Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

HHS Public Access

Author manuscript

Behav Res Methods. Author manuscript; available in PMC 2018 February 01.

Published in final edited form as:

Behav Res Methods. 2017 February ; 49(1): 282–293. doi:10.3758/s13428-015-0697-6.

Examining the effect of initialization strategies on the performance of Gaussian mixture modeling

Emilie Shireman[1], **Douglas Steinley**[1], and **Michael J. Brusco**[2]

[1]University of Missouri, Columbia, MO, USA

[2]Florida State University, Tallahassee, FL, USA

Abstract

Mixture modeling is a popular technique for identifying unobserved subpopulations (e.g., components) within a data set, with Gaussian (normal) mixture modeling being the form most widely used. Generally, the parameters of these Gaussian mixtures cannot be estimated in closed form, so estimates are typically obtained via an iterative process. The most common estimation procedure is maximum likelihood via the expectation-maximization (EM) algorithm. Like many approaches for identifying subpopulations, finite mixture modeling can suffer from locally optimal solutions, and the final parameter estimates are dependent on the initial starting values of the EM algorithm. Initial values have been shown to significantly impact the quality of the solution, and researchers have proposed several approaches for selecting the set of starting values. Five techniques for obtaining starting values that are implemented in popular software packages are compared. Their performances are assessed in terms of the following four measures: (1) the ability to find the best observed solution, (2) settling on a solution that classifies observations correctly, (3) the number of local solutions found by each technique, and (4) the speed at which the start values are obtained. On the basis of these results, a set of recommendations is provided to the user.

Keywords

Mixture modeling; EM algorithm; Algorithm initialization

Finding subgroups in data is a popular analytic goal in exploratory data analysis. Although many techniques are available to partition a data set, one of the most commonly implemented is *mixture modeling* (also known as *latent profile analysis* or *finite mixture modeling*). The most common iterative algorithm to find maximum likelihood parameters for a mixture model is the expectation-maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977), which relies upon a set of initial starting values. A poor set of starting values for the EM algorithm can significantly impact the quality of the resulting solution (Biernacki, Celeux, & Govaert, 2003; Seidel, Mosler, & Alker, 2000). Considering that mixture models also suffer from a prevalent local-optima problem (Hipp & Bauer, 2006; McLachlan & Peel, 2000; Shireman, Steinley, & Brusco, 2015), the choice of starting values in mixture modeling is paramount to finding a solution that is accurate. The goal of this

Correspondence to: Emilie Shireman.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

article is to assess the quality of a variety of commonly implemented techniques to generate starting values for estimating a mixture model.

To gain a deeper understanding of the properties of the methods used to initialize mixture models, four different characteristics of starting values will be examined: (1) the abilities of the techniques to provide values that lead to the best observed solution (noting that the true solution, or *global optimum*, is unknowable); (2) the abilities of the techniques to provide values that lead to a solution with an accurate classification of observations; (3) their propensities to find locally optimal solutions; and (4) the speeds at which the techniques provide initial values. We will begin by describing mixture modeling and the EM algorithm, followed by a description of the initialization strategies included in this study. The simulations and their results are then explained in detail, followed by a discussion.

Mixture model

Mixture modeling is the process of fitting a chosen number of probability density functions (pdfs) to data. If $\mathbf{X}$ is an $n \times p$ matrix of observed variables, then the population density,

$$f(X;\Theta)=\sum_{k=1}^{K}\pi_k f_k(X;\vartheta_k),$$

is a mixture of K independent pdfs (also referred to as *components*, *mixtures*, *classes*, or *groups*), f_k of which each contain a proportion, $\boldsymbol{\pi}_k$, of the total population. The proportions must sum to unity ($\sum_{k=1}^{K}\pi_k=1$), and each must be greater than zero (i.e., there are no empty classes; $\boldsymbol{\pi}_k > 0$, $k = 1, \ldots, k$). In the common case of mixtures of multivariate normal distributions (e.g., Gaussian mixture modeling), each f_k is the Gaussian density, and $\boldsymbol{\vartheta}_k$ contain the $p \times 1$ mean vector $\boldsymbol{\mu}_k$ and the $p \times p$ covariance matrix $\boldsymbol{\Sigma}_k$ for the kth cluster. For a full treatment of Gaussian mixture modeling and several variants, we refer readers to the definitive text by McLachlan and Peel (2000).

EM algorithm

As we mentioned above, the EM algorithm is commonly employed to obtain maximum likelihood estimates for the parameters of the mixture model (for more details on the EM algorithm, we refer readers to McLachlan & Krishnan, 1997, and McLachlan & Peel, 2000). After obtaining initial parameter estimates, the EM algorithm alternates between (a) estimating the posterior probability of an observation belonging to each mixture by assuming that the parameters are fixed (the E-step), and (b) updating estimates of the parameters by fixing the posterior probabilities of class membership (the M-step). The algorithm continues until the change in the likelihood criterion is below a set tolerance level, or until a maximum number of iterations has been reached.

The EM algorithm is guaranteed to arrive at a local maximizer of the likelihood function (i.e., a local optimum, although in rare instances the algorithm can settle on a saddle point;

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

McLachlan & Krishnan, 2008). Whether or not this value is the globally optimal solution is unknown.[1] Additionally, researchers have typically explored several models, with different numbers of clusters and increasing numbers of covariance parameters (like the nine covariance decompositions used in the mixture modeling program "mclust"; Fraley & Raftery, 2006). This leads to a large number of models to choose from,[2] and selecting among these models is typically done with a fit statistic, the most common of which is the Bayes information criterion (BIC; Schwarz, 1978).

The EM algorithm is deterministic, meaning that upon repeated initializations, a given set of starting values will necessarily converge to the same solution. Consequently, given its influence on the final solution, appropriately choosing starting values is paramount. At this point, the influence of starting values on the performance measures mentioned above (the goodness of fit and quality of the resulting solution, as well as the speed to return the start values) has been underevaluated. Given the propensity for mixture models to arrive at local solutions (Shireman, Steinley, & Brusco, 2015), it is unlikely that any particular solution (regardless of the initialization strategy) will be globally optimal. However, by generating data with a known cluster structure, the quality of any given initialization strategy can be assessed in terms of its recovery of that structure (Steinley & Brusco, 2011a). Steinley and Brusco (2007) used a similar dual assessment to determine which initialization strategy for *k*-means clustering was best, in terms of both the objective function and cluster recovery (see also Steinley, 2006b). The need for such guidance is made clear by the fact that Steinley and Brusco (2007) is the second most cited article (107 Citations) in the *Journal of Classification* since the article's original publication.[3]

Initialization strategies

The existing techniques for initialization can be thought to follow two distinct philosophies. We will refer to the first as "uninformed" by the data; it contends that the use of a brute-force search technique will find the globally optimal solution by examining the entire landscape of the likelihood function. The other will be referred to as "informed" by the data, and it uses a directed-search approach to find the globally optimal solution (perhaps more quickly than an uninformed technique).

This article expands prior work investigating the efficacy of different initialization techniques for mixture models (Biernacki, Celeux, & Govaert, 2003; Karlis & Xekalaki,

[1]Global optimality, although it is the main yardstick by which developers of initialization techniques create their algorithms, is perhaps not the "correct" solution (Gan & Jiang, 1999). However, we assume that if a researcher observes results that have a higher likelihood (or higher value on a fit index), the researcher will more than likely prefer it to any model with a lower likelihood. Therefore, we use the evaluation of "fit" using the Bayes information criterion (BIC), which does not completely correspond to the mathematical goal of global optimality, but in the case that the researcher has found the globally optimal likelihood, we assume that this would be the preferred, and thus the interpreted, solution. We refer to this as the "goal of global optimality," and present the results of the best observed BIC to put our study into the context of the literature of other works examining initialization techniques. However, we believe that our inclusion of other measures of solution quality, such as the classification of individuals in the sample into groups, attenuates this focus on global optimality and widens the discussion of what should be considered the "best" initialization technique. We thank an anonymous reviewer for bringing up this important point.

[2]With one to four clusters and nine covariance matrix types, there are already 36 models to compare, not considering potential local solutions.

[3]This citation information was taken from a Google Scholar search on December 4, 2014, searching for articles published in the *Journal of Classification* from 2007 to 2014.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

2003; Melnykov & Melnykov, 2012) and clustering algorithms (Steinley & Brusco, 2007). Biernacki et al. examined the performance of initialization techniques for mixture modeling, but the techniques examined were limited to alternative instantiations of the EM algorithm. Karlis and Xekalaki, and Melnykov and Melnykov, investigated specific initialization approaches under a limited range of settings, with the former focusing only on *univariate* Gaussian mixtures and the main focus of the latter was determining the number of clusters present in the data. Steinley and Brusco (2007) examined 12 procedures for initializing the *k*-means clustering algorithm. *k*-means clustering is a similar analysis procedure to a mixture model (and is equivalent in the case of homogeneous, spherical clusters; see Steinley & Brusco, 2011b, for a discussion). Their examination included both techniques that were uninformed by the data (Faber, 1994; SPSS, 2003; Steinley, 2003) and that were informed by the data (Bradley & Fayyad, 1998; Milligan, 1980), concluding that an uninformed technique was sufficient for *k*-means clustering. The following investigation expands the work by Biernacki et al. by examining initialization techniques that were not included in their simulation but that are commonly employed in available software, while incorporating the proposed method of Karlis and Xekalaki in a broader comparison below (viz., an iteratively constrained technique).

What follows is a description of each of the initialization strategies, an illustration with real data, and two simulation comparisons of the techniques: one in which the amount of time to generate starting values was constrained, and another in which the techniques were allowed to take as long as required. The initialization strategies chosen for the comparative investigation are those that appear in the most commonly implemented software (two R packages “mclust” and “mixture,” and the statistical computing software LatentGOLD and Mplus); in fact, one would be hard-pressed to find software that did not utilize one of the following five approaches.

1. Random starting values—A common initialization technique in methodological and substantive research is the use of random values to initialize a mixture model. *Random* starting values in this context refer to randomly classifying each observation to one of *k* clusters, each cluster receiving an equal number of observations. The parameters (both the group-specific mean vectors and covariance matrices) estimated from this initial cluster assignment begins the first E-step of the EM algorithm. Considering the high prevalence of local solutions in mixture models, two solutions resulting from this technique are likely to be quite different from one another. For this reason, a researcher initializing a mixture model with random values should start many times[4] and interpret the best solution obtained by using a fit criterion such as the BIC. Random starting values have been shown to have performance similar to or better than other initialization techniques involving restricted iterations of an alternative form of the EM algorithm (such as the classification EM or

[4]It has been shown that a minimal number of starting values in order to arrive at an adequate solution should exceed 1,000 for some data types; see Steinley and Brusco (2007) and Shireman, Steinley, and Brusco (2015). However, given the breadth of the present simulation study, such a large number of initializations was infeasible; consequently, 100 random initializations were used. There exist so-called “global” optimizing algorithms to fit a mixture model that are designed to remove the guesswork of selecting the number of initializations required (De Boer, Kroese, Mannor, & Rubinstein, 2005; Heath, Fu, & Jank, 2009; Hu, Fu, & Marcus, 2007). However, these techniques are not yet implemented in popular software (e.g., SAS, 2012; the “mclust” package for the R computing software, described by Fraley & Raftery, 2006; or M*plus*, Muthén & Muthén, 2012), nor are they immune from arriving at locally optimal solutions.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

stochastic EM, the details of which are omitted here; see McLachlan & Peel, 2000, and Biernacki et al., 2003).

There are ways of generating random starting values alternative to the one studied here. Many involve generating starting parameter values from a uniform distribution around some bounds informed by the data (see Hipp & Bauer, 2006; Muthén & Muthén, 2012); however, in the context of *k*-means clustering, Steinley (2003) found that random starting values that used only one observation per cluster increased the chances of being located in a "poor" part of the parameter space. Similarly, Karlis and Xekalaki (2003) found that randomly drawing parameters from their respective distributions exhibited poor performance.

Although random values have been shown to have decent performance, the main weakness of this initialization technique stems from the fact that many random starts must be implemented before a solution can be settled on. In addition, no existing technique can determine when the number of initializations is sufficient to ensure a full examination of the likelihood function.

2. Iteratively constrained EM—This technique makes the assumption that given a set of random values to initialize the algorithm, the quickest increase in likelihood will provide a better solution than random initializations that increase less rapidly. The technique involves running several iteratively constrained EM algorithms (themselves initialized with random starting values), thereby constraining the search space for the global optimum (see Lubke & Muthén, 2007, for a simulation application). The parameters of the best-fitting of these restricted solutions is used to start values in a mixture model that is allowed to iterate until convergence. A similar technique is used as the default for providing starting values in the statistical software M*plus* Version 7 (Muthén & Muthén, 2012).[5] This approach is a variant to the short-EM algorithm proposed in Biernacki et al. (2003), which was found to be the best-performing method.

There are a few drawbacks with the choice of an iteratively constrained technique. First, the user must choose the number of iterations in the initial-stage EM. The value chosen could then be sufficient for some likelihood functions, but perhaps not others (this is in contrast to initial values provided by an EM with large convergence criteria, which would not be as susceptible to this problem; see, e.g., Biernacki et al., 2003). An additional drawback is the selection of the number of initial-stage EM algorithms to run, which greatly increases the amount of time required for computation. M*plus* defaults to 20 random starts, constrained to ten iterations of the EM algorithm. Finally, the same drawbacks from using random values still apply to the use of an iteratively constrained technique, since this technique is still susceptible to local optima and, by construction, only searches a limited part of the parameter space.

3. K-means clustering—The *k*-means clustering technique uses an alternative clustering solution from the data to provide its starting values. Considering the demonstrated

[5]M*plus* differs from the described examination in that it generates its initial random starting values from uniform distributions (see Hipp & Bauer, 2006, for a discussion) and also uses an accelerated quasi-Newton algorithm in conjunction with the typical EM algorithm for estimation.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

agreement between the results of a *k*-means cluster analysis and a mixture model (Steinley & Brusco, 2011b), the results from the *k*–means algorithm can provide close to accurate parameters when they are used as initial values for the EM algorithm. This technique is implemented in the R package "mixture" (Browne, ElSherbiny, & McNicholas, 2014) and has been recommended by McLachlan and Peel (2000) as a viable initialization strategy.

This clustering initialization is implemented by performing a *k*-means cluster analysis a number of times, perhaps hundreds, and using the classifications from the solution with the lowest sum of squared errors (the most cohesive clusters, a typical way of choosing the best *k*-means solution) as initial values in the first M-step of the EM algorithm. For more detail about the computation of a *k*-means cluster analysis and the issues therein, see Steinley (2006a).

As with the other techniques, there are drawbacks to the *k*-means initialization technique. *K*-means clustering has been criticized for its inclination to find spherical clusters (e.g., the clustering counterpart of a latent profile mixture model) and its use of an ad-hoc clustering criterion (Lubke & Muthén, 2005). When the clusters are highly heterogeneous and non-spherical, the results of a *k*-means clustering may not be an accurate representation of the data, and thus may not provide adequate starting values. In addition, *k*-means clustering is subject to locally optimal solutions, so it shares some of the difficulty of the methods previously described, that the number of initializations (and the number of times to start the *k*–means within the procedure) is still a subjective determination made by the researcher.

4. Agglomerative hierarchical clustering—Another common way to generate starting values using an alternative clustering of the data is with agglomerative hierarchical clustering. A hierarchical cluster analysis is performed, and a partition is provided of the chosen number of clusters. The parameters of these groups are used for the first, E-step of the mixture model estimation (for more detail on hierarchical clustering, see Everitt, Landau, Leese, & Stahl, 2011). A program utilizing hierarchical clustering for starting values is the "mclust" package, provided by Fraley and Raftery (2006) for the computing software R (www.r-project.org).

Hierarchical clustering may be an accurate way to describe data that are organized hierarchically (e.g., biological taxonomies), but some researchers have disputed the accuracy of classifications based on this form of cluster analysis (Edelbrock, 1979). It has been demonstrated, however, that starting values based on hierarchical clustering classifications perform well for a *k*-means cluster analysis (Dasgupta & Raftery, 1998; Fraley, 1998; Steinley & Brusco, 2007). In addition, the use of only one set of starting values can be a negative in some cases, since it restricts the search for a solution to only one possible result.

5. Sum scores—This technique finds initial class memberships by utilizing the sum score as a representation of the data. After calculating a sum score for each individual, the data are split into *k* equally sized ordered groups, and these groups are used for the M-step of the EM algorithm (for more detail, see Bartholomew, Knott, & Moustaki, 2011).[6]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Although, in the case of relatively small sample sizes and data dimensionality, this technique is simple and quick to implement, the drawbacks to using it for initialization are the same as those for hierarchical clustering and *k*-means—the sum score may not be an adequate representation of the data. If the clusters are not organized by increasing means on the manifest variables, these clusters will not be nearly accurate. However, this technique's merit lies in its speed. A sum score will take a very short time to calculate, and since there is only one possible set of starting values, there is no need for several initializations.

Assessment of performance

To examine the ability of each initialization procedure to obtain an adequate solution, the performance for the following simulation was assessed in four ways: (1) the ability for the techniques to find the best-fitting solution (i.e., the best BIC), (2) the ability for the techniques to find a correct classification of individuals, (3) the propensity for the techniques to arrive at locally optimal solutions (i.e., solutions that have BICs differing by more than 1×10^{-16}), and (4) the time the techniques require to provide initial parameter values.

Best-fitting solutions—Although optimizing algorithms like the EM algorithm cannot guarantee the finding of a globally optimal solution, but only a locally optimal one, the highest BIC (Schwarz, 1978) observed in all initializations across techniques is the best estimation we have of this theoretical value. The quality of the initialization techniques will be assessed in terms of the percentages of times the best solution found by a technique was the best found by all techniques within a given data set over all initializations. That is, if the best BIC found over all techniques for all initializations of the EM algorithm on a single data set was −4,567, only those techniques that had found a solution corresponding to this BIC would receive a "1" for this data set; all others would receive a "0." Note that all techniques might receive a "1" if they all obtained a solution with the maximum BIC, but for each data set only one technique must necessarily receive a "1." The proportion of "1"s for a given factor condition would be used as the measure of the propensity of the technique to find "optimal" solutions (the maximum BIC being a proxy for the optimal solution). We will refer to this measure as the "percent of maximum BICs," or $p_{\max}$. The $p_{\max}$ for the *j*th technique on a factor condition of interest is formally given as

$$p_{\max}(j)=\frac{\sum_{i=1}^{F}\Pi_{ij}}{F},$$
$$\Pi_{ij}(\theta)=f(x)=\begin{cases}1, & \max(BIC_{ij})=\max(BIC_i)\\ 0, & else\end{cases}$$
$$BIC_{ij}=2\log L\,(x;\theta_{ij})+d\log(n),$$

where F is the number of data sets with the factor condition of interest, d is the number of estimated parameters, n is the sample size, $L(\mathbf{x}, \Theta)$ is the log likelihood, i indexes over each factor condition, and j indexes over each technique.

[6]This technique was implemented in a previous edition of the latent variable modeling software Latent GOLD (Vermunt & Magidson, 2000).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Classification assessment—Initialization techniques for mixture modeling have previously been evaluated on their ability to arrive at the global optimum (Biernacki et al., 2003). A common application of a mixture model is to use the modeled densities to classify the individuals in a sample to one of the k substantively interesting groups. An initialization technique might find a solution with accurate classifications, but not optimal fit. In addition, although a brute-force searching algorithm might find the globally optimal solution, using the data to create informed starting values could result in a solution with types of clusters that have more substantive validity (see Steinley & Hubert, 2008).

Following Steinley and Brusco (2007, 2011a), the classification quality of the mixture model solution will be evaluated with the Hubert–Arabie adjusted Rand index (ARI; Hubert & Arabie, 1985; Steinley, 2004). The ARI is a measure of cluster solution agreement with a maximum value of 1, indicating identical solutions. An ARI of 0 indicates agreement between two solutions that is equivalent to chance. The ARI is computed as

$$\frac{\binom{N}{2}(a+d)-[(a+b)(a+c)+(c+d)(b+d)]}{\binom{N}{2}^2-[(a+b)(a+c)+(c+d)(b+d)]},$$

where a is the number of pairs of individuals in the same group in both solutions, d is the number of pairs in different groups in both solutions, and b and c are the numbers of pairs that are discordant between the two solutions (same–different and different–same, respectively). We note that Steinley (2004) has demonstrated that the ARI avoids many of the pitfalls associated with misclassification rates, providing more support for its use. In the simulations to follow, the results will be compared with the true cluster memberships, so the ARI can be interpreted as the accuracy of the classifications.

Local optima—Local optima is a useful measure for evaluating the solution of a clustering algorithm, and a large number of local optima can be indicative that the algorithm has not been able to find the optimal solution (Shireman, Steinley & Brusco, 2015; Steinley, 2004). The percent of unique solutions is referred to as p_{lo} and is calculated as follows:

$$p_{lo}=\frac{\#\,Unique\,BICs}{I}$$

where I is the number of initializations on the same data set. Not all of the initialization procedures are subject to unique solutions. The initialization strategies of sum scores and hierarchical clustering only produce one set of starting values, and thus only lead to a single solution, so those techniques will not be included in the analysis of unique solutions.

Time—Elapsed time can be used to examine the efficiency of each technique. This is an important consideration in the applicability of initialization techniques, because as computation time increases, the likelihood that applied researchers will be willing or able to

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

conduct repeated initializations decreases. The time in CPU seconds was measured from when the initialization technique began until it provided the initial classifications.

Simulation I: Timed initializations

The first simulation examined the differential performance of each of the techniques in a variety of data situations, while constraining each technique to create its starting values within 20 s. A limit of 20 s was chosen because it gives more than sufficient time for the quick techniques (sum score and random values) and is just sufficient for the longer initialization techniques of hierarchical clustering and iteratively constrained EM. The goal of this first simulation was to examine each technique's performance in terms of identical conditions.

The data were generated with a known cluster structure and a mixture model that was fit 100 times for each initialization technique (assuming that the correct number of components is known but the covariance model is not). A model was fit to the nine covariance matrix decompositions provided in "mclust," which describe restrictions on the spectral decomposition of $\mathbf{\Sigma}$ (see Celeux & Govaert, 1995, for more description).

Data generation

The data for these simulations were generated using a MATLAB (MathWorks, 2012) adaptation of the cluster generation procedure described by Qiu and Joe (2006a). This program is flexible, in that (1) it can incorporate varying degrees of separation of the clusters, based on univariate projections (described in Qiu & Joe, 2006b); (2) it does not force the isolation of clusters on any manifest variable (a constraint imposed by a previously famous cluster generation procedure described in Milligan, 1985); and (3) it allows for general shapes of the clusters (i.e., no constraints on the equality of the covariance matrices). These qualities make the data generated from this program more realistic and allow the simulation results to generalize more easily to real-world data situations.

Factors

To gain an understanding of how each initialization technique can be affected by different data structures likely to be encountered in real analyses, several factors were varied in the data generation process.

The factors varied in this simulation were (1) the number of clusters (two, four, and six), (2) the number of variables (two, four, and six), (3) the sample size (100, 500, and 800), (4) the relative cluster density, and (5) the minimum cluster separation.

The *relative cluster density* refers to the percentage of the total observations generated in each cluster. Milligan (1980), Milligan and Cooper (1988), and Steinley (2003, 2006b) showed that the relative densities of each cluster can affect the recovery of clustering algorithms. The cluster densities were varied so that (1) all clusters contained equal numbers of observations, (2) one cluster contained 10% of the observations and the rest of the clusters were equally sized, and (3) one cluster contained 60% of the observations and the rest of the clusters were equally sized.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Qiu and Joe (2006b) proposed a cluster separation measure (alternatively referred to as *cluster overlap*) based on a set of one-dimensional projections used in conjunction with their data generation program, the technical details of which are omitted here. This separation index has a minimum of −1 (the separation of one cluster from itself) and a theoretical maximum of 1 (two clusters whose means are separated by an infinite distance). Qiu and Joe (2006b) also laid out conventions for this measure, in which $J = .01$ indicates close clusters and $J = .342$ indicates well-separated clusters. Their cluster generation procedure allows the user to indicate a minimum separation of the clusters. This minimum separation was varied in our simulation to be $J = .01$, .171, and .342, indicating close, moderately separated, and well-separated clusters.

A completely balanced design resulted in 3 (number of clusters) × 3 (number of variables) × 3 (sample size) × 3 (relative cluster density) × 3 (cluster separation) = 243 unique data conditions. Three data sets of each type were created, resulting in 729 data sets. Furthermore, for each data set, nine candidate mixture models were considered (each with a different covariance structure, but all assuming the correct number of clusters/components)—for a total of 6,561 mixture models to be fit to the data. Each model was fit with each initialization procedure, resulting in 32,805 models by initialization conditions. However, for three of the initialization conditions (random starts, iteratively constrained, and *k*-means clustering), 100 random initializations were used. The other two initialization conditions (sum scores and hierarchical clustering) were deterministic conditional on the observed data. This resulted in millions of mixture models being estimated, making this by far the largest and most extensive examination of initializations of finite mixture modeling to date. All analyses were conducted in MATLAB r2012a (MathWorks, 2012).

Results from Simulation I

When it comes to finding the global optimum, the technique with the highest $p_{\max}$ is the iteratively constrained technique, $p_{\max} = .55$ (see Table 1). That is, for 55% of the data sets, the best solution arrived at by the iteratively constrained technique is the best solution found by all techniques. However, we do note that achieving the best solution only 55% of the time leaves much room for improvement. The only factors that led the iteratively constrained technique not to have the highest $p_{\max}$ were well-separated clusters and a small sample size. In these two conditions, random initializations provided the highest $p_{\max}$.

The techniques of iteratively constrained EM and random values both resulted in the best classifications (see Table 2), averaging over all data types (ARI = .47). The differences in classification quality between the techniques were minute, however. Performance for all techniques improved when the clusters become more separated (as should be expected). The only exception to the overall performance patterns was the sum score technique, whose performance decreased when one cluster was larger than the rest, whereas the other techniques were less affected by this change. As we noted regarding the cluster generation technique, the structure of these groups was most realistic, in terms of not imposing arbitrary constraints on their shapes and/or overlap. As a result, many of the ARI values are considered low (Steinley, 2004, indicated that a “good” solution would have an ARI of at least .60 or .65). This difficulty that mixture models have with finding a known cluster

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

structure has been noted in several other studies (Shireman, Steinley, & Brusco, 2015; Steinley & Brusco, 2008, 2011b).

The iteratively constrained and random-value techniques have much larger p_{lo}s (see Table 3). These techniques are designed to examine the landscape of the likelihood function, so this discrepancy is not surprising. However, the p_{lo} for the iteratively constrained technique was consistently lower than that for random values, indicating that it was indeed constraining the search space for the global optimum. Furthermore, the greater performance of random starting values and iteratively constrained starting values, when compared to starting values provided by *k*-means clustering, hierarchical clustering, and sum scores, is not surprising, since the latter methods have already constrained the search space to an even more narrow region.

Simulation II: Initializations without time constraint

The previous simulation constrained each technique to find the best solution within 20 s. This was a disadvantage mostly to the iteratively constrained EM and hierarchical clustering techniques, which require relatively long computation times. The next simulation removed this constraint, allowing each technique to take as long as it needed to produce starting values. Considering that the differences in the results within each factor did not vary a great deal from technique to technique, the factors were cut down to increase the speed of computation. For the following simulation, the data sets were varied according to the following factors: (1) variables (two, four, and six), (2) sample size (100, 500, and 800), (3) relative cluster densities (equal-sized clusters, one with 10%, and one with 60%), and (4) minimum cluster separation (close, moderately separated, and well-separated). The number of clusters was kept constant at 2. In the case of iteratively constrained EM, the iteratively constrained stage was again allowed ten iterations of the EM algorithm, but it was given 20 random starts for this stage (the default in M*plus*). For the *k*-means initialization technique, the start values were the results from the best of 100 runs of the algorithm.

Results from Simulation II

The ordering by performance of the techniques did not vary by factor (i.e., there were no interactions between techniques and factors). In lieu of a table of the results for each data type, the average p_{max}, ARI, p_{lo}, and elapsed CPU seconds for each technique are given in Table 4.

As in Simulation I, the techniques of random values and iteratively constrained EM provided the most unique solutions. When the iteratively constrained technique was allowed 20 starting values to choose from, it found fewer local optima than when it was allowed as many as it could within the 20-s time limit, indicating that the search space for the global optimum was constrained further. This increase in the number of “initial”-stage starts did not, however, result in better solutions (i.e., a higher p_{max}) than in Simulation I. A supplemental simulation was conducted to examine the impact of increasing the number of starts, and showed that as the number of initial-stage starts increased (to > 30), the iteratively constrained technique consistently found the best solution more often than did the random-

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

values technique. However, at this level, the iteratively constrained technique required more than 1,000 times as much computing time to find starting values.

The techniques of random values and sum score take, on average, less than a hundredth of a second to generate their starting values, and the initialization technique of *k*-means cluster analysis takes on average less than half a second. The iteratively constrained and hierarchical clustering techniques take significantly more time to generate their starting values. Hierarchical clustering required 20 s on average, and the iteratively constrained EM over 30 s on average. However, the iteratively constrained technique used its extra time to buy more optimality: The iteratively constrained technique found the best solution 36% more often than the hierarchical clustering technique ($p_{\max}$ = .10 for hierarchical clustering vs. .46 for iteratively constrained EM).

On the other hand, if the primary concern is cluster recovery, then a strong recommendation can be made to initialize the mixture model with *k*-means clustering, since it obtains the same ARI as both the random-initialization approach and the iteratively constrained EM algorithm. Even though it only achieves the best found solution 7% of the time, it still yields the same quality of final partitioning, but 60 times faster than the iteratively constrained approach.

Empirical illustration

To illustrate the practical impact of starting values on the resulting solutions, the above initialization techniques were implemented on data from the National Epidemiologic Survey of Alcohol and Related Conditions (NESARC; Grant et al., 2003).

Participants and procedure

NESARC contains data concerning the etiology and comorbidity of substance use disorders, measured at two time points (e.g., waves). The first wave of the NESARC consists of data from 43,093 population-representative individuals collected in face-to-face interviews between 2001 and 2002. The analyses here utilized the 11 dichotomous criteria for alcohol use disorder outlined in the *Diagnostic and Statistical Manual of Mental Disorders* (DSM-5; American Psychiatric Association, 2013). For these analyses, Wave 2 data were used, with men[7] and past-year abstainers from alcohol excluded (a final sample size of N = 11,782). The 11 criteria and their descriptions are given in Table 5.

Analysis

A latent class analysis (LCA; a mixture model assuming heterogeneous clusters for dichotomous data) was fit 1,000 times to these data for each of the initialization techniques. Only the three-class solution will be examined, since previous analyses of the 11 diagnostic criteria has shown support for the three-class solution (Beseler, Taylor, Kraemer, & Leeman, 2012; Chung & Martin, 2001; Sacco, Bucholz, & Spitznagel, 2009).

[7]Men were excluded on the basis of significant differential item functioning between genders (Saha, Chou, & Grant, 2006).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Results

The results for this illustration are given in Table 6. The best BIC of the 1,000 initializations, the proportion of local optima, and the times that the program required to generate starting values are given. The technique that found the best BIC was iteratively constrained EM. The technique with the fewest locally optimal solutions was *k*-means initialization, and the technique that took the shortest amount of time to generate its starting values was the random-values technique. Of the techniques that needed to calculate information from the data, the fastest technique was sum score, taking only two hundredths of a second, on average, to generate starting values. Clearly, the technique at a disadvantage with this measure was hierarchical clustering, which, due to the large sample size, took 418 s to create starting values (although this solution was nearer to the best BIC found than was the sum score or *k*–means technique).

Table 7 shows the agreement between the final solutions provided by each initialization scheme in terms of the ARI. Random values and iteratively constrained EM resulted in identical solutions. This is striking, especially considering the difference in the times these techniques require to generate the start values: Random starting values took less than a hundredth of a second, and iteratively constrained EM required over 12 s per initialization, but the solutions were identical. In addition, the classifications agree poorly between the "uninformed" and "informed" initialization techniques (ARI = .69 or worse). This indicates that although the classifications were of comparable quality in the preceding simulations, these classifications were not similar.

Discussion

Generally, although investigations of competing methods (in this case, the initialization of mixture models) via simulation are helpful, it is necessary to understand their applicability to a "real-world" setting. As George Box famously said, "Essentially all models are wrong, but some are useful"; in this case, it is incumbent to determine whether the results of the study presented here can aid in ensuring that the mixture model will be a useful, replicable technique. As Milligan and Cooper (1988) indicated when comparing approaches for choosing the number of clusters, for an approach to be seriously considered and used by research analysts, it is necessary that it exhibit good performance in cases with well-defined cluster structures. In this case, the simulation represents an idealized world where the true number of classes and their membership are known, and the performance of competing initialization approaches was assessed on the basis of that information alone.

Unfortunately, that is almost never the case when analyzing an empirical data set. First, it is rare that the true number of classes (e.g., mixtures) is known. Second, even if the number of classes is known, it is exceedingly rare—almost never—that the distributional form of each of the mixtures is known. As such, the results presented herein should be considered the best-case scenario, in terms of the overall performance of the mixture model itself. As the model and data begin to diverge, one would often observe an increasing number of local optima (see Steinley, 2003, 2006b). In those cases, it appears that initializing mixture models via *k*-means clustering is most effective in finding the best solution. This finding is

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

supported by the results of Steinley and Brusco (2011a, b) showing that *k*-means clustering is less sensitive to departures from the model assumptions.

The potential disconnect between model and data raises another interesting question. Namely, if the data do not conform to either mixtures or the correct distributional form of the mixtures themselves, how useful is the mixture model? And by extension, how useful is knowing the best way to initialize the model? McLachlan and Peel (2000) noted that mixture modeling can be used for two primary purposes: (1) trying to identify the true, underlying set of groups present in the data (the most common use), and (2) approximating the density of a multivariate data set through a series of mixtures, most typically Gaussian mixtures. In the latter case, the results presented here will always be relevant, since it is desirable to choose the initialization procedure that leads to the best goodness-of-fit statistic. In the former case, we conjecture that the results will be broadly relevant and robust, since researchers will be interested in using the results of the mixture model for predictions of external covariates.

There is no theoretical reason to believe that the recommendations outlined below will change depending on the model being fit to a particular data set. If the number of mixtures is either under- or overspecified, that will lead to an increase in the number of locally optimal solutions (Shireman, Steinley, & Brusco, 2015) and result in the general recommendation of using *k*-means clustering to initialize the mixture model (discussed further in the Recommendations below). Furthermore, candidate mixture models should be subjected to the appropriate tests of reliability and generalizability. We do not see choosing the appropriate initializing method as obviating that need; rather, it is the first step in a broader series of steps to ensure that the best model (in terms of fit indices and/or the quality of the solution) is subjected to those tests of reliability and generalizability. Disregarding the method of initialization as an important choice will likely introduce greater error to the modeling process.

Recommendations

The choice of starting values for a mixture model is not a decision to be taken lightly or to be left to the defaults of any given program. Although the classification results do not vary a great deal between the different techniques, the ability to find the globally optimal solution (or at least the best observed solution) varies a great deal on the basis of the choice of starting values. In terms of specific recommendations, we offer the following guidelines:

- We recommend the use of the iteratively constrained approach tested herein to generate initial values. Additionally, if it is computationally feasible, we recommend that the number of starting configurations be set to 30, to ensure that the quality is more likely to surpass that of a random-values technique. Although M*plus* uses an iteratively constrained initialization technique, its parameter estimates are generated using uniform distributions, which was not explicitly tested here and (considering the results from Steinley, 2003) may provide worse start values.
- A nearly equivalent form of initialization is the random-values technique, and it is recommended in the case when analyses need to be run quickly. Although the

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

iteratively constrained EM technique had the best performance, it only found the best observable BIC in 55% of data sets in the first simulation. It increases computation time by many orders of magnitude, but buys relatively little in terms of optimality.

- If it seems that there are numerous local optima, it may also be worthwhile to search more of the space with the *k*-means initialization approach, since it finds fewer local optima but settles on a solution that is of comparable quality.
- On the basis of the simulation results, caution should be employed when using the popular "mclust" package in R. The primary caution comes from its utilization of a hierarchical-clustering method as its initialization strategy; this approach was shown to be one of the least effective in terms of converging to the best observed BIC. A similar caution should be applied when using the classic method of sum scores for initializing mixture models.

Future research should focus on exploring the relationship between a broader range of fit statistics and cluster recovery.

References

American Psychiatric Association. Diagnostic and statistical manual of mental disorders. 5. Arlington, VA: American Psychiatric Publishing; 2013.

Bartholomew, DJ., Knott, M., Moustaki, I. Latent variable models and factor analysis: A unified approach. 3. West Sussex, UK: Wiley; 2011.

Beseler CL, Taylor LA, Kraemer DT, Leeman RF. A latent class analysis of DSM-IV alcohol use disorder criteria and binge drinking in undergraduates. Alcoholism: Clinical and Experimental Research. 2012; 36:153–161.

Biernacki C, Celeux G, Govaert G. Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models. Computational Studies and Data Analysis. 2003; 41:561–575.

Bradley, PS., Fayyad, UM. Refining initial points for *k*-means clustering. Proceedings 15th International Conference on Machine Learning; San Francisco, CA: Morgan Kaufmann; 1998. p. 91-99.

Browne, RP., ElSherbiny, A., McNicholas, PD. mixture: Mixture models for clustering and classification. 2014. (R package version 1.3). Retrieved from https://cran.r-project.org/web/packages/mixture/index.html

Celeux G, Govaert G. Gaussian parsimonious clustering models. Pattern Recognition. 1995; 28:781–793.

Chung T, Martin CS. Classification and course of alcohol problems among adolescents in addictions treatment programs. Alcoholism: Clinical and Experimental Research. 2001; 25:1734–1742.

Dasgupta A, Raftery AE. Detecting features in spatial point processes with clutter via model-based clustering. Journal of the American Statistical Association. 1998; 93:294–302.

De Boer PT, Kroese DP, Mannor S, Rubinstein RY. A tutorial on the cross-entropy method. Annals of Operations Research. 2005; 134:19–67.

Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society: Series B. 1977; 39:1–38.

Edelbrock C. Mixture model tests of hierarchical clustering algorithms: The problem of classifying everybody. Multivariate Behavioral Research. 1979; 14:367–384. DOI: 10.1207/s15327906mbr1403_6 [PubMed: 26821856]

Everitt, B., Landau, S., Leese, M., Stahl, D. Cluster analysis. 5. West Sussex, UK: Wiley; 2011.

Faber V. Clustering and the continuous *k*-means algorithm. Los Alamos Science. 1994; 22:138–144.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Fraley C. Algorithms for model-based Gaussian hierarchical clustering. SIAM Journal on Scientific Computing. 1998; 20:270–281.

Fraley, C., Raftery, AE. MCLUST version 3: An R package for normal mixture modeling and model-based clustering. Seattle, WA: Washington University, Department of Statistics; 2006.

Gan L, Jiang J. A test for global maximum. Journal of the American Statistical Association. 1999; 94:847–854.

Grant, BF., Moore, TC., Kaplan, KD. Source and accuracy statement: Wave 1 National Epidemiologic Survey on Alcohol and Related Conditions (NESARC). Bethesda, MD: National Institute on Alcohol Abuse and Alcoholism; 2003.

Heath JW, Fu MC, Jank W. New global optimization algorithms for model-based clustering. Computational Statistics and Data Analysis. 2009; 53:3999–4017.

Hipp JR, Bauer DJ. Local solutions in the estimation of growth mixture models. Psychological Methods. 2006; 11:36–53. DOI: 10.1037/1082-989X.11.1.36 [PubMed: 16594766]

Hu J, Fu MC, Marcus SI. A model reference adaptive search method for global optimization. Operations Research. 2007; 55:549–568.

Hubert L, Arabie P. Comparing partitions. Journal of Classification. 1985; 2:193–218.

Karlis D, Xekalaki E. Choosing initial values for the EM algorithm for finite mixtures. Computational Statistics and Data Analysis. 2003; 41:577–590.

Lubke GH, Muthén B. Investigating population heterogeneity with factor mixture models. Psychological Methods. 2005; 10:21–39. DOI: 10.1037/1082-989X.10.1.21 [PubMed: 15810867]

Lubke G, Muthén BO. Performance of factor mixture models as a function of model size, covariate effects, and class-specific parameters. Structural Equation Modeling. 2007; 14:26–47. DOI: 10.1080/10705510709336735

MathWorks. MATLAB user's guide. Natick, MA: Author; 2012.

McLachlan, GJ., Krishnan, T. The EM algorithm and extensions. New York, NY: Wiley; 1997.

McLachlan, GJ., Krishnan, T. The EM algorithm and extensions. 2. New York, NY: Wiley; 2008.

McLachlan, G., Peel, DA. Finite mixture models. New York, NY: Wiley; 2000.

Melnykov V, Melnykov I. Initializing the EM algorithm in Gaussian mixture models with an unknown number of components. Computational Statistics and Data Mining. 2012; 56:1381–1395.

Milligan GW. The validation of four ultrametric clustering algorithms. Pattern Recognition. 1980; 12:41–50.

Milligan GW. An algorithm for generating artificial test clusters. Psychometrika. 1985; 50:123–127.

Milligan GW, Cooper MC. A study of variable standardization. Journal of Classification. 1988; 5:181–204.

Muthén, LK., Muthén, BO. Mplus user's guide. 7. Los Angeles, CA: Muthén & Muthén; 2012.

Qiu W, Joe H. Generation of random clusters with specified degree of separation. Journal of Classification. 2006a; 23:315–334.

Qiu W, Joe H. Separation index and partial membership for clustering. Computational Statistics and Data Analysis. 2006b; 50:585–603.

Sacco P, Bucholz KK, Spitznagel EL. Alcohol use among older adults in the national epidemiologic survey on alcohol and related conditions: A latent class analysis. Journal of Studies on Alcohol and Drugs. 2009; 70:829–839. [PubMed: 19895759]

Saha TD, Chou SP, Grant BF. Toward an alcohol use disorder continuum using item response theory: Results from the National Epidemiologic Survey on Alcohol and Related Conditions. Psychological Medicine. 2006; 36:931–941. DOI: 10.1017/S003329170600746X [PubMed: 16563205]

Schwarz G. Estimating the dimension of a model. Annals of Statistics. 1978; 6:461–464.

Seidel W, Mosler K, Alker M. A cautionary note on likelihood ratio tests in mixture models. Annals of the Institute of Statistical Mathematics. 2000; 52:481–487.

Shireman EM, Steinley D, Brusco M. Local optima in mixture modeling. 2015 Manuscript submitted for publication.

SPSS. SPSS 12.0 command syntax reference. Chicago, IL: SPSS, Inc; 2003.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Steinley D. Local optima in *k*-means clustering: What you don't know may hurt you. Psychological Methods. 2003; 8:294–304. [PubMed: 14596492]

Steinley D. Properties of the Hubert–Arabie adjusted Rand index. Psychological Methods. 2004; 9:386–396. DOI: 10.1037/1082-989X.9.3.386 [PubMed: 15355155]

Steinley D. *K*-means clustering: A half-century synthesis. British Journal of Mathematical and Statistical Psychology. 2006a; 59:1–34. [PubMed: 16709277]

Steinley D. Profiling local optima in *K*-means clustering: Developing a diagnostic technique. Psychological Methods. 2006b; 11:178–192. DOI: 10.1037/1082-989X.11.2.178 [PubMed: 16784337]

Steinley D, Brusco MJ. Initializing *k*-means batch clustering: A critical evaluation of several techniques. Journal of Classification. 2007; 24:99–121.

Steinley D, Brusco MJ. Selection of variables in cluster analysis: An empirical comparison of eight procedures. Psychometrika. 2008; 73:125–144.

Steinley D, Brusco MJ. Choosing the number of clusters in *K*-means clustering. Psychological Methods. 2011a; 16:285–297. DOI: 10.1037/a0023346 [PubMed: 21728423]

Steinley D, Brusco MJ. Evaluating mixture modeling for clustering: Recommendations and cautions. Psychological Methods. 2011b; 16:63–79. DOI: 10.1037/a0022673 [PubMed: 21319900]

Steinley D, Hubert L. Order-constrained solutions in *K*-means clustering: even better than being globally optimal. Psychometrika. 2008; 73:647–664.

Vermunt, J., Magidson, J. Latent GOLD 4.5 user's guide. Belmont, MA: Statistical Innovations; 2000.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 1

All p_{nax} results for Simulation I

	RV	ICEM	KM	AHC	SS
Overall	.40	**.55**	.11	.05	.02
Clusters					
2	.43	**.47**	.05	.10	.04
4	.42	**.62**	.11	.04	.01
6	.35	**.57**	.16	.01	.00
Variables					
2	.36	**.49**	.11	.09	.04
4	.40	**.60**	.07	.02	.01
6	.44	**.56**	.14	.04	.00
SS					
100	**.58**	.43	.15	.06	.01
500	.36	**.59**	.09	.02	.02
800	.26	**.63**	.09	.06	.01
Density					
Equal-sized	.43	**.51**	.10	.05	.04
1 Small	.35	**.57**	.17	.07	.00
1 Large	.42	**.58**	.05	.02	.01
Separation					
Close	.25	**.62**	.10	.06	.01
Moderate	.40	**.56**	.14	.06	.01
Well	**.56**	.48	.09	.02	.02

Each cell is the average proportion of solutions in which the resulting BIC was the best found for that data set. Bold values indicate the best ARI for the factor condition. RV, random values; ICEM, iteratively constrained EM; KM, *k*-means clustering; AHC, agglomerative hierarchical clustering; SS, sum score.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 2

Adjusted Rand index (ARI) results for Simulation I

	RV	ICEM	KM	AHC	SS
Overall	**.47**	**.47**	.46	.46	.44
Clusters					
2	**.43**	**.43**	.42	**.43**	.42
4	**.52**	.51	.51	.50	.49
6	**.45**	**.45**	**.45**	.44	.41
Variables					
2	.52	.52	.52	**.53**	.51
4	.42	.42	**.43**	.40	.40
6	.45	**.45**	.43	.44	.41
SS					
100	**.39**	**.39**	**.39**	**.39**	.37
500	**.48**	**.48**	.47	.47	.45
800	**.53**	**.53**	.52	.51	.50
Density					
Equal-sized	.44	.42	.44	.44	.43
1 Small	.46	**.47**	.46	**.47**	.45
1 Large	**.50**	**.50**	.48	.46	.43
Separation					
Close	.26	**.26**	.25	**.26**	.24
Moderate	.53	**.54**	.53	.52	.51
Well	**.60**	**.60**	**.60**	.59	.56

Each cell is the average ARI of the classifications of the resulting solution and the true generated cluster structure. Bold values indicate the best ARI for the factor condition. RV, random values; ICEM, iteratively constrained EM; KM, *k*-means clustering; AHC, agglomerative hierarchical clustering; SS, sum score.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 3

All p_{lo} results for Simulation I

	RV	ICEM	KM
Overall	.74	.59	**.23**
Clusters			
2	.62	.32	**.04**
4	.73	.62	**.22**
6	.86	.83	**.42**
Variables			
2	.73	.63	**.11**
4	.78	.63	**.25**
6	.70	.51	**.32**
SS			
100	.64	.38	**.21**
500	.77	.69	**.26**
800	.80	.70	**.21**
Density			
Equal-sized	.71	.57	**.21**
1 Small	.73	.58	**.19**
1 Large	.78	.62	**.28**
Overlap			
Close	.91	.75	**.32**
Moderate	.69	.55	**.19**
Well	.61	.47	**.17**

Each cell is the average proportion of solutions with a unique BIC. Bold values indicate the lowest p_{lo} for the factor condition. RV, random values; ICEM, iteratively constrained EM; KM, *k*-means clustering; AHC, agglomerative hierarchical clustering; SS, sum score.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 4

Results for Simulation II

	RV	ICEM	KM	AHC	SS
p_{max}	.46	.46	.07	.10	.05
ARI	.41	.41	.41	.40	.40
p_{lo}	.58	.25	.02	N/A	N/A
Time	0.00	30.66	0.49	20.31	0.00

Each cell is the average of each measure over all factor conditions. p_{max}, proportion of solutions with the maximum BIC; ARI, average classification agreement; p_{lo}, average proportion of solutions with a unique BIC; Time, average CPU seconds to generate start values; RV, random values; ICEM, iteratively constrained EM; KM, *k*-means clustering; AHC, agglomerative hierarchical clustering; SS, sum score.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 5

DSM-5 alcohol use disorder criteria

Abbreviation	Description
TOL	Tolerance of alcohol
WITHD	Withdrawal after quitting or cutting down
LARGERLONGER	Substance used more than intended
CUTD	Desire or efforts to cut down on use
TIME	Spent time using substance or recovering from effects
GIVEUP	Important activities given up because of use
CONTINUE	Continued use despite health problems
ROLE	Failure to fulfill role obligations
HAZARD	Use in physically hazardous situations
SOCIAL	Use despite social/interpersonal problems
CRAVING	Craving or a strong desire or urge to use alcohol

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 6

NESARC illustration results

	RV	ICEM	KM	AHC	SS
BIC	−33,791.0	**−33,786.5**	−35,422.9	−33,991.1	−36,268.7
p_{Io}	1.00	.99	.00		
Time	0.00	12.22	99.44	418.56	0.03

BIC Bayes information criterion; p_{IO} average proportion of solutions with a unique BIC; *Time* average CPU seconds to generate start values; *RV* random values; *ICEM* iteratively constrained EM; *KM k*-means clustering; *AHC* agglomerative hierarchical clustering; SS, sum score

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 7

Agreement between solutions by initialization schemes

	RV	ICEM	KM	AHC	SS
RV	–	1.00	.69	.53	.50
ICEM	–	–	.69	.53	.50
KM	–	–	–	.76	.75
AHC	–	–	–	–	.97
SS	–	–	–	–	–

Each cell shows the pairwise adjusted Rand index for each solution. *RV* random values; *ICEM* iteratively constrained EM; *KM* *k*-means clustering; *AHC* agglomerative hierarchical clustering; *SS* sum score