Chapter 8

Estimation of Mixture Models

The previous chapters discussed two standard approaches to analyzing repeated measures/longitudinal data. The estimating equation approach is based on estimating the mean response conditional on observed explanatory risk factors. A mixed model uses continuous latent variables to specify the error distribution in addition to the mean model. The latent variables, called random effects, are used to allow for individual-specific deviations and often assumed to follow a normal model. This structure helps us take dependence between repeated measures into account by modeling within-individual variability. However, standard random effects fail to model the between-individual variability that cannot explained by observed explanatory variables. In this chapter, we introduce the finite mixture model, which is a discrete latent variable model that can provide a more data-driven approach to model heterogeneity in the mean response over time.

A finite mixture model is composed of a fixed number of subgroups in the population, each with its own distribution. To illustrate the general model in a univariate setting,

let's imagine we are interested the height of individuals at age 25. There is quite a bit of variability in heights and the distribution looks non-normal and a bit bimodal. You might already be thinking of possible subgroups that could explain the shape of the distribution, but rather than imposing our a priority knowledge, which could be flawed, we can assume that our population is made up of subgroups with their own height distribution in terms of location and spread. A mixture model is a weighted combination of these distributions, weighted according to their relative size. Visually, Figure 8.1 shows the population distribution with two equally-sized subgroups each with their own Gaussian distribution.

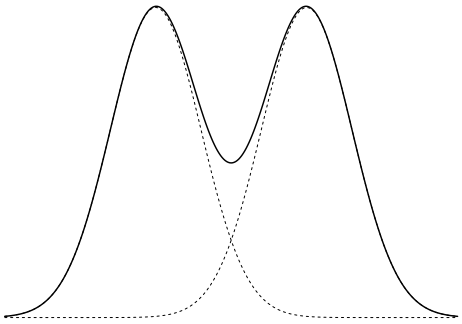

Figure 8.1: Gaussian mixture model involving two subgroups with different means and equal spread.

In addition to modeling non-normal histograms, the subgroup structure of a mixture model can also be generalized to accommodate multiple different relationships between explanatory variables and the response or in the case of longitudinal data, different development patterns over time. This model is particularly useful when the goals are threefold: 1) clustering similar subjects into subgroups as a way to explain the variability in the

response, 2) estimating the mean regression coefficients of each group, and 3) estimating the relationship between baseline variables and group membership. One advantage of this model is that the group structure can approximate complex distributions and can easily flexibility accommodate non-linear relationships between baseline risk factors and the mean response over time.

Similar to the mixed model, we specify the entire distribution of the data and we are interested in the individual-level value of the latent variables in addition to the mean relationships within group. However, with the introduction of the group structure, we no longer have mathematical luck that we had with mixed effects models and GEE. The estimation of the mean coefficients is no longer unbiased when the correlation structure is misspecified [Heggeseth and Jewell(2013), Gray(1994)]. The magnitude of the bias depends on how close the model is (in terms of distribution and covariance structure) to the truth and the degree to which the distributions of the subgroups overlap. Consequently, we must pay attention to how model the dependence within repeated measures and not treat it purely as a nuisance.

In this chapter, we start by discussing the general framework for mixture models and the estimation procedure. Then, we will introduce extensions of the model to increase the utility for longitudinal models. Finally, we discuss two popular versions of this model that are currently implemented in software.

8.1 General Framework for Mixture Models

Mixture models have a long history in the statistical literature. They were originally used to model outliers [Newcomb(1886)] as well as measurements from two groups [Pearson(1894)]. Now, they are used for density estimation, model-based clustering, and analysis of heterogeneous populations. See mixture models books for a comprehensive survey of the history and applications of mixture models [McLachlan and Peel(2000), Titterington et al.(1985)Titterington, Smith, and
We present the multivariate version of model in the context of longitudinal data when there are repeated measures on individual subjects.

8.1.1 The Likelihood

Let Y_{ij} be the jth outcome for individual i observed at time t_{ij}, $i = 1, ..., m$ and $j = 1, ..., n_i$. Without any covariates, we will assume the underlying model for $\mathbf{Y}_i$ is a finite mixture model with K subgroups,

$$f(\mathbf{Y}_i) = \sum_{k=1}^{K} \pi_k f_k(\mathbf{Y}_i).$$

where $0 \leq \pi_k \leq 1$ for all $k = 1, ..., K$ and $\sum_{k=1}^{K} \pi_k = 1$. This weighted sum is made up of the product of group frequencies or prior probabilities, π_k, and subgroup densities, f_k. Usually, f_k's are assumed to be of parametric form i.e. $f_k(\mathbf{Y}_i) = f_k(\mathbf{Y}_i|\boldsymbol{\theta}_k)$, where the distributional form is completely known but for the parameter vector $\boldsymbol{\theta}_k$. Given m independent identically distributed observations of an outcome vector, the likelihood for a

mixture model with K subgroups is

$$L(\boldsymbol{\theta}, \boldsymbol{\pi}) = \prod_{i=1}^{m} \sum_{k=1}^{K} \pi_k f_k(\mathbf{Y}_i | \boldsymbol{\theta}_k)$$

where $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_K)$ and $\boldsymbol{\pi} = (\pi_1, ..., \pi_{K-1})$. For continuous outcomes, the densities are most commonly assumed multivariate Gaussian with a group-specific mean, $\boldsymbol{\mu}_k$, and covariance matrix, $\boldsymbol{\Sigma}_k$. If the outcome is categorical or discrete, then Binomial or Poisson distributions should be used. Semi- or Non-parametric methods may also be used in a general mixture model, but we will focus on parametric forms in this text.

In order to estimate the parameters of the assumed model, we often apply the expectation maximization (EM) algorithm [Dempster et al.(1977)Dempster, Laird, and Rubin] to find the maximum likelihood estimates of $\boldsymbol{\theta}$ and $\boldsymbol{\pi}$. This iterative algorithm is applied when the model includes discrete latent variables such as unknown group labels.

Let z_{ik} be a variable indicating the group membership for individual i such that

$$z_{ik} = \begin{cases} 1 & \text{if individual } i \text{ is in group } k \\ 0 & \text{otherwise.} \end{cases}$$

If these group labels are known, they can be incorporated into the likelihood by allowing individuals to contribute only to their originating group. Taking advantage of the binary nature of z_{ik}, the likelihood can be written in term of products of terms taken to the z_{ik}th power. Given m i.i.d. observations of the outcome vector and their group labels, the

complete data likelihood is written as

$$L_C(\boldsymbol{\theta}, \boldsymbol{\pi}) = \prod_{i=1}^{m} \prod_{k=1}^{K} (\pi_k f_k(\mathbf{Y}_i|\boldsymbol{\theta}_k))^{z_{ik}}.$$

The observed-data likelihood, $L(\boldsymbol{\theta}, \boldsymbol{\pi})$, can be obtained by integrating $L_C(\boldsymbol{\theta}, \boldsymbol{\pi})$ over the group labels. Therefore, the ML estimates calculated via the EM algorithm based on the complete log likelihood maximize the observed-data likelihood.

The EM algorithm consists of two alternating steps. The expectation step (E-step) involves replacing group indicators, z_{ik}, with current values of their conditional expectation, the posterior probability of group membership, written as

$$\alpha_{ik} = \frac{\pi_k f_k(\mathbf{Y}_i|\boldsymbol{\theta}_k)}{\sum_{j=1}^{K} \pi_j f_j(\mathbf{Y}_i|\boldsymbol{\theta}_j)}$$

for $i = 1, ..., n_i$ and $k = 1, ..., K$ using current estimates of the parameters. In the maximization step (M-step), the parameters estimates are updated by maximizing the expected complete log likelihood from the E-step. Typically, this is done using numerical optimization. However, for a simple multivariate Gaussian mixture, there are closed form solutions. Specifically, for $k = 1, ..., K$, the updated weights are

$$\hat{\pi}_k^{new} = \frac{\sum_{i=1}^{m} \alpha_{ik}}{m},$$

the estimates of the means are

$$\hat{\boldsymbol{\mu}}_k^{new} = \frac{\sum_{i=1}^{m} \alpha_{ik} \mathbf{Y}_i}{\sum_{i=1}^{m} \alpha_{ik}},$$

and the kth new covariance matrix estimate is

$$\hat{\boldsymbol{\Sigma}}_k^{new} = \frac{\sum_{i=1}^m \alpha_{ik}(\mathbf{Y}_i - \boldsymbol{\mu}_k^{new})(\mathbf{Y}_i - \boldsymbol{\mu}_k^{new})^T}{\sum_{i=1}^m \alpha_{ik}}.$$

These two steps alternate until convergence. The EM algorithm guarantees convergence of the likelihood to a local maximum [Dempster et al.(1977)Dempster, Laird, and Rubin]. Global convergence may be attained by starting the algorithm multiple times with initial random group assignments and using the parameter estimates with highest associated likelihood. Robust standard error estimates for the can be obtained using the Huber-White sandwich estimator introduced in Chapter 6.

The EM algorithm provides estimates of the parameter vectors for the group means and covariance, $\boldsymbol{\theta}$, as well as prior probabilities, $\boldsymbol{\pi}$. These are useful for exploring the longitudinal response and its relationship with other factors in the data as well as the group frequencies. Additionally, the posterior probabilities from the last E-step can be used to classify individuals into subgroups or clusters by choosing the group with the largest posterior probability for a particular individual. Beyond hard clustering individuals into one and only one group, the posterior probability can serve as an indication of uncertainty in the group membership. For example, if the largest posterior probability for an individual is 95%, then we can be fairly certain about our group assignments in contrast to a probability of 60% ,which would indicate more uncertainty about into which group an individual should be placed. A histogram of the maximum posterior probabilities can be used as a evaluation tool for the resulting groups. Large values near 100% indicate well-separated groups and high certainty in the classification while many lower values indicate there are not well-

defined groups that are uniqued distinct.

8.2 Longitudinal Extensions

The general multivariate model can be directly applied to repeated measures longitudinal data. However, there are many ways to extend the model to make it more useful in answer research questions in longitudinal studies. There are two types of research questions that come up with longitudinal data. The first is about the relationship between a time-varying explanatory variable and the response over time, i.e. relationship between body mass index and age. The second is about the effect of baseline risk factors on the response over time, i.e. the effect of early life exposure on the relationship between age and body mass index. We can extend the mixture model by incorporating a generalized linear model for the group means and for the prior probabilities to help us answer these two types of questions. We address the second question first.

8.2.1 Multinomial Logistic Regression

With a mixture model, each subgroup can represent different response distributions or as we will see different relationships between our explanatory variables such as our measurement of time and the response. To determine if baseline factors impact group membership, a naive but tempting approach would be to translate the posterior probabilities into group labels and then complete a secondary regression analysis using the group labels as the response variable. This approach is easy and straightforward to carry out, but it does not

take into account the uncertainty of group membership provided by the posterior probabilities. The loss of uncertainty information can have negative consequences on estimation and interpretation of the relationship of interest. Similar information loss occurs when transforming quantitative variables into distinct intervals that are then treated as categories in further analysis [MacCallum et al.(2002)MacCallum, Zhang, Preacher, and Rucker]. The resulting impact on possible conclusions highly depends on the variability within the new categories or groups.

An alternative model-based approach that preserves the uncertainty involves modeling the group probabilities, π_k, in the mixture. Using baseline risk factors in a regression structure, all model parameters are estimated simultaneously using maximum likelihood. We notate the observed vector of baseline factors as $\mathbf{w}_i$ for individual i and denote the probability of the ith individual being in the kth group conditional on these factors as $\pi_k(\mathbf{w}_i)$. Using multinomial logistic regression model, which is a generalization of simple logistic regression, we use a linear model to predict the log ratio of group probabilities. Specifically, the log ratio of the probability of being in the kth group relative to the probability of being in the Kth group, which we will use as a reference group, is a linear function of baseline factors,

$$\log\left(\frac{\pi_k(\mathbf{w}_i)}{\pi_K(\mathbf{w}_i)}\right) = \boldsymbol{\gamma}_k \mathbf{w}_i,$$

for $k = 1, ..., K-1$. By exponentiating both sides, solving for the probabilities, and using

the fact that the probabilities have to add to one, we get

$$\pi_K(\mathbf{w}_i) = \frac{1}{1 + \sum_{k=1}^{K-1} \exp(\boldsymbol{\gamma}_k \mathbf{w}_i)} \tag{8.1}$$

and then plugging this into equation (8.1) and solving for $\pi_k(\mathbf{w}_i)$, we get

$$\pi_k(\mathbf{w}_i) = \frac{\exp(\boldsymbol{\gamma}_k \mathbf{w}_i)}{1 + \sum_{j=1}^{K-1} \exp(\boldsymbol{\gamma}_j \mathbf{w}_i)}$$

for $k = 1, ..., K-1$. We use these probabilities in our likelihood function in place of π_k and estimate the multinomial regression coefficients, $\boldsymbol{\gamma}_k$, by adding a numerical optimization step in the M-step of the EM algorithm.

8.2.2 Regression Mean

Let's return to the first research question. To study the relationship between a time-varying explanatory variable and a response, we can incorporate a regression structure into the mean model. Similar to mixed effects models and marginal models, we can write the expected value of the response of an individual in group k as a linear function of explanatory variables. For example, given a time-varying explanatory variable, $\mathbf{X}$, the expected value of a response can be written as

$$E(Y_{ij}|X_{ij}) = \beta_{k0} + \beta_{k1} X_{ij}.$$

To generalize this to categorical or count responses, logic and log link functions can be used here.

In many situations, we are interested in how the response changes over time. For that case, the time-varying explanatory variable, X_{ij} would be the observation time t_{ij}, or another measure of time such as age, time since baseline, or year depending on the research question. A standard linear model assumes that the change of the mean over time is constant. If the functional form of the relationship between the $E[Y_{ij}|X_{ij}]$ and X_i is not a straight line, you can use a polynomial form by adding higher order terms such as X_{ij}^2 or X_{ij}^3 to the mean model. An alternative procedure is to take a nonparametric approach and use a spline functional basis such as B-splines to model the nonlinear relationship. B-spline functions are piecewise polynomials of a chosen order that are continuous at the knots or break-points. We leave the details of B-splines to other references [Curry and Schoenberg(1966), De Boor(1976)], but given the desired degree of the polynomials, the number of internal knots, and data, most statistical software packages can produce a design matrix for a given variable in the regression model.

8.3 Continuous Response Models

Assuming there are K underlying groups, each with a different relationship between a continuous response and time-varying covariates, a general model for the vector of responses from individual i from the kth group is

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta}_k + \mathbf{Z}_i\boldsymbol{\beta}_{ik} + \boldsymbol{\epsilon}_i$$

where $\mathbf{X}_i$ is a design matrix for the time-varying covariates, $\boldsymbol{\beta}_k$ is a vector of fixed effect parameters, $\mathbf{Z}_i$ is a subset (could be an empty subset) of columns from $\boldsymbol{X}_i$ specified to

have a random coefficient, $\boldsymbol{\beta}_{ik}$ is a vector of random variables with a distribution specified within group k, and the vector of errors within a subject i is multivariate normal, or $\boldsymbol{\epsilon}_i \sim MVN(0, \boldsymbol{\Sigma}_{ik})$. Similar to linear mixed effects models, the joint distribution of $\boldsymbol{\beta}_{ik}$ has to be defined and in most cases, it is assumed to be multivariate normal with mean 0 and covariance $\mathbf{G}_k$. In practice, the size of $\boldsymbol{\beta}_{ik}$ is restricted since adding one parameter to a mixture model increases the total number of parameters by K. If not $\mathbf{Z}_i$ is non-empty, the model is called a growth mixture model and it can be estimated using the Mplus software [Muthén and Muthén(1998-2010)] and the lcmm package in R **??**. With this structure, we allow parameters to randomly vary between individuals, i.e. random effects within groups. However, in practice, often $\mathbf{Z}_i$ only includes a column of 1's for a random intercept model for the sake of parsimony. The likelihood of an observation from subject i is $\mathbf{Y}_i \sim MVN(\mathbf{X}_i\boldsymbol{\beta}_k, \mathbf{V}_i)$ where the covariance matrix of the responses on subject i in group k is $\mathbf{V}_i = \mathbf{Z}_i\mathbf{G}_k\mathbf{Z}_i^T + \boldsymbol{\Sigma}_{ik}$. In practice, it is not unusual to only allow random intercepts due to model complexity and computational time; this is equivalent to using the exchangeable correlation structure and allowing the covariance parameters to vary between groups.

Now, in contrast to mixed models and GEE, the quality of the estimates of the mean model parameters relies on correctly specifying the dependence structure of the data. Besides random effects, another way to take the dependence between repeated measures into account is directly through the covariance matrix of the error within a subject, $\boldsymbol{\Sigma}_{ik}$. Since the observations in longitudinal data are often sparse and irregular, simplifying assumptions are almost always necessary. Therefore, a trade-off occurs between correctly approximating

the dependency and the need for parsimony due to sample size. One simplified correlation structure for continuous outcomes is conditional independence; given the group membership labels, individual observations are assumed to be independent of each other. This structure is often paired with an assumption of constant variance across time within an individual ($\mathbf{\Sigma}_{ik} = \sigma_k^2 \mathbf{I}_{n_i \times n_i}$). With independence, we could allow the magnitude of the variance of the errors to vary between groups or if sample size is relatively small, we might restrict the variance to be the same across groups ($\mathbf{\Sigma}_{ik} = \sigma^2 \mathbf{I}_{n_i \times n_i}$). The specification of independence and constant variance within and between groups is used in group-based trajectory modeling or latent class growth analysis implemented in the Proc Traj package in SAS and Stata [Jones et al.(2001)Jones, Nagin, and Roeder] and lcmm package in R [Proust-Lima et al.(2014)Proust-Lima, Philipps, Diakite, and Liquet].

Other common correlation structures, such as exchangeable, exponential, and autoregressive, discussed earlier in Chapter INSERTREFERENCE can be used to allow dependencies between repeated measures while limiting the total number of parameters needed. An exponential model with group-varying parameters may be more appropriate if the serial dependency decreases with increasing time lag, or $\mathbf{\Sigma}_{ik} = \sigma_k^2 \mathbf{R}_{ik}$ where the correlation between the jth and the lth observation is $[\boldsymbol{R}_{ik}]_{jl} = \exp(-|t_j - t_l|/r_k)$. Other correlation structures can also be used as long as the resulting covariance matrix is positive-definite.

8.3.1 Example: BMI in NLSY

We can apply a mixture model with these extensions on a continuous outcome from a national longitudinal data set, body mass index from the National Longitudinal Survey of Youth (NLSY). About 12,000 young men and women between the ages of 14 and 22 years old were selected to participate in the NLSY in 1979 organized by the Bureau of Labor Statistics. While the main goal of the survey is to identify characteristics defining the transition that youth make from school to the labor market as adults, the survey also included health information that we can use to study weight gain and obesity development in this cohort.

Body mass index (BMI) is one measure of weight relative to height calculated as weight in kg divided by squared height in m^2. This measure, albeit potentially an inaccurate measure of body fat content, is often used as a crude indicator of obesity. To study the development over time, our main explanatory variable is age of the individual and potential baseline characteristics that could explain some of the variability in development patterns could be sex and childhood socioeconomic status.

Figure **??** shows the BMI over time for a random selection of 500 individuals who had at least 3 BMI measurements. It is hard to decipher the shape of the patterns from a visual representation of the data due to the number of individuals in the data set and variability within individuals. Therefore, in order to flexibly model the mean patterns over time, we use a B-spline basis of degree 2 with one internal knot at the median age to create the design matrix $\mathbf{X}_i$. In practice, the degree of the basis could be increased and the number

of internal knots can be increased to model more complex development patterns.

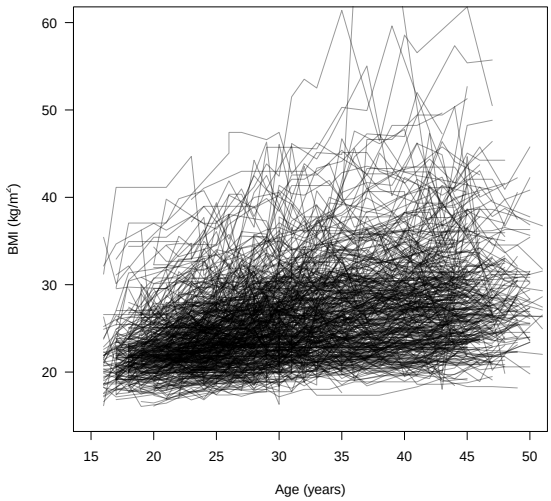

Figure 8.2: Body mass index of 500 randomly selection individuals from NLSY who had at least 3 observed BMI measurements.

Two other aspects of the mixture model specification need to be chosen before we can estimate the parameter: the number of groups, K, and the covariance structure. To choose the number of groups, you can choose a maximum number of groups, $K_{max} < n$, and fit a mixture model for each $K = 2, ..., K_{max}$. Then, you choose the value of K that minimizes a chosen criterion such as the BIC, which has been shown to work well in practice even though it is not a consistent estimation due to lack of regularity conditions [Fraley and Raftery(1998)]. Fitting the model also requires a specified covariance structure. For this data set, we plot the BIC for different values of K for a variety of correlation structures in Figure 8.3 and choose the correlation structure with the lowest BIC value and then within that structure choose the smallest K such that the difference between BIC for K and $K+1$ are small relative to prior differences to give us a parsimonious model. Therefore,

choosing between independence, autoregressive (AR) and random intercept (exchangeable) correlation structures, we choose AR(1) and $K = 3$.

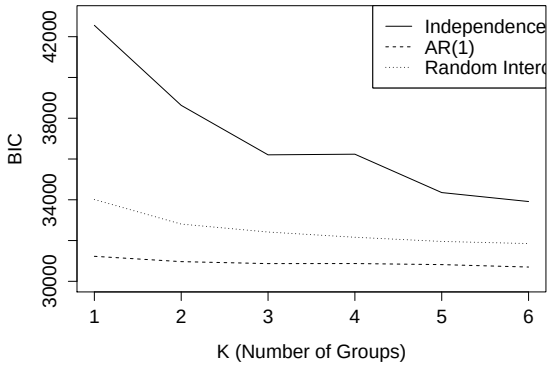

Figure 8.3: Bayesian Information Criterion (BIC) for various values of K and correlation structure fit to body mass index data of 500 randomly selection individuals from NLSY who had at least 3 observed BMI measurements.

Once the parameters are estimated, group memberships can be assigned based on the posterior probabilities. Figure 8.4 shows the raw BMI trajectories classified into three groups based on the group in which they have the highest posterior probability on the left and the group mean trajectories on the right. The most populous group with almost 90% of the individuals seems to have more normal BMI levels that on average increase linearly over time. The other two groups capture more abnormal patterns. One very small group has very high BMI levels and saw great increases in their lives up until age 50. The last group of individuals, which has higher BMI levels, generally have a higher rate of BMI growth prior to age 35 than the majority of the individuals, but then the trajectory seems to flatten out on average. It is important to notice that the groups seem to be homogeneous in shape as well as level.

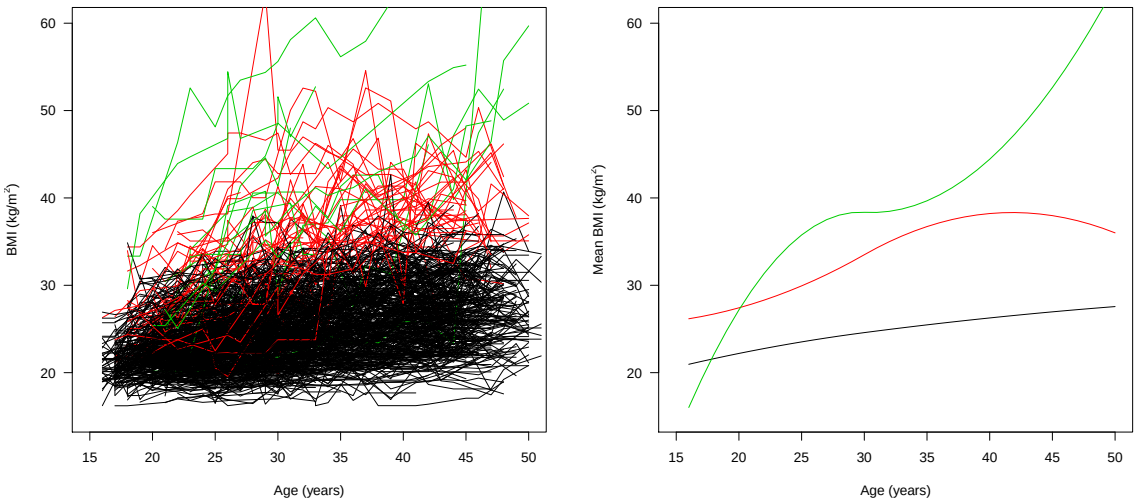

Figure 8.4: Group membership labels (left) and group mean body mass index (right) based on a mixture model with $K = 3$ and AR(1) correlation structure fit to 500 randomly selection individuals from NLSY who had at least 3 observed BMI measurements.

If we were to want to directly focus on the development pattern over time, we could vertically scale every trajectory by their individual mean and get different groups. Figure 8.5 shows the difference in group membership and means if you subtract the means and focus on shape. This simple action of removing the individual-specific level impacts the correlation so we switched to using independence to fit the mixture on the transformed data. Now the most populous group, which we will call Group 1, has a low rate of change over time but one that decreases and stabilizes by age 30. The smallest group, which we will call Group 2, has the greatest rate of change during this time period but the individuals in this group have a variety of BMI levels at age 15-20 years. Lastly, the second biggest group, which we will call Group 3, has a trajectory pattern in between the two extremes that is associated with a variety of BMI levels.

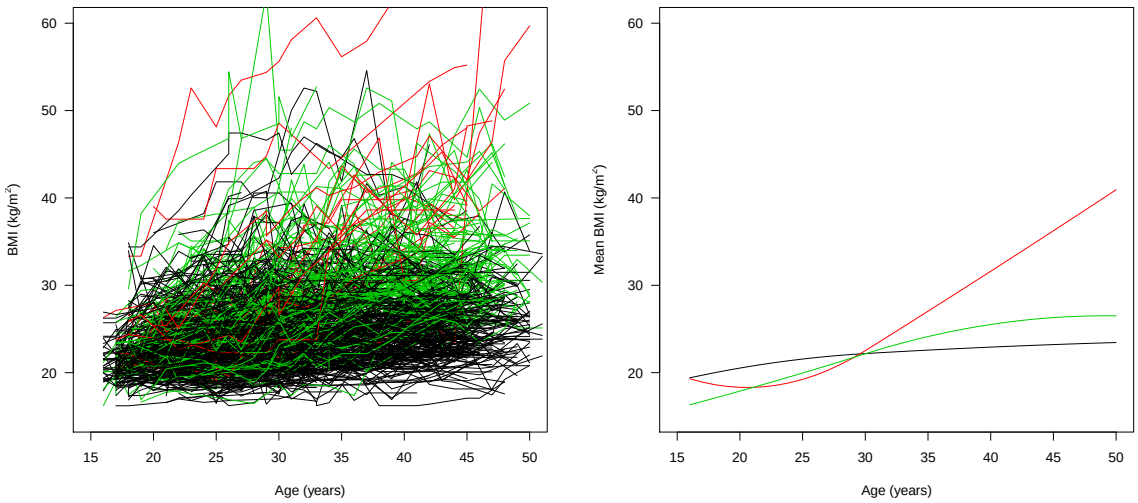

Figure 8.5: Group membership labels (left) and group mean body mass index (right) based on a mixture model after subtracting the mean with $K = 3$ and independence correlation structure fit to 500 randomly selection individuals from NLSY who had at least 3 observed BMI measurements.

Now that we have explored the BMI trajectories, the next question to ask is what factors impact the individual trajectory. In other words, what determines which group you will be in? We can address this question by adding baseline, non-time-varying factors to the multinomial logistic regression for the prior probabilities. In this circumstance, we might be interested in sex and race have any relationship with this data-driven groups. Adding them to the mixture model based on the shifted data, we find that sex has moderate to little relationship with the groups in that being male does not determine your BMI trajectory; however, race does have some impact. Relative to individuals self-categorized as Black or Hispanics, those self-categorized as White are more likely to follow the general slow, linear growth pattern. Table **??** gives the estimated multinomial coefficients with the associated standard errors, Wald test statistic, and p-value, treating Group 3 as the reference group.

	Coefficient	SE	Wald	p-value
Intercept Group 1	1.46	0.21	6.84	< 0.00001
Intercept Group 2	-1.49	0.42	-3.56	0.00038
Male Group 1	0.30	0.25	1.23	0.21782
Male Group 2	-1.01	0.62	-1.62	0.10554
Black Group 1	-1.38	0.27	-5.09	< 0.00001
Black Group 2	-0.77	0.66	-1.16	0.24447
Hispanic Group 1	-0.79	0.34	-2.29	0.02201
Hispanic Group 2	0.35	0.65	0.54	0.58843

In order to interpret these point estimates of the coefficients, we should look back to the model structure. For each individual, we use a regression structure to for the logarithm of the ratio of two probabilities. With three indicator variables as our baseline covariates, we have

$$\log\left(\frac{\pi_k(\mathbf{w})}{\pi_3(\mathbf{w})}\right) = \gamma_{0k} + \gamma_{1k}Male + \gamma_{2k}Black + \gamma_{3k}Hispanic$$

for Group $k = 1, 2$. Then, to compare between men and women, we let $\mathbf{w}^*$ include the value of 1 for Male indicator and $\mathbf{w}^*$ include the value of 0 for the Male indicator with all other variables constant. Then, the difference in the two equations equals

$$\log\left(\frac{\pi_k(\mathbf{w}^*)}{\pi_3(\mathbf{w}^*)}\right) - \log\left(\frac{\pi_k(\mathbf{w})}{\pi_3(\mathbf{w})}\right) = \gamma_{1k},$$

which can be simplified to

$$\frac{\pi_k(\mathbf{w}^*)/\pi_3(\mathbf{w}^*)}{\pi_k(\mathbf{w})/\pi_3(\mathbf{w})} = \exp(\gamma_{1k})$$

for Group $k = 1, 2$ comparing to reference Group 3. To illustrate, the relative probability of Group 1 as compared to Group 3 is $\exp(0.30) = 1.35$ times greater for males than for females. In contrast, the probability of Group 2 relative to Group 3 is $\exp(-1.01) = 0.36$

times less for males than for females. If you compare these estimates with the variability given by their standard errors, they don't seem to be that big. The only significant relationships are with Group 1 relative to Group 3 comparing between race/ethnicity.

8.4 Binary Outcome

As mentioned earlier in the chapter, if our outcome of interest is a categorical variable rather than continuous, we use an appropriate model for the distribution of the data such as Bernoulli or Poisson model within each group. If time-dependent explanatory variables explain some variability in the outcome and there are a finite number of subgroups with their own relationships, an appropriate link function between the mean and explanatory variables could be incorporated into a finite mixture model. In the case of a binary variable, the distribution for each group would be modeled using a logistic model with group-specific parameters,

$$f_k(\mathbf{Y}_i|\boldsymbol{\theta}_k) = \prod_{j=1}^{n_i} p_{ijk}^{y_{ij}}(1-p_{ijk})^{1-y_{ij}}$$

with probabilities

$$p_{ijk} = \frac{\exp(\mathbf{x}_{ij}^T\boldsymbol{\beta}_k + \mathbf{z}_{ij}^T\boldsymbol{\beta}_{ik})}{1+\exp(\mathbf{x}_{ij}^T\boldsymbol{\beta}_k + \mathbf{z}_{ij}^T\boldsymbol{\beta}_{ik})}.$$

For count data, a Poisson model would be used with a log link function,

$$f_k(\mathbf{Y}_i|\boldsymbol{\theta}_k) = \prod_{j=1}^{n_i} \frac{\lambda_{ijk}^{y_{ij}} e^{-\lambda_{ijk}}}{y_{ij}!}$$

with mean

$$\lambda_{ijk} = \exp(\mathbf{x}_{ij}^T \boldsymbol{\beta}_k + \mathbf{z}_{ij}^T \boldsymbol{\beta}_{ik}).$$

Estimation of these parameters can be completed via maximum likelihood. Software packages that fit mixture models will have slightly different available model specifications as there isn't one known best algorithm for the general mixture model. Parameter estimates can be interpreted in the same manner as if there were only one group in the mixture model.

Bibliography

[Curry and Schoenberg(1966)] Curry, H. B., Schoenberg, I. J., 1966. On Pólya frequency functions iv: the fundamental spline functions and their limits. Journal d'Analyse Mathématique 17 (1), 71–107.

[De Boor(1976)] De Boor, C., 1976. Splines as linear combinations of B-splines. A survey. In: Lorentz, G. G., Chui, C. K., Schumaker, L. L. (Eds.), Approximation Theory II. Academic Press, New York, pp. 1–47.

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

[Fraley and Raftery(1998)] Fraley, C., Raftery, A. E., 1998. How many clusters? which clustering method? answers via model-based cluster analysis. The Computer Journal 41 (8), 578–588.
URL http://comjnl.oxfordjournals.org/content/41/8/578.abstract

[Gray(1994)] Gray, G., 1994. Bias in misspecified mixtures. Biometrics 50 (2), 457–470.

[Heggeseth and Jewell(2013)] Heggeseth, B. C., Jewell, N. P., 2013. The impact of covariance misspecification in multivariate Gaussian mixtures on estimation and inference: An application to longitudinal modeling. Statistics in Medicine 32 (16), 2790–2803. URL 10.1002/sim.5729

[Jones et al.(2001)Jones, Nagin, and Roeder] Jones, B. L., Nagin, D. S., Roeder, K., 2001. A SAS procedure based on mixture models for estimating developmental trajectories. Sociological Methods & Research 29 (3), 374–393.

[MacCallum et al.(2002)MacCallum, Zhang, Preacher, and Rucker] MacCallum, R. C., Zhang, S., Preacher, K. J., Rucker, D. D., 2002. On the practice of dichotomization of quantitative variables. Psychological methods 7 (1), 19.

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

[Muthén and Muthén(1998-2010)] Muthén, L. K., Muthén, B. O., 1998-2010. Mplus user's guide.

[Newcomb(1886)] Newcomb, S., 1886. A generalized theory of the combination of observations so as to obtain the best result. American Journal of Mathematics, 343–366.

[Pearson(1894)] Pearson, K., 1894. Contributions to the mathematical theory of evolution. Philosophical Transactions of the Royal Society of London. A, 71–110.

[Proust-Lima et al.(2014)Proust-Lima, Philipps, Diakite, and Liquet] Proust-Lima, C., Philipps, V., Diakite, A., Liquet, B., 2014. lcmm: Estimation of extended mixed

models using latent classes and latent processes. R package version 1.6.4.

URL `http://CRAN.R-project.org/package=lcmm`

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