MODEL SELECTION IN MIXTURE MODELING

A Dissertation

Presented to

The Faculty of the Graduate School

At the University of Missouri

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

---

by

EMILIE SHIREMAN

Dr. Douglas Steinley, Dissertation Supervisor

May 2016

The undersigned, appointed by the dean of the Graduate School, have examined the dissertation entitled

MODEL SELECTION IN MIXTURE MODELING

presented by Emilie Shireman,
a candidate for the degree of doctor of philosophy of quantitative psychology,
and hereby certify that, in their opinion, it is worthy of acceptance.

Professor Douglas Steinley

Professor Phillip K. Wood

Professor Kenneth J. Sher

Professor Lori Thombs

DEDICATION

I dedicate this dissertation to my husband, Robert Shireman, my parents, Barbara McCulloch and Michael Rausch, and Isabelle and Toki Ann, who provided endless love and support that made this dissertation possible.

I also dedicate this dissertation to the many people who have served as a mentor for me at some point during my education. I am forever grateful to these people, who selflessly took time out of their lives to help me along my way, giving near limitless advice and going to bat for me countless times. Without these people, none of this is was possible.

ACKNOWLEDGEMENTS

I would like to thank Douglas Steinley, Kenneth Sher, Phillip Wood, and Lori Thombs for serving on my dissertation committee. I am also grateful to Michaela Hoffman for detailed comments on sections of this dissertation.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

Tables

Figures

Introduction

Mixture Modeling

Mixture modeling (also called latent profile analysis or model-based cluster analysis) is a form of density estimation and classification that finds a number of specified subgroups in data (also referred to as clusters or classes). These groups are modeled as independent components:

$$f(\boldsymbol{x};\ \boldsymbol{\Psi}) = \sum_{k=1}^{K} \pi_k f_k(\boldsymbol{x};\ \psi_k)$$

where f_k is the distribution of class k, $\mathbf{x}$ is an $n \times p$ data matrix, $\boldsymbol{\Psi}$ contains the parameters of the entire mixture model (i.e., $\boldsymbol{\Psi} = \{\pi_1, \dots, \pi_K;\ \psi_1, \dots, \psi_k\}$), π_k is the population proportion of class k, and ψ_k contains the parameter estimates of the k^{th} cluster ($k = \{1, \dots, K\}$). In the social sciences, these classes are assumed to each have a Gaussian distribution (i.e., $f_k = f$ and $\psi_k = \{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}$) or:

$$f(\boldsymbol{x};\ \boldsymbol{\Psi}) = \sum_{k=1}^{K} \pi_k f(\boldsymbol{x};\ \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

where f is the Gaussian density, $\boldsymbol{\mu}_k$ is the $1 \times p$ mean vector, and $\boldsymbol{\Sigma}_k$ is the $p \times p$ covariance matrix for the k^{th} cluster.

Expectation-Maximization (EM) Algorithm)

Typically, the parameters of the mixture model are estimated via maximum likelihood using the Expectation-Maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977). After a set of initial parameter estimates are found ($\widehat{\boldsymbol{\Psi}}^{(0)} =$

$\{\widehat{\boldsymbol{\pi}}^{(0)}, \widehat{\boldsymbol{\mu}}^{(0)}, \widehat{\boldsymbol{\Sigma}}^{(0)}\}$), the posterior probability of class assignment is calculated for each individual *i* and each cluster k ($i = 1, \dots, n; k = 1, \dots, K$):

$$\tau_{i,k}{}^{(0)} = \frac{\hat{\pi}_k^{(0)} f(\boldsymbol{x};\ \widehat{\boldsymbol{\mu}}_k^{(0)}, \widehat{\boldsymbol{\Sigma}}_k^{(0)})}{\sum_{l=1}^{K} \hat{\pi}_l^{(0)} f(\boldsymbol{x};\ \widehat{\boldsymbol{\mu}}_l^{(0)}, \widehat{\boldsymbol{\Sigma}}_l^{(0)})}$$

$$\widehat{\boldsymbol{\mu}}_k^{(1)} = \frac{\sum_{i=1}^{n} \tau_{i,k}{}^{(0)} \boldsymbol{x}_i}{\sum_{i=1}^{n} \tau_{i,k}{}^{(0)}}$$

$$\widehat{\boldsymbol{\Sigma}}_k^{(1)} = \frac{\sum_{i=1}^{n} \tau_{i,k}{}^{(0)} (\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}_k^{(0)})(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}_k^{(0)})^T}{\sum_{i=1}^{n} \tau_{i,k}{}^{(0)}}$$

$$\hat{\pi}_k^{(1)} = \frac{\sum_{i=1}^{n} \tau_{i,k}{}^{(0)}}{n}$$

An iteration concludes with calculating the log likelihood, or:

$$\ell\big(\widehat{\boldsymbol{\Psi}}\big)^{(1)} = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \hat{\pi}_k^{(1)} f(\boldsymbol{x}_i; \widehat{\boldsymbol{\mu}}_k^{(1)}, \widehat{\boldsymbol{\Sigma}}_k^{(1)})$$

The EM algorithm continues to iterate between estimating the posterior probabilities $\tau_{i,k}$ and updating the parameter estimates ($\widehat{\boldsymbol{\Psi}}^{(m)}, \widehat{\boldsymbol{\Psi}}^{(m+1)}, \dots$) until the difference in log likelihood between consecutive iterations (i.e., $\ell\big(\widehat{\boldsymbol{\Psi}}\big)^{(m)} - \ell\big(\widehat{\boldsymbol{\Psi}}\big)^{(m+1)}$) is below a set tolerance or a maximum number of iterations is reached.

Covariance Specification

The previous section provided the calculations for $\widehat{\boldsymbol{\Sigma}}_k$ in the case that every unique element of each cluster's covariance matrix is estimated (i.e., $Kp(p + 1)/2$ estimated covariance parameters). However, we can alternatively express the within-cluster covariance matrix as its spectral decomposition:

$$\boldsymbol{\Sigma}_k = \lambda \boldsymbol{DAD}^T$$

where λ is a constant, $\boldsymbol{D}$ is a diagonal matrix with each nonzero entry an eigenvalue of $\boldsymbol{\Sigma}_k$, and $\boldsymbol{A}$ is a matrix of eigenvectors of $\boldsymbol{\Sigma}_k$. Using this decomposition, a set of parsimonious covariance models can be described (Celeux & Govaert, 1995). These models are created by constraining to the identity matrix, constraining to equality between clusters, or freely estimating the subparts of the spectral decomposition above. The result is mixtures of classes which differ in size (by varying λ), shape (by varying $\boldsymbol{A}$), and orientation (by varying $\boldsymbol{D}$). The 14 parsimonious models are provided in Table 1, and range from equal-identity-identity, abbreviated EII (i.e., clusters that are homogeneous in size, spherical in shape, and identity orientation), to variable-variable-variable or VVV (i.e., clusters which have heterogeneous sizes, shapes, and orientations).

Table 1. Gaussian Parsimonious Clustering Models

Number	Model	Volume	Shape	Orientation	Σ_k	Free covariance parameters
1	EII	Equal	Spherical	–	$\lambda\mathbf{I}$	1
2	VII	Variable	Spherical	–	$\lambda_k\mathbf{I}$	K
3	EEI	Equal	Equal	Axis-Aligned	$\lambda\mathbf{B}$	p
4	VEI	Variable	Equal	Axis-Aligned	$\lambda_k\mathbf{B}$	$p+K-1$
5	EVI	Equal	Variable	Axis-Aligned	$\lambda\mathbf{B}_k$	$pK-K+1$
6	VVI	Variable	Variable	Axis-Aligned	$\lambda_k\mathbf{B}_k$	pK
7	EEE	Equal	Equal	Equal	$\lambda\mathbf{DAD}^T$	$p(p+1)/2$
8	EEV	Equal	Equal	Variable	$\lambda\mathbf{D}_k\mathbf{A}\mathbf{D}_k^T$	$Kp(p+1)/2-(K-1)p$
9	VEV	Variable	Equal	Variable	$\lambda_k\mathbf{D}_k\mathbf{A}\mathbf{D}_k^T$	$Kp(p+1)/2-(K-1)(p-1)$
10	VVV	Variable	Variable	Variable	$\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}_k^T$	$Kp(p+1)/2$
11	EVE	Equal	Variable	Equal	$\lambda_k\mathbf{D}\mathbf{A}_k\mathbf{D}^T$	$p(p+1)/2+(K-1)(p-1)$
12	VVE	Variable	Variable	Equal	$\lambda_k\mathbf{D}\mathbf{A}_k\mathbf{D}^T$	$p(p+1)/2+(K-1)p$
13	VEE	Variable	Equal	Equal	$\lambda_k\mathbf{DAD}^T$	$p(p+1)/2+(K-1)$
14	EVV	Equal	Variable	Variable	$\lambda\mathbf{D}_k\mathbf{A}_k\mathbf{D}_k^T$	$Kp(p+1)/2-(K-1)$

Note: $\boldsymbol{\Sigma}_k$ - covariance matrix of the kth cluster, K – Number of clusters, p – number of variables, $\mathbf{I}$ - $p \times p$ identity matrix, $\mathbf{B}$ - $p \times p$ diagonal matrix, $\boldsymbol{DAD'}$ - spectral decomposition of covariance matrix, where $\boldsymbol{D}$ contains the eigenvalues and $\mathbf{A}$ the eigenvectors of $\boldsymbol{\Sigma}_k$.

Initialization Techniques

The EM algorithm is deterministic, meaning that a set of initial parameter estimates $\widehat{\boldsymbol{\Psi}}^{(0)}$ necessarily converges to the same solution (McLachlan & Peel, 2000). There are a handful of different techniques to obtain initial estimates that are implemented in popular software, including random initializations (McLachlan & Peel, 2000), *K*-means clustering (McLachlan & Peel, 2000), hierarchical clustering (Milligan, 1980), and a restricted EM algorithm (Biernacki, Celeux, & Govaert, 2003). The choice between these techniques can affect the goodness-of-fit and classification accuracy of the resulting model (Shireman, Steinley, & Brusco, 2016).

Random initializations involve either (1) randomly generating parameter values and starting the first stage of the algorithm calculating posterior probabilities, or (2) generate random classifications or posterior probabilities and begin the first stage of the algorithm calculating the parameter estimates. There are benefits and drawbacks to both of these techniques. Depending on the distribution from which initial parameters are drawn and the number of observations in the data, Technique 1 may have the ability to discover a wider variety of solutions, increasing the chances that the globally optimal solution is found. The maximum number of unique start values in Technique 2 is the same as the number of unique ways to classify the sample[1], or $\frac{1}{N!}\sum_{i=1}^{N}(-1)^{n-i}\binom{n}{i}i^K$ (Steinley, 2003). With any proper sample size, this value becomes very large, but for a continuous parameter generating distribution, the maximum number of unique start values is infinite. For Technique 1, however, the researcher (or software developer) needs to specify the distribution from which the start values are drawn. A poor choice in

[1] Note that this value is the maximum, not the expected, number of unique ways to classify the sample. The expected number of different ways to classify the sample and also have alternative parameter estimates is empirically determined by the data, but will only be maximally large as this constant.

distribution could lead to so-called "spurious" solutions, or solutions that exist on the boundary of the parameter space (e.g., an empty or nearly empty group) which would result in an inflated likelihood value (McLachlan & Peel, 2000).

Model Selection

The specification of both the covariance model (M) and the number of clusters (K) in a mixture model leads to a large set of potential models to compare. For instance, if a researcher is interested in comparing 1- though 6-class models and all fourteen covariance matrix types, this leads to a collection of 84 models to compare. When evaluating mixture models with increasing numbers of clusters and covariance parameters, the typical likelihood ratio statistic (i.e., the ratio of likelihoods corrected by their degrees of freedom) does not follow a χ^2 distribution with the degrees of freedom equal to the difference in parameters between the two models being compared (McLachlan, 1987), making it difficult to determine whether additional parameters significantly improve the fit of the mixture model. Therefore, researchers typically rely on absolute fit indices and adjustments to the likelihood ratio test.

Fit indices (or fit statistics) are measures of fit that are not typically comparisons of a candidate model over a null model, and thus are useful for non-nested model comparison. Fit statistics are optimal at the model that achieves the desired goal of the specific statistic, which is (in most cases) maximization of fit with a penalty to account for parsimony. The strength of this penalty is the main way that these statistics differ. The most common fit indices are the Bayesian Information Criterion (BIC; Schwarz, 1978), Akaike Information Criterion (AIC; Akaike, 1974), and the Sample-Size Adjusted

Bayesian Information Criterion (SSA-BIC; Rissanen, 1978). However, many other fit statistics are available or calculable by hand given a handful of summary measures of the model. Many simulation and empirical studies have been undertaken to examine the tendency for these statistics to underfit or overfit the number of components and/or the covariance model of the mixtures (to be discussed in detail in the following chapter; Andrews, Ainslie, & Currim, 2002; Bauer & Curran, 2003; Biernacki & Govaert, 1999; Biernacki & Govaert, 1997; McLachlan & Peel, 2000; Nylund, Asparouhov, & Muthén, 2007; Oliver, Baster, & Wallace, 1996; Xiang & Gong, 2005; Yang & Yang, 2007). Most have shown that the BIC is among most accurate fit statistics, though it tends to underfit the number of components when it errs. These studies also consistently agree that the AIC is quite inaccurate, and tends to overfit the number of components.

Adjusted likelihood ratio tests attempt to test the increase in fit balanced with the increase in estimated parameters in the case when the typical likelihood ratio test is not asymptotically χ^2. Vuong (1989) developed a test for non-nested models (mainly in multiple regression) where the likelihood ratio statistic is distributed according to the sum of weighted χ_1^2 (i.e., one degree of freedom). Lo, Mendell, and Rubin (2001) adapted this test for comparing mixture models with *k* and *k-1* classes (referred to as the LMR-LRT). McLachlan (1987) developed a similar bootstrapping test to compare *K* with *K-1* classes based on bootstrapping the log likelihood value (referred to as the BLRT). The BLRT has been shown to outperform the LMR-LRT in a large simulation study (Nylund et al., 2007).

Heuristics for Model Selection

Model selection for mixture models in the social sciences is not a standardized process. Most studies include the BIC, SSA-BIC, and/or the BLRT. However, many studies also present the AIC, the LMR-LRT, and Vuong Lo-Mendell-Rubin Likelihood Ratio Test (VLMR-LRT), despite the studies showing the lower accuracy of these statistics compared to the BIC and the BLRT. Furthermore, many will examine measures of classification certainty, such as the entropy (Biernacki, Celeux, & Govaert, 2000), the average probability of out-of-class assignment (i.e., the average $\tau_{i,l}$, where *l* indexes over all posterior probabilities except the maximum), and the smallest class proportion in the model. And finally, researchers will frequently cite theoretical concerns in model selection, examining the means of competing models for interpretability and alignment with existing theory (Marsh, Lüdtke, Trautwein, & Morin, 2009; Muthen, 2002). As an illustrative example of the variety of procedures in model selection, take for example Borden et al. (2014) and Gomez, Gomez, Winther, & Vance (2014). Both of these studies present latent profile analyses or LPAs (i.e., mixture models with homogeneous, spherical covariance matrices) and were published in the *Journal of Abnormal Child Psychology* in the same year. However, they differ substantially in their method of model selection.

Borden et al (2014) fit LPAs with 1-5 classes and compared the fit of these competing models with the BIC, SSA-BIC, VLMR-LRT, and Entropy. The BIC and SSA-BIC are optimal at the maximum number of clusters (five), the VLMR-LRT is only significant when comparing one cluster versus two, and the Entropy is maximal at two clusters. However, the researchers select the four-cluster solution for interpretation (a model not optimal on any examined criteria). The authors justify their model selection by stating that the means in the five-cluster solution were “uninterpretable”, and the four-cluster solution is preferred to the three cluster solution using the BLRT.

Gomez et al (2014) also fit LPAs with 1-5 classes. In addition to the BIC, SSA-BIC, VLMR-LRT, and Entropy, the researchers present the LMR-LRT. The BIC and Entropy are optimal at the 3-class solution, the SSA-BIC is optimal at the 4-class solution, and the VLMR-LRT, and LMR-LRT are significant up to the comparison of the 3- and 4-class solutions. The researchers select the 3-class model by citing the difference in entropy (.64 in the 4-class solution and .74 in the 3-class solution). Therefore, the selected model was optimal on 2 of the statistics they examined, and the researchers did not report examining the means of the clusters in the data to influence their decision.

To summarize, one study cites the interpretability of the solution to justify the selection of clusters, selecting a model that is not optimal on any criterion (not an uncommon occurrence, see examples in Adams et al., 2011; Borden et al., 2014; DiStefano, 2006; Flensborg Damholdt, Shevlin, Borghammer, Larsen, & Østergaard, 2012; Geiser et al., 2014; Giang & Graham, 2008; Grunschel, Patrzek, & Fries, 2013; Hill, Degnan, Calkins, & Keane, 2006; Miller, Turner, & Henderson, 2009; Mokros et al., 2015; Morin et al., 2011; Pastor, Barron, Miller, & Davis, 2007; Turner, Miller, & Henderson, 2008), and another selects between models based on a change in criteria (also not uncommon, see examples in Au, Dickstein, Comer, Salters-Pedneault, & Litz, 2013; Chen, 2012; Gerber, Jonsdottir, Lindwall, & Ahlborg, 2014; Giang & Graham, 2008; Martinson et al., 2011; Rajendran, O’Neill, Marks, & Halperin, 2015; Scheier, Ben Abdallah, Inciardi, Copeland, & Cottler, 2008; Vaughn, Perron, & Howard, 2007). There are a variety of other model selection heuristics that have been implemented, however, including the rate of decrease in the BIC or log likelihood (Giang & Graham, 2008; Maynard, Salas-Wright, Vaughn, & Peters, 2012; Vaughn, DeLisi, Beaver, & Howard, 2008), and the size of classes (Hall, Howard, & McCabe, 2010; Keefer, Parker, & Wood, 2012; Merz & Roesch, 2011). What all of the above heuristics share is that none are based on selection of a model that exhibits the best value on an established information criterion, and also that each adds an amount of subjectivity in model selection. Additionally, none have been systematically examined for efficacy or supported by statistical theoretical work.

Local Optima

The estimation of mixture models is occasionally problematic due to the existence of local optima, or local maximizers of the likelihood function. It is sometimes difficult, and other times impossible, to determine whether any one solution is globally optimal. This is due to the fact that locally optimal solutions result from two conditions: either (1) out of the initializations of the EM algorithm, none may happen to find the globally optimal solution (see a technique for determining global optimality in Gan & Jiang, 1999), but also that (2) the global optimum may not exist, and thus every solution is a locally optimal solution. The global optimum does not exist when the covariance matrices are heterogeneous (i.e., any covariance model excluding EII, EEI, and EEE), as the likelihood function is unbounded (McLachlan & Peel, 2000). The goal, then, is to initialize a requisite number of times to find a solution that closely approximates the globally optimal solution.

Hipp and Bauer (2006) examined the issue of locally optimal solutions in the growth mixture model (GMM), a longitudinal extension of the mixture model which typically assumes a functional form of growth over time. The researchers examined the relationship between the number of estimated parameters, the convergence of the algorithm, the number of unique solutions, the percent of initializations for which the best likelihood is found, and the percent of datasets where the modal solution has the best likelihood. As the number of clusters in the model increases, they found that the number of local optima also increases monotonically. Additionally, as the overlap of the classes increases, as does the number of local optima found. These results imply that an increase in local optima should be expected as K increases and as these classes become closer together.

In a similar examination, Steinley (2006) examined the factors which influence local optima in *k*-means clustering. *K*-means clustering is equivalent to mixture modeling when the covariance matrices are homogeneous-spherical (i.e., $\boldsymbol{\Sigma}_k = \boldsymbol{\Sigma} = \lambda \boldsymbol{I}$ or model EII; Steinley & Brusco, 2011; Steinley & McDonald, 2007). Steinley (2006) found a strong relationship between a dataset that has a large number of local optima and the classification quality of the solution, and developed a diagnostic technique to determine when a dataset has so many local solutions that the resulting partition is likely to be poor.

Shireman, Steinley, & Brusco (2016a) extend Steinley's (2006) work on locally optimal solutions to mixture models. The authors find that, even when the number of clusters is correctly specified, mixture models show a high propensity to find locally optimal solutions. Additionally, the authors find a moderate relationship between greater numbers of local optima and poor classification results, with local optima being more likely to occur in data where clusters have a high amount of overlap and when the data includes variables that are unrelated to cluster structure.

Model Nonconvergence

In addition to local optima, researchers fitting a mixture model have to navigate around nonconvergent models. Models are typically considered "converged" if the log likelihood difference at successive iterations of the EM algorithm falls below a tolerance level (i.e., $\ell\left(\widehat{\boldsymbol{\Psi}}^{(m)}\right) - \ell\left(\widehat{\boldsymbol{\Psi}}^{(m-1)}\right) < T$, where T is a small number set by the software program). Additionally, the model can be considered nonconvergent if, at any point during the EM algorithm, the estimated covariance matrix is singular.

There is some inconsistency between software programs in how they indicate a model is converged. First, many differ to a great degree on the default minimum log likelihood difference for the EM algorithm—ranging from machine-precision equality (SAS®, 2004) to 1E-05 (R package **mclust**; Fraley & Raftery, 2006). Therefore, a model that does not converge in one software program may converge using a program with a larger tolerance. Additionally, only some software programs check for singularity of the covariance matrices: the statistical computing software M*plus* (Muthén & Muthén, 2007) does not quit the iteration process when a cluster covariance matrix nears singularity, while the R package **mclust** does (Fraley & Raftery, 2006). This error alerts the user to an empty or near-empty cluster (i.e., a potential spurious solution).

Computational Indicators of Model Misspecification

In a variety of studies, models which either do not converge or find a large number of local optima are typically omitted from consideration for model selection. However, it is difficult to tell without detailed model fitting procedures whether researchers ran into convergence issues that are more indicative of increased complexity of the fitted model, or true model misspecification. As an illustrative example, consider Wolf et al. (2012) where the authors omit the best-fitting model from consideration, stating that "*The best log likelihood value was not replicated, despite increasing the number of random starts. This can be an indication that too many classes were extracted and/or that a local maxima* [sic] *was reached and the parameter estimates may not be reliable or replicable*" (Table 1, table note *a*). This table note was used to justify the selection of a suboptimal solution—a change in BIC from −20311.698 in the convergent model to −20568.202 in the model with one more class (where lower BICs are to be preferred, an absolute difference in BIC of 205.504). I will now discuss this table note in detail, which exhibits typical reasoning for omitting models for model selection due to computational difficulties (Morin, Morizot, Boudrias, & Madore, 2011; Pastor, Barron, Miller, & Davis, 2007; Stapleton, Turrisi, Hillhouse, Robinson, & Abar, 2010), and explain why I believe that it leaves out essential information for determining whether the model was truly over-specified and also demonstrates a lack of understanding of the way mixture models are fit.

First, the note begins by stating: "*The best log likelihood value was not replicated, despite increasing the number of random starts.*" The researchers are stating that among all the initializations, the solution with the best log likelihood occurred only once. This could occur when each initialization is a local optimum. The researchers do not provide the number of intializations they started with, and the amount to which they increased, to determine whether they fully attempted to find all possible local optima. It could be the case that their data have so many local optima that finding each one is not possible with the default number of initializations (and perhaps many more than that, see Shireman, Steinley, & Brusco, 2016a, 2016b).

The note continues to state that the lack of replication of the log likelihood value "*can be an indication that too many classes were extracted"*. They are referring to the phenomenon that the optimal model is more likely to be found frequently if the model is correctly specified (Hipp & Bauer, 2006). However, without providing the number of local optima and number of initializations, it is difficult to tell whether the best solution occurred only once due to the preponderance of local optima, or whether the model is overspecified. The number of unique solutions is partly based on data conditions (Shireman, Steinley, & Brusco, 2016b) and partly analysis conditions (i.e., the number of clusters fitted to the data; Hipp & Bauer, 2006), so considering the high likelihood of local optima even in a correctly specified mixture model, it is not immediately clear that misspecification is the sole cause for the lack of replication of the log likelihood value.

The authors continue to state that the best model only occurring once can also indicate "*that a local maxima* [sic] *was reached*". As is discussed above, it is sometimes impossible to tell whether a solution is globally or locally optimal. Additionally, the initialization technique for their analyses is not explicitly outlined, which would provide further essential information to determine whether the lack of replication is due to a spurious solution, misspecification, or a preponderance of unique solutions. Nuances in the initialization procedure could cause the researcher to find a wide variety of solutions, or the same solution twice, depending on the distributions from which the parameters are generated. Therefore, finding a solution frequently should not contribute evidence toward a model being the correct specification or the globally optimal solution. Furthermore, even if the globally optimal solution were to be found, there is no guarantee that it is the "correct" solution (Gan & Jiang, 1999).

Finally, Wolf (2012) states that the lack of replicability of the log likelihood is important because "*the parameter estimates may not be reliable or replicable*". I believe the phenomenon to which the authors are referring is the tendency for mixture models to arrive at "spurious" solutions, or solutions with very large likelihoods that are on the boundary of the parameter space. These inflated likelihoods can result from poor initial values, but also when one or more data points are perfectly described by a cluster's parameters, which occurs if a cluster is defined over very few points. Many software programs deal with this issue by checking whether covariance matrices are nearing singularity at each iteration of the algorithm. If a solution is spurious because of a small cluster, this can be determined by looking closely at the parameter estimates. However, the researchers provide no information on whether the parameter estimates of this solution indicate that it is spurious.

Wolf et al (2012) is only one of many studies that omit one or several models from consideration in model comparison due to local optima or lack of replicability of the optimal solution (Geiser, Okun, & Grano, 2014; Hori et al., 2014; Klonsky & Olino, 2008; Morin et al., 2011; Pastor, Barron, Miller, & Davis, 2007; Stapleton et al., 2010). In fact, the table note in Wolf et al (2012) closely approximates the wording in an error message output from the software M*plus* (Muthén & Muthén, 2007). This error message explicitly states: "*WARNING: THE BEST LOGLIKELIHOOD VALUE WAS NOT REPLICATED. THE SOLUTION MAY NOT BE TRUSTWORTHY DUE TO LOCAL MAXIMA. INCREASE THE NUMBER OF RANDOM STARTS*". M*plus*, however, lacks certain programming that would be useful for understanding the specific modeling difficulties that researchers are encountering when they receive this error. For instance, M*plus* does not clearly outline the parameter generation technique that underlies their mixture model initialization nor provides information on how many local optima were found or the frequency of each local optimum. This information would give the user an idea of whether there is a preponderance of local optima, or whether the best solution truly is spurious. Additionally, M*plus* does not check for singular covariance matrices during computation, which would additionally assist the user to determine whether a solution is spurious. Without this essential programming, it is difficult to tell whether researchers who remove solutions due to difficulty in computation are doing so due to model misspecification, or due to the typical difficulty in mixture model estimation that occurs even when the model is correctly specified.

Majority Vote Model Selection

Majority Vote Model Selection

As was discussed above, it is commonplace in psychological research to present mixture model results for competing models using several different criteria for model selection. When selecting a final model, researchers will frequently holistically consider all of the fit indices that a software program provides. Therefore, it is useful to understand whether there is an increase in model selection accuracy when examining many fit indices at once. Considering that it is known that some fit indices tend to overestimate the number of clusters (e.g., AIC) and some to underestimate (e.g., BIC), when these statistics agree it could hypothetically represent a greater amount of evidence toward that model than solely using the most accurate statistic.

To test whether there are true benefits in model selection accuracy from using this common heuristic, I perform a large simulation and real data demonstration that systematically assesses (1) the accuracy of 14 fit statistics for mixture modeling used singularly, and (2) the accuracy of using a “majority vote” model selection heuristic, or selecting the model that is optimal on the most criteria. I calculate two majority vote statistics, first using all of the 14 criteria examined, and another with only the most common statistics (AIC, BIC, and SSA-BIC). What follows is a description of 14 information criteria, previous simulation comparisons of criteria accuracy, and the current examination.

1. Bayesian Information Criterion (BIC). The BIC (equivalent to the Minimum Description Length in the machine learning literature) is the most widely used and most heavily studied information criterion (for more detail and a theoretical background in the context of mixture modeling, see McLachlan & Peel, 2000, Ch 6.9). The objective of the BIC (as is the objective of many fit statistics) is to maximize the probability of observing the data given the model, while controlling for the capitalization on chance by penalizing the likelihood by a function of the number of parameters used to estimate the mixture model. It also follows the structure that most information criteria follow: *Model Fit - Parameter Penalization*. The BIC is formulated

$$BIC = 2\,\ell\big(\widehat{\boldsymbol{\Psi}}_{K,M}\big) - d\log n$$

where d is the number of estimated parameters, n is the sample size, $\ell(\cdot)$ is the log likelihood as defined in Chapter 1, and $\widehat{\boldsymbol{\Psi}}_{K,M}$ denotes jointly the data, $\boldsymbol{x}$, and the parameters of the model with K clusters and covariance model M, where M is one of the 14 parsimonious clustering models in Table 1.

2. Sample-size Adjusted Bayes Information Criterion (SSA-BIC). The SSA-BIC is the BIC with an alternative penalization of the log likelihood (Rissanen, 1978), designed to overcome the BIC's tendency to underfit the number of components. It adjusts the sample size by replacing n with $N^* = (n+2)/24$.

$$SSA\ BIC = 2\,\ell\big(\widehat{\boldsymbol{\Psi}}_{K,M}\big) - d\log N^*$$

3. Akaike's Information Criterion (AIC). The second most popular information criterion for model selection is the AIC (Akaike, 1973). The AIC's popularity is mainly rooted in theoretical justification (see Bozdogan, 2000). It follows the same structure as the BIC, but they differ in the parameter penalization:

$$AIC = 2\,\ell\big(\widehat{\boldsymbol{\Psi}}_{K,M}\big) - 2d$$

4. Consistent Akaike's Information Criterion (CAIC). Bozdogan (1987) proposed a consistent AIC that aims to overcome the AIC's tendency to choose models which are too complex. It is calculated:

$$CAIC = 2\,\ell\big(\widehat{\boldsymbol{\Psi}}_{K,M}\big) - d\log(n+1)$$

5. Sample Size Adjusted Consistent Akaike's Information Criterion (SSA-CAIC). The SSA-CAIC is the same as the CAIC but with the same sample size adjustment used in the SSA-BIC (Yang & Yang, 2007):

$$SSA\ CAIC = \ 2\,\ell\left(\widehat{\mathbf{\Psi}}_{K,M}\right) - d\log(N^* + 1)$$

6. Completed Likelihood-Akaike's Information Criterion (CL-AIC). In yet another correction on the AIC, the AIC is altered by utilizing the completed likelihood, or the likelihood function which includes the vector of class labeling (Xiang & Gong, 2005). Explicitly, the completed likelihood of $\widehat{\mathbf{\Psi}}_{K,M}$ is:

$$L_c\left(\widehat{\mathbf{\Psi}}_{K,M}\right) = \ \sum_{k=1}^{K} \hat{\pi}_k f\left(\boldsymbol{x};\ \widehat{\boldsymbol{\mu}}_k, \widehat{\mathbf{\Sigma}}_k\right) + \ \sum_{i=1}^{n}\sum_{k=1}^{K} z_{i,k}\log\tau_{i,k}$$

where $\boldsymbol{\mu}_k$ and $\mathbf{\Sigma}_k$ are the parameter estimates for the kth cluster, $\tau_{i,k}$ is the posterior probability that individual *i* is classified into cluster *k*, and the $z_{i,k}$ are binary cluster labels such that

$$z_{i,k} = \begin{cases} 1 & if\ \ \max\{\tau_{i,1}, \dots, \tau_{i,K}\} = \ \tau_{i,k} \\ 0 & else \end{cases}$$

The CLAIC is calculated:

$$CL\ AIC = \ 2\,L_c\left(\widehat{\mathbf{\Psi}}_{K,M}\right) - d$$

7. Akaike's Information Criteria "3" (AIC3). The AIC3, sometimes referred to simply as the AIC, was created to correct the AIC with a stricter parameter penalization (Bozdogan & Sclove, 1984). Its formulation is given below:

$$AIC3 = \ 2\,\ell\left(\widehat{\mathbf{\Psi}}_{K,M}\right) - 3d$$

8. Akaike's Information Criterion-Bozdogan (AIC-BOZ). The AIC-BOZ is a criterion very similar to the AIC3, but with an alternative calculation of the constant used to scale the likelihood value (Bozdogan, 1993). The best model maximizes the following:

$$AIC\ BOZ = 2n^{-1}(n - 1 - dK^{-1} - K2^{-1})\ell\left(\widehat{\boldsymbol{\Psi}}_{\mathrm{K,M}}\right) - 3d$$

9. Integrated Completed Likelihood (ICL). The ICL was developed to make use of the completed likelihood (Biernacki et al., 2000; Biernacki & Govaert, 1997):

$$ICL = L_c(\widehat{\boldsymbol{\Psi}}_{\mathrm{K,M}}) - d2^{-1}\log n$$

10. Integrated Completed Likelihood-Bayes Information Criterion (ICL-BIC). The ICL-BIC makes use of the estimated entropy of the solution as an additional way to account for parsimony (Biernacki et al., 2000; McLachlan & Peel, 2000). The ICL-BIC is formulated as follows:

$$ICL\ BIC = 2\,\ell(\widehat{\boldsymbol{\Psi}}_{K,M}) + 2\,EN(\boldsymbol{\tau}) - d\log n$$

where $\boldsymbol{\tau}$ is the matrix of posterior probabilities ($\tau_{i,k}$; $i = 1, \dots, N, k = 1, \dots, K$) and $EN(\boldsymbol{\tau})$ is the estimated entropy:

$$EN(\boldsymbol{\tau}) = -\sum_{k=1}^{K}\sum_{i=1}^{n}\tau_{i,k}\log\tau_{i,k}$$

11. Approximate Weight of Evidence (AWE). The AWE was developed as an approximation of the exact Bayes solution (Banfield & Raftery, 1993). It also uses the completed likelihood and a correction factor similar to the BIC, and it is calculated as follows:

$$AWE = 2\,L_c\left(\widehat{\boldsymbol{\Psi}}_{K,M}\right) - d(3/2 + \log n)$$

12. Information Complexity (ICOMP). ICOMP was also developed to select a model without the characteristic overfit for which the AIC is known (Bozdogan, 1990, 1993). However, the ICOMP takes into consideration the complexity and interrelatedness of the data to provide a statistic that demonstrates the trade-off between complexity and likelihood:

$$ICOMP = 2\,\ell\left(\widehat{\boldsymbol{\Psi}}_{K,M}\right) + d\log d^{-1}\,tr\left(\mathbb{I}^{-1}\left(\widehat{\boldsymbol{\Psi}}_{K,M}\right)\right) - \log \mathbb{I}^{-1}\left|\widehat{\Psi}_{K,M}\right|$$

where $\mathbb{I}\left(\widehat{\boldsymbol{\Psi}}_{K,M}\right)$ is the expected information matrix, with the estimation process described in Bozdogan (1993).

13. Partition Coefficient (PC). The PC is focused on finding a solution which maximizes the posterior probabilities of class assignment (Roberts, Husmeier, Rezek, & Penny, 1998; Windham & Cutler, 1992), which was originally described in Bezek (1981). It is defined as:

$$PC = n^{-1}\sum_{i=1}^{n}\sum_{k=1}^{K}\tau_{i,k}^{2}$$

14. Fuzzy Hypervolume (FHV). FHV defines as optimal the classes that have the lowest total volume (Gath & Geva, 1989; Roberts et al., 1998). It is defined below:

$$FHV = \sum_{k=1}^{K}\sqrt{\left|\widehat{\boldsymbol{\Sigma}}_{k}\right|}$$

where $|\cdot|$ indicates the matrix determinant.

Previous Simulation Comparisons

This section describes previous comparisons of fit indices for mixture modeling and related techniques. Although not exhaustive, this list provides a useful background on the comparisons of fit indices conducted previously and how the current work adds to this line of research. Information contained in this section is summarized in Tables 2 and 3.

Table 2. Previous Simulation Comparisons of Fit Indices: Fit Indices Studied

Study	BIC	SSA-BIC	AIC	CAIC	SSA-CAIC	CL-AIC	AIC3	AIZ-BOZ	ICL	ICL-BIC	AWE	ICOMP	PC	FHV
AC03	X		X	X			X					X		
BC03	X		X	X						X				
BG97	X		X				X				X	X		
BG99	X		X				X					X		
BCG00	X								X					
MP00	X		X						X	X				
NAM07	X	X	X	X										
OBW96	X		X					X				X	X	
XG06	X		X			X			X					
YY07	X	X	X	X	X		X							

Note: AC03- Andrews & Currim, 2007; BC03 - Bauer & Curran, 2003; BG97- Biernacki & Govaert, 1997; BG99- Biernacki & Govaert, 1999; BCG00- Biernacki, Celeux, & Govaert, 2000; MP00- McLachlan & Peel, 2000; NAM07- Nylund, Asparouhov, & Muthen, 2007; OBW96- Oliver, Baxter, & Wallace, 1996; XG06- Xiang & Gong, 2006; YY07- Yang & Yang, 2007

Andrews & Currim (2003). Concerned specifically with nominal data in the field of marketing segmentation research, Andrews and Currim (2003) compare several information criteria in their selection of the correct number of market “segments” (i.e., groups or clusters). In their simulation comparison, accuracies are compared in estimating the number of clusters for many statistics including the AIC, AIC3, ICOMP, BIC, and CAIC (all defined identically as in the previous section).

In the simulation comparison, data were generated with 2 or 3 clusters, with small or large cluster separation (mean separations of 1.0, 1.5, and 2.0), sample sizes varying between small (100) or large (300), 3 or 6 choice alternatives (i.e., number of unique values of the simulated variables), and minimum mixing proportion (5-10%, 10-20%, or 20-30%), replicated three times per cell. In addition, they generate 100 individuals to serve as a validation sample.

The authors find that the best performance overall comes from the AIC3 and the worst performance by the CAIC. The BIC never overfit the number of components; however, it was very likely to underfit.

Bauer & Curran (2003). Bauer and Curran (2003) focused on the distributional properties of a growth mixture model, examining the propensity for the mixture model to

Table 3. Previous Simulation Comparisons of Fit Indices: Simulation Conditions

Article	Focus of Assessment	Clusters	Number of Variables	Sample Size	Mixing Proportions	Variances	Overlap	Additional Simulation Factors	Fully crossed?
AC03	Number of clusters	2 or 3	Generated from $\Gamma(5)$ or $\Gamma(10)$	100 or 300	.5-.10, .10-.20, or .20-.30	Standard Normal	Small, medium, large	3 or 6 choice alternatives	Y
BC03	Number of clusters	1	5	200 or 600	N/A	Equal	N/A	Increasing skew and kurtosis	Y
BG97	Number of clusters	2 or 3	1 or 2	30	Equal or unequal	Equal or unequal	Small, medium, large		N
BG99	Covariance model and classification accuracy	2	2	40 or 200	(.50,.50) or (.3,.7)	14 GPCM	Small, medium, large		Y
BCG00	Number of clusters and covariance model	1 or 3	2	400	Equal	Equal or unequal	Small or large	Non-normality in the single-class condition	N
MP00	Number of clusters	2, 3, or 4	2, 3, or 4	100, 300, or 625	Equal or unequal	Equal or unequal λ	Small or large		N
NAM07	Number of clusters	3 or 4	Between 3 and 15	200, 500, or 1000	Equal or unequal	Standard Normal	Low		Y
OBW96	Number of clusters	1,2, or 3	2	20, 40, 80, or 100	Equal	Equal	Low or Very High		N
WC92	Number of clusters	3	2	300	Equal	Standard Normal	Low to high		Y
XG06	Number of clusters	5	2	From 25 to 1000 by 25s	Unequal	Unequal	Severely	Degree of within-class deviations from normality	Y
YY07	Number of clusters	1, 2, 3, or 4	Between 1 and 7	200, 400, 600, 800, or 1000	Equal		Low on two dimensions	Additionally considered a MIMIC LCA	Y

Note: AC03- Andrews & Currim, 2007; BC03 - Bauer & Curran, 2003; BG97- Biernacki & Govaert, 1997; BG99- Biernacki & Govaert, 1999; BCG00- Biernacki, Celeux, & Govaert, 2000; MP00- McLachlan & Peel, 2000; NAM07- Nylund, Asparouhov, & Muthen, 2007; OBW96- Oliver, Baxter, & Wallace, 1996; WC92- Windham & Cutler, 1992; XG06- Xiang & Gong, 2006; YY07- Yang & Yang, 2007 overfit components to a single, non-normal distribution. Several fit statistics are compared to see if any is more likely to overfit, including the AIC, BIC, CAIC, and ICL-BIC.

The data were always generated with only one true group, at sample sizes of 200 and 600, and with various distributional properties: distributed normally, skewed a small amount with low kurtosis (skew=1, kurtosis=1), and skewed more heavily with high kurtosis (skew=1.5, kurtosis=6). Two mixture models were fit to each simulated data set: a 1-cluster and 2-cluster model. The researchers assessed overfit by those that selected the 2-class over the 1-class solution.

When the data are normally distributed, the fit statistics on average chose the correct, one-class solution. The AIC was the most likely to select the two-class model in the normally-distributed case. When the data are non-normal, the AIC, BIC, and CAIC overfit the number of classes for nearly every data condition. The ICL-BIC is the most resistant to choosing the two-class model.

Biernacki & Govaert (1997, 1999). Biernacki and Govaert published two large simulation comparisons of several fit statistics which are included in these analyses. Their 1997 comparison included the AIC, AIC3, BIC, ICOMP, AWE, and the log likelihood. The data in the 1997 simulation were varied by the mixing proportions (equal, not equal) and the separation as defined by the average probability of misclassification (poor, middle, and strong separation). They find that the AIC3 and the BIC perform the best.

In 1999, Biernacki and Govaert followed up with an additional simulation. They included the log likelihood, AIC, AIC3, BIC, and ICOMP. This examination was unique because it examined the complexity (i.e., number of total parameters) of the chosen model for each statistic. In addition, models were evaluated based on the resulting "misclassification", or the average posterior probabilities that points were allocated to clusters other than the one they are classified into. Data were varied by mixing proportion, misclassification rate (i.e., cluster separation), and covariance matrices. This examination found that the log likelihood best approximates the true misclassification rate, however this statistic overfits the number of parameters. The AIC most nearly approximates the true mean number of parameters, however this statistic tended to overfit.

Biernacki, Celeux, & Govaert (2000). Biernacki, Celeux, and Govaert (2000) propose the Integrated Completed Likelihood (ICL), a fit statistic which is growing in popularity in the statistics literature (see recent uses in Andrews, McNicholas, & Subedi, 2011; Steane, McNicholas, & Yada, 2012). To demonstrate the ICL's efficacy, Biernacki and colleagues compare it to the BIC in a limited simulation and real data demonstration assessing the ability to choose the correct number of clusters and the covariance model. The simulation varied the orientation of the clusters as well as the overlap (well-separated and overlapped). The BIC shows better performance than the ICL in selecting the correct number of clusters when the clusters are overlapped, and slightly worse performance when the clusters are well-separated.

McLachlan & Peel (2000), Ch 6.11. In McLachlan and Peel's (2000) discussion of selection of the number of classes in the mixture model, a small simulation was conducted examining the accuracy several fit indices including the AIC, BIC, ICL, and ICL-BIC. Three datasets are generated that vary in the sample size (625, 300, and 200), means, and covariances. The ICL-BIC and ICL choose the correct number of classes in each case. The AIC and BIC overestimated for at least one data set.

Nylund, Asparouhov, & Muthen (2007). Nylund, Asparouhov, and Muthen (2007) compared the performance of the AIC, CAIC, BIC, and SSA-BIC in selecting the model in a latent class analysis (i.e., mixtures with binary indicators), a mixture of factor analyzers (McLachlan & Peel, 2000), and a growth mixture model (Ram & Grimm, 2009). Data varied by the sample size (200, 500, and 1000), number of variables (between 3 and 15), number of classes (3 or 4), mixing proportions (equal, unequal), and "structure", or the differentiation of the classes on their item probabilities (simple or complex). 100 replications (500 in the case of binary indicators) of each data type was generated.

Their results show that the AIC underperforms consistently, but the rest vary in their performance depending on the data complexity, concluding that the BIC is superior. The SSA-BIC, although having adequate results, loses its utility at medium sample sizes.

Oliver, Baxter, & Wallace (1996). Oliver et al. (1996) compared the efficacy of the Minimum Message Length (MML, a computationally equivalent statistic to the BIC developed by Wallace & Freeman, 1987), AIC-BOZ, PC, and ICOMP. The data generation in their simulation comparison varied sample sizes of 20, 40, 80, and 160, increasing variances, 2 or 3 clusters, and homogeneity of variances.

The MML (BIC) is found to be the best at selecting the number of clusters, tending slightly to underfit when the variance of the classes is large. The AIC-BOZ tended to overfit the number of classes, and the PC favored the 2-cluster solution. The ICOMP frequently overfit at larger sample sizes.

Xiang & Gong (2006). Xiang and Gong (2006) developed the CL-AIC and used a simulation to demonstrate the efficacy of the CL-AIC as compared to other popular indices. Their simulation comparison included the BIC, ICL, and AIC. The simulated data varied by sample size (between 25 and 100 by increments of 25) and the normality of the within-class distributions (Gaussian or slightly non-Gaussian). All data were generated with uniform noise on each dimension. The CL-AIC was shown to overfit when the sample size is small. However, when the sample size increases and when the data are non-Gaussian, it outperforms all other statistics.

Yang & Yang (2007). Yang and Yang's (2007) proposed the SSA-CAIC as well as compared its efficacy of the AIC, AIC3, BIC, CAIC, and SSA-BIC. Data were generated with 1, 2, 3, or 4 latent classes, with sample sizes of 200, 400, 600, 800, and 1000, and one or two variables (i.e., only univariate or bivariate data). “Distinctness” (i.e., separation) of the latent classes was varied between easy, moderate, and hard structures. All proportions are set to be equal (i.e., $\pi_1 = \pi_2 = \cdots = \pi_K = 1/K$) and each data set was replicated 500 times. The AIC, BIC, and CAIC were calculated for each model, including their sample size-adjusted counterparts.

As sample size increased, all statistics improved in choosing the right number of clusters except for the AIC, which decreased in performance. SSA-BIC and AIC3 were the most accurate, but recovery did not reach a high level (90%) until the sample size was over 600. Although all statistics decreased in performance when the data contain more classes, the BIC and CAIC were most affected. Adding the sample size adjustment showed a considerable improvement to CAIC's performance.

Simulation Comparison: Individual Statistics

Examining Tables 2 and 3 shows a great deal of sparseness in the inclusion of a variety of statistics—although most include the AIC and BIC, some which are shown to be among the most accurate when they are examined (e.g., SSA-CAIC, CL-AIC, and AIC-BOZ). Additionally, many of the data conditions represent nearly optimal conditions. For instance, of the 11 studies whose results are discussed above, five are restricted to the performance of the fit indices using bivariate data. Additionally, most simulations examine the accuracy of the in the number of clusters, two examine the accuracy in selecting the covariance model, but none examine the accuracy of fit indices in selecting the correct number of clusters, covariance model, and a model with suitable classifications.

Therefore, to replicate and extend previous research in the accuracy of fit statistics, I perform a large simulation that examines the accuracy of the 14 statistics to individually select a model. The data vary according to the following factors: (1) the number of clusters (*C* = 2 and 3), (2) the number of variables (*V* = 4 and 6), (3) the relative cluster density (also called mixing proportions) as described in Milligan (1980) and Milligan and Cooper (1988) and implemented in Steinley (2003, 2006) (*D* = 1, equally sized clusters or $\pi_1 = \pi_2 = \cdots = \pi_K = 1/K$, and *D* = 2, one small cluster containing 10% of the sample or $\pi_1 = .1,\ \pi_2 = \cdots = \pi_K = .9/(K-1)$), (4) the sample size (*S* = 500 and 1000), (5) the average cluster overlap (*O* = Low and High, defined below), and (6) the covariance model (*M* = EII, EEE, and VEV, see Table 1). The data were generated with the package **MixSim** (Melnykov, Chen, & Maitra, 2012) for the statistical computing software R (R Development Core Team, 2008)

The R package **MixSim** generates overlapped clusters by varying the probability of misclassification (Melnykov & Maitra, 2010). The average probability of misclassification ($\bar{\omega}$), the maximum probability of misclassification ($\breve{\omega}$), or both can be controlled in the generation of data. For the following comparisons, the clusters were generated solely using $\bar{\omega}$ and this was varied to be $\bar{\omega}$ = .001 (considered to be "low" overlap, see Melnykov & Matria, 2010), and $\bar{\omega}$ = .2 ("high" overlap).

The factor conditions were fully crossed, and each unique data type was replicated 100 times (96 unique conditions, total of 9600 data sets). *K*-means is used to generate starting values for the EM algorithm (Steinley, 2006). For each data set, a mixture is fit using the R package **mixture** (Browne, ElSherbiny, & McNicholas, 2015) with 1-5 clusters and each of the 14 covariance matrix decompositions (i.e., 70 mixture models per data set or 672,000 mixture models total). Initial values are created by randomly generating posterior probabilities for each individual in the data (analogous to random partitioning), and calculating the first set of parameter estimates using those probabilities. For each mixture model, convergence of the EM algorithm is based on the difference between the log likelihood and the asymptotic estimate of the log likelihood at that iteration or Aitken acceleration (Aitken, 1926). If this difference is less than 1E-8, the algorithm is determined to have converged. Then, each fit statistic is calculated using the same results. That is, for each generated dataset and each fit index, we obtain results in the form of an 5×14 matrix (1-5 clusters, 1-14 covariance models). For instance, the matrix of BICs for each model on the same dataset is in the form of:

$$\boldsymbol{BIC} = \begin{bmatrix} BIC(\widehat{\boldsymbol{\Psi}}_{1,1}) & \cdots & BIC(\widehat{\boldsymbol{\Psi}}_{1,5}) \\ \vdots & \ddots & \vdots \\ BIC(\widehat{\boldsymbol{\Psi}}_{14,1}) & \cdots & BIC(\widehat{\boldsymbol{\Psi}}_{14,5}) \end{bmatrix}$$

An analogous matrix is created for each of the 14 fit indices examined.

The first set of results focus on the accuracy of the indices to select the correct number of classes, in the case when the correct covariance model is fit (i.e., using the row in the matrix above that corresponds to the true generated covariance model and comparing values within that row vector). The second set of results examine the accuracy of the statistics in selecting the covariance model, when the number of classes is correctly specified (i.e., finding the optimal value only within column that corresponds to the true generated number of clusters). The third set of results examine the accuracy of the classifications in the solution selected between all numbers of clusters and covariance models (i.e., finding the optimal value in the entire matrix).

As is common in practice, an observation's group membership was determined by selecting the group for which the posterior probability was the greatest. Accuracy is assessed by comparing the classification results of the selected model and the true generated partition using the Adjusted Rand Index or ARI (Hubert & Arabie, 1985), formulated as follows:

$$ARI = \frac{\binom{n}{2}(a+d) - \big((a+b)(a+c)+(c+d)(b+d)\big)}{\binom{n}{2}^2 - \big((a+b)(a+c)+(c+d)(b+d)\big)}$$

where *a* is the number of pairs of individuals classified into the same group in both solutions, *d* is the number of pairs classified into different groups in both solutions, and *b* and *c* are the numbers of pairs which have discordant classifications between the two solutions (same-different and different-same, respectively). Steinley (2004) outlined conventions for the size of ARI that indicate the magnitude of agreement between the two compared solutions: >.90 indicates “excellent” recovery, >.80 “good”, >.65 “moderate”, and < .65 "poor". The ARI has a maximum of 1, indicating perfect agreement. In the following simulations, the classifications resulting from each model selected by the fit indices will be compared to the true generated cluster assignments. In this context, an ARI of 1 indicates a solution that classified every observation correctly. An ARI of 0 indicates the resulting solution agreed at random with the correct classifications.

Table 4. Accuracy Results for Individual Fit Indices: Clusters

	BIC	SSA-BIC	AIC	CAIC	SSA-CAIC	CL-AIC	AIC3	AIC-BOZ	ICL	ICL-BIC	AWE	ICOMP	PC	FHV
Overall	0.72	0.62	0.51	0.70	0.74	0.54	0.66	0.68	0.59	0.47	0.11	0.22	0.44	0.06
K														
2	0.90	0.69	0.52	0.90	0.85	0.77	0.77	0.83	0.83	0.60	0.17	0.25	0.90	0.02
3	0.59	0.58	0.50	0.56	0.65	0.36	0.58	0.58	0.41	0.38	0.08	0.20	0.11	0.09
V														
2	0.61	0.53	0.52	0.61	0.64	0.48	0.53	0.55	0.50	0.39	0.09	0.06	0.33	0.05
4	0.88	0.75	0.50	0.83	0.88	0.60	0.83	0.88	0.71	0.58	0.15	0.44	0.58	0.08
D														
10%	0.61	0.50	0.33	0.59	0.61	0.48	0.56	0.59	0.52	0.43	0.13	0.22	0.46	0.07
Equal	0.82	0.73	0.67	0.80	0.85	0.58	0.75	0.77	0.65	0.52	0.10	0.22	0.42	0.05
SS														
200	0.57	0.48	0.45	0.55	0.58	0.45	0.55	0.57	0.53	0.43	0.10	0.13	0.40	0.08
1000	0.89	0.78	0.57	0.87	0.91	0.63	0.78	0.81	0.65	0.52	0.13	0.31	0.48	0.04
O														
Low	0.84	0.67	0.56	0.84	0.84	0.82	0.75	0.82	0.86	0.86	0.18	0.14	0.54	0.07
High	0.60	0.58	0.46	0.56	0.63	0.25	0.56	0.54	0.32	0.09	0.05	0.30	0.33	0.05
M														
EII	0.79	0.71	0.66	0.76	0.82	0.63	0.74	0.76	0.66	0.47	0.13	0.18	0.53	0.05
EEE	0.61	0.53	0.39	0.58	0.61	0.42	0.53	0.58	0.42	0.42	0.11	0.11	0.39	0.03
VEV	0.76	0.63	0.47	0.76	0.79	0.55	0.71	0.71	0.68	0.53	0.11	0.37	0.39	0.11

Note: Each cell is the proportion of times the fit statistics chose a model with correct number of clusters. In all cases, the covariance model is correctly specified. *K* - Number of clusters, *V* - number of variables, *D* - Cluster density, *SS* - Sample size, *O* - Overlap, *M* - True covariance model. *Equal* - all clusters have an equal population proportion, *10%* - the smallest cluster contains 10% of the population, *EII* - equal volumes, spherical shapes, *EEE* - equal volumes, equal shapes, equal orientations, *VEV* - variable volumes, equal shapes, variable orientations

Simulation Results: Individual Statistics

The percent accuracy of the statistics in selecting the correct number of clusters is given in Table 4. Several results validate earlier comparisons; for example, the BIC is one of the most accurate statistics (72% accurate over all factor conditions), and the AIC is one of the least accurate (51% accurate). However, the most accurate statistic overall is the SSA-CAIC (74%). For nearly all statistics, the performance diminishes when the data have more clusters, more variables, equal mixing proportions, a larger sample size, and less overlap. For most statistics (BIC, CAIC, SSA-CAIC, CL-AIC, ICL, ICL-BIC, ICOMP, and PC) the performance is high for the covariance model EII condition, low for the EEE condition, and then improves in the VEV condition.

Table 5. Accuracy Results for Individual Fit Indices: Covariance Model

	BIC	SSA-BIC	AIC	CAIC	SSA-CAIC	CL-AIC	AIC3	AIC-BOZ	ICL	ICL-BIC	AWE	ICOMP	PC	FHV
Overall	0.69	0.48	0.39	0.71	0.59	0.26	0.54	0.54	0.57	0.42	0.28	0.60	0.06	0.08
K														
2	0.75	0.52	0.38	0.77	0.67	0.27	0.60	0.63	0.67	0.56	0.42	0.63	0.08	0.04
3	0.65	0.45	0.41	0.67	0.53	0.26	0.48	0.48	0.50	0.32	0.18	0.58	0.05	0.11
V														
2	0.61	0.30	0.23	0.62	0.42	0.15	0.36	0.36	0.41	0.33	0.20	0.65	0.05	0.09
4	0.81	0.73	0.63	0.83	0.81	0.42	0.77	0.79	0.79	0.54	0.40	0.52	0.08	0.06
D														
10%	0.65	0.44	0.35	0.67	0.52	0.26	0.50	0.50	0.56	0.39	0.30	0.56	0.04	0.09
Equal	0.73	0.52	0.43	0.75	0.65	0.27	0.57	0.58	0.58	0.45	0.27	0.63	0.08	0.07
SS														
200	0.65	0.43	0.38	0.65	0.52	0.28	0.53	0.55	0.53	0.38	0.25	0.45	0.02	0.08
1000	0.74	0.54	0.41	0.78	0.67	0.24	0.54	0.54	0.61	0.46	0.31	0.76	0.11	0.07
O														
Low	0.70	0.47	0.42	0.70	0.65	0.42	0.54	0.56	0.72	0.70	0.19	0.60	0.09	0.14
High	0.68	0.49	0.37	0.72	0.53	0.11	0.53	0.53	0.42	0.14	0.37	0.60	0.04	0.02
M														
EII	0.89	0.58	0.47	0.89	0.76	0.32	0.68	0.68	0.74	0.50	0.32	0.92	0.08	0.16
EEE	0.79	0.42	0.32	0.79	0.55	0.24	0.47	0.47	0.63	0.50	0.45	0.58	0.00	0.00
VEV	0.39	0.45	0.39	0.45	0.45	0.24	0.45	0.47	0.34	0.26	0.08	0.29	0.11	0.08

Note: Each cell is the proportion of times the fit statistics chose a model with correct covariance model. In all cases, the number of clusters is correctly specified. *K* - Number of clusters, *V* - number of variables, *D* - Cluster density, *SS* - Sample size, *O* - Overlap, *M* - True covariance model. *Equal* - all clusters have an equal population proportion, *10%* - the smallest cluster contains 10% of the population, *EII* - equal volumes, spherical shapes, *EEE* - equal volumes, equal shapes, equal orientations, *VEV* - variable volumes, equal shapes, variable orientations

If we alternatively use the statistics to select amongst the covariance models, assuming the true number of clusters is known, we see a few important differences in the performances of the techniques (proportion accurate over each factor condition is given in Table 5). The BIC is quite accurate (69% overall), and the AIC is again one of the least accurate (39% overall). However, the statistic that performed the best overall is now the CAIC (71%) rather than the SSA-CAIC. Nearly all statistics' accuracy decreases with increasing clusters, decreasing variables, a small population proportion, a small sample size, and high overlap. In contrast to the accuracy in selecting the right number of clusters, however, nearly all statistics decrease in accuracy as the generated covariance model becomes less homogeneous. The most notable exception being the ICL, which has the best accuracy at the EEE-generated covariance model.

Table 6. Accuracy Results for Individual Fit Indices: Classifications

	BIC	SSA-BIC	AIC	CAIC	SSA-CAIC	CL-AIC	AIC3	AIC-BOZ	ICL	ICL-BIC	AWE	ICOMP	PC	FHV
Overall	0.74	0.71	0.65	0.74	0.76	0.60	0.75	0.76	0.62	0.58	0.48	0.60	0.60	0.58
K														
2	0.76	0.70	0.61	0.76	0.78	0.58	0.77	0.78	0.61	0.54	0.47	0.51	0.71	0.56
3	0.72	0.72	0.69	0.72	0.74	0.63	0.74	0.74	0.63	0.60	0.48	0.66	0.52	0.59
V														
2	0.74	0.73	0.68	0.75	0.76	0.64	0.75	0.76	0.66	0.60	0.49	0.61	0.59	0.57
4	0.73	0.69	0.61	0.72	0.75	0.56	0.74	0.75	0.58	0.54	0.45	0.57	0.60	0.59
D														
10%	0.74	0.70	0.62	0.74	0.76	0.59	0.76	0.76	0.61	0.58	0.49	0.60	0.58	0.52
Equal	0.73	0.72	0.68	0.73	0.75	0.62	0.74	0.75	0.64	0.57	0.46	0.60	0.61	0.62
SS														
200	0.70	0.65	0.63	0.70	0.73	0.59	0.73	0.73	0.62	0.59	0.48	0.58	0.58	0.55
1000	0.78	0.78	0.67	0.78	0.78	0.62	0.78	0.78	0.63	0.56	0.47	0.61	0.61	0.60
O														
Low	0.98	0.93	0.87	0.99	0.97	0.95	0.97	0.98	0.96	0.96	0.89	0.80	0.78	0.76
High	0.49	0.49	0.43	0.48	0.54	0.26	0.53	0.53	0.28	0.19	0.06	0.39	0.41	0.39
M														
EII	0.78	0.70	0.63	0.77	0.78	0.59	0.77	0.77	0.62	0.59	0.50	0.68	0.63	0.61
EEE	0.68	0.71	0.65	0.68	0.74	0.54	0.74	0.74	0.54	0.51	0.50	0.56	0.62	0.55
VEV	0.75	0.72	0.67	0.75	0.75	0.69	0.74	0.76	0.72	0.62	0.43	0.54	0.54	0.56

Note: Each cell is the average ARI. In all cases, the number of clusters is correctly specified. *K* - Number of clusters, *V* - number of variables, *D* - Cluster density, *SS* - Sample size, *O* - Overlap, *M* - True covariance model. *Equal* - all clusters have an equal population proportion, *10%* - the smallest cluster contains 10% of the population, *EII* - equal volumes, spherical shapes, *EEE* - equal volumes, equal shapes, equal orientations, *VEV* - variable volumes, equal shapes, variable orientations

Finally, the accuracy of these statistics was assessed in selecting a model with accurate classifications, without assuming the true number of clusters or the true covariance model are known. Although statistics that find accurate classifications are likely to correspond closely to those that are accurate in estimating the number of clusters and the covariance model, statistics that overfit the covariance parameters may be at an advantage, as the model may be flexible to allow classifications to be closer to the generated structure.

Table 6 shows the average ARI of each fit statistic where the model was able to choose between K = 1, ..., 5, and all 14 models (i.e., all 70 possible models). The best average ARI over all factor conditions was ARI = .76, which was obtained by the SSA-CAIC and the AIC-BOZ, with the AIC3 having a nearly as high average ARI (.75). The AIC has a much lower overall average ARI (.65). The generated covariance model had far less of an effect on the average ARI than it did on the accuracy in the previous results.

Additionally, the overlap of the clusters had a far stronger effect on the ARI than it did on the accuracy of selection of the number of clusters and the covariance model.

Simulation Comparison: Majority Vote

Individually, the statistics are shown above to have moderate to good accuracy in selecting a model. However, as discussed earlier, researchers will typically present many indices for several competing models, and somehow combine the information provided by the results to choose a model. To test systematically whether more fit indices will truly indicate a better model than indices individually, I create a “majority vote” model selection heuristic, where each fit index “votes” for a candidate model, and the model with the most votes is chosen. Majority vote is examined two ways, first by allowing each of the 14 statistics examined in this simulation to have a vote. As this requires the use of fit indices not typically used in practice, the second majority vote will be calculated only using those statistics that are commonly implemented (AIC, BIC, and SSA-BIC). This will be referred to as the “most popular” majority vote. In the case when no statistics agree on a model, the BIC is used to select the model. The BIC was chosen as the deciding vote for two reasons: (1) it is the most commonly used fit index in mixture modeling, thus in case there was no clearly optimal model in empirical research this would be a likely fit index to use, and (2) it was one of the most accurate of the three indices in the majority vote heuristic based on the above simulation.

Simulation Results: Majority Vote

The proportion of datasets for which each majority vote heuristic accurately selected the number of clusters (assuming the covariance model is known) is given in the leftmost panels in Table 7. The average accuracy of all the statistics (.69) is lower than the accuracy of the most popular statistics (.72). Both statistics have diminished performance when there are more clusters, fewer variables, one small cluster, and more overlap. However, high overlap tends to impact the performance of the heuristic using the most popular statistics to a lesser degree than using all 14 criteria.

Table 7. Accuracy Results for Majority Vote

	Clusters		Covariance Model		Classifications	
	All	Most Popular	All	Most Popular	All	Most Popular
Overall	0.69	0.72	0.60	0.53	0.72	0.71
K						
2	0.88	0.90	0.69	0.50	0.72	0.70
3	0.56	0.59	0.53	0.55	0.72	0.71
V						
2	0.58	0.61	0.44	0.38	0.72	0.73
4	0.85	0.88	0.81	0.73	0.72	0.64
D						
10%	0.61	0.61	0.52	0.44	0.72	0.69
Equal	0.77	0.82	0.67	0.60	0.72	0.71
SS						
200	0.55	0.57	0.53	0.43	0.69	0.64
1000	0.85	0.89	0.67	0.63	0.76	0.78
O						
Low	0.84	0.84	0.70	0.54	0.98	0.92
High	0.54	0.60	0.49	0.51	0.46	0.48
M						
EII	0.76	0.79	0.76	0.68	0.73	0.69
EEE	0.55	0.61	0.58	0.47	0.68	0.69
VEV	0.76	0.76	0.45	0.42	0.76	0.72

Note: Clusters - the proportion of times the heuristics chose a model with correct number of clusters when the covariance model is correctly specified, *Covariance Model* - the proportion of times the heuristics chose a model with the correct covariance model when the number of clusters is correctly specified, *Classifications* - the average ARI of the model selected by the heuristics over all clusters and covariance models. *K* - Number of clusters, *V* - number of variables, *D* - Cluster density, *SS* - Sample size, *O* - Overlap, *M* - True covariance model. *Equal* - all clusters have an equal population proportion, *10%* - the smallest cluster contains 10% of the population, *EII* - equal volumes, spherical shapes, *EEE* - equal volumes, equal shapes, equal orientations, *VEV* - variable volumes, equal shapes, variable orientations

When the majority vote is used to select a covariance model, the accuracy of the heuristic using the most popular statistics is now worse than using the most popular statistics (53% using just the most popular, 60% using all). The pattern in performance, however, is nearly identical across factor conditions between the two heuristics. The exception is when the number of clusters goes from 2 to 3, where the statistics have opposite reactions to the increase in clusters in the data. When using all indices, the performance for K = 3 is worse than in the K = 2 condition, but performance improves when moving from these conditions if only the most popular fit indices are used for majority vote. Additionally, using only the most popular fit indices improves slightly on the majority vote using all indices when the overlap is high (51% accurate when only using most popular fit indices, 49% when using all indices).

Finally, the majority vote heuristic was used to select a model between all 70 unique combinations of numbers of clusters and covariance models, and the classification accuracy (measured with the Adjusted Rand Index or ARI) is measured. When the most popular indices are used to select a model, the accuracy is nearly the same (overall average ARI = .71) than when all statistics are used for model selection (.72). The ARIs again do not vary a great deal between factor condition, with the exception of the overlap level of the clusters, which has a dramatic impact on the ability for both heuristics to find adequate classifications.

Simulation Summary

The results of the simulation are summarized in Table 8, where the statistics are assigned ranks for their performance in finding the number of clusters, covariance model, and the model with accurate classifications. The ranks are averaged to create a summary measure of their performance. In addition to the statistics individually, the accuracy is compared with the majority vote results (labeled “All” for majority vote using all statistics, “Most Popular” for majority vote using just the AIC, BIC, and SSA-BIC). The SSA-CAIC and BIC tie for the highest average rank, with the CAIC coming in third. Although the “All” model selection heuristic is better than using many statistics examined in this simulation, including the SSA-BIC and the AIC, it is not in the top tier of accuracy. Additionally, using only the most popular statistics was more accurate than many statistics (average rank of 6.33). However, this is still less accurate than using only the BIC.

Table 8. Average Rank in Performance for Fit Indices and Majority Vote Heuristics

	BIC	SSA-BIC	AIC	CAIC	SSA-CAIC	CL-AIC	AIC3	AIC-BOZ
Clusters	2	8	11	4	1	10	7	6
Model	2	10	12	1	5	14	8	7
ARI	4	7	9	5	2	11	3	1
Average	*2.67*	*8.33*	*10.67*	*3.33*	*2.67*	*11.67*	*6.00*	*4.67*
	ICL	**ICL-BIC**	**AWE**	**ICOMP**	**PC**	**FHV**	**All**	**Most Popular**
Clusters	9	12	15	14	13	16	5	2
Model	6	11	13	3	16	15	3	9
ARI	10	15	16	13	12	14	6	8
Average	*8.33*	*12.67*	*14.67*	*10.00*	*13.67*	*15.00*	*4.67*	*6.33*

Note: Each cell is the rank in performance for the statistic evaluated on selecting the right number of clusters, the right covariance model, and accurate classifications (measured with the Adjusted Rand Index). *Average* - Average rank over all three evaluations

Conclusion

This chapter systematically examined the common model selection heuristic implemented by psychologists attempting to choose a mixture model solution: combining the results of several fit indices to choose a model that optimizes the most fit indices. To do this, I calculated two "majority vote" model selection heuristics: one using a variety of statistics used for model selection in psychology, statistics, and machine learning, and another using only the statistics most likely to be presented in an empirical analysis. This is the first time such a wide range of statistics were combined to compare performance in a single study (note the sparseness of Table 2) providing a unification of different aspects of the mixture modeling literature and allowing for a broader comparison between fit indices. I found that the most accurate individual statistics are the SSA-CAIC (Yang & Yang, 2006) and the BIC (Schwarz, 1978). This is a significant finding because the SSA-CAIC is rarely studied, and is not included in popular software for fitting mixture models (M*plus*, Muthen & Muthen, 2012; R-package **mixture**, Browne, ElShirbiny, & McNicholas, 2014; LatentGOLD, Vermunt & Magidson, 2005).

More significantly, I found that the use of combining the results of many indices to select a model that is optimal on many statistics does not improve the performance of model selection. However, it is important to note that this was a relatively difficult test: I included many indices in the determination of majority vote that are known to not be accurate (i.e., the AIC in the "Most Popular" majority vote, the AWE, PC, and FHV in "All" majority vote). The performance of majority vote could undoubtedly be improved by using only those optimal statistics (i.e., inclusion of SSA-CAIC, CAIC, and AIC-BOZ). Furthermore, the model selection could potentially be improved by considering test-based models selection statistics, like the BLRT and LMR-LRT. These are topics of future research in this subject.

Change in Fit

In addition to selecting models based on a group of fit indices, another common heuristic by which psychologists select a mixture model is to examine the difference in fit between two competing models. Typically, this is done using the BIC (Au et al., 2013; Chen, 2012b; Giang & Graham, 2008; Miller et al., 2009; Rajendran et al., 2015; Scheier et al., 2008; Vaughn et al., 2007). Raftery (1995) discussed statistical issues related to this practice—more specifically, when researchers present a large number of models and several appear to have adequate fit. When models that are being compared are non-nested (the case for mixture models with different numbers of clusters), this determination is even more difficult, as there is no standard “null” hypothesis to gauge evidence against. Raftery warned against selecting the best model strictly using *p*-values, but rather to take into account the inherent uncertainty involved in model selection by use of a Bayesian style of model selection. In a Bayesian model selection process, belief in a current model is weighed against the evidence from the data at hand, and the uncertainty in both are combined to adjust, rather than wholly replace, the current model.

If there are several models to compare (the preponderance of analyses using mixture models in psychology), Raftery suggests that a large set of models should be reduced by excluding models if their BIC difference is too large from the optimal model found and, additionally, if they are not parsimonious (referred to as “Occam’s window”). He shows that this is more accurate than using *p*-values from a traditional likelihood ratio test to select a model. Raftery, however, lines out these rules for model selection in the case of linear regression and structural equation models, and the conditions that are required for this theory[2] are not satisfied by mixture models (McLachlan & Peel, 2000).

The same conditions that are required for using the difference in BIC are required for the theoretical underpinnings of using the BIC itself for model selection. Regardless, the BIC has been shown to be quite efficacious in model selection, translating across studies and fields, and even evidenced in the previous chapter. Therefore, the lack of

[2] More specifically, in Raftery (1995) section 4.1, the expansion to equation (15) relies on the parameters to be identifiable, which is not satisfied by a mixture model. Also see McLachlan & Peel (2000, Ch 6.9.3).

theoretical justification may not impede the ability for the difference in BIC to be a useful model selection heuristic.

The following simulation systematically examines the common model selection heuristic of selecting between models that have similar fit values on two different dimensions: (1) similarity in classifications of the sample, and (2) similarity in the accuracy of the classifications of the sample. If there is a relationship in either of these dimensions, then it can be used to create similar thresholds as those outlined in Raftery (1995) that indicate that the change in BIC is not large enough to constitute selection of the more parameterized over the less parameterized model, as the two are either likely to be classifying the sample similarly, or of similar classification accuracy.

Simulation

To systematically test the impact of selecting between solutions with a small difference in fit, I conduct a large simulation. The data are again generated using the R package **MixSim** (Melnykov et al., 2012). To determine whether data conditions impact the relationship between the fit of the model and the similarity of classifications, the following data conditions are varied: clusters in the data (*C* = 2, 3, or 4), variables (*V* = 4, 8, or 16), overlap of the clusters (*O* = .001, .2, or.3, corresponding to “low”, “high”, or “very high” overlap), sample size (*SS* = 200 or 1000). All clusters were generated with heterogeneous shapes and sizes and with clusters of equal density (i.e., $\boldsymbol{\pi} = \{\pi_1 = \pi_2 = \cdots \pi_K = 1/K\}$). To generalize the results here to those from a variety of software programs, several factors in the analysis are varied, generated to be reminiscent of a variety of mixture model routines in popular software, including SAS®, M*plus* (Muthén

& Muthén, 2007), and the R package **mclust** (Fraley & Raftery, 2006). The analysis factors varied are: convergence tolerance of the EM algorithm (T = 1E-16, 1E-08, or 1E-06), initialization technique (I = K-means or random initialization), and the maximum number of iterations of the EM algorithm (MI = 50, 100, 500, or 1000).

To simplify the comparison between models, all covariance matrices will be fully unconstrained. These mixture models are more easily fit using R package **EMMIX** (Mclachlan, Wang, Ng, & Peel, 2013). A mixture model is fit for 1-5 clusters within each data set and analysis condition, or 6480 results total, using the best of 100 initializations (i.e., 648000 mixture models fit total).

The Adjusted Rand Index (ARI) is calculated in two different capacities in the following simulation: (1) between each pairwise solution on each dataset, and (2) between the solutions and the true classifications. In the previous simulation, the ARI was interpreted as the accuracy of a certain solution. When the two solutions being compared are two competing mixture model solutions on the same data, the ARI is interpreted as the agreement between the partitions, and will be referred to as $ARI_{i,j}$. When the ARI is calculated using the true classifications as the alternative partition, is interpreted as in the previous simulation: the accuracy of classifications. This will be referred to as $ARI_{i,True}$. The 1-class solution is omitted from the results as its classification correspondence with any other solution is always necessarily zero[3]. The absolute difference in $ARI_{i,True}$ is calculated for every pairwise solution for each data set:

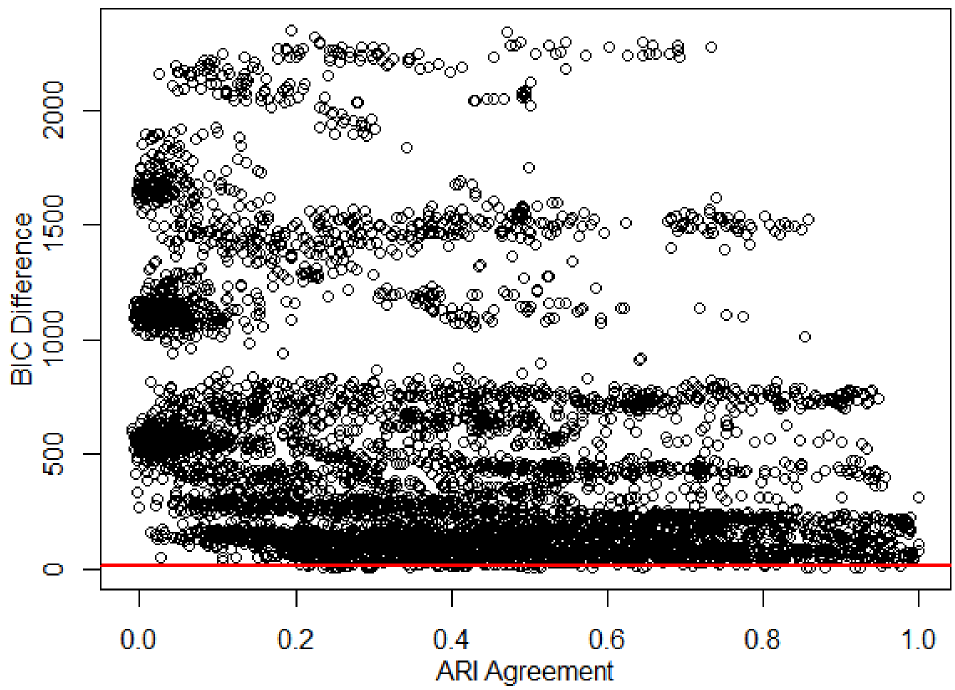
$$\Delta ARI = |ARI_{i,True} - ARI_{j,True}|$$

[3] See Appendix for proof

Additionally, the absolute pairwise difference in BIC is calculated for every model i and j ($i = 2, \dots, 5; j = 2, \dots, 5; \forall\, i \neq j$):

$$\Delta BIC = \left|BIC_i - BIC_j\right|$$

Figure 1. BIC and ARI Agreement



Simulation Results

Overall, the correlation between $ARI_{i,j}$ and ΔBIC is moderately negative (-.31), which indeed indicates that a smaller difference between BIC values is indicative of a larger agreement between the two solutions. However, Figure 1 displays the plot of every ΔBIC with every $ARI_{i,j}$, with a line indicating those solutions which would be retained using Raftery's Occam's window technique. It is clear from examining this plot that, among those solutions with small BIC differences, the solutions range from perfect agreement ($ARI_{i,j} = 1$) to near random agreement ($ARI_{i,j} = .2$).

Figure 2. BIC and ARI Difference

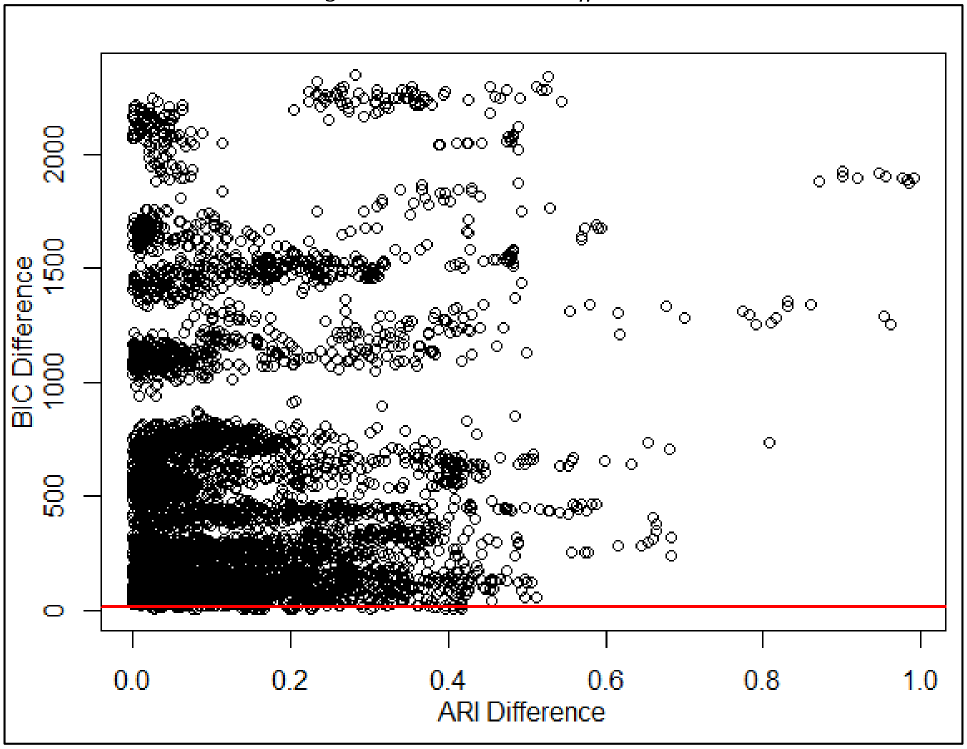
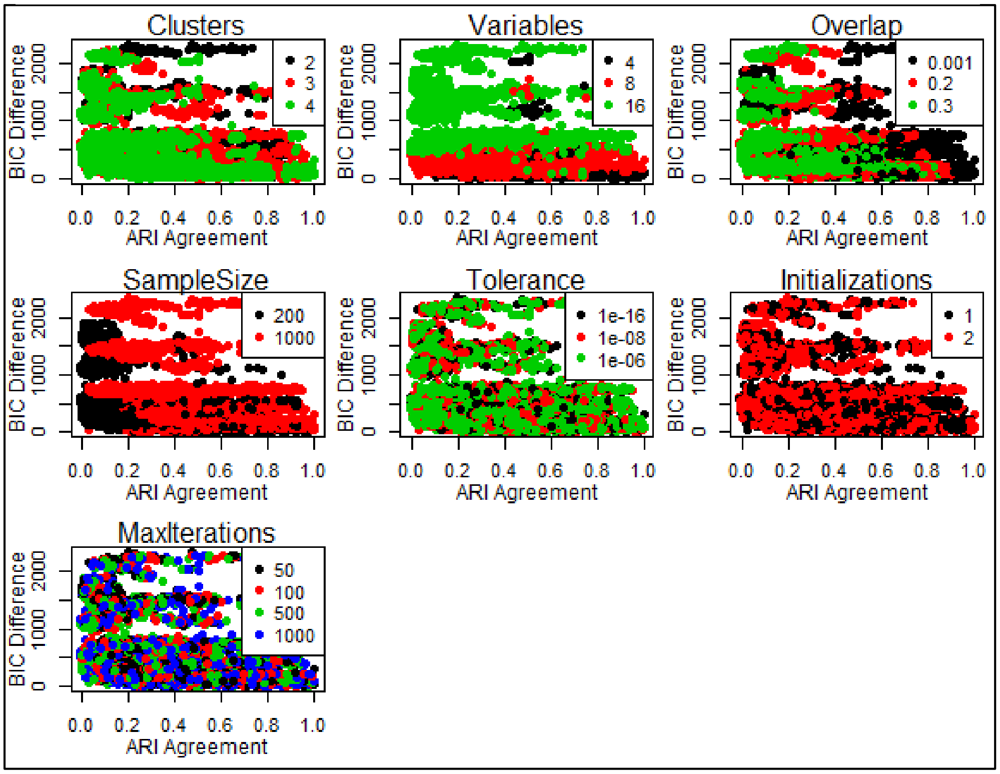


Figure 2 displays the ΔARI, or the difference in classification accuracy between the two solutions, with the ΔBIC. The correlation between these two measures is small and positive (.16), meaning that there is a slight tendency for the solutions with larger BIC differences to have differentially accurate solutions. However, it is influenced by the fact that nearly all the solutions have small differences in ARI. Furthermore, among those solutions with small ARI differences, the BIC differences range from very small (less than 10) to extremely large (over 2000).

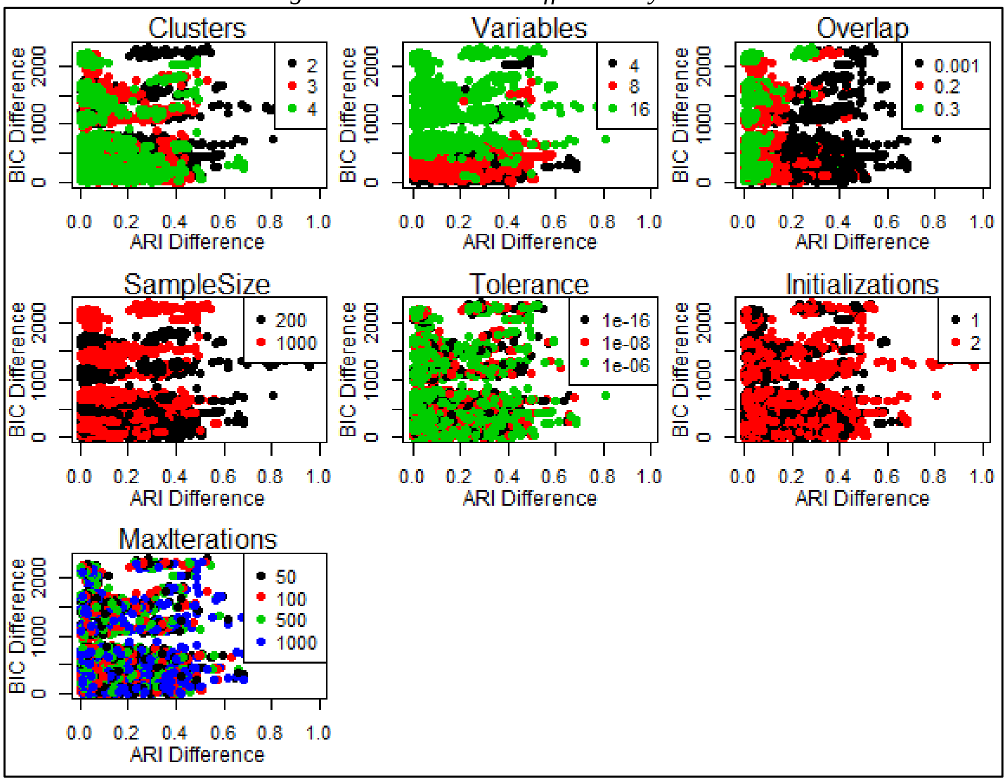
To determine whether these relationships are partially due to data or analysis conditions, Figure 3 presents the $ARI_{i,j}$ and ΔBIC by factor condition, and Figure 4 the ΔARI and ΔBIC by factor condition. For nearly all conditions, there is no separation between different levels of factor conditions (number of clusters, tolerance, type of initialization, and maximum iterations). For the others, there is some separation between conditions. For instance, when the data contain 16 variables, most of the BIC differences are quite large (>500). This is likely due to the BIC taking a greater parsimony penalty for the highly parameterized 16-variable cluster model, where consecutive models have a difference in parameters of $K + K * p + \frac{Kp(p+1)}{2} = 153 * k$, whereas the 4-variable data have an increase in parameters of only $20 * k$. Additionally, when the overlap is low, there is greater agreement between solutions and greater ARI differences. This is likely due to the

Figure 3. BIC and ARI Agreement by Factor



fact that the solutions are better on average when the overlap is low, making a more dramatic difference between the correctly and incorrectly specified solutions.

Figure 4. BIC and ARI Difference by Factor



Empirical Data Demonstration

To determine whether the same patterns that were found in the simulation replicate in a set of real data, the same analyses as above were conducted on five real datasets. All datasets are available for download at the University of California-Irvine Machine Learning Repository (*archive.ics.uci.edu*, Lichman, 2013).

Crabs. The Leptograpsus Crabs dataset (Campbell & Mahon, 1974) contains five measurements on 200 crabs corresponding to measurements in millimeters of: (1) width of frontal lip, (2) rear width, (3) length along the mid-line of the carapace, (4) maximum width of the carapace, and (4) body depth (Campbell & Mahon, 1974). These crabs are known to come from four groups corresponding to (a) orange males, (b) orange females, (c) blue males, and (d) blue females (K = 4, $\boldsymbol{\pi}$ = [.25 .25 .25 .25]).

Iris. The iris data is a well-known dataset which is used to illustrate the performance of clustering algorithms (Fisher, 1936). The data contain four measurements in centimeters of: (1) sepal length, (2) sepal width, (3) petal length, and (4) petal width. The data contains three distinct species of iris: (a) *Setosa*, (b) *Versicolor*, and (c) *Virginica* on 150 flowers (K = 3, $\boldsymbol{\pi}$ = [.33 .33 .33]).

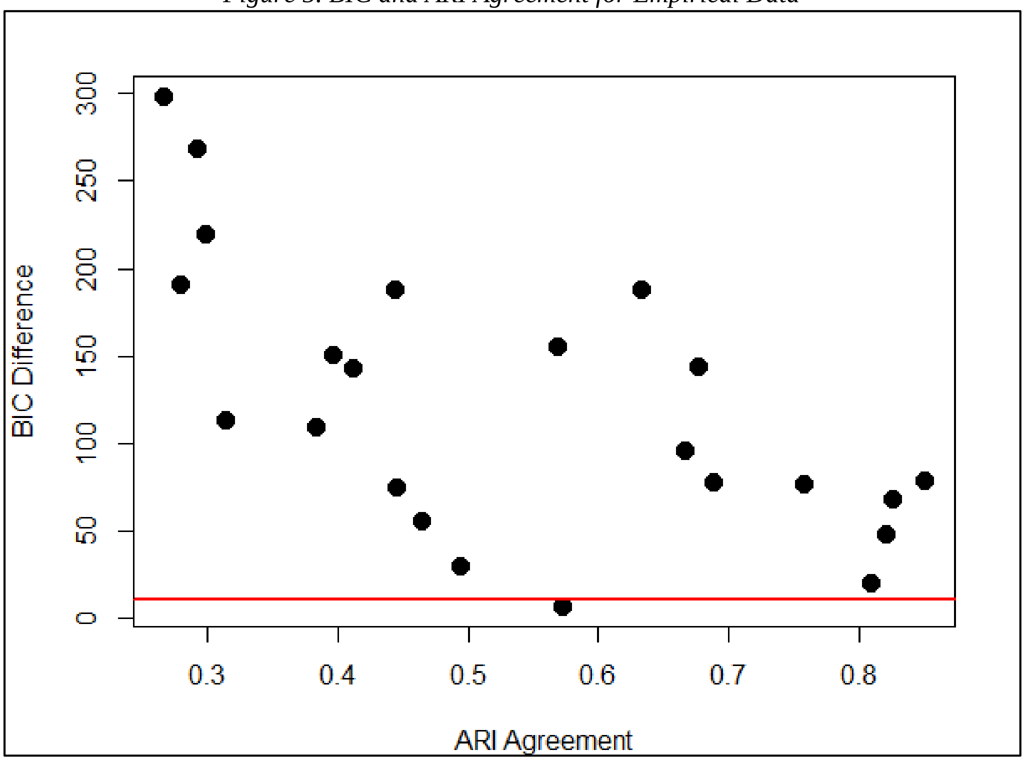
User. The user dataset (Kahraman, Sagiroglu, & Colak, 2013) contains six measures of user knowledge of a web-based task on 403 individuals: (1) study time of goal object materials, (2) repetitions of goal object materials, (3) study time of related object materials, (4) exam performance for related objects, (5) exam performance of user for goal objects. The researchers classified individuals into groups which describe their level of knowledge: (a) high, (b) middle, (c) low, and (d) very low (K = 4, $\boldsymbol{\pi}$ = [.25 .30 .32 .12]).

Seeds. The seeds dataset (Charytanowicz et al., 2010) contains measurements on 210 wheat kernels on the following attributes: (1) area (2) perimeter, (3) compactness, (4) length of kernel, (5) width of kernel, (6) asymmetry coefficient, and (7) length of kernel groove. These kernels are known to belong to three different varieties of wheat: (a) Kama, (b) Rosa, and (c) Canadian (K = 3, $\boldsymbol{\pi}$ = [.33 .33 .33]).

Empirical Data Demonstration Results

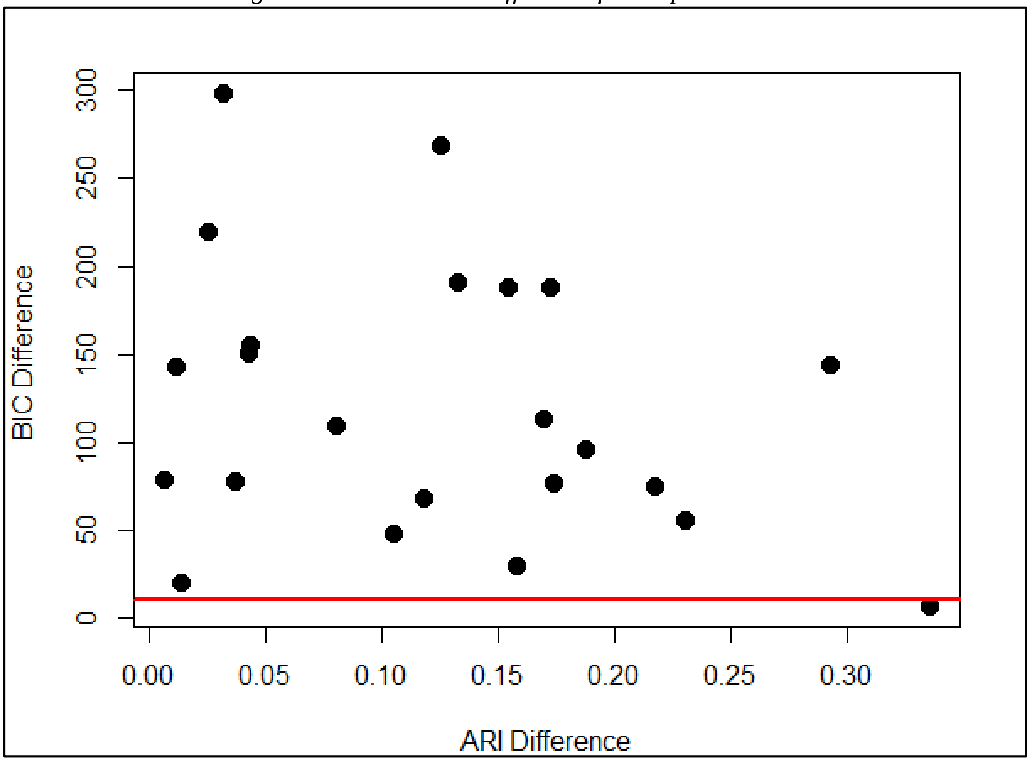
Contrary to the simulation results, correlation between $ARI_{i,j}$ and ΔBIC in the empirical data results is strongly negative (-.66), which recall is the direction that this relationship is hypothesized to be in (i.e., a greater difference in BIC leads to a smaller agreement in the classifications). Figure 5 plots the classification agreement with the change in BIC, and there is a clear downward trend as the agreement between the solutions increases. Additionally, there are no large outliers in these data (see Figure 5), but if we use the same threshold as was used in the simulation to indicate "small" differences between solutions (i.e., $\Delta BIC < 10$), then there is only one solution under this threshold, with an agreement of around .60.

Figure 5. BIC and ARI Agreement for Empirical Data



The relationship between the absolute difference in ARI with the true solutions, ΔARI, and ΔBIC is small-moderate and negative, (correlation of -.31), which is against the hypothesized direction of the relationship (see Figure 6). That is, as models become more differentiated, they are more likely to have similar fit values. This relationship, however, does not show the distinct negative relationship as the $ARI_{i,j}$. The solution that is under the BIC difference, which should indicate very similarly-classified solutions, actually has the largest difference in solution accuracy (.34).

Figure 6. BIC and ARI Difference for Empirical Data



Conclusion

This chapter systematically examined the efficacy of the commonly used model selection heuristic whereby a large class of competing models are reduced based on similarity to the optimal BIC. A large simulation and data demonstration examined whether a small difference in BIC is indicative of either (1) a solution that similarly classifies the sample or (2) a solution that is similar in its accuracy. In both the simulated and real data, there was shown to be little to no relationship between the BIC difference between solutions and the classification agreement or the accuracy of the classifications in the results. In fact, there are several cases when a small BIC difference was associated with large, rather than small, differences in accuracy of the solution, which is in direct opposition to the practice of model selection in the social sciences.

Model Nonconvergence and Local Optima

As was discussed previously, psychologists will frequently omit certain models from consideration due to convergence or local optima issues. Omitting models due to non-convergence would be a very useful heuristic to determine the number of clusters in the data, as it would not require comparison of fit by model selection statistics, some of which were just shown to have limited accuracy in some data conditions. There is some evidence that, when the model is correctly specified, the best solution tends to happen more often (Hipp & Bauer, 2006). This may be a single component of an overall easier estimation process. However, the same simulation found that the number of local optima increases with increasing numbers of clusters fit to the data. Additionally, there is evidence that certain data conditions like the sample size and overlap of the clusters can affect the number of local optima found by the mixture model (Shireman, Steinley, & Brusco, 2016a). Therefore, it may be difficult to separate the increase in estimation difficulty that is due to data and analysis conditions to estimation difficulty that is due to model misspecification.

In a similar theme as the previous chapters, I will systematically examine the commonly used model omission technique of removing solutions for which there is an increased difficulty in estimation. Two forms of model estimation difficulty are examined: (1) an increased number of iterations until convergence, and (2) an increased number of local optima.

Simulation

Data are generated identically to the simulation examining the difference in fit, using the R package **MixSim** (Melnykov et al., 2012). The following data conditions are varied, which are all consistent from the previous simulation: true clusters in the data (*C* = 2, 3, or 4), variables (*V* = 4, 8, or 16), overlap of the clusters (*O* = .001, .2, or.3, corresponding to “low”, “high”, or “very high” overlap), sample size (*SS* = 200 or 1000). Additionally, several factors in the analysis are varied: convergence tolerance of the EM algorithm (*T* = 1E-16, 1E-08, or 1E-06), initialization technique (*I* = *K*-means or random initialization), and the maximum number of iterations of the EM algorithm (*MI* = 50, 100, 500, or 1000). Mixture models will be fit with 1-5 clusters again using the R-package **EMMIX** (Mclachlan et al., 2013) assuming fully unconstrained covariance structures.

To test the degree to which local optima and the iterations until convergence can indicate that the model is misspecified, two statistics are calculated that relate to the change in the iterations until convergence and the number of local optima:

$$\Delta IT_k = |\#\, Iterations_k - \#\, Iterations_{k+1}|$$

$$\Delta LO_k = |\#\, Local\ Optima_k - \#\, Local\ Optima_{k+1}|$$

where k refers to the number of clusters ($k = \{2, \ldots, 5\}$) and $|\cdot|$ indicates the absolute value. $\#\, Iterations_k$ is the number of iterations of the EM algorithm until the difference in log likelihood from the previous to current iteration is less than the specified tolerance. The number of iterations until convergence was used to approximate a situation in which the maximum iterations is reached in the model estimation. $\#\,Local\ Optima_k$ is the number of local optima in the k-cluster model over 100 initializations. A solution is considered a local optimum when the log likelihood after convergence is unique, up to machine-precision equality or 1E-16.

A model will be selected by maximizing the above measures, and will select a model based on the degree to which model estimation becomes more difficult as the number of clusters moves from the current value to one greater. Both measures correct for the natural tendency for the mixture model to have increased difficulty as the number of clusters increases, due to the fact that each measure is calculated relating the current number of clusters to one greater, rather than comparing a model with a small number of clusters to one with many. To assess the degree to which examining the estimation difficulty of the model improves on current techniques for model selection, results using the new model selection heuristics will be tested in comparison to the Bayesian Information Criterion or BIC (Schwarz, 1978). This statistic was selected not only because it is one of the most commonly used, but also because it was shown above to be one of the most accurate in selecting a model.

Results

The results by factor are given in Table 9. Overall, the ΔLO performs better than either the BIC or the ΔIT (55% accurate for ΔLO, 52% accurate for BIC, and 45% accurate for ΔIT). No statistics were substantially impacted by the different analysis conditions (i.e., the convergence tolerance, type of initialization, and maximum iterations of the EM algorithm). However, several data conditions impact the performance of the

Table 9. Misspecification Model Selection: Results by Factor

	ΔIT	ΔLO	BIC
Overall	0.45	0.55	0.52
K			
2	0.66	0.88	1.00
3	0.38	0.37	0.28
4	0.30	0.39	0.28
V			
4	0.43	0.61	0.62
8	0.45	0.64	0.50
16	0.47	0.39	0.44
O			
Low	0.52	0.77	0.78
High	0.42	0.49	0.44
Very High	0.40	0.39	0.33
SS			
200	0.36	0.45	0.41
1000	0.53	0.65	0.63
T			
1.E-16	0.44	0.54	0.52
1.E-08	0.44	0.55	0.52
1.E-06	0.46	0.56	0.52
I			
K-means	0.44	0.54	0.52
Random	0.46	0.55	0.52
MI			
50	0.42	0.55	0.52
100	0.49	0.54	0.52
500	0.48	0.56	0.52
1000	0.40	0.54	0.52

Note: Each cell is the percent of times the model selection heuristic selected the right number of clusters. ΔIT - Change in iterations until convergence, ΔLO - change in local optima, *K* - Number of classes, *V* - Number of variables, *O* - Overlap of clusters, *SS* - sample size, *T* - convergence tolerance, *I* -

measures. All have improved performance *K* = 2, when the overlap of the clusters is low, and when the sample size is high. However, the measures have dissimilar relationships with the number of variables in the data—ΔIT increases in performance slightly as the number of variables increases, BIC decreases in performance substantially, and ΔLO performs well when *V* = 4, has slightly improved performance when *V* = 8, and has a significant decrease in performance when *V* = 16.

Empirical Data Demonstration

To examine the degree to which the patterns evidenced in the simulated data extend to real data, the same four empirical datasets that were analyzed in the previous chapter are utilized here: the Crabs (Campbell & Mahon, 1974), Iris (Fisher, 1936), User (Kahraman et al., 2013), and Seeds (Charytanowicz et al., 2010) datasets. Models were fit with 2-5 clusters in each dataset, and since the analysis conditions were shown to not impact the performance of each the model selection heuristics, the analysis conditions remained consistent with convergence tolerance of 1E-08, *K*-means initialization, and maximum iterations of 500.

Table 10. Misspecification Model Selection: Empirical Data Results

	True K	ΔIT	ΔLO	BIC
User	4	3	3	3
Seeds	3	4	3	4
Iris	3	3	3	2
Crabs	4	4	2	4
Percent Accurate:		*0.50*	*0.50*	*0.25*

The empirical data demonstration results partially replicate what was seen in the simulation. The cluster selection and average accuracy of each statistic in choosing the number of clusters for each of the example datasets is given in Table 10. The ΔIT and ΔLO select the correct number of clusters 50% of the time (Iris and Crabs were correctly selected by the ΔIT, Seeds and Iris were correctly selected by the ΔLO). The BIC, however, only selects the correct model once (Crabs).

Conclusion

In the previous simulation and real data demonstration, I systematically examined the ability for the change in local optima and the change in iterations until convergence to indicate when the model has moved from the true number of classes to one more than the true number of classes. This technique was shown to be quite efficacious, and slightly more so than using only the BIC to select a model. Surprisingly, these techniques were not shown to be substantially affected by the change in analysis conditions (e.g., convergence tolerance). This suggests that this technique may be easily implemented in many software programs even using the default conditions. However, these conditions should be considered to be a best-case scenario, as in each case it was known that the true number of clusters was within the boundary of fitted numbers of clusters.

These results do not completely vindicate the removal of models due to software errors that indicate a preponderance of locally optimal solutions, like the error produced by the software M*plus* (see Chapter 1). Omitting models that receive an error message indicating all unique solutions would likely have worse performance than the results shown here. In the case when the number of local optima correctly selected the number of clusters, only half reached a level of local optima where every initialization resulted in a unique solution in the $K = 5$ condition, where nearly all datasets had monotonic increases in local optima as K increased. Therefore, observing a dataset where all initializations lead to a unique solution would have assuredly led to many false negatives, as some datasets showed so few local optima that increasing K past the true number of clusters would not have resulted in all unique solutions.

Summary

Conclusion

The previous chapters presented a three-part systematic examination of commonly used heuristics for model selection in mixture models: (1) selecting a model based on a holistic examination of several information criteria, (2) reducing the collection of competing models based on similarity with the optimal fit value found, and (3) determining when to cease adding additional clusters to the model by examining difficulties in estimation.

In psychological research, it is common to examine candidate models on several information criteria, each of which penalizes the fit of the model with a different correction for the number of estimated parameters. The researchers will typically consider all of these statistics to select a model. The motivation behind this technique may be to combine the results of statistics which tend to underfit, and some that tend to overfit, as a model for which both types of statistics tend to agree may signify overwhelming evidence for a model. In a large simulation, selecting a model based on a “majority vote”, or allowing a group of fit indices to vote for a candidate model, was shown to be less efficacious than many different fit indices used individually. This is significant because it suggests that this common practice may be a flawed way to select a final model, and researchers would be better served to ignore those statistics that are not accurate individually.

An additional common heuristic to reduce the set of potential models for consideration is to consider the difference in BIC of solutions. Raftery (1995) outlined differences in BIC that constitute differences in fit that represents “strong” versus “weak” evidence for a potential model. These thresholds rely on regularity conditions that are not satisfied by the mixture model. Considering, though, that these conditions are the same that underlie the justification for the BIC (which does not impede the BIC’s ability to select a model), there still may be efficacy in this technique. To test whether the choice between mixture models of comparable BICs represents an arbitrary selection, I conducted a large simulation and data demonstration and assessed the classification agreement between competing solutions and differences in classification accuracy. These results showed that there is little relationship between the difference in BIC and the classification agreement or the difference in classification accuracy, suggesting that there is no threshold by which the choice between mixture model solutions is arbitrary.

The last heuristic examined in this dissertation examines using the computational difficulties in the model as an indicator that the model is overspecified. Researchers will commonly consider only those models without estimation difficulties, but considering that mixture model estimation necessarily gets more difficult as the number of clusters increases, it may be a poor indicator of misspecification. To assess the degree to which model estimation difficulty indicates model misspecification, the difference in local optima and iterations until convergence is assessed for models with consecutive clusters as K increases. Using these indices was surprisingly accurate in determining the number of clusters in both a large simulation and data demonstration. In fact, using the maximum change in the number of local optima to select a model demonstrated better accuracy than the BIC.

Future Research

The previous chapters assessed three techniques of model selection for mixture modeling; however, each of these can be further extended or altered to maximize accuracy or further examine nuances of the theory and data conditions that impact the accuracy of these heuristics. For instance, in the majority vote model selection, the simulation results can be used to determine an optimal set of fit indices, which can be customized to give the most accurate group of indices given the conditions that each of these tends to be accurate. Majority vote may also be improved by use of adjusted model comparison tests, which were omitted from the comparison. Additionally, although Raftery's threshold of a BIC difference of 10 was not efficacious in separating similar from dissimilar solutions, there may be different fit indices that can find separation between solutions with accurate and inaccurate classifications. To determine whether the change in local optima and iterations until convergence are still accurate at higher numbers of clusters, a simulation comparison should be expanded.

Another future direction of this research is to examine whether these same patterns hold for extensions of the traditional finite mixture model. This includes the growth mixture model (Grimm, Steele, & Ram, 2013) and mixtures of factor analyzers (McLachlan & Peel, 2000). In these models, there is an equal level of model selection subjectivity that persists, due to the lack of clarity regarding the optimal model selection techniques. It would also be useful to determine whether these patterns persist in less parameterized mixture models, like the latent profile analysis (i.e., a constrained covariance structure) or mixtures of binary or nominal variables.

References

Adams, M. A., Sallis, J. F., Kerr, J., Conway, T. L., Saelens, B. E., Frank, L. D., … Cain, K. L. (2011). Neighborhood environment profiles related to physical activity and weight status: A latent profile analysis. *Preventive Medicine*, *52*(5), 326–331. http://doi.org/10.1016/j.ypmed.2011.02.020

Aitken, A. . (1926). On Bernoullis numerical solution of algebraic equations. *Proceedings of the Royal Society of Edinburgh*, *46*, 289–305.

Akaike, H. (1973). *Information theory and an extension of the maximum likelihood principle*. *Selected Papers of Hirotugu Akaike*. http://doi.org/10.1007/978-1-4612-1694-0

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6). http://doi.org/10.1109/TAC.1974.1100705

Andrews, J. L., McNicholas, P. D., & Subedi, S. (2011). Model-based classification via mixtures of multivariate t-distributions. *Computational Statistics and Data Analysis*, *55*(1), 520–529. http://doi.org/10.1016/j.csda.2010.05.019

Andrews, R. L., Ainslie, A., & Currim, I. S. (2002). An Empirical Comparison of Logit Choice Models with Discrete Versus Continuous Representations of Heterogeneity. *Journal of Marketing Research*, *39*(4), 479–487. http://doi.org/10.1509/jmkr.39.4.479.19124

Au, T. M., Dickstein, B. D., Comer, J. S., Salters-Pedneault, K., & Litz, B. T. (2013). Co-occurring posttraumatic stress and depression symptoms after sexual assault: a latent profile analysis. *Journal of Affective Disorders*, *149*(1-3), 209–216. http://doi.org/10.1016/j.jad.2013.01.026

Banfield, J., & Raftery, A. (1993). Model-based Gaussian and Non-Gaussian Clustering. *Biometrics*. http://doi.org/10.2307/2532201

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

Bezek, J. C. (1981). *Pattern recognition with Fuzzy Objective Function Algorithms.* New York: Plenum.

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., Celeux, G., & Govaert, G. (2003). Choosing starting values for the EM algorithm for getting the highest likehood in multivariate Gaussian mixture models. *Computational Statistics and Data Analysis*, *41*(3-4), 561–575. http://doi.org/10.1016/S0167-9473(02)00163-9

Biernacki, C., & Govaert, G. (1997). Biernacki & Govaert (1997b).pdf. *Computing Science and Statistics*, 451–457.

Biernacki, C., & Govaert, G. (1997). Using the classification likelihood to choose the number of clusters. In *Computing Science and Statistics: Proceedings of the 28th Symposium on the Interface* (pp. 451–457).

Biernacki, C., & Govaert, G. (1999). Choosing models in model-based clustering and discriminant analysis. *Journal of Statistical Computation and Simulation*, *64*(1), 49–71. http://doi.org/10.1080/00949659908811966

Borden, L. a., Herman, K. C., Stormont, M., Goel, N., Darney, D., Reinke, W. M., &

Webster-Stratton, C. (2014). Latent profile analysis of observed parenting behaviors in a clinic sample. *Journal of Abnormal Child Psychology*, *42*, 731–742. http://doi.org/10.1007/s10802-013-9815-z

Bozdogan, H. (1987). Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, *52*(3), 345–370. http://doi.org/10.1007/BF02294361

Bozdogan, H. (1990). On the information-based measure of covariance complexity and its application to the evaluation of multivariate linear models. *Communications in Statistics - Theory and Methodology*, *19*(1), 221–278.

Bozdogan, H. (1993). Choosing the number of component clusters in the mixture-model using a new informational complexity criterion of the inverse-Fisher information matrix. *Information and Classification*, (1988), 40–54. http://doi.org/10.1007/978-3-642-50974-2_5

Bozdogan, H. (2000). Akaike's Information Criterion and Recent Developments in Information Complexity. *Journal of Mathematical Psychology*, *44*(1), 62–91. http://doi.org/10.1006/jmps.1999.1277

Bozdogan, H., & Sclove, S. L. (1984). Multi-sample cluster analysis using Akaike's Information Criterion. *Annals of the Institute of Statistical Mathematics*, *36*, 163–180.

Browne, R., ElSherbiny, A., & McNicholas, P. D. (2015). Package " mixture ."

Campbell, N., & Mahon, R. (1974). A multivariate study of variation in two species of rock crab of the genus Leptograpsus. *Australian Journal of Zoology*, *22*(3), 417. http://doi.org/10.1071/ZO9740417

Celeux, G., & Govaert, G. (1995). Gaussian parsimonious clustering models. *Pattern Recognition*, *28*(5), 781–793. http://doi.org/10.1016/0031-3203(94)00125-6

Charytanowicz, M., Niewczas, J., Kulczycki, P., Kowalski, P. A., Lkasik, S., & Zak, S. (2010). A Complete Gradient Clustering Algorithm for Features Analysis of X-ray Images. In E. Pietka & J. Kawa (Eds.), *Information Technologies in Biomedicine* (pp. 15–24). Berlin-Heidelberg: Springer-Verlag.

Chen, S.-K. (2012a). Internet use and psychological well-being among college students: A latent profile approach. *Computers in Human Behavior*, *28*(6), 2219–2226. http://doi.org/10.1016/j.chb.2012.06.029

Chen, S.-K. (2012b). Internet use and psychological well-being among college students: A latent profile approach. *Computers in Human Behavior*, *28*(6), 2219–2226. http://doi.org/10.1016/j.chb.2012.06.029

Dempster, A. P., Laird, N. M., & D.B., R. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B (Methodological)*, *39*(1), 1–38.

DiStefano, C. (2006). Investigating Subtypes of Child Development: A Comparison of Cluster Analysis and Latent Class Cluster Analysis in Typology Creation. *Educational and Psychological Measurement*, *66*(5), 778–794. http://doi.org/10.1177/0013164405284033

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, *7*, 179–188.

Flensborg Damholdt, M., Shevlin, M., Borghammer, P., Larsen, L., & Østergaard, K. (2012). Clinical heterogeneity in Parkinson’s disease revisited: A latent profile analysis. *Acta Neurologica Scandinavica*, *125*(5), 311–318. http://doi.org/10.1111/j.1600-0404.2011.01561.x

Fraley, C., & Raftery, A. E. (2006). MCLUST Version 3 for R: Normal Mixture Modeling and Model-Based Clustering. *In: Report No. 504*, *Department*(504), 1–57. http://doi.org/Technical Report No. 504

Gan, L., & Jiang, J. (1999). A test for global maximum. *Journal of the American Statistical …*, *94*(447), 847–854. http://doi.org/10.2307/2669999

Gath, I., & Geva, a. B. (1989). Unsupervised optimal fuzzy clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *11*(7), 773–780. http://doi.org/10.1109/34.192473

Geiser, C., Okun, M. A., & Grano, C. (2014). Who is motivated to volunteer ? A latent profile analysis linking volunteer motivation to frequency of volunteering. *Psychological Test and Assessment Modeling*, *56*(1), 3–24.

Gerber, M., Jonsdottir, I. H., Lindwall, M., & Ahlborg, G. (2014). Physical activity in employees with differing occupational stress and mental health profiles: A latent profile analysis. *Psychology of Sport and Exercise*, *15*(6), 649–658. http://doi.org/10.1016/j.psychsport.2014.07.012

Giang, M. T., & Graham, S. (2008). Using latent class analysis to identify aggressors and victims of peer harassment. *Aggressive Behavior*, *34*(2), 203–13. http://doi.org/10.1002/ab.20233

Gomez, R., Gomez, R. M., Winther, J., & Vance, A. (2014). Latent profile analysis of working memory performance in a sample of children with ADHD. *Journal of Abnormal Child Psychology*, *42*(8), 1367–79. http://doi.org/10.1007/s10802-014-

9878-5

Grimm, K., Steele, J., & Ram, N. (2013). Exploratory Latent Growth Models in the Structural Equation Modeling Framework. *Structural Equation Modeling: A Multidisciplinary Journal*, *20*(4), 568–591. http://doi.org/10.1080/10705511.2013.824775

Grunschel, C., Patrzek, J., & Fries, S. (2013). Exploring different types of academic delayers: A latent profile analysis. *Learning and Individual Differences*, *23*, 225–233. http://doi.org/10.1016/j.lindif.2012.09.014

Hall, M. T., Howard, M. O., & McCabe, S. E. (2010). Subtypes of adolescent sedative/anxiolytic misusers: A latent profile analysis. *Addictive Behaviors*, *35*(10), 882–889. http://doi.org/10.1016/j.addbeh.2010.05.006

Hill, A. L., Degnan, K. A., Calkins, S. D., & Keane, S. P. (2006). Profiles of externalizing behavior problems for boys and girls across preschool: The roles of emotion regulation and inattention. *Developmental Psychology*, *42*(5), 913–928. http://doi.org/10.1037/0012-1649.42.5.913

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

Hori, H., Teraishi, T., Sasayama, D., Matsuo, J., Kinoshita, Y., Ota, M., … Kunugi, H. (2014). A latent profile analysis of schizotypy, temperament and character in a nonclinical population: Association with neurocognition. *Journal of Psychiatric Research*, *48*(1), 56–64. http://doi.org/10.1016/j.jpsychires.2013.10.006

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, *2*(1), 193–218. http://doi.org/10.1007/BF01908075

Kahraman, H. T., Sagiroglu, S., & Colak, I. (2013). Developing intuitive knowledge classifier and modeling of users' domain dependent data in web. *Knowledge Based Systems*, *37*, 283–295.

Keefer, K. V., Parker, J. D. a., & Wood, L. M. (2012). Trait Emotional Intelligence and University Graduation Outcomes: Using Latent Profile Analysis to Identify Students at Risk for Degree Noncompletion. *Journal of Psychoeducational Assessment*, *30*(4), 402–413. http://doi.org/10.1177/0734282912449446

Klonsky, E. D., & Olino, T. M. (2008). Identifying clinically distinct subgroups of self-injurers among young adults: a latent class analysis. *Journal of Consulting and Clinical Psychology*, *76*(1), 22–27. http://doi.org/10.1037/0022-006X.76.1.22

Lichman, M. (2013). UCI Machine Learning Repository. Retrieved March 27, 2016, from http://archive.ics.uci.edu/ml

Marsh, H. W., Lüdtke, O., Trautwein, U., & Morin, A. J. S. (2009). *Classical Latent Profile Analysis of Academic Self-Concept Dimensions: Synergy of Person- and Variable-Centered Approaches to Theoretical Models of Self-Concept*. *Structural Equation Modeling: A Multidisciplinary Journal* (Vol. 16). http://doi.org/10.1080/10705510902751010

Martinson, B. C., VazquezBenitez, G., Patnode, C. D., Hearst, M. O., Sherwood, N. E., Parker, E. D., … Lytle, L. (2011). Obesogenic family types identified through latent profile analysis. *Annals of Behavioral Medicine : A Publication of the Society of Behavioral Medicine*, *42*(2), 210–20. http://doi.org/10.1007/s12160-011-9286-9

Maynard, B. R., Salas-Wright, C. P., Vaughn, M. G., & Peters, K. E. (2012). Who Are

Truant Youth? Examining Distinctive Profiles of Truant Youth Using Latent Profile Analysis. *Journal of Youth and Adolescence*, *41*(12), 1671–1684. http://doi.org/10.1007/s10964-012-9788-1

McLachlan, G. J., & Peel, D. (2000). *Finite Mixture Models*. New York, NY: John Wiley & Sons, Inc.

Mclachlan, G. J., Wang, K., Ng, A., & Peel, D. (2013). EMMIX: The EM Algorithm and Mixture Models Version 1.0.1.

Melnykov, V., Chen, W.-C., & Maitra, R. (2012). MixSim : An R Package for Simulating Data to Study Performance of Clustering Algorithms. *Journal of Statistical Software*, *51*(12), 1–25.

Melnykov, V., & Maitra, R. (2010). Finite mixture models and model-based clustering. *Statistics Surveys*, *4*(0), 80–116. http://doi.org/10.1214/09-SS053

Merz, E. L., & Roesch, S. C. (2011). A latent profile analysis of the Five Factor Model of personality: Modeling trait interactions. *Personality and Individual Differences*, *51*(8), 915–919. http://doi.org/10.1016/j.paid.2011.07.022

Miller, H. a., Turner, K., & Henderson, C. E. (2009). Psychopathology of Sex Offenders: A Comparison of Males and Females Using Latent Profile Analysis. *Criminal Justice and Behavior*, *36*(8), 778–792. http://doi.org/10.1177/0093854809336400

Mokros, A., Hare, R. D., Neumann, C. S., Santtila, P., Habermeyer, E., Mokros, A., … Nitschke, J. (2015). Journal of Abnormal Psychology Variants of Psychopathy in Adult Male Offenders : A Latent Profile Analysis Variants of Psychopathy in Adult Male Offenders : A Latent Profile Analysis. *Journal of Abnormal Psychology*.

Morin, a. J. S., Morizot, J., Boudrias, J.-S., & Madore, I. (2011). A Multifoci Person-

Centered Perspective on Workplace Affective Commitment: A Latent Profile/Factor Mixture Analysis. *Organizational Research Methods*, *14*(1), 58–90. http://doi.org/10.1177/1094428109356476

Muthen, B. (2002). Latent variable mixture modeling: Latent profile analysis. Retrieved from http://www.statmodel.com/discussion/messages/13/115 .html?1114649917

Muthén, L., & Muthén, B. (2007). Mplus user's guide (version 7.0). *Los Angeles: Author*. Retrieved from http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Mplus+user+guide#8

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. http://doi.org/10.1080/10705510701575396

Oliver, J., Baxter, R., & Wallace, C. (1996). Unsupervised Learning Using MML. *In Machine Learning: Proceedings of the Thirteenth International Conference (ICML 96}*, 364–372.

Pastor, D. a., Barron, K. E., Miller, B. J., & Davis, S. L. (2007). A latent profile analysis of college students' achievement goal orientation. *Contemporary Educational Psychology*, *32*(1), 8–47. http://doi.org/10.1016/j.cedpsych.2006.10.003

Pastor, D. A., Barron, K. E., Miller, B. J., & Davis, S. L. (2007). A latent profile analysis of college students' achievement goal orientation. *Contemporary Educational Psychology*. *Special Issue: Applications of Latent Variable Modeling in Educational Psychology Research. Vol 32(1)*, *32*(2007), 8–47. Retrieved from http://csaweb106v.csa.com/ids70/view_record.php?id=12&recnum=0&log=from_res&SID=pfju0rofq9a6i5442dcebqnkf3&mark_id=search%3A12%3A15%2C0%2C1

Pastor, D., Barron, K., Miller, B., & Davis, S. (2007). A latent profile analysis of college students' achievement goal orientation. *Contemporary Educational Psychology*, *32*(1), 8–47. http://doi.org/10.1016/j.cedpsych.2006.10.003

Raftery, A. E. (1995). Bayesian Model Selection in Social Research. *Sociological Methodology*. http://doi.org/10.2307/271063

Rajendran, K., O'Neill, S., Marks, D. J., & Halperin, J. M. (2015). Latent profile analysis of neuropsychological measures to determine preschoolers' risk for ADHD. *Journal of Child Psychology and Psychiatry*, *56*(9), 958–965. http://doi.org/10.1111/jcpp.12434

Ram, N., & Grimm, K. J. (2009). Methods and Measures: Growth mixture modeling: A method for identifying differences in longitudinal change among unobserved groups. *International Journal of Behavioral Development*, *33*(6), 565–576. http://doi.org/10.1177/0165025409343765

Rissanen, J. (1978). Modeling by shortest data description. *Automatica*.

Roberts, S. J., Husmeier, D., Rezek, I., & Penny, W. (1998). Bayesian approaches to Gaussian mixture modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*(11), 1133–1142. http://doi.org/10.1109/34.730550

Scheier, L. M., Ben Abdallah, A., Inciardi, J. A., Copeland, J., & Cottler, L. B. (2008). Tri-city study of Ecstasy use problems: a latent class analysis. *Drug and Alcohol Dependence*, *98*(3), 249–63. http://doi.org/10.1016/j.drugalcdep.2008.06.008

Schwarz, G. (1978). Estimating a Dimension of a Model. *The Annals of Statistics*.

Shireman, E., Steinley, D., & Brusco, M. J. (2016). Examining the effect of initialization strategies on the performance of Gaussian mixture modeling. *Behavior Research Methods*. http://doi.org/10.3758/s13428-015-0697-6

Stapleton, J., Turrisi, R., Hillhouse, J., Robinson, J. K., & Abar, B. (2010). A comparison of the efficacy of an appearance-focused skin cancer intervention within indoor tanner subgroups identified by latent profile analysis. *Journal of Behavioral Medicine*, *33*(3), 181–190. http://doi.org/10.1007/s10865-009-9246-z

Steane, M. a., McNicholas, P. D., & Yada, R. Y. (2012). Model-based classification via mixtures of multivariate t-factor analyzers. *Communications in Statistics - Simulation and Computation, 41*(4), 510–523. http://doi.org/10.1080/03610918.2011.595984

Steinley, D. (2003). Local optima in K-means clustering: what you don't know may hurt you. *Psychological Methods*, *8*(3), 294–304. http://doi.org/10.1037/1082-989X.8.3.294

Steinley, D. (2004). Properties of the Hubert-Arabie adjusted Rand index. *Psychological Methods*, *9*(3), 386–396. http://doi.org/10.1037/1082-989X.9.3.386

Steinley, D. (2006). K-means clustering: a half-century synthesis. *The British Journal of Mathematical and Statistical Psychology*, *59*(Pt 1), 1–34. http://doi.org/10.1348/000711005X48266

Steinley, D., & Brusco, M. J. (2011). Evaluating mixture modeling for clustering: recommendations and cautions. *Psychological Methods*, *16*(1), 63–79. http://doi.org/10.1037/a0022673

Steinley, D., & McDonald, R. P. (2007). Examining Factor Score Distributions to

Determine the Nature of Latent Spaces. *Multivariate Behavioral Research*, *42*(1), 133–156. http://doi.org/10.1080/00273170701341217

Team, R. D. C. (2008). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.r-project.org

Turner, K., Miller, H. a., & Henderson, C. E. (2008). Latent Profile Analyses of Offense and Personality Characteristics in a Sample of Incarcerated Female Sexual Offenders. *Criminal Justice and Behavior*, *35*(7), 879–894. http://doi.org/10.1177/0093854808318922

Vaughn, M. G., DeLisi, M., Beaver, K. M., & Howard, M. O. (2008). Toward a quantitative typology of burglars: a latent profile analysis of career offenders. *Journal of Forensic Sciences*, *53*(6), 1387–92. http://doi.org/10.1111/j.1556-4029.2008.00873.x

Vaughn, M. G., Perron, B. E., & Howard, M. O. (2007). Variations in social contexts and their effect on adolescent inhalant use: a latent profile investigation. *Drug and Alcohol Dependence*, *91*(2-3), 129–33. http://doi.org/10.1016/j.drugalcdep.2007.05.012

Vermunt, J. K., & Magidson, J. (2005). Latent GOLD 4.0 User's Guide. Belmont, Massachusetts: Statistical Innovations Inc.

Wallace, C. S., & Freeman, P. R. (1987). Estimation and Inference by Compact Coding. *Journal of the Royal Statistical Society. Series B (Methodological)1*, *49*(3), 240–265.

Windham, M. P., & Cutler, a. (1992). Information Ratios For Validating Mixture

Analyses. *Journal of the American Statistical Association*, *87*(420), 1188–1192. http://doi.org/10.2307/2290659

Wolf, E., Miller, M., Reardon, A., Ryabchenko, K., Castillo, D., & Freund, R. (2012). A latent class analysis of dissociation and PTSD: evidence for a dissociative subtype. *Archives of General …*, *69*(7), 698–705. http://doi.org/10.1001/archgenpsychiatry.2011.1574.A

Xiang, T., & Gong, S. (2005). *Unsupervised Learning of Visual Context Using Completed Likelihood AIC*.

Yang, C. C., & Yang, C. C. (2007). Separating latent classes by information criteria. *Journal of Classification*, *24*(2), 183–203. http://doi.org/10.1007/s00357-007-0010-1

Appendix: ARI with One-Class Solution is Necessarily Zero

The numerator of the ARI is:

$$\binom{n}{2}(a+d)-[(a+b)(a+c)+(c+d)(b+d)$$

where a is the number of pairs of individuals classified into the same group in both solutions, d is the number of pairs classified into different groups in both solutions, and b and c are the numbers of pairs which have discordant classifications between the two solutions. When one of the partitions being compared has only one class, both d and either b or c are necessarily zero (we will assume c is zero). This simplifies the numerator to:

$$\binom{n}{2}a-[(a+b)(a+c)+(c+d)(b+d)]$$

$$\binom{n}{2}a-[(a+b)a]$$

Therefore, we need to show that $\binom{n}{2}=(a+b)$. If we rewrite a and b in terms of their generating equations, where t_{rc} is the number of objects in subsets r and c:

$$a+b=\frac{\sum_{r=1}^{R}\sum_{c=1}^{C}t_{rc}^{2}-n}{2}+\frac{\sum_{r=1}^{R}t_{r+}^{2}-\sum_{r=1}^{R}\sum_{c=1}^{C}t_{rc}^{2}}{2}$$

$$=\frac{-n+\sum_{r=1}^{R}t_{r}^{2}}{2}$$

Since the second subset contains all the observations, the equation can be further simplified to:

$$=\frac{-n+n^{2}}{2}$$

$$=\frac{n(n-1)}{2}$$

$$a+b=\binom{n}{2}$$

Therefore, the numerator of the ARI necessarily reduces to zero when one of the solutions is a one-class partition.

Vita

Emilie Shireman (neé Rausch) was born in Jacksonville. Illinois on June 5, 1989. After graduating from Jacksonville High School in 2007, she studied Psychology at the University of Missouri, receiving her BA 2010. She received an MA in Psychology in May 2013, and an MA in Statistics in December 2014, both from the University of Missouri.