ED255C
Assignment#5

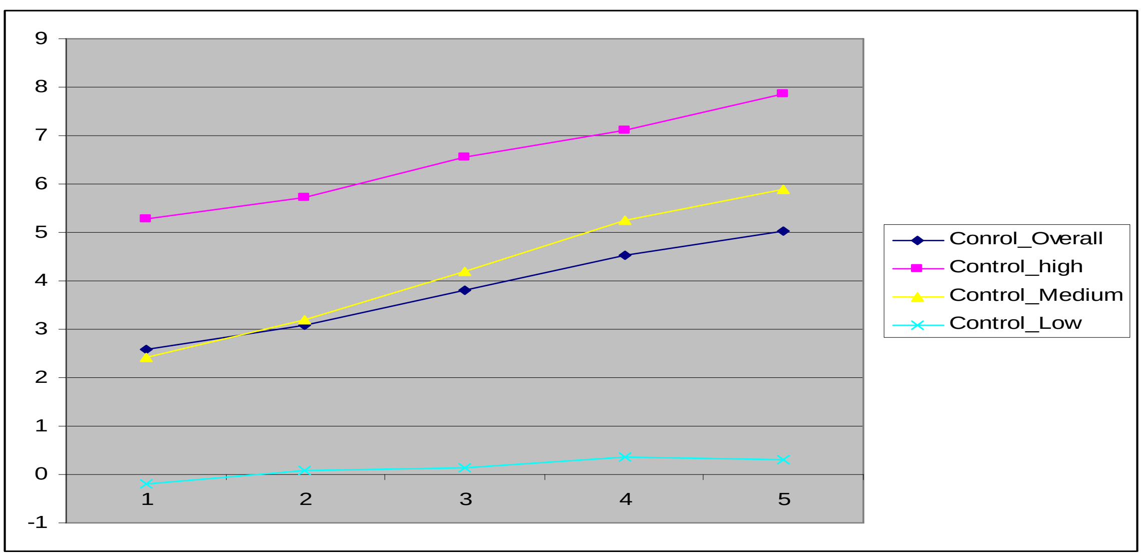
Analysis for Control Group using Growth Mixture Modeling

Introduction

This report uses growth mixture modeling to analyze, over time, the reading achievement of the control group of an artificial dataset, which was generated by the Mplus Monte Carlo facility. In addition to the data from 512 subjects in the control group, the artificial data also consists of data from 488 subjects in an intervention group as well. The reading achievement of the subjects in this dataset was repeatedly measured five times – one pre-intervention measurement as the baseline measurement and four post-intervention measurements. By analyzing the control group only, I will be able to detect the normative growth pattern of reading achievement over time.

Figure 1 displays the overall mean of reading achievement over the five timepoints for the control group and provides evidence of an overall linear increasing trend. Compared to the baseline, reading achievement increases by more than one standard deviation by time 5. In addition, Figure 1 shows the change in mean over time for three arbitrary sub-groups (Low, Medium and High), which were obtained by dividing the groups into the first through third quartiles at the first timepoint[1]. The difference in trends between certain subgroups suggests there might be qualitatively different subpopulations in the control group in terms of growth patterns over time. To capture the characteristics of the subpopulations in the control group in more detail, a growth mixture modeling analysis was conducted.

Figure 1 Reading achievement of overall mean and three subgroups based on the baseline values



[1] The values of the 25% and 75% quantiles were 0.83 and 4.04 respectively. Subgroups were demarcated as follows: first, the range of the "Low" group was defined as any value lower than .83; second, the "Medium" group was defined as values between .83 and 4.04; and finally, the "High" group contained all values greater than 4.04.

Model Selection

Step1: Analysis without Covariates

1) As the first step of a Growth Mixture Modeling analysis, I allowed only the growth factor means to be different across classes while other parameters – the variances and covariances of the growth factors and the residual variances – were held equal across classes. Next, to determine the number of classes, several growth mixture models were explored; all tests concluded that a three-class model provided the best solution. The likelihood ratio test suggests that a three-class solution fits the data best. The BIC (Bayesian Information Criterion) agrees with a three-class solution since the BIC reaches its lowest point at three classes as well. Since the p-value for comparison of a three versus four class model is 0.0602, the Vuong-Lo-Mendell-Rubin likelihood ratio test (LMR-LRT) also supports a three-class solution. However, the entrophy for a three-class solution is slightly smaller than that of a two or four-class solution. This might be improved by allowing class-specific variance.

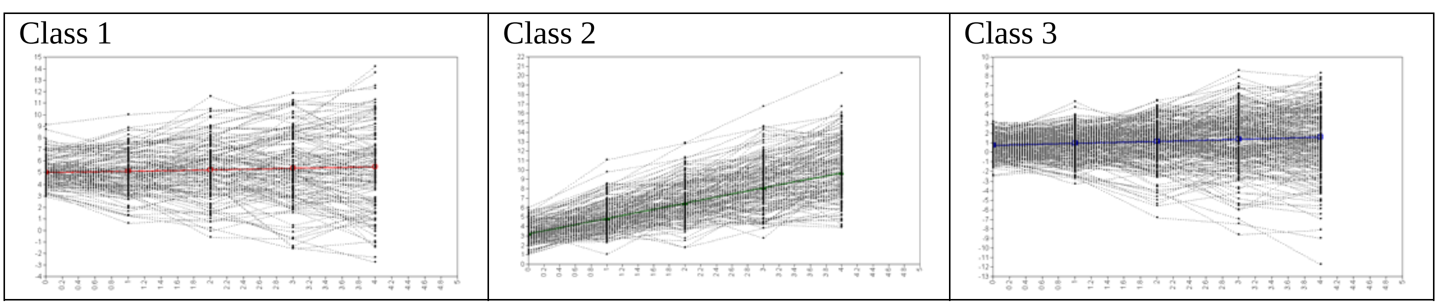
Table 1 GMM without a covariate

Number of Classes	Loglikelihood	#parameters	BIC	Entrophy	LMR-LRT p-value	-2*Difference
1	-5286.451	10	10635.284			
2	-5259.091	13	10599.280	0.765	0.0000	54.72
3	-5237.921	16	10575.656	0.698	0.0004	42.34
4	-5234.637	19	10587.803	0.770	0.0602	6.568

The three-class solution produces three subgroups in terms of growth patterns – 24.2% were in a class with a high baseline which grew slowly (Class 1), 31.3% were in a class with a moderate baseline which grew rapidly (Class 2), and finally, 44.4% were in a class with a low baseline which grew slowly (Class3).

2) To improve the three-class model, class specific variance was investigated by looking at the plots of the estimated mean and observed individual trajectories (Figure 2). Class 2 seems to have a different intercept and slope variation from other classes indicating that Class 2 might have a class-specific variance. Thus, 1) the three-class GMM allowing class-specific intercept variance and 2) the three-class GMM with class-specific intercept and slope variance were fitted.

Figure 2 Estimated means and observed individual trajectories for three-class GMM with no variation across classes and no covariates



The GMM with class-specific intercept variance concluded that the intercept variance for Class 1 seemed to be different from the intercept variances for other classes. However, the likelihood ratio test from the three-class invariant GMM model did not provide any strong evidence of class-specific intercept variance ($\chi_4^2 = 2.72$, p>0.05). In fact, the variance for each class appeared to be close – the variances for Class 1, Class 2 and Class 3 were 1.660, 1.04 and 0.978 respectively. Since the GMM allowing for class specific variances and covariances for both intercept and slope yielded a non-positive definite latent variable covariance matrix, the GMM with only class-specific variances for both intercept and slope was fitted. Allowing class specific variance for intercept and slope did not seem to significantly improve the class invariant GMM ($\chi_4^2 = 2.67$, p>0.05), which suggests that the model that allows for only growth factor means to vary across classes should be selected.

Step2: Analysis with Covariates

1) In this step, a covariate, the poverty status of the student's family was included in the model. As in the analysis without covariates, to decide the number of classes, a couple of growth mixture models, holding effects of the poverty status on growth factors invariant across class, were explored. All model fit indices suggest a three-class GMM solution (Table2). First, the likelihood ratio test suggests that a three-class solution fits the data best ($\chi_4^2 = 49.594$, p<0.001). The BIC also agrees with a three-class solution because the BIC dips at three classes. Since the p-value for comparison of a three versus four-class model is 0.5895, LMR-LRT also supports a three class solution.

Compared to the three-class GMM without the new covariate, a higher Entropy (0.717) of the three-class solution model, in particular, indicates that adding a covariate to the model seems to improve the quality of classification. The three-class solution seems to produce more distinct classes by reducing the variance of the initial status growth factor. The class invariant regression of growth factors on the covariate indicates that the covariate is significantly and negatively associated with the initial growth factor. In contrast, the effect of the covariate on the growth rate factor was not found to be significant.

Table 2 GMM with a covariate

Number of Classes	Loglikelihood	#parameters	BIC	Entrophy	LMR-LRT p-value	-2*Difference
1	-5239.679	12	10554.219			
2	-5209.867	16	10519.547	0.777	0.0000	59.624
3	-5185.070	20	10494.907	0.717	0.0006	49.594
4	-5182.121	24	10513.962	0.710	0.5895	5.898

2) In the next step, I let the effect of the covariate on the growth factors vary across class to see whether the regression slopes of growth factors on the covariate vary across classes. It seems that the covariate significantly influences the initial growth factor for Class 1 and Class 3 while it did not have any significant effect on the growth rate factor for Class2. Although the effect of the covariate on the initial growth factor was slightly varying across classes, the likelihood test did not provide any strong evidence for class-varying effects of the covariate on growth factors ($\chi_4^2 = 5.432$, p>0.05, Table 3). Hence, I conclude that the three-class GMM with covariates while holding the effect of a covariate on growth factors class-invariant seems the best model if a covariate is introduced to the model.

Table 3 Class invariant Model with a covariate versus Class-variant Model with a covariate

Model	**Loglikelihood**	**#parameters**	**BIC**	**Entropy**	**-2*Difference**
3 class GMM with a class invariant covariate	-5185.070	20	10494.907	0.717	
3 class GMM with a class variant covariate	-5182.354	24	10514.427	0.721	5.432

Step3: Analysis with a Distal outcome

In this step, I included a distal outcome in order to more specifically investigate the relationship between growth patterns of subpopulations and a negative future event. To decide the number of classes, first, the analysis for the effects of class membership on a distal outcome was carried out. Unlike the selection for a GMM with covariates only, model fit indices did not agree with the decision for the number of classes (in Table 4). Since the critical value of is χ_5^2 at a significance level of 0.05 is 11.07, the likelihood ratio test rejects a three-class model in favor of a four-class model. However, LMR-LRT and BIC point to the three-class solution because the p-value for testing three versus four classes in LMR-LRT is 0.1957 and the BIC dips at three classes. Although a likelihood ratio test supports a four-class solution, the loglikelihood values for the four-class solution were not stable, indicating local maxima. Moreover, the four-class solution did not seem to yield a good quality of classification. One class has a very small number of individuals compared to the rest of the three classes – the proportion for each class is 9.4%, 42.0%, 28.9% and 19.6%. These point out that there might be too many classes in the model. Therefore, I selected the three-class GMM with distal outcome.

Next, I tested whether or not there is a class-invariant effect on distal outcomes by comparing a class-varying model to a class invariant model. Since I hypothesized that a group which has a low starting point and low growth rate i.e. a “poor developmental” group might have a higher probability of having a negative distal outcome occur, I also tested whether this group is different from other classes. The likelihood ratio test suggested that the model which allows the probability of distal outcomes to vary between the poor developmental class and the other two classes appeared to fit the data best ($\chi_1^2 = 39.398$, p<0.001, Table 5). This indicates that the poor developmental group seems to have a different probability of having a distal outcome. Through all these steps, the three-class GMM with a class-invariant covariate and a class-varying distal outcome has been selected as my final model.

Table 4 GMM with a covariate and a distal outcome

Number of Classes	Loglikelihood	#parameters	BIC	Entrophy	LMR-LRT p-value	-2*Difference
1	-5560.532	14	11208.400			
2	-5520.827	19	11160.183	0.717	0.0000	79.41
3	-5485.193	24	11120.106	0.735	0.0000	71.268
4	-5478.575	29	11138.062	0.721	0.1957	13.236

Table 5 Class invariant Model with a distal outcome versus Class-variant Model with a distal outcome

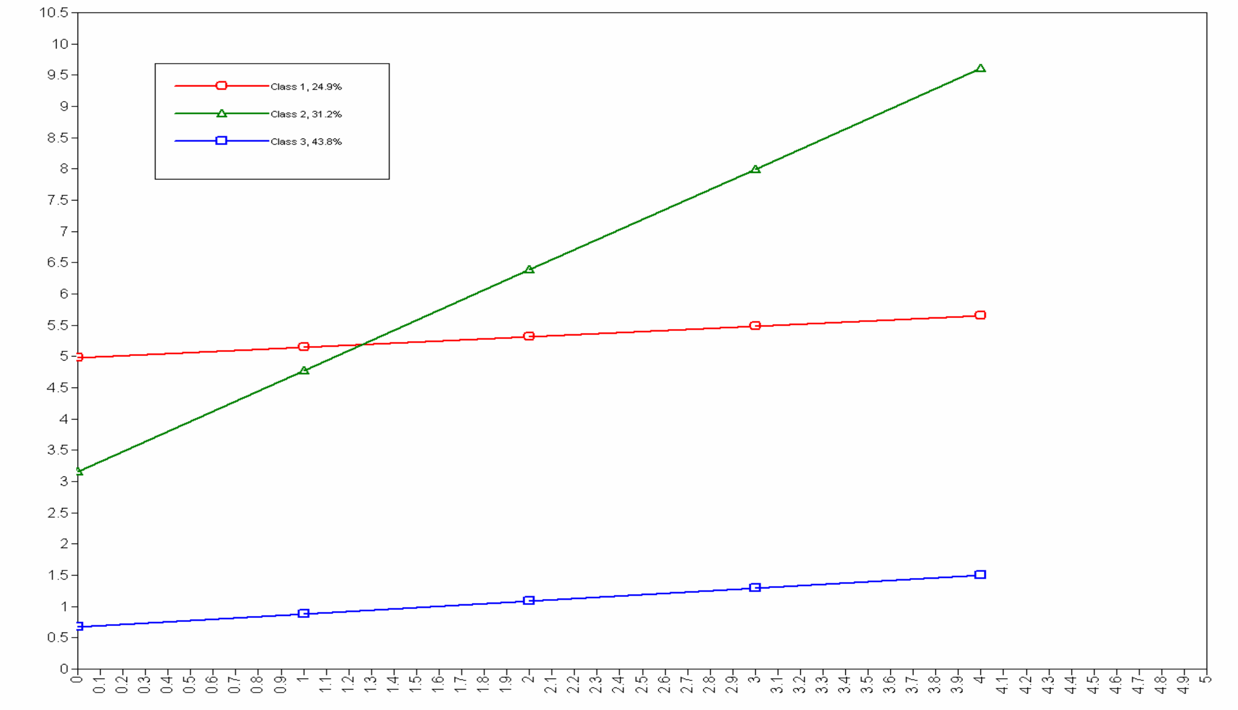
Model	Loglikelihood	#parameters	BIC	Entrophy	-2*Difference
3 class GMM with class invariant distal outcome	-5505.923	22	11149.088	0.717	
3 class GMM with 1 class varying distal outcome	-5486.224	23	11115.930	0.733	39.398
3 class GMM with all class varying distal outcome	-5485.193	24	11120.106	0.735	2.06

Interpretation of the Final Model – Three-class GMM with a covariate and a distal outcome

Figure 4 provides the estimated means of the three classes and the proportion of individuals in each class. A similar growth pattern for each class was obtained in the final model as the growth pattern found earlier in the three-class GMM without a covariate and without a distal outcome. The three-class solution suggests the existence of three groups in terms of growth patterns of reading achievement: a class with a high baseline which grew slowly (Class 1), a class with a moderate baseline which grew rapidly (Class 2), and finally, a class with a low baseline which grew slowly (Class3).

There are approximately 24.9% of the students classified in Class 1. The estimated means of reading achievement at initial status for this class was 4.916 and the average growth rate was 0.177. This class seems to start off with a high level of reading aptitude which does not change very much across time. Next, 31.2% of the students were classified in Class 2. The estimated means of reading achievement at initial status for this class was 3.028 and the mean of the growth rate was 1.627. This class seems to start off at a medium range of reading ability which increases rapidly across time compared to other classes. Lastly, there are 43.8% of the students classified in Class 3. The estimated means of reading achievement at initial status for this class was 0.822 and the average growth rate was 0.189. This class seems to start off at a low level of reading performance which increases very little across time, which indicates consistent low performance and possibly disengagement in school. Thus, I would expect the probability of having the negative distal outcome in this class to be higher than any other class.

Figure 3 Estimated means for the 3 class GMM with a covariate and a distal outcome



The effect of the covariate, the poverty status of the student's family, on growth factors was also examined. It was found that the poverty status of the student's family is significantly and negatively associated with the initial growth factor. This implies that groups of students with a greater level of poverty appear to have lower initial status in terms of reading achievement trajectories than students with a smaller level of poverty. However, poverty status itself did not seem to influence the growth rate. Interestingly, the amount of poverty provides additional information regarding the characteristics of the poor developmental class. The multinomial logistic regression for class membership indicates that relative to the class with a high level of reading ability, the odds of membership in the class with a low level of reading achievement ($e(0.833) = 2.300$) are significantly increased by having a greater level of poverty.

Adding a negative distal outcome seemed to help to characterize the poor developmental class in more detail. The overall distal outcome occurrence rate in the control group is 43%. Table 6 displays the probability of the negative distal outcome occurrence for each class when the covariate equals zero and this probability was obtained by as follows:

$$P(u=1 \mid c_i = K, x_i) = \frac{1}{1+e^{-(-\tau_k + \beta_k x_i)}} = \frac{1}{1+e^{-(-\tau_k)}} \ ,$$

The probability of the distal outcome occurrence in the poor developmental class is much higher than in the normative developmental class. The odds of this poor developmental class having a negative distal outcome are 4.55 times larger than for the normative developmental groups (Class 1 and Class 2). This indicates that the negative distal outcome is more likely happen to those who are in the poor developmental class than those who are in the normative developmental group.

Table 6 Occurrence of Distal outcome as a function of poor developmental reading class

Class	Threshold	Probability	Odds	OR
Class 3 (Poor developmental)	-0.498	0.621989	1.645427	4.55
Class 1 & Class 2 (Normal developmental)	1.017	0.265612	0.361678	1

Discussion

This report started with the question that there might be qualitatively different groups in terms of their reading achievement growth patterns. The three-class GMM with a covariate and a distal outcome found three subgroups which have the following reading achievement developmental trajectories: a class with a high baseline which grew slowly (Class 1), a class with a moderate baseline which grew rapidly (Class 2), and finally, a class with a low baseline which grew slowly (Class3). Table 7 compares the growth patterns of subgroups from the three-class GMM with subgroups obtained by the baseline of reading achievement. Although using only the baseline to obtain subgroups can suggest the possibility of the existence of subgroups in terms of the growth pattern over time, it is clear that that method cannot capture the specific growth patterns of subgroups over time which the three-class GMM is able to describe.

The three-class GMM also found that as the greater the level of poverty, the higher the probability of being in a poor developmental group. Moreover, those who are in a poor developmental group are more likely to have a negative distal outcome than those who are in normal developmental groups. It will be of interest to find out how an intervention could affect the development patterns of these three groups – whether prevention boosts the growth rate of Class 2, the rapidly growing group, or increases the rate growth rate of Class 3, the poor developmental group.

Table 7 Comparison subgroups by baseline versus subgroups by 3 class GMM