Psychological Methods
2011, Vol. 16, No. 1, 82–88

1082-989X/11/$12.00 DOI: 10.1037/a0020144

COMMENT

K-Means May Perform as Well as Mixture Model Clustering but May Also Be Much Worse: Comment on Steinley and Brusco (2011)

Jeroen K. Vermunt
Tilburg University

Steinley and Brusco (2011) presented the results of a huge simulation study aimed at evaluating cluster recovery of mixture model clustering (MMC) both for the situation where the number of clusters is known and is unknown. They derived rather strong conclusions on the basis of this study, especially with regard to the good performance of *K*-means (KM) compared with MMC. I agree with the authors' conclusion that the performance of KM may be equal to MMC in certain situations, which are primarily the situations investigated by Steinley and Brusco. However, a weakness of the paper is the failure to investigate many important real-world situations where theory suggests that MMC should outperform KM. This article elaborates on the KM–MMC comparison in terms of cluster recovery and provides some additional simulation results that show that KM may be much worse than MMC. Moreover, I show that KM is equivalent to a restricted mixture model estimated by maximizing the classification likelihood and comment on Steinley and Brusco's recommendation regarding the use of mixture models for clustering.

Keywords: mixture of multivariate normal distributions, number of clusters, cluster recovery, classification EM, adjusted Rand index

Steinley and Brusco (2011) performed an extended simulation study aimed at evaluating the performance of mixture models with multivariate normal components as clustering tools under a variety of conditions. As far as I know, this was the first large-scale simulation study on mixture model clustering (MMC) that focused on cluster recovery—that is, on whether subjects are assigned to the correct cluster. As a measure for agreement between true and assigned cluster membership, Steinley and Brusco used the adjusted Rand index (ARI), which is a general measure for comparing cluster solutions that takes on values between 0 (agreement is no larger than expected by chance) and 1 (agreement is perfect). Steinley and Brusco investigated situations where the number of clusters is known (Simulations I and II) and unknown (Simulation III). They not only compared nine possible specifications of the MMC with one another but also compared MMC with *K*-means (KM) clustering. In fact, most of their conclusions concern the latter comparison.

Steinley and Brusco (2011) performed a huge simulation study and derived rather strong conclusions on the basis of this study. Simulation Study I suggested that *K*-means performs as well as MMC with data generated from local independence populations; Study II suggested that it performs slightly better with data generated from local dependence populations; and Study III suggested that it performs better when the number of clusters is unknown, with data generated from local independence populations. I agree with the authors' conclusion that the performance of KM may be equal to MMC in certain situations, which are primarily the situations they investigated. However, a weakness of the paper is the failure to investigate many important real-world situations where theory suggests that MMC should outperform KM.

In any large simulation study, it is always a challenge to select the most relevant conditions to investigate, because the number of such conditions is potentially unbounded. However, because KM itself can be shown to be a restricted form of mixture modeling, theory can be used to help structure the design of the KM–MMC comparison by including conditions where the more flexible, more realistic forms of mixture modeling would be expected to yield improved performance with real data. Unfortunately, these conditions were largely not considered by Steinley and Brusco (2011). Before elaborating on the KM–MMC comparison in terms of cluster recovery and providing some additional simulation results, I show that KM is equivalent to a restricted mixture model estimated by maximizing the classification likelihood and comment further on Steinley and Brusco's recommendations on how to use mixture models for clustering.

Equivalence Between KM and MMC: The Classification Expectation-Maximization (EM) Approach

Steinley and Brusco (2011) state that KM is strongly related to a mixture model with a covariance structure they call S1. In the S1 specification, the residual variances are assumed to be equal across variables and clusters; moreover, the variables are assumed to be independent within clusters, meaning that covariances are fixed to 0. The S1 mixture model can be formulated as follows:

Correspondence concerning this article should be addressed to Jeroen K. Vermunt, Department of Methodology and Statistics, Tilburg University, P.O. Box 90153, Tilburg, the Netherlands 5000 LE. E-mail: j.k.vermunt@uvt.nl

$$f(x_i) = \sum_{k=1}^{K} \alpha_k \prod_{j=1}^{J} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x_{ij} - \mu_{jk}}{\sigma^2}\right)^2\right],$$

where the product over the J variables follows from the local independence assumption and where the free parameters are the cluster proportions (α_k), the cluster-specific means for all variables (μ_{jk}), and the single residual variance (σ^2). In the MMC context, these parameters are typically estimated by maximizing the marginal log-likelihood function $\sum_{i=1}^{N} \log f(\mathbf{x}_i)$, say by means of the EM algorithm.

However, within the mixture modeling literature, alternative estimation methods have been proposed, one of which involves maximizing the classification log-likelihood using an algorithm referred to as classification EM (CEM; Celeux & Govaert, 1992). CEM yields a hard partitioning of the sample under study, similar to KM. It can be shown that KM is exactly equivalent to CEM with a S1 covariance structure and two additional constraints. More specifically, the CEM procedure maximizes the following function:

$$\begin{aligned}
&\log L_c \\
&= \sum_{k=1}^{K} \sum_{i \in C_k} \log(\alpha_k) + \sum_{k=1}^{K} \sum_{i \in C_k} \sum_{j=1}^{J} \log \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x_{ij} - \mu_{jk}}{\sigma^2}\right)^2\right] \\
&= \sum_{k=1}^{K} \sum_{i \in C_k} \log(\alpha_k) - NJ \log(\sigma) - NJ \log(\sqrt{2\pi}) \\
&\quad - \frac{1}{2\sigma^2} \sum_{k=1}^{K} \sum_{i \in C_k} \sum_{j=1}^{J} (x_{ij} - \mu_{jk})^2,
\end{aligned}$$

where $i \in C_k$ means that the summation concerned is over the subjects belonging to the kth cluster. Recall that KM minimizes the residual sum of squares (RSS):

$$RSS = \sum_{k=1}^{K} \sum_{i \in C_k} \sum_{j=1}^{J} (x_{ij} - \mu_{jk})^2.$$

It can now be easily seen that if α_k and σ^2 are fixed, say $\alpha_k = 1/K$, and $\sigma^2 = 1$, maximizing $\log L_c$ is equivalent to minimizing RSS. In other words, KM is not only related to MMC, it *is* a mixture model in which only the cluster-specific means are free parameters that are estimated by maximizing the classification log-likelihood rather than the marginal log-likelihood function.

Not only the CEM procedure but also the standard maximum likelihood approach can be equated to KM. This is achieved by setting $\alpha_k = 1/K$ and by fixing σ^2 to a small value. The latter rescales the posterior membership probabilities in such a way that the largest one becomes 1 and the others become 0, yielding again the CEM-type hard partitioning.

Recommendations on the Use of Mixture Models for Clustering

I agree with most of Steinley and Brusco's (2011) recommendations on how to apply MMC in practice. Their simulation study showed that using too complex a covariance structure—the most complex one is S9, which has a free covariance matrix for each cluster—may cause substantial deterioration in the performance of MMC. This is especially true when the number of variables is large, the number of clusters is large, and the total sample size or the sample size in some of the clusters is small; that is, when the number of parameters and number of cases available to estimate the parameters are far out of balance. Steinley and Brusco recommended starting a MMC procedure with a parsimonious covariance structure; that is, with a structure S5 corresponding to the classical latent profile model, which assumes local independence but imposes no further restrictions on the variances across classes or variables. S5 is indeed a good choice, because it is very important to account for unequal variances across classes and variables (see also below)—one reason why the more restricted covariance structures (S1–S4) are usually less suitable as starting points for MMC.

When using model type S5, violations of the local independence assumption can easily be diagnosed by inspecting bivariate residuals, which are provided by, for example, the Latent GOLD (Vermunt & Magidson, 2005) and Mplus (Muthén & Muthén, 2006) mixture modeling software packages. If one or more bivariate residuals are large, one may decide either to increase the number of clusters or to allow for within-cluster correlations. It should be noted that the latter does not necessarily imply that the fully unrestricted model S9 should be used. Instead, one may use, for example, a model with some free covariances or with a factor analytic structure for the covariances. Steinley and Brusco (2011) did not investigate such models. Instead, they focused on the nine models that can be derived from the parameterization proposed by Banfield and Raftery (1993). However, I have never seen applications of this parameterization in psychological research. From these nine models, S3, S5, S6 and S9 are the only ones used, although they are not parameterized in the complex manner shown by Steinley and Brusco. In addition, as indicated above, restricted variants between S3 and S6 and between S5 and S9 may be formulated by allowing some but not all covariance to be free or by using a specific covariance structure model.

As a final remark on the use of mixture models for clustering, I would like to stress that one should also be critical with regard to the within-cluster normal distribution assumption. This is especially relevant with discrete data—for example, when the clustering is based on a set of 5- or 7-point Likert items. My advice would be not to use a mixture model for continuous responses but instead a mixture model for discrete responses assuming multinomial instead of a normal within-cluster distributions. For ordinal variables, it is possible to restrict the multinomial response probabilities by relating the cluster membership to the response using an ordinal regression model.

When Do Differences Between MMC and KM Matter?

My main point of criticism of the Steinley and Brusco (2011) article is the lack of explanation for why KM's performance is equal to MMC's under the studied conditions. Without the use of a relevant theory to structure the KM–MMC comparisons, there can be no specification of situations in which MMC might be expected to perform better than KM. The authors failed to formulate clear hypotheses about key factors that might be expected to

affect the differential performance of KM and MMC. Some of the main features that differentiate MMC from the more restrictive KM are that MMC (a) allows residual variances to differ across variables and/or across clusters; (b) allows residuals to be correlated, possibly in a different way across clusters; and (c) yields a soft instead of hard partitioning, which is a way to account for the fact that we are usually not fully certain about a subject's cluster membership. I speculate below on the possible effects of these differences both for cluster recovery when the number of clusters is known (the topic of Simulations I and II) and for decisions about the number of clusters (the topic of Simulation III).

Cluster Recovery

My hypothesis is that the key difference between KM and MMC for cluster recovery with known K occurs under Option (a)—the option of allowing residual variances to differ across variables and across clusters—only under the condition that clusters are jointly overlapping or near one another. I focus on differences in variances across clusters, because differences in variances across variables may at least partially be repaired by transforming them prior to the actual cluster analysis. Magidson and Vermunt (2002b) provided a KM–MMC comparison showing that cluster recovery deteriorates by wrongly assuming equality of variances across variables.

What is the role of the error variances in MMC? In fact, the contribution of variable j in the computation of the distance of subject i to the center of cluster k is scaled by the inverse of the error variance concerned. To give an example, suppose subject i's squared distance for variable j is 2 for Cluster 1 and 4 for Cluster 2, and the cluster-specific residual variances are 2 and 8, respectively. This implies that as far as variable j is concerned, subject i is closer to Cluster 2 than Cluster 1 ($4/8 < 2/2$). Such a rescaling that takes into account cluster heterogeneity does not occur in KM clustering.

It should be noted that the assumption that variances are equal across clusters is not very likely to hold in practice. In other words, the S5 model usually fits real data much better than the KM-like S1 model. Despite the fact that it may not fit and may not apply the appropriate scaling of distances, the S1 model (and thus also KM) may perform very well in terms of cluster recovery for a fixed K. This is the case when clusters do not overlap and are, moreover, far enough from one another so that an incorrect scaling of the distances does not change their order.

Ignoring the within-cluster correlations can be expected to have little impact on the cluster recovery for a fixed K, which is more strongly affected by the correct estimation of the cluster centers and variances. This means that using S5 for cluster recovery may be acceptable even if the correct model is S9, and the same applies to using S3 instead of S6 (when the covariance matrix is equal across clusters).

My hypothesis on deteriorated cluster recovery when ignoring the fact that error variances are unequal was only partially investigated by Steinley and Brusco (2011). One of the distributions used in Simulation I assumed unequal variances, although the text is somewhat unclear about this. The authors stated that not only the normal distributions with unequal variances but also the uniform and triangular distributions were of the S5 type (thus with unequal variances across clusters), but it is unclear how this was achieved. Moreover, about the normal with unequal variances they stated, "The fourth condition relaxes the equality of variances constraint in the previous condition, allowing variables to assume different variances (however, they were still equal across clusters, resulting in hyperellipsoids in multidimensional space)" (Steinley & Brusco, 2011, p. 67). From the remark in parentheses, I conclude that this yields a model of the S3 type. Nevertheless, the results in their Table 1 show that under the "fourth condition," the S5 model performs better than S1 and KM. In my opinion, this effect is deflated because it is an average over the other conditions and thus also over all nonoverlap conditions, whereas my hypothesis stated that the effect of ignoring unequal variances occurs only when there is overlap (or clusters are very near). It is also unclear how unequal the variances were across clusters, which one needs to know to interpret the results properly.

The settings in Simulation II clearly were not suited to detect differences between KM and MMC. They confirmed my second hypothesis that ignoring the covariances does not deteriorate cluster recovery. But to show differences between KM and MMC, one should also look at the situation in which variances and covariances are cluster specific—a condition that is not included in Simulation II.

In the context of this commentary, it is not possible to set up a large simulation study. Instead, I generated two data sets to illustrate my hypotheses about the effect of unequal variances and the absence of an effect of within-cluster covariances on cluster recovery for known K. These two generated data sets are not just toy examples but are constructed to be similar to an existing data set, namely the diabetes data set used by several authors to illustrate MMC. I took the estimates from the final model for this data set reported by Vermunt and Magidson (2002), rounded these estimates (always in such a way that clusters became slightly more overlapping), and rescaled the three observed variables by dividing them by 20, 10, and 50, respectively, to make their scales more similar to each other. Table 1 presents the population values for the assumed three-cluster model. Means, variances, and covariances vary across clusters. I generated two data sets with 250 subjects belonging to Cluster 1, 500 to Cluster 2, and 250 to Cluster 3. The first data set is generated from an S5 population model (covariances are fixed to 0) and the second from a restricted S9 model. In the latter, labeled S9(x_1–x_2), the covariance between x_1 and x_2 is free and set to values corresponding to correlations of 0.5, 0.0, and 0.9 for Clusters 1, 2, and 3, respectively. Figures 1 and 2 depict the two generated data sets. They show that there is quite some joint overlap between the clusters. Moreover, the impact of unequal variances across clusters and within-cluster correlations on the clusters' shapes can be clearly observed.

Table 1
Definition of the S5 and S9(x_1–x_2) Populations

Parameter	Cluster		
	1	2	3
Mean x_1	10.00	10.00	20.00
Mean x_2	10.00	8.00	20.00
Mean x_3	15.00	7.50	5.00
Variance x_1	3.00	1.00	50.00
Variance x_2	6.40	1.20	32.00
Variance x_3	62.50	6.25	6.25
Covariance x_1–x_2 in S9(x_1–x_2) population	2.20	0.00	36.00

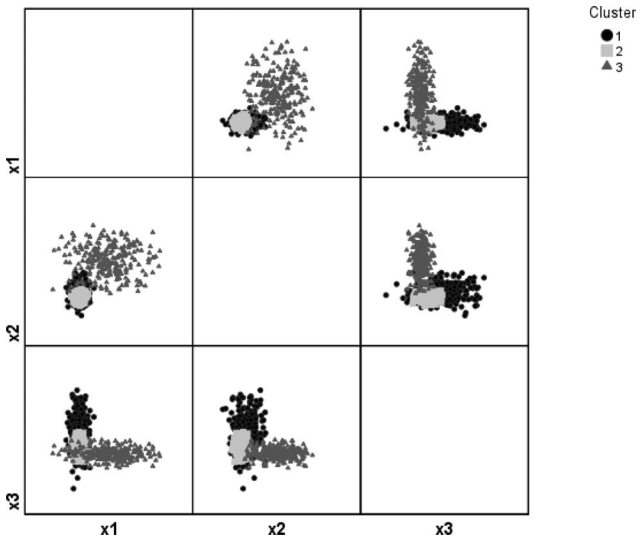

Figure 1. Generated S5 data set.

I estimated eight three-class mixture models with these two data sets, where the main distinction is between models with equal (co)variances across clusters—S1, S4, S6(x_1–x_2), and S6—and variants of these models with unequal (co)variances across Clusters—S2, S5, S9(x_1–x_2), S9—and also performed a KM analysis. Tables 2 and 3 report the ARI values for the estimated models. The largest values are .83 and .79, respectively, which are obtained when exactly the right model is specified. It can also be seen that the other models with cluster-specific (co)variances perform well too. This means that ignoring the (very large) correlation has little impact on the ARI (Model S5 in Table 3). We also see that using a model that is too complex may cause the performance to deteriorate slightly (Model S9 in Table 2), an effect that Steinley and Brusco (2011) showed to be larger with a larger number of variables and clusters and smaller sample sizes. What can also be observed is that Model S2 performs worse that the other three models with unequal (co-)variances, which is the result of the incorrect equal variance assumption across variables.

All four mixture models with equal (co)variances perform very poorly, which illustrates the importance of taking into account that variances differ across clusters. The fact that KM has larger ARI values than these mixtures models suggests that it may be more robust for violations of the equal (co)variance assumption. Overall these results confirm the hypothesis formulated above: Whereas KM and mixture models with equal variances may perform very well in certain situations (such as in those investigated by Steinley and Brusco, 2011), they will perform much worse than correctly specified mixture models when clusters have different variances and are overlapping.

Tables 2 and 3 also report the typical indices that are used for model selection in MMC. It is very reassuring to see that irrespective of the index one uses, the models with unequal (co)variances are always preferred over their equal (co)variance variants. Moreover, the model used to generate the data is identified as the best one, which shows that the risk of overfitting is not as large as Steinley and Brusco (2011) suggest.

Determining the Number of Clusters

Simulation III is aimed at comparing the performance of KM and MMC when the number of clusters is unknown. For this study, Steinley and Brusco (2011) also neglected to formulate hypotheses regarding the possible impact of specific differences between these methods on their ability to determine the correct number of clusters.

In general, it can be expected that misspecification of the within-cluster covariance structure will result in an overestimation of the number of clusters. One type of misspecification that may have a large impact is wrongly assuming that variances are equal across clusters, when in fact they are unequal. The reason for this is that a model with equal variances can only capture the larger volume of a cluster with a larger variance by splitting it up into several clusters with the smaller equal variance. Ignoring the presence of within-cluster correlations may also have a strong impact on the estimated number of clusters. The shape of a cluster with a moderate to high correlation between two variables is very different from the shape of a local independence cluster (compare, for example, the shape of Cluster 3 in the x_1–x_2 scatter plots of Figures 1 and 2). As a result, several local independence clusters may be needed to capture one local dependence cluster. Most probably, the effects of ignoring between-cluster difference in variances and ignoring covariances are stronger when clusters are overlapping.

Again, in my opinion, Steinley and Brusco (2011) failed to investigate several of the most relevant conditions to check these hypotheses. More specifically, they used only one covariance structure, namely the one with zero covariances and unequal variances across clusters (though some confusion about this remains, and it is unclear how different these variance were); moreover, they looked only at the situation in which there is no joint overlap or, as they state, "each of the clusters will have some 'empty space' between them in the joint multivariate space" (p. 73).

As indicated earlier, in the context of the current commentary, it is not possible to perform a complete simulation study on the effect of unequal variances, within-cluster correlations, and overlap be-

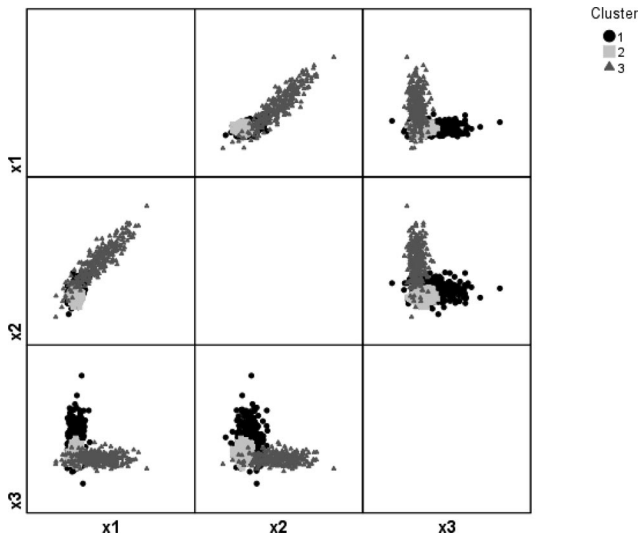

Figure 2. Generated S9(x_1–x_2) data set.

Table 2
BIC, AIC, AIC3, CAIC, and ARI Values Assuming K = 3 and Estimated Values of K by BIC and CH for the Generated S5 Data Set

Equation	Label	BIC	AIC	AIC3	CAIC	ARI	Estimated *K*	
							By BIC	By CH
σ_kI	S2	16,268	16,199	16,213	16,282	0.73	6 or more	2
D_k	S5	15,315	15,217	15,237	15,335	0.83	3	2
$D_k + \sigma_{12k}$	S9(x_1–x_2)	15,333	15,220	15,243	15,356	0.83	3	2
Σ_k	S9	15,371	15,229	15,258	15,400	0.82	3	2
σI	S1	17,187	17,129	17,141	17,199	0.54	6 or more	2
D	S3	17,081	17,013	17,027	17,095	0.39	6 or more	2
$D + \sigma_{12}$	S6(x_1–x_2)	17,069	16,995	17,010	17,084	0.39	6 or more	3
Σ	S6	17,075	16,991	17,008	17,092	0.40	6 or more	2
I	KM					0.63		3

Note. BIC = Bayesian information criterion; AIC = Akaike information criterion; AIC3 = Akaike information criterion (with 3 as penalizing factor); CAIC = consistent Akaike information criterion; ARI = adjusted Rand index; CH = Calinski and Harabasz (1974) index; KM = *K*-means.

tween clusters on the estimation of the number of clusters. However, the results obtained with the two simulated data sets reported in the last two columns of Tables 2 and 3 give an indication of the possible effects of these factors. When data are generated from a population that is in agreement with Model S5, then model S5 itself as well as the slightly too complex models S9(x_1–x_2) and S9 identify the right number of clusters on the basis of the Bayesian information criterion (BIC). This appears to be in contrast with the Steinley and Brusco (2011) result: "Although the covariance structure used to generate the data corresponded to S5, MMC/BIC correctly chose that structure 0% of the time" (p. 74). All other (overly restricted) models come up with six or more clusters. Table 3 shows a similar pattern: The correct S9(x_1–x_2) model and the more complex S9 model find the right number of clusters, and all other too restricted models come up with six or more classes. These results confirm my hypothesis that misspecifying the (co)variance structure may result in an overestimation of the number of clusters. Similar results on the effects of ignoring unequal variances and within-cluster correlations were reported by Magidson and Vermunt (2002a).

The results obtained with the information criteria show that Steinley and Brusco's (2011) conclusion, "it becomes apparent that making choices based on different functions of the likelihood (i.e., BIC, AIC) does not serve as a proxy for cluster recovery" (p. 76), is much too strong. Of course, many factors affect the performance of the various information criteria (such as sample size and separation levels), but the two examples showed that these measures may perform very well under rather difficult conditions, as long as the correct covariance structure is used.

Steinley and Brusco (2011) proposed using the Calinski and Harabasz (CH; 1974) index as a measure for determining the correct value of *K*, not only in KM but also in MMC. It is not fully clear from their text how they defined the within and between sum of squares for the CH index in the case of a soft partitioning such as is obtained in MMC. I obtained these using the posterior cluster membership probabilities (p_{ik}) as weights. That is,

$$CH_{MMC-K} = \frac{\sum_{k=1}^{K}\sum_{i=1}^{N}\sum_{j=1}^{J} p_{ik}\,(\mu_{jk} - \mu_j)^2 \Big/ (K-1)}{\sum_{k=1}^{K}\sum_{i=1}^{N}\sum_{j=1}^{J} p_{ik}\,(x_{ij} - \mu_{jk})^2 \Big/ (N-K)}.$$

Tables 2 and 3 (last column) indicate which solution would be selected on the basis of this index. When KM is used, CH selects

Table 3
BIC, AIC, AIC3, CAIC, and ARI Values Assuming K = 3 and Estimated Values of K by BIC and CH for the Generated S9(x_1–x_2) Data Set

Equation	Label	BIC	AIC	AIC3	CAIC	ARI	Estimated *K*	
							By BIC	By CH
σ_kI	S2	16,203	16,135	16,149	16,217	0.62	6 or more	3
D_k	S5	15,320	15,222	15,242	15,340	0.74	6 or more	2
$D_k + \sigma_{12k}$	S9(x_1–x_2)	14,936	14,823	14,846	14,959	0.76	3	2
Σ_k	S9	14,972	14,829	14,858	15,001	0.76	3	3
σI	S1	16,674	16,615	16,627	16,686	0.44	6 or more	3
D	S3	16,388	16,319	16,333	16,402	0.32	6 or more	3
$D + \sigma_{12}$	S6(x_1–x_2)	16,154	16,081	16,096	16,169	0.40	6 or more	3
Σ	S6	16,158	16,075	16,092	16,175	0.39	6 or more	3
I	KM					0.48		3

Note. BIC = Bayesian information criterion; AIC = Akaike information criterion; AIC3 = Akaike information criterion (with 3 as penalizing factor); CAIC = consistent Akaike information criterion; ARI = adjusted Rand index; CH = Calinski and Harabasz (1974) index; KM = *K*-means.

the correct number of clusters in both data sets (though it should be noted that in the first data set, the CH values for the two- and three-cluster models were very close to each other). On the basis of Steinley and Brusco's (2011) result, I had hoped that CH might be a good alternative to the information criteria that are typically used in MMC; that is, a measure that is less dependent on whether the within-cluster covariance is misspecified. This turned out not to be the case in the two examples. Depending on the specified covariance structure, CH selects two or three clusters, but there is no clear pattern in its preference for two or three. In the data set generated from the S5 population, we see that only the S2 model identifies the right number of clusters, whereas in the other data set almost all models identify the right *K* but not the model used to generate the data. Thus, it appears that CH is not a good alternative to BIC, although it is clear that the latter works properly only if the correct covariance structure is specified, where slight overfitting is not a problem.

Steinley and Brusco (2011) also stated, "If retaining the mixture modeling framework, we recommend using the CH index in conjunction with the covariance matrix of S5 to estimate the number of clusters" (p. 76). However, the above results showed that the CH index may not yield the right number of clusters, even if S5 is the correct model. Another problem with the Steinley and Brusco recommendation is that it seems to be based on the assumption that the right number of clusters can also be found when the covariance structure S5 is incorrect. Our second example showed that this assumption is completely invalid. As I indicated earlier, Model S5 is indeed a good starting point, but blindly increasing the number of clusters so long as BIC improves is not a good strategy. It is important to inspect residuals to see whether the misfit is caused by violations of the local independence assumption. If so, one may consider relaxing this assumption instead of increasing the number of clusters.

Final Remarks

Steinley and Brusco (2011) compared the differences between KM and MMC in terms of cluster recovery and ability to detect the correct number of clusters. In this commentary, I have shown that MMC may be preferred over KM under several important real-world conditions that were not investigated by Steinley and Brusco. These are conditions in which the greater flexibility provided by MMC is needed to obtain the correct results.

Whereas the two simulated data sets supported the initial a priori predictions concerning key factors affecting cluster recovery and determining the number of clusters, a more extended simulation study is needed to be able to generalize these results to different simulated conditions. Moreover, although in the two simulated data sets the appropriate within-cluster model was found by all fit measures, this is not always the case in practice, which implies that the MMC may be misspecified and therefore perform less well. KM seems to be more robust for such misspecifications; that is, it performs better than MMC with a misspecified model.

Besides being a more flexible clustering tool in terms of the covariance structures it allows, mixture modeling has several other advantages compared with KM clustering. One of these is that it can be used with categorical variables as well as with variables of mixed scale types (Vermunt & Magidson, 2002). Another advantage of using a model-based approach is that it can be tailored to specific data and application types. For example, mixture growth models (Muthén, 2004; Nagin, 1999) and latent Markov models (Collins & Wugalter, 1992; Van de Pol & De Leeuw, 1986; Vermunt, Tran, & Magidson, 2008) are mixture models for the analysis of longitudinal data; multilevel mixture models (Vermunt, 2003) can be used for the analysis of data sets with a hierarchical structure; mixture item response theory and mixture factor models can be used to separate discrete and continuous heterogeneity (De Boeck, Wilson, & Acton, 2005; Lubke & Neale, 2006); and mixture regression models can be used to cluster individuals based on differences in predictor effects on an outcome variable (Vermunt & Van Dijk, 2001; Wedel & DeSarbo, 1994).

References

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

Calinski, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics, 3,* 1–27.

Celeux, G., & Govaert, G. (1992). A classification EM algorithm for clustering and two stochastic versions. *Computational Statistics & Data Analysis, 14,* 315–332.

Collins, L. M., & Wugalter, S. E. (1992). Latent class models for stage sequential dynamic latent variables. *Multivariate Behavioral Research, 27,* 131–157.

De Boeck, P., Wilson, M., & Acton, G. S. (2005). A conceptual and psychometric framework for distinguishing categories and dimensions. *Psychological Review, 112,* 129–158.

Lubke, G. H., & Neale, M. C. (2006). Distinguishing between latent classes and continuous factors: Resolution by maximum likelihood? *Multivariate Behavioral Research, 41,* 499–532.

Magidson, J., & Vermunt J. K. (2002a, March). Latent class modeling as a probabilistic extension of *K*-means clustering. *Quirk's Marketing Research Review, 20,* 77–80.

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

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

Muthén, L. K., & Muthén, B. O. (2006). *Mplus user's guide* (4th ed.). Los Angeles, CA: Muthén & Muthén.

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

Steinley, D., & Brusco, M. J. (2011). Evaluating Mixture Modeling for Clustering: Recommendations and cautions. *Psychological Methods, 16,* 63–79.

Van de Pol, F., & De Leeuw, J. (1986). A latent Markov model to correct for measurement error. *Sociological Methods and Research, 15,* 118–141.

Vermunt, J. K. (2003). Multilevel latent class models. *Sociological Methodology, 33,* 213–239.

Vermunt, J. K., & Magidson, J. (2002). Latent class cluster analysis. In J. A. Hagenaars & A. L. McCutcheon (Eds.), *Applied latent class analysis* (pp. 89–106). Cambridge, England: Cambridge University Press.

Vermunt, J. K., & Magidson, J. (2005). *Technical guide for Latent GOLD 4.0: Basic and advanced.* Belmont, MA: Statistical Innovations.

Vermunt, J. K., Tran, B., & Magidson, J. (2008). Latent class models in longitudinal research. In S. Menard (Ed.), *Handbook of longitudinal research: Design, measurement, and analysis* (pp. 373–385). Burlington, MA: Elsevier.

Vermunt, J. K., & Van Dijk, L. (2001). A nonparametric random-coefficients approach: The latent class regression model. *Multilevel Modelling Newsletter, 13,* 6–13.

Wedel, M., & DeSarbo, W. A. (1994). A review of recent developments in latent class regression models. In R. P. Bagozzi (Ed.), *Advanced methods of marketing research* (pp. 352–388). Cambridge, MA: Blackwell.

Received October 7, 2009
Revision received March 10, 2010
Accepted March 22, 2010 ■