Growth mixture models outperform simpler clustering algorithms when detecting longitudinal heterogeneity, even with small sample sizes

Daniel P. Martin & Timo von Oertzen

University of Virginia

Department of Psychology

Author Note

A special thanks should be given to Andreas Brandmaier and Steven Boker for providing feedback on earlier versions of this work. Correspondence concerning this paper should be addressed to Daniel P. Martin, University of Virginia, dm4zz@virginia.edu. The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant #R305B090002 to the University of Virginia. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

Abstract

Identifying sub-populations based on longitudinal trajectories can provide new avenues to answer theoretically interesting research questions. While many techniques to accomplish this task exist, a common method used in psychology is the growth mixture model. Recent simulations have found that this analytic method shows a decline in performance for smaller sample sizes commonly found in psychological research (Kim, 2012; Peugh & Fan, 2012). This raises the question: are there better methods available for smaller sample sizes? Monte Carlo simulations were used to explicitly compare growth mixture models with other clustering methods, ranging on a spectrum from not-informed to very-informed, across different simulation conditions. To compare results both between and within analytic method, Kullback-Leibler divergence is introduced as a measure of cluster solution misfit. Results show that despite this decreased performance for smaller sample sizes, growth mixture models still outperform simpler, more general clustering methods.

Keywords: longitudinal clustering, growth heterogeneity, comparative simulation

Growth mixture models outperform simpler clustering algorithms when detecting longitudinal heterogeneity, even with small sample sizes

Who we are now is not the same person as who we were a year, a month, or even a day ago. In a complex world, the use of longitudinal measurement allows psychological researchers to measure this individual change systematically, while such information would remain unknown if measured using a cross-sectional approach (Nesselroade & Ram, 2004). With the goal of evaluating interindividual differences in intraindividual change, longitudinal data are commonly analyzed using a homogeneous growth curve framework (Nesselroade, 1991; MacCallum & Austin, 2000; Raudenbush & Bryk, 2002).

While such a framework is very powerful and intuitive, it makes the statistical assumption that all observations from a particular sample follow a specified longitudinal trend with normally distributed variance. This may not always be the case. For example, Muthén (2004) examined heterogeneity in longitudinal trajectories of high school mathematics achievement and found that three subpopulations existed within the larger sample: poor, moderate, and good development. These groups differed in attrition rates, with low dropout rates in the moderate (8%) and good (1%) development groups and a much higher rate (69%) in the poor development group (Muthén, 2004). Unless this longitudinal heterogeneity was modeled by potential moderators, such an interesting effect with important substantive implications would have remained undiscovered.

Person-Centered Approaches to Model Growth Trajectories

In a more general sense, taking an unsupervised clustering approach can result in identifying such homogeneous subpopulations within a larger heterogeneous sample, thus indicating the need to uncover potential moderators in future research projects. Such methods can also be used in a purely exploratory way to discover new findings in previously collected data. This is particularly useful, given the cost-intensive nature of longitudinal research designs (Swartout, Swartout, & White, 2011).

Longitudinal, person-centered approaches have seen increased popularity in a variety of sub-domains within psychology, such as child development (Jordan, Kaplan, Nabors Oláh, & Locuniak, 2006), psychopathology (Steinman, Hunter, & Teachman, 2013), and substance abuse (Malone, Van Eck, Flory, & Lamis, 2010) to name only a few. This may be due to the increased availability of a recent technique, growth mixture modeling, in popular statistical software packages such as MPlus, SAS, and R (Jung & Wickrama, 2008). Despite this recent increase in popularity for growth mixture modeling, techniques to identify heterogeneity in longitudinal growth trajectories have been around for much longer. For instance, Tucker (1966) introduced Tuckerized growth curves to identify heterogeneity in learning trajectories. Over the years, many different longitudinal clustering techniques have been developed and used successfully to answer questions across a variety of substantive areas (e.g., Brandmaier, von Oertzen, McArdle, & Lindenberger, 2013). A question that remains understudied is how the relative performance of these techniques compare to one another.

Previous Research Evaluating Longitudinal Person-Centered Approaches

The evaluation of longitudinal, person-centered approaches as stand-alone methods are quite common in the literature. The majority of these focus on assessing criteria to choose the correct number of clusters for the analysis in question. For example, Tibshirani, Walther, and Hastie (2001) developed a metric to determine the number of clusters in a k-means algorithm and showed via simulation that it outperformed other widely used metrics for datasets with well-separated clusters. Simulations using traditional cluster algorithms (i.e., k-means, partitioning around medoids, etc.) have received much attention over the years. While the age of the techniques differ, similar simulation research for other person-centered techniques, such as latent variable mixture models, are much more limited. Given the recent popularity of latent variable mixture models, an increased focus is being placed on evaluating the effectiveness of different cluster choice metrics for these methods.

Tofighi and Enders (2008) simulated longitudinal samples with three latent classes and found that the sample size adjusted Bayesian Information Criterion (BIC) identified the correct number of clusters most often. Nylund, Asparouhov, and Muthén (2007) examined metrics for latent class analysis, factor mixture models, and growth mixture models, and found that the bootstrap likelihood ratio test consistently outperformed all metrics, though the BIC performed the best out of the available information criteria. Finally, Peugh and Fan (2012) focused on both homogeneous and heterogeneous samples of longitudinal growth trajectories. Results indicated considerable variability for all metrics depending upon dataset manipulations such as: number of clusters, sample size, and whether trajectory sample size proportions were equivalent (Peugh & Fan, 2012).

More recently, researchers exploring the performance of growth mixture modeling using Monte Carlo simulations have begun to focus on parameter bias and sample size requirements. Depaoli (2013) compared various estimation techniques for GMMs, such as maximum likelihood using the EM algorithm and Bayesian estimation using various priors, for longitudinal data that included nonlinear trends. Results indicated that both accurate information and prior knowledge estimation techniques showed less bias in the recovery of growth parameters, while the other methods resulted in poorer results (Depaoli, 2013). Kim (2012) examined the impact of varying number of indicators and sample size had on performance of both single and multi-phasic growth mixture models. It was shown that large samples were often neccessary to obtain accurate estimates. On cases with low cluster separation and a large number of clusters, sample sizes over 1,200 were cited as being the low threshold (Kim, 2012). Similar conclusions regarding large sample sizes for growth mixture modeling were reached in other studies as well (Peugh & Fan, 2012).

These sample size recommendations are not surprising as latent variable models, such as growth mixture models, are widely regarded as large sample techniques (Kline, 2005). However, other methods that can be used to detect longitudinal heterogeneity, such as k-means cluster analysis, have no such "rule of thumb" for suggested sample size. (Mooi &

Sarstedt, 2011). This raises the question: could other methods with weaker requirements for sample size outperform growth mixture models for smaller sample sizes?

While there are many studies evaluating the effectiveness of individual techniques, both simulations and applications directly comparing different longitudinal person-centered approaches are limited in number. Brossart, Parker, and Willson (1998) compared the application of k-means clustering and Tuckerized growth curves on trajectories for a group conflict scale and found similar results between the two methods. Muthén (2004) used latent variable mixture modeling to compare latent class growth analysis to growth mixture modeling. In a dataset measuring the development of adolescent delinquency, results between the two methods indicated the possibility of different substantive interpretations (Muthén, 2004). Thus, results between these methods have the potential to yield more or less accurate approximations to the data, potentially resulting in different conclusions being formed.

In addition, the evaluation criteria used most often in these longitudinal clustering simulations, proportion of correct cluster choice and parameter bias, have potential limitations. For example, proportion of correct cluster choice yields a binary response of whether a particular clustering enumeration index chose the correct number of clusters *a priori*. In sum, this dichotomization assumes that the cluster choice is either "right" or "wrong." This ignores the possibility that a solution which chooses the wrong number of clusters could actually approximate the true cluster solution well, as well as the possibility that a solution which chooses the correct number of clusters might approximate the true cluster solution poorly. When using parameter bias, it becomes difficult to compute an overall bias metric for all parameters for the purposes of comparison across simulation conditions or analysis methods. A more desirable metric would address these two shortcomings by evaluating a cluster solution regardless of whether the solution chose the correct number of clusters to yield a measure of bias for all parameters in a single number. For such a metric, we propose the use of Kullback-Leibler (KL) divergence. This value

represents the amount of information lost when using one distribution to approximate another (Kullback & Leibler, 1951).

This paper looks to explicitly compare the performance of clustering methods on detecting longitudnal heterogeneity across a variety of sample sizes, with the clustering methods ranging from not-informed (i.e., k-means) to very-informed (i.e., growth mixture models) with regard to longitudinal information. Additionally, both cluster choice accuracy and KL-divergence will be used as evaluation criteria to address the concerns mentioned above. Given the previous research citing larger sample sizes being necessary for growth mixture models, it was hypothesized that simpler, less-informed methods would outperform growth mixture models for very small sample sizes (N = 150). However, as the sample size increased (N > 600), it was also hypothesized that the benefit of explicitly specifying growth would result in better performance for growth mixture models over less-informed methods. To foreshadow the end result, while the traditional evaluation metric of proportion correct cluster choice was somewhat inconclusive, using KL-divergence as a new evaluation metric yielded a much more clear result, with growth mixture models consistently outperforming the other methods, regardless of sample size.

Method

First, the analytic methods used in this comparative simulation will be outlined. After that, the simulation conditions and evaluation criteria of proportion correct cluster choice and KL-divergence will be explained.

Statistical Techniques and Implementation

Each dataset in this simulation was subjected to an automated implementation of all four different clustering methods described below. While each method involved different metrics, the logic remained essentially the same. First, each method used one metric to determine the number of clusters to extract *a priori*. Note that this study was not focused on evaluating enumeration indices, and thus the authors simply chose a widely used

enumeration index for each technique. Two to five clusters were considered for each method. Then, both true and estimated cluster membership were appended to the dataset and evaluated based on two criteria. All simulations and analyses were conducted with R, version 3.0.1 (R Core Team, 2013).

K-Means. K-means is an iterative, unsupervised learning algorithm with the goal of partitioning a particular dataset into k clusters, where k is specified *a priori* (MacQueen, 1967). The iteration follows a simple, four step process:

1. Once k is specified, k cluster centers are placed with random starting values.
2. Each data point is then assigned to the nearest cluster center[1].
3. The centroid of each cluster becomes the new cluster center.
4. Repeat steps 2 and 3 until the algorithm converges.

Due to its simplicity and quick convergence, k-means is a very popular clustering technique in the data mining community, although not necessarily in a longitudinal framework. In its application to longitudinal data, it represents an observed-space, model-free approach. Regardless of implementation, this technique has some potential drawbacks. Due to the nature of the iteration involving cluster centroids, the algorithm typically finds clusters with equivalent covariance matrices. Also, this technique converges to different results for different starting values defined in the first step of the algorithm. Thus, multiple runs with different starting values are needed to ensure the best solution has been found is necessary. For this simulation, the k-means algorithm was implemented twice: once using all five time points, and once using only the first (baseline) time point to serve as a naïve baseline condition to compare to the other analytic techniques. Both applications of k-means followed the same implementation.

Silhouette width was used to determine how many clusters to extract (Rousseeuw, 1987). For each datapoint, i, in a sample, this metric can be calculated using:

$$s(i) = \frac{b(i) - a(i)}{max\left\{a(i), b(i)\right\}},$$

[1]For this simulation, Euclidean distance was used.

where *a(i)* is the within-cluster cohesion of each datapoint with its respective cluster, and *b(i)* is the between cluster separation of each datapoint with the nearest cluster. Silhouette width is then defined as the mean of *s(i)* over the entire sample. Because a goal of clustering is to both minimize within-cluster distance and maximize between-cluster distance, the k cluster solution that yields the largest value of silhouette width is used. Both silhouette width and the k-means algorithm were implemented using the fpc package in R, which utilizes multiple starting values to avoid local minima (Hennig, 2013).

Finite Mixture Models. As mentioned previously, a weakness of k-means is the tendency to extract clusters with similar covariance structures. Finite mixture models (FMMs), and more specifically Gaussian mixture models,[2] are more flexible than k-means, as they include the ability to model different covariance structures among the variables and clusters of interest. Another added benefit of Gaussian mixture models is that they are model-based, and thus yield model-based fit statistics (such as information criteria) to compare the various cluster solutions of varying k (Fraley & Raftery, 2002). As such, in the context of detecting longitudinal heterogeneity, Gaussian mixture models can be thought of as an observed-space model-based approach.

Gaussian mixture models with k latent classes are of the form:

$$p(x) = \sum_{k=1}^{K} \pi_k \, \mathcal{N}(\mu_k, \Sigma_k)$$

where $\mathcal{N}(\mu_k, \Sigma_k)$ represents the Gaussian distribution for each class and π_k represents the mixture weight of one of the distributions and the sum of the mixture weights equals one. Mixture weights and the parameters of each component Gaussian distribution are estimated from the observed variables using an Expectation-Maximization (EM) procedure. A Gaussian mixture model is considered to be a fuzzy clustering technique, because each individual is assigned a probability of belonging to each mixture component.

[2]FMM is being used instead of GMM to avoid confusing Gaussian mixture models with growth mixture models

This is often cited as an advantage over hard clustering techniques such as k-means, where each observation is assigned to only one cluster.

In estimating each Gaussian mixture, the optimal model was chosen according to whichever k and covariance structure yielded the lowest Bayesian Information Criterion (BIC; Schwarz, 1978). The BIC measures the fit of a model using likelihood with a penalty term to reduce overfitting, and is defined by:

$$BIC = -2 \ln L + k \ln n$$

where L is the maximized value of the likelihood function, k is the number of parameters to be estimated in the model, and n is the sample size. BIC has been shown to perform well in choosing the number of latent classes correctly for such models (Jedidi, Jagpal, & DeSarbo, 1997; Roeder & Wasserman, 1997). The mclust package in R was used to estimated finite mixture models in this simulation (Fraley, Raftery, Murphy, & Scrucca, 2012).

Growth Mixture Models. Originally introduced by Muthén and Shedden (1999), growth mixture modeling (GMM) is an extension of growth curve modeling, and allows the estimation of parameters across unobserved subpopulations within the dataset (Muthén, 2004). Through this conceptualization, a growth mixture model is mathematically identical to a multi-group latent growth curve. However, because the population heterogeneity is unobserved, the number of subpopulations needs to be specified by the researcher. A growth mixture model can be considered to be a latent-space model-based approach, and is defined by the following formula:

$$Y_{ti} = \eta_{k0i} + \lambda_t \eta_{k1i} + \epsilon_{kti}$$

$$\eta_{k0i} = \mu_{k0} + \zeta_{k0i}$$

$$\eta_{k1i} = \mu_{k1} + \zeta_{k1i}$$

where Y_{ti} is the observed value for person i at time t, μ_{k0} and μ_{k1} are the mean intercept and slope for each latent subpopulation k and ϵ_{kti} denotes the residual error term for

person i at time t. Additionally, ζ_{k0i} and ζ_{k1i} represents each individual's deviation from the mean intercept and slope, respectively, in a particular latent class k. These deviations have variances σ^2_{k0} and σ^2_{k1} in addition to σ_{k01} as the covariance between them. Without group membership, this is the specification for a latent growth curve model. An EM algorithm is used to simultaneously estimate the specified hierarchical and class membership probabilities (Wang & Bodner, 2007). Again, because this technique involves assigning probabilities to individuals for belonging in each latent class, growth mixture modeling is also a fuzzy clustering technique. Note that discrete class membership can be inferred with maximum likelihood.

Since the data for this simulation are all based on linear change, this was the model that was specified for all GMMs. Coming up with one metric to evaluate different clustering solutions for GMMs is difficult, as there is still no commonly accepted enumeration index for choosing the correct number of clusters. Nylund et al. (2007) have shown via simulation that the BIC performed the best among fit statistics, though it was outperformed by the bootstrap likelihood ratio test. Other simulation studies have reported more inconclusive results, with no index acting clearly superior (Peugh & Fan, 2012). Again, as this study was not focused on evaluating enumeration indices, the BIC was used as an individual metric to select the number of clusters. Each model was built and estimated using the lcmm package in R (Proust-Lima, Philipps, Diakite, & Liquet, 2013). Additionally, because GMMs can converge to local minima, each analysis was run multiple times with random starting values to increase the probability that a global minimum had been reached.

Evaluation Criteria

Evaluating correct choice of cluster number via various enumeration indices is a popular method to evaluate the effectiveness of clustering techniques (Tibshirani et al., 2001; Nylund et al., 2007; Peugh & Fan, 2012). However, focusing on this evaluation criterion leaves the following question unanswered: How close might a potential result

approximate the "true" cluster solution despite choosing the incorrect number of clusters *a priori*? One can certainly imagine a situation where choosing the incorrect number of clusters might result in a fairly good approximation, or another where choosing the correct number of clusters might actually result in a poor approximation. For this reason, both correct cluster choice and cluster solution misfit were included as evaluation criteria and are detailed below.

Choice of Number of Clusters. For all conditions, clustering was performed with two to five clusters and the method specific index was used to identify the optimal number of clusters. An overall percentage was calculated for each condition indicating the proportion of datasets within that condition that yielded the correct choice for number of clusters.

Cluster Solution Misfit. Cluster solution misfit was measured for intercept and slope parameter estimates using Kullback-Leibler (KL) divergence, a measure of relative entropy. For distributions P and Q of a continuous random variable, KL divergence is defined as:

$$D_{KL}(P||Q) = \int_{-\infty}^{+\infty} ln\left(\frac{p(x)}{q(x)}\right) p(x)dx,$$

where $p(x)$ represents the population distribution and $q(x)$ represents a model, or approximation, of $p(x)$. The KL metric can then be interpreted as the information lost when $q(x)$ is used to approximate $p(x)$, and thus lower values of KL-divergence indicate a better approximatation of $p(x)$ (Kullback & Leibler, 1951). Because the interest of this article is clustering, both $p(x)$ and $q(x)$ can be assumed to be a Gaussian mixture model, where $p(x)$ is the dataset with all datapoints correctly classified and $q(x)$ is the estimated cluster solution. Unfortunately, no closed form of KL-divergence exists for Gaussian mixtures. However, $D_{KL} = \mathbb{E}_p[log(\frac{p(x)}{q(x)})]$, and so it can be estimated using a Monte Carlo approximation with the following steps:

1. Draw n independent samples from $p(x)$

2. Compute $\widehat{D_{KL}}(P||Q) = \frac{1}{n}\sum_{i=1}^{n} log\left(\frac{p(x)}{q(x)}\right)$

Note that very large sample sizes are neccessary to calculate an accurate approximate for KL-divergence, and so $3 \cdot 10^6$ independent samples were drawn from the population distribution for each dataset (Hershey & Olsen, 2007).

Simulation Conditions

Monte Carlo simulations were used in order to evaluate two evaluation criteria in relation to four different clustering methods. Datasets were simulated from a heterogeneous longitudinal population with two or three clusters. The total sample size was varied between conditions to be either 150, 300, 600, 900, 1200, or 1500. As such, within-cluster sample size in each condition was $\frac{N}{K}$, where N is the total sample size and K is the number of clusters. It was hypothesized that all clustering methods would perform best in conditions with less clusters and larger sample sizes. In addition, there were three cases of growth trajectory patterns in the heterogeneous conditions adapted from Peugh and Fan (2012): the same intercepts and different slopes, the same slopes and different intercepts, and different intercepts and different slopes. The three cases are presented in Figure 1.

XXXXXXXXXXXXX

X Figure 1 about here

XXXXXXXXXXXXX

All observations were simulated to follow a linear growth curve (McArdle & Epstein, 1987) with five time points and no missing observations[3]. The simulated model is presented in a path diagram in Figure 2. Intercept variance was fixed to 0.3, slope variance was fixed to 0.2, and the correlation between intercept and slope was fixed to 0.2 (and so the covariance between intercept and slope is approximately 0.05) for all simulation

[3]Some methods investigated here have no straightforward way to hangle missingness. Thus, we assumed full data for this study.

conditions. Additionally, error variance was simulated to be 0.2 for all five time points[4].

XXXXXXXXXXXXX

X Figure 2 about here

XXXXXXXXXXXXX

The degree of separation between clusters was manipulated in multivariate Mahalanobis distance (M distance) units, defined as:

$$D^2 = (\mu_{Trajectory_1} - \mu_{Trajectory_2})^T \Sigma^{-1} (\mu_{Trajectory_1} - \mu_{Trajectory_2}),$$

where μ is intercept and slope vector for each cluster trajectory, and Σ is the covariance matrix between them. Clusters were separated by $D = 1.5$ or $D = 3$. In the three cluster case, the nearest clusters would vary by either 1.5 or 3 D units as well, indicating that the furthest clusters would vary by 3 or 6 D units, respectively[5]. Because M distance represents a multivariate effect size between clusters, it was expected that a larger value would result in a higher likelihood of choosing the correct number of classes and less cluster misfit measured by KL-divergence.

In total, this simulation had four different manipulations (sample size, M distance, number of clusters, and intercept-slope relationship) implemented in a fully crossed design. This resulted in the creation of 72 (6 x 2 x 2 x 3) unique dataset conditions. Each condition was then simulated 200 times for a total of 14,400 sample datasets.

Results

First, simulation results will be presented for all conditions by both method and sample size. Then, results will be presented by condition manipulation (number of clusters, M distance, and intercept-slope relationship). Both metrics of correct cluster choice and KL-divergence will be presented together.

[4]Both slope and error variance were manipulated in previous simulations, but were found to have no differential effect on method and were not included.

[5]Note these differences are expressed in D units, not D^2 units.

Overall

As expected, the silhouette width metric for k-means on full data was stable regardless of sample size when determining the correct number of clusters. The other three methods showed increases along with sample size. While growth mixture models performed best for sample size conditions <600, finite mixture models performed better for larger sample sizes. Given that growth mixture models had the most longitudinal information, this result is surprising and will be discussed in conjunction with the KL-divergence evaluation metric below. In terms of the cluster choice metric, both finite and growth mixture models performed the best. Refer to Figure 3 for the results for correct cluster choice by sample size and method over all conditions.

XXXXXXXXXXXXX

X Figure 3 about here

XXXXXXXXXXXXX

KL-divergence depicts a much clearer picture of the relative performance of these methods. Overall, all methods follow a similar trend, in that they show better approximation to the population cluster solution as the sample size increases. Contrary to the original hypothesis, k-means on baseline performs the worst, followed by k-means on full data, finite mixture models, and growth mixture models. At no point are growth mixture models inferior to the other three methods across sample size. This is an unsurprising finding given that growth mixture models have the most longitudinal information available with a correctly specified growth model. However, despite being outperformed by finite mixture models with regard to the correct cluster choice metric, growth mixture models show relatively stable KL-divergence results across sample size conditions and consistently outperform all other methods. This results indicates that even with the incorrect number of clusters, growth mixture models still approximate the true cluster distribution the best. Refer to Figure 4 for these results.

XXXXXXXXXXXXX

X Figure 4 about here

XXXXXXXXXXXXX

Number of Clusters

A clear dichotmomy emerges when examining the results by the number of cluster conditions. In conditions where only two clusters exist in the true solution, k-means on full data, finite mixture models, and growth mixture models all perform at similar levels, choosing the correct number greater than 90% of the time. This changes substantially, however, on conditions with three clusters. Both k-means on baseline and full data show very poor performance on the cluster choice metric for these conditions, with both methods actually getting worse as sample size increases. Both finite and growth mixture models, on the other hand, show modest increases, reaching approximately 30-40% in the larger sample size conditions. Again a similar pattern emerges where growth mixture models yield the best performance on the cluster choice metric until the sample size exceeds 600, when finite mixture models then has the best performance. Refer to Figure 5 for these results.

XXXXXXXXXXXXX

X Figure 5 about here

XXXXXXXXXXXXX

When examining KL-divergence by number of cluster, the same pattern emerges. In this case, both k-means on baseline and k-means on full data show remarkably similar performance in the two cluster condition. Both finite and growth mixtures are more separated, but stable as sample size increases. As expected, all four methods show worse KL-divergence when there are three clusters rather than two. Again, the same pattern exists for the three cluster solution as well. Refer to Figure 6 for these results.

XXXXXXXXXXXXX

X Figure 6 about here

XXXXXXXXXXXXX

Mahalanobis Distance

For the correct cluster choice metric, the Mahalanobis distance manipulation shows a similar pattern compared to the number of clusters manipulation in relation to sample size. That is, in the smaller cluster separation condition (M = 1.5), all methods with the exception of *k*-means on baseline are relatively stable. This changes in the larger cluster separation condition (M = 3), where only finite mixtures and growth mixtures show increases with sample size, while both *k*-means methods are again stable. An interesting result here is how both *k*-means methods are relatively stable across sample size with a proportion cluster correct of about 0.50. This indicates that, regardless of cluster separation, the silhouette width metric is most likely going to choose the smallest k available (two in this case). This differs from BIC, which while stable across N for smaller cluster separation, shows substantial gains as N increases for the larger cluster separation condition. Finite mixture models actually get close to 100% cluster correct in the largest sample size, larger cluster separation condition. Refer to Figure 7 for these results.

XXXXXXXXXXXXX

X Figure 7 about here

XXXXXXXXXXXXX

The same pattern for KL-divergence is seen when examining results by the cluster separation condition, with one exception: *k*-means on baseline actually outperformed *k*-means on full data in the smaller cluster separation. In addition, both *k*-means on baseline and *k*-means on full data actually show worse cluster misfit in the larger cluster separation condition compared to the smaller separation condtion. Both finite mixture models and growth mixture models, on the other hand, showed better cluster misfit in the larger cluster separation condition, especially in those conditions with a larger sample size. Again, despite being outperformed by finite mixture models in the correct cluster choice metric with larger samples, growth mixtures again shows a much better approximation in the large cluster separation condition. Refer to Figure 8 for these results.

XXXXXXXXXXXXX

X Figure 8 about here

XXXXXXXXXXXXX

Intercept-Slope Relationship

For the correct cluster choice metric, finite mixture models and growth mixture models show similar patterns found in previous conditions, where growth mixture models perform the best up to the sample size condition of 600, where finite mixture models are then performing best. Both k-means on baseline and k-means on full data are also showing similar patterns in previous conditions, such that k-means on full data is relatively stable at approximately 0.50 proportion correct, while k-means on baseline is much worse for smaller sample sizes, but increases to the same level in larger sample size conditions. k-means on baseline actually outperforms k-means on full data when the clusters in the dataset have different intercepts and the same slopes. Refer to Figure 9 for these results.

XXXXXXXXXXXXX

X Figure 9 about here

XXXXXXXXXXXXX

Again, all methods show similar pattern for KL-divergence when examining the manipulation of intercept-slope relationship. Not surprisingly, k-means on baseline performs very poorly when cluster intercepts are simulated to be the same. However, when cluster intercepts are the same and cluster slopes are different, k-means on baseline shows a similar amount of misfit to finite mixture models. Growth mixture models are relatively stable across both sample size and intercept-slope relationship as before, and consistently outperform all other methods. Refer to Figure 10 for these results.

XXXXXXXXXXXXX

X Figure 10 about here

XXXXXXXXXXXXX

Discussion

The purpose of this simulation was to investigate the relative performance of growth mixture models compared to other methods across a variety of sample sizes. As Peugh and Fan (2012) note, sample sizes under 500 are considered to be small for homogeneous structural equation modeling in some situations, and so heterogeneous structural equation modeling might require even larger samples to yield accurate parameter estimates. As such, it was hypothesized the growth mixture models would be outperformed by other methods for smaller sample sizes due to the complexity of the estimation procedure. This simulation was also the first to employ KL-divergence as a means to yield an aggregated metric with the ability to evaluate cluster solution misfit, even when the number of clusters chosen *a priori* were chosen incorrectly.

Overall, there were three main findings of note. First, results indicated that growth mixture models clearly outperformed the other three methods across all sample sizes based on the two metrics of proportion cluster choice correct and cluster solution misfit measured via a Monte Carlo approximation of KL-divergence. However, these two metrics told somewhat different stories. The cluster choice metric implied that growth mixture models performed the best for smaller sample sizes, while finite mixture chose to correct number of clusters for larger sample sizes. Contrary to this, KL divergence concluded that growth mixture models consistently yielded the lowest KL-divergence value, indicating that growth mixtures held the best approximation of the covariance matrix between the intercept and slope parameter estimates. KL-divergence for growth mixture models was also shown to be relatively stable as sample size increased, providing some evidence that even for smaller sample sizes, growth mixture models yield fairly accurate parameter estimates.

Second, the silhouette width metric for both k-means on baseline and k-means on full data typically chose the smallest number of clusters available, regardless of sample size. Both finite mixture models and growth mixture models similarly showed poor performance under datasets with a simulated three cluster solution, but only for smaller sample sizes.

The BIC metric used for both showed substantial gains in performance as sample size increased. This replicates previous, cross-sectional findings that show that probabilistic clustering methods, such as Gaussian mixture models, typically outperform more traditional clustering methods, such as k-means (Magidson & Vermunt, 2002).

Third, growth mixture models actually show a moderate decline in choosing the correct number of cluster for larger sample sizes, contrary to what has been found in previous research (Peugh & Fan, 2012). This is most likely due the software used for this simulation, rather than an artifact for the method of growth mixture modeling. For larger sample sizes, the lcmm package yielded missing values for the KL-divergence calculation in about 4% of datasets. This was due to a non-positive-definite covariance matrix calculated because of a very low sample size in one of the clusters, which indicates a higher rate of solutions converging to local minima. However, excluding these exact same datasets for the other three methods yielded no difference in results shown in this paper.

Limitations and Future Directions

In order to limit this simulation from getting too unwieldy, many simulation conditions of interest were not investigated. For instance, both cluster sample size ratio and covariance structure within each dataset were simulated to be equivalent. A much more common occurrence in substantive applications of these methods is to have clusters of varying sizes and shapes. If this were to be the case, it would be expected that both mixture modeling methods would continue to outperform k-means, as latent variable mixture models have the added flexibility of modeling different covariance structures for each cluster within a dataset. Additionally, only one cluster choice metric was used to determine the correct number of clusters *a priori*. In this case, silhouette width was used for k-means and BIC was used for both finite and growth mixture models. In reality, many different indices exist for these methods and are used in tandem with one another to determine the best choice.

As such, a future research idea that emerged from this simulation is to utilize KL-divergence as an evaluation metric to revisit old research questions. For example, Peugh and Fan (2012) compared various enumeration indices to determine which ones yielded the largest proportion correct cluster choice correct and found the results varied depending upon the specific simulation condition. Researchers could revisit these enumeration indices to see if KL-divergence shows clearer patterns, much like the overall result in this paper.

Additionally, data were simulated based on a latent growth curve with a specified linear trend. A benefit of both k-means and finite mixture models is that they can detect and identify possible nonlinear trends with no requirement of model specification, because both ignore the time-dependency in the data. Growth mixture models, on the other hand, require a particular growth model to be specified first. In this study, growth mixture models had the added benefit of being correctly specified in all conditions. Future research could examine the impact mispecification could have on growth mixture models, by comparing how misspecification affects both parameter estimates and cluster choice. Of particular interest would be to examine the use of finite mixture models as a more exploratory method of detecting longitudinal heterogeneity when researchers are unsure if heterogeneity exists, and compare it to the estimation of growth mixture models with a free latent basis (Grimm, 2007)

Conclusion

While the number of simulations regarding GMMs seems to increase as the popularity of the method also increases, many other techniques to identify potential heterogeneity in longitudinal growth trajectories also exist. This simulation showed that even with the difficulties of growth mixture models with smaller sample sizes noted in previous research, growth mixture models still consistently outperform other methods that set out to accomplish the same thing. Additionally, KL-divergence emerged as a valuable metric to evaluate clustering solutions for methods to detect longitudinal heterogeneity

that should prove useful in future simulation research on this topic. Still, more research must be done in relation to how these methods perform in comparison to one another under different simulated dataset conditions. By continuing to uncover this information, researchers will have a better idea of what types of analyses to use in particular situations to yield more accurate predictions.

References

Brandmaier, A. M., von Oertzen, T., McArdle, J. J., & Lindenberger, U. (2013). Structural equation model trees. *Psychological Methods*, *18*(1), 71–86.

Brossart, D., Parker, R., & Willson, V. (1998). A comparison of two methods for analyzing longitudinal data: Tuckerized growth curves and an application of K means analysis. *Learning and Individual Differences*, *1*(2), 121–136.

Depaoli, S. (2013). Mixture class recovery in GMM under varying degrees of class separation: Frequentist versus Bayesian estimation. *Psychological Methods*, *18*(2), 186–219.

Fraley, C., Raftery, A., Murphy, T., & Scrucca, L. (2012). mclust version 4 for R: Normal mixture modeling for model-based clustering, classification, and density estimation. *Technical Report No. 597*.

Fraley, C., & Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, *97*(458), 611–631.

Grimm, K. J. (2007). Multivariate longitudinal methods for studying developmental relationships between depression and academic achievement. *International Journal of Behavioral Development*, *31*(4), 328–339.

Hennig, C. (2013). fpc: Flexible procedures for clustering [Computer software manual]. Retrieved from `http://cran.r-project.org/package=fpc`

Hershey, J. R., & Olsen, P. A. (2007). Approximating the Kullback-Leibler divergence between Gaussian mixture models. In *Acoustics, speech and signal processing* (Vol. 4, pp. IV–317).

Jedidi, K., Jagpal, H., & DeSarbo, W. (1997). Finite-mixture structural equation models for response-based segmentation and unobserved heterogeneity. *Marketing Science*, *16*(1), 39–59.

Jordan, N. C., Kaplan, D., Nabors Oláh, L., & Locuniak, M. N. (2006). Number sense

growth in kindergarten: A longitudinal investigation of children at risk for mathematics difficulties. *Child development*, *77*(1), 153–175.

Jung, T., & Wickrama, K. A. S. (2008). An introduction to latent class growth analysis and growth mixture modeling. *Social and Personality Psychology Compass*, *2*(1), 302–317.

Kim, S.-Y. (2012). Sample size requirements in single- and multiphase growth mixture models: A Monte Carlo simulation study. *Structural Equation Modeling: A Multidisciplinary Journal*, *19*(3), 457–476.

Kline, R. B. (2005). *Principles and practice of structural equation modeling* (2nd ed.). New York: Guilford Press.

Kullback, S., & Leibler, R. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, *22*(1), 79–86.

MacCallum, R., & Austin, J. (2000). Applications of structural equation modeling in psychological research. *Annual Review of Psychology*, 201–226.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In L. M. Le Cam & J. Neyman (Eds.), *Proceedings of the fifth berkeley symposium on mathematical statistics and probability* (pp. 281–297). Berkeley, CA: University of California Press.

Magidson, J., & Vermunt, J. (2002). Latent class models for clustering: A comparison with k-means. *Canadian Journal of Marketing*, *20*(1), 36–43.

Malone, P. S., Van Eck, K., Flory, K., & Lamis, D. a. (2010). A mixture-model approach to linking ADHD to adolescent onset of illicit drug use. *Developmental Psychology*, *46*(6), 1543–1555.

McArdle, J., & Epstein, D. (1987). Latent Growth Curves within Developmental Structural Equation Models. *Child development*, *58*(1), 110–133.

Mooi, E., & Sarstedt, M. (2011). *A concise guide to market research*. Berlin, Germany: Springer-Verlag.

Muthén, B. O. (2004). Growth mixture modeling and related techniques for longitudinal data. In D. Kaplan (Ed.), *The sage handbook of quantitative methodology for the social sciences* (pp. 345–368). Thousand Oaks, CA: Sage Publications.

Muthén, B. O., & Shedden, K. (1999). Finite mixture modeling with mixture outcomes using the EM algorithm. *Biometrics*, *55*(2), 463–469.

Nesselroade, J. R. (1991). Interindividual differences in intraindividual change. In L. A. Collins & J. L. Horn (Eds.), *Best methods for the analysis of change* (pp. 92–106). Washington, DC: American Psychological Association.

Nesselroade, J. R., & Ram, N. (2004). Studying intraindividual variability: What we have learned that will help us understand lives in context. *Research in Human Development*, *1*(1-2), 9–29.

Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(4), 535–569.

Peugh, J., & Fan, X. (2012). How well does growth mixture modeling identify heterogeneous growth trajectories? A simulation study examining GMM's performance characteristics. *Structural Equation Modeling: A Multidisciplinary Journal*, *19*(2), 204–226.

Proust-Lima, C., Philipps, V., Diakite, A., & Liquet, B. (2013). lcmm: Estimation of latent class mixed models, joint latent class mixed models and mixed models for curvilinear outcomes [Computer software manual]. Retrieved from `http://cran.r-project.org/package=lcmm`

R Core Team. (2013). R: A Language and Environment for Statistical Computing [Computer software manual]. Vienna, Austria. Retrieved from `http://www.r-project.org/`

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Roeder, K., & Wasserman, L. (1997). Practical Bayesian density estimation using mixtures of normals. *Journal of the American Statistical Association*, *92*(439), 894–902.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464.

Steinman, S. A., Hunter, M. D., & Teachman, B. A. (2013). Do patterns of change during treatment for panic disorder predict future panic symptoms? *Journal of Behavior Therapy and Experimental Psychiatry*, *44*(2), 150–157.

Swartout, A. G., Swartout, K. M., & White, J. W. (2011). What your data didn't tell you the first time around: Advanced approaches to longitudinal data analyses. *Violence Against Women*, *17*(3), 309–321.

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B*, *63*(2), 411–423.

Tofighi, D., & Enders, C. K. (2008). Identifying the correct number of classes in growth mixture models. In G. R. Hancock & K. M. Samuelsen (Eds.), *Advances in latent variable mixture models* (pp. 317–341). Charlotte, NC: Information Age.

Tucker, L. R. (1966). Learning theory and multivariate experiment: Illustration by determination of generalized learning curves. In R. B. Cattell (Ed.), *Handbook of multivariate experimental psychology* (pp. 476–501). New York: Rand McNally.

Wang, M., & Bodner, T. E. (2007). Growth mixture modeling: Identifying and predicting unobserved subpopulations with longitudinal data. *Organizational Research Methods*, *10*(4), 635–656.

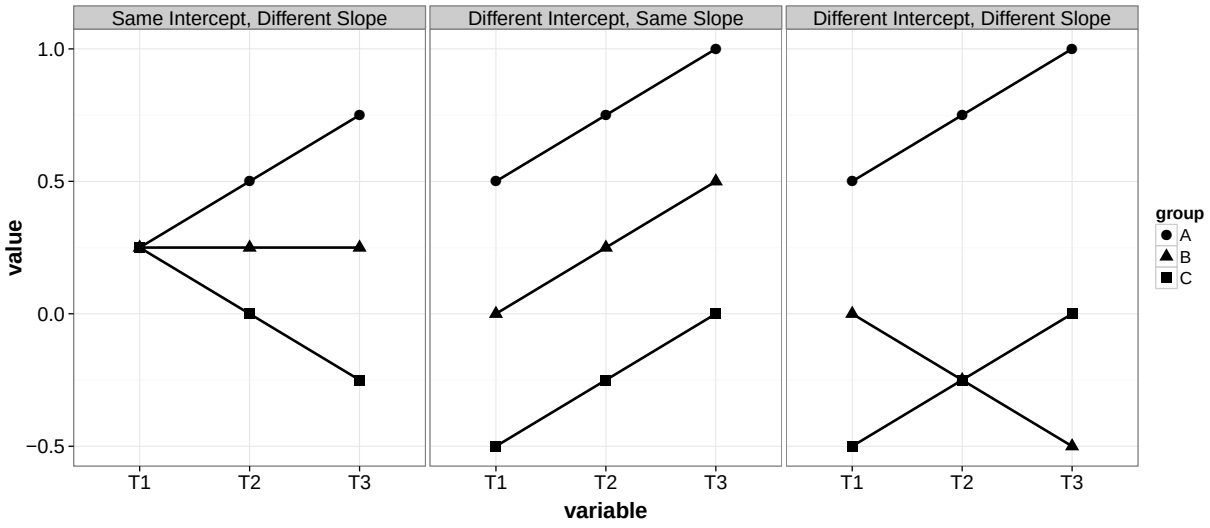

Figure 1. The three simulated intercept-slope relationships.

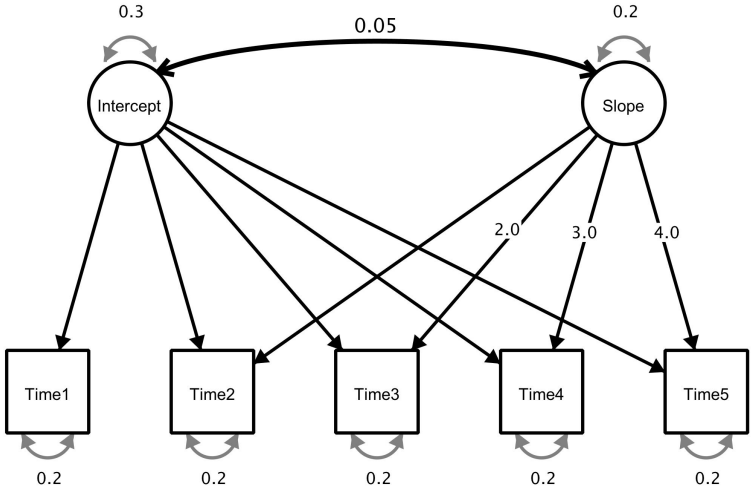

Figure 2. The simulated latent growth model. Note that the correlation was fixed to be 0.2, and so the covariance was fixed to be approximately 0.05

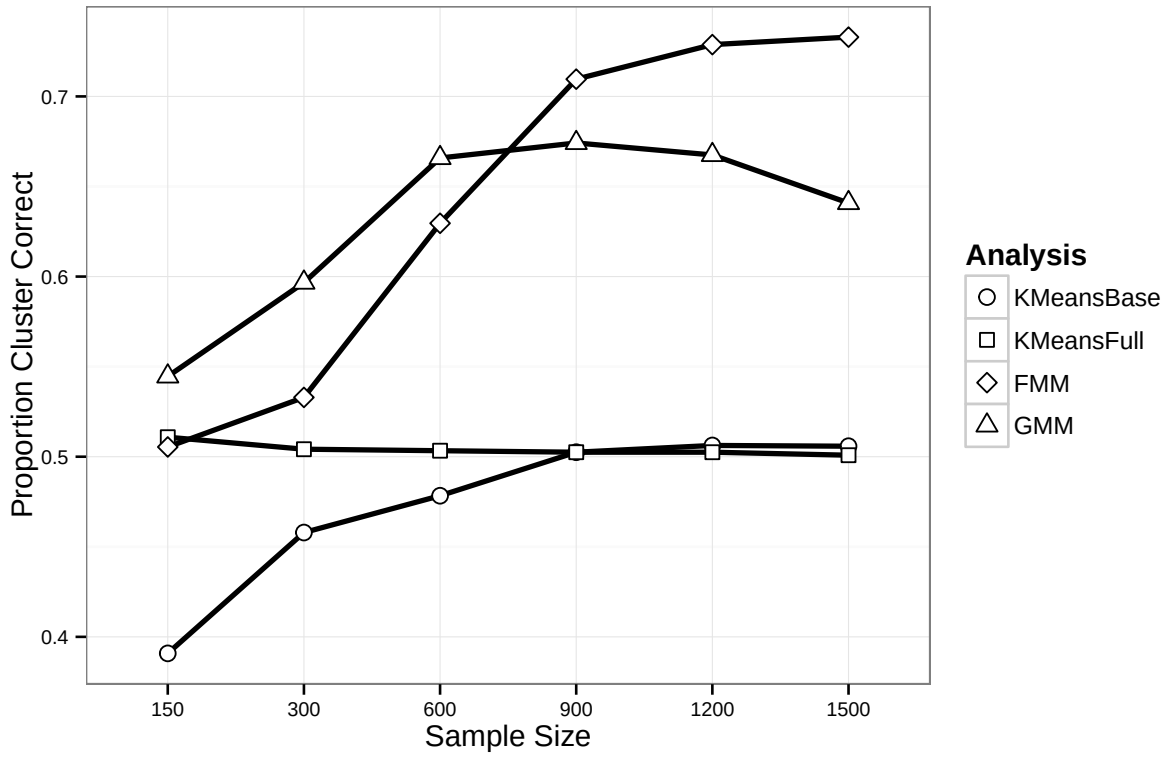

Figure 3. Cluster choice correct metric by sample size and analytic method over all 72 dataset conditions.

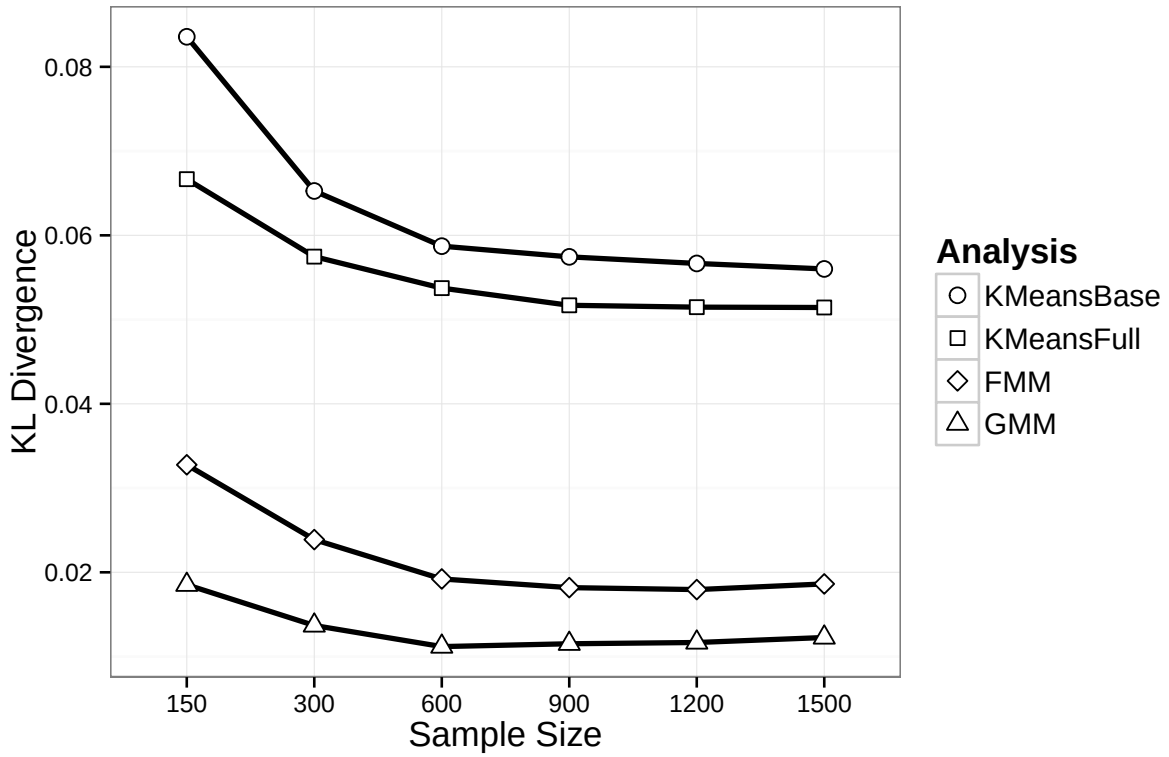

Figure 4. KL-divergence metric by sample size and analytic method over all 72 dataset conditions.

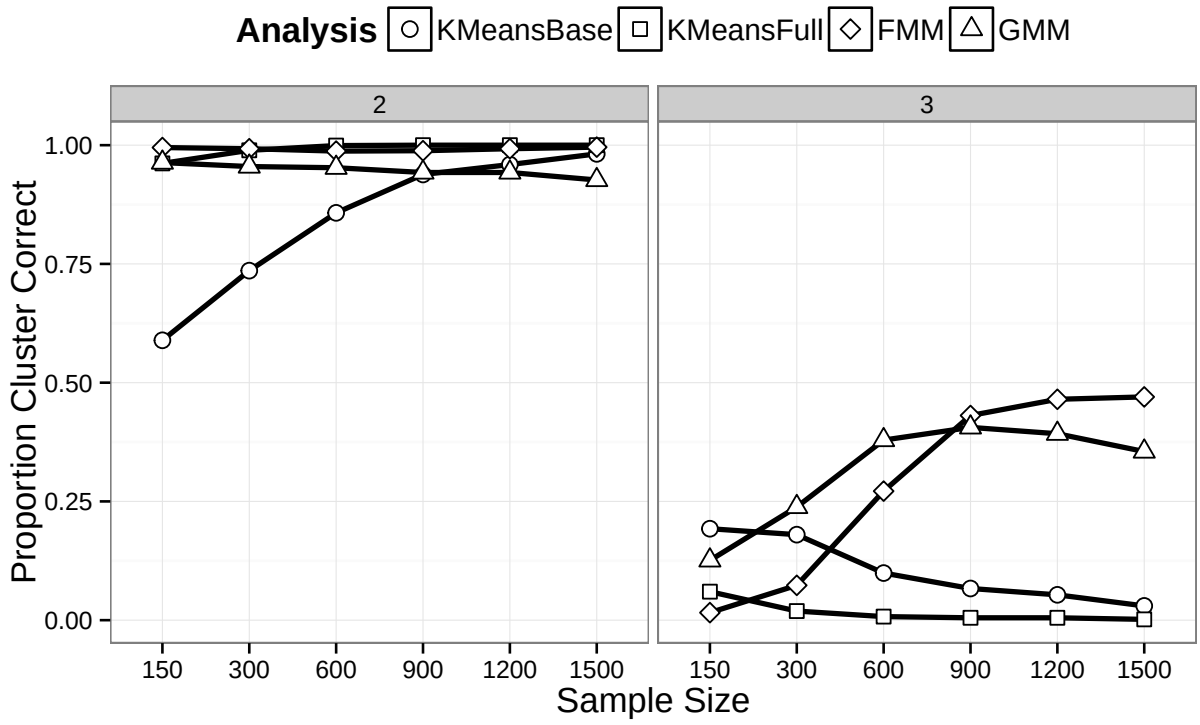

Figure 5. Cluster choice correct metric by sample size, analytic method, and number of clusters.

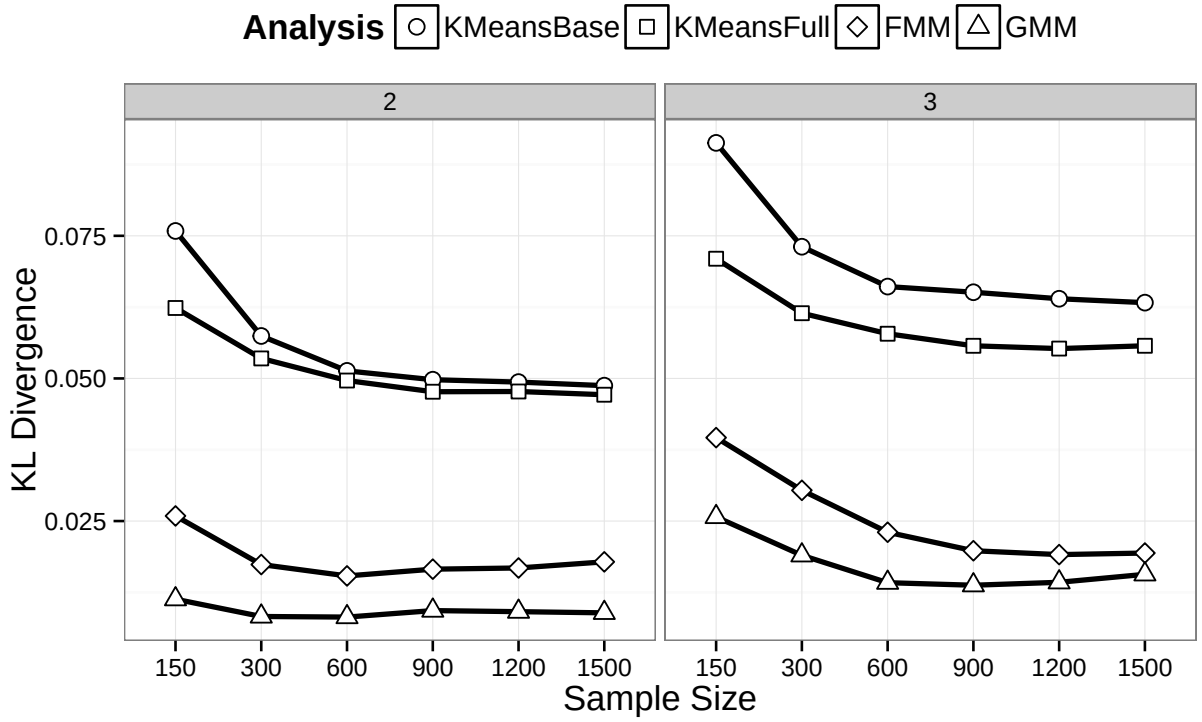

Figure 6. KL-divergence metric by sample size, analytic method, and number of clusters.

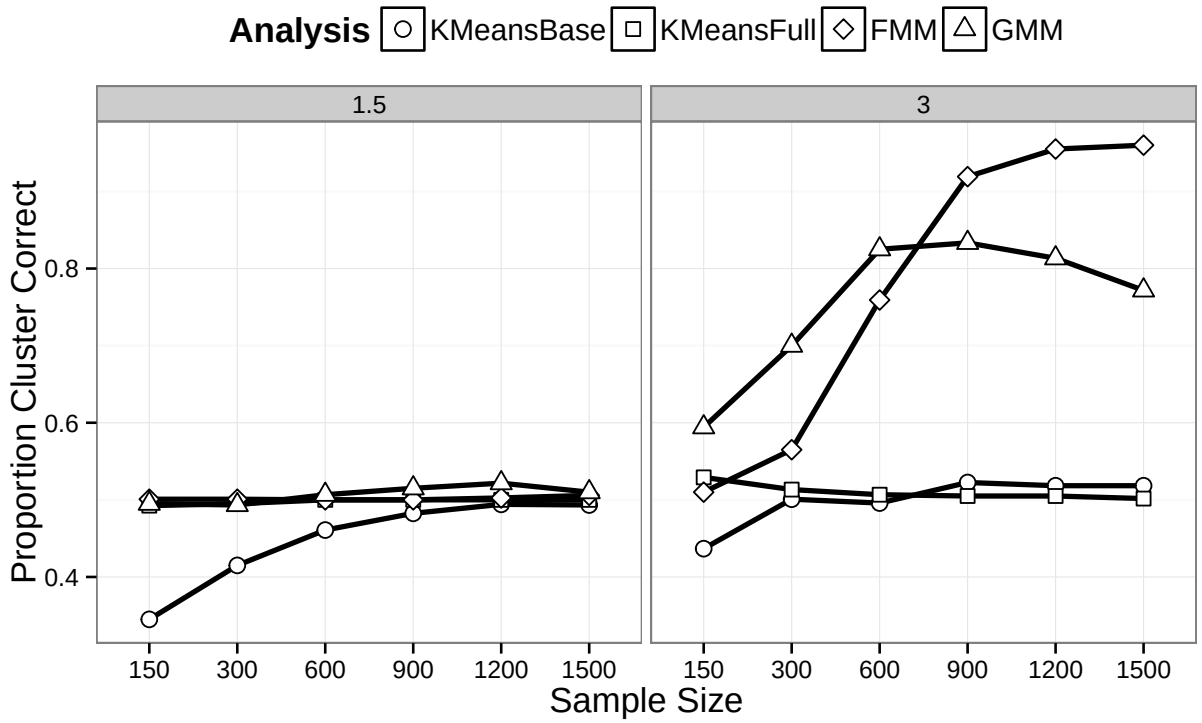

Figure 7. Cluster choice correct metric by sample size, analytic method, and Mahalanobis distance.

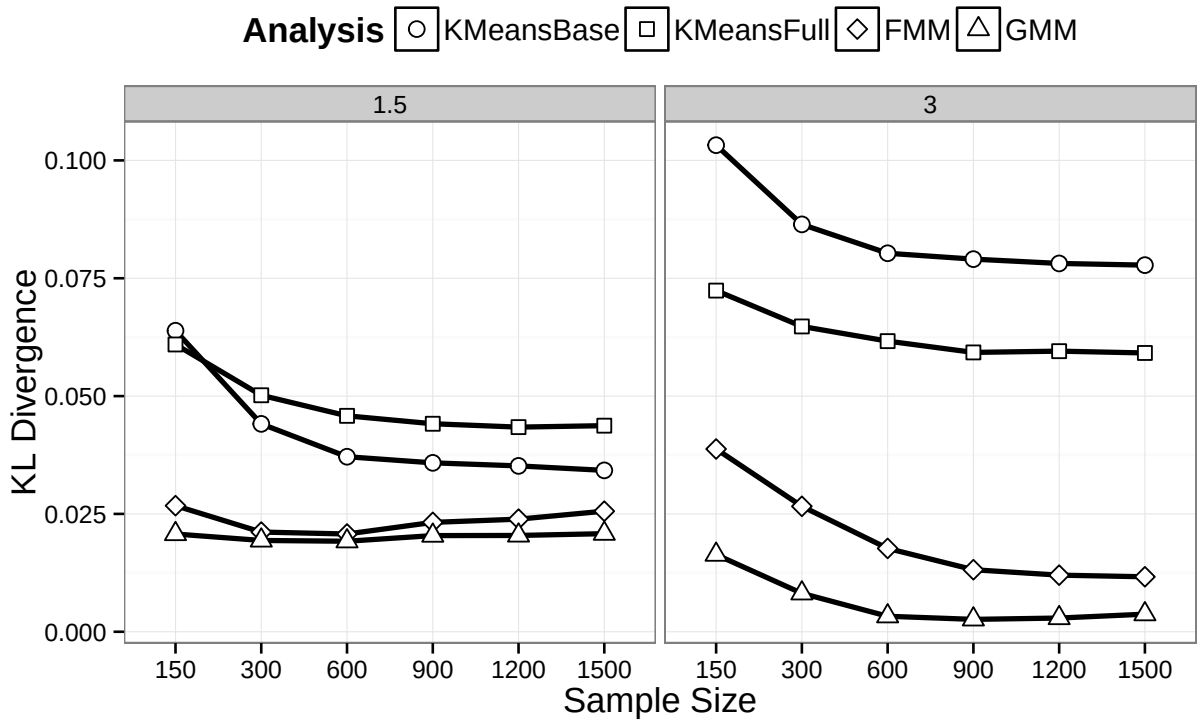

Figure 8. KL-divergence metric by sample size, analytic method, and Mahalanobis distance.

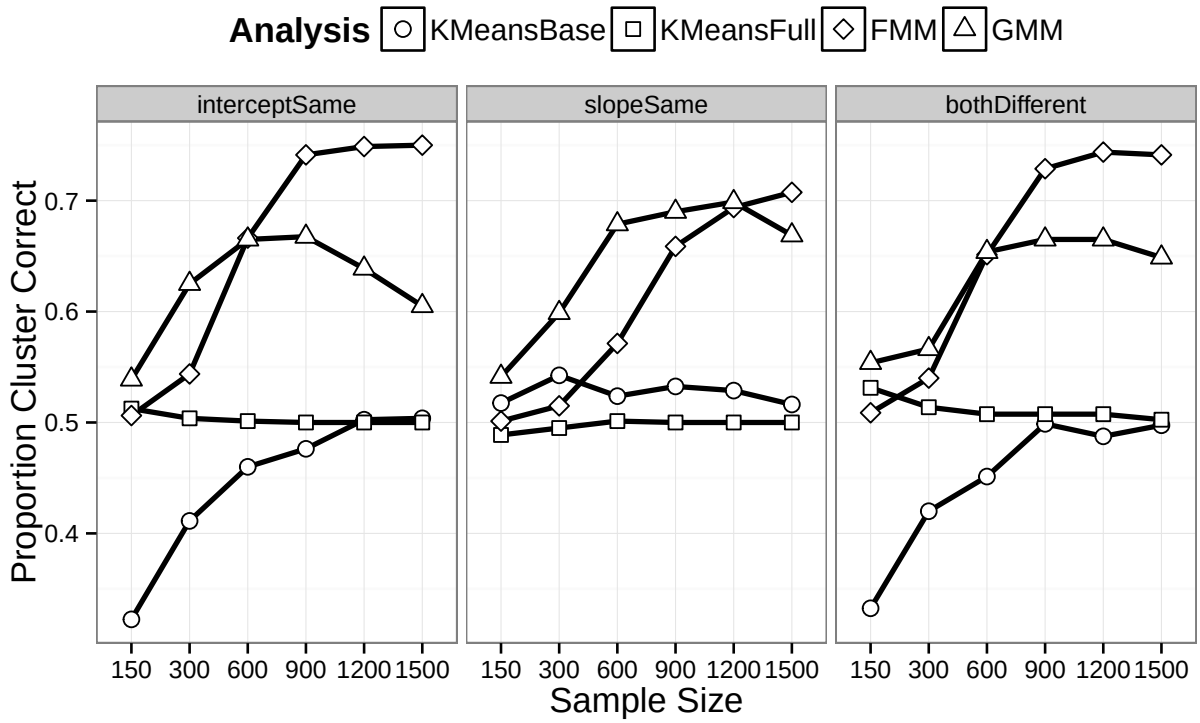

Figure 9. Cluster choice correct metric by sample size, analytic method, and intercept-slope relationship.

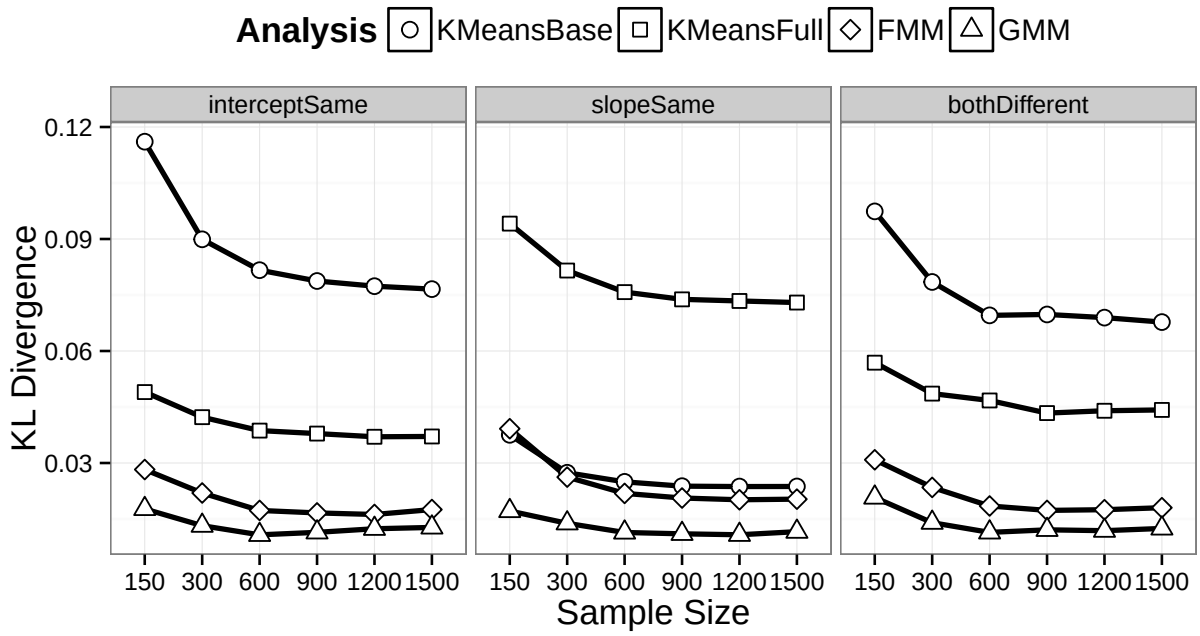

Figure 10. KL-divergence metric by sample size, analytic method, and intercept-slope relationship.