Impact of Likelihoods on Class Enumeration in Bayesian Growth Mixture Modeling ⋆

Xin Tong[1][0000−0003−3050−1554], Seohyun Kim[1][0000−0002−2119−6044], and Zijun Ke[2][0000−0002−3162−3871]

[1] Department of Psychology, University of Virginia, Charlottesville, VA 22903, USA
xt8b@virginia.edu
sseohyun@gmail.com
[2] Department of Psychology, Sun Yat-sen University, Guangzhou, China
keziyun@mail.sysu.edu.cn

Abstract. Bayesian methods have been widely used to estimate models with complex structures. To assess model fit and compare different models, researchers typically use model selection criteria such as Deviance Information Criteria (DIC), Watanabe-Akaike Information Criteria (WAIC) and leave-one-out cross validation (LOO-CV), the calculation of which is based on the likelihoods of the models. When models contain latent variables, the likelihood is often specified as conditional on the latent variables in popular Bayesian software (e.g., BUGS, JAGS, and Stan). Although this practice reduces computation work and does not affect model estimation, the previous literature has shown that model comparisons based on the conditional likelihood could be misleading. In contrast, marginal likelihoods can be obtained by integrating out the latent variables and be used to calculate model selection criteria. In this study, we evaluate the effect of using conditional likelihoods and marginal likelihoods in model selection for growth mixture models. Simulation results suggest that marginal likelihoods are much more reliable and should be generally used for growth mixture modeling.

Keywords: Marginal likelihood · conditional likelihood · DIC · WAIC · LOO-CV.

1 Introduction

Growth mixture modeling (GMM) is a method for identifying multiple unobserved subgroups in a population, describing longitudinal change within each subgroup, and examining differences in change among those subgroups (Ram & Grimm, 2009). GMM has been increasingly used in social and behavioral sciences (e.g., Frankfurt et al., 2016; McDermott et al., 2018; Smith & Ehlers, 2020;

Ren et al., 2021) to flexibly model growth trajectories with substantial individual variations. Despite the popularity of GMM, several issues are involved in its model estimation (Bauer & Curran, 2003; Hipp & Bauer, 2006), including violation of distributional assumptions, obtaining local solutions, nonconvergence, etc. Researchers have made efforts in addressing these issues, many of which are in the Bayesian framework (e.g., Depaoli, 2013; Kim et al., 2021a; Lu et al., 2011). Bayesian approaches are relatively flexible in accounting for the nonnormality in data and enable incorporating prior information into model estimation to help yield converged results when there are not enough samples or latent classes are not well separated (Depaoli, 2014; Kim et al., 2021b). In addition, data augmentation and Markov chain Monte Carlo (MCMC) techniques can be naturally applied in the Bayesian framework to reduce the mathematical demands for complex model estimation.

Deciding the appropriate number of latent classes (i.e., unobserved subgroups) is critical in GMM and is typically achieved by comparing models with different number of latent classes and selecting the best fitting model. In Bayesian statistics, model comparison and selection can be performed using the Bayes factor which is the ratio of the posterior odds to the prior odds of two competing models. Since the calculation of the Bayes factor is often difficult and greatly influenced by the priors, model comparison in GMM is typically conducted using information criteria and cross validation which estimate out-of-sample predictive accuracy using within-sample fits. The calculation of the model selection criteria is based on the likelihoods of the models.

In popular Bayesian software (e.g., BUGS, JAGS, and Stan), the likelihood of GMM is often specified as conditional on the latent variables. However, recent studies (e.g., Kim et al., 2021a; Merkle et al., 2019) reported that model selection and comparison based on the conditional likelihood in latent variable modeling can be misleading. Instead, marginal likelihoods were used where latent variables were integrated out in the likelihood functions. Although conditional and marginal likelihoods do not make differences in terms of model estimation after all Markov chains converge, the distinction between them in model selection is substantial but is often overlooked. As far as we are aware, only Merkle et al. (2019) has particularly studied the difference between conditional and marginal likelihoods and recommended use of marginal likelihood based information criteria in Bayesian latent variable analysis.

Due to the complexity of GMM and unique challenges associated with it, in this study, we will evaluate the performance of conditional and marginal likelihood in GMM class enumeration. We focus on two information criteria: Deviance Information Criterion (DIC; Spiegelhalter et al., 2002) and Widely Applicable Information Criterion (Watanabe-Akaike Information Criterion, WAIC; Watanabe, 2010), and one cross validation approach: leave-one-out cross validation (LOO-CV; Gelman et al., 2013; Vehtari et al., 2017). Their performance based on different likelihoods in GMM class enumeration will be investigated. The paper is organized as follows. We first briefly review growth mixture models, introduce the associated conditional and marginal likelihoods, and different model

selection criteria. Then we use a simulation study to assess the impact of conditional and marginal likelihoods on GMM model selection. Recommendations are provided at the end of the article.

2   Bayesian GMM Model Selection

2.1   A brief review of growth mixture models

Growth mixture models extend growth curve models by assuming that a population consists of a number of latent classes (i.e., unobserved subgroups) and each latent class is characterized by a unique growth trajectory. Suppose that a population consisted of G latent classes that have distinct patterns of change. Let $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{iT_i})'$ denote a vector of T_i repeated observations for individual i ($i \in \{1, \ldots, N\}$). A general form of growth mixture models can be expressed as

$$\boldsymbol{y}_i | (z_i = g) = \boldsymbol{\Lambda}_i \boldsymbol{b}_{ig} + \boldsymbol{\epsilon}_i, \tag{1}$$

where the subscript g indicates that a corresponding parameter or variable is class-specific. In this model, z_i represents a class indicator for individual i with mixing proportion for class g being $P(z_i = g) = \pi_g$, $\boldsymbol{\Lambda}_i$ is a $T_i \times q$ matrix of factor loadings that determines the shape of the growth trajectories, $\boldsymbol{b}_{ig}$ is a $q \times 1$ vector of latent factors for class g ($g \in \{1, \ldots, G\}$), and $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \ldots, \epsilon_{iT_i})'$ is a $T_i \times 1$ vector of intraindividual measurement errors. The latent factors are often assumed to follow multivariate normal distributions such that $\boldsymbol{b}_{ig} \sim MN(\boldsymbol{\beta}_g, \boldsymbol{\Psi}_g)$, where $\boldsymbol{\beta}_g$ is the mean of $\boldsymbol{b}_{ig}$ and $\boldsymbol{\Psi}_g$ is the covariance matrix of $\boldsymbol{b}_{ig}$. The measurement errors are also assumed to be normally distributed, $\boldsymbol{\epsilon}_i \sim MN(\boldsymbol{0}, \boldsymbol{\Sigma}_g)$, leading the conditional mean of $\boldsymbol{y}_i$ given $\boldsymbol{b}_{ig}$ to be $E(\boldsymbol{y}_i | \boldsymbol{b}_{ig}) = \boldsymbol{\Lambda}_i \boldsymbol{b}_{ig}$. In practice, it is common to further assume that the intraindividual measurement errors have equal variances and are independent across time, so that $\boldsymbol{\Sigma}_g = \sigma_g^2 \boldsymbol{I}$, where σ_g^2 is a scale parameter for class g. We assumed this measurement error structure for the rest of this study.

2.2   Conditional and marginal likelihoods of growth mixture models

With the normality assumption, the likelihood function of the model in Equation (1) can be specified. As stated previous, popular Bayesian software often specify the likelihood as conditional on the latent variables. That is,

$$L_C(\boldsymbol{b}_{ig}, z_i, \sigma_g^2 | \mathbf{y}) = p(\mathbf{y} | \boldsymbol{b}_{ig}, z_i = g, \sigma_g^2), \tag{2}$$

where $p(\mathbf{y}_i | \boldsymbol{b}_{ig}, z_i = g, \sigma_g^2)$ is the density function of the multivariate normal distribution $MN(\boldsymbol{\Lambda}_i \boldsymbol{b}_{ig}, \sigma_g^2 \boldsymbol{I})$.

To obtain the marginal likelihood of Model (1), the latent variables $\boldsymbol{b}_{ig}$ and z_i have to be integrated out of the conditional likelihood. The marginal likelihood for the normal-distribution-based GMM has a closed form:

$$L_M(\boldsymbol{\beta}_g, \boldsymbol{\Psi}_g, \pi_g, \sigma_g^2 | \mathbf{y}) = p(\mathbf{y} | \boldsymbol{\beta}_g, \boldsymbol{\Psi}_g, \pi_g, \sigma_g^2) = \prod_{i=1}^{N} \sum_{g=1}^{G} \pi_g p(\mathbf{y}_i | \boldsymbol{\beta}_g, \boldsymbol{\Psi}_g, \sigma_g^2), \tag{3}$$

where $p(\mathbf{y}_i|\boldsymbol{\beta}_g, \boldsymbol{\Psi}_g, \sigma_g^2)$ is the density function of the multivariate normal distribution $MN(\boldsymbol{\Lambda}_i\boldsymbol{\beta}_g, \boldsymbol{\Lambda}_i\boldsymbol{\Psi}_g\boldsymbol{\Lambda}_i' + \sigma_g^2\boldsymbol{I})$.

Given the likelihood functions, model selection criteria can be computed. Based on the model selection criteria, we can compare GMMs with different number of latent classes and select the best fitting model.

2.3   Model comparison criteria

In this paper, we use DIC, WAIC, and LOO-CV to select the optimal number of latent classes for GMM. We briefly introduce the three model comparison criteria below.

DIC was proposed by Spiegelhalter et al. (2002). Although it has received much criticism (e.g., Celeux et al., 2006), it is widely used in Bayesian model selection. DIC is defined as the sum of the expected deviance over the parameter space and the effective number of model parameters,

$$DIC = \bar{D} + p_D.$$

The expected deviance is

$$\bar{D} = E_{\boldsymbol{\Theta}}[-2log\ p(\mathbf{y}|\boldsymbol{\Theta})|\mathbf{y}] + C,$$

where $\boldsymbol{\Theta}$ is a set of model parameters, and C is a constant that can be canceled out when comparing models. $\bar{D}$ is calculated as the posterior mean of the deviance. The effective number of parameters, p_D, measures the complexity of the model and is defined as

$$p_D = \overline{D} - \hat{D},$$

where $\hat{D}$ is the deviance calculated at the posterior mean of $\boldsymbol{\Theta}$. Models with smaller DICs are preferred.

WAIC was proposed more recently and have been shown to have advantages over DIC (Vehtari et al., 2017). WAIC uses the entire posterior distribution, is asymptotically equal to Bayesian cross validation, is invariant to parameterization, and works for singular models. We used the following definition of WAIC (Gelman et al., 2013).

$$WAIC = -2\sum_{i=1}^{N}\log\left(\frac{1}{S}\sum_{s=1}^{S}p(y_i|\boldsymbol{\Theta}^{(s)})\right) + 2\sum_{i=1}^{N}Var_{s=1}^{S}log\ p(y_i|\boldsymbol{\Theta}^{(s)})),$$

where S is the number of MCMC iterations, $\boldsymbol{\Theta}^{(s)}$ is a draw from the posterior distribution at the sth iteration, and $Var_{s=1}^{S}$ represents the sample variance, $Var_{s=1}^{S}a_s = \frac{1}{S-1}\sum_{s=1}^{S}(a_s - \bar{a})^2$. Models with smaller WAICs are preferred.

LOO-CV evaluates the model fit based on an estimate of the log predictive density of the hold-out data. Each data point is taken out at a time to cross

validate the model that is fitted based on the remaining data. LOO-CV is defined as

$$LOO = -2 \sum_{i=1}^{N} \log \int p(y_i|\Theta)p(\Theta|y_{-i})d\Theta,$$

and in practice, it can be approximately calculated as

$$\widehat{LOO} = -2 \sum_{i=1}^{N} \log \frac{1}{\frac{1}{S}\sum_{s=1}^{S} \frac{1}{p(y_i|\boldsymbol{\theta}^{(s)})}}.$$

Vehtari et al. (2017) showed that although WAIC is asymptotically equal to LOO-CV, LOO-CV is more robust in the finite case with weak priors or influential observations.

3   A Simulation Study

We now present a simulation study to evaluate the impact of conditional and marginal likelihoods based model selection criteria on GMM class enumeration.

We generated data from a two-class linear growth mixture model with 4 equally spaced measurement occasions. Namely, in Equation (1), $G = 2$, $T_i = 4$, $\boldsymbol{\Lambda}_i = ((1,1,1,1)', (0,1,2,3)')$, the latent intercept and slope for class 1, $\boldsymbol{b}_{i1} \sim MN(\boldsymbol{\beta}_1, \boldsymbol{\Psi}_1)$ and the latent intercept and slope for class 2, $\boldsymbol{b}_{i2} \sim MN(\boldsymbol{\beta}_2, \boldsymbol{\Psi}_2)$. The covariance matrix of the latent intercepts and slopes were set to be $\boldsymbol{\Psi}_g = \boldsymbol{\Psi} = \begin{pmatrix} 0.25 & 0 \\ 0 & 0.04 \end{pmatrix}$ for $g = 1$ and 2, and the intraindividual measurement error variance was set at $\sigma^2 = 0.2$. These variance and covariance parameters were assumed to be the same across the two latent classes. We manipulated three factors that could potentially influence the performance of GMM in the simulation study: sample size, class separation, and class proportions. Two different sample sizes were considered ($N = 300$ or 500). Class separation was characterized using Mahalanobis distance, which can be calculated as $MD = \sqrt{(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)'\boldsymbol{\Psi}^{-1}(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)}$, where $\boldsymbol{\beta}_1$ represented the means of latent intercepts and latent slopes for the first latent class, and $\boldsymbol{\beta}_2$ represented the means of latent intercepts and latent slopes for the second latent class. We evaluated the influence of a high class separation and a relatively low class separation. For the high separation, the first class had an average latent intercept of 2 and an average slope of 0.5, $\boldsymbol{\beta}_1 = (2, 0.5)'$, so in general, the scores were increasing over time. The second class had an average latent intercept of 1 and an average slope of 0, $\boldsymbol{\beta}_2 = (1, 0)'$, indicating that the overall trajectory was a flat line. This setting yielded a Mahalanobis distance $MD = 3.2$. For the low class separation, $\boldsymbol{\beta}_1 = (1.5, 0.5)'$ and $\boldsymbol{\beta}_2 = (1, 0)'$, which had $MD = 2.7$. The class proportions were set to be either unbalanced (25% from the first class and 75% from the second class) or balanced (50% from both latent classes).

For each simulation condition, 200 datasets were generated. For each dataset, we fit growth mixture models with one class ($G = 1$), two classes ($G = 2$), and three classes ($G = 3$). Bayesian estimation of GMM was conducted using JAGS with the *rjags* R package (Plummer, 2017). JAGS is a Bayesian data analysis program that uses MCMC algorithms (e.g., Gibbs sampler) for generating samples from the posterior distribution of model parameters. In JAGS, we obtained posterior samples of the model parameters by augmenting the latent variables ($\boldsymbol{b}_{ig}$ and z_i). With the sampled parameters and the likelihood of the model, DIC, WAIC, and LOO-CV can be calculated. Since the likelihood can be calculated in Equation (2) or Equation (3) when a model contains latent variables, DIC, WAIC, and LOO-CV were calculated based on the conditional likelihood and the marginal likelihood, separately. We then assessed the performance of the model comparison criteria based on different likelihoods in class enumeration.

The following priors were used for model inferences as these priors had little information about the parameters: $p(\boldsymbol{\beta}_g) = MN(0, 10^3 \times \boldsymbol{I})$ for $g = 1, ..., G$, $p(\boldsymbol{\Psi}) = InvWishart(2, \boldsymbol{I}_2)$, $p(\sigma^2) = InvGamma(.01, .01)$, and $p(\boldsymbol{\pi}) \sim Dirichlet(10\boldsymbol{j}_G)$, where G is the total number of latent classes, and $\boldsymbol{j}_G$ is a $G \times 1$ vector that has 1 for all components for $G > 1$. The number of MCMC iterations was set to $10,000$, and the first half of the iterations were discarded for burn-in. Although our pilot study showed that the $10,000$ iterations were enough for the chains to converge, to guarantee convergence, we also allowed up to 10 different starting values for each model estimation to obtain converged results.
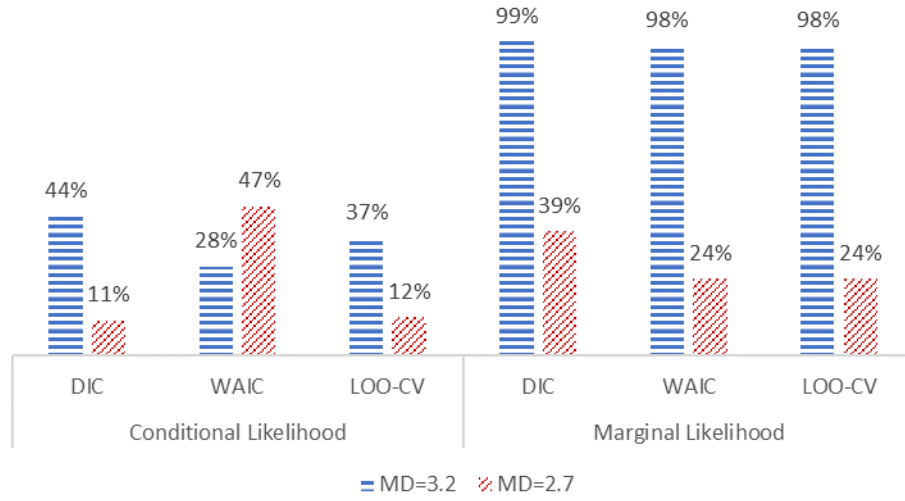
3.1   Results

Figures 1(a)-(b) summarize the model selection results based on DIC, WAIC, and LOO-CV when class proportions are 25% and 75%. For balanced classes, the relative performance of the model comparison criteria based on conditional and marginal likelihoods has the same pattern and thus is not repeated in this section. Figures 1(a) and 1(b) report the results for $N = 300$ and $N = 500$, respectively. For each figure, the vertical axis (i.e., height of the bars) represents the probability of the correct model (2-class growth mixture model) being selected.

From Figure 1, it can be seen that, in general, model selection based on DIC, WAIC, and LOO-CV is more likely to be correct when sample size is larger and the class separation is higher. When the class separation is relatively low (e.g., $MD = 2.7$), increasing sample size raises the chance to select the correct number of latent classes in GMM. For example, the marginal likelihood based DIC has 39% of the chances to select the correct model when $N = 300$. This percentage increases to 87% when the sample size is 500.

In addition, model selection criteria based on the marginal likelihood are more reliable as the bars on the right panel of Figures 1(a)-(b) (marginal likelihood based criteria) are generally taller than the bars on the left panel of the figures (conditional likelihood based criteria). Although conditional likelihood based model selection criteria may perform well under some conditions (e.g., when $N = 500$, DIC calculated with the conditional likelihood has 87% of the chance to

(a) $N = 300$

(b) $N = 500$
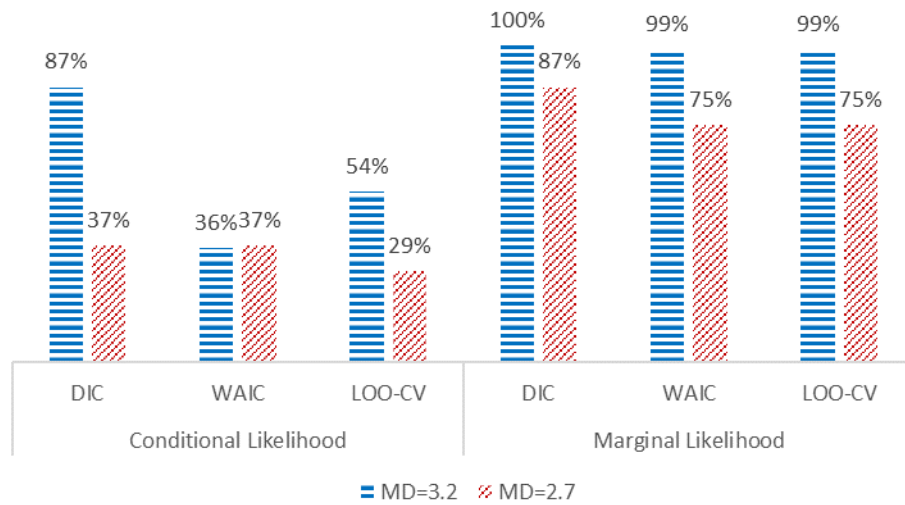
Fig. 1. Model selection results based on DIC, WAIC, and LOO-CV when class proportions are 25% and 75%.

select the correct model), they are unstable in general. In practice, it is difficult to tell whether the conditional likelihood based model selection criteria are reliable or not for the study setting. Therefore, the conditional likelihood should not be used to calculate model selection criteria for model comparison. We have further investigated the probability of each model being selected for different data conditions. As demonstrated in Figure 2, even when the class separation is relatively low, DIC, WAIC, and LOO-CV calculated based on the marginal likelihood almost always select the correct 2-class model. In contrast, the model selection criteria calculated based on the conditional likelihood tend to prefer simpler models (i.e., 1-class model) under this condition.
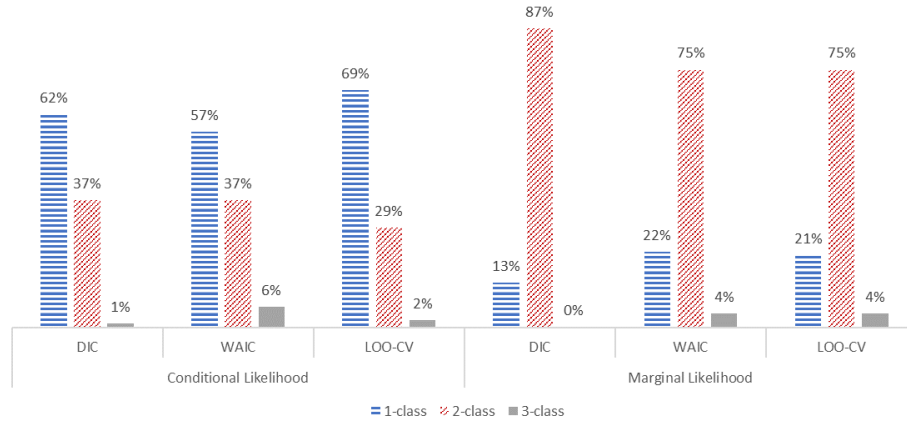
Fig. 2. The comparison between the conditional likelihood and marginal likelihood in selecting different growth mixture models when $N = 500$ and $MD = 2.7$. The height of the bar represents the percentage that the corresponding model is selected.

Moreover, when the marginal likelihood is used, DIC, WAIC, and LOO-CV provide similar values. WAIC and LOO-CV, in particular, provide almost identical values. However, when the conditional likelihood is used, DIC, WAIC, and LOO-CV are not as similar as those we get from the marginal likelihood. In addition, WAIC and LOO-CV seem to perform slightly worse than DIC, especially when the class separation is relatively low.

4   Discussion

Bayesian methods have been increasingly used for GMM model estimation because of their flexibility and capability to handle model with complex structures. An important task of GMM is to determine the number of latent classes, and is typically conducted by model comparisons. Commonly used Bayesian model

comparison criteria are calculated based on the likelihood of the model. In this paper, we evaluated the impact of the conditional likelihood and the marginal likelihood on the performance of different model comparison criteria using a simulation study. We would like to note that the simulation results showed a very salient pattern and were robust against the simulation settings. Our study echoed the previous literature and emphasized the use of marginal likelihood for the calculation of Bayesian model selection criteria when models contain latent variables.

We want to point out that when data are normally distributed, the marginal likelihood is recommended to use and DIC, WAIC, and LOO-CV calculated based on the marginal likelihood can almost guarantee the correct class enumeration. However, the performance of the model selection criteria based on different likelihoods were not systematically evaluated when data are contaminated by outliers. Previous research (e.g., Kim et al., 2021a) suggested the application of robust methods for dealing with the nonnormality in class enumeration in GMM. We expect that combining robust methods with marginal likelihood based model selection criteria may improve the model selection accuracy. Future research needs to be conducted towards this direction.

We also would like to note that in our study, the normality assumption is applied to the growth mixture model, with which a close form of the marginal likelihood is available. When a close form of the marginal likelihood cannot be obtained (e.g., a robust model using Student's t distributions), we need to numerically integrate the conditional likelihood with respect to the latent variables. Since numerical integration takes time, the entire class enumeration procedure may be slowed down. It is worth investigating ways to solve numerical integrations faster.

Bibliography

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. doi:10.1037/1082-989X.8.3.338.

Celeux, G., Forbes, F., Robert, C. P., & Titterington, D. M. (2006). Deviance information criteria for missing data models. *Bayesian Analysis*, 1, 651–673. DOI: 10.1214/06-ba122.

Depaoli, S. (2013). Mixture class recovery in GMM under varying degrees of class separation: Frequentist versus Bayesian estimation. *Psychological Methods*, 18, 186–219. DOI: 10.1037/a0031609.

Depaoli, S. (2014). The impact of inaccurate âinformativeâ priors for growth parameters in Bayesian growth mixture modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, 21(2), 239–252. DOI: 10.1080/10705511.2014.882686.

Frankfurt, S., Frazier, P., Syed, M., & Jung, K. R. (2016). Using group-based trajectory and growth mixture modeling to identify classes of change trajectories. *The Counseling Psychologist*, 44(5), 622–660. DOI:10.1177/0011000016658097.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis*. Boca Raton, FL: Chapman and Hall/CRC.

Hipp, J. R. & Bauer, D. J. (2006). Local solutions in the estimation of growth mixture models. *Psychological methods*, 11(1), 36–53. doi:10.1037/1082-989X.11.1.36.

Kim, S., Tong, X., & Ke, Z. (2021a). Exploring class enumeration in Bayesian growth mixture modeling based on conditional medians. *Frontiers in Education*. A special research topic on Advances in Mixture Modeling. DOI: 10.3389/feduc.2021.624149.

Kim, S., Tong, X., Zhou, J., & Boichuk, J. P. (2021b). Conditional median based Bayesian growth mixture modeling for nonnormal data. *Behavior Research Methods*. DOI: 10.3758/s13428-021-01655-w.

Lu, Z., Zhang, Z., & Lubke, G. (2011). Bayesian inference for growth mixture models with latent class dependent missing data. *Multivariate Behavioral Research*, 46, 567–597.

McDermott, P. A., Rovine, M. J., Reyes, R. S., Chao, J. L., Scruggs, R., Buek, K., & Fantuzzo, J. W. (2018). Trajectories of early education learning behaviors among children at risk: A growth mixture modeling approach. *Psychology in the Schools*, 55(10), 1205–1223. DOI:10.1002/pits.22145.

Merkle, E. C., Furr, D., & Rabe-Hesketh, S. (2019). Bayesian comparison of latent variable models: conditional versus marginal likelihoods. *Psychometrika*, 84(3), 802–829. DOI: 10.1007/s11336-019-09679-0.

Plummer, M. (2017). Jags version 4.3. 0 user manual.

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. DOI: 10.1177/0165025409343765.

Ren, L., Tong, X., Xu, W., Wu, Z., Zhou, X., & Hu, B. Y. (2021). Distinct patterns of organized activity participation and their associations with school readiness among Chinese preschoolers. *Journal of School Psychology.* DOI: 10.1016/j.jsp.2021.03.007.

Smith, K. V. & Ehlers, A. (2020). Cognitive predictors of grief trajectories in the first months of loss: A latent growth mixture model. *Journal of consulting and clinical psychology*, 88(2), 93–105. DOI:10.1037/ccp0000438.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Linde, A. v. d. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4), 583–639.

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Journal of Statistics and Computing.*

Watanabe, S. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11, 3571–3594.