active Sleep	Increased Quiet/Active Waking
Firsttime Mother?	54.0	58.5	52.3	50.3	67.4
Multiple Birth?	26.3	28.5	29.1	19.8	19.1
Cesarean	60.6	63.0	59.0	62.0	45.5

- This class had the highest mean birth weight of any class by far (1354 grams) - the next highest was 1197 grams.
- This class had the highest gestational age of any class by far (30.3 weeks) - the next highest was 28.7 weeks.
- This class had the lowest mean APGAR scores (5.1 at 1 minute, 7.0 at 5 minutes).
- While counts with developmental data were small for the development scores, the class scored among the highest based on the few measures available.

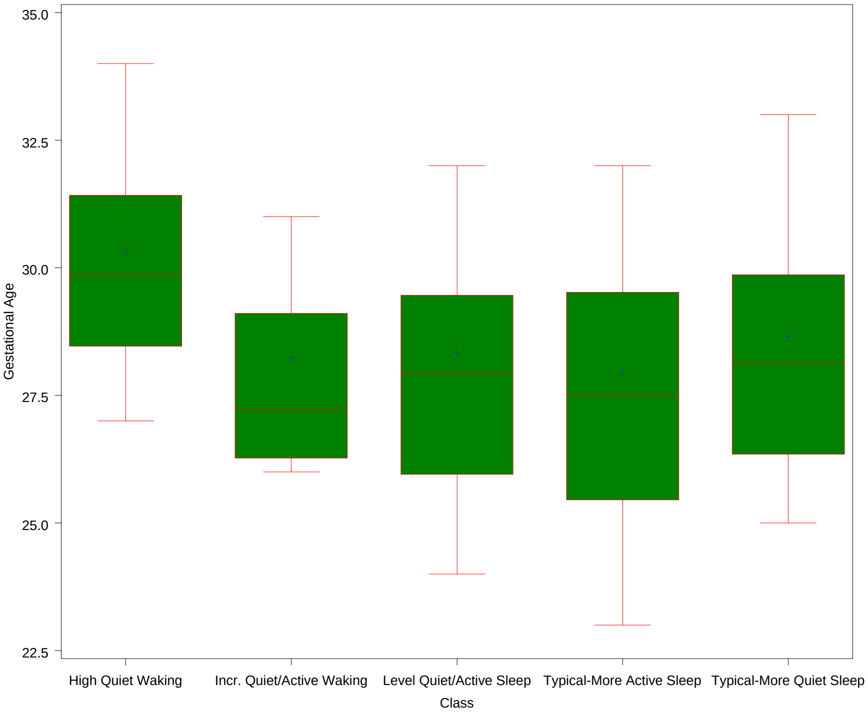
Purple - "Increased Quiet Waking and Active Waking"

- This class was average in terms of birthweight (1133.7 grams) and gestational age (28.2 weeks).
- This class had the highest APGAR scores (6.1 at 1 minute, 7.4 at 5 minutes).
- This class had the youngest mean age of the mother by far at 25.6 years - the next closest was 27.1.
- This class had the highest percent of first-time mothers (67.4 percent) and lowest percent of cesarean sections (45.5 percent).
- This class had among the lowest developmental scores by many measures.

It is apparent from the class descriptions that developmental score is not a measure that simply reflects the birthweight and/or gestational age of the infants. There is a more complex process at

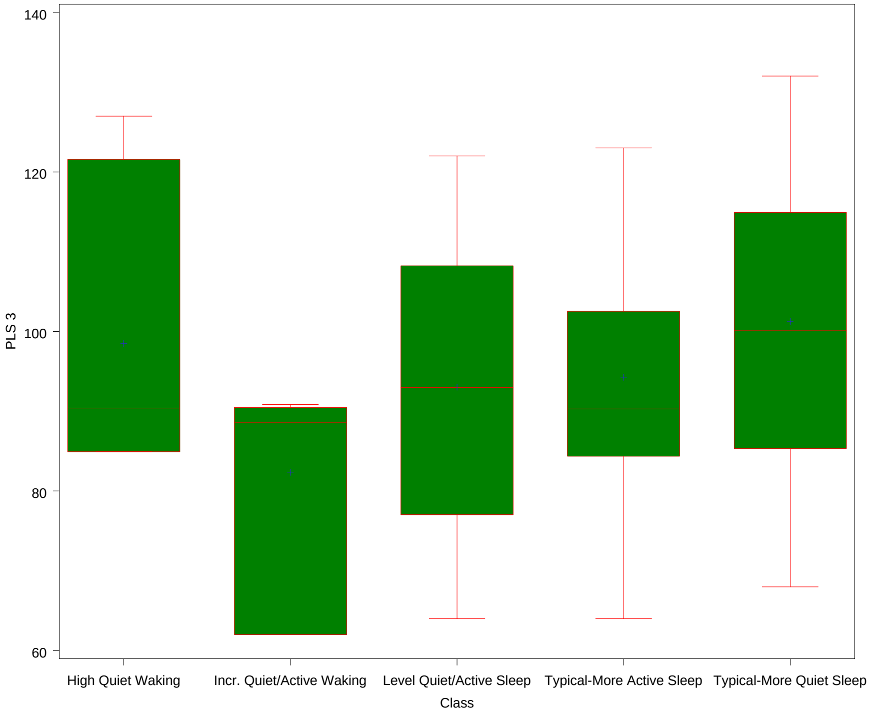


(a) Birthweight (grams)

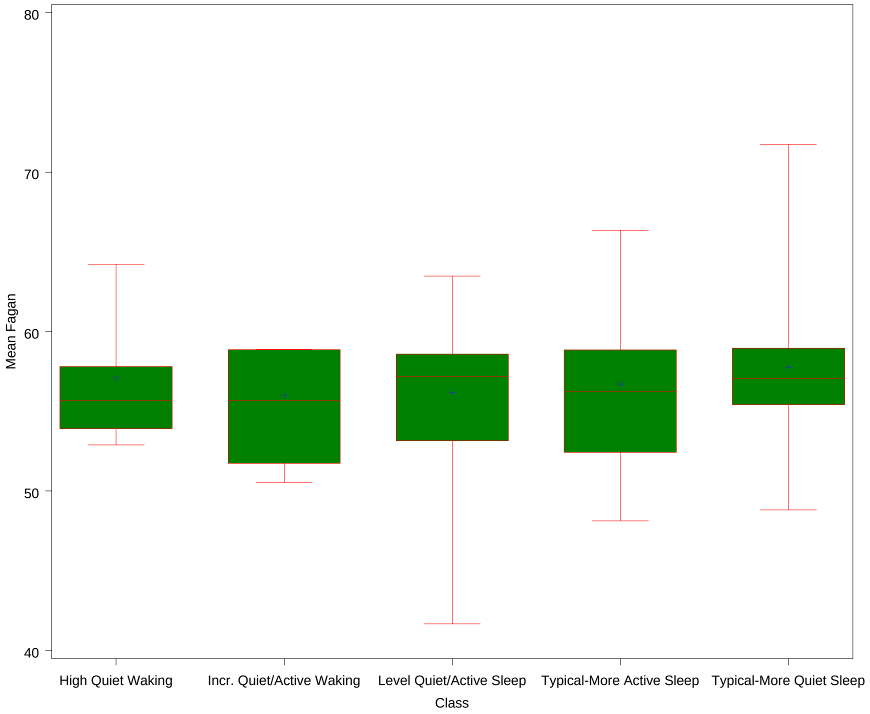


(b) Gestational Age (weeks)

Figure 3.9: Summary of Characteristics of Identified Classes - Part 1



(a) PLS 3



(b) Fagan

Figure 3.10: Summary of Characteristics of Identified Classes - Part 2

work, and analysis of the sleep-wake patterns of these infants may provide an earlier assessment of their brain development. This example has demonstrated how the LCLMM allows the statistician to both divide the population into subpopulations and simultaneously fit the best LMM for the data in each group. In situations where there are several longitudinal measures, this can be a very complex task, and the LCLMM provides a useful tool in this regard.

3.6 Example 3 - Atherosclerosis Risk in Communities Study (ARIC) Revisited

The Atherosclerosis Risk in Communities Study (ARIC) is a prospective epidemiologic study designed to investigate the etiology and natural history of atherosclerosis, the etiology of clinical atherosclerotic diseases, and the variation in cardiovascular risk factors, medical care and disease by race, gender, location, and date. The study was conducted in four communities in the United States - Forsyth County, NC, Jackson, MS, the northwestern suburbs of Minneapolis, MN, and Washington County, MD. Each ARIC field center randomly selected and recruited a cohort sample of approximately 4,000 individuals aged 45-64 from a defined population in their community. A total of 15,792 individuals participated. Study participants received an extensive examination, including medical, social, and demographic data. These participants were reexamined every three years with the first screen (baseline) occurring in 1987-89, the second in 1990-92, the third in 1993-95, and the fourth and final exam in 1996-98. Follow-up occurs yearly by telephone to maintain contact with participants and to assess health status of the cohort.

A subset of 2,066 members of the ARIC study cohort participated in the Carotid MRI Substudy in 2004-2005. The goal was to recruit 1,200 participants with high values of maximum carotid artery wall thickness at their last ultrasound examination, and 800 individuals randomly sampled from the remainder of the carotid artery wall thickness distribution. Participants had a maximum of the four ARIC cohort examinations, plus one Carotid MRI Substudy examination. Measurement protocols were identical at all five visits.

Fasting blood samples were collected at each examination and assayed for total cholesterol, HDL cholesterol, and triglycerides. LDL cholesterol was calculated according to the Friedewald formula. LDL cholesterol, HDL cholesterol, and triglycerides are commonly considered risk factors for coronary artery disease (CAD) and other related diseases or events and are also routinely measured at annual physicals in the general population. Ballantyne [1998] notes that "clinical trials with statins and other lipid-regulating therapies have conclusively shown that lowering LDL cholesterol decreases both morbidity and mortality from CAD and other vascular diseases." Kwiterovich [1998] notes that "the Framingham Heart Study produced compelling epidemiologic evidence indicating that a low level of HDL cholesterol was an independent predictor of coronary artery disease." And Gotto [1998] points out that "the current evidence argues compellingly for including triglycerides in the evaluation of patient risk for CAD. ... The revived attention to hypertriglyceridemia with respect to increased CAD risk represents an important step in assessing a patient's global risk for developing CAD."

For purposes of analysis, lab lipid data for the 2,066 individuals in the ARIC Carotid MRI substudy, as well as information related to medication use for controlling cholesterol levels were utilized in the models discussed below. Since the distribution of triglycerides is non-normal, a log transformation for this lab value was used in the analyses. The goal of this application is to demonstrate the usefulness of the LCLMM for modeling the data for LDL cholesterol, HDL cholesterol, and triglycerides simultaneously. Section 1.5 already discussed an analysis in which the class probabilities were determined by the relative fit of the model from each underlying latent class. In order to illustrate another useful way to specify a LCLMM, this data is analyzed a second time here. In this revised analysis, class membership is instead structured such that individuals are classified into groups based on a limited number of prespecified factors.

As in Section 1.5, the LMM is fit with an intercept and linear and quadratic terms for age for each of the three lipid parameters. Two indicator variables - one for LDL cholesterol and one for HDL cholesterol - are also included to account for the effect of cholesterol medication on these measures. Random intercepts are included for all three lipid parameters, and a random slope is fit for LDL (the variances for the random slopes for HDL and triglycerides were close to 0). The random intercept and slope for LDL are fit as being correlated, and the error terms are assumed to be uncorrelated, with separate variances for each lipid parameter. For the LCLMM, the intercept and linear and quadratic age terms for HDL, LDL, and triglycerides are allowed to vary for each latent class or subpopulation. However, the effect of the cholesterol-lowering medication on HDL and LDL is fit as being the same for each latent class. Random intercepts for all three lipids and a random slope variable for LDL cholesterol are fit in the LCLMM as in the LMM, with correlation permitted between the LDL intercept and slope. The LMM (1 class) as well as the LCLMM with 2-5 classes were fit. In the LCLMM, the variances were permitted to differ by latent class, while the correlation parameter was fit as being the same for each latent class. The main difference in this application is that class membership is modeled via a logistic regression model with an intercept and the following covariates:

- Did the subject ever smoke cigarettes? (1=Yes/0=No)
- Does the subject exercise or play sports? (1=Yes/0=No)
- Has the subject ever been treated for high blood pressure? (1=Yes/0=No)
- Has the subject ever been treated for high cholesterol? (1=Yes/0=No)
- Baseline BMI
- Baseline HDL Cholesterol
- Baseline Glucose
- Baseline Triglycerides
- Baseline Heart Rate
- Number of glasses of wine consumed per week

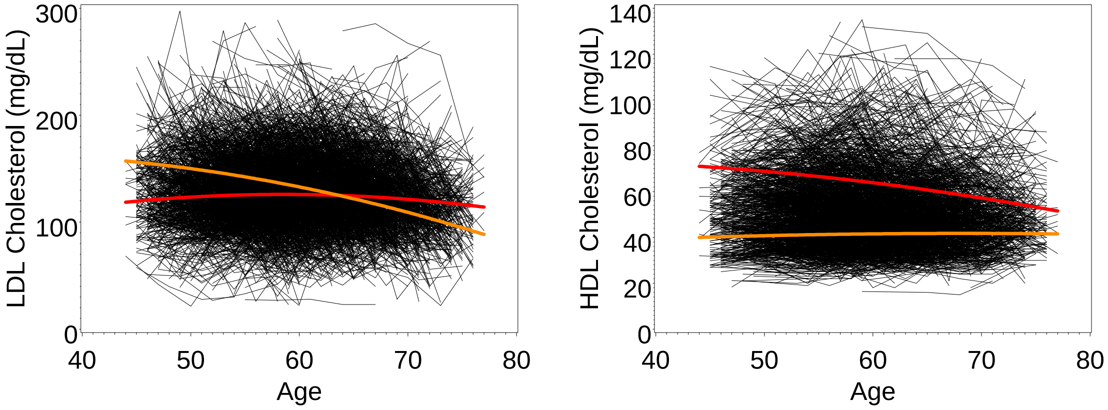
Selected information-based criteria and the within-subject MSE from each of the models are shown in Table 3.10. Note that the likelihood-based criteria all choose the highest number of classes. This is consistent with a result found in the simulation study presented in Section 2.8, which found that the usual likelihood-based criteria tend to overestimate the number of classes that should be fit. Note that the ICL shows a big improvement in going from one class to two classes, then has a somewhat smaller drop from two to three classes, and then decreases more slightly as classes four and five are added. This is evidence that perhaps the 2- or 3-class models fit reasonably well and there is not a substantial benefit of adding classes four and five. In addition, note that entropy jumps when the fourth class is added, indicating more uncertainty in class assignments after three classes. Another measure to examine is the within-subject MSE. The above-mentioned simulation study showed that the within-subject MSE provided a useful measure by which to choose from competing LCLMMs. In this case, the within-subject MSE shows an improvement in model fit for HDL when the second class is added, a slight improvement for triglycerides, and a slightly worse fit for LDL. Adding the third and subsequent classes does not appear to improve the model fit based on within-subject MSEs. Given the above, the 2-class model was selected in this case. This is indeed consistent with the ARIC study data. Recall that the 2,066 subjects were selected as two subsets - approximately 1,200 participants with high values of maximum carotid artery wall thickness at their last ultrasound examination and 800 individuals randomly sampled from the remainder of the carotid artery wall thickness distribution.

Table 3.10: Example 3 - Information Criteria and MSEs

	Number of Classes Fit				
Measure	1	2	3	4	5
Observed Likelihood	-82670.94	-80649.81	-79882.32	-79393.77	**-79016.73**
AIC	165379.87	161391.63	159910.64	158987.54	**158287.45**
CAIC	165423.86	161498.12	160079.64	159219.05	**158581.47**
BIC	165486.91	161650.76	160321.87	159550.88	**159002.89**
Entropy	0.00	**119.37**	223.17	470.79	494.06
C	165341.88	161538.37	160210.99	159729.12	**159021.58**
ICL	165486.91	161889.51	160768.22	160492.46	**159991.01**
MSE (Within-Subject)					
- HDL Cholesterol	47.669	**45.393**	46.204	46.204	46.481
- LDL Cholesterol	**365.078**	368.808	369.400	369.400	369.756
- Triglycerides	0.073	0.072	0.072	0.072	**0.071**
Number of Parameters (v)	19	46	73	100	127

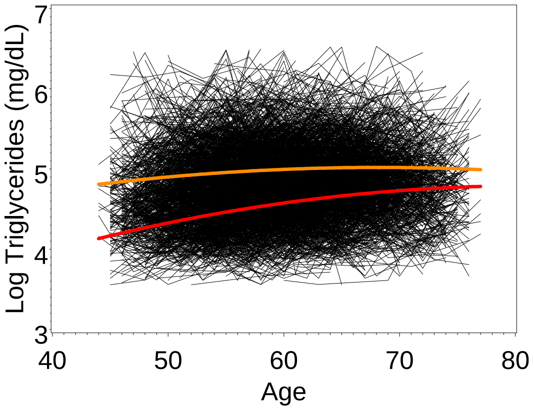
The fitted population trajectories for each lab parameter are displayed in Figure 3.11. In addition, the observed subject trajectories for subjects identified in each latent class are presented in Figures 3.12-3.13, along with the fitted population trajectories for each class. The final parameter estimates are displayed in Table 3.11. Similar to the 2-class model fit from the example in Section 1.5, the model detects a class with higher HDL and lower triglycerides than the other - this class is labeled as 'Optimal' for discussion here, while the class with lower HDL and higher triglycerides is labeled as 'At-Risk'. The 'Optimal' class also tends to have LDL that is average and steady, while the 'At-Risk' class has LDL that starts out higher but then decreases. This is associated with the fact that while both groups tended to be on cholesterol-lowering meds at baseline in equal percentages, a greater percentage of the 'At-Risk' group tended to be on cholesterol-lowering meds by visit 5.

The main difference between this model and the model discussed in Section 1.5 is that an individual's mixture probabilities are determined by a logistic regression model which has a limited number of variables. This can be a very powerful feature since it allows the statistician to simultaneously divide the population into classes based on a set of class membership parameters while also fitting the best



(a) LDL Cholesterol

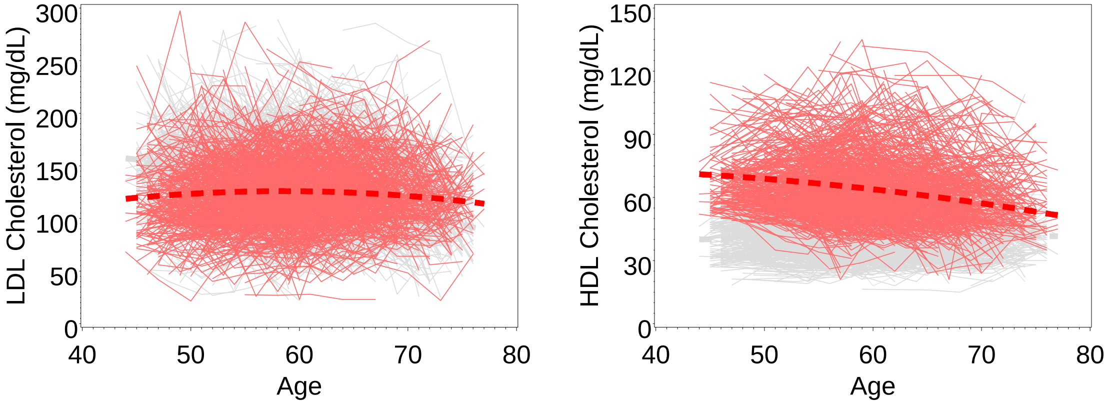
(b) HDL Cholesterol



(c) Triglycerides

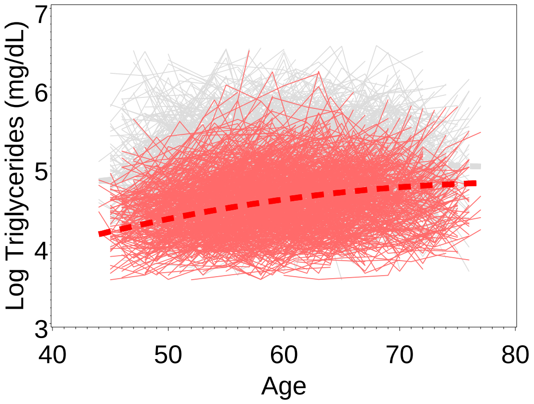
Figure 3.11: Example 3 - Fitted LCLMM (2 Classes)

Red = Optimal
Orange = At-Risk



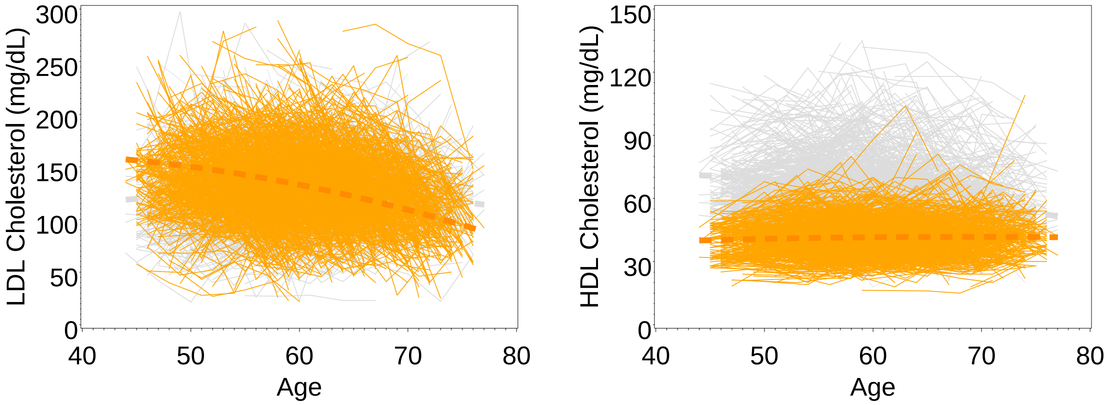
(a) LDL Cholesterol

(b) HDL Cholesterol

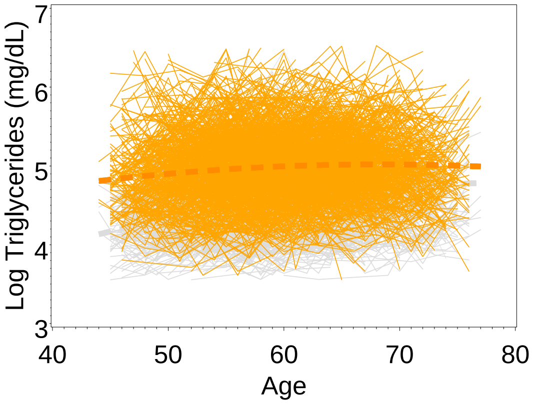


(c) Triglycerides

Figure 3.12: Example 3 - Fitted LCLMM (2 Classes) - Optimal Class



(a) LDL Cholesterol (b) HDL Cholesterol



(c) Triglycerides

Figure 3.13: Example 3 - Fitted LCLMM (2 Classes) - At-Risk Class

Table 3.11: Example 3 - Parameter Estimates for 2-Class LCLMM with Class-Specific Variances

	Optimal		At-Risk	
Parameter	Estimate	SE	Estimate	SE
Effect of Chol. Meds on LDL	-17.03	0.6919	⟶	
Effect of Chol. Meds on HDL	-0.17	0.2172	⟶	
LDL - Intercept	130.57	1.0473	141.45	0.9825
LDL - Linear (Age)	-0.04	0.0858	-1.98	0.0890
LDL - Quadratic (Age)	-0.03	0.0074	-0.03	0.0075
HDL - Intercept	64.49	0.5595	41.56	0.2700
HDL - Linear (Age)	-0.57	0.0355	0.05	0.0193
HDL - Quadratic (Age)	-0.01	0.0032	-0.00	0.0017
Triglyc. - Intercept	4.55	0.0127	4.99	0.0125
Triglyc. - Linear (Age)	0.02	0.0009	0.01	0.0009
Triglyc. - Quadratic (Age)	-0.00	0.0001	-0.00	0.0001
Variance (Random Int - LDL)	675.90	40.1351	804.85	42.4745
Variance (Random Slope - LDL)	1.47	0.2945	2.53	0.3398
Variance (Random Int - HDL)	162.57	9.2756	44.37	2.3423
Variance (Random Int - Triglyc)	0.09	0.0056	0.12	0.0060
Variance (Error - LDL)	443.47	12.9086	502.510	12.9419
Variance (Error - HDL)	93.90	2.8177	33.059	0.8661
Variance (Error - Triglyc)	0.075	0.0019	0.097	0.0021
Corr (Rnd Int LDL/Slope LDL)	-0.499	0.0406	⟶	
Class-Membership Parameters:				
- Intercept	0.0	-	26.439	4.7331
- Ever Smoked Cigarettes?	0.0	-	0.919	0.5215
- Exercise or Play Sports?	0.0	-	-0.536	0.6093
- Ever Been Treated for High Blood Pressure?	0.0	-	-0.549	0.5540
- Ever Been Treated for High Cholesterol?	0.0	-	2.608	0.6250
- Baseline BMI	0.0	-	-0.073	0.0553
- Baseline HDL Cholesterol	0.0	-	-0.672	0.0988
- Baseline Glucose	0.0	-	-0.011	0.0125
- Baseline Triglycerides	0.0	-	0.062	0.0106
- Baseline Heart Rate	0.0	-	0.066	0.0285
- Number of Glasses of Wine Per Week	0.0	-	-0.129	0.0791

LMMs to the data for each of the classes. The fitted odds ratios of being classified in the At-Risk class versus the Optimal class are shown in Table 3.12, including point estimates and 95% confidence intervals. Note that baseline LDL was not included in this model since LDL cholesterol values can be modified with medication use. Therefore, the indicator for whether a subject was treated for high cholesterol was included instead.

Table 3.12: Example 3 - Odds Ratios

	Odds Ratio of Being in At-Risk Class	
Covariate	Point Estimate	95% CI
Ever Been Treated for High Cholesterol?	13.57	3.99 - 46.20
Ever Smoked Cigarettes?	2.51	0.90 - 6.97
Baseline Heart Rate	1.07	1.01 - 1.13
Baseline Triglycerides	1.06	1.04 - 1.09
Baseline Glucose	0.99	0.97 - 1.01
Baseline BMI	0.93	0.83 - 1.04
Number of Glasses of Wine Per Week	0.88	0.75 - 1.03
Exercise or Play Sports?	0.59	0.18 - 1.93
Ever Been Treated for High Blood Pressure?	0.58	0.19 - 1.71
Baseline HDL Cholesterol	0.51	0.42 - 0.62

As you might expect, the indicator for whether a subject was treated for high cholesterol is the most significant of the covariates, with subjects treated for high cholesterol found to have 13.57 times the odds of being in the At-Risk class as subjects not treated for high cholesterol. Subjects who have smoked were found to have 2.5 times the odds of being in the At-Risk class, although this was not quite statistically significant. Increases by 1 unit in baseline heart rate and triglycerides were associated with a 6-7% increase in the odds of being in the At-Risk class. Baseline glucose and BMI did not have a statistically significant result. Exercising/playing sports and drinking wine were associated with decreases in one's odds of being in the At-Risk class (again not statistically significant). Finally, the odds of being classified in the At-Risk class halved for each single-unit increase in baseline HDL cholesterol level - a statistically significant and powerful association!

Lastly, a note of caution is warranted related to the use of the logistic model for determination of class probabilities. When comparing the plots for the 2-class models from this section with those from Chapter 1, it is apparent that they are nearly identical. When distinguishable patterns are present in the longitudinal data (as there are here), and the class membership parameters are sufficiently informative so as to differentiate the subjects, then the resulting models would be expected to be comparable. In such a case, the model from this section has the added benefit of simultaneously obtaining estimates for the parameters in the class membership model. However, when the class membership model is uninformative and there is little information upon which to differentiate subjects, this model has the potential to fit the data for certain subjects poorly - in fact worse than had the LMM been run. As an example, the LCLMM in the above example was refit with only an intercept and baseline BMI in the class membership model. While the likelihood values do improve slightly when compared with the LMM, the within-subject MSEs for this model worsen. The within-subject MSE for this model was 49.9 for HDL cholesterol (vs. just 47.7 for the LMM), 381.1 for LDL cholesterol (vs. just 365.1 for the LMM), and 0.074 for triglycerides (vs. 0.073 for the LMM). Given the simplicity of the class membership model, the class probabilities for each individual will be fit solely based on baseline BMI. However, given so little information, it is likely that some subjects will be misclassified. For example, selected subjects may have a high BMI and are therefore classified into the At-Risk class while all of their other characteristics may indicate a healthy lifestyle. In these cases, the subject will receive a fitted value representative of the At-Risk class even though their longitudinal data may be representative of the Optimal class. One way to examine whether this is the case is to compare the fitted class probabilities based on the class membership model with the relative fit of each underlying LMM (see Equation 3.10). Cross-tabulations of the two sets of class probabilities are shown in Tables 3.13 and 3.14 for the more informative and less informative class membership models, respectively. In the more informative case, more than 1,700 of the 2,066 subjects have similar class probabilities based on both sets of probabilities. This is not the case in the less informative case, where only about 600 subjects have similar class probabilities. In fact, in the less informative case, there are 165 subjects who have a 70-100% probability of being in the At-Risk class based on the logistic model, but would in fact be fit better under the Optimal class.

This example, in which the ARIC data was reanalyzed with a structured class-membership model, demonstrates the utility as well as the potential pitfalls of the approach. While the model provides the researcher with added insight into the factors which drive subjects to classes, it also may lead to a poor model fit for selected subjects. In cases where the class probabilities based on the logistic model and the relative fit of the underlying LMMs do not agree, the statistician should limit interpretation to the subpopulation level and avoid commentary on individual trajectories.

Table 3.13: Example 3 - Cross-Tabulation of Class Probabilities - More Informative Class Membership Model

		Prob(AtRisk) Based on Relative Fit			
Frequency Percent Row Percent Column Percent		0%-30%	30%-70%	70%-100%	Total
Prob(AtRisk) Based on Logistic Model	0-30%	720 34.85 85.61 91.49	96 4.65 11.41 39.67	25 1.21 2.97 2.41	841 40.71
	30%-70%	28 1.36 24.35 3.56	38 1.84 33.04 15.70	49 2.37 42.61 4.73	115 5.57
	70%-100%	39 1.89 3.51 4.96	108 5.23 9.73 44.63	963 46.61 86.76 92.86	1110 53.73
	Total	787 38.09	242 11.71	1037 50.19	2066 100.00

Table 3.14: Example 3 - Cross-Tabulation of Class Probabilities - Less-Informative Class Membership Model

		Prob(AtRisk) Based on Relative Fit			
Frequency Percent Row Percent Column Percent		0%-30%	30%-70%	70%-100%	Total
Prob(AtRisk) Based on Logistic Model	0-30%	8	0	0	8
		0.39	0.00	0.00	0.39
		100.00	0.00	0.00	
		1.15	0.00	0.00	
	30%-70%	522	151	695	1368
		25.27	7.31	33.64	66.21
		38.16	11.04	50.80	
		75.11	66.23	60.80	
	70%-100%	165	77	448	690
		7.99	3.73	21.68	33.40
		23.91	11.16	64.93	
		23.74	33.77	39.20	
	Total	695	228	1143	2066
		33.64	11.04	55.32	100.00

3.7 Conclusions and Further Research

The Latent Class Linear Mixed Model provides a flexible approach to more effectively model longitudinal data in the presence of unknown subpopulations. In addition to the usual assumptions necessary to run the linear mixed model, the statistician must make several additional ones. These include specifying the number of latent classes, how class membership should be modeled, and which parameters are permitted to vary by latent class. The resulting parameter estimates offer the statistician important information related to the makeup of the subpopulations as well as their associated trajectories. The examples presented in this chapter illustrate possible uses of the methods and can serve as a reference for statisticians as they begin to utilize these methods in their every day research.

Future research should examine a hybrid of the relative-fit and structured class membership models. Since there may be instances in which the class membership model fits a subject as being in one class while the subject may have longitudinal data which would be better fit by another, a model which fits each subject's class probabilities based on a combination of the two would offer an intermediate and perhaps more reliable result.

Appendix A

Quantities Used in Derivative Calculations

$$
\begin{aligned}
\mathbf{x}_{ij} &= j^{th} \text{ column of } \mathbf{X}_i \\
\mathbf{w}_{ij} &= j^{th} \text{ column of } \mathbf{W}_i \\
\mathbf{e}_{ik} &= \mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{W}_i\boldsymbol{\lambda}_k \\
\boldsymbol{\Sigma}_{ik} &= \mathbf{Z}_i\mathbf{D}_k\mathbf{Z}_i' + \mathbf{R}_{ik} \\
M_{1ik} &= -\frac{1}{2}\mathbf{e}_{ik}'\boldsymbol{\Sigma}_{ik}^{-1}\mathbf{e}_{ik} \\
\frac{\partial M_{1ik}}{\partial \boldsymbol{\beta}} &= \mathbf{X}_i'\boldsymbol{\Sigma}_{ik}^{-1}\mathbf{e}_{ik} \\
\frac{\partial M_{1ik}}{\partial \boldsymbol{\lambda}_k} &= \mathbf{W}_i'\boldsymbol{\Sigma}_{ik}^{-1}\mathbf{e}_{ik} \\
\left[\frac{\partial M_{1ik}}{\partial \boldsymbol{\theta}}\right]_r &= \frac{1}{2}\mathbf{e}_{ik}'\boldsymbol{\Sigma}_{ik}^{-1}\frac{\partial \boldsymbol{\Sigma}_{ik}}{\partial \boldsymbol{\theta}_r}\boldsymbol{\Sigma}_{ik}^{-1}\mathbf{e}_{ik} \\
\left[\frac{\partial^2 M_{1ik}}{\partial \boldsymbol{\beta}\partial \boldsymbol{\theta}}\right]_{jr} &= -\mathbf{x}_{ij}'\boldsymbol{\Sigma}_{ik}^{-1}\frac{\partial \boldsymbol{\Sigma}_{ik}}{\partial \boldsymbol{\theta}_r}\boldsymbol{\Sigma}_{ik}^{-1}\mathbf{e}_{ik} \\
\left[\frac{\partial^2 M_{1ik}}{\partial \boldsymbol{\lambda}_k\partial \boldsymbol{\theta}}\right]_{jr} &= -\mathbf{w}_{ij}'\boldsymbol{\Sigma}_{ik}^{-1}\frac{\partial \boldsymbol{\Sigma}_{ik}}{\partial \boldsymbol{\theta}_r}\boldsymbol{\Sigma}_{ik}^{-1}\mathbf{e}_{ik} \\
\left[\frac{\partial^2 M_{1ik}}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}}\right]_{rs} &= \frac{1}{2}\mathbf{e}_{ik}'\left\{\begin{array}{l} \boldsymbol{\Sigma}_{ik}^{-1}\left[\left(-\frac{\partial \boldsymbol{\Sigma}_{ik}}{\partial \boldsymbol{\theta}_r}\boldsymbol{\Sigma}_{ik}^{-1}\frac{\partial \boldsymbol{\Sigma}_{ik}}{\partial \boldsymbol{\theta}_s}\boldsymbol{\Sigma}_{ik}^{-1}\right) + \left(\frac{\partial^2 \boldsymbol{\Sigma}_{ik}}{\partial \boldsymbol{\theta}_r\partial \boldsymbol{\theta}_s}\boldsymbol{\Sigma}_{ik}^{-1}\right)\right] \\ + \left[-\boldsymbol{\Sigma}_{ik}^{-1}\frac{\partial \boldsymbol{\Sigma}_{ik}}{\partial \boldsymbol{\theta}_s}\boldsymbol{\Sigma}_{ik}^{-1}\frac{\partial \boldsymbol{\Sigma}_{ik}}{\partial \boldsymbol{\theta}_r}\boldsymbol{\Sigma}_{ik}^{-1}\right] \end{array}\right\}\mathbf{e}_{ik} \\
M_{2ik} &= (2\pi)^{n_i/2}\left|\boldsymbol{\Sigma}_{ik}\right|^{1/2} \\
\left[\frac{\partial M_{2ik}}{\partial \boldsymbol{\theta}}\right]_r &= \frac{1}{2}(2\pi)^{n_i/2}\left|\boldsymbol{\Sigma}_{ik}\right|^{1/2}\operatorname{tr}\left(\boldsymbol{\Sigma}_{ik}^{-1}\frac{\partial \boldsymbol{\Sigma}_{ik}}{\partial \boldsymbol{\theta}_r}\right) \\
\left[\frac{\partial^2 M_{2ik}}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}}\right]_{rs} &= \frac{1}{2}(2\pi)^{n_i/2}\left\{\begin{array}{l} \Big[\left|\boldsymbol{\Sigma}_{ik}\right|^{1/2} tr\,\Big(\boldsymbol{\Sigma}_{ik}^{-1}\frac{\partial^2 \boldsymbol{\Sigma}_{ik}}{\partial \boldsymbol{\theta}_r\partial \boldsymbol{\theta}_s} - \boldsymbol{\Sigma}_{ik}^{-1}\frac{\partial \boldsymbol{\Sigma}_{ik}}{\partial \boldsymbol{\theta}_s}\boldsymbol{\Sigma}_{ik}^{-1}\frac{\partial \boldsymbol{\Sigma}_{ik}}{\partial \boldsymbol{\theta}_r}\Big)\Big] \\ + \Big[\frac{1}{2}\left|\boldsymbol{\Sigma}_{ik}\right|^{1/2}\ \operatorname{tr}\ \Big(\boldsymbol{\Sigma}_{ik}^{-1}\frac{\partial \boldsymbol{\Sigma}_{ik}}{\partial \boldsymbol{\theta}_s}\Big)\ \operatorname{tr}\ \Big(\boldsymbol{\Sigma}_{ik}^{-1}\frac{\partial \boldsymbol{\Sigma}_{ik}}{\partial \boldsymbol{\theta}_r}\Big)\Big] \end{array}\right\} \\
\mathbf{R}_{ik} &= \left[M_{2ik}\exp\left(M_{1ik}\right)\frac{\partial M_{1ik}}{\partial \boldsymbol{\theta}} - \exp\left(M_{1ik}\right)\frac{\partial M_{2ik}}{\partial \boldsymbol{\theta}}\right] \\
\frac{\partial \mathbf{R}_{ik}}{\partial \boldsymbol{\beta}} &= M_{2ik}\exp\left(M_{1ik}\right)\left[\frac{\partial M_{1ik}}{\partial \boldsymbol{\beta}}\frac{\partial M_{1ik}}{\partial \boldsymbol{\theta}} + \frac{\partial^2 M_{1ik}}{\partial \boldsymbol{\theta}\partial \boldsymbol{\beta}}\right] - \exp\left(M_{1ik}\right)\frac{\partial M_{1ik}}{\partial \boldsymbol{\beta}}\frac{\partial M_{2ik}}{\partial \boldsymbol{\theta}} \\
\frac{\partial \mathbf{R}_{ik}}{\partial \boldsymbol{\lambda}_k} &= M_{2ik}\exp\left(M_{1ik}\right)\left[\frac{\partial M_{1ik}}{\partial \boldsymbol{\lambda}_k}\frac{\partial M_{1ik}}{\partial \boldsymbol{\theta}} + \frac{\partial^2 M_{1ik}}{\partial \boldsymbol{\theta}\partial \boldsymbol{\lambda}_k}\right] - \exp\left(M_{1ik}\right)\frac{\partial M_{1ik}}{\partial \boldsymbol{\lambda}_k}\frac{\partial M_{2ik}}{\partial \boldsymbol{\theta}}
\end{aligned}
$$

Appendix B

Derivatives for the Gradient and Hessian-Based Algorithms

The derivatives of the quantities Q_{1i}, Q_{2ik}, and Q_{3ik} (defined in Equation 1.22) with respect to the parameters are presented below. Note that in cases where the variances are class-specific, some first and second derivatives will be equal to $\mathbf{0}$.

First Derivatives:

$$
\begin{aligned}
\frac{\partial Q_{1i}}{\partial \boldsymbol{\beta}} &= \sum_{k=1}^{K} \pi_{ik} \frac{\partial Q_{3ik}}{\partial \boldsymbol{\beta}} \\
\frac{\partial Q_{1i}}{\partial \boldsymbol{\lambda}_k} &= \pi_{ik} \frac{\partial Q_{3ik}}{\partial \boldsymbol{\lambda}_k} \\
\frac{\partial Q_{1i}}{\partial \boldsymbol{\theta}} &= \sum_{k=1}^{K} \pi_{ik} \frac{\partial Q_{3ik}}{\partial \boldsymbol{\theta}} \\
\frac{\partial Q_{1i}}{\partial \boldsymbol{\alpha}_k} &= \sum_{j=1}^{K} Q_{3ij} \frac{\partial Q_{2ij}}{\partial \boldsymbol{\alpha}_k} \\
\frac{\partial Q_{2ij}}{\partial \boldsymbol{\alpha}_k} &= \begin{cases} \mathbf{t}_i \times \pi_{ik} \left(1 - \pi_{ik}\right) & \text{if } j = k \\ \mathbf{t}_i \times \left(-\pi_{ij} \pi_{ik}\right) & \text{if } j \neq k \end{cases} \\
\frac{\partial Q_{3ik}}{\partial \boldsymbol{\beta}} &= Q_{3ik} \times \left(\mathbf{X}_i' \boldsymbol{\Sigma}_{ik}^{-1} \mathbf{e}_{ik}\right) \\
\frac{\partial Q_{3ik}}{\partial \boldsymbol{\lambda}_k} &= Q_{3ik} \times \left(\mathbf{W}_i' \boldsymbol{\Sigma}_{ik}^{-1} \mathbf{e}_{ik}\right) \\
\frac{\partial Q_{3ik}}{\partial \boldsymbol{\theta}} &= \frac{\mathbf{R}_{ik}}{\left(M_{2ik}\right)^2}
\end{aligned}
$$

Second Derivatives:

$$
\begin{aligned}
\frac{\partial^2 Q_{2ij}}{\partial \boldsymbol{\alpha}_r \partial \boldsymbol{\alpha}_p} &= \begin{cases}
\mathbf{t}_i \mathbf{t}_i' \times (1 - 2\pi_{ij}) \left(\pi_{ij} - \pi_{ij}^2\right) & \text{if } j = r = p \\
\mathbf{t}_i \mathbf{t}_i' \times (1 - 2\pi_{ij}) (-\pi_{ij}\pi_{ip}) & \text{if } j = r \neq p \\
\mathbf{t}_i \mathbf{t}_i' \times -\pi_{ij} \left[\left(\pi_{ir} - \pi_{ir}^2\right) - \pi_{ir}\pi_{ip}\right] & \text{if } p = r \neq j \\
\mathbf{t}_i \mathbf{t}_i' \times (1 - 2\pi_{ij}) (-\pi_{ij}\pi_{ir}) & \text{if } j = p \neq r \\
\mathbf{t}_i \mathbf{t}_i' \times 2\pi_{ij}\pi_{ir}\pi_{ip} & \text{if } j \neq r \neq p
\end{cases} \\
\frac{\partial^2 Q_{3ik}}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}} &= Q_{3ik} \times \left(-\mathbf{X}_i' \boldsymbol{\Sigma}_{ik}^{-1} \mathbf{X}_i\right) + Q_{3ik} \times \left(\mathbf{X}_i' \boldsymbol{\Sigma}_{ik}^{-1} \mathbf{e}_{ik}\right) \left(\mathbf{X}_i' \boldsymbol{\Sigma}_{ik}^{-1} \mathbf{e}_{ik}\right)' \\
\frac{\partial^2 Q_{3ik}}{\partial \boldsymbol{\lambda}_k \partial \boldsymbol{\lambda}_j} [\text{for } j = k] &= Q_{3ik} \times \left(-\mathbf{W}_i' \boldsymbol{\Sigma}_{ik}^{-1} \mathbf{W}_i\right) + Q_{3ik} \times \left(\mathbf{W}_i' \boldsymbol{\Sigma}_{ik}^{-1} \mathbf{e}_{ik}\right) \left(\mathbf{W}_i' \boldsymbol{\Sigma}_{ik}^{-1} \mathbf{e}_{ik}\right)' \\
[\text{for } j \neq k] &= \mathbf{0} \\
\frac{\partial^2 Q_{3ik}}{\partial \boldsymbol{\lambda}_k \partial \boldsymbol{\beta}} &= Q_{3ik} \times \left(-\mathbf{W}_i' \boldsymbol{\Sigma}_{ik}^{-1} \mathbf{X}_i\right) + Q_{3ik} \times \left(\mathbf{W}_i' \boldsymbol{\Sigma}_{ik}^{-1} \mathbf{e}_{ik}\right) \left(\mathbf{X}_i' \boldsymbol{\Sigma}_{ik}^{-1} \mathbf{e}_{ik}\right)' \\
\frac{\partial^2 Q_{3ik}}{\partial \boldsymbol{\theta} \partial \boldsymbol{\beta}} &= \frac{\partial \mathbf{R}_{ik}}{\partial \boldsymbol{\beta}} / \left(M_{2ik}\right)^2 \\
\frac{\partial^2 Q_{3ik}}{\partial \boldsymbol{\theta} \partial \boldsymbol{\lambda}_k} &= \frac{\partial \mathbf{R}_{ik}}{\partial \boldsymbol{\lambda}_k} / \left(M_{2ik}\right)^2 \\
\frac{\partial^2 Q_{3ik}}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}} &= 2 \times \ \exp\left(M_{1ik}\right) \left(M_{2ik}\right)^{-3} \frac{\partial M_{2ik}}{\partial \boldsymbol{\theta}} \frac{\partial M_{2ik}}{\partial \boldsymbol{\theta}} \\
&- \ \exp\left(M_{1ik}\right) \left(M_{2ik}\right)^{-2} \frac{\partial^2 M_{2ik}}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}} \\
&- \ \exp\left(M_{1ik}\right) \frac{\partial M_{1ik}}{\partial \boldsymbol{\theta}} \left(M_{2ik}\right)^{-2} \frac{\partial M_{2ik}}{\partial \boldsymbol{\theta}} \\
&+ \ \left(M_{2ik}\right)^{-1} \frac{\partial M_{1ik}}{\partial \boldsymbol{\theta}} \ \exp\left(M_{1ik}\right) \frac{\partial M_{1ik}}{\partial \boldsymbol{\theta}} \\
&+ \ \left(M_{2ik}\right)^{-1} \frac{\partial^2 M_{1ik}}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}} \ \exp\left(M_{1ik}\right) \\
&- \ \left(M_{2ik}\right)^{-2} \frac{\partial M_{2ik}}{\partial \boldsymbol{\theta}} \frac{\partial M_{1ik}}{\partial \boldsymbol{\theta}} \ \exp\left(M_{1ik}\right)
\end{aligned}
$$

Appendix C

Derivatives for the Random Effects Calculations

First Derivatives:

$$\frac{\partial P_{3ik}}{\partial \mathbf{b}_{ik}} = P_{3ik} \times \left[\mathbf{Z}_i' \mathbf{R}_{ik}^{-1} \left(\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\beta} - \mathbf{W}_i \boldsymbol{\lambda}_k - \mathbf{Z}_i \mathbf{b}_{ik}\right) - \mathbf{D}_k^{-1} \mathbf{b}_{ik}\right]$$

Second Derivatives:

$$\begin{aligned} \frac{\partial P_{3ik}}{\partial \mathbf{b}_{ik} \partial \mathbf{b}_{ik}} &= P_{3ik} \times \left\{ \left[-\mathbf{Z}_i' \mathbf{R}_{ik}^{-1} \mathbf{Z}_i - \mathbf{D}_k^{-1}\right] + \right. \\ & \left[\mathbf{Z}_i' \mathbf{R}_{ik}^{-1} \left(\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\beta} - \mathbf{W}_i \boldsymbol{\lambda}_k - \mathbf{Z}_i \mathbf{b}_{ik}\right) - \mathbf{D}_k^{-1} \mathbf{b}_{ik}\right] \\ & \left. \left[\mathbf{Z}_i' \mathbf{R}_{ik}^{-1} \left(\mathbf{y}_i - \mathbf{X}_i \boldsymbol{\beta} - \mathbf{W}_i \boldsymbol{\lambda}_k - \mathbf{Z}_i \mathbf{b}_{ik}\right) - \mathbf{D}_k^{-1} \mathbf{b}_{ik}\right]' \right\} \end{aligned}$$

Appendix D

Proof of Hathaway's Condition with Regard to Variance Constraints

Recall Hathaway's Condition (HC) was: all eigenvalues of $\mathbf{\Sigma}_j \mathbf{\Sigma}_{j*}^{-1}$ must be $\geq$ some $c > 0$.

Theorem: Bounding the variances of the random effects in $\mathbf{D}_j$ and the residual error variances in $\mathbf{R}_j$ where $\mathbf{\Sigma}_j = \mathbf{Z}\mathbf{D}_j\mathbf{Z} + \mathbf{R}_j$, as proposed in Section 1.4.2, results in a variance-covariance matrix $\mathbf{\Sigma}_j$ which satisfies Hathaway's condition.

Proof:

$\mathbf{\Sigma}_j$ is symmetric implies

$$\begin{aligned}
\mathbf{\Sigma}_j &= \mathbf{\Sigma}_j^T \\
\left(\mathbf{\Sigma}_j^{-1}\right)^T &= \mathbf{\Sigma}_j^{-1} \\
\left(\mathbf{\Sigma}_j \mathbf{\Sigma}_{j*}^{-1}\right)^T &= \left(\mathbf{\Sigma}_{j*}^{-1}\right)^T \left(\mathbf{\Sigma}_j\right)^T = \mathbf{\Sigma}_{j*}^{-1}\mathbf{\Sigma}_j.
\end{aligned}$$

Being symmetric, $\mathbf{\Sigma}_j \mathbf{\Sigma}_{j*}^{-1}$ and $\mathbf{\Sigma}_{j*}^{-1}\mathbf{\Sigma}_j$ have the same spectral decomposition. Thus, if $\mathbf{u}$ and λ are an eigenvector and eigenvalue pair, then

$$\begin{aligned}
\mathbf{\Sigma}_j \mathbf{\Sigma}_{j*}^{-1}\mathbf{u} &= \lambda \mathbf{u} \\
\mathbf{\Sigma}_{j*}^{-1}\mathbf{\Sigma}_j \mathbf{u} &= \lambda \mathbf{u} \\
\mathbf{\Sigma}_j \mathbf{u} &= \lambda \mathbf{\Sigma}_{j*}\mathbf{u}
\end{aligned}$$

Alternatively, one could write

$$\frac{\mathbf{u}^T\mathbf{\Sigma}_j\mathbf{u}}{\mathbf{u}^T\mathbf{\Sigma}_{j*}\mathbf{u}} = \lambda$$

$$\frac{\mathbf{u}^T\mathbf{\Sigma}_{j*}^{-1}\mathbf{\Sigma}_j\mathbf{u}}{\mathbf{u}^T\mathbf{u}} = \lambda = \frac{\mathbf{u}^T\mathbf{\Sigma}_j\mathbf{\Sigma}_{j*}^{-1}\mathbf{u}}{\mathbf{u}^T\mathbf{u}}$$

It is known that

$$\underset{\mathbf{u}\neq\mathbf{0}}{\text{minimum}}\ \left\{\frac{\mathbf{u}^T\mathbf{\Sigma}_j\mathbf{u}}{\mathbf{u}^T\mathbf{\Sigma}_{j*}\mathbf{u}}\right\} = \lambda_n = \text{ the smallest eigenvalue of } \mathbf{\Sigma}_{j*}^{-1}\mathbf{\Sigma}_j$$

$$\boxed{\text{Thus, HC is } R_{j,j*}(\mathbf{x}) = \frac{\mathbf{x}^T\mathbf{\Sigma}_j\mathbf{x}}{\mathbf{x}^T\mathbf{\Sigma}_{j*}\mathbf{x}} \geq \text{ c } \forall\mathbf{x}\neq\mathbf{0}}$$

If $\mathbf{\Sigma}_j$ is of the form $\mathbf{\Sigma}_j = \mathbf{Z}\mathbf{D}_j\mathbf{Z} + \mathbf{R}_j$, then $\boldsymbol{\Sigma_j}$ has a linear covariance structure and

$$\mathbf{\Sigma}_j = \textstyle\sum_{h=1}^{H}\mathbf{G}_h\boldsymbol{\tau}_{jh},$$

in which $\mathbf{G}_h$ is a known positive semidefinite (PSD) matrix and $\boldsymbol{\tau}_j = [\boldsymbol{\tau}_{j1}\cdots\boldsymbol{\tau}_{jH}]$ are the variance-covariance parameters for class j.

Then $R_{j,j*}(\mathbf{x}) = \dfrac{\sum\limits_{h=1}^{H}(\mathbf{x}_T\mathbf{G}_h\mathbf{x})\,\boldsymbol{\tau}_{jh}}{\sum\limits_{h=1}^{H}(\mathbf{x}_T\mathbf{G}_h\mathbf{x})\,\boldsymbol{\tau}_{j*h}} = \frac{\mathbf{w}^T\boldsymbol{\tau}_j}{\mathbf{w}^T\boldsymbol{\tau}_{j*}}$ in which $\mathbf{w}_h = \frac{\mathbf{x}^T\mathbf{G}_h\mathbf{x}}{\mathbf{x}^T\mathbf{x}} \geq 0$ by the PSD property of $\mathbf{G}_h$.

Thus HC is $\frac{\mathbf{w}^T\boldsymbol{\tau}_j}{\mathbf{w}^T\boldsymbol{\tau}_{j*}} \geq$ c $\forall\mathbf{w}\geq\mathbf{0}$ (i.e. $w_h \geq 0\ \forall h$) or equivalently $0 \leq \mathbf{w}^T(\boldsymbol{\tau}_j - c\boldsymbol{\tau}_{j*})\,\forall\mathbf{w}$ such that $w_h \geq 0\ \forall h$

Hathaway's condition is satisfied by the proposed bounds on both the variances in $\mathbf{D}_j$ and the residual error variances in $\mathbf{R}_j$.

Bibliography

Aitkin, M. and Aitkin, I. (1996). A Hybrid EM/Gauss-Newton Algorithm for Maximum Likelihood in Mixture Distributions. Statistics and Computing 6, 127-130.

ARIC Investigators (1989). The Atherosclerosis Risk in Communities (ARIC) Study: design and objectives. Am J Epidemiol 129, 687-702.

Arcidiacono, P. and Jones, J. B. (2003). Finite Mixture Distributions, Sequential Likelihood and the EM Algorithm. Econometrica 71, 933-946.

Ballantyne, C. M. (1998). Low-Density Lipoproteins and Risk for Coronary Artery Disease. American Journal of Cardiology 82, 3Q-12Q.

Banfield, J. D. and Raftery, A. E. (1993). Model-Based Gaussian and Non-Gaussian Clustering. Biometrics 49, 803-821.

Basford, K. E. and McLachlan, G. J. (1985). Likelihood Estimation with Normal Mixture Models. Applied Statistics 34, 282-289.

Beale, E. M. L. (1972). A Derivation of Conjugate Gradients. In: Lootsma, F. A. (editor). Numerical Methods for Nonlinear Optimization. London: Academic Press.

Biernacki, C., and Govaert, G. (1999). Choosing Models in Model-Based Clustering and Discriminant Analysis. Journal of Statistical Computation and Simulation 64, 49-71.

Biernacki, C., Gilles, C., and Govaert, G. (2000). Assessing a Mixture Model for Clustering with the Integrated Completed Likelihood. IEEE Transactionson Pattern Analysis and Machine Intelligence 22, 719-725.

Biernacki, C., Celeux, G., and Govaert, G. (2003). Choosing Starting Values for the EM Algorithm for Getting the Highest Likelihood in Multivariate Gaussian Mixture Models. Computational Statistics and Data Analysis 41, 561-575.

Biernacki, C., Gilles, C., Govaert, G., and Langrognet, F. (2005). Model-Based Cluster and Discriminant Analysis with the MIXMOD Software. Preprint submitted to Elsevier Science.

Bohning, D., and Seidel, W. (2003). Editorial: Recent Developments in Mixture Models. Computational Statistics and Data Analysis 41, 349-357.

Borghese, I. F., Minard, K. L., and Thoman, E. B. (1995). Sleep rhythmicity in premature infants: implications for developmental status. Sleep 18, 523-530.

Bozdogan, H. (1987). Model Selection and Akaikes Information Criterion (AIC): The General Theory and Its Analytical Extensions. Psychometrika 52, 345370.

Broyden, C. G. (1965). A class of methods for solving nonlinear simultaneous equations. Math. Computat. 19, 577-593.

Celeux, G. and Govaert, G. (1992). A Classification EM Algorithm for Clustering and Two Stochostic Versions. Computational Statistics and Data Analysis 14, 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 Modeling 5, 243-267.

Coleman, D. A. and Woodruff, D. L. (2000). Cluster Analysis for Large Datasets: An Effective Algorithm for Maximizing the Mixture Likelihood. Journal of Computational and Graphical Statistics 9, 672-688.

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, Series B (Methodological), 39, 1-38.

Edwards, L. J. (2000). Modern Statistical Techniques for the Analysis of Longitudinal Data in Biomedical Research. Pediatric Pulmonology 30, 330-344.

Everitt, B. S. (1984). Maximum Likelihood Estimation of the Parameters in a Mixture of Two Univariate Normal Distributions: A Comparison of Different Algorithms. The Statistician 33, 205-215.

Fraley, C. and Raftery, A. E. (2002). Model-Based Clustering, Discriminant Analysis, and Density Estimation. Journal of the American Statistical Association 97, 611-631.

Fraley, C., Raftery, A., and Wehrens, R. (2003). Incremental Model-Based Clustering for Large Datasets With Small Clusters. Technical Report 439, Department of Statistics, University of Washington.

Fraley, C. and Raftery, A. E. (2006). MCLUST Version 3: An R Package for Normal Mixture Modeling and Model-Based Clustering. Technical Report 504, Department of Statistics, University of Washington.

Gertner, S., Greenbaum, C. W., Sadeh, A., Dolfin, A., Sirota, L., and Ben-Nun, Y. (2002). Sleep-wake patterns in preterm infants and 6 month's home environment: implications for early cognitive development. Early Hum Dev 68, 93-102.

Gotto, A. M., Jr. (1998). Triglyceride as a Risk Factor for Coronary Artery Disease. American Journal of Cardiology 82, 22Q-25Q.

Grebe. M. T., Schoene, E., Schaefer, C. A., Boedeker, R. H., Kemkes-Matthes, B., Voss, R., and Tillmanns, H. H. (2007). Elevated Lipoprotein(a) does not promote early atherosclerotic changes of the carotid arteries in young, healthy adults. Atherosclerosis 190, 194-198.

Grun, B. and Leisch, F. (2007). FlexMix: An R Package for Finite Mixture Modelling. R News 7, 8-13.

Gusnanto, A., Ploner, A., and Pawitan, Y. (2005). Fold-Change Estimation of Differentially Expressed Genes Using Mixture Mixed Model. Statistical Applications in Genetics and Molecular Biology 4, 1-22.

Halpern, L. F., MacLean W. E., and Baumeister, A. A. (1995). Infant sleep-wake characteristics: relation to neurological status and the prediction of developmental outcome. Dev Rev 15, 255-291.

Hathaway, R. J. (1985). A Constrained Formulation of Maximum-Likelihood Estimation for Normal Mixture Distributions. The Annals of Statistics 13, 795-800.

Holditch-Davis, D., Scher, M., Schwartz, T., and Hudson-Barr, D. (2004). Sleeping and waking state development in preterm infants. Early Human Development 80, 43-64.

Ingrassia, S. (2004). A likelihood-based constrained algorithm for multivariate normal mixture models. Statistical Methods and Applications 13, 151-166.

Jamshidian, M. and Jennrich, R. I. (1993). Conjugate Gradient Acceleration of the EM Algorithm. Journal of the American Statistical Association 88, 221-228.

Jamshidian, M. and Jennrich, R. I. (1997). Acceleration of the EM Algorithm by Using Quasi-Newton Methods. Journal of the Royal Statistical Society, Series B 59, 569-587.

Jamshidian, M. and Jennrich, R. I. (2000). Standard Errors for EM Estimation. Journal of the Royal Statistical Society Series B 62, 257-270.

Jennrich, R. I. and Schluchter, M. D. (1986). Unbalanced Repeated-Measures Models with Structured Covariance Matrices. Biometrics 42, 805-820.

Karlis, D. and Xekalaki, E. (2003). Choosing Initial Values for the EM Algorithm for Finite Mixtures. Computational Statistics and Data Analysis 41, 577-590.

Kwiterovich, P. O., Jr. (1998). The Antiatherogenic Role of High-Density Lipoprotein Cholesterol. American Journal of Cardiology 82, 13Q-21Q.

Laird, N. M. and Ware, J. H. (1982). Random-Effects Models for Longitudinal Data. Biometrics 38, 963-974.

Laird, N., Lange, N., and Stram, D. (1987). Maximum Likelihood Computations with Repeated Measures: Application of the EM Algorithm. Journal of the American Statistical Association 82, 97-105.

Lange, K. (1995). A Quasi-Newton Acceleration of the EM Algorithm. Statistica Sinica 5, 1-18.

Lange, K. (1995). A Gradient Algorithm Locally Equivalent to the EM Algorithm. Journal of the Royal Statistical Society, Series B 57, 425-437.

Li, J. (2005). Clustering Based on a Multilayer Mixture Model. Journal of Computational and Graphical Statistics 14, 547-568.

Lin, H., McCulloch, C. E., Turnbull, B. W., Slate, E. H., and Clark, L. C. (2000). A Latent Class Mixed Model for Analyzing Biomarker Trajectories with Irregularly Scheduled Observations. Statistics in Medicine 19, 1303-1318.

Lin, H., Turnbull, B. W., McCulloch, C. E., and Slate, E. H. (2002). Latent Class Models for Joint Analysis of Longitudinal Biomarker and Event Process Data: Application to Longitudinal Prostate-Specific Antigen Readings and Prostate Cancer. Journal of the American Statistical Association 97, 53-65.

Lin, H., McCulloch, C. E., and Mayne, S. T. (2002). Maximum likelihood estimation in the joint analysis of time-to-event and multiple longitudinal variables. Statistics in Medicine 21, 2369-2382.

Lin, H., McCulloch, C. E., and Rosenheck, R. A. (2004). Latent Pattern Mixture Models for Informative Intermittent Missing Data in Longitudinal Studies. Biometrics 60, 295-305.

Louis, T. A. (1982). Finding the Observed Information Matrix when Using the EM Algorithm. Journal of the Royal Statistical Society Series B 44, 226-233.

McCulloch, C. E. (1998). Review of the EM Algorithm and Its Extensions. Journal of the American Statistical Association 93, 403-404.

McCulloch, C. E. (2003). Joint Models in the Analysis of Complex HIV Cohort Studies. Technical Report 62, Department of Epidemiology and Biostatistics, University of California - San Francisco.

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

McLachlan, G. J. and Krishnan, T. (1997). The EM Algorithm and Extensions. New York: John Wiley.

Meng, X. and Rubin, D. B. (1991). Using EM to Obtain Asymptotic Variance-Covariance Matrices: The SEM Algorithm. Journal of the American Statistical Association 86, 899-909.

Meng, X. and Rubin, D. B. (1993). Maximum Likelihood Estimation via the ECM Algorithm: A General Framework. Biometrika 80, 267-278.

Muthen, B. O. (1998-2004). Mplus Technical Appendices. Los Angeles: Muthen and Muthen.

Muthen, B. and Shedden, K. (1999). Finite Mixture Modeling with Mixture Outcomes Using the EM Algorithm. Biometrics 55, 463-469.

Muthen, B. O. (2002). Beyond SEM: General Latent Variable Modeling. Behaviormetrika 29, 81-117.

Muthen, B., Brown, C. H., Masyn, K., Jo, B., Khoo, S., Yang, C., Wang, C., Kellam, S. G., Carlin, J. B., and Liao, J. (2002). General Growth Mixture Modeling for Randomized Preventive Interventions. Biostatistics 3, 459-475.

Nityasuddhi, D. and Bohning, D. (2003). Asymptotic Properties of the EM Algorithm Estimate for Normal Mixture Models with Component Specific Variances. Computational Statistics and Data Analysis 41, 591-601.

Peel, D. and McLachlan, G.J. (2000). Robust mixture modeling using the t distribution. Statistics and Computing 10, 339-348.

Powell, J. M. D. (1977). Restart Procedures for the Conjugate Gradient Method. Math. Prog. 12, 241-254.

Proust, C., and Jacqmin-Gadda, H. (2005). Estimation of Linear Mixed Models with a Mixture of Distribution for the Random Effects. Computer Methods and Programs in Biomedicine 78, 165-173.

Rai, S. N. and Matthews, D. E. (1993). Improving the EM Algorithm. Biometrics 49, 587-591.

Redner, R. A. and Walker, H. A. (1984). Mixture Densities, Maximum Likelihood, and the EM Algorithm. SIAM Review 26, 195-239.

Roy, J. and Lin, X. (2000). Latent Variable Models for Longitudinal Data with Multiple Continuous Outcomes. Biometrics 56, 1047-1054.

Roy, J. (2003). Modeling Longitudinal Data with Nonignorable Dropouts Using a Latent Dropout Class Model. Biometrics 59, 829-836.

SAS Institute Inc. (2008). SAS®/OR 9.2 Users Guide: Mathematical Programming. Cary NC: SAS Institute Inc.

Scher, M. S., Steppe, D. A., and Banks, D. L. (1994). Lower neurodevelopmental performance at 2 years in healthy preterm neonates. Pediatric Neurology 11, 121-122.

Searle, S. R., Casella, G., and McCulloch, C. E. (1992). Variance Components. New York: John Wiley.

Spiessens, B., Verbeke, G., and Komarek, A. (2002). A SAS® Macro for the Classification of Longitudinal Profiles Using Mixtures of Normal Distributions in Nonlinear and Generalised Linear Mixed Models. Technical Report, Biostatistical Center, Catholic University of Leuven, Belgium.

Thoman, E. B. (1982). A biological perspective and a behavioral model for assessment of premature infants. In: Bond L. A. and Joffee J. M., editors. Primary prevention of psychopathology: facilitating infant and early childhood development, Vol. 6. Hanover NH: University Press of New England, 159-179.

Verbeke, G., and Lesaffre, E.. (1996). A Linear Mixed-Effects Model With Heterogeneity in the Random Effects Population. Journal of the American Statistical Association 91, 217-221.

Verbeke, G., and Molenberghs, G. (2000). Linear Mixed Models for Longitudinal Data. New York: Springer-Verlag.

Whitney, M. P. and Thoman, E. B. (1993). Early sleep patterns of premature infants are differentially related to later developmental disabilities. J Dev Behav Pediatr 14, 71-80.

Xu, W. and Hedeker, D. (2001). A Random-Effects Mixture Model for Classifying Treatment Response in Longitudinal Clinical Trials. Journal of Biopharmaceutical Statistics 11, 253-273.

Yakowitz, S. J. and Spragins, J. D. (1968). On the Identifiability of Finite Mixtures. The Annals of Mathematical Statistics 39, 209-214.

Yau, K. K. W., Lee, A. H., and Ng, A. S. K. (2003). Finite Mixture Regression Model with Random Effects: Application to Neonatal Hospital Length of Stay. Computational Statistics and Data Analysis 41, 359-366.