Nonconvergence, Covariance Constraints, and Class Enumeration in Growth Mixture Models

Daniel McNeish[1], Jeffrey R. Harring[2], & Daniel J. Bauer[3]

[1]Arizona State University
Department of Psychology
Tempe, AZ, USA

[2]University of Maryland, College Park
Department of Human Development and Quantitative Methodology
College Park, MD, USA

[3]University of North Carolina, Chapel Hill
Department of Psychology and Neuroscience
Chapel Hill, NC, USA

Accepted at *Psychological Methods*

Contact Information
Daniel McNeish, PO Box 871104, Arizona State University, Tempe, AZ 85287. Email: dmcneish@asu.edu

Acknowledgements
The authors were supported by the Institute of Educational Sciences program in statistical and research methodology, project number R305D190011. The content is solely the responsibility of the authors and does not represent the official views of the Institute of Educational Sciences.

The data in the empirical were supplied by the original authors with a Creative Common BY 4.0 International license that allows use provided that proper attribution is provided to the original authors. We thank these authors for their transparency and commitment to making scientific research an open process.

Supplemental files including all data simulation and analysis files can be found on the Open Science Framework page for this project, https://osf.io/eay5v/.

Abstract

Growth mixture models (GMMs) are a popular method to identify latent classes of growth trajectories. One shortcoming of GMMs is nonconvergence, which often leads researchers to apply covariance equality constraints to simplify estimation, though this may be a dubious assumption. Alternative model specifications have been proposed to reduce nonconvergence without imposing covariance equality constraints that perform well when the correct number of classes is known, but research has not yet examined their use when the number of classes is unknown. Given the importance of selecting the number of classes, more information about class enumeration performance is crucial to assess the potential utility of these methods. We conduct an extensive simulation to explore class enumeration and classification accuracy of model specifications that are more robust to nonconvergence. Results show that the typical approach of applying covariance equality constraints performs quite poorly. Instead, we recommended covariance pattern GMMs because they (a) had the highest convergence rates, (b) were most likely to identify the correct number of classes, and (c) had the highest classification accuracy in many conditions, even with modest sample sizes. An analysis of empirical PTSD data is provided to show that the typical 4-Class solution found in many empirical PTSD studies may be an artefact of the covariance equality constraint method that has permeated this literature.

Nonconvergence, Covariance Constraints, and Class Enumeration in Growth Mixture Models

Growth models are a common group of statistical methods applied to repeated measures data when the interest is in quantifying mean change and individual differences in growth trajectories (Hedeker & Gibbons, 2006; Nesselroade, 1991; Ram & Grimm, 2007). One way to model growth trajectory heterogeneity is to identify unobserved, latent classes of growth trajectories from the data (Jung & Wickrama, 2008). The goal is similar to including a moderator for growth like sex or treatment condition to allow different growth trajectories for different types of people. However, the moderator in this case is latent and not known a priori. Discrete latent classes of growth trajectories are determined from the data by combining latent class analysis with growth modeling in what have generally been deemed *growth mixture models* (GMMs; Muthén & Shedden, 1999; Verbeke & Lesaffre, 1996)

A key aspect of latent class analysis broadly (including GMMs) revolves around identifying the number of latent classes underlying the data (e.g., Steinley & Brusco, 2011); however, because the classes are unobserved, class enumeration is a difficult process (Tofighi & Enders, 2008, p. 316). The difficulty of class enumeration is heightened in GMMs by the fact that nonconvergence issues are omnipresent (Jung & Wickrama, 2008). Diallo, Morin, & Lu (2016) note that models allowing the mean and covariance parameters to be completely class-specific commonly result in inadmissible estimates (e.g., nonpositive definite covariance matrices) or nonconvergence of optimization routines. A common remedy (and M*plus* default parameterization) is to apply equality constraints to covariance parameters across classes (Wickrama et al., 2016).

Though constraining covariance parameters is effective for reducing nonconvergence, observed groups often differ in their variance so it stands to reason that one should not, by default, assume latent groups are homoskedastic (Bauer & Curran, 2003). Previous studies have shown that the number of classes selected from models with constrained covariance parameters are rarely correct

(Diallo et al., 2016; Kreuter & Muthén, 2008) and additional aspects of the model are adversely affected when covariance parameters are constrained such as bias in the estimated growth trajectories in each class (Davies et al., 2017; Heggeseth & Jewell, 2013; Sijbrandij et al., 2020), the meaning of the classes (Bauer & Curran, 2004; Shiyko et al., 2012), and assignment of people to the proper class (Infurna & Luthar, 2016).

Recent research has tried to address the issue of how to improve GMM convergence while avoiding constraining all covariance parameters to be equal across all classes. Extensions of the proportional covariance method of Liu and Rubin (1998) to GMMs have been one suggestion whereby all covariance parameters in a referent class are freely estimated (Proust-Lima, Philipps, & Liquet, 2017). Then, the covariance matrix in all other classes is equal to the referent class multiplied by a class-specific coefficient, resulting in equal correlation matrices but not equal covariance matrices across classes (Barnard, McCulloch, & Meng, 2000; Manly & Rayner, 1987).

As another possible approach, McNeish and Harring (2020) note that GMMs typically follow the random effects tradition but that the marginal model tradition – covariance pattern models, specifically – may be able to help address some of the nonconvergence issues encountered with GMMs. The process of partitioning the covariance into within-person and between-person components in the random effect tradition can encumber estimation when latent classes are added (Hipp & Bauer, 2006). McNeish and Harring (2020) argue that estimation of the between-person variance provides little substantive utility in models with latent classes because the main interest is in quantifying differences between *classes* rather than quantifying differences between *individuals* within the same class. They demonstrate that modeling the marginal covariance of the repeated measures directly in each class is computationally simpler and improves convergence without requiring covariance parameter constraints across classes. Allowing the model to retain the flexibility to estimate all parameters as class-specific was shown to reduce the bias of the estimated class trajectories and improve assignment of people to the correct class, provided that the correct number of classes was fit (McNeish & Harring, 2021). This flexibility is more in line with the spirit of GMMS in which all parameters are permitted to be class-specific.

A looming question – however – is whether *proportional* GMMs or *covariance pattern* GMMs can select the correct number of classes. The goal of this paper is to extensively study class enumeration in a simulation whose characteristics are based on real data where nonconvergence is common. To outline the structure of the remainder of the paper, we first compare and contrast marginal and random effect traditions to growth modeling. Then, we discuss how models from each tradition – not just the random effects tradition – can be extended by adding latent classes. We use the post-traumatic stress disorder (PTSD) literature to inform conditions of a simulation study interested in class enumeration performance of different types of GMMs. We present an empirical example on PTSD symptoms to demonstrate differences in the results and interpretation of different methods with real data. We conclude with a discussion of practical implications, limitations, and future directions.

2. Random Effects vs. Marginal Traditions for Growth Modeling

Repeated measures data are characterized by a violation of the traditional independence assumption because residuals of repeated measures from the same person are more related to each other than they are to residuals from another person (Hedeker & Gibbons, 2006). Models for repeated measures data must therefore account for the dependence among repeated measures for inferences to be valid (Diggle, Heagerty, Liang, & Zeger, 2002). Multiple approaches can be taken to accomplish this, which has led to debate in the longitudinal data analysis literature about random effect versus marginal approaches (Zeger, Liang, & Albert, 1988) and many pedagogical papers have been written to guide researchers through the differences (Burton, Gurrin, & Sly, 1998; Hanley, Negassa, Edwardes, & Forrester, 2003; Hubbard et al., 2010; McNeish et al., 2017).

The defining characteristic of random effects models is that a unique growth trajectory is formed for each person. The presence of person-specific growth trajectories partitions the covariance between repeated measures from the same person into within-person and between-person sources (Curran, Obeidat, & Losardo, 2010). Between-person sources capture heterogeneity in the growth factors defining the growth trajectory and within-person sources capture the variability in the observed data around the person-specific trajectory. These two sources are estimated separately but they can be combined to form the marginal covariance of the repeated measures. In a random effects model, the different components of the covariance are theoretically interesting and are on equal ground to the regression coefficients that describe mean changes over time (Gardiner, Luo, & Roman 2009).

On the other hand, marginal models do not provide unique growth trajectories for each person in the data. Instead, they acknowledge the covariance among repeated measures by directly estimating elements of the covariance matrix, which is done separate from the growth factors (i.e., the regression coefficients are not modeled as random). The result is that covariance is not partitioned into between-person and within-person sources with marginal models. Rather, marginal models estimate the average growth trajectory while directly estimating the covariance between repeated measures. This approach does not provide person-specific growth trajectories; however, the absence of random effects makes estimation simpler while requiring fewer assumptions. Parameter estimates and their standard errors account for the covariance between repeated measures, but the covariance is not a focus and is treated as a nuisance to be accommodated to obtain valid inferences rather than substantive interest.

2.1 Latent Growth Models

The standard latent growth model falls under the random effect tradition and can be written

$$\begin{aligned} \mathbf{y}_i &= \boldsymbol{\Lambda}_i \boldsymbol{\eta}_i + \boldsymbol{\varepsilon}_i \\ \boldsymbol{\eta}_i &= \boldsymbol{\alpha} + \boldsymbol{\zeta}_i \end{aligned} \tag{1}$$

In the first expression, $\mathbf{y}_i$ is a $t_i \times 1$ vector of responses where t_i is the number of observed repeated measures provided by person *i*, $\boldsymbol{\Lambda}_i$ is a $t_i \times q$ matrix of loadings for *q* the number of latent growth factors, $\boldsymbol{\eta}_i$ is a $q \times 1$ vector of person-specific latent growth factor scores for individual *i*, and $\boldsymbol{\varepsilon}_i$ is a $t_i \times 1$ vector of residuals where $\boldsymbol{\varepsilon}_i \sim MVN\left(\mathbf{0}, \boldsymbol{\Theta}_i\right)$ and $\boldsymbol{\Theta}_i$ is a $t_i \times t_i$ residual covariance matrix. In the second expression, the person-specific latent growth factor scores $\boldsymbol{\eta}_i$ are equal to a $q \times 1$ vector of growth factor means $\boldsymbol{\alpha}$ plus a $q \times 1$ vector of between-person random effects, $\boldsymbol{\zeta}_i \sim MVN\left(\mathbf{0}, \boldsymbol{\Psi}\right)$.

The unconditional model-implied mean and marginal covariance structures are $\boldsymbol{\mu}_i = \boldsymbol{\Lambda}_i \boldsymbol{\alpha}$ and $\boldsymbol{\Sigma}_i = \boldsymbol{\Lambda}_i \boldsymbol{\Psi} \boldsymbol{\Lambda}_i^{\mathrm{T}} + \boldsymbol{\Theta}_i$, respectively. Note that the marginal covariance matrix ($\boldsymbol{\Sigma}_i$) is a combination of the between-person random effect covariance matrix ($\boldsymbol{\Psi}$) and the within-person residual covariance matrix ($\boldsymbol{\Theta}_i$).

2.2 Covariance Pattern Model

The covariance pattern model is one type of marginal model for continuous outcomes (Jennrich & Schluchter, 1986; Schluchter, 1988) and can be written as

$$\mathbf{y}_i = \boldsymbol{\Lambda}_i \boldsymbol{\alpha} + \boldsymbol{\varepsilon}_i \tag{2}$$

Note that there are no person-specific growth factors in Equation 2, only the marginal growth curve is captured by the parameters in $\boldsymbol{\alpha}$ such that the model-implied mean structure is also $\boldsymbol{\mu}_i = \boldsymbol{\Lambda}_i \boldsymbol{\alpha}$. There are no random effects in the model and the covariance matrix is not partitioned into different sources. Instead, the marginal covariance is directly patterned as a function of parameters in the $\boldsymbol{\theta}$ vector, which can include autoregressive parameters, correlations, or variances such that $\boldsymbol{\varepsilon}_i \sim MVN\left(\mathbf{0}, \boldsymbol{\Sigma}_i\left(\boldsymbol{\theta}\right)\right)$. For instance, to model a marginal compound symmetric covariance matrix,

$$\boldsymbol{\theta} = \begin{bmatrix} \sigma^2 \\ \rho \end{bmatrix} \text{such that } \boldsymbol{\Sigma}_i\left(\boldsymbol{\theta}\right) = \sigma^2 \begin{bmatrix} 1 & & & \\ \rho & \ddots & & \\ \vdots & \ddots & \ddots & \\ \rho & \cdots & \rho & 1 \end{bmatrix}.$$

3. Extending Growth Models with Latent Classes

GMMs are often thought as an extension of latent growth models (McArdle & Epstein, 1987; Meredith & Tisak, 1990) with a discrete latent moderator (B. O. Muthén, 2002). Essentially, GMMs are a multiple-group growth model where the grouping variable happens to be latent and unobserved (Bauer & Curran, 2003; Ram & Grimm, 2009). So whereas groups are split based on a known variable in a multiple group growth model, GMMs probabilistically uncover the groups into which the data are split (Grimm et al., 2016, p. 138).

If it is hypothesized that the data are comprised of a small number of distinct classes, the latent growth model from Equation 1 can be augmented with *k* subscripts to represent the *K* different latent classes,

$$\begin{aligned}\mathbf{y}_i &= \boldsymbol{\Lambda}_i \boldsymbol{\eta}_{ik} + \boldsymbol{\varepsilon}_i \\ \boldsymbol{\eta}_{ik} &= \boldsymbol{\alpha}_k + \boldsymbol{\zeta}_i\end{aligned} \tag{3}$$

where $\boldsymbol{\zeta}_i \sim MVN\left(\mathbf{0}, \boldsymbol{\Psi}_k\right)$ and $\boldsymbol{\varepsilon}_i \sim MVN\left(\mathbf{0}, \boldsymbol{\Theta}_{ik}\right)$. The parameters characterizing change and variability are now class-specific and indexed by *k* (where $k = 1, ..., K$ for *K* the total number of classes as selected by the researcher). This *unconstrained* GMM implies that every class will have its own growth factor means ($\boldsymbol{\alpha}_k$), between-person random effect covariance matrix ($\boldsymbol{\Psi}_k$), and within-person residual covariance matrix ($\boldsymbol{\Theta}_{ik}$).

The unconditional model-implied mean and covariance structures for class *k* are $\boldsymbol{\mu}_{ik} = \boldsymbol{\Lambda}_i \boldsymbol{\alpha}_k$ and $\boldsymbol{\Sigma}_{ik} = \boldsymbol{\Lambda}_i \boldsymbol{\Psi}_k \boldsymbol{\Lambda}_i^{\mathrm{T}} + \boldsymbol{\Theta}_{ik}$, respectively. Pooling over classes, the density of $\mathbf{y}_i$ can be written as the mixture distribution $f\left(\mathbf{y}_i \mid \boldsymbol{\varphi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}\right) = \sum_{k=1}^{K} \left[\varphi_k f_k\left(\mathbf{y}_i \mid \boldsymbol{\mu}_{ik}, \boldsymbol{\Sigma}_{ik}\right)\right]$ where f_k is the component normal probability density function for the *k*th class, and φ_k is the proportion of people in the *k*th class where $0 \le \varphi_k \le 1$ and $\varphi_K = 1 - \sum_{k=1}^{K-1} \varphi_k$.

3.1 Nonconvergence with Unconstrained GMMs

The GMM allows for between-person variability (captured by $\boldsymbol{\Psi}_k$) around the class-specific mean trajectory (defined by $\boldsymbol{\Lambda}_i\boldsymbol{\alpha}_k$) and within-person variability (denoted by $\boldsymbol{\Theta}_{ik}$) around the person-specific growth trajectory (defined by $\boldsymbol{\Lambda}_i\boldsymbol{\eta}_{ik}$). This helps to fully partition the variability in between-person and within-person sources, but it also means that there are many latent variable covariance parameters to estimate in the model.

In growth models, covariance parameters are the most difficult to estimate (Kiernan, 2018) and allowing each class to have its own class-specific covariance parameters quickly increases the quantity of difficult-to-estimate parameters in the model and can amplify nonconvergence issues (Liu & Hancock, 2014; Pastor & Gagné, 2013). Often, an unconstrained GMM will not converge even when it is the exact model from which data were generated (McNeish & Harring, 2020). In other words, the model can sometimes be too complex to fit, even when it is the true model (Kim, 2012).

A key problem is that the presence of so many covariance parameters across latent classes can create singularities in the likelihood surface whereby the likelihood spikes to infinity (Hipp & Bauer, 2006). If an estimation algorithm encounters one of these singularities, the algorithm will simply fail to converge because the gradient is undefined. Another more subtle issue with singularities occurs when the estimation algorithm encounters values near the singularity, which often result in local maxima that terminate the estimation algorithm at a solution that does not represent the global maximum of the likelihood surface (Biernacki, 2005; McLachlan & Peel, 2004). A common remedy is to reduce the number of covariance parameters, typically by adopting a homoskedastic specification.

3.2 Homoskedastic Growth Mixture Models

Homoskedastic growth mixture models (HGMMs) remove the *k* subscripts from $\boldsymbol{\Psi}$ and $\boldsymbol{\Theta}_i$ such that $\boldsymbol{\Psi}_1 = \boldsymbol{\Psi}_2 = \cdots = \boldsymbol{\Psi}_K = \boldsymbol{\Psi}$ and $\boldsymbol{\Theta}_{i1} = \boldsymbol{\Theta}_{i2} = \cdots = \boldsymbol{\Theta}_{iK} = \boldsymbol{\Theta}_i$. Doing so retains the concept behind the unconstrained GMM in Equation 3 while reducing the number of covariance parameters. The logic is that, if the number of covariance parameters make estimation difficult, then estimating fewer of these parameters should address the problem (Banfield & Raftery, 1993). The advantage with this approach is that the model retains person-specific trajectories within each class but is optimization is much easier (Diallo et al., 2016; Hipp & Bauer, 2006). HGMMs are thus commonly implemented in empirical settings to aid convergence (e.g., HGMMs are the default in the M*plus* software; Infurna & Grimm, 2018).

The HGMM specification has, however, also been widely criticized. Bauer and Curran (2003) explicitly question the choice to apply constraints across classes by stating,

> Although [constraints across classes] are statistically expedient, we do not regard these equality constraints as optimal from a theoretical standpoint, and in our experience, they are rarely found to be tenable in practice. Indeed, implementing these constraints is in some ways inconsistent with the spirit of the analysis, because one is forcing the majority of the parameter estimates to be the same over classes (permitting only mean differences in the within-class trajectories) (p. 346).

That is, homogeneity constraints are often applied in response to nonconvergence rather than theory (Infurna & Grimm, 2018; Infurna & Jayawickreme, 2019). The constrained model will then attempt to classify individuals while satisfying the assumption that classes vary in equal amounts. To the extent that the covariance structure differs across classes, enumeration and classification errors will increase (Diallo et al., 2016; Gilthorpe et al., 2014; Heggeseth & Jewell, 2013; Infurna & Luthar, 2016; Kooken et al., 2019).

One case in which this is especially problematic is when responses are uniformly near the ceiling or floor and show minimal change (e.g., studies on substance use that contain abstainers). Members in such a class will have minimal variability but constraining the variance to be equal to other classes will result in the “floor effect” or “ceiling effect” class necessarily expanding and consuming data that would be assigned elsewhere were the covariance parameters unconstrained across classes (Infurna & Jayawickreme, 2019). Nonetheless, when researchers face nonconvergence issues, there are relatively few other options to consider (though see van de Schoot et al., 2018 for a discussion of Bayesian approaches), so the HGMM specification remains common in empirical studies.

3.3 Proportional Growth Mixture Models

The number of covariance parameters can alternatively be reduced without imposing homogeneity of covariance structures across classes with a *proportional growth mixture model* (PGMM). PGMMs freely estimate all the covariance parameters in one referent class and then constrain the covariance parameters in all other classes to be proportional to the referent class (Barnard, McCulloch, & Meng, 2000). The covariance structure in all non-referent classes only requires one additional parameter that controls the proportional increase or decrease in class *k* is relative to the referent class (Banfield & Raftery, 1993). This results in identical *correlation* matrices across classes but unique dispersion in each class so that the *covariance* matrices are different across classes. This can be a parsimonious option for areas such as alcohol consumption or cigarette smoking where higher mean values are associated with higher variability (e.g., Maher, Ra, Leventhal, Hedeker, Huh, Chou, & Dunton, 2018; Pugach, Hedeker, Richmond, Sokolovsky, & Mermelstein, 2014; Weinstein & Mermelstein, 2013).

This approach is not popular with M*plus* users presumably because it requires manually imposing many model constraints (one per covariance parameter, per class) but is more common in latent class software in R and can be implemented in the `lcmm` package (Proust-Lima, Philipps, & Liquet, 2017). In statistical notation, the PGMM mean structure is the same as Equation 3 but the covariance submodel changes such that $\boldsymbol{\zeta}_i \sim MVN\left(\mathbf{0}, \omega_k \boldsymbol{\Psi}\right)$ and $\boldsymbol{\varepsilon}_i \sim MVN\left(\mathbf{0}, \omega_k \boldsymbol{\Theta}_i\right)$ where $\boldsymbol{\Psi}$ is the estimated growth factor covariance matrix in the referent class, $\boldsymbol{\Theta}_i$ is the estimated residual covariance matrix in the referent class, and ω_k is the covariance proportion parameter for class *k* (which is fixed to 1 for the referent class). The performance of the PGMM has not been studied in as much depth as the HGMM in the context of nonconvergence, so less is known about its statistical properties – class enumeration especially – but it remains appealing for using similar logic while being less restrictive.

3.4 Latent Class Growth Models

Another computationally simpler approach is *latent class growth analysis* (LCGA; Nagin, 1999; Nagin & Tremblay, 2001). Nagin and colleagues note that interpreting *N* separate trajectories as in GMMs can be unnecessary and that *N* trajectories can be reduced to a handful of prototypical trajectories. Instead of fitting a model with continuous, normally distributed random effects to capture between-person variance, between-person differences are captured by discrete latent classes. The result is that people are classified to the discrete class whose trajectory most closely matches what the person-specific trajectory would have been, rather than estimating a unique trajectory for each person. The model does not allow for between-person differences within classes and people within a class are considered interchangeable, so any deviation from the class-specific trajectory is absorbed into the residual term (Bauer & Curran, 2004).

The LCGA can be written as

$$\mathbf{y}_i = \boldsymbol{\Lambda}_i \boldsymbol{\alpha}_k + \boldsymbol{\varepsilon}_i \tag{4}$$

where $\boldsymbol{\varepsilon}_i$ is a $t_i \times 1$ vector of residuals such that $\boldsymbol{\varepsilon}_i \sim MVN\left(0, \sigma_k^2 \mathbf{I}_{t_i}\right)$, meaning that the residuals are homoskedastic and independent across time. There is no within-class variation across people and the only source of heterogeneity is through the latent classes, so the goal of LCGAs is therefore to provide a semiparametric representation of the growth trajectories (Nagin & Tremblay, 2005). With no random effects within classes, computational difficulties are rarely encountered (Kreuter & Muthén, 2007) .

A drawback is that the covariance structure among repeated measures is quite simple (Fitzmaurice et al., 2012; B. O. Muthén, 2004). This is intentional because the LCGA conceptualizes classes differently from GMMs. LCGAs define a class as a collection of people who follow a similar and distinct trajectory whereas GMMs define a class as a heterogeneous set of people that can be described by a single probability distribution (Nagin & Tremblay, 2005, p. 895). Consequently, additional classes are often extracted with a LCGA relative to a corresponding GMM (Bauer & Curran, 2004; Kreuter & Muthén, 2008). This is not to say that the LCGA is incorrect, but rather that the solution represents a different definition of what constitutes a "class". LCGAs have merit for their own theoretical considerations, but they do not align with the goal of GMMs and can be a tenuous substitute for GMMs when combating nonconvergence because LCGAs often lead to different solutions and interpretations than GMMs (Sijbrandij et al., 2019; Twisk & Hoekstra, 2012).

3.5 Covariance Pattern Growth Mixture Models

The random effect approach is dominant with GMMs but the reason is difficult to pinpoint and the need for random effects in these models has been questioned (Anderlucci & Viroli, 2015; Henderson & Rathouz, 2018). This dominance is peculiar because the interest in GMM applications is almost universally between-*class* differences rather than between-*person* differences captured by the within-class random effects (Cole & Bauer, 2016; Sterba & Bauer, 2010, 2014). This suggests that the between-person variability within classes is a feature to accommodate rather than a direct research interest (Lee & Nelder, 2004). In such cases, Heagerty and Zeger (2000) explicitly recommended against a random effect approach, stating "if the primary objective of analysis is to make inference regarding the mean response … then a marginalized model may be preferred" (p. 17).

Just as latent classes can be added to latent growth models to create a GMM, latent classes can also be added to covariance pattern models to form a *covariance pattern growth mixture model* (CPGMM) such that

$$\mathbf{y}_i = \mathbf{\Lambda}_i \boldsymbol{\alpha}_k + \boldsymbol{\varepsilon}_i \tag{5}$$

where $\boldsymbol{\varepsilon}_i \sim MVN\left(\mathbf{0}, \mathbf{\Sigma}_{ik}\left(\boldsymbol{\theta}_k\right)\right)$. There are no random effects, so $\mathbf{\Sigma}_{ik}$ combines within- and between-person sources of variability in class *k*, whose structure is patterned by class-specific parameters, $\boldsymbol{\theta}_k$.

Note that the LCGA in Equation 4 is a special case of the CPGMM when $\boldsymbol{\theta}_k = \left[\sigma_k^2\right]$ such that $\boldsymbol{\Sigma}_{ik}\left(\boldsymbol{\theta}_k\right) = \sigma_k^2 \mathbf{I}_{t_i}$. CPGMMs extend LCGAs by allowing for more complex covariance patterns among repeated measures, which adopts the computational advantages of LCGAs but applies the GMM definition of “class” by fully modeling the probability distribution of the repeated measures within each class (McNeish & Harring, 2020).

In essence, the goal of CPGMMs is to arrive at the same marginal covariance in each class as an unconstrained GMM but to do so without having to go through the arduous step of partitioning the covariance. The limited research on CPGMMs has shown that they improve convergence, reduce bias of the class-specific growth trajectories, eliminate the need to constrain parameters across classes, and perform better with smaller sample sizes (McNeish & Harring, 2020, 2021). Whether CPGMMs can select the proper number of classes has not been comprehensively studied, which is why this important step in the GMM process is the focus of the current paper.

As disadvantages, researchers are responsible for selecting the structure of the marginal covariance, which can be more difficult than selecting the structure of partitioned covariance structures. The approach is also only applicable to continuous outcomes and a different approach like latent class generalized estimating equations would be needed to accommodate discrete outcomes (Rosen, Jiang, &Tanner, 2000; Tang & Qu, 2016). Additionally, the ability to obtain person-specific trajectories and the ability to differentiate between-person and within-person sources of variance is lost. However, as reported in a literature review in Appendix A, this information is not typically reported in empirical studies and sacrificing this information to improve nonconvergence may be negligible in the context of answering relevant research questions in most cases.

The next section provides a simulation study with conditions inspired by a review of empirical PTSD studies using GMMs reported in Appendix A to examine (a) how class constraints employed to address nonconvergence impact class enumeration and (b) whether CPGMMs, PGMMs, or LCGAs are able to improve class enumeration in conditions where nonconvergence with traditional GMMs is prevalent. After the simulation, we model empirical data from a PTSD study to demonstrate how the findings from the simulation may affect interpretations and conclusions of empirical studies.

4. Simulation Design

4.1 Data Generation Models

We use two data generation models, each with different number of latent classes. The first data generation model has $K = 4$ latent classes and the class trajectories follow the "Cat's Cradle" pattern that often emerges in PTSD and substance use research (Bonanno, 2004; Sher, Jackson, & Steinley, 2011). The four classes are

(1) A "Chronic" class that starts at high values and maintains high values
(2) A "Resilient" class that starts low and maintains low values
(3) A "Recovery" class that starts high and decreases over time
(4) A "Delayed Onset" class that starts low and increases over time

The Resilient group comprised 63% of the population, the Recovery class 12%, the Chronic class 19%, and the Delayed Onset class 6% to mirror allocations found in empirical PTSD applications of GMMs reviewed in Appendix A. The second data generation model has $K = 3$ latent classes and represents the first three classes of the 4-Class model but without the Delayed Onset class. In this model, the Resilient group comprised 60% of the population, the Recovery class 20%, and the Chronic class 20%.

Substantively, the rationale for choosing this pair of models revolves around the debate in the psychiatry literature about the existence of the Delayed Onset class and whether it may be a statistical artifact rather than substantively meaningful phenotype of PTSD (Andrew, Brewin, Philpott, & Stewart, 2007; Frueh, Grubaugh, Yeager, & Magruder, 2009; Infurna & Jayawickreme, 2019; Infurna & Grimm, 2018; McNally, 2003; Spitzer, First, & Wakefield, 2007). Generating data from a 3-Class model without the Delayed Onset class can help address whether such a class is extracted when it is not present or whether criteria used for class enumeration can accurately distinguish between models with and without a small Delayed Onset class.

Both models will feature 5 repeated measures. Time is coded as: 0 (baseline), 1, 10, 18, and 26, which follows the timing employed in motivating the empirical data we use later in the paper. From the reviewed PTSD studies in Appendix A, the range for the number of repeated measures was 3 to 7, so 5 repeated measures were chosen because it was the midpoint of the range. The number of repeated measures was not manipulated in the simulation.

The data generation model is an unconstrained quadratic GMM with different covariance structures in each class. The generating models allow the intercept and linear slope to vary across individuals within classes. The quadratic slope variance was constrained to zero in the data generation model because quadratic variance is difficult to estimate due to scale differences (Diallo et al., 2014) and we did not wish to inadvertently favor models that do not feature random effects based on how we generated the data. Table 1 shows the model equations and covariance structures that were used to generate the data. Table 1 has two sets of covariance matrices that satisfy the manipulated simulation conditions because we also manipulate class separation (discussed shortly).

4.2 Manipulated Conditions

We manipulated three conditions in our simulation design. First, we generated data from three different sample sizes (100, 300, and 500). These conditions were selected from the distribution of the sample sizes observed in the review of the PTSD literature in Appendix A; the 10^{th} percentile was near 100, the 35^{th} percentile was near 300, and the median was near 500. Our focus is on situations where an unconstrained GMM will not converge – even if it is the true model – so our simulation conditions focus on the lower 50% of the sample size distribution where these issues are most likely to exist.

Second, we manipulate the separation between the latent classes. Sufficient sample size requirements in GMMs concern the interaction between number of people and class separation (e.g., Depaoli, 2013; Kim, 2012; Tein et al., 2013). If the classes are completely distinguishable, then the model will easily identify the number of classes and estimate the growth trajectory in each class, even at very small sample sizes. For example, Verbeke and Molenberghs (2009, p. 186) provide a 2-Class example of height changes in 20 schoolgirls without issues because the classes are highly separated. On the other hand, samples of 1000 or more may be insufficient for estimation and class assignment when classes are poorly separated (Tofighi & Enders, 2008; Tueller & Lubke, 2010).

Class separation is straightforward to define with 2-Class solutions (e.g., Mahalanobis distance), but it is not as straightforward to quantify with more than two classes as in the current simulation design. Relative entropy has been used in past studies by manipulating the population values of $\boldsymbol{\Psi}_k$ and $\boldsymbol{\Theta}_{ik}$ (Diallo et al., 2017; Tofighi & Enders, 2008). Following this precedent, our low separation condition featured relative entropy of 0.70 for the population model and our high separation condition resulted in relative entropy of 0.90. These values also mirror the 25th and 75th percentiles found in the PTSD review.

Each generated dataset will be fit with four different models, (1) a HGMM where covariance matrices are constrained to be equal across classes, (2) a PGMM with proportional growth factor and residual covariance matrices, (3) a LCGA with a class-specific residual variance, and (4) a CPGMM with a class-specific compound symmetric marginal covariance matrix. Note that a compound symmetric structure is somewhat misspecified because the data generation includes random slopes, resulting in a non-uniform pattern of correlations among repeated measures. A compound symmetric structure was chosen for the CPGMM based on the sample covariance between repeated measures rather than being based on the data generation model to represent real-life practice more closely; the off-diagonal elements of the sample covariance matrix were roughly equivalent because the scale of the intercept variance was larger than the linear slope variance. The mean structure was properly specified for each class in each model and featured both linear and quadratic effects for time. Because the population value for the quadratic slope variance was 0, HGMM and PGMM conditions did not include a quadratic slope random effect. The covariance between the intercept and linear slope random effect was estimated for HGMM and PGMM conditions despite being null in the population.

For data generated from the K = 4 model, we will fit HGMM, PGMM, LCGA, and CPGMM models with 2 through 5 latent classes. For data generated from the K = 3 model, we will fit HGMM, PGMM, LCGA, and CPGMM models with 2 through 4 latent classes. All four fitted models are misspecified to some degree because they sacrifice some information for the sake of reduced computational complexity. We fit the true unconstrained model only for conditions with the correct number of classes as a reference point for convergence rates.

We ran 500 replications in each of the 168 conditions in the simulation design (2 data generation models × 3 sample sizes × 2 class separation × 4 fitted models × (3 or 4) different latent class solutions). Data were generated and analyzed in M*plus* Version 8.3 using robust maximum likelihood estimation and output files were collated with a SAS macro. Given the focus on nonconvergence, we followed suggestions from the methodological literature (e.g., Li, Harring, & Macready, 2014; Liu & Hancock, 2014; Shireman et al., 2016, 2017) and altered the M*plus* defaults related to optimization and convergence criteria. This includes,

1. Increasing the number initial stage starts to 100 and 10 final stage optimizations
2. Increasing the number of initial stage iterations from 10 to 100
3. Increasing the numbers of Quasi-Newton iterations from 20 to 250
4. Using 500 EM algorithm iterations
5. Using an EM algorithm convergence criteria of 1E-5

All data generation and analysis files can be found on the first author's Open Science Framework page (https://osf.io/eay5v/).

4.3 Simulation Outcomes

Bias of class-specific growth trajectories has been reported in previous studies when the true number of classes are fit (McNeish & Harring, 2020, 2021), so we focus on convergence and class enumeration for different class solutions. We consider a replication nonconvergent if (a) different final stage optimizations were unable to replicate the same loglikelihood [e.g., suggesting the solution is a local maximum], (b) if the parameter estimates were inadmissible [e.g., negative variances], (c) one or more classes were empty, (d) the maximum number of iterations was exceeded, or (e) the Hessian matrix evaluated at the final parameters estimates was nonpositive definite.

For class enumeration, we track the proportion of replications in which different models select the correct number of classes across conditions. To assess this, we track 9 different information criteria to select between different non-nested models. These include:

1. Bayesian Information Criterion (BIC; Schwarz, 1978)
2. Sample-size Adjusted BIC (SABIC; Sclove, 1987)
3. Draper BIC (DBIC; Draper, 1995).
4. Integrated Completed Likelihood BIC (ICL-BIC; Biernacki, Celeux, & Govaert, 2000)
5. Sample-size Adjusted ICL-BIC (ICL-SABIC; Peugh & Fan, 2012)
6. ICL-DBIC
7. Classification Likelihood Criterion (CLC; Biernacki & Govaert, 1997)
8. Hurvich-Tsai Akaike Information Criteria (HT-AIC; Hurvich & Tsai, 1989)
9. Hannan-Quinn Akaike Information Criteria (HQ-AIC; Hannan & Quinn, 1979)

Table 2 shows the equations for each of these 9 criteria. Importantly, "entropy" in Table 2 is different from the *relative* entropy value reported in the M*plus* output that is normed to be between 0 and 1. The relationship between entropy (E) and relative entropy (RE) is $E = -N\left(RE - 1\right)\ln\left(K\right)$.

BIC and SABIC have been reported to perform well for enumerating latent class in general (Nylund et al., 2007) and are output by M*plus* by default. BIC is a staple information criterion that applies a penalty term based on the ratio of sample size to the number of estimated parameters to address effects of overfitting. The penalty can become too severe when *N* is not much larger than the number of estimated parameters, so Sclove (1987) suggested the SABIC with an adjustment to *N* for smaller sample sizes. SABIC has been noted to perform particularly well with challenging data characteristics like poor class separation and smaller samples (Kim, 2014). DBIC is a related criterion that includes an additional penalty term to penalize uncertainty in structural assumptions of the model (e.g., normal distributional assumptions; Draper, 1995).

ICL-BIC and ICL-SABIC are related to BIC and SABIC but include a second penalty term for entropy, which is a measure of separation amongst class probabilities. In its raw metric (rather than relative version reported by M*plus*), entropy near zero indicates better separation, meaning the additional $2 \times \text{Entropy}$ penalty in the ICL versions of BIC rewards models with better class separation. Previous studies have suggested that ICL versions of BIC maintain reasonable class enumeration performance for smaller samples ($N = 500$) for SEM mixture models (Henson, Reise, & Kim, 2007). The ICL-DBIC is a new metric devised for this study whereby the DBIC is substituted for BIC in the ICL-BIC formula such that an entropy-based penalty is applied to DBIC.

Outside of BIC-based criteria, the CLC has an entropy penalty but does not penalize for the number of parameters. Henson et al. (2007) found good performance for class enumeration of CLC across sample sizes. HT-AIC and HQ-AIC are based on AIC which penalizes only for the number of parameters (but not relative to sample size) but includes adjustments that are appropriate for autocorrelated data and have been shown to perform better with smaller sample sizes. These criteria have been studied in latent profile analysis (Peugh & Fan, 2013) and in GMMs with larger samples when the model is correct (Peugh & Fan, 2012) but not in the common practical context of when the model is necessarily misspecified because the true model is too complex to converge. The Open Science Framework page includes a spreadsheet for calculating these different indices from values found in M*plus* output.

Information criteria will be compared across all class solutions in each condition and the class solution with the lowest information criteria for each replication will be considered the number of classes selected for that replication. To coincide with how enumeration is conducted in empirical settings, replications will be retained if at least one model solution converged. For instance, if the 2-, 3-, and 4-Class solutions converge but the 5-Class solution does not, then the selected number of classes is based on the lowest information criteria among only the 2-, 3-, and 4-Class solutions.

As a rough metric, we consider selecting the correct number of classes in at least 50% of replications to be reasonable performance. We report this metric across all replications (converged and non-converged) and also only for converged replications. We emphasize the results across all replications when discussing the results. In the text, we provide results for the percentage of replications selecting the correct number of classes. Tables B1 through B8 in Appendix B show the percentage of replications selecting each number of classes to provide more detail about conditions or metrics that are more likely to over-extract or under-extract latent classes.

Though the main interest is in class enumeration, we also track the adjusted Rand Index (ARI) to assess classification accuracy (Hubert & Arabie, 1985). The ARI is used to compare agreement between two different data partitions, adjusting for chance level of agreement (Milligan & Cooper, 1986; Steinley, 2004). In the case of a simulation study, the two partitions being compared are the population class assignments and the estimated class assignments (based on the highest posterior probability) such that the ARI assesses how well the estimated classes recover the population classes (Steinley & Brusco, 2018). The ARI is also invariant to permutations of cluster labels, which is useful for simulation studies where class labels may switch across different replications (e.g., Hullermeier & Rifqi, 2009). An ARI value of 0 indicates chance level agreement and a value of 1 indicates perfect agreement. The ARI has been found to perform best among competing methods for comparing data partitions (Steinley et al., 2016) and was calculated in the mclust R package (Scrucca et al., 2016). We calculate the ARI for converged replications fitting the true number of classes to keep the presentation of results manageable, although it is possible to compute the index when the number of classes in the population and fitted model are incongruent.

4.4 Outcomes Not Included

There are metrics for enumerating classes that we did not include. Though we included adjusted versions of AIC, AIC itself was not studied because it has been found to perform very poorly in previous simulation studies (Diallo et al., 2016; Li & Hser, 2011; Nylund et al., 2007; Yang, 2006). We also focus only on information criteria and did not include inferential statistics for class enumeration, most notably the bootstrapped likelihood ratio test (BLRT; McLachlan, 1987) and the Lo-Mendell-Rubin test (Lo, Mendell, & Rubin, 2001). Both of these tests provide a *p*-value for the likelihood ratio test comparing a model with k classes to a model with $k-1$ classes; models with different numbers of classes are non-nested, so the proper *p*-value is not straightforward to obtain.

The rationale for focusing on information criteria is that inferential statistics for class enumeration may not be powerful enough for the sample size and class separation conditions included in our simulation (Dziak et al., 2014; Tekle et al., 2016). Diallo et al. (2016, 2017) also found that information criteria outperformed inferential tests for class enumeration when the covariance structure was misspecified via constraints or under-specification, which is a main focus of our simulation (HGMMs and PGMMs are misspecified via constraints; CPGMMs and LCGAs are misspecified via under-specification).

5. Simulation Results

5.1 Convergence Results

Table 3 shows the percent of replications that converged for each fitted model across simulation conditions. As a reference and to demonstrate how nonconvergence is rampant for models under these conditions, we also include the convergence rate for the unconstrained GMM when the number of classes is correct, which corresponds to the true data generation model. When the data generation included a small fourth class, the unconstrained GMM was too complex for the data, only converging in 2%, 17%, and 36% of replications for $N = 100$, $N = 300$, and $N = 500$, respectively, when classes were well separated. Again, this nonconvergence occurs when the *true* model is being fit. Convergence was even worse when the unconstrained GMM was fit to the 4-Class data with low class separation and did not break 10%. The true unconstrained GMM fared better for the 3-Class data generation model, but convergence rates remained under 70% under $N = 500$ and high class separation. Put simply, with modest sample sizes, even fitting the exact model used to generate the data far from assures convergence.

HGMMs are a common strategy to address nonconvergence of unconstrained GMMs and the reason is understandable given the relative increase in convergence rates. Nonetheless, HGMMs are not a foolproof strategy for improving nonconvergence in these conditions as convergence remained far from 100% in most conditions. With the 3-Class high separation data generations conditions, HGMM convergence was slightly worse than the unconstrained GMM. We double-checked to verify the source of this oddity and found that HGMM failed to converge in these conditions almost exclusively because of local maximum and an inability to replicate the best loglikelihood from different final stage optimizations (as opposed to the unconstrained GMM where nonconvergence was more often attributable to inadmissible estimates). Convergence with PGMMs was only marginally better or sometimes worse than unconstrained GMMs when class separation was high. However, PGMMs converged more often with low rather than high class separation across conditions and provided a marked improvement over the unconstrained GMM.

Convergence rates for CPGMMs and LCGAs were notably better than HGMMs or PGMMs and offered vast improvements over unconstrained GMMs. Both models simplify the estimation by eliminating computationally difficult random effects, which made it possible to consistently obtain admissible estimates even when $N = 100$. Of course, being able to produce admissible estimates is necessary but not sufficient, and a more important aspect is whether these different models can correctly identify the number of classes present in the data. Class enumeration results are discussed in detail in the subsections that follow.

5.2 Class Enumeration: High Separation, 4-Class Data Generation

Table 4 shows the percentage of all replications (top) and the percentage of converged replications (bottom) in which different information criteria selected the correct 4-Class solution across sample size conditions when the classes were well-separated. The table also shows the average relative entropy estimate across converged replications for the given condition to assess how well the model estimated class separation. For the 4-Class data generation model, when factoring in convergence, only CPGMMs were able to select the correct number of classes in at least 50% of all replications and had the best performance among the four methods.

Highlights of the CPGMM results include,

1. HT-AIC, HQ-AIC, and DBIC identified the correct number of classes in a majority of replications with $N = 100$.
2. BIC, ICL-SABIC, and CLC, all selected the correct number of classes in at least 75% of all replications with $N = 300$ and 90% of all replications with $N = 500$.
3. DBIC was the only criteria to select the correct number of classes in at least 50% of all replications for all sample sizes.
4. The relative entropy for the CPGMM was quite accurate with estimates between 0.91 and 0.92 across conditions compared to the 0.90 population value.
5. CPGMMs had the best ARI in each condition, indicating that CPGMMs was the best at assigning observations to the correct classes.

When conditioning only on replications that converged, PGMMs were able to regularly detect the true number of classes. When coupled with PGMMs and conditional on convergence being achieved, SABIC often selected the correct four class solution across sample size conditions and DBIC, HT-AIC, and HQ-AIC all selected the correct number of classes in about least 85% of converged replications when $N \geq 300$. The major drawback of PGMMs was the low convergence rates – the number of converged replications never exceed 21%. Although the correct number of classes could be reliably selected when the model converged in these conditions, convergence was achieved so infrequently that the overall ability to select the proper number of classes ended up being quite poor based on the top half of Table 4 when all replications were considered.

HGMMs performed poorly in these conditions when compared to CPGMMs and PGMMs. None of the criteria were able to identify the correct number of classes in at least 50% of all replications at any sample size and criteria with entropy penalties fared particularly poorly when coupled with HGMMs. SABIC, DBIC, HT-AIC, and HQ-AIC were able to select the correct number of classes in more than 50% of converged replications. However, these percentages were below both CPGMMs and PGMMs. The relative entropy for HGMMs was inaccurate and underestimated the true separation of the latent classes while the ARI was much lower than CPGMMs.

The near-perfect convergence rate with LCGAs in Table 3 did not translate into selecting the correct number of classes or accurately assigning observations to the correct class. The correct 4-Class solution was typically selected only when the 5-Class solution did not converge, supporting the conclusion of Kreuter and Muthén (2008) that the different definition of classes within LCGAs typically results in additional classes compared to GMMs.

5.3 Class Enumeration: High Separation, 3-Class Data Generation

Table 5 shows the class enumeration results for the 3-Class data generation model based on all replications (top) and only converged replications (bottom). The pattern of the results is similar to the 4-Class data generation model results shown in Table 4, so we will only briefly discuss the major points to avoid redundancy.

CPGMMs resulted in the best performance – they converged most often, selected the correct number of classes in a high percentage of replications, and had the highest ARI. PGMMs continued to perform well provided that convergence was achieved. However, although convergence was considerably higher in Table 5 than in Table 4, convergence rates continued to be much lower than CPGMMs (82-100% for CPGMMs; 51-60% in PGMMs). A sizeable discrepancy between the performance of HGMMs and the performance of CPGMMs or PGMMs remained and the discrepancy was wider than in Table 4. LCGAs only identified the correct number of classes when models with more classes failed to converge.

5.4 Class Enumeration: Low Separation, 4-Class Data Generation

Table 6 shows the percentage of all replications (top) and the percentage of converged replications (bottom) in which different information criteria selected the correct 4-Class solution across sample size conditions when the classes were poorly separated. As expected, class enumeration is much more difficult with smaller sample sizes and poorly separated classes, especially in the 4-Class condition when the data generation model contains a small class containing only 6% of the population.

Among the competing methods, CPGMMs performed the best relative to other methods and had reasonably high convergence along with at least one criterion that could consistently select the correct number of classes, though the ARI was worse than with HGMMs or PGMMs. Though CPGMMs were the only method to select the correct number of classes in more than 50% of replications in any condition, no criteria were able to identify the correct number of classes in all sample sizes. This blotchy set of results does not reveal consistent patterns upon which to build recommendations. Unlike with high class separation, criteria with entropy penalties did not perform well with CPGMMs when class separation was low. CPGMMs did manage to converge more often than HGMMs or PGMMs in these conditions.

With HGMMs, no criteria were able to consistently identify the correct number of classes with poor separation, regardless of whether considering all or only converged replications. The same was true of PGMMs which performed poorly when considering only converged replications and especially poorly when considering all replications. PGMMs had a higher ARI relative to CPGMMs, though the improvement is mostly nullified by lower convergence and inability to select the correct number of classes with regularity. CPGMMs and PGMMs were able to estimate relative entropy within ±0.10 of the 0.70 population value. HGMMs overestimated the relative entropy in these conditions, suggesting that the class separation was much better than it truly was. The LCGA performed rather poorly, which is expected because the data were generated with "class" being defined as a probability distribution, which is counter to the definition of "class" used in the LCGA.

Although the class enumeration performance leaves a lot to be desired with low class separation, CPGMMs and PGMMs were at least able to accurately estimate relative entropy and report that class separation is poor, which is useful for helping researchers gauge the difficulty of extracting latent classes from their data. HGMMs and LCGAs perform poorly both for class enumeration and estimation of relative entropy (assuming the probability distribution definition of class), which distorts important information needed to evaluate the possibility or trustworthiness of fitting GMMs with modest sample sizes. Although CPGMMs performed best in relative terms, in absolute terms CPGMMs were not great, and the results reinforce the difficulty of fitting GMMs when samples are modest and separation is poor.

5.5 Class Enumeration: Low Separation, 3-Class Data Generation

Table 7 shows the percentage of all replications (top) and the percentage of only converged replications (bottom) in which different information criteria selected the correct 3-Class solution across sample size conditions when the classes were poorly separated. Similar to the 4-Class low separation results in Table 6, all methods had difficulty identifying the correct number of classes. As in other conditions, CPGMMs had high convergence rates and were the only method to identify the correct number of classes in at least 50% of replications in the top half of the table. In the 3-Class low separation condition, there did appear to be some consistency that DBIC with CPGMMs had a reasonable ability to consistently selected the correct number of classes, although the ARI remained rather low. The 3-Class data generation model results clarify that the difficulty seen in the 4-Class data generation condition is not completely attributable to the presence of a small class.

PGMMs did not select the correct number of classes when considering all replications and only selected the correction number of classes in more than 50% of converged replications in one condition. CPGMMs and PGMMs both estimated relative entropy within ±0.10 of the population 0.70 value and accurately reflected the degree of separation among the classes in the population. The ARI with PGMMs was the best among methods, but poor convergence and poor enumeration with PGMMs in these conditions make this superior classification accuracy moot.

With HGMMs, no information criteria were able to select the correct number of classes in at least 50% of all replications. The correct number of classes was selected in at least 50% of converged with HGMMs in a few conditions, but there were no consistent patterns. Relative entropy continued to be poorly estimated with HGMMs in these conditions and was never within ±0.10 of the 0.70 population value, again making HGMMs a liability for assessing class separation and determining appropriateness of the analysis with modest samples. In these conditions, the LCGA almost never detected the correct number of classes and did not accurately estimate relative entropy.

5.6 Simulation Result Summary

An unconstrained GMM can simply be too complex to fit to some data, even when it is the true data generation model. HGMMs help convergence somewhat (though not as much as other methods), but rarely select the proper number of classes and assign observations to the correct class less often. Criteria with no entropy penalty coupled with PGMMs were able to select the correct number of classes when classes were highly separated, but PGMMs generally had the worst convergence rates of the four competing methods in the simulation. Given that a motivating reason for using each of the models is to address nonconvergence in unconstrained GMMs, the lower convergence rates of PGMMs is a vulnerability that hampers its utility.

CPGMMs and LCGAs greatly improve convergence, but the definition of “class” with LCGAs differs from GMMs and – expectedly – LCGAs rarely recover the correct number of classes based on GMM’s definition of “class” and also performed worst at assigning observations to the correct class. The definition of “class” in CPGMMs is consistent with GMMs and CPGMMs can model class-specific covariance structures more easily, which is essential for proper class enumeration. CPGMMs were the only method that could simultaneously improve convergence and select the correct number of classes. When class separation was high, CPGMMs also displayed the highest accuracy in assigning observations to the correct class.

Specific notable findings are summarized below.

1. The ability to select the proper number of classes at realistic sample sizes is heavily dependent on class separation. Successful class enumeration is possible at modest samples, but only when the classes are relatively well separated (relative entropy near .90 or higher). Detecting the presence of small classes with modest samples did not appear to be problematic with CPGMMs.
2. Permitting each class to have a unique covariance structure as with CPGMMs – and PGMMs, to a lesser extent – is highly beneficial, even if the covariance structure in each class is not exactly right. In other words, properly specifying the covariance structure is less important than allowing the covariance structure to be class-specific.
3. With high class separation, CPGMMs could often select the proper number of classes and accurately assign observations to those classes. DBIC was the most consistent criteria across conditions, but DBIC did not necessarily have the single highest percentage of replications with the correct number of classes. HT-AIC and HQ-AIC tended to be better in the smallest sample condition whereas BIC, CLC, ICL-SABIC, and ICL-DBIC were better with the larger samples included in the simulation.

4. CPGMMs and PGMMs accurately estimate relative entropy values and allow for accurate assessment of class separation. HGMMs do not accurately estimate relative entropy and make it difficult to assess whether class separation is sufficient to trust the class solution.
5. LCGAs use a different definition of what constitutes a “class” and do not recover classes consistent with the GMM definition. LCGA should not substituted for GMMs to avoid nonconvergence issues or to simplify estimation. CPGMMs follow the same approach of LCGA by removing random effects with similar improvement in convergence, but the definition of “class” with CPGMMs is consistent with GMMs.
6. Criteria with entropy penalties coupled with CPGMMs were consistently among the best at selecting the correct number of classes when class separation was high but were universally poor with low class separation and consistently selected too few classes. Perhaps development of a different or varying penalty for entropy would help generalize the strong performance of entropy-based criteria (similar to the logic of BIC versus AIC).
7. HGMMs performed quite poorly across all conditions included in the study. Even though HGMMs are a commonly applied approach when convergence is an issue or when defaults are applied indiscreetly, our results add to the existing literature showing that this method is quite poor and ill-advised. This method should be avoided given that CPGMMs demonstrate superior performance for convergence, class enumeration, and classification accuracy (in addition to other metrics as reported in previous research) while retaining the flexibility to allow parameters – especially covariance structures – to be unconstrained across classes.

6. Empirical Example

Empirical data come from a study on PTSD with 301 burn victims who were admitted to a burn center between 1997 and 2000 (van Loey et al., 2003). Participants completed the Impact of Event Scale (Horowitz et al., 1979) following a traumatic incident where higher scores indicate more severe symptoms. The first two waves were taken two and three weeks after the traumatic incident, respectively. The next two waves were collected in 2-month intervals thereafter to assess PTSD symptoms 4 months following the incident (i.e., time is coded 0, 1, 9, 17). The data include additional future waves, but we do not include them to more closely match the characteristics of the simulation and the studies reviewed to inform the simulation conditions. The observed covariance matrix of the raw data is $\mathbf{S} = \begin{bmatrix} 289.55 & & & \\ 207.30 & 274.13 & & \\ 184.39 & 197.26 & 307.02 & \\ 173.50 & 193.41 & 239.49 & 312.33 \end{bmatrix}$, which shows that the variances appear to grow over time and the covariances appear to diminish at each subsequent lag.

We first tried to fit an unconstrained quadratic GMM such that the growth factor variances and time-point specific residual variances were freely estimated across each class. Written statistically,

$$\begin{aligned} \mathbf{y}_i &= \mathbf{\Lambda}_i \boldsymbol{\eta}_{ik} + \boldsymbol{\varepsilon}_i \\ \boldsymbol{\eta}_{ik} &= \boldsymbol{\alpha}_k + \boldsymbol{\zeta}_i \\ \boldsymbol{\zeta}_i &\sim MVN\left(\mathbf{0}, \mathbf{\Psi}_k\right) \\ \boldsymbol{\varepsilon}_i &\sim MVN\left(\mathbf{0}, \sigma_{tk}^2 \mathbf{I}_4\right) \end{aligned} \tag{6}$$

We fit models with between 2 and 5 latent classes in M*plus* Version 8.3 with robust maximum likelihood estimation using the same recommended estimation options as used in the simulation (Li et al., 2014; Liu & Hancock, 2014; Shireman et al., 2016, 2017). As anticipated from our simulation results, this model was unable to converge with between 2 and 5 classes. This occurred regardless of whether the quadratic slope variance was estimated or constrained to 0. If the residual variances were modeled as homoskedastic such that $\boldsymbol{\varepsilon}_i \sim MVN\left(\mathbf{0}, \sigma_k^2 \mathbf{I}_4\right)$, the 2-Class model converged but models with 3 through 5 latent classes did not.

Given that the unconstrained GMM did not converge, we then fit a PGMM with between 2 and 5 classes. Consistent with the simulation results in Table 3, the PGMM did not converge for any class solutions with heteroskedastic or homoskedastic residuals. For each class solution, nonconvergence was attributable to a nonpositive definiteness. A HGMM and CPGMM both were able to converge, the results of which are covered in a subsection dedicated to each model.

6.1 Constrained Growth Mixture Model

Consistent with current practice, we fit a quadratic HGMM with the covariance structures constrained across classes (the same model as in Equation 6 except $\mathbf{\Psi}$ and $\mathbf{\Theta}$ have no k subscripts). The HGMM models did not initially converge if the residuals were heteroskedastic but did converge with homoskedastic residuals. Table 8 shows the nine information criteria explored earlier in the simulation study and the relative entropy for models with 2 through 5 classes. The class separation appears to be rather good with relative entropy values mostly in the high 0.80s and nearly all criteria suggest 5 classes, except the ICL-BIC which narrowly suggests 3 classes over 5 classes. The ICL-BIC selected the proper number of classes in 1% or fewer of replications in the simulation with HGMM for all high separation conditions, so it appears safe to conclude that the HGMM is suggesting 5 classes for these data.

The class trajectories estimated by the HGMM are compared to the empirical data for people assigned to each class in Figure 1. The extracted classes largely resembled the "Cat's Cradle" pattern commonly observed in empirical HGMM analyses of PTSD data. Class 1 represents the Recovery class such that the curve is decreasing at the end of the observation window. Class 2 is the Resilient class where people start with subclinical scores (the clinical cutoff for this measure is 33) and are essentially flat over time. Class 3 represents the Chronic class where people begin at or near the clinical cutoff and increase over time. Class 4 and 5 appear to be components of the Delayed Onset class whereby curves are increasing at the end of the observation window. Note that the variance around the trajectory in each class is about the same across each class – this is imposed by the HGMM to reduce the number of covariance parameters to facilitate estimation. However, this assumption can be relaxed with a CPGMM, which is fit next and results in a rather different solution.

6.2 Covariance Pattern Growth Mixture Model

Next, we consider a quadratic CPGMM. Because the variances appear to grow and the off-diagonal elements diminish as a function of the lag, we use an unrestricted marginal covariance structure that uniquely estimates all 10 covariance parameters per class to model the marginal covariance. The model can be written

$$\begin{gathered}\mathbf{y}_i = \boldsymbol{\Lambda}_i \boldsymbol{\alpha}_k + \boldsymbol{\varepsilon}_i \\ \boldsymbol{\varepsilon}_i \sim MVN\left(\begin{bmatrix}0\\0\\0\\0\end{bmatrix}, \begin{bmatrix}\theta_{11k} & & & \\ \theta_{21k} & \theta_{22k} & & \\ \theta_{31k} & \theta_{32k} & \theta_{33k} & \\ \theta_{41k} & \theta_{42k} & \theta_{43k} & \theta_{44k}\end{bmatrix}\right)\end{gathered} \tag{7}$$

Equation 7 with 2 through 5 classes converges without any additional constraints or alterations and the best loglikelihood was replicated with different final stage optimizations. Fit criteria and relative entropy for each class are shown in the bottom half of Table 8. The relative entropy estimates were a little different than the HGMM, but similarly suggest that class separation is rather good and is in the mid-0.80s to low-0.90s. This is consistent with the simulation where the HGMM underestimated relative entropy when separation was high. ICL-BIC, ICL-SABIC, and ICL-DBIC suggests 3 classes while BIC, CLC, HT-AIC, HQ-AIC, and DBIC suggest 4 classes. SABIC was the only criteria to suggest 5 classes.

In the simulation, BIC tended to select the number of classes well when N =300; however, in this example, the BIC is extremely close between 3-Class and 4-Class solutions (8,678 vs 8,674), which suggests limited evidence in favor or either model based on guidelines in Rafferty (1995). From the simulation, if K = 4 with high separation (results from Table 4), it was unlikely that HT-AIC would detect the true number of classes when N =300 (4 classes selected in 41% of replications) while ICL-SABIC and ICL-DBIC would select too few classes (19% and 26%, respectively, as reported in the Appendix Table B1). Conversely, if K = 3 with high separation (results from Table 5), it was rather likely that that ICL-SABIC and ICL-DBIC selected the correct number of classes

(89% and 88%, respectively) while HT-AIC selected too many classes (70% as reported in the Appendix Table B5). Based on this pattern on information criteria, it seems that the CPGMM is suggesting that 3 classes are appropriate for this data. Also note that the CPGMM criteria are uniformly lower than the corresponding HGMM criteria, suggesting more parsimonious fit despite the CPGMMs having many more parameters when modeling unrestricted marginal covariance structures unique to each class.

The estimated class trajectories from the CPGMM are compared to the empirical data for people assigned to each class in Figure 2. Note that the growth trajectories are much flatter than the corresponding HGMM but that the real difference in the classes lies in the *variability*. Class 1 is an Erratic class composed of people who appear to have difficulty managing their symptoms. Many of the data points are above the clinical cutoff, and individuals bounce above and below the cutoff across repeated measures. This class has very high residual variance and the marginal class trajectory does a poor job representing the data in this class because the symptoms for people in this class are unpredictable. Class 2 corresponds a Subclinical class. These people appear modestly affected by the traumatic incident they endured and show a slight decrease in symptoms over time. The residual variance in this class is small but nonzero as the empirical data bounce around the marginal class trajectory and people show minor struggles with symptoms over time. Despite small volatility, these people appear to largely have their symptoms under control and do not approach the clinical cutoff. Class 3 is the Resilient class composed of people who show essentially no symptoms. This class has essentially no residual variance because everyone in this class starts at 0 and their symptoms do not grow over time, so the marginal class trajectory almost perfectly fits the data in this class.[1] Essentially, this class solution is defined by *management* of symptoms, not *change* in symptoms.

6.3 Practical Implications

[1] Bauer and Curran (2003) allude to the possibility that this “class” may be attributable to skew in the response distribution and simply captures the floor effect of the scale rather than a substantively meaningful class.

HGMMs are quite common in empirical studies – in the PTSD literature review in Appendix A, only 11% of studies (2/19) reporting application of a GMM on observed repeated measures were able to fit a model with the covariance structures unconstrained across classes. Of the studies that constrained covariance structures and applied the HGMM, 65% (11/17) reported 4 or more classes whereas neither of the 2 studies without constraints found 4 or more classes. The empirical example seemed to follow this same pattern – models that allow the covariance structure to be unique to each class (like CPGMMs) often come to different conclusions regarding the number of classes when compared to models with constrained covariance structures (like HGMMs).

Note that the CPGMM class solution in the empirical example is strongly influenced by the variability, not by the mean growth trajectories. Approaches that constrain the variance to be equal across classes like HGMM – by design – are unable to detect classes following this type of pattern. Similarly, the Resilient classes that emerges with the CPGMM (Figure 2, Class 3) looks very different and is much smaller when compared to the HGMM Resilient class (Figure 1, Class 2) because a class of horizontal trajectories at 0 cannot exist by definition when variances are constrained across classes. Therefore, the Resilient and Subclinical classes from the CPGMM (Figure 2, Classes 2 and 3) are merged in the HGMM (Figure 1, Class 2) to satisfy a potentially unwarranted between-class homoskedasticity assumption imposed to avoid nonconvergence, completely losing the distinction between Resilient and affected but Subclinical individuals. This finding is not unique to this data and has previously been noted in Infurna and Jayawickreme (2019). Furthermore, the simulation results suggest that if the “Cat’s cradle” pattern were the true model, the CPGMM would have been able to detect it with high probability under these data conditions. As noted earlier, the existence of the delayed onset PTSD subgroup found in empirical GMM analyses has been questioned in the psychiatry literature and similar observations regarding classes extracted from empirical GMM not being observed in clinical settings has also been observed in the substance use literature (Jacob et al., 2005).

Of course, the results here represent a single empirical example, and it is not possible to know which (or neither) solution is correct, so these results should not be reified either. Nonetheless, an increasing volume of simulation and methodological research continues to demonstrate the importance of modeling variability generally (e.g., Hedeker et al., 2012; Hoffman, 2007; Nesselroade & Ram, 2004; Williams et al., 2020) and the importance of giving equal consideration to the covariance and mean structure in GMMs (Diallo et al., 2016; Diallo et al., 2017; McNeish & Harring, 2021). Researchers should strive to keep both structures class-specific to uphold the spirit of mixture modeling and to optimize the performance of their models.

7. Discussion

7.1 Limitations and Future Directions

Although we believe there are many strengths to the current study, it is also important to recognize its limitations. First, a weakness of simulations focused on mixture modeling in general is that the entire enterprise operates under the presumption that there is a "true" number of substantively meaningful classes to find, which assumes a *direct application* of mixture modeling. In reality, it is equally possible that the classes do not correspond to substantively meaningful classes and the classes are merely a mathematical device to approximate a complex reality (Dolan & van der Mass, 1998; Titterington et al., 1985), which would correspond to an *indirect application* of mixture modeling. In an indirect application, there is no "correct" number of latent classes to find, but only varying degrees of approximation to the observed data where one seeks to neither underfit the data (i.e., take too few classes to capture the primary features) nor overfit the data (i.e., include classes that capitalize on chance irregularities).

Reliably differentiating between these two types of applications empirically is not currently possible (Bauer & Curran, 2004). Further, Bauer (2007) has argued it may be most realistic to regard the majority of applications of GMMs in psychology as indirect, even if they were not originally motivated as such. There is then a lack of concordance between the direct application context assumed in class enumeration simulations and the actual application of GMMs in empirical research. Nevertheless, it is reasonable to conjecture that the same models and criteria that perform well in direct applications will also balance under- and over-fitting in indirect applications.

Second, our simulation results focused on the ability to select the proper number of classes. However, a reasonable follow-up question may be whether the growth trajectories in classes that are extracted from these models accurately reflect the generated class trajectories. Though we do not report this in the current paper, McNeish and Harring (2021) is fully dedicated to this question and finds that – assuming the number of classes is correct – the CPGMM class trajectories are faithfully recovered across converged replications whereas the HGMM class trajectories were less similar to the data generation model. We rely on previous studies to address this question rather than recreate and report such results here alongside these more novel class enumeration results.

Third, our simulation conditions were based upon data characteristics observed from the PTSD literature. GMMs are applied in many content areas in which the data characteristics may look quite different. To the extent that characteristics differ from the PTSD literature; convergence and class enumeration performance noted in our simulation may not carryover to other adjacent areas of application.

Fourth, we focused on the lower half of the sample size distribution where nonconvergence issues tend to be most rampant. CPGMMs performed well in these conditions within our simulation but the disparity between CPGMMs and other methods may diminish at larger sample sizes. A future direction would be studying the performance of CPGMMs in large samples to determine whether the performance increase is limited to modest samples or whether CPGMMs may be a preferable approach to combine latent class analysis with growth modeling more generally.

Fifth, the PTSD literature tends to have a fairly small number of repeated measures, which was reflected in our simulation design. However, modeling the marginal covariance with a CPGMM can increase in difficulty when data contain more repeated measures. For instance, with 10 repeated measures, it may be challenging to select a parsimonious marginal covariance structure, especially if the repeated measures are taken in highly unequal intervals. If an unrestricted marginal covariance is used with many repeated measures, this can require many parameters to be estimated.

Sixth, we assumed that normality assumptions were met. When normality is violated, class enumeration can be more difficult because it can confound classes that may be substantively meaningful and classes that are mathematical approximations for a complex distribution (e.g., Bauer, 2007; Depaoli et al., 2019; Guerra-Peña & Steinley, 2016). Robustness to non-normality for more recently developed and less well-studied methods like CPGMMs and PGMMs is a potential future direction. CPGMMs may fare better than PGMMs because there are no random effects and therefore fewer distributional assumptions to satisfy. These methods could also be extended to skew-normal or *t*-distributions as suggested by Muthén and Asparouhov (2015) to mitigate some effects of non-normality.

7.2 Concluding Remarks

Class enumeration is a critical but difficult aspect of modeling heterogeneity in growth trajectories with latent classes. Current approaches make an already difficult processes even more difficult by employing models that are unnecessary complex for the research question of interest as many research questions do not require within-class random effects. Modeling within-class random effects piles latent classes on top of latent growth trajectories and estimation algorithms can only extract so much latent information from a few observed repeated measures. Because of rigid disciplinary preferences for growth modeling with the random effects tradition, researchers often must choose between convergence and flexibly modeling the covariance structure in each class. However, expanding one's perspective beyond models in the random effects tradition can provide alternatives beyond this dichotomy.

Typically, the focus of GMMs is squarely on the latent classes and not on the person-specific growth trajectories. This interest can be reflected in the model by disregarding the within-class random effects and instead modeling the marginal covariance in each class directly with CPGMMs. We found this approach to improve convergence while also permitting class-specific covariance structures. This combination led to higher accuracy when enumerating classes in a variety of conditions, especially when the classes are reasonably well separated.

References

Anderlucci, L., & Viroli, C. (2015). Covariance pattern mixture models for the analysis of multivariate heterogeneous longitudinal data. *The Annals of Applied Statistics*, *9*(2), 777-800.

Andrews, B., Brewin, C. R., Philpott, R., & Stewart, L. (2007). Delayed-onset posttraumatic stress disorder: a systematic review of the evidence. *American Journal of Psychiatry*, *164*(9), 1319-1326.

Banfield, J. D., & Raftery, A. E. (1993). Model-based gaussian and non-gaussian clustering. *Biometrics*, *49*(3), 803–821.

Barnard, J., McCulloch, R., & Meng, X. L. (2000). Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statistica Sinica*, *10*, 1281-1311.

Bauer, D. J. (2007). Observations on the use of growth mixture models in psychological research. *Multivariate Behavioral Research*, *42*(4), 757-786.

Bauer, D. J., & Curran, P. J. (2004). The integration of continuous and discrete latent variable models: potential problems and promising opportunities. *Psychological Methods*, *9*(1), 3–29.

Bauer, D. J., & Curran, P. J. (2003). Distributional assumptions of growth mixture models: implications for overextraction of latent trajectory classes. *Psychological Methods*, *8*(3), 338–363.

Biernacki, C. (2005). Testing for a global maximum of the likelihood. *Journal of Computational and Graphical Statistics*, *14*(3), 657–674.

Biernacki, C., Celeux, G., & Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(7), 719-725.

Biernacki, C., & Govaert, G. (1997). Using the classification likelihood to choose the number of clusters. *Computing Science and Statistics*, *29*, 451-457.

Bonanno, G. A. (2004). Loss, trauma, and human resilience: Have we underestimated the human capacity to thrive after extremely aversive events? *American Psychologist*, *59*(1), 20–28.

Burton, P., Gurrin, L., & Sly, P. (1998). Extending the simple linear regression model to account for correlated responses: an introduction to generalized estimating equations and multi-level mixed modelling. *Statistics in Medicine*, *17*(11), 1261-1291.

Celeux, G., & Soromenho, G. (1996). An entropy criterion for assessing the number of clusters in a mixture model. *Journal of Classification*, *13*(2), 195-212.

Cole, V. T., & Bauer, D. J. (2016). A note on the use of mixture models for individual prediction. *Structural Equation Modeling*, *23*(4), 615–631.

Curran, P. J., Obeidat, K., & Losardo, D. (2010). Twelve frequently asked questions about growth curve modeling. *Journal of Cognition and Development*, *11*(2), 121-136.

Davies, C. E., Glonek, G. F., & Giles, L. C. (2017). The impact of covariance misspecification in group-based trajectory models for longitudinal data with non-stationary covariance structure. *Statistical Methods in Medical Research*, *26*(4), 1982–1991.

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

Depaoli, S., Winter, S. D., Lai, K., & Guerra-Peña, K. (2019). Implementing continuous non-normal skewed distributions in latent growth mixture modeling: An assessment of specification errors and class enumeration. *Multivariate Behavioral Research*, *54*(6), 795-821.

Diallo, T. M., Morin, A. J., & Lu, H. (2017). The impact of total and partial inclusion or exclusion of active and inactive time invariant covariates in growth mixture models. *Psychological Methods*, *22*(1), 166-190.

Diallo, T. M. O., Morin, A. J. S., & Lu, H. (2016). Impact of misspecifications of the latent variance–covariance and residual matrices on the class enumeration accuracy of growth mixture models. *Structural Equation Modeling*, *23*(4), 507–531.

Diallo, T. M., Morin, A. J., & Parker, P. D. (2014). Statistical power of latent growth curve models to detect quadratic growth. *Behavior Research Methods*, *46*(2), 357-371.

Diggle, P., Diggle, P. J., Heagerty, P., Heagerty, P. J., Liang, K. Y., & Zeger, S. (2002). *Analysis of longitudinal data*. New York: Oxford University Press.

Dolan, C. V., & van der Maas, H. L. J. (1998). Fitting multivariate normal finite mixtures subject to structural equation modeling. *Psychometrika, 63*, 227–253.

Draper, D. (1995). Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society: Series B*, *57*(1), 45-70.

Dziak, J. J., Lanza, S. T., & Tan, X. (2014). Effect size, statistical power, and sample size requirements for the bootstrap likelihood ratio test in latent class analysis. *Structural Equation Modeling*, *21*(4), 534-552.

Enders, C. K., & Tofighi, D. (2008). The impact of misspecifying class-specific residual variances in growth mixture models. *Structural Equation Modeling*, *15*(1), 75–95.

Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2012). *Applied Longitudinal Analysis*. Wiley.

Frueh, B. C., Grubaugh, A. L., Yeager, D. E., & Magruder, K. M. (2009). Delayed-onset post-traumatic stress disorder among war veterans in primary care clinics. *The British Journal of Psychiatry*, *194*(6), 515–520.

Gardiner, J. C., Luo, Z., & Roman, L. A. (2009). Fixed effects, random effects and GEE: what are the differences?. *Statistics in Medicine*, *28*(2), 221-239.

Gilthorpe, M. S., Dahly, D. L., Tu, Y.-K., Kubzansky, L. D., & Goodman, E. (2014). Challenges in modelling the random structure correctly in growth mixture models and the impact this has on model mixtures. *Journal of Developmental Origins of Health and Disease*, *5*(3), 197–205.

Grimm, K. J., Ram, N., & Estabrook, R. (2016). *Growth modeling: Structural equation and multilevel modeling approaches*. New York: Guilford Publications.

Guerra-Peña, K., & Steinley, D. (2016). Extracting spurious latent classes in growth mixture modeling with normal errors. *Educational and Psychological Measurement, 76* (6), 933-953.

Hanley, J. A., Negassa, A., Edwardes, M. D. D., & Forrester, J. E. (2003). Statistical analysis of correlated data using generalized estimating equations: an orientation. *American Journal of Epidemiology*, *157*(4), 364-375.

Hannan, E. J., & Quinn, B. G. (1979). The determination of the order of an autoregression. *Journal of the Royal Statistical Society: Series B*, *41*(2), 190-195.

Heagerty, P. J., & Zeger, S. L. (2000). Marginalized multilevel models and likelihood inference. *Statistical Science*, *15*(1), 1-26.

Hedeker, D., & Gibbons, R. D. (2006). *Longitudinal Data Analysis*. Wiley.

Hedeker, D., Mermelstein, R. J., & Demirtas, H. (2012). Modeling between-subject and within-subject variances in ecological momentary assessment data using mixed-effects location scale models. *Statistics in Medicine*, *31*(27), 3328-3336.

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.

Henderson, N. C., & Rathouz, P. J. (2018). AR (1) latent class models for longitudinal count data. *Statistics in Medicine*, *37*(29), 4441-4456.

Henson, J. M., Reise, S. P., & Kim, K. H. (2007). Detecting mixtures from structural model differences using latent variable mixture modeling: A comparison of relative model fit statistics. *Structural Equation Modeling*, *14*(2), 202-226.

Hipp, J. R., & Bauer, D. J. (2006). Local solutions in the estimation of growth mixture models. *Psychological Methods*, *11*(1), 36-53.

Hoffman, L. (2007). Multilevel models for examining individual differences in within-person variation and covariation over time. *Multivariate Behavioral Research*, *42*(4), 609-629.

Horowitz, M., Wilner, N., & Alvarez, W. (1979). Impact of Event Scale: A measure of subjective stress. *Psychosomatic Medicine*, *41*(3), 209-218.

Hubbard, A. E., Ahern, J., Fleischer, N. L., Van der Laan, M., Satariano, S. A., Jewell, N., ... & Satariano, W. A. (2010). To GEE or not to GEE: comparing population average and mixed models for estimating the associations between neighborhood risk factors and health. *Epidemiology*, *21*, 467-474.

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, *2*(1), 193-218.

Hullermeier, E., & Rifqi, M. (2009). A fuzzy variant of the Rand index for comparing clustering structures. *Joint 2009 International Fuzzy Systems Association World Congress and 2009 European Society of Fuzzy Logic and Technology Conference, IFSA-EUSFLAT 2009* (pp. 1294-1298).

Hurvich, C. M., & Tsai, C. L. (1989). Regression and time series model selection in small samples. *Biometrika*, *76*(2), 297-307.

Infurna, F. J., & Grimm, K. J. (2018). The use of growth mixture modeling for studying resilience to major life stressors in adulthood and old age: Lessons for class size and identification and model selection. *The Journals of Gerontology: Series B*, *73*(1), 148–159.

Infurna, F. J., & Jayawickreme, E. (2019). Fixing the growth illusion: New directions for research in resilience and posttraumatic growth. *Current Directions in Psychological Science*, *28*(2), 152–158.

Infurna, F. J., & Luthar, S. S. (2016). Resilience to major life stressors is not as common as thought. *Perspectives on Psychological Science*, *11*(2), 175–194.

Jacob, T., Bucholz, K. K., Sartor, C. E., Howell, D. N., & Wood, P. K. (2005). Drinking trajectories from adolescence to the mid-forties among alcohol dependent males. *Journal of Studies on Alcohol*, *66*(6), 745-755.

Jennrich, R. I., & Schluchter, M. D. (1986). Unbalanced repeated-measures models with structured covariance matrices. *Biometrics*, *42*(4), 805–820.

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.

Kiernan, K. (2018). Insights into using the GLIMMIX procedure to model categorical outcomes with random effects (SAS2179–2018). SAS Institute.

Kim, S.-Y. (2014). Determining the number of latent classes in single- and multiphase growth mixture models. *Structural Equation Modeling, 21*, 263-279.

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

Kooken, J., McCoach, D. B., & Chafouleas, S. M. (2019). The impact and interpretation of modeling residual noninvariance in growth mixture models. *The Journal of Experimental Education*, *87*(2), 214–237.

Kreuter, F., & Muthén, B. (2008). Analyzing criminal trajectory profiles: bridging multilevel and group-based approaches using growth mixture modeling. *Journal of Quantitative Criminology*, *24*(1), 1–31.

Kreuter, F., & Muthén, B. O. (2007). Longitudinal modeling of population heterogeneity: Methodological challenges to the analysis of empirically derived criminal trajectory profiles. In G.R.

Hancock & K. M. Samuleson (Eds.), *Advances in latent variable mixture models* (pp. 53–75). Information Age Publishing: Charlotte, NC.

Lee, Y., & Nelder, J. A. (2004). Conditional and marginal models: Another view. *Statistical Science*, *19*(2), 219–238.

Li, M., Harring, J. R., & Macready, G. B. (2014). Investigating the feasibility of using Mplus in the estimation of growth mixture models. *Journal of Modern Applied Statistical Methods*, *13*(1), 484–513.

Li, L., & Hser, Y. I. (2011). On inclusion of covariates for class enumeration of growth mixture models. *Multivariate Behavioral Research*, *46*(2), 266-302.

Liu, M., & Hancock, G. R. (2014). Unrestricted mixture models for class identification in growth mixture modeling. *Educational and Psychological Measurement*, *74*(4), 557–584.

Liu, C., & Rubin, D. B. (1998). Ellipsoidally symmetric extensions of the general location model for mixed categorical and continuous data. *Biometrika*, *85*(3), 673-688.

Lo, Y., Mendell, N. R., & Rubin, D. B. (2001). Testing the number of components in a normal mixture. *Biometrika*, *88*(3), 767-778.

Maher, J. P., Ra, C. K., Leventhal, A. M., Hedeker, D., Huh, J., Chou, C. P., & Dunton, G. F. (2018). Mean level of positive affect moderates associations between volatility in positive affect, mental health, and alcohol consumption among mothers. *Journal of Abnormal Psychology*, *127*, 639-649.

Manly, B. F., & Rayner, J. C. W. (1987). The comparison of sample covariance matrices using likelihood ratio tests. *Biometrika*, *74*(4), 841-847.

McArdle, J. J., & Epstein, D. (1987). Latent growth curves within developmental structural equation models. *Child Development*, *58*(1), 110–133.

McLachlan, G. J., & Peel, D. (2004). *Finite mixture models*. Wiley.

McLachlan, G. J. (1987). On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. *Journal of the Royal Statistical Society: Series C*, *36*(3), 318-324.

McNally, R. J. (2003). Progress and controversy in the study of posttraumatic stress disorder. *Annual Review of Psychology*, *54*(1), 229–252.

McNeish, D. (2021, February 22). Nonconvergence, Covariance Constraints, and Class Enumeration in Growth Mixture Models. Open Science Framework page. Retrieved from osf.io/eay5v

McNeish, D. & Harring, J.R. (2021). Improving convergence in growth mixture models without covariance structure constraints. *Statistical Methods in Medical Research.*

McNeish, D. & Harring, J.R. (2020). Covariance pattern mixture models: Eliminating random effects to improve convergence and performance. *Behavior Research Methods, 52*, 947-979.

McNeish, D., Stapleton, L. M., & Silverman, R. D. (2017). On the unnecessary ubiquity of hierarchical linear modeling. *Psychological Methods*, *22*(1), 114–140.

Meredith, W., & Tisak, J. (1990). Latent curve analysis. *Psychometrika*, *55*(1), 107–122.

Milligan, G. W., & Cooper, M. C. (1986). A study of the comparability of external criteria for hierarchical cluster analysis. *Multivariate Behavioral Research, 21* (4), 441– 458.

Muthén, B. O. (2004). Latent variable analysis: Growth mixture modeling and related techniques for longitudinal data. In D. Kaplan (Ed.), *Handbook of quantitative methodology for the social sciences* (pp. 345–368). Sage.

Muthén, B. O. (2002). Beyond SEM: General latent variable modeling. *Behaviormetrika*, *29*(1), 81–117.

Muthén, B., & Asparouhov, T. (2015). Growth mixture modeling with non-normal distributions. *Statistics in Medicine*, *34*(6), 1041-1058.

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

Nagin, D. S. (1999). Analyzing developmental trajectories: A semiparametric, group-based approach. *Psychological Methods*, *4*(2), 139–157.

Nagin, D. S., & Tremblay, R. E. (2005). Developmental trajectory groups: Fact or a useful statistical fiction?. *Criminology*, *43*(4), 873–904.

Nagin, D. S., & Tremblay, R. E. (2001). Analyzing developmental trajectories of distinct but related behaviors: A group-based method. *Psychological Methods*, *6*(1), 18–34.

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.

Nesselroade, J. R. (1991). Interindividual differences in intraindividual change. In *Best methods for the analysis of change: Recent advances, unanswered questions, future directions* (pp. 92–105). American Psychological Association.

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*, *14*(4), 535–569.

Pastor, D. A., & Gagné, P. (2013). Mean and covariance structure mixture models. In *Structural equation modeling: A second course, 2nd ed* (pp. 343–393). IAP Information Age Publishing.

Peugh, J., & Fan, X. (2013). Modeling unobserved heterogeneity using latent profile analysis: A Monte Carlo simulation. *Structural Equation Modeling*, *20*(4), 616-639.

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*, *19*(2), 204-226.

Proust-Lima, C., Philipps, V., & Liquet, B. (2017) Estimation of extended mixed models using latent classes and latent processes: The R package lcmm. *Journal of Statistical Software*, *78*(2), 1-56.

Pugach, O., Hedeker, D., Richmond, M. J., Sokolovsky, A., & Mermelstein, R. (2013). Modeling mood variation and covariation among adolescent smokers: application of a bivariate location-scale mixed-effects model. *Nicotine & Tobacco Research*, *16*, S151-S158.

Ram, N., & Grimm, K. J. (2009). Growth mixture modeling: A method for identifying differences in longitudinal change among unobserved groups. *International Journal of Behavioral Development*, *33*(6), 565–576.

Ram, N., & Grimm, K. (2007). Using simple and complex growth models to articulate developmental change: Matching theory to method. *International Journal of Behavioral Development*, *31*(4), 303–316.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, *66*(336), 846-850.

Rosen, O., Jiang, W., & Tanner, M. A. (2000). Mixtures of marginal models. *Biometrika*, *87*(2), 391-404.

Sclove, S. L. (1987). Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika*, *52*(3), 333-343.

Schluchter, M. D. (1988). Analysis of incomplete multivariate data using linear models with structured covariance matrices. *Statistics in Medicine*, *7*(1–2), 317–324.

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

Scrucca L, Fop M, Murphy TB, Raftery AE (2016). mclust 5: Clustering, classification and density estimation using Gaussian finite mixture models. *The R Journal*, 8, 289–317.

Sher, K. J., Jackson, K. M., & Steinley, D. (2011). Alcohol use trajectories and the ubiquitous cat's cradle: Cause for concern? *Journal of Abnormal Psychology*, *120*(2), 322–335.

Shireman, E., Steinley, D., & Brusco, M. J. (2017). Examining the effect of initialization strategies on the performance of mixture modeling. *Behavior Research Methods, 49* (1), 282-293.

Shireman, E. , Steinley, D. & Brusco, M. J. (2016). Local optima in mixture modeling. *Multivariate Behavioral Research, 51* (4), 466-481.

Shiyko, M. P., Ram, N., & Grimm, K. J. (2012). An overview of growth mixture modeling: A simple nonlinear application in OpenMx. In *Handbook of structural equation modeling* (pp. 532–546). The Guilford Press.

Sijbrandij, J. J., Hoekstra, T., Almansa, J., Peeters, M., Bültmann, U., & Reijneveld, S. A. (2020). Variance constraints strongly influenced model performance in growth mixture modeling: a simulation and empirical study. *BMC Medical Research Methodology*, *20*(1), 1-15.

Sijbrandij, J. J., Hoekstra, T., Almansa, J., Reijneveld, S. A., & Bültmann, U. (2019). Identification of developmental trajectory classes: Comparing three latent class methods using simulated and real data. *Advances in Life Course Research*, *42*, 100288.

Spitzer, R. L., First, M. B., & Wakefield, J. C. (2007). Saving PTSD from itself in DSM-V. *Journal of Anxiety Disorders*, *21*(2), 233–241.

Steinley, D. (2004). Properties of the Hubert-Arable Adjusted Rand Index. *Psychological Methods*, *9*(3), 386-396.

Steinley, D., & Brusco, M. J. (2018). A note on the expected value of the Rand index. *British Journal of Mathematical and Statistical Psychology*, *71*(2), 287-299.

Steinley, D., & Brusco, M. J. (2011). Evaluating mixture modeling for clustering: recommendations and cautions. *Psychological Methods*, *16*(1), 63-79.

Steinley, D., Brusco, M. J, & Hubert, L. (2016). The variance of the adjusted Rand index. *Psychological Methods, 21* (2), 261-272.

Sterba, S. K., & Bauer, D. J. (2010). Matching method with theory in person-oriented developmental psychopathology research. *Development and Psychopathology*, *22*(2), 239–254.

Sterba, S. K., & Bauer, D. J. (2014). Predictions of individual change recovered with latent class or random coefficient growth models. *Structural Equation Modeling*, *21*(3), 342–360.

Tang, X., & Qu, A. (2016). Mixture modeling for longitudinal data. *Journal of Computational and Graphical Statistics*, *25*(4), 1117-1137.

Tein, J. Y., Coxe, S., & Cham, H. (2013). Statistical power to detect the correct number of classes in latent profile analysis. *Structural Equation Modeling*, *20*(4), 640-657.

Tekle, F. B., Gudicha, D. W., & Vermunt, J. K. (2016). Power analysis for the bootstrap likelihood ratio test for the number of classes in latent class models. *Advances in Data Analysis and Classification*, *10*(2), 209-224.

Titterington, D. M., Smith, A. F. M., & Makov, U. E. (1985). *Statistical analysis of finite mixture distributions*. Chichester, England: Wiley

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

Tueller, S. J., Drotar, S., & Lubke, G. H. (2011). Addressing the problem of switched class labels in latent variable mixture model simulation studies. *Structural Equation Modeling*, *18*(1), 110-131.

Tueller, S., & Lubke, G. (2010). Evaluation of structural equation mixture models: Parameter estimates and correct class assignment. *Structural Equation Modeling*, *17*(2), 165–192.

Twisk, J., & Hoekstra, T. (2012). Classifying developmental trajectories over time should be done with great caution: a comparison between methods. *Journal of Clinical Epidemiology*, *65*(10), 1078-1087.

Van Loey, N. E. E., Maas, C. J. M., Faber, A. W., & Taal, L. A. (2003). Predictors of chronic posttraumatic stress symptoms following burn injury: results of a longitudinal study. *Journal of Traumatic Stress*, *16*(4), 361-369.

van de Schoot, R., Sijbrandij, M., Depaoli, S., Winter, S. D., Olff, M., & Loey, N. E. van. (2018). Bayesian PTSD-trajectory analysis with informed priors based on a systematic literature search and expert elicitation. *Multivariate Behavioral Research*, *53*(2), 267–291.

van Loey, N. E. E., Maas, C. J. M., Faber, A. W., & Taal, L. A. (2003). Predictors of chronic posttraumatic stress symptoms following burn injury: Results of a longitudinal study. *Journal of Traumatic Stress*, *16*(4), 361–369.

Vavrek, M. J. (2011). Fossil: palaeoecological and palaeogeographical analysis tools. *Palaeontologia Electronica*, *14*(1), 16.

Verbeke, G., & Lesaffre, E. (1996). A linear mixed-effects model with heterogeneity in the random-effects population. *Journal of the American Statistical Association*, *91*(433), 217–221.

Verbeke, G., & Molenberghs, G. (2009). *Linear mixed models for longitudinal data*. Springer

Weinstein, S. M., & Mermelstein, R. J. (2013). Influences of mood variability, negative moods, and depression on adolescent cigarette smoking. *Psychology of Addictive Behaviors*, *27*(4), 1068-1078.

Wickrama, K. K. A. S., Lee, T. K., O'Neal, C. W., Lorenz, F. O., Lee, T. K., O'Neal, C. W., & Lorenz, F. O. (2016). *Higher-Order growth curves and mixture modeling with Mplus: A practical guide*. Routledge.

Williams, D. R., Mulder, J., Rouder, J. N., & Rast, P. (2020). Beneath the surface: Unearthing within-person variability and mean relations with Bayesian mixed models. *Psychological Methods*, advance online publication.

Yang, C. C. (2006). Evaluating latent class analysis models in qualitative phenotype identification. *Computational Statistics & Data Analysis*, *50*(4), 1090-1104.

Zeger, S. L., Liang, K. Y., & Albert, P. S. (1988). Models for longitudinal data: a generalized estimating equation approach. *Biometrics*, *44*, 1049-1060.

Table 1

Data generation equations for population model

Class	Model Equation	Covariance Structures, Low Separation	Covariance Structures, High Separation
1 (Resilient)	$y_{ti} = \eta_{0i} + \eta_{1i} t + \varepsilon_{ti}$ $\eta_{0i} = 15.0 + \zeta_{0i}$ $\eta_{1i} = -0.15 + \zeta_{1i}$	$\mathbf{\Theta} = 28\mathbf{I}_5$ $\mathbf{\Psi} = \begin{bmatrix} 40 & \\ 0 & 0.04 \end{bmatrix}$	$\mathbf{\Theta} = 25\mathbf{I}_5$ $\mathbf{\Psi} = \begin{bmatrix} 25 & \\ 0 & 0.02 \end{bmatrix}$
2 (Recovering)	$y_{ti} = \eta_{0i} + \eta_{1i} t + \eta_{2i} t^2 + \varepsilon_{ti}$ $\eta_{0i} = 38.0 + \zeta_{0i}$ $\eta_{1i} = -0.10 + \zeta_{1i}$ $\eta_{2i} = -0.008$	$\mathbf{\Theta} = 72\mathbf{I}_5$ $\mathbf{\Psi} = \begin{bmatrix} 160 & \\ 0 & 0.04 \end{bmatrix}$	$\mathbf{\Theta} = 10\mathbf{I}_5$ $\mathbf{\Psi} = \begin{bmatrix} 20 & \\ 0 & 0.01 \end{bmatrix}$
3 (Chronic)	$y_{ti} = \eta_{0i} + \eta_{1i} t + \eta_{2i} t^2 + \varepsilon_{ti}$ $\eta_{0i} = 41.0 + \zeta_{0i}$ $\eta_{1i} = 0.12 + \zeta_{1i}$ $\eta_{2i} = -0.002$	$\mathbf{\Theta} = 60\mathbf{I}_5$ $\mathbf{\Psi} = \begin{bmatrix} 120 & \\ 0 & 0.04 \end{bmatrix}$	$\mathbf{\Theta} = 15\mathbf{I}_5$ $\mathbf{\Psi} = \begin{bmatrix} 15 & \\ 0 & 0.015 \end{bmatrix}$
4 (Delayed Onset)	$y_{ti} = \eta_{0i} + \eta_{1i} t + \eta_{2i} t^2 + \varepsilon_{ti}$ $\eta_{0i} = 18.0 + \zeta_{0i}$ $\eta_{1i} = 0.20 + \zeta_{1i}$ $\eta_{2i} = 0.015$	$\mathbf{\Theta} = 60\mathbf{I}_5$ $\mathbf{\Psi} = \begin{bmatrix} 72 & \\ 0 & .035 \end{bmatrix}$	$\mathbf{\Theta} = 12\mathbf{I}_5$ $\mathbf{\Psi} = \begin{bmatrix} 10 & \\ 0 & 0.015 \end{bmatrix}$

Note: The 3-class model uses the same equation for Classes 1 through 3 but omits Class 4 and changes the class proportions; t = 0, 1, 10, 18, 26

Table 2
Equations for information criteria considered in the simulation

Criteria	Equation
BIC	$-2\ell + p\ln(N)$
SABIC	$-2\ell + p\ln\left[(N+2)/24\right]$
DBIC	$-2\ell + p\left[\ln(N) - \ln(2\pi)\right]$
ICL-BIC	$-2\ell + p\ln(N) + 2\text{Entropy}$
ICL-SABIC	$-2\ell + p\ln\left[(N+2)/24\right] + 2\text{Entropy}$
ICL-DBIC	$-2\ell + p\left[\ln(N) - \ln(2\pi)\right] + 2\text{Entropy}$
CLC	$-2\ell + 2\text{Entropy}$
HT-AIC	$-2\ell + 2p + \left[2(p+1)(p+2)\right]/(N-p-2)$
HQ-AIC	$-2\ell + 2p\left[\ln(\ln(N))\right]$

Note: ℓ = loglikelihood, p = number of estimated parameters, N = sample size. Entropy is not the same as the relative entropy value output by M*plus*; $Entropy = -N(RelEntropy - 1)\ln(K)$ for K the number of classes.

Table 3

Percentage of replications converging by model type and simulation condition

		4-Class Generation Model														
		GMMU			CPGMM			HGMM			PGMM			LCGA		
Class Separation	Class Solution	100	300	500	100	300	500	100	300	500	100	300	500	100	300	500
High	2-Class				69	67	63	100	100	100	100	100	100	99	100	100
	3-Class				82	91	95	58	67	63	40	32	34	97	100	100
	4-Class	2	17	36	66	95	98	35	48	60	7	15	21	93	99	100
	5-Class				35	68	83	21	41	46	1	0	0	82	95	98
Low	2-Class				99	100	100	96	100	100	99	100	100	99	100	100
	3-Class				97	99	100	71	79	80	79	92	97	96	100	100
	4-Class	0	3	6	63	79	88	51	75	74	35	50	50	96	99	100
	5-Class				42	57	67	27	61	68	31	51	54	87	99	100
		3-Class Generation Model														
		GMMU			CPGMM			HGMM			PGMM			LCGA		
		100	300	500	100	300	500	100	300	500	100	300	500	100	300	500
High	2-Class				97	100	100	99	100	100	100	100	100	99	100	100
	3-Class	19	44	66	82	99	100	46	37	27	51	52	60	46	96	97
	4-Class				49	79	87	26	37	38	6	11	14	26	99	100
Low	2-Class				97	100	100	91	100	100	98	100	100	100	100	100
	3-Class	11	22	30	95	99	100	64	79	76	72	81	73	97	100	100
	4-Class				74	74	80	50	79	83	32	47	45	97	99	100

Note: GMMU = Unconstrained Growth Mixture Model, CPGMM = Covariance Pattern Growth Mixture Model, HGMM = Homoskedastic Growth Mixture Model, LCGA = Latent Class Growth Analysis. The GMMU is only fit with the correct number of classes to serve as a reference.

Table 4

Percent of total replications (top) and percentage of converged replications (bottom) identifying the correct number of classes for the 4-class data generation model with high class separation (population relative entropy = 0.90)

	All Replications											
	CPGMM			HGMM			PGMM			LCGA		
	100	300	500	100	300	500	100	300	500	100	300	500
BIC	28	**90**	**92**	1	11	35	0	5	20	21	5	2
SABIC	43	43	28	22	30	38	6	14	21	16	5	2
DBIC	**52**	**77**	**63**	11	29	43	3	13	21	17	5	2
ICL-BIC	24	**61**	**80**	0	0	0	3	2	1	32	11	2
ICL-SABIC	46	**75**	**90**	2	0	0	2	2	5	21	7	2
ICL-DBIC	41	**71**	**87**	1	0	0	1	1	3	25	8	2
CLC	45	**77**	**91**	5	1	0	1	9	12	18	6	2
HT-AIC	**50**	41	19	10	30	33	2	14	21	18	5	2
HQ-AIC	**50**	**70**	42	9	30	42	2	13	21	17	5	2
	Only Converged Replications											
	CPGMM			HGMM			PGMM			LCGA		
	100	300	500	100	300	500	100	300	500	100	300	500
BIC	42	**95**	**94**	3	23	58	0	33	**95**	23	5	2
SABIC	**65**	45	29	**63**	**63**	**63**	**86**	**93**	**100**	17	5	2
DBIC	**79**	**81**	**64**	31	**60**	**72**	43	**87**	**100**	18	5	2
ICL-BIC	36	**64**	**82**	0	0	0	43	13	5	34	11	2
ICL-SABIC	**70**	**79**	**92**	6	0	0	29	13	24	23	7	2
ICL-DBIC	**62**	**75**	**89**	3	0	0	14	7	14	27	8	2
CLC	**68**	**81**	**93**	14	2	0	14	**60**	**57**	19	6	2
HT-AIC	**76**	43	19	29	**63**	**55**	29	**93**	**100**	19	5	2
HQ-AIC	**76**	**74**	43	26	**63**	**70**	29	**87**	**100**	18	5	2
Convergence %	66	95	98	35	48	60	7	15	21	93	99	100
Relative Entropy	.92	.91	.91	.84	.80	.79	.92	.89	.89	.93	.91	.90
Adjusted Rand Index	.80	.87	.88	.57	.60	.60	.57	.75	.82	.49	.48	.48

Note: CPGMM = Covariance Pattern Growth Mixture Model, HGMM = Homoskedastic Growth Mixture Model, PGMM = Proportional Growth Mixture Model, LCGA = Latent Class Growth Analysis. Entries for the "Relative Entropy" row are the average entropy across all replications within the respective condition. Bold entries indicate that the information criteria identified the correct number of classes in at least 50% of replications.

Table 5

Percent of total replications (top) and percentage of converged replications (bottom) identifying the correct number of classes for 3-class data generation model with high class separation (population relative entropy = 0.90)

	All Replications											
	CPGMM			HGMM			PGMM			LCGA		
	100	300	500	100	300	500	100	300	500	100	300	500
BIC	**61**	**95**	**88**	4	7	10	11	42	**59**	28	1	0
SABIC	43	35	25	26	8	14	48	**50**	**57**	28	1	0
DBIC	**68**	**75**	**54**	17	13	15	37	**50**	**59**	28	1	0
ICL-BIC	40	**81**	**92**	1	0	0	3	2	1	28	1	0
ICL-SABIC	**60**	**89**	**96**	3	0	0	23	8	6	28	1	0
ICL-DBIC	**57**	**88**	**95**	1	0	0	14	4	4	28	1	0
CLC	**58**	**87**	**96**	6	0	0	33	16	10	28	1	0
HT-AIC	**74**	30	15	19	6	11	35	**50**	**53**	28	1	0
HQ-AIC	**71**	**65**	38	13	13	15	32	**50**	**58**	28	1	0
	Only Converged Replications											
	CPGMM			HGMM			PGMM			LCGA		
	100	300	500	100	300	500	100	300	500	100	300	500
BIC	**74**	**96**	**88**	9	19	37	22	**81**	**98**	61	1	0
SABIC	6	35	25	**57**	22	**52**	**94**	**96**	**95**	61	1	0
DBIC	**70**	**76**	**54**	37	35	**56**	**73**	**96**	**98**	61	1	0
ICL-BIC	48	**82**	**92**	2	0	0	6	4	2	61	1	0
ICL-SABIC	**54**	**90**	**96**	7	0	0	45	15	10	61	1	0
ICL-DBIC	**63**	**89**	**95**	2	0	0	27	8	7	61	1	0
CLC	38	**88**	**96**	13	0	0	**65**	31	17	61	1	0
HT-AIC	**85**	30	15	41	16	41	**69**	**96**	**88**	61	1	0
HQ-AIC	**78**	**66**	38	28	35	**56**	**63**	**96**	**97**	61	1	0
Convergence %	82	99	100	46	37	27	51	52	60	46	96	97
Relative Entropy	.92	.91	.91	.84	.80	.79	.92	.90	.89	.96	.95	.95
Adjusted Rand Index	.86	.89	.90	.69	.72	.70	.86	.86	.88	.69	.71	.74

Note: CPGMM = Covariance Pattern Growth Mixture Model, HGMM = Homoskedastic Growth Mixture Model, PGMM = Proportional Growth Mixture Model, LCGA = Latent Class Growth Analysis. Entries for the “Relative Entropy” row are the average entropy across all replications within the respective condition. Bold entries indicate that the information criteria identified the correct number of classes in at least 50% of replications.

Table 6

Percent of total replications (top) and percentage of converged replications (bottom) identifying the correct number of classes for 4-class data generation model with low class separation (population relative entropy = 0.70)

	All Replications											
	CPGMM			HGMM			PGMM			LCGA		
	100	300	500	100	300	500	100	300	500	100	300	500
BIC	0	2	18	12	8	17	0	0	0	29	1	0
SABIC	35	48	**60**	30	36	33	20	12	7	12	1	0
DBIC	12	27	**57**	13	30	38	3	1	1	13	1	0
ICL-BIC	1	1	0	8	0	0	0	0	2	49	22	17
ICL-SABIC	17	2	0	14	2	1	6	1	0	18	10	6
ICL-DBIC	8	1	0	5	1	0	2	0	0	29	14	11
CLC	25	6	2	25	11	2	10	4	2	14	5	3
HT-AIC	7	**52**	42	11	37	26	1	17	25	15	1	0
HQ-AIC	7	35	**68**	9	34	38	1	2	3	14	1	0
	Only Converged Replications											
	CPGMM			HGMM			PGMM			LCGA		
	100	300	500	100	300	500	100	300	500	100	300	500
BIC	0	3	20	24	11	23	0	0	0	30	1	0
SABIC	**56**	**61**	**68**	**59**	48	45	**57**	24	14	13	1	0
DBIC	19	34	**65**	25	40	**51**	9	2	2	14	1	0
ICL-BIC	2	10	0	16	0	0	0	0	4	**51**	22	17
ICL-SABIC	27	1	0	27	3	1	17	2	0	19	10	6
ICL-DBIC	13	1	0	10	1	0	6	0	0	30	14	11
CLC	40	8	2	49	15	3	29	8	4	15	5	3
HT-AIC	11	**66**	48	22	49	35	3	34	**50**	16	1	0
HQ-AIC	11	44	**77**	18	45	51	3	4	6	15	1	0
Convergence %	63	79	88	51	75	74	35	50	50	96	99	100
Relative Entropy	.78	.69	.66	.89	.86	.85	.80	.75	.72	.91	.88	.88
Adjusted Rand Index	.44	.45	.43	.57	.59	.59	.57	.58	.60	.32	.31	.31

Note: CPGMM = Covariance Pattern Growth Mixture Model, HGMM = Homoskedastic Growth Mixture Model, PGMM = Proportional Growth Mixture Model, LCGA = Latent Class Growth Analysis. Entries for the “Relative Entropy” row are the average entropy across all replications within the respective condition. Bold entries indicate that the information criteria identified the correct number of classes in at least 50% of replications.

Table 7

Percent of total replications (top) and percentage of converged replications (bottom) identifying the correct number of classes for 3-class data generation model with low class separation (population relative entropy = 0.70)

	All Replications											
	CPGMM			HGMM			PGMM			LCGA		
	100	300	500	100	300	500	100	300	500	100	300	500
BIC	9	35	**72**	10	35	48	1	0	0	3	1	0
SABIC	42	**50**	34	35	26	15	46	23	16	3	1	0
DBIC	42	**69**	**71**	36	45	30	12	5	2	3	1	0
ICL-BIC	8	1	0	7	4	1	1	0	0	6	1	0
ICL-SABIC	25	3	1	21	8	2	14	4	2	4	1	0
ICL-DBIC	17	1	0	16	5	1	6	2	1	4	1	0
CLC	29	7	1	24	13	4	23	10	5	3	1	0
HT-AIC	45	44	23	38	21	13	11	32	35	3	1	0
HQ-AIC	36	**68**	**54**	31	42	21	7	8	5	3	1	0
	Only Converged Replications											
	CPGMM			HGMM			PGMM			LCGA		
	100	300	500	100	300	500	100	300	500	100	300	500
BIC	9	35	**72**	16	44	**63**	1	0	0	3	1	0
SABIC	44	**51**	34	**55**	33	20	**64**	28	22	3	1	0
DBIC	44	**70**	**71**	**56**	57	39	17	6	3	3	1	0
ICL-BIC	8	1	0	11	5	1	1	0	0	6	1	0
ICL-SABIC	26	3	1	33	10	3	19	5	3	4	1	0
ICL-DBIC	18	1	0	25	6	1	8	2	1	4	1	0
CLC	31	7	1	38	16	5	32	12	7	3	1	0
HT-AIC	47	44	23	59	27	17	15	40	48	3	1	0
HQ-AIC	38	**69**	**54**	48	**53**	28	10	10	7	3	1	0
Convergence %	95	99	100	64	79	76	72	81	73	97	100	100
Relative Entropy	.76	.69	.66	.87	.84	.83	.80	.74	.72	.92	.90	.90
Adjusted Rand Index	.46	.47	.45	.55	.56	.56	.54	.60	.62	.39	.37	.36

Note: CPGMM = Covariance Pattern Growth Mixture Model, HGMM = Homoskedastic Growth Mixture Model, LCGA = Latent Class Growth Analysis. Entries for the "Relative Entropy" row are the average entropy across all replications within the respective condition. Bold entries indicate that the information criteria identified the correct number of classes in at least 50% of replications.

Table 8
Information criteria for empirical PTSD data with between 2 and 5 latent classes

HGMM					
	$k = 2$	$k = 3$	$k = 4$	$k = 5$	Classes
BIC	9,076	9,042	9,030	**9,011**	5
SABIC	9,041	8,995	8,970	**8,938**	5
DBIC	9,056	9,015	8,995	**8,969**	5
ICL-BIC	9,159	**9,128**	9,156	9,133	3
ICL-SABIC	9,124	9,080	9,095	**9,060**	5
ICL-DBIC	9,139	9,100	9,121	**9,090**	5
CLC	9,096	9,042	9,047	**9,001**	5
HT-AIC	9,036	8,989	8,963	**8,931**	5
HQ-AIC	9,051	9,009	8,988	**8,960**	5
Relative Entropy	.80	.87	.85	.88	
CPGMM					
	$k = 2$	$k = 3$	$k = 4$	$k = 5$	Classes
BIC	8,766	8,678	**8,674**	8,710	4
SABIC	8,681	8,548	8,499	**8,491**	5
DBIC	8,717	8,603	**8,573**	8,583	4
ICL-BIC	8,806	**8,753**	8,808	8,890	3
ICL-SABIC	8,720	**8,623**	8,634	8,672	3
ICL-DBIC	8,756	**8,677**	8,707	8,764	3
CLC	8,652	8,519	**8,494**	8,497	4
HT-AIC	8,672	8,540	**8,496**	8,498	4
HQ-AIC	8,706	8,587	**8,552**	8,557	4
Relative Entropy	.91	.89	.84	.81	

Note: CPGMM = Covariance Pattern Growth Mixture Model, HGMM = Homoskedastic Growth Mixture Model, bold entries indicate that the information criterion was the lowest across the different class solutions

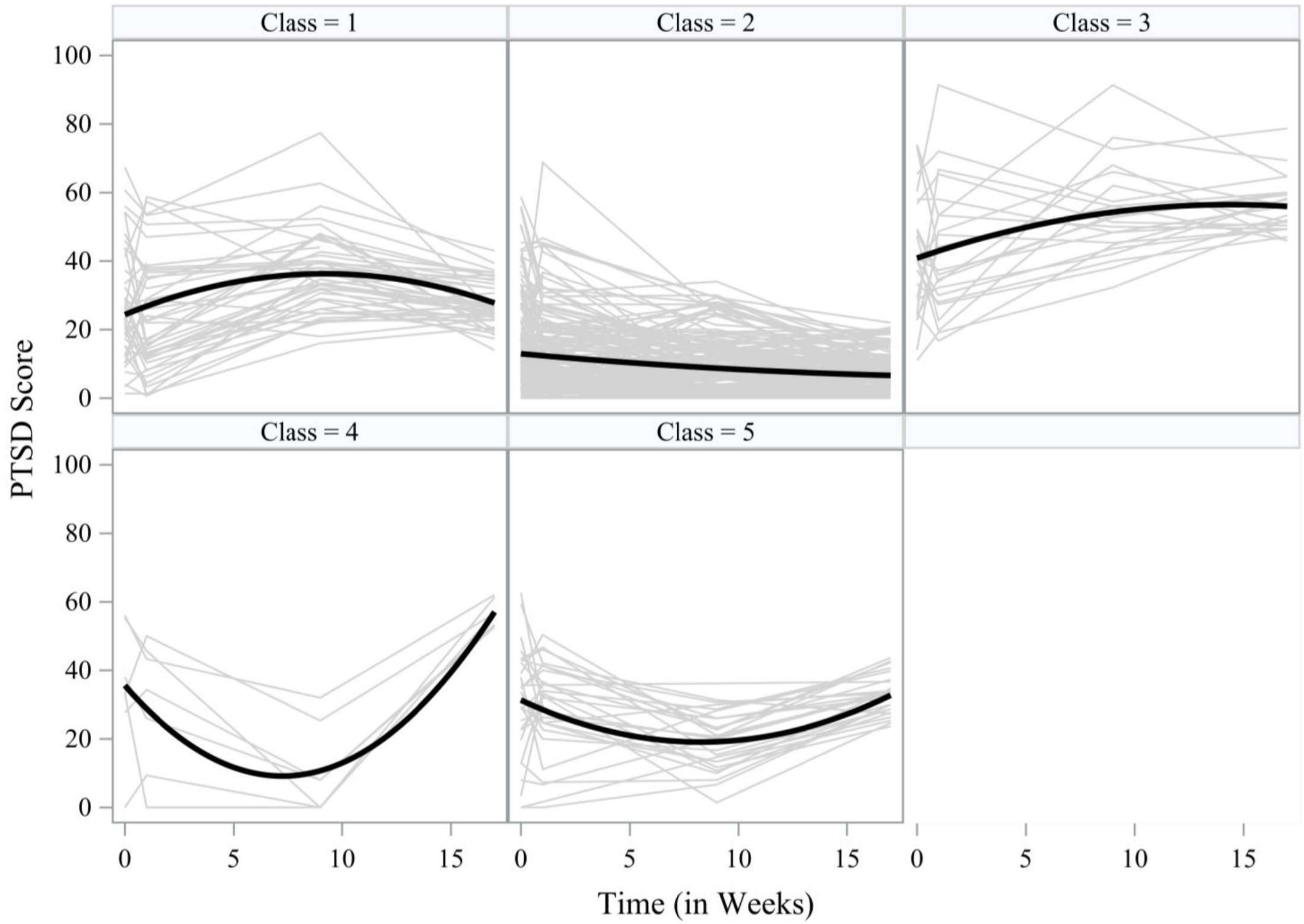

Figure 1. 5-Class HGMM class-specific trajectories (in black) with the empirical data of the people assigned to each class shown in grey

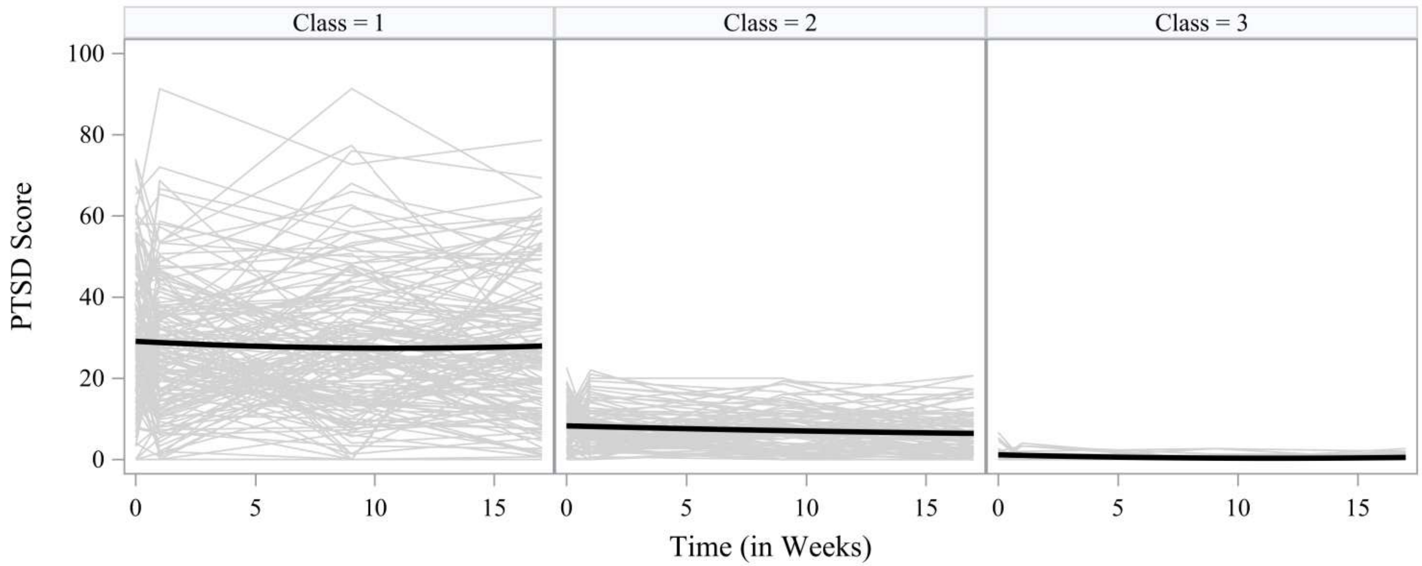

Figure 2. 3-Class CPGMM class-specific trajectories (in black) with the empirical data of the people assigned to each class shown in grey

Appendix A

Evidence from PTSD Literature to Inform Simulation Conditions

To provide context and evidence for the nonconvergence in empirical studies and to gauge data characteristics seen in practice, we consider findings from the PTSD literature where GMM applications are commonplace. To facilitate this review, we use the work of van de Schoot et al. (2018) as a baseline, whose review screened 11,395 papers that satisfied keywords, ultimately whittling down to 34 papers containing 38 unique studies that all used the Impact of Event Scale (Horowitz et al., 1979). The original goal of their review was to create informative prior distributions for a Bayesian GMM analysis. To address the interest of the current paper, we re-reviewed these studies with a focus on the characteristics of the data and the modeling decisions used to arrive at the final model (aspects not tracked in the original review).

The modeling decisions of these studies were telling for how researchers dealt with nonconvergence. First, only 2 papers (6%) reported using an unconstrained GMM with class-specific covariance structures whereas 9 studies (27%) reported a HGMM. Another 9 studies (27%) did not provide enough information in the reports to determine if there were covariance structure constraints across classes. Presumably, these 9 studies used a HGMM because this is the default setting in the M*plus* software reported. An additional 14 studies (41%) used a LCGA and no studies reported using a PGMM, presumably because M*plus* rather than R is the most popular software in this area. A breakdown of the number of classes extracted in these studies by model type is provided in Table A1.

The prevalence of model choices reflecting nonconvergence issues is less surprising when looking at the attributes of the data in these studies. The median sample size was 517, with the first and third quartiles being 207 and 835, respectively (range: 70 to 16,488). In the methodological literature on GMMs, 300 is a typical lower bound for sample size in simulations with complete data (Diallo et al., 2016; Enders & Tofighi, 2008). Additionally, 23 studies reported the relative entropy of the final model; the median was 0.85 with the first and third quartiles being 0.72 and 0.93, respectively (range: 0.49 to 0.97). High class separation (as measured by relative entropy) can offset estimation difficulties with smaller sample sizes, but it appears that a number of studies may have relatively poor separation (loosely defined as relative entropy below 0.80; Celeux & Soromenho, 1996).

Table A1
Class enumeration of 34 PTSD studies by model characteristics

Classes	All Studies	GMM	LCGA	HGMM
2	9%	10%	8%	0%
3	32%	38%	23%	42%
4	41%	43%	38%	47%
5	12%	5%	23%	5%
6	6%	5%	8%	5%

Appendix B

Percentage of total replications in each condition selecting a particular number of classes

Table B1

Full class enumeration results for K = 4 data generation model with high separation for CPGMM and HGMM

	CPGMM											
	N=100				N=300				N=500			
	2	3	4	5	2	3	4	5	2	3	4	5
BIC	14%	57%	28%	1%	0%	5%	90%	5%	0%	0%	92%	8%
SABIC	3%	21%	43%	33%	0%	2%	43%	55%	0%	0%	28%	71%
DBIC	4%	30%	52%	14%	0%	2%	77%	21%	0%	0%	63%	37%
ICL-BIC	25%	50%	24%	1%	2%	36%	61%	0%	0%	20%	80%	0%
ICL-SABIC	8%	29%	46%	17%	1%	19%	75%	5%	0%	10%	90%	1%
ICL-DBIC	14%	38%	41%	7%	1%	26%	71%	2%	0%	13%	87%	1%
CLC	4%	25%	45%	26%	0%	9%	77%	13%	0%	5%	91%	4%
HT-AIC	4%	40%	50%	5%	0%	2%	41%	57%	0%	0%	19%	80%
HQ-AIC	5%	35%	50%	10%	0%	2%	70%	28%	0%	0%	42%	57%
	HGMM											
	N=100				N=300				N=500			
	2	3	4	5	2	3	4	5	2	3	4	5
BIC	88%	10%	1%	0%	61%	26%	11%	1%	31%	26%	35%	7%
SABIC	46%	19%	22%	13%	30%	13%	30%	27%	20%	8%	38%	34%
DBIC	65%	21%	11%	2%	38%	22%	29%	11%	23%	16%	43%	18%
ICL-BIC	97%	3%	0%	0%	97%	3%	0%	0%	97%	3%	0%	0%
ICL-SABIC	87%	10%	2%	0%	92%	8%	0%	0%	94%	6%	0%	0%
ICL-DBIC	93%	6%	1%	0%	95%	5%	0%	0%	95%	5%	0%	0%
CLC	78%	13%	5%	3%	88%	12%	1%	0%	90%	10%	0%	0%
HT-AIC	66%	23%	10%	1%	29%	12%	30%	29%	20%	5%	33%	41%
HQ-AIC	69%	20%	9%	2%	34%	21%	30%	16%	21%	10%	42%	27%

Table B2

Full class enumeration results for K = 4 data generation model with high separation for PGMM and LCGA

	PGMM											
	N=100				N=300				N=500			
	2	3	4	5	2	3	4	5	2	3	4	5
BIC	95%	5%	0%	0%	85%	10%	5%	0%	68%	13%	20%	0%
SABIC	62%	31%	6%	1%	64%	22%	14%	0%	57%	22%	21%	0%
DBIC	80%	16%	3%	0%	72%	16%	13%	0%	62%	17%	21%	0%
ICL-BIC	97%	3%	0%	0%	98%	2%	0%	0%	95%	4%	1%	0%
ICL-SABIC	84%	14%	2%	0%	90%	8%	2%	0%	87%	8%	5%	0%
ICL-DBIC	90%	9%	1%	0%	94%	5%	1%	0%	90%	7%	3%	0%
CLC	73%	21%	5%	1%	79%	12%	9%	0%	77%	11%	12%	0%
HT-AIC	82%	16%	2%	0%	63%	23%	14%	0%	55%	24%	21%	0%
HQ-AIC	84%	14%	2%	0%	70%	17%	13%	0%	60%	19%	21%	0%
	LCGA											
	N=100				N=300				N=500			
	2	3	4	5	2	3	4	5	2	3	4	5
BIC	0%	1%	21%	78%	0%	0%	5%	95%	0%	0%	2%	98%
SABIC	0%	1%	16%	82%	0%	0%	5%	95%	0%	0%	2%	98%
DBIC	0%	1%	17%	82%	0%	0%	5%	95%	0%	0%	2%	98%
ICL-BIC	0%	2%	32%	66%	0%	0%	11%	89%	0%	0%	2%	97%
ICL-SABIC	0%	1%	21%	78%	0%	0%	7%	93%	0%	0%	2%	98%
ICL-DBIC	0%	1%	25%	74%	0%	0%	8%	92%	0%	0%	2%	98%
CLC	0%	1%	18%	81%	0%	0%	6%	94%	0%	0%	2%	98%
HT-AIC	0%	1%	18%	81%	0%	0%	5%	95%	0%	0%	2%	98%
HQ-AIC	0%	1%	17%	82%	0%	0%	5%	95%	0%	0%	2%	98%

Table B3

Full class enumeration results for K = 4 data generation model with low separation for CPGMM and HGMM

	CPGMM											
	N=100				N=300				N=500			
	2	3	4	5	2	3	4	5	2	3	4	5
BIC	87%	13%	0%	0%	44%	54%	2%	0%	6%	77%	18%	0%
SABIC	1%	29%	35%	35%	0%	24%	48%	28%	0%	9%	60%	31%
DBIC	33%	50%	12%	4%	7%	63%	27%	3%	0%	39%	57%	4%
ICL-BIC	85%	14%	1%	0%	91%	8%	1%	0%	95%	5%	0%	0%
ICL-SABIC	37%	34%	17%	13%	84%	12%	2%	2%	92%	8%	0%	0%
ICL-DBIC	62%	27%	8%	3%	89%	9%	1%	0%	93%	7%	0%	0%
CLC	16%	31%	25%	28%	71%	19%	6%	4%	84%	13%	2%	1%
HT-AIC	37%	56%	7%	0%	0%	17%	52%	31%	0%	6%	42%	52%
HQ-AIC	46%	45%	7%	2%	4%	56%	35%	5%	0%	18%	68%	14%
	HGMM											
	N=100				N=300				N=500			
	2	3	4	5	2	3	4	5	2	3	4	5
BIC	87%	12%	1%	0%	62%	29%	8%	1%	35%	43%	17%	5%
SABIC	19%	33%	30%	18%	5%	20%	36%	39%	3%	10%	33%	54%
DBIC	50%	34%	13%	3%	23%	35%	30%	12%	9%	24%	38%	29%
ICL-BIC	91%	8%	1%	0%	96%	4%	0%	0%	98%	2%	0%	0%
ICL-SABIC	58%	23%	14%	5%	88%	8%	2%	2%	96%	3%	1%	0%
ICL-DBIC	79%	15%	5%	1%	92%	7%	1%	1%	96%	3%	0%	0%
CLC	38%	27%	25%	10%	69%	14%	11%	6%	89%	7%	2%	2%
HT-AIC	51%	36%	11%	2%	4%	15%	37%	44%	3%	7%	26%	64%
HQ-AIC	58%	31%	9%	2%	16%	32%	34%	18%	5%	14%	38%	43%

Table B4

Full class enumeration results for K = 4 data generation model with low separation for PGMM and LCGA

	PGMM											
	N=100				N=300				N=500			
	2	3	4	5	2	3	4	5	2	3	4	5
BIC	98%	2%	0%	0%	94%	6%	0%	0%	81%	19%	0%	0%
SABIC	21%	47%	20%	12%	26%	57%	12%	5%	17%	74%	7%	2%
DBIC	74%	23%	3%	1%	65%	34%	1%	0%	45%	55%	1%	0%
ICL-BIC	96%	3%	0%	0%	97%	3%	0%	0%	98%	2%	0%	0%
ICL-SABIC	73%	18%	6%	4%	89%	10%	1%	0%	96%	4%	0%	0%
ICL-DBIC	88%	9%	2%	1%	93%	7%	0%	0%	97%	3%	0%	0%
CLC	49%	26%	10%	15%	81%	13%	4%	2%	91%	7%	2%	1%
HT-AIC	76%	22%	1%	0%	19%	58%	17%	7%	5%	59%	25%	11%
HQ-AIC	83%	16%	1%	0%	57%	40%	2%	0%	29%	69%	3%	0%
	LCGA											
	N=100				N=300				N=500			
	2	3	4	5	2	3	4	5	2	3	4	5
BIC	0%	1%	29%	70%	0%	0%	1%	99%	0%	0%	0%	100%
SABIC	0%	1%	12%	87%	0%	0%	1%	99%	0%	0%	0%	100%
DBIC	0%	1%	13%	86%	0%	0%	1%	99%	0%	0%	0%	100%
ICL-BIC	0%	6%	50%	44%	0%	0%	22%	78%	0%	0%	17%	83%
ICL-SABIC	0%	1%	18%	80%	0%	0%	10%	90%	0%	0%	6%	94%
ICL-DBIC	0%	2%	29%	69%	0%	0%	14%	86%	0%	0%	11%	89%
CLC	0%	1%	14%	85%	0%	0%	5%	99%	0%	0%	3%	97%
HT-AIC	0%	1%	15%	84%	0%	0%	1%	99%	0%	0%	0%	100%
HQ-AIC	0%	1%	14%	85%	0%	0%	1%	99%	0%	0%	0%	100%

Table B5

Full class enumeration results for K = 3 data generation model with high separation for CPGMM and HGMM

	CPGMM								
	N=100			N=300			N=500		
	2	3	4	2	3	4	2	3	4
BIC	33%	61%	5%	0%	95%	4%	0%	88%	12%
SABIC	9%	43%	48%	0%	35%	64%	0%	25%	75%
DBIC	12%	68%	20%	0%	75%	25%	0%	54%	46%
ICL-BIC	58%	40%	2%	18%	81%	0%	8%	92%	0%
ICL-SABIC	23%	60%	17%	7%	89%	4%	3%	96%	1%
ICL-DBIC	35%	57%	8%	10%	88%	1%	4%	95%	1%
CLC	14%	58%	28%	3%	87%	10%	1%	96%	4%
HT-AIC	14%	74%	12%	0%	30%	70%	0%	15%	85%
HQ-AIC	15%	71%	14%	0%	65%	34%	0%	38%	62%
	HGMM								
	N=100			N=300			N=500		
	2	3	4	2	3	4	2	3	4
BIC	96%	4%	0%	92%	7%	1%	87%	10%	2%
SABIC	55%	26%	20%	66%	8%	26%	55%	14%	31%
DBIC	77%	17%	5%	78%	13%	10%	67%	15%	18%
ICL-BIC	99%	1%	0%	100%	0%	0%	100%	0%	0%
ICL-SABIC	99%	1%	0%	100%	0%	0%	100%	0%	0%
ICL-DBIC	99%	1%	0%	100%	0%	0%	100%	0%	0%
CLC	91%	6%	3%	100%	0%	0%	100%	0%	0%
HT-AIC	77%	19%	4%	65%	6%	29%	52%	11%	37%
HQ-AIC	84%	13%	3%	72%	13%	15%	59%	15%	26%

Table B6

Full class enumeration results for K = 3 data generation model with high separation for PGMM and LCGA

	PGMM								
	N=100			N=300			N=500		
	2	3	4	2	3	4	2	3	4
BIC	89%	11%	0%	57%	42%	1%	38%	59%	3%
SABIC	46%	48%	5%	43%	50%	7%	36%	57%	8%
DBIC	62%	37%	1%	46%	50%	5%	36%	59%	6%
ICL-BIC	97%	3%	0%	98%	2%	0%	99%	1%	0%
ICL-SABIC	75%	23%	2%	92%	8%	0%	94%	6%	0%
ICL-DBIC	85%	14%	0%	96%	4%	0%	96%	4%	0%
CLC	65%	33%	2%	83%	16%	1%	89%	10%	1%
HT-AIC	64%	35%	1%	43%	50%	7%	36%	53%	11%
HQ-AIC	67%	32%	1%	44%	50%	5%	36%	58%	7%
	LCGA								
	N=100			N=300			N=500		
	2	3	4	2	3	4	2	3	4
BIC	46%	28%	27%	0%	1%	99%	0%	0%	100%
SABIC	46%	28%	27%	0%	1%	99%	0%	0%	100%
DBIC	46%	28%	27%	0%	1%	99%	0%	0%	100%
ICL-BIC	46%	28%	27%	0%	1%	99%	0%	0%	100%
ICL-SABIC	46%	28%	27%	0%	1%	99%	0%	0%	100%
ICL-DBIC	46%	28%	27%	0%	1%	99%	0%	0%	100%
CLC	46%	28%	27%	0%	1%	99%	0%	0%	100%
HT-AIC	46%	28%	27%	0%	1%	99%	0%	0%	100%
HQ-AIC	46%	28%	27%	0%	1%	99%	0%	0%	100%

Table B7

Full class enumeration results for K = 3 data generation model with low separation for CPGMM and HGMM

	CPGMM								
	N=100			N=300			N=500		
	2	3	4	2	3	4	2	3	4
BIC	91%	9%	0%	64%	35%	1%	25%	72%	3%
SABIC	4%	42%	54%	2%	50%	48%	0%	34%	66%
DBIC	48%	42%	10%	18%	69%	13%	2%	71%	27%
ICL-BIC	92%	8%	1%	99%	1%	0%	100%	0%	0%
ICL-SABIC	58%	25%	18%	96%	3%	1%	99%	1%	0%
ICL-DBIC	77%	17%	6%	98%	1%	1%	100%	0%	0%
CLC	37%	29%	34%	90%	7%	3%	99%	1%	0%
HT-AIC	50%	45%	5%	1%	44%	54%	0%	23%	77%
HQ-AIC	58%	36%	6%	12%	68%	21%	1%	54%	46%
	HGMM								
	N=100			N=300			N=500		
	2	3	4	2	3	4	2	3	4
BIC	89%	10%	1%	58%	35%	7%	26%	48%	26%
SABIC	23%	35%	43%	6%	26%	68%	4%	15%	80%
DBIC	52%	36%	12%	17%	45%	39%	6%	30%	64%
ICL-BIC	93%	7%	1%	96%	4%	0%	99%	1%	0%
ICL-SABIC	65%	21%	14%	91%	8%	1%	98%	2%	0%
ICL-DBIC	79%	16%	5%	94%	5%	1%	99%	1%	0%
CLC	46%	24%	30%	80%	13%	7%	94%	4%	1%
HT-AIC	52%	38%	10%	6%	21%	73%	4%	13%	83%
HQ-AIC	60%	31%	9%	13%	41%	46%	4%	21%	75%

Table B8

Full class enumeration results for K = 3 data generation model with low separation for PGMM and LCGA

	PGMM								
	N=100			N=300			N=500		
	2	3	4	2	3	4	2	3	4
BIC	99%	1%	0%	100%	0%	0%	100%	0%	0%
SABIC	34%	46%	20%	70%	23%	7%	82%	16%	2%
DBIC	86%	12%	2%	95%	5%	0%	98%	2%	0%
ICL-BIC	99%	1%	0%	100%	0%	0%	100%	0%	0%
ICL-SABIC	81%	14%	4%	96%	4%	0%	97%	2%	0%
ICL-DBIC	94%	6%	0%	98%	2%	0%	99%	1%	0%
CLC	64%	23%	12%	87%	10%	3%	93%	5%	2%
HT-AIC	88%	11%	1%	58%	32%	10%	53%	35%	12%
HQ-AIC	92%	7%	1%	91%	8%	1%	95%	5%	1%
	LCGA								
	N=100			N=300			N=500		
	2	3	4	2	3	4	2	3	4
BIC	0%	3%	97%	0%	1%	99%	0%	0%	100%
SABIC	0%	3%	97%	0%	1%	99%	0%	0%	100%
DBIC	0%	3%	97%	0%	1%	99%	0%	0%	100%
ICL-BIC	0%	6%	94%	0%	1%	99%	0%	0%	100%
ICL-SABIC	0%	4%	96%	0%	1%	99%	0%	0%	100%
ICL-DBIC	0%	4%	96%	0%	1%	99%	0%	0%	100%
CLC	0%	3%	97%	0%	1%	99%	0%	0%	100%
HT-AIC	0%	3%	97%	0%	1%	99%	0%	0%	100%
HQ-AIC	0%	3%	97%	0%	1%	99%	0%	0%	100%