Behav Res (2013) 45:782–791
DOI 10.3758/s13428-012-0293-y

CHull as an alternative to AIC and BIC in the context of mixtures of factor analyzers

Kirsten Bulteel · Tom F. Wilderjans · Francis Tuerlinckx · Eva Ceulemans

Published online: 10 January 2013

Abstract Mixture analysis is commonly used for clustering objects on the basis of multivariate data. When the data contain a large number of variables, regular mixture analysis may become problematic, because a large number of parameters need to be estimated for each cluster. To tackle this problem, the mixtures-of-factor-analyzers (MFA) model was proposed, which combines clustering with exploratory factor analysis. MFA model selection is rather intricate, as both the number of clusters and the number of underlying factors have to be determined. To this end, the Akaike (AIC) and Bayesian (BIC) information criteria are often used. AIC and BIC try to identify a model that optimally balances model fit and model complexity. In this article, the CHull (Ceulemans & Kiers, 2006) method, which also balances model fit and complexity, is presented as an interesting alternative model selection strategy for MFA. In an extensive simulation study, the performances of AIC, BIC, and CHull were compared. AIC performs poorly and systematically selects overly complex models, whereas BIC performs slightly better than CHull when considering the best model only. However, when taking model selection uncertainty into account by looking at the first three models retained, CHull outperforms BIC. This especially holds in more complex, and thus more realistic, situations (e.g., more clusters, factors, noise in the data, and overlap among clusters).

Keywords Mixture analysis · Model selection · AIC · BIC · CHull

K. Bulteel · T. F. Wilderjans · F. Tuerlinckx · E. Ceulemans
KU Leuven, Leuven, Belgium

E. Ceulemans (✉)
Methodology of Educational Sciences Research Group,
KU Leuven, Andreas Vesaliusstraat 2,
3000 Leuven, Belgium
e-mail: Eva.Ceulemans@ppw.kuleuven.be

In the behavioral sciences, researchers often cluster multivariate (i.e., object-by-variable) data in order to capture the heterogeneity that is present in the population. The resulting clusters can differ with regard to their level and/or covariance structure. A first example pertains to the case in which a number of children are scored on certain psychopathological symptoms. The aim then is to discern different groups and to describe the differences between the groups in terms of the strength of the symptoms and/or of their linear covariation. A second example is a consumer psychologist who wants to identify different groups of consumers on the basis of their appraisals of a wide range of food products.

A commonly used clustering method is mixture analysis (McLachlan & Peel, 2000). In this method, each cluster is described by a different multivariate distribution, and every object belongs to each cluster with a particular probability. As a result, the full data follow a mixture of multivariate distributions. In practice, because of their computational simplicity, multivariate normal distributions are often assumed (McLachlan, Peel, & Bean, 2003), implying that each cluster is characterized by a mean vector and a covariance matrix.

When the number of variables increases, such a mixture of multivariate normals may become problematic, in that a large number of variance and covariance parameters need to be estimated for each cluster [i.e., for J variables, $J(J+1)/2$ variances and covariances need to be determined]. This problem is aggravated when the sample size is small (i.e., few objects) and/or when the clusters differ considerably in size, because in these cases, the information available to estimate the parameters is too limited.

To tackle this problem, the mixtures-of-factor-analyzers (MFA) model (Ghahramani & Hinton, 1997; McLachlan & Peel, 2000) was developed. This method combines clustering with exploratory factor analysis (EFA). In particular, the

variance–covariance matrix of each cluster is modeled by an EFA model, implying that the variables are reduced to Q cluster-specific latent factors. As a consequence, the number of parameters that need to be estimated is decreased drastically (i.e., for each cluster, only the parameters of the EFA model need to be estimated), which is advantageous in two respects. First, the estimation of the unrestricted mixture model may lead to numerical problems, because it is so richly parameterized (i.e., near-singular variance–covariance matrices; see McLachlan, Baek, & Rthnayake, 2011). Such numerical problems may be avoided when using the MFA model. A second advantage is enhanced interpretability. Instead of comparing entire variance–covariance matrices [with $J(J + 1)/2$ elements each], the main structural differences and similarities between the clusters can be studied by comparing the $J \times Q$ factor loadings per cluster, which reflect the cluster-specific linear relationships among the variables.

When using MFA, the model selection problem becomes intricate, in that one has to determine both the number of clusters and the number of underlying factors. To address the MFA model selection problem, McLachlan et al. (2003) suggested using information criteria such as the well-known Akaike information criterion (AIC; Akaike, 1974) and Bayesian information criterion (BIC; Schwarz, 1978). Both information criteria look for an optimal balance between model fit (or misfit), on the one hand, and model complexity, in terms of the number of estimated parameters, on the other. To the best of our knowledge, the performances of AIC and BIC in the context of MFA have not yet been evaluated thoroughly and systematically in an extensive simulation study. Yet simulation results for Gaussian mixture models have revealed that AIC tends to overestimate the number of clusters, while BIC may underestimate this number (Celeux & Soromenho, 1996). For factor mixture models, which generally constrain the latent factors to be identical across clusters, BIC mostly outperforms AIC (Henson, Reise, & Kim, 2007; Lubke & Neal, 2006; Nylund, Asparouhov, & Muthén, 2007; for an overview, see Vrieze, 2012). However, the performance of BIC deteriorates, and can even be worse than the performance of AIC, in more complex situations (e.g., large differences in cluster sizes or a wider range of models, including EFA and regular mixtures).

As an alternative model selection strategy, one may consider the use of the CHull method, which was proposed for tackling complex model selection problems in the context of deterministic clustering and/or dimension reduction methods (Ceulemans & Kiers, 2006). Relying on the same logic as AIC and BIC (i.e., finding an optimal balance between model fit and complexity), the CHull method generalizes and automates the idea behind the scree test (Cattell, 1966). In particular, a measure of model fit (or misfit) is plotted against a measure of model complexity, and by means of a numerical procedure, the model is identified after which the gain in fit by adding more parameters levels off. Simulation studies have shown that the CHull strategy performs well for a variety of clustering and component models (Ceulemans & Kiers, 2009; Ceulemans, Timmerman, & Kiers, 2011; Ceulemans & Van Mechelen, 2005; Schepers, Ceulemans, & Van Mechelen, 2008). Especially interesting in the present context is the study of Lorenzo-Seva, Timmerman, and Kiers (2011), who concluded that CHull outperforms AIC and BIC in the context of EFA.

The goal of the present article is twofold. First, the performance of AIC and BIC as model selection methods for MFA will be investigated by means of a simulation study. In addition, CHull will be proposed as an alternative model selection strategy for MFA, and its performance will be compared to that of AIC and BIC.

The remainder of this article starts with a section on MFA, in which the model as well as the parameter estimation will be treated. In the following section, AIC, BIC, and CHull are discussed. The fourth section presents the design and results of the simulation study. In the final section, we offer a summary of the results and outline some concluding remarks.

Mixtures of factor analyzers

Model

Exploratory factor analysis In EFA, the linear relations among J variables, measured for I objects, are explained by means of Q latent factors. Specifically, the J-variate vector $\mathbf{y}_i$ of object i is modeled as follows:

$$\mathbf{y}_i = \boldsymbol{\mu} + \mathbf{B}\mathbf{u}_i + \mathbf{e}_i \text{ (with } i = 1, \ldots, I), \quad (1)$$

with $\boldsymbol{\mu}$ being a J-variate vector containing the means of the J variables, $\mathbf{B}$ ($J \times Q$) the matrix of factor loadings, $\mathbf{u}_i$ a Q-dimensional vector containing the factor scores of object i, and $\mathbf{e}_i$ the J-variate residual vector. The following assumptions are made: (a) the $\mathbf{u}_i$ are independent and identically distributed (i.i.d.) as $N(0, \mathbf{I}_Q)$, with $\mathbf{I}_Q$ being the $Q \times Q$ identity matrix; (b) the $\mathbf{e}_i$ are i.i.d. as $N(0, \mathbf{D})$, with $\mathbf{D}$ being a $J \times J$ diagonal matrix containing the J unique variances σ_j^2 ($j = 1, \ldots, J$); and (c) $\mathbf{u}_i$ and $\mathbf{e}_i$ are independent. The model implies that the $\mathbf{y}_i$ are i.i.d. as $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with

$$\boldsymbol{\Sigma} = \mathbf{B}\mathbf{B}^{\mathrm{T}} + \mathbf{D}. \quad (2)$$

To count the number of free parameters in the EFA model, one should take the rotational freedom (i.e., the common factors may be rotated orthogonally [e.g., by varimax; Kaiser,

1958] or obliquely) into account: Given the orthonormal matrix $\mathbf{R}$ (i.e., $\mathbf{RR}^{\mathrm{T}} = \mathbf{I}$), the variance–covariance matrix $\mathbf{\Sigma}$ can be decomposed equally well as in Eq. 2, but now using a transformed loading matrix $\mathbf{B^*} = \mathbf{BR}$: $\mathbf{\Sigma} = \mathbf{B^*B^{*\mathrm{T}}} + \mathbf{D} = \mathbf{BRR^{\mathrm{T}}B^{\mathrm{T}}} + \mathbf{D} = \mathbf{BB}^{\mathrm{T}} + \mathbf{D}$. Therefore, the number of free parameters *fp* for an EFA model equals $2J + JQ - (1/2)Q(Q-1)$. Note that the rotational freedom may be used to facilitate the interpretation of the latent factors.

Mixtures of factor analyzers As described in the introduction, MFA is an extension of mixture analysis (MA). Given the scores of I objects on J variables, MA finds K object clusters in the data, with each cluster having its own density function. Assuming multivariate normal densities, MA models the density of the data as

$$f(\mathbf{y};\Psi) = \sum_{k=1}^{K} \pi_k \cdot \varphi(\mathbf{y};\boldsymbol{\mu}_k, \mathbf{\Sigma_k}), \tag{3}$$

where $\varphi(\mathbf{y};\boldsymbol{\mu}_k, \mathbf{\Sigma}_k)$ designates the cluster-specific multivariate normal density function with mean vector $\boldsymbol{\mu}_k$ and covariance matrix $\mathbf{\Sigma}_k$, π_k the a priori probability that an object i belongs to cluster k (with these probabilities being restricted to sum to one), and K the number of clusters. The vector $\boldsymbol{\psi}$ collects all mixing probabilities π_k and the means and covariances in $\boldsymbol{\mu}_k$ and $\mathbf{\Sigma}_k$ ($k = 1, \ldots, K$), respectively. As each cluster-specific covariance matrix has $(1/2)J(J+1)$ elements, the model is highly parameterized (McLachlan et al., 2003).

In order to reduce the number of parameters, MFA models each cluster-specific covariance matrix $\mathbf{\Sigma}_k$ with a separate EFA model, reducing the J variables to Q factors. Consequently, the density function of the mixture of factor analyzers has the same form as the mixture probability density function in Eq. 3, but it restricts the covariance matrices $\mathbf{\Sigma}_k$ to

$$\mathbf{\Sigma}_k = \mathbf{B}_k\mathbf{B}_k^{\mathrm{T}} + \mathbf{D}_k \ (k = 1, \ldots, K), \tag{4}$$

with $\mathbf{B}_k$ ($J \times Q$) being the factor loading matrix of cluster k, and $\mathbf{D}_k$ a diagonal matrix containing the unique variances for the kth cluster (i.e., σ^2_{jk}). The $\boldsymbol{\psi}$ vector in Eq. 3 now consists of all elements in $\boldsymbol{\mu}_k$, $\mathbf{B}_k$, and $\mathbf{D}_k$, as well as the mixing probabilities π_k.

In each cluster, the factor loadings may be orthogonally or obliquely rotated in order to facilitate the interpretation of the cluster-specific factors. This implies that the number of free parameters *fp* is reduced to $K[2J + JQ - (1/2)Q(Q-1)] + (K-1)$ (McLachlan et al., 2003). Note that, in order to further reduce the number of parameters, additional constraints may be imposed on the MFA model. For instance, $\mathbf{B}_k$ (Baek, McLachlan, & Flack, 2010), $\mathbf{D}_k$, and/or $\mathbf{\Sigma}_k$ may be constrained to be equal across clusters.

Parameter estimation

In order to estimate the MFA model parameters, given a specified number of clusters K and factors Q, the following log likelihood function is maximized (McLachlan & Peel, 2000):

$$\log\ L(\psi) = \sum_{i=1}^{I} \log\left[\sum_{k=1}^{K} \pi_k \cdot \varphi(\mathbf{y}_i;\boldsymbol{\mu}_k, \mathbf{\Sigma}_k)\right], \tag{5}$$

with

$$\varphi(\mathbf{y}_i;\boldsymbol{\mu}_k, \mathbf{\Sigma}_k) = \frac{1}{\sqrt{(2\pi)^J|\mathbf{\Sigma}_k|}} e^{-1/2(\mathbf{y}_i - \boldsymbol{\mu}_k)'\mathbf{\Sigma}_k^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_k)} \tag{6}$$

To this end, often an expectation-maximization algorithm, or further extensions thereof (McLachlan & Peel, 2000), is used. In this article, the MFA model will be fitted by means of the EMFAC software package (McLachlan et al., 2003), which adopts the alternating expectation-conditional maximization approach.

Model selection methods for MFA

In general, the numbers of clusters and factors that underlie a given data set are unknown. To assess these numbers, one usually fits a set of MFA models with different numbers of clusters and factors to the data. Next, for each fitted model, a goodness-of-fit measure (or misfit value) is computed, along with a measure of the complexity of the model. The final step consists of selecting the model with the best fit–complexity balance. For this last step, a number of methods are available, three of which will be discussed and compared in this article: AIC, BIC, and CHull.

AIC and BIC

For selecting the optimal number of MFA clusters K and factors Q, McLachlan et al. (2003) recommended the use of information criteria. The most popular information criteria, which therefore will be studied in this article, are AIC and BIC. Both criteria look for a model that describes the data well and that is parsimonious at the same time. To this end, the model fit (i.e., minus two times the log likelihood of the model) is penalized by the model complexity, which equals the number of free parameters $fp = K[2J + JQ - (1/2)Q(Q-1)] + (K-1)$. In particular, AIC is computed as follows (Akaike, 1974):

$$\mathrm{AIC} = -2\log\ \mathrm{L} + 2 \times fp, \tag{7}$$

where the model with the lowest AIC value is the optimal one. AIC is an estimate of the expected relative Kullback–Leibler

divergence and selects the model that asymptotically tries to minimize this distance (Vrieze, 2012).

The use of AIC implies that one assumes that each observation provides new, independent information regarding the underlying model, which may be unrealistic when the sample under study grows large. Therefore, Schwarz (1978) proposed the BIC criterion, which takes the sample size into account, as an alternative:

$$\mathrm{BIC} = -2\log\ \mathrm{L} + \log(\mathrm{I}) \times fp. \tag{8}$$

The BIC has its roots in Bayesian statistics, but paradoxically, it is almost always applied in a frequentist (i.e., non-Bayesian) model selection context. It can be shown that the difference between the BICs of two models can be seen as an asymptotic approximation to the logarithm of the Bayes factor of these models (see Claeskens & Hjort, 2008), which gives an indication of the support for the first model against the second one.[1] An attractive property of the BIC is that it is consistent: The BIC selects the correct model with a probability that goes to 1 as I grows large (see Claeskens & Hjort, 2008). This consistency property will hold given that the true model is among the set of models to choose from and that some other regularity assumptions are satisfied (i.e., the number of parameters is finite and does not grow with I; for more information, see Vrieze, 2012).

On the basis of the theoretical differences between AIC and BIC, Vrieze (2012) presented some rules of thumb for choosing between the model selection methods. In the case that one is interested in identifying the true model (i.e., operating under a zero/one loss function) and that the true model has a fixed and finite number of parameters, BIC is preferable (in the asymptotic situation, because of its consistency property mentioned above). However, when the true model is not in the choice set and/or when this true model is too complex for parametric modeling (e.g., an unknown, highly nonlinear model), AIC will minimize (asymptotically) the mean squared error of estimation more efficiently than BIC. Given that our interest in this article is in selecting the true model, the latter situation of an unknown model not being represented in the choice set is not of immediate interest to us.

It should be noted, however, that these asymptotic results do not simply transfer to finite samples (Vrieze, 2012). In such cases, simulation studies are needed to evaluate and compare the behavior of AIC and BIC. Regarding such simulations, comparing the formulas of AIC and BIC, it can be predicted that AIC will often select more complex models (i.e., models with large *fp* values) than BIC does (see also Vrieze, 2012). Indeed, the only important difference between the two criteria is that, unless the sample size is small, AIC penalizes adding parameters to the model less than BIC does. Consequently, AIC will often select a too complex model, which is a tendency of AIC that has already been observed in the context of many models (e.g., Celeux & Soromenho, 1996; Kass & Raftery, 1995).

[1] The Bayes factor for comparing Models 1 and 2 is the ratio of the posterior odds (of Model 1 vs. Model 2) versus the prior odds (of Model 1 vs. Model 2): $\mathrm{BF}_{12}=[p(M_1|y)/p(M_2|y)]/[p(M_1)/p(M_2)]$. The Bayes factor expresses how much the odds in favor of Model 1 change when observing the data y. If we take the prior odds as being equal to 1 [such that $p(M_1)=p(M_2)=.5$], then BF_{12} equals the posterior odds and expresses how much evidence there is for Model 1 as compared to Model 2 (i.e., if $\mathrm{BF}_{12}>1$, then Model 1 is to be preferred, and vice versa).

CHull

Like AIC and BIC, CHull searches for the model with the best balance of model fit and model complexity. Being a generalized and automated version of the well-known scree test of Cattell (1966), CHull first singles out the models at the higher boundary of the convex hull of a log-likelihood-versus-*fp* plot, as these models have a better balance between model fit and model complexity than do the other models. Second, an optimal model is identified by selecting the hull model after which the gain in fit by adding extra parameters levels off (i.e., the elbow in the scree plot). A clear advantage of the automated CHull method over the scree test is that the elbow in the plot is not determined visually, which always implies a subjective judgment, but numerically.

The CHull procedure allows the researcher to define the fit and complexity measure, so that researchers can tailor these measures according to the research question at hand (Ceulemans & Kiers, 2006). As a consequence, CHull is more general than AIC and BIC, in that CHull can also be used for model selection in deterministic models in which no likelihood can be computed, because no particular assumptions regarding the noise in the data are imposed (e.g., cluster and component analysis), whereas AIC and BIC are restricted to stochastic models only. In the case of MFA, the log likelihood of the model will be taken as the fit measure, and the number of free parameters *fp* as a measure of model complexity.

Specifically, the CHull model selection strategy consists of the following six steps (Ceulemans & Kiers, 2006). First, when different models exist that have the same number of free parameters *fp*, only retain the model with the largest log likelihood. Second, rank the N retained models according to their numbers of free parameters and denote them by s_n ($n = 1, \ldots, N$). Third, omit all models s_n for which there exists a less complex model s_j ($j < n$) that has a larger log likelihood value. Fourth, considering each triplet of adjacent models, exclude the middle model if it is located below or on the line connecting the other two models of the triplet in a plot of the log likelihood versus the number of free parameters. In

addition, all models for which the gain in fit is less than 1 % of the fit value of the preceding (less complex) model are also discarded (because such small fit gains may unduly influence the results; Wilderjans, Ceulemans, & Meers, in press). Repeat this step until no more models can be discarded. These first four CHull steps yield the models situated on the upper boundary of the convex hull. In order to find the optimal hull model, the fifth step comprises computing the following scree test value *st* for each hull model:

$$st_n = \frac{\left\{\frac{\log L_n - \log L_{n-1}}{fp_n - fp_{n-1}}\right\}}{\left\{\frac{\log L_{n+1} - \log L_n}{fp_{n+1} - fp_n}\right\}}, \quad (9)$$

with $\log L_n$ and fp_n denoting the log likelihood and the number of free parameters, respectively, for the *n*th model. In Eq. 9, the numerator and denominator represent the slopes of successive parts of the upper boundary of the convex hull (Ceulemans et al., 2011) and pertain to the (average) increase in likelihood per added parameter. A large st_n ratio then implies that model s_n fits the data clearly better than solution s_{n-1} (i.e., a relatively large increase in model fit per added parameter), whereas s_{n+1} only implies a relatively small increase in model fit per added parameter. Note that no *st* value can be computed for the least and most complex models on the boundary of the convex hull. The sixth and final step consists of selecting the solution associated with the largest *st* value. Note that free software for applying the CHull procedure is available from http://ppw.kuleuven.be/okp/software/chull/ (Wilderjans et al., in press).

Uncertainty in the model selection statistics

Besides selecting an optimal model, the presented model selection strategies also allow for a ranking of interesting models. Moreover, the AIC, BIC, and *st* values are in fact statistics computed from the data, and they are subject to sampling fluctuations. To take into account this uncertainty in the model selection statistics, we may try to quantify (by analytical tools or by simulation) the uncertainty by means of, for instance, a standard error for the AIC, BIC, or *st* values. Such an approach would be rather intricate, and therefore we will use a crude way of accounting for uncertainty in the model selection criteria by considering the three best models as identified by each strategy (as has been advised by Ceulemans & Kiers, 2006). It further should be noted that, in practice, it is not recommended to base one's decision regarding the optimal model on an automated model selection procedure solely, but to also take substantive arguments and the interpretability of the results into account (Ceulemans & Kiers, 2006; Wilderjans et al., in press).

Application to an example data set

To illustrate the different model selection strategies, they will be applied to the iris data (Anderson, 1935, 1936), which contain the scores of 150 irises for the following four morphologic variables: length of the sepal, width of the sepal, length of the petal, and width of the petal. Each iris belongs to one of the following three species of iris flowers: Setosa, Versicolor, and Virginica (for didactic reasons, we use data with a known cluster allocation).

Table 1 displays the number of free parameters *fp*, the log likelihood, and the AIC, BIC, and *st* values for MFA models of the iris data, with the number of clusters varying from one to five and the number of factors varying from one to three. On the basis of AIC, the model with five clusters and two factors would be selected, since this model has the lowest AIC value (i.e., 423.5). The clustering corresponding to this selected model (by assigning each iris to the cluster with the largest posterior probability) shows that for all species, the flowers are distributed over three or more clusters. In contrast to the overestimation by AIC, the lowest BIC value (i.e., 582.8) is encountered for the model with three clusters and one factor. The clustering based on this model coincides with the three species of irises, with the exception of only three flowers.

When the first four CHull steps are applied to the MFA models in Table 1, the following four models appear to be

Table 1 Numbers of free parameters (*fp*), log likelihoods, AIC and BIC values, and *st* values (only for hull models) for MFA models fitted to the iris data, with the number of clusters varying from one to five and the number of factors from one to three

Number of Clusters	Number of Factors	Number of Free Parameters (*fp*)	Log Likelihood	AIC	BIC	*st* Value
1	1	12	−423.9	871.8	907.9	–*
	2	15	−391.0	812.0	857.1	
	3	17	−379.6	793.3	844.5	
2	1	25	−232.6	515.2	590.4[b]	5.3[a]
	2	31	−217.9	497.8	591.2[c]	
	3	35	−215.2	500.3	605.7	
3	1	38	−196.2	468.4	582.8[a]	1.8[b]
	2	47	−182.9	459.8	601.3	
	3	53	−181.0	468.0	627.6	
4	1	51	−180.7	463.4	616.9	
	2	63	−163.5	453.0	642.6	
	3	71	−157.8	457.6	671.4	
5	1	64	−159.9	447.8[c]	640.5	
	2	79	−132.7	423.5[a]	661.3	–*
	3	89	−134.6	447.2[b]	715.2	

[a] Best-fitting model. [b] Second-best-fitting model. [c] Third-best-fitting model. * For the first and last hull models, the *st* value is not defined

located on the upper boundary of the convex hull (located on the red line in Fig. 1): (a) a model having one cluster and one factor, (b) a model with two clusters and one factor, (c) a model with three clusters and one factor, and (d) a model with five clusters and two factors. Given the associated *st* values for these hull models, which can be found in Table 1, CHull selects the model with two clusters and one factor (i.e., an *st* value of 5.3), with the next-best-fitting model being the one with three clusters and one factor (i.e., an *st* value of 1.8). Note that CHull selects the model with three clusters (i.e., three species of irises) only as the second-best model, whereas BIC identifies this model as the best one.

Simulation study

Research questions

In this simulation study, we will examine to what extent AIC, BIC, and CHull succeed in correctly identifying the numbers of underlying clusters and factors. As suggested above, for each model selection strategy, we will discuss the results concerning the single best model, but we will also look at the best three models to account for uncertainty in the model selection statistics. From the theoretical discussion on AIC and BIC (Vrieze, 2012; Wagenmakers & Farrell, 2004), we formulate the following expectation about the performance of both model selection methods: Since the true model is always under consideration and the number of parameters of the true model is fixed and finite, we conjecture that BIC will outperform AIC.

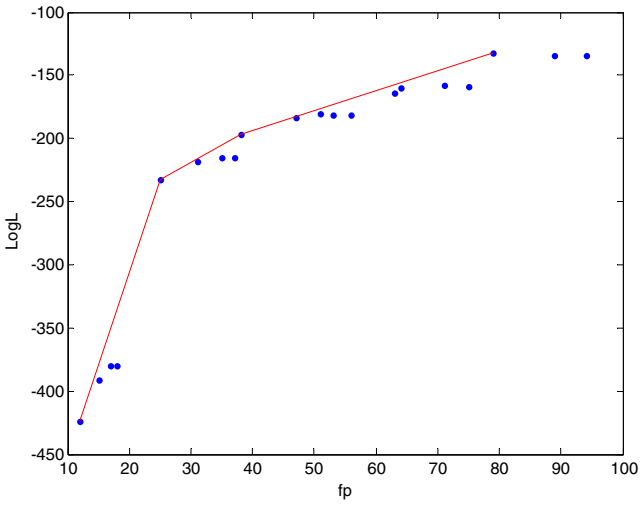

Fig. 1 Numbers of free parameters *fp* plotted against the log likelihood Log *L* for the MFA models that were fitted for the iris data, with the number of clusters ranging from one to five and the number of factors varying from one to three (see also Table 1). The upper boundary of the convex hull is indicated by a red line

Moreover, the effect of the following five data characteristics on the performance of the different model selection strategies will be investigated: (a) the number of clusters, (b) the distribution of the objects across the clusters, (c) the number of factors, (d) the amount of noise in the data, and (e) the degree of overlap between the clusters. In simulations of different clustering techniques, it appears that retrieving the optimal clustering becomes harder when the number of clusters increases and when (many) small clusters exist (e.g., Milligan & Cooper, 1985; Schepers et al., 2008). Furthermore, we expect that when the number of underlying factors becomes larger (Lorenzo-Seva et al., 2011), when the data contain a large amount of noise (Ceulemans & Kiers, 2006), and when the clusters overlap more in mean level or covariance structure (De Roover, Ceulemans, & Timmerman, 2012), model selection performance will decrease and the differences between the model selection methods under study will become more pronounced (see, e.g., Schepers et al., 2008).

Design and procedure

In this simulation study, data sets with 12 variables and 200 observations were constructed. The following five factors were systematically manipulated in a completely randomized design:

1. The number of clusters, at two levels: two and four;
2. The distribution of objects across the clusters, at three levels: clusters of equal size (i.e., equal-size condition), one cluster containing 60 % of the objects and the remaining objects being equally distributed over the other clusters (i.e., one-large-cluster condition), and one cluster containing 10 % of the objects and the remaining objects being equally distributed across the remaining clusters (i.e., one-small-cluster condition);
3. The number of factors, at two levels: two and four;
4. The noise level ε, at three levels: .05, .15, and .25;
5. The amount of overlap between the clusters, at two levels: a small amount of overlap and a moderate amount of overlap.

To construct a data set, the following procedure was applied. First, the true mixing probabilities $\boldsymbol{\pi}^{\text{true}}$ ($1 \times K$) were determined on the basis of the number of clusters and the distribution of the objects across clusters, and the objects were accordingly assigned to one of the clusters. Second, for the objects that belong to a particular cluster, factor scores $\mathbf{u}_k^{\text{true}}$ and residual $\mathbf{e}_{ik}$ ($1 \times J$) values were drawn from a multivariate normal distribution—with the 0-vector as the mean vector and the identity matrix as the covariance matrix—and were orthogonalized (ensuring the independence of $\mathbf{u}_k^{\text{true}}$ and $\mathbf{e}_{ik}$) and standardized. Third, cluster-specific true loadings $\mathbf{B}_k^{\text{true}}$ ($k = 1, \ldots, K$) were generated

by sampling numbers from a uniform distribution on the interval [−1, 1]. Fourth, these cluster-specific loadings were rescaled in order to make their sum of squares per variable equal to $1 - \varepsilon$. Fifth, to manipulate the amount of overlap, cluster-specific true mean vectors $\boldsymbol{\mu}_k^{\text{true}}$ $(1 \times J)$ were specified, as is shown in Table 2. Sixth, combining the cluster-specific true loadings $\mathbf{B}_k^{\text{true}}$, the cluster-specific true factor scores $\mathbf{u}_k^{\text{true}}$, the cluster-specific true means $\boldsymbol{\mu}_k^{\text{true}}$, and the true cluster memberships $\boldsymbol{\pi}^{\text{true}}$ yielded true data $\mathbf{X}^{\text{true}}$. Seventh, the $\mathbf{e}_{ik}$ values were rescaled such that their variances equalled ε per variable. In a final step, the noise was added to the true data.

In order to quantify for each generated data set the amount of overlap between the clusters, we computed for each pair of clusters $k1$ and $k2$ the amount of overlap by using the following overlap measure from Lu, Smith, and Good (1989), assuming multivariate normality:

$$\lambda = \frac{2|\boldsymbol{\Sigma}_{k1}\boldsymbol{\Sigma}_{k2}|^{1/2}}{\left|\left(\frac{1}{2}\right)(\boldsymbol{\Sigma}_{k1} + \boldsymbol{\Sigma}_{k2})\right|^{1/2}\left(|\boldsymbol{\Sigma}_{k1}|^{1/2} + |\boldsymbol{\Sigma}_{k2}|^{1/2}\right)} \times e^{\left[-\left(\frac{1}{2}\right)(\boldsymbol{\mu}_{k1}-\boldsymbol{\mu}_{k2})'(\boldsymbol{\Sigma}_{k1}+\boldsymbol{\Sigma}_{k2})^{-1}(\boldsymbol{\mu}_{k1}-\boldsymbol{\mu}_{k2})\right]}, \quad (10)$$

with $\boldsymbol{\mu}_{k1}$ and $\boldsymbol{\mu}_{k2}$ being specified as in Table 2, and $\boldsymbol{\Sigma}_{k1}$ and $\boldsymbol{\Sigma}_{k2}$ computed according to Eq. 4. The λ measure varies between 0 and 1, with 0 meaning the absence of overlap and 1 indicating perfect overlap. Next, a total overlap measure δ was computed by taking the average of the pairwise overlap values. In the conditions with a small amount of overlap, the overlap values δ ranged from 0 to .000394, with a mean value of .000005 (SD = .000031). In the moderate overlap conditions, the overlap values ranged from 0 to .1083, with a mean of .0239 (SD = 0.0280). Note that, when considering each combination of the levels of the four other variables separately, data sets from the moderate-overlap condition clearly had a larger amount of overlap than did data sets from the small-overlap condition.

For each of the 2 × 3 × 2 × 3 × 2 = 72 conditions, ten replications were generated, yielding 720 simulated data sets. Next, MFA was applied to each simulated data set, with the numbers of clusters and factors varying from one to seven. To this end, the EMFAC software package (McLachlan et al., 2003) was used, with the following analysis options: (a) 100 random starts (with 70 % of the data being used to determine random starting values), (b) ten K-means starts, (c) no standardization of the variables (since the procedure already yields standardized data), (d) no restrictions on the covariance matrices $\boldsymbol{\Sigma}_k$ or the diagonal matrices $\mathbf{D}_k$ across clusters, (e) not starting the algorithm from a user-defined initial clustering of the objects, (f) using normal distributions for the mixture components, and (g) using a rational initialization for the cluster-specific loadings. The simulation study was programmed in MATLAB R2011b and conducted on a supercomputer consisting of INTEL XEON L5420 processors with a clock frequency of 2.5 GHz and with 8 GB RAM.

Table 2 True cluster-specific mean vectors $\boldsymbol{\mu}_k^{\text{true}}$ for the different combinations of the levels of the number of clusters and the amount of overlap between the clusters

Number of Clusters	Amount of Cluster Overlap*	
	Small Amount	Moderate Amount
2	[1 –1]; [−1 1]	[.25 –.25]; [−.25 .25]
4	[1 –1]; [−1 1]; [1 1]; [−1 −1]	[.25 –.25]; [−.25 .25]; [.25 .25]; [−.25 –.25]

* [1 –1] indicates that the means for the first six variables equal 1 and the means for the next six variables equal −1

Results

When examining how often the correct numbers of underlying clusters and factors were identified by the three model selection methods, it could first be concluded that AIC has a low success rate. In particular, only considering the first choice, AIC succeeded in selecting the correct model for only 10 % of the data sets (see Table 3). Taking the first three choices into account, the success rate rose to about 25 % (i.e., the correct model was among the three best AIC models in only 173 of the 720 data sets). When investigating the model selection mistakes more closely, it appears that AIC tended to overfit by selecting a too complex model. In particular, the least complex model of the three choices of AIC was more complex (i.e., had a higher *fp* value) than did the correct model for 546 of the 547 data sets in which AIC failed to select the correct model. Because of the poor performance of AIC, in the remainder of this discussion, the focus will be on the comparison among BIC and CHull.

First choice only When considering the first choice only, BIC performed a bit better than CHull. In particular, the best BIC model was the correct one in 630 of the 720 data sets (87.5 %), whereas CHull retrieved the correct model in 612 cases (85 %). In order to further compare BIC and CHull, in Table 4 a cross-classification is presented of the model

Table 3 Numbers of data sets for which the correct model was among the first three models that were retained by AIC, BIC, and CHull

	First Choice	Second Choice	Third Choice	Total
AIC	72	68	33	173
BIC	630	25	4	659
CHull	612	59	17	688

Table 4 Cross-classification of the selection performance of CHull and BIC, when considering either the first choice only or the first three choices

	Choice		BIC	
			Incorrect	Correct
CHull	First choice only	Incorrect	63	45
		Correct	27	585
	First three choices	Incorrect	29	3
		Correct	32	656

selection results of both methods. Regarding the first choice, it appears that both methods retrieved the correct model in 81 % of the cases and that both of them failed for 9 % of the data sets. The latter 63 data sets mainly belonged to the conditions with four underlying clusters and factors (i.e., 51 cases, or 81.0 % of the failures) and with large amounts of noise in the data (i.e., 51 cases, or 81.0 %, were located in the 25 %-noise condition). For these data sets, both BIC and CHull tended to underestimate the complexity of the optimal model. In particular, CHull selected a too simple model (i.e., too small *fp* value) in 60 out of the 63 cases, and BIC in 62. When focusing on the 72 data sets for which only one method failed on the first choice, one can see in Table 4 that BIC performed better in 45 cases (62 %) and CHull in 27 cases. The 45 data sets for which BIC outperformed CHull were mainly located in the conditions with larger amounts of noise (i.e., only three data sets—7 %—belonged to the 5 %-error conditions) and spread out (in a nonsystematic way) over the levels of the other manipulated data characteristics. In most cases (i.e., 30 out of the 45 data sets) CHull selected a model with a lower *fp* value than the correct model, indicating underestimation of the model complexity. The 27 data sets for which CHull outperformed BIC mainly belonged to the more difficult conditions (i.e., four clusters and/or four factors, a large amount of noise, and a moderate amount of cluster overlap), with BIC in all cases underestimating the complexity of the model. In Table 5, the numbers of data sets for which the model selection tools succeeded in identifying the correct numbers of underlying clusters and factors are presented for each level of the five manipulated characteristics. For most levels, BIC performed better than CHull. When the underlying data contained four clusters, CHull outperformed BIC.

First three choices When taking the first three choices into account, CHull clearly outperformed BIC, in that CHull had a success rate of 95.6 % (i.e., 688 of the 720 data sets), whereas BIC identified the correct model in 659 of the 720 cases (91.5 %). In order to check whether the better performance of CHull over BIC (considering the first three choices) varied over the manipulated data characteristics, in Table 6, for each level of the five manipulated characteristics, the numbers of data sets are presented for which the correct model was among the three retained models. From this table, one can see that for all characteristics, CHull outperformed BIC, with this effect being more pronounced when the data became more challenging for all methods (i.e., larger numbers of underlying clusters and factors, larger amounts of noise, and stronger cluster overlap). Therefore, it is not surprising that the cross-classification in Table 4 (for the first three choices) shows that CHull succeeded and BIC failed for 32 cases, mostly because BIC underestimated the number of clusters and/or factors, whereas the opposite (i.e., BIC succeeding and CHull failing) was true for three data sets only. Finally, note that considering additional models hardly changed the results, as the fourth- or fifth-best models according to BIC and CHull were the correct ones for very few data sets (i.e., at most six).

Table 5 Numbers of data sets for which the three model selection methods selected the correct model, taking only the first choice into account, for all levels of the five manipulated data characteristics

Characteristic	Level	AIC	BIC	CHull
Number of clusters	2	0	352	325
	4	72	278	287
Distribution of objects across clusters	Equal size	21	223	223
	One large cluster	23	201	197
	One small cluster	28	206	192
Number of factors	2	35	352	343
	4	37	278	269
Amount of noise	0.05	2	240	237
	0.15	20	219	207
	0.25	50	171	168
Amount of cluster overlap	Small	40	331	320
	Moderate	32	299	292

Table 6 Numbers of data sets for which the three model selection methods selected the correct model, taking the first three choices into account, for all levels of the five manipulated data characteristics

Characteristic	Level	AIC	BIC	CHull
Number of clusters	2	0	357	359
	4	173	302	329
Distribution of objects across clusters	Equal size	58	229	239
	One large cluster	56	205	218
	One small cluster	59	225	231
Number of factors	2	87	355	359
	4	86	304	329
Amount of noise	0.05	10	240	240
	0.15	69	230	231
	0.25	94	189	217
Amount of cluster overlap	Small	91	347	353
	Moderate	82	312	335

Discussion and concluding remarks

The present study aimed at evaluating three model selection methods in the context of MFA: AIC, BIC, and CHull. The simulation study showed overall poor performance of AIC, in that it almost always selected an overly complex model. In general, BIC and CHull performed well. Comparing AIC and BIC, the results were in line both with theory and simulation results: In the case of a zero/one loss function (such that the true model was in the choice set), BIC showed its consistency property and AIC tended to overfit (see Vrieze, 2012).

Considering only the first choices of the model selection strategies, BIC was the most accurate model selection statistic. However, when taking model selection uncertainty into account by examining the three best models, as retained by each model selection procedure, CHull outperformed BIC. Inspecting the effect of the manipulated data characteristics, CHull was better than BIC in the difficult—but more realistic—conditions (i.e., more underlying clusters and factors, large amounts of noise and cluster overlap). In these cases, BIC appeared to underestimate the numbers of underlying clusters and factors. The better performance of CHull in the difficult simulation conditions may be caused by the fact that BIC, unlike CHull, takes the sample size into account, and therefore penalizes model complexity too harshly.

An important remark concerning the presented simulation study is that the generated data deviated from the stochastic model in Eq. 3, in that each sample that was generated had characteristics (e.g., orthogonality of the factor scores and residuals) that are expected to be true for the stochastic model (i.e., when J goes to infinity) only. An advantage of this approach is that exact manipulation of the five factors became possible. A disadvantage, however, is that sampling fluctuations are not taken into account, presumably making it slightly easier for the model selection strategies to identify the correct numbers of underlying clusters and factors. In future research, the effect of sampling fluctuations on model selection performance could be investigated by incorporating a sampling approach.

On the basis of the simulation study, it can be concluded that considering the first three choices provides a proper balance between taking uncertainty in the model selection statistics into account and determining the (optimal) numbers of underlying clusters and factors (which is the purpose of applying model selection statistics in the first place). Concerning the rankings of the models implied by the different model selection strategies, a clear difference exists between AIC and BIC, on the one hand, and CHull, on the other. Even though all three model selection methods provide a ranking of interesting models, CHull is stricter, in that it only ranks those models that show a good balance between model fit and model complexity (i.e., hull models) by means of the *st* values. Furthermore, these *st* values give an indication of the relative differences in quality of the hull models. Similar features, however, are also offered by extensions of AIC and BIC. Specifically, Akaike weights can be calculated and used to compute for each model the probability that this model is the optimal one, given a range of alternative models (Wagenmakers & Farrell, 2004).

Another remark pertains to defining the complexity of an MFA model as the number of free parameters *fp*. First of all, as was argued by Pitt, Myung, and Zhang (2002), model complexity is also influenced by the functional form of the model. Second, defining complexity in this way implies that each parameter, regardless of its type (i.e., factor loading, unique variance, variable-specific mean, or a priori probability), has the same weight when determining the complexity of a model. In the context of MFA, this may seem like comparing apples and oranges, in that it is not clear whether or not a factor loading should be given the same weight as a unique variance or an a priori probability parameter. Therefore, it may be worthwhile to investigate whether and how such adding of different types of model parameters can be avoided and how the functional form of the model can be taken into account.

Finally, with this study we aimed to provide a thorough and systematic evaluation of the performances of AIC, BIC, and CHull in the context of MFA. It would be interesting, however, to evaluate how useful CHull is for picking the correct model out of a wider range of models—including EFA, MFA, and regular mixture models—for instance, using the simulation design of Lubke and Neale (2006).

Author note The research reported in this article was partially supported by the Fund for Scientific Research (FWO)–Flanders (Belgium), Project No. G.0477.09 awarded to E.C., Marieke Timmerman, and Patrick Onghena, and by the Research Council of KU Leuven (Grant No. GOA/2010/02). T.F.W. is a postdoctoral fellow of the FWO–Flanders (Belgium).

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, AC-19,* 716–723. doi:10.1109/TAC.1974.1100705

Anderson, E. (1935). The Irises of the Gaspé peninsula. *Bulletin of the American Iris Society, 59,* 2–5.

Anderson, E. (1936). The species problem in Iris. *Annals of the Missouri Botanical Garden, 23,* 457–509.

Baek, J., McLachlan, G. J., & Flack, L. K. (2010). Mixtures of factor analyzers with common factor loadings: Applications to the clustering and visualization of high-dimensional data. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 32,* 1298–1309. doi:10.1109/TPAMI.2009.149

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research, 1,* 245–276. doi:10.1207/s15327906mbr0102_10

Celeux, G., & Soromenho, G. (1996). An entropy criterion for assessing the number of clusters in a mixture model. *Journal of Classification, 13,* 195–212. doi:10.1007/BF01246098

Ceulemans, E., & Kiers, H. A. L. (2006). Selecting among three-mode principal component models of different types and complexities: A numerical convex hull based method. *British Journal of Mathematical and Statistical Psychology, 59,* 133–150. doi:10.1348/000711005X64817

Ceulemans, E., & Kiers, H. A. L. (2009). Discriminating between strong and weak structures in three-mode principal component analysis. *British Journal of Mathematical and Statistical Psychology, 62,* 601–620. doi:10.1348/000711008X369474

Ceulemans, E., Timmerman, M. E., & Kiers, H. A. L. (2011). The CHull procedure for selecting among multilevel component solutions. *Chemometrics and Intelligent Laboratory Systems, 106,* 12–20. doi:10.1016/j.chemolab.2010.08.001

Ceulemans, E., & Van Mechelen, I. (2005). Hierarchical classes models for three-way three-mode binary data: Interrelations and model selection. *Psychometrika, 70,* 461–480. doi:10.1007/s11336-003-1067-3

Claeskens, G., & Hjort, N. L. (2008). *Model selection and model averaging*. Cambridge, U.K.: Cambridge University Press.

De Roover, K., Ceulemans, E., & Timmerman, M. E. (2012). How to perform multiblock component analysis in practice. *Behavior Research Methods, 44,* 41–56. doi:10.3758/s13428-011-0129-1

Ghahramani, Z., & Hinton, G. E. (1997). *The EM algorithm for mixtures of factor analyzers* (Technical Report CRG-TR-96-1). Retrieved from http://mlg.eng.cam.ac.uk/zoubin/papers/tr-96-1.pdf

Henson, J. M., Reise, S. P., & Kim, K. H. (2007). Detecting mixtures from structural model differences using latent variable mixture modeling: A comparison of relative model fit statistics. *Structural Equation Modeling, 14,* 202–226. doi:10.1080/10705510709336744

Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika, 23,* 187–200. doi:10.1007/BF02289233

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association, 90,* 773–795. doi:10.1080/01621459.1995.10476572

Lorenzo-Seva, U., Timmerman, M. E., & Kiers, H. A. L. (2011). The Hull method for selecting the number of common factors. *Multivariate Behavioral Research, 46,* 340–364. doi:10.1080/00273171.2011.564527

Lu, R.-P., Smith, E. P., & Good, I. J. (1989). Multivariate measures of similarity and niche overlap. *Theoretical Population Biology, 35,* 1–21. doi:10.1016/0040-5809(89)90007-5

Lubke, G. H., & Neale, M. C. (2006). Distinguishing between latent classes and continuous factors: Resolution by maximum likelihood? *Multivariate Behavioral Research, 41,* 499–532. doi:10.1207/s15327906mbr4104_4

McLachlan, G. J., Baek, J., & Rthnayake, S. I. (2011). Mixtures of factor analysers for the analysis of high-dimensional data. In K. L. Mengersen, C. P. Robert, & D. M. Titterington (Eds.), *Mixtures: Estimation and application* (pp. 189–212). Chichester, U.K.: Wiley.

McLachlan, G. J., & Peel, D. (2000). *Finite mixture models*. New York, NY: Wiley.

McLachlan, G. J., Peel, D., & Bean, R. W. (2003). Modelling high-dimensional data by mixtures of factor analyzers. *Computational Statistics and Data Analysis, 41,* 379–388. doi:10.1016/S0167-9473(02)00183-4

Milligan, G. W., & Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika, 50,* 159–179.

Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling, 14,* 535–569. doi:10.1080/10705510701575396

Pitt, M. A., Myung, I. J., & Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review, 109,* 472–491. doi:10.1037/0033-295X.109.3.472

Schepers, J., Ceulemans, E., & Van Mechelen, I. (2008). Selecting among multi-mode partitioning models of different complexities: A comparison of four model selection criteria. *Journal of Classification, 25,* 67–85. doi:10.1007/s00357-008-9005-9

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6,* 461–464. doi:10.1214/aos/1176344136

Vrieze, S. I. (2012). Model selection and psychological theory: A discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). *Psychological Methods, 17,* 228–243. doi:10.1037/a0027127

Wagenmakers, E.-J., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin & Review, 11,* 192–196. doi:10.3758/BF03206482

Wilderjans, T. F., Ceulemans, E., & Meers, K. (in press). CHull: A generic convex-hull-based model selection method. *Behavior Research Methods*. doi:10.3758/s13428-012-0238-5