Model selection strategies for determining the optimal number of overlapping clusters in additive overlapping partitional clustering

Rossbroich, J.; Durieux, J.; Wilderjans, T.F.

Journal of Classification (2022) 39:264–301
https://doi.org/10.1007/s00357-021-09409-1

Model Selection Strategies for Determining the Optimal Number of Overlapping Clusters in Additive Overlapping Partitional Clustering

Julian Rossbroich[1] · Jeffrey Durieux[2,3] · Tom F. Wilderjans[2,3,4,5]

Accepted: 17 December 2021 / Published online: 17 January 2022

Abstract

In various scientific fields, researchers make use of partitioning methods (e.g., *K*-means) to disclose the structural mechanisms underlying object by variable data. In some instances, however, a grouping of objects into clusters that are allowed to overlap (i.e., assigning objects to multiple clusters) might lead to a better representation of the underlying clustering structure. To obtain an overlapping object clustering from object by variable data, Mirkin's ADditive PROfile CLUStering (ADPROCLUS) model may be used. A major challenge when performing ADPROCLUS is to determine the optimal number of overlapping clusters underlying the data, which pertains to a model selection problem. Up to now, however, this problem has not been systematically investigated and almost no guidelines can be found in the literature regarding appropriate model selection strategies for ADPROCLUS. Therefore, in this paper, several existing model selection strategies for *K*-means (a.o., CHull, the Caliński-Harabasz, Krzanowski-Lai, Average Silhouette Width and Dunn Index and information-theoretic measures like AIC and BIC) and two cross-validation based strategies are tailored towards an ADPROCLUS context and are compared to each other in an extensive simulation study. The results demonstrate that CHull outperforms all other model selection strategies and this especially when the negative log-likelihood, which is associated with a minimal stochastic extension of ADPROCLUS, is used as (mis)fit measure. The analysis of a post hoc AIC-based model selection strategy revealed that better performance may be obtained when a different—more appropriate—definition of model complexity for ADPROCLUS is used.

✉ Tom F. Wilderjans
t.f.wilderjans@fsw.leidenuniv.nl

[1] Friedrich Miescher Institute for Biomedical Research, Basel, Switzerland

[2] Methodology and Statistics Research Unit, Institute of Psychology, Faculty of Social and Behavioral Sciences, Leiden University, Pieter de la Court Building, Leiden, The Netherlands

[3] Leiden Institute for Brain and Cognition (LIBC), Leids Universitair Medisch Centrum (LUMC), Leiden, The Netherlands

[4] Research Group of Quantitative Psychology and Individual Differences, Faculty of Psychology and Educational Sciences, KU Leuven, Leuven, Belgium

[5] Department of Clinical Psychology, Faculty of Behavioural and Human Movement Sciences, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

Keywords Overlapping clustering · Additive profile clustering · ADPROCLUS · Model selection · Choosing the number of clusters · CHull · AIC · BIC

1 Introduction

In various domains of research, clustering models are used to summarize and reveal the structural mechanisms underlying data regarding a set of objects (e.g., object by variable data or object by object similarity data). Empirical and theoretical arguments may support the hypothesis that these underlying mechanisms can be captured by a grouping of the objects into homogenous clusters. A psychological researcher, for example, might investigate a patient population and collect data about the extent to which these patients express a number of clinical symptoms, like sleep disturbances, irritability, concentration problems, psychomotor retardation, or weight loss. In this example, the underlying mechanisms may be disclosed by grouping the patients into clusters such that patients from the same cluster show similar symptom profiles; the underlying mechanisms then pertain to the syndrome(s), like major depressive disorder or generalized anxiety disorder, that each patient group suffers from. As a second example, take network data, which are commonly encountered in computer, biological or social sciences (e.g., the Worldwide Web, power grids, food webs, metabolic networks, acquaintance networks and collaboration networks; Girvan & Newman, 2002; Kanaya et al., 2014; Scott, 2000), in which for each object pair information is available on the strength to which objects are connected (e.g., similarity data). In this type of data, natural divisions and community structures among the objects may be revealed by grouping strongly (inter)connected objects.

Three types of cluster structures (for a graphical representation of these three types, see Fig. 1 of Wilderjans et al., 2011) can be distinguished that may underlie such data (a detailed review of the three clustering types can be found in Van Mechelen et al., 2004). A first type of clustering is partitioning, which separates the objects into a set of exhaustive and mutually exclusive clusters, which implies that clusters are not allowed to overlap (i.e., no object can belong to more than a single cluster). Well-known partitioning methods include *K*-means (Forgy, 1965; Hartigan & Wong, 1979; Jancey, 1966; Macqueen, 1967; Steinheus, 1956), latent clustering (Lazarsfeld & Henry, 1968) and model-based clustering (Fraley & Raftery, 2002; Mclachlan & Chang, 2004). Although model-based clustering methods yield probabilistic or fuzzy clusterings (i.e., for each data object, the methods estimate the probability of belonging to each cluster), which could be interpreted as overlapping clusterings, these methods assume an underlying object partition. A basic assumption underlying the mixture model in a model-based cluster analysis is that each data object is generated from only a single mixture component (i.e., cluster), with the posterior probabilities indicating the probability that a specific data object is generated from each particular mixture component. Therefore, the assumptions of model-based clustering methods imply a partitioning of the data and do not assume that clusters overlap in the sense that objects can belong to more than one cluster. Hence, we categorize model-based clustering methods under the first type of cluster structures.

A second type of clustering pertains to hierarchical clustering methods (Duda et al., 2001; Gordon, 1987, 1996; Sneath & Sokal, 1973). These methods need a (dis)similarity object by object matrix as input. When originally having an object by variable data matrix, this matrix can be transformed into a (dis)similarity matrix by, for instance, computing correlations or Euclidean distances between the data objects. Hierarchical clustering methods yield

a set of nested clusters in which each cluster can be obtained by merging objects and/or smaller clusters. The nestedness of the clusters implies that, for each pair of clusters, the intersection of two clusters consists of the smaller of both clusters or equals the empty cluster. Note that in case of the former, this indicates that object clusters overlap, which implies that some objects belong to multiple clusters. Most commonly used hierarchical clustering methods are agglomerative (i.e., repeatedly merging objects/clusters that are most similar) and divisive (i.e., recursively splitting clusters into smaller ones) methods (Manning et al., 2008).

Lastly, overlapping clustering methods can be distinguished as a third type of clustering. In this type of clustering, as is true for hierarchical clustering, an object may be a member of multiple clusters at the same time. The main difference between both types of clustering is that hierarchical methods only allow for cluster overlap in terms of nested clusters (i.e., two clusters are only allowed to overlap in the sense that one cluster is a subset of the other one), whereas in overlapping clustering methods the cluster overlap is not restricted in any way (see Fig. 1 of Wilderjans et al., 2011). The advantage of overlapping clustering, compared to the other two clustering types, is the possibility to account for more complex underlying cluster structures which may lead to more accurate representations of the sometimes intricate reality. In the patient by symptom data set, for example, in which clusters represent subjects with the same clinical syndrome, a patient might suffer from more than one syndrome at the same time (i.e., syndrome co-morbidity; Khanmohammadi et al., 2017). Such a patient should logically be assigned to multiple clusters. In the case of community selection in network data, an object might be strongly connected to two or more communities and it would therefore be sensible to allow the object to be a member of both clusters (Ding et al., 2016).

Besides interesting from a theoretical point of view (see above), the necessity of using overlapping clusters has also been acknowledged in empirical research (Banerjee et al., 2005). First, overlapping clustering is often successfully used in community detection problems (Tang & Liu, 2009; Wang et al., 2010; Fellows et al., 2011; Bonchi et al., 2013). Take as an example social networks in which subjects could be a member of multiple communities (Azaouzi et al., 2019), biological protein networks where proteins are part of various protein complexes simultaneously (Palla et al., 2008), or genetic networks, in which genes influence multiple cellular functions and as such belong to several metabolic pathways (Segal et al., 2003; Battle et al., 2004; Hastie et al., 2000). Second, in video classification, overlapping clustering is a necessary requirement to meaningfully group multi-genre movies based on the defining themes (e.g., a movie may simultaneously belong to the genres of action, horror and science fiction; Snoek et al., 2006). Third, in emotion detection, overlapping clustering methods are used to group text documents or music pieces based on the emotions they elicit (Wieczorkowska et al., 2006). Fourth, text clustering makes use of overlapping clustering methods to cluster text documents based on the topics discussed in them (Gil-GarcÍa & Pons-Porrata, 2010; PèRez-Suàrez et al., 2013b) or to learn patterns of cross-posting to several newsgroups (Lang, 1995). Finally, in the context of clustering in graphs, Khandekar et al. (2012) found that overlapping clusterings outperform non-overlapping ones in terms of optimizing the conductance over the clusters. In all of these examples, non-overlapping clusterings are not able to fully capture the rich cluster structure underlying the data. For instance, very similar multi-genre movies that, for example, contain both science fiction and horror themes may arbitrarily get allocated to separate clusters. Indeed, some of these movies may be placed in a science fiction cluster and separated from similar movies in which the science fiction theme is less pronounced compared to the horror theme. Indeed,

any partitional algorithm will arbitrarily split an overlapping multi-genre movie cluster into two or more separate genre clusters, herewith obscuring the meaningful overlapping cluster structure. In this paper, we will only pay attention to methods for overlapping clustering.

In the literature, several methods were proposed to obtain overlapping clusters (for an overview, see Ben N'Cir et al., 2015; Baadel et al., 2016; Xie et al., 2013). In his review, Ben N'Cir et al. (2015) notes that most existing overlapping clustering methods can be conceived as extensions of more classical clustering methods, which according to the authors can be categorized into four classes of methods (i.e., method families): hierarchical, graph-based, generative and partitional methods. Besides these four main classes of methods, other methods were recently proposed that tackle the problem of overlapping clustering in alternative ways. For example, extensions of correlation clustering (Bonchi et al., 2013) and topological maps (Cleuziou, 2013), the latter being an extension of Self-Organizing-Maps (SOM; Kohonen, 1995). The four main classes of overlapping clustering methods will be briefly discussed here. First, hierarchical clustering methods aim to decrease the discrepancies between the original dissimilarities between the data objects and the dissimilarities predicted by the hierarchical structure. An example of an overlapping variant of the hierarchical clustering methods is the weak-hierarchies methods of Bertrand and Janowitz (2003). These methods are, in general, too restrictive regarding the overlap that is allowed. Second, graph-theory based methods, an example of which is Overlapping Clustering based on Relevance (OClustR; PėRez-Suàrez et al., 2013a), are mostly used in the context of community detection in complex networks (Zhang et al., 2007; Davis & Carley, 2008; Wang & Fleury, 2011; Fellows et al., 2011). Although relevant in their applications, a clear disadvantage of these methods is their large computational complexity. Third, generative methods search for the unknown—overlapping—clusters in the data by fitting a mixture model. Mixture models for overlapping clusters (Banerjee et al., 2005; Heller & Ghahramani, 2007; Fu & Banerjee, 2008; Zhong & Ghosh, 2003), in which it is assumed that a data object is generated by simultaneously activating multiple mixture components, are flexible in the choice of probability distribution for each cluster. As a disadvantage, however, mixture models for overlapping clustering do not allow the user to easily exert control over the type and amount of overlap between clusters.

A fourth class of clustering methods are partitional methods that belong to the most commonly used methods for clustering. Within partitional methods, Ben N'Cir et al. (2015) makes a distinction between uncertain memberships and hard membership models. In uncertain memberships partitional models, a fuzzy, possibilistic or evidential framework is adopted to represent the clusters' memberships of each data object, which implies that each data object can belong to a larger or smaller extent to a (multiple) cluster(s). Hard membership partitional methods, on the contrary, rely on a binary function to model clusters' memberships (i.e., hard clustering), which implies that objects belong or do not belong to one (or more) cluster(s). Examples of uncertain memberships methods for overlapping clustering are fuzzy C-means (FCM; Lingras and West, 2004; Zhang et al., 2007; Höppner et al., 1999; Hruschka et al., 2009; Bezdek, 1981), possibilistic C-means (PCM; Krishnapuram & Keller, 1993; Krishnapuram & Keller, 1996) and possibilistic fuzzy C-means (PFCM; Pal et al., 2005), which modify the cluster results from a standard non-overlapping method into overlapping clusters. Other examples are Evidential c-means (ECM; Masson & Denoeux, 2008) and Belief c-means (BCM; Liu et al., 2012), which tailor the objective criterion in order to model cluster overlap. For hard membership partitional methods, a further distinction is made by Ben N'Cir et al. (2015) between additive and geometrical based methods. Additive methods model cluster overlap as the sum of representative objects of the clusters

in question. As a consequence, these methods aim at minimizing the sum of squared residuals between the scores of each observation and the sum of clusters' representatives scores to which the observation in question belongs. Examples of additive overlapping clustering methods are Principal Cluster Analysis (PCL; Mirkin, 1987, 1990) and a related version of PCL that uses the Alternating Least Square algorithm (ALS; Depril et al., 2008), Lowdimensional Additive Overlapping Clustering (Depril et al., 2012) and Bi-clustering ALS (Wilderjans et al., 2013). Geometrical methods, on the contrary, conceive cluster overlap in terms of the barycenter on the related cluster representatives. As such, these methods minimize the sum of distances between each observation and the average—instead of the sum—of clusters' representatives to which the observation belongs to. Examples of geometrical methods for overlapping clustering are Overlapping *K*-means (OKM; Cleuziou, 2008), Parameterized R-OKM (Ben N'Cir et al., 2013), Overlapping k-Medoid (OKMED; Cleuziou, 2009), Weighted Overlapping *K*-means (WOKM; Huang et al., 2005; Cleuziou, 2009), Overlapping Partitioning Cluster (OPC; Chen & Hu, 2006), Multi-Cluster Overlapping *K*-means Extension (MCOKE; Baadel et al., 2015) and Kernel Overlapping *K*-means (KOKM; Ben N'Cir et al., 2010; Ben N'Cir & Essoussi, 2012).

In the remainder of this paper, we will only focus on partitional methods for overlapping clustering that imply a hard membership model (i.e., binary memberships) and that conceive cluster overlap as additive. Moreover, we will limit our attention to methods that can be applied to object by variable data. To obtain an overlapping (additive) object (hard) clustering based on object by variable data, Mirkin (1987, 1990) developed the ADditive PROfile CLUStering (ADPROCLUS) method, which operates on the object by variable data directly. Note that the original ADCLUS model of Shepard and Arabie (1979) can only be used for analyzing an object by object (dis)similarity matrix. To improve the Principal Cluster Analysis algorithm of Mirkin (1987, 1990), Depril et al. (2008) proposed an Alternating Least Squares (ALS) algorithm to estimate an ADPROCLUS model from a given data set, and Wilderjans et al. (2011) provided easy-to-use software to perform ADPROCLUS.

An important question when performing any cluster analysis concerns the identification of the optimal number of clusters underlying the data, which boils down to a problem of model selection. To tackle this problem, one strategy consists of, first, computing a number of cluster solutions with an increasing number of clusters. Next, the optimal number of clusters is determined by means of a model selection procedure that searches for an optimal compromise between model fit and clustering complexity, with the latter being related to the number of clusters. For partitioning and hierarchical clustering methods, a broad range of such model selection procedures has been proposed (e.g., Akaike, 1974; Caliński and Harabasz, 1974; Hannan & Quinn, 1979; Schwarz, 1978; Sugar & James, 2003; Tibshirani et al., 2001b; Wang, 2010; Steinley & Brusco, 2011) and these approaches have been compared to each other in extensive simulation studies (e.g., Steinley & Brusco, 2011; Milligan & Cooper, 1985).

For overlapping clustering, however, almost no strategies for model selection have been proposed. Moreover, up to now, no systematic study has been conducted in which various model selection strategies for determining the optimal number of overlapping clusters underlying a given object by variable data set are compared. To address this gap in the literature, in this paper, various commonly used model selection strategies for other purposes than ADPROCLUS are adapted to or specifically tailored towards ADPROCLUS, and these strategies are evaluated in an extensive simulation study. The model selection strategies included in this study are as follows: Akaike's Information Criterion (AIC, Akaike, 1974), corrected AIC (AICc, Hurvich & Tsai, 1989), Bayesian Information Criterion (BIC,

Schwarz, 1978), Hannan-Quinn Measure (HQM, Hannan & Quinn, 1979), CHull (Ceulemans et al., 2011; Wilderjans et al., 2012), Steinley and Brusco's Lower Bound Technique (LBT; Steinley & Brusco, 2011), the Caliński-Harabasz Index (CH; Caliński & Harabasz, 1974), the Dunn Index (DI; Dunn, 1973, 1974), the Krzanowski-Lai Index (KL; Krzanowski & Lai, 1988) and the Average Silhouette Width (ASW; Rousseeuw, 1987). Further, we also tailored two cross-validation based procedures for model selection in clustering to an ADPROCLUS context.

The remainder of this paper is organized in four main sections: In Section 2, we will present Mirkin's ADPROCLUS model (Mirkin 1987, 1990), the associated loss function and the Alternating Least Squares (ALS) algorithm of Depril et al. (2008) to fit ADPROCLUS to given object by variable data. Next, the model selection strategies that will be compared in the simulation study are introduced in Section 3. In this section, we will also discuss the derivation of a likelihood function for a minimal stochastic extension of the ADPROCLUS model, which is needed for the calculation of information criteria for ADPROCLUS solutions. In Section 4, the design, procedure and results of an extensive simulation study to evaluate the performance of the model selection strategies to determine the correct number of overlapping clusters in ADPROCLUS will be presented and discussed. Finally, some concluding remarks are presented in Section 5.

2 Additive Profile Clustering (ADPROCLUS)

2.1 Model

The additive profile clustering (ADPROCLUS) model, as introduced by Mirkin (1987, 1990), approximates an $I \times J$ object by variable data matrix $\mathbf{X}$ by an $I \times J$ model matrix $\mathbf{M}$ that can be decomposed into an $I \times K$ binary cluster membership matrix $\mathbf{A}$ and a $K \times J$ real-valued cluster profile matrix $\mathbf{P}$, with K indicating the number of overlapping clusters. In particular, $\mathbf{M}$ is decomposed as:

$$\mathbf{M} = \mathbf{AP} \tag{1}$$

The entries a_{ik} in matrix $\mathbf{A}$ represent whether object i belongs to cluster k ($a_{ik} = 1$) or not ($a_{ik} = 0$), with each object being allowed to belong to one cluster, to multiple clusters or to no cluster at all. The entries p_{kj} of matrix $\mathbf{P}$ denote the prototypical value of variable j in cluster k. The k-th row of $\mathbf{P}$ therefore contains the variable profile for the k-th object cluster ($k = 1, \ldots, K$). The predicted/reconstructed data entries m_{ij} based on the ADPROCLUS model can be calculated as:

$$m_{ij} = \sum_{k=1}^{K} a_{ik} p_{kj}. \tag{2}$$

As a consequence, the estimated variable profile for a particular object i can be obtained by summing the cluster-specific variable profiles (i.e., rows of $\mathbf{P}$)—hence the name *additive profile clustering* —associated with the clusters to which object i belongs (for an illustration of the ADPROCLUS model, see Wilderjans et al., 2011).

2.2 Loss Function

In practice, as data always contain (considerable amounts of) noise, no model matrix $\mathbf{M}$ that can be decomposed into $\mathbf{A}$ and $\mathbf{P}$ with K clusters (and $K < \min(I,J)$) can be found

that reconstructs **X** perfectly. As a consequence, discrepancies between **X** and **M** are allowed:

$$\mathbf{X} = \mathbf{M} + \mathbf{E} = \mathbf{AP} + \mathbf{E}, \tag{3}$$

with **E** being an $I \times J$ matrix of residuals denoting the discrepancies between **X** and **M**.

The aim of an ADPROCLUS analysis is therefore, given a fixed number of clusters K, to estimate a binary membership matrix **A** and a real-valued profile matrix **P** such that model matrix **M = AP** reconstructs data matrix **X** as close as possible in least squares sense (i.e, sum of squared residuals). In particular, the following least squares loss function needs to be minimized:

$$lf(\mathbf{A}, \mathbf{P}) = \| \mathbf{X} - \mathbf{AP} \|_F^2, \tag{4}$$

where $\| \ . \ \|_F$ indicates the Frobenius matrix norm (i.e., the Euclidean norm applied to the entries of the matrix). The loss function in (4) computes the sum of squared residuals between **X** and **M**, which becomes clear when the loss function is written element-wise:

$$lf(\mathbf{A}, \mathbf{P}) = \sum_{i=1}^{I} \sum_{j=1}^{J} \left(x_{ij} - \sum_{k=1}^{K} a_{ik} p_{kj} \right)^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} e_{ij}^2, \tag{5}$$

with e_{ij} being the elements of the residual matrix **E**.

2.3 Algorithm

2.3.1 Alternating Least Squares Algorithm

Depril et al. (2008) proposed three algorithms for fitting the ADPROCLUS model to data and showed that these algorithms outperform the initial Principal Cluster Analysis (PCL) algorithm, which is also denoted as SEFIT, of Mirkin (1987, 1990). In this paper, we will use the ALS_{lf_2}-algorithm of Depril et al. (2008), which is an alternating least squares (ALS) algorithm for the minimization of loss function (4). This algorithm was adopted as, overall, this algorithm performs better or at least as good than the other algorithms presented in Depril et al. (2008). Moreover, this algorithm is straightforward to program and is not as time-consuming as the other algorithms from Depril et al. (2008). In this algorithm, starting from an initial random or (pseudo-)rational estimate of **A** (see Section 2.3.2), the following two steps are alternated until convergence: re-estimating **P** conditional upon **A** and updating **A** conditional upon **P**. The algorithm is considered converged when updating **A** and **P** does only yield a negligible decrease in the value of the loss function (e.g., a relative decrease smaller than some pre-determined tolerance value, like .000001).

As re-estimating **P** with **A** kept fixed boils down to a multivariate multiple regression problem, the update of **P** is given by the following closed-form expression:

$$\mathbf{P} = (\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'\mathbf{X}, \tag{6}$$

where $(\mathbf{A}'\mathbf{A})^{-1}$ denotes the matrix inverse of $\mathbf{A}'\mathbf{A}$. Note that the Moore-Penrose pseudo-inverse (Moore, 1920) needs to be used for inverting $\mathbf{A}'\mathbf{A}$ when **A** is not of full column rank (for more details, see Depril et al., 2008). **A** can be re-estimated conditionally upon **P** row by row as the loss function is separable as follows Chaturvedi and Carroll (1994):

$$lf(\mathbf{A}, \mathbf{P}) = \sum_{j} \left(x_{1j} - \sum_{k=1}^{K} a_{1k} p_{kj} \right)^2 + \cdots + \sum_{j} \left(x_{Ij} - \sum_{k=1}^{K} a_{Ik} p_{kj} \right)^2, \tag{7}$$

where, when **P** is kept fixed, each term only depends on a different row of **A**. As a consequence, loss function (4) can be decreased globally by decreasing each term in (7)

separately. To update a row of **A** for a given **P**, all possible binary membership patterns (i.e., 2^K in total) are evaluated and the binary pattern yielding the lowest loss value is retained. This procedure is applied such that each row of **A** is updated once.

2.3.2 Initialization Strategies for Membership Matrix A

To obtain an initial estimate of membership matrix **A**, which is used to start the ALS algorithm, three strategies can be used: a random, a rational or pseudo-rational strategy (for more information on these types of initial starting configurations, see Ceulemans et al., 2007). A random initial **A** is generated by drawing its entries independently from a Bernoulli distribution with parameter $\pi = .5$ (i.e., a 50% chance for a 0 or a 1). A rational estimate of the initial **A** can be determined in two ways. First, by using the overlapping clustering that is obtained by running the SEFIT-algorithm of Mirkin (1987, 1990); we will call this the SEFIT-rational start. Second, by taking a random sample of K objects (i.e., rows) from data matrix **X**, collecting these samples in a matrix **P** and computing the optimal **A** conditionally upon this **P**; we will call this the $\mathbf{A}^{(\mathbf{P})}$-rational start. Finally, a pseudo-rational initial **A** can be obtained by slightly perturbating one of both rationally determined initial **A**'s. In particular, a small percentage, like 20%, of the entries of a rationally obtained initial **A** are switched from zero to one or vice versa.

2.3.3 Multi-start Procedure and Data Preprocessing

As noted by Depril et al. (2008), loss function (4) has multiple local optimal solutions, implying that the solution obtained with the previously discussed ALS algorithm may strongly depend on the initial membership matrix **A** chosen. To minimize the risk of the algorithm retaining a suboptimal solution, a multi-start procedure is recommended (a similar multi-start procedure is also recommended for K-means, see Steinley, 2003). In such a procedure, the algorithm is run multiple times, each time starting with a different initial **A**. The best solution, in terms of minimizing loss function (4), across all runs of the algorithm is retained as the final solution. Regarding the choice of initial **A**'s, Depril et al. (2008) suggest to use a mix of multiple random and (pseudo-)rational starts.

The clustering solutions obtained by the ADPROCLUS algorithm are strongly influenced by preprocessing steps, like multiplicative (e.g., normalization) and additive (e.g., centering) transformations of the variables. Based on the arguments discussed in Wilderjans et al. (2011), it is advisable to preprocess the raw data before performing ADPROCLUS. Two possible forms of preprocessing are as follows: First, divide the variables by their range or standard deviation to account for large between-variable differences in variance (Milligan & Cooper, 1988). Second, when required, convert raw scores into deviation scores from a mean, a reference level or a normative score (for more information, see Wilderjans et al., 2011).

3 Model Selection Strategies

In this section, we describe the various model selection strategies that will be compared in their capacity to determine the correct number of overlapping clusters for ADPROCLUS. First, six existing strategies are described that are (often) used in the context of partitioning and that quite easily can be adapted for use in ADPROCLUS. Next, four commonly encountered model selection strategies for partitioning are presented that cannot in a

straightforward way be adapted for use in ADPROCLUS as they rely on the extent to which between-cluster distances are large relative to the within-cluster distances: the Caliński-Harabasz (CH; Caliński & Harabasz, 1974), the Dunn Index (DI; Dunn, 1973; Dunn, 1974), the Krzanowski-Lai Index (KL; Krzanowski & Lai, 1988) and the Average Silhouette Width (ASW; Rousseeuw, 1987). As such, these four strategies only make sense when clusters are sought that are internally cohesive and externally isolated (Cormack, 1971), which is not the case for overlapping clusters. Therefore, we tailored these methods in such a way that they can also be applied in the context of overlapping clustering. Finally, two (cross-) validation based strategies are introduced that were specifically tailored to an ADPROCLUS context. The total number of estimated ADPROCLUS parameters (i.e., IK binary memberships a_{ik} and JK real-valued profile values p_{jk}) is taken as complexity measure fp for an ADPROCLUS solution with K clusters:

$$fp = (I + J) \times K + 1. \tag{8}$$

Note that an extra parameter is added, which represents the residual variance σ_e^2 that has to be estimated when fitting the minimal stochastic extension of ADPROCLUS (see Section 3.1.1).

3.1 Existing Model Selection Strategies Adapted to ADPROCLUS

Some of the strategies that will be presented, like strategies based on information criteria, depend on the fit of the ADPROCLUS model in terms of a (log-)likelihood value. As the ADPROCLUS model is a deterministic model in which no distributional assumptions are made regarding (the model parameters and) the residuals e_{ij}, no such likelihood value can be calculated. Therefore, a minimal stochastic extension of the ADPROCLUS model is proposed that allows the construction of a likelihood function for the ADPROCLUS model. It will turn out that the cluster solution optimizing this likelihood function equals the solution optimizing the original least squares loss function (4), which justifies the use of this log-likelihood value as a measure of model (mis)fit for ADPROCLUS.

3.1.1 Likelihood Function

Combining (2) and (3) results in

$$x_{ij} = \sum_{k=1}^{K} a_{ik} p_{kj} + e_{ij} = m_{ij} + e_{ij}. \tag{9}$$

A minimal stochastic extension of the ADPROCLUS model can be constructed by assuming that the residuals e_{ij} are independently drawn from $N(0, \sigma_e^2)$; this implies that the data entries x_{ij} are assumed to be distributed as

$$x_{ij} \overset{i.i.d.}{\sim} N(m_{ij}, \sigma_e^2), \tag{10}$$

with $\sigma_e^2 = \mathrm{VAR}(e_{ij})$ being the residual error variance. Assuming all x_{ij}'s being independent from each other (and identically distributed), results in the negative log-likelihood $-l(\mathbf{X}|\mathbf{M})$ of the whole data set $\mathbf{X}$ being equal to:

$$-l(\mathbf{X}|\mathbf{M}) = \frac{n}{2}\log(2\pi) + n\log(\sigma_e) + \frac{1}{2\sigma_e^2}\sum_{i=1}^{I}\sum_{j=1}^{J}(x_{ij} - m_{ij})^2, \tag{11}$$

where $n = IJ$ refers to the number of terms in the likelihood function, which equals the number of data points x_{ij}. Acknowledging that the error variance σ_e^2 can be estimated as

$$\sigma_e^2 = \frac{\sum_{i=1}^{I}\sum_{j=1}^{J}(x_{ij} - m_{ij})^2}{n}, \tag{12}$$

the log-likelihood can be simplified as

$$-l(\mathbf{X}|\mathbf{M}) = \frac{IJ}{2}\log(2\pi) + \frac{IJ}{2}\Big(1 - \log(IJ)\Big) + \frac{IJ}{2}\log\left(\sum_{i=1}^{I}\sum_{j=1}^{J}(x_{ij} - m_{ij})^2\right). \tag{13}$$

It appears that the negative log-likelihood (13) can be optimized by minimizing its last term only (i.e., all other terms are constants), which fully depends on $\sum_{i=1}^{I}\sum_{j=1}^{J}(x_{ij} - m_{ij})^2$. As a consequence, minimizing the ADPROCLUS least squares loss function (4) is equivalent to minimizing the negative log-likelihood function (11) associated with the proposed minimal stochastic extension of ADPROCLUS. In other words, the optimal ADPROCLUS solution minimizing the original least squares loss function (4) equals the solution optimizing the (negative) log-likelihood function (11).

3.1.2 CHull

The CHull method (Ceulemans et al., 2011; Wilderjans et al., 2012) is an automated procedure for determining the elbow in a scree-like plot (Cattell, 1966). Such a plot is often used for model selection purposes, like determining the number of components in Principal Components Analysis (PCA) or the number of clusters in a cluster analysis. In a scree plot, the (mis)fit f of various obtained solutions, which, depending on the type of analysis, may be eigenvalues or the amount of explained variance, is plotted against a measure fp of the complexity of these solutions, like the number of components or clusters. In the case of ADPROCLUS, the loss value (4) of the optimal solution may serve as a model misfit measure and the number of estimated parameters (8) as a model complexity value. To determine the optimal number of overlapping clusters, one searches for an elbow point in the scree plot in which the loss value (4) is displayed against the fp value (8) for solutions with increasing numbers of overlapping clusters $k = 1, \ldots, K$.

CHull can be used to identify the elbow point in the scree plot in an automated way instead of visually, which may be quite subjective. In the CHull procedure, first, the solutions are determined that are located on the boundary of the convex hull of the plotted points. Depending on whether the fit measure pertains to model fit, like percentage explained variance, or model misfit, like sum of squared residuals, the upper or lower boundary of the convex hull, respectively, is sought for. Next, the solution on the convex hull is selected that has an optimal balance between model fit and complexity (for an overview of the six-step CHull procedure, see Wilderjans et al., 2012). Note that CHull can never select the least and most complex model.

In our simulation study, we will use CHull with $fp = ((I + J) \times K) + 1$ (see (8)) as complexity measure. For the fit measure, we will evaluate two CHull strategies: one using the value of the least squares loss function (4) and one based on the value of the negative log-likelihood function (11), which are in the following referred to as CHull LSQ and CHull NLL, respectively.

3.1.3 Akaike Information Criterion (AIC)

Information criteria are commonly used to tackle model selection problems in a variety of statistical models, such as mixture analysis (Solka et al., 1998; Steele & Raftery, 2009), exploratory factor analysis (Preacher et al., 2013), generalized linear (mixed) models (Mclachlan & Peel, 2000; Müller & Stadtmüller, 2005) and *K*-means (Steinley & Brusco, 2011). One of the most commonly used information criterion is the Akaike Information Criterion (AIC, Akaike, 1974), in which twice the negative log-likelihood *-l* of the model is penalized by two times the number of estimated parameters fp in the model:

$$\text{AIC} = 2(\text{-}l) + 2 \times fp. \tag{14}$$

The optimal model is the model with the smallest AIC value.

3.1.4 Corrected AIC (AICc)

As observed by Hurvich and Tsai (1989), AIC does not perform well in applications where the sample size is small (for a thorough discussion regarding the bias in AIC, see Bozdogan, 2000) as in these situations AIC has the tendency to select a too complex model (for a similar observation in the context of mixture of factor analyzers, see Bulteel et al., 2013). To correct for this bias, Hurvich and Tsai (1989) developed the corrected Akaike Information Criterion (AICc) and demonstrated its advantage in small-sample applications. Essentially, the AICc extends the AIC by penalizing additional parameters more heavily:

$$\text{AICc} = \text{AIC} + \frac{2fp(fp+1)}{n - fp - 1}, \tag{15}$$

where n indicates the sample size. Note that for ADPROCLUS, we take $n = IJ$ as the data matrix $\mathbf{X}$ contains IJ values (and, as a consequence, IJ terms separately contribute to the likelihood function). Note that when the sample size becomes much larger than fp, the correction term approaches 0 and therefore becomes negligible. As a consequence, the AICc should generally be preferred over AIC (Burnham & Anderson, 2004) as it offers serious advantages in small-sample applications and no substantial disadvantages in large-sample applications.

3.1.5 Bayesian Information Criterion (BIC)

A central assumption to AIC is that each observation in the data set provides new and independent information about the underlying model (Bozdogan, 1987). However, Schwarz (1978) argued that this assumption may be unrealistic for large data sets and therefore proposed the Bayesian Information Criterion (BIC), which takes the sample size into account, as an alternative to AIC:

$$\text{BIC} = 2(\text{-}l) + \log(n) \times fp. \tag{16}$$

A significant advantage of the BIC over AIC is that its probability to select the true model is increasing as the sample size gets larger, as long, of course, the true model is among the candidate models. In situations where the number of estimated parameters does not depend on the sample size, the probability of the BIC to select the true model approaches 1 (for more information, see Vrieze, 2012).

3.1.6 Hannan-Quinn Measure (HQM)

Similar to AICc and BIC, the Hannan-Quinn Measure (HQM, Hannan and Quinn, 1979) is proposed to alleviate AIC's tendency of selecting a too complex model. Although HQM was originally introduced in the context of autoregressive modelling, later it was adapted to other likelihood-based models (see, Claeskens & Hjort, 2008). The HQM can be calculated as

$$\text{HQM} = 2(\text{-}l) + 2 \times fp \times \log(\log(n)). \tag{17}$$

In the simulation study, all four information criteria will be included with (*-l*) being the negative log-likelihood value (11), *fp* equaling $((I + J) \times K) + 1$ and $n = IJ$.

3.1.7 Lower Bound Technique (LBT)

Steinley and Brusco (2011) proposed the Lower Bound of SSE Technique (LBT) for determining the optimal number of clusters in *K*-means clustering. Note that the ADPROCLUS and the *K*-means clustering model are mathematically very similar. In particular, there are only two differences between both clustering models. A first difference is that in *K*-means the binary membership matrix is restricted to take the form of a partition matrix, which implies that each row of the membership matrix is constrained to sum to one, whereas no such constraint is imposed on the binary membership matrix of the ADPROCLUS model. A second difference between *K*-means and the ADPROCLUS model is that in *K*-means the profile matrix contains the cluster-specific variable means, whereas this is not the case for ADPROCLUS due to the cluster overlap. Because of this mathematical similarity between the *K*-means and the ADPROCLUS model, the LBT can be applied without modification to an overlapping clustering context. Generally, the LBT is a normalized index that is defined as

$$\text{LBT}_K = \frac{\text{SSE}_K - \text{SSE}_{\text{min}}^K}{\text{SST}}, \tag{18}$$

where SST refers to the sum of the squared entries of $\mathbf{X}$ after centering the variables by subtracting the variable mean from each data point, and SSE_K equals the loss value for a solution with *K* clusters (for ADPROCLUS: loss function value (4)). $\text{SSE}_{\text{min}}^K$ indicates the lower bound for a solution with *K* clusters and is calculated as:

$$\text{SSE}_{\text{min}}^K = \text{trace}(\mathbf{X}'\mathbf{X}) - \sum_{i=1}^{K} \lambda_i^{(\mathbf{XX}')}, \tag{19}$$

with $\lambda^{(\mathbf{XX}')}$ denoting the eigenvalues of $\mathbf{XX}'$, ordered from largest to smallest ($\lambda_1^{(\mathbf{XX}')} \geq \lambda_2^{(\mathbf{XX}')} \geq \cdots \geq \lambda_I^{(\mathbf{XX}')}$). As such, LBT compares the observed SSE_K-value for a *K*-cluster solution to a minimal value $\text{SSE}_{\text{min}}^K$ that is associated with that particular number of clusters *K*. The optimal solution is the solution for which $|\text{LBT}_K|$ is minimal. In an extensive simulation study on *K*-means, the LBT outperformed traditional model selection strategies, such as the CH Index, the BIC, and Wilk's Λ (Steinley & Brusco, 2011).

3.1.8 The Caliński-Harabasz (CH) Index

The Caliński-Harabasz Index (CH; Caliński & Harabasz, 1974), which is also known as the variance ratio criterion, was originally proposed to determine the number of

clusters in the context of a non-overlapping clustering problem. The CH Index is computed as:

$$\mathrm{CH}_k = \frac{\mathrm{SS}_\mathrm{B}^{(k)}}{\mathrm{SS}_\mathrm{W}^{(k)}} \frac{N-k}{k-1}, \tag{20}$$

where k is the number of (non-overlapping) clusters, N the total number of data points, $\mathrm{SS}_\mathrm{W}^{(k)}$ the overall within-cluster variance (i.e., total within-cluster sum of squared differences) and $\mathrm{SS}_\mathrm{B}^{(k)}$ the overall between-cluster variance (i.e., total between-cluster variance in the cluster centroids) for the best solution with k clusters. As a small SS_W—which always will become smaller when k increases—indicates that clusters are homogeneous and that cluster centroids are nicely spread out, which implies a large SS_B, the largest value for the CH Index points at the optimal number of clusters k that should be retained.

In the context of overlapping clusters, however, the CH Index is not a good measure to determine the optimal number of—overlapping—clusters as the CH Index searches for a clustering with homogeneous clusters that are clearly spread out, which does not make sense for overlapping clusters. As a solution, in order to adapt the CH Index to the context of overlapping clustering, we make use of the fact that each clustering with k overlapping clusters—at least in the case of a hard clustering (i.e., binary memberships)—can always be transformed into a clustering with maximally 2^k non-overlapping clusters (e.g., for a clustering with $k = 2$ overlapping clusters, one out of $2^k = 2^2 = 4$ possible membership patterns—00, 10, 01 or 11—is allocated to each object, which implies a clustering of the objects into—maximally—four non-overlapping clusters). Now, the CH Index can be computed by considering the non-overlapping clustering that is derived from the overlapping clustering, herewith, however, using a value of 2^k for the number of clusters k (e.g., for 2, 3, and 4 clusters, we used a value of k of 4, 8 and 16, respectively) in formula (20) and calculating $\mathrm{SS}_\mathrm{W}^{(k)}$ and $\mathrm{SS}_\mathrm{B}^{(k)}$ based on the non-overlapping clusters and associated cluster centroids.

3.1.9 The Dunn Index (DI)

The Dunn Index (DI; Dunn, 1973, 1974) is a cluster validity index that aims at identifying the value for k that yields a solution with compact and well separated clusters, where the variance between cluster members (i.e., the within-cluster variance) is small and the cluster means are sufficiently far apart from each other. The Dunn Index can be computed as:

$$\mathrm{DI}_k = \frac{\min_{i \neq j} \langle \delta(\mathrm{C}_i, \mathrm{C}_j) \rangle}{\max_{1 \leqslant l \leqslant k} \langle \Delta(\mathrm{C}_l) \rangle}, \tag{21}$$

with C_i, C_j and C_l denoting cluster i, j and l, $\delta(C_i, C_j) = \min_{x \in C_i, y \in C_j} \langle d(x, y) \rangle$, $\Delta(C_l) = \max_{x,y \in C_l} \langle d(x, y) \rangle$, $d(x, y)$ the Euclidean distance between objects x and y and k the considered number of clusters. The optimal number of clusters k is identified by choosing the solution with k clusters for which the Dunn Index (21) is maximal. As the Dunn Index—for the same reason as for the CH Index—can only be used in the context of non-overlapping clusters, we adapted the Dunn Index in a similar way as the CH Index by computing DI for the non-overlapping clustering with 2^k clusters that can be derived from the overlapping clustering with k clusters (see Section 3.1.8).

3.1.10 The Krzanowski-Lai Index (KL)

Similar to the CH Index, the Krzanowski-Lai Index (KL; Krzanowski & Lai, 1988) uses the within-cluster sum of squares as a criterion for determining the optimal number of clusters k. The KL Index for a solution with k clusters is defined as

$$\mathrm{KL}_k = \left| \frac{\mathrm{DIFF}_k}{\mathrm{DIFF}_{k+1}} \right|, \tag{22}$$

where DIFF_k denotes a scaled difference between the within-cluster sum of squares (SS_W) of the two sequential clustering solutions with $k-1$ and k clusters:

$$\mathrm{DIFF}_k = (k-1)^{2/J}\ \mathrm{SS}_{\mathrm{W}}^{(k-1)} - k^{2/J}\ \mathrm{SS}_{\mathrm{W}}^{(k)}, \tag{23}$$

where J is the number of variables. The value of k maximizing the KL Index (22) is regarded as the optimal number of clusters. Note that the KL Index in formula (22) shows some similarity with the st-ratio used in CHull. In particular, the st-ratio equals $\frac{\mathrm{DIFF}_k}{\mathrm{DIFF}_{k+1}}$, with $\mathrm{DIFF}_k = \frac{1}{c_k - c_{k-1}}(f_k - f_{k-1})$, f_k being a (mis)fit value for an ADPROCLUS solution with k clusters (e.g., SS_{W} or the value of the least squares loss function or the negative log-likelihood function) and c_k denoting the complexity of such a solution (e.g., k or fp). CHull is a flexible method as the user can decide which (mis)fit and complexity measure to use. It should be noted that due to the need for sequential clustering solutions, the KL index is not defined for the smallest and largest numbers of clusters k considered. To adapt the KL Index to additive overlapping clusterings with ADPROCLUS, we compute the KL Index for the non-overlapping clustering with 2^k clusters that can be derived from the overlapping clustering with k clusters (see Section 3.1.8), with 2^k used as value for k in (23).

3.1.11 The Average Silhouette Width (ASW)

The Average Silhouette Width, or Silhouette Index (ASW; Rousseeuw, 1987) was introduced as a graphical method for model selection in partitional clustering, specifically for K-means and K-medians clustering. In this vein, the ASW was designed to be useful in the case of compact, clearly separated and roughly spherical clusters. The ASW for a solution with k clusters is defined as

$$\mathrm{ASW}_k = \frac{\sum_i^N s_i}{N}, \tag{24}$$

where the Silhouette value of each object s_i ($i = 1, \ldots, N$) is defined as

$$s_i = \frac{b_i - a_i}{\max\{a_i, b_i\}} \quad \text{if } |C_i| > 1, \tag{25}$$

and

$$s_i = 0 \quad \text{if } |C_i| = 1, \tag{26}$$

where

$$a_i = \frac{1}{|C_i| - 1} \sum_{j \in C_i, i \neq j} d(i, j) \tag{27}$$

is a measure of the average dissimilarity of object i to all other objects in the same cluster C_i. $|C_i|$ is the number of data points assigned to cluster C_i and $d(i, j)$ is the (Euclidean) distance between data points i and j in cluster C_i. The smaller the value of a_i, the better

we can regard the assignment of object i to the cluster C_i (i.e., object i is quite close to the other objects of C_i). Similarly,

$$b_i = \min_{k \neq i} \frac{1}{|C_k|} \sum_{j \in C_k} d(i, j) \tag{28}$$

calculates the smallest average distance between object i and all objects in any other cluster (of which object i is not a member). We can regard b_i as a measure of how far object i is from the (objects in the) closest neighboring cluster. The optimal number of clusters k is chosen as the k for which the ASW in (24) is maximized.

Note that the computation of a_i (i.e., distance to objects in the same cluster) and b_i (i.e., distance to objects in other clusters) only makes sense when an object partition with more or less compact clusters is obtained. We therefore, in order to be able to use ASW for model selection for ADPROCLUS, consider the partition with 2^k non-overlapping clusters that can be derived from the solution with k overlapping clusters (see Section 3.1.8) when computing the ASW statistic.

3.2 Tailored Model Selection Strategies

Cross-validation has been widely used in the literature to determine an optimal prediction model (Stone, 1974; Browne, 2000). Tibshirani et al. (2001a) adapted the idea of cross-validation to a cluster analysis context and, based on this, Wang (2010) proposed a modified cross-validation procedure for determining the correct number of clusters in K-means clustering that is based on a clustering instability criterion. Wang (2010) defined clustering instability as

$$s(\Psi, K, n) = E[d\{\Psi(Z^n; K), \Psi(Z^{*n}; K)\}], \tag{29}$$

with E being the expected value and Z^n and Z^{*n} indicating a pair of independent samples of size n; $\Psi(Z^n; K)$ and $\Psi(Z^{*n}; K)$ denote the cluster solution with K clusters that is obtained by applying the clustering algorithm under study to Z^n and Z^{*n}, respectively; d represents a distance metric between cluster solutions that is interpreted as a measure for clustering instability (i.e., when clusterings become more different, their distance becomes larger). Inspired by Wang (2010), we created a clustering instability measure for ADPROCLUS and used this measure in two different (cross-)validation scenario's, with the first being computationally less expensive than the second one.

3.2.1 Simplified (Cross-)validation

We propose the following four-step procedure to measure cluster instability. First, an $I \times J$ data matrix $\mathbf{X}$ is split into three parts, $\mathbf{X}_1$, $\mathbf{X}_2$ and $\mathbf{X}_3$ with m, m and $I - 2m$ observations, respectively. A sensible choice for m would be an integer value close to $\frac{I}{3}$. Second, ADPROCLUS is applied to $\mathbf{X}_1$ and $\mathbf{X}_2$, resulting in cluster membership matrices $\mathbf{A}_1$ and $\mathbf{A}_2$ and profile matrices $\mathbf{P}_1$ and $\mathbf{P}_2$. Third, cluster membership matrices $\mathbf{A}_3^{(\mathbf{P}_1)}$ and $\mathbf{A}_3^{(\mathbf{P}_2)}$ for the observations in $\mathbf{X}_3$ are computed that are conditionally optimal upon the profile matrices $\mathbf{P}_1$ and $\mathbf{P}_2$, respectively (see Section 2.3 for information on how to compute the conditional optimal $\mathbf{A}$ given $\mathbf{P}$). Finally, the clustering instability measure s is calculated as follows:

$$s = |\, \| \mathbf{X}_3 - \mathbf{A}_3^{(\mathbf{P}_1)}\mathbf{P}_1\|_F^2 - \| \mathbf{X}_3 - \mathbf{A}_3^{(\mathbf{P}_2)}\mathbf{P}_2\|_F^2 |, \tag{30}$$

with $\mathbf{A}_3^{(\mathbf{P}_1)}\mathbf{P}_1$ and $\mathbf{A}_3^{(\mathbf{P}_2)}\mathbf{P}_2$ indicating the predicted scores for $\mathbf{X}_3$ based on $\mathbf{P}_1$ and $\mathbf{P}_2$, respectively, and $|z|$ denoting the absolute value of the number z. This instability measure s equals

the difference in (mis)fit for $\mathbf{X}_3$ when using the profile matrices $\mathbf{P}_1$ and $\mathbf{P}_2$. Note that a stable clustering is indicated by a small s-value.

3.2.2 Complex Cross-validation

In the "simplified" (cross-)validation procedure, only one part of the data is used as training set (i.e., $\mathbf{X}_1$ and $\mathbf{X}_2$) and the other part as validation set (i.e., $\mathbf{X}_3$), without interchanging (i.e., crossing-over) the roles of both data parts. As such, the specific choice of which part of the data figures as training and which part as validation set influences the estimate of the cross-validation error. To lower the variance in this estimate, we also developed a cross-validation procedure for ADPROCLUS that resembles fivefold cross-validation. In this procedure, data set $\mathbf{X}$ is first randomly split into five parts, called folds, with $\frac{I}{5}$ observations each; these five sets are labeled as $\mathbf{X}_3^1, \ldots, \mathbf{X}_3^5$. Subsequently, for each $\mathbf{X}_3^n$ ($n = 1, \ldots, 5$), the remaining 80% of observations are equally divided into two training sets, $\mathbf{X}_1^n$ and $\mathbf{X}_2^n$, for which ADPROCLUS solutions with varying K are obtained. The clustering instability criterion (30) is calculated for each K and each of the five folds of the data and a mean score $s_{av}(K)$ across the five folds is obtained for each K. The optimal number of clusters K^* is then defined as the K for which $s_{av}(K)$ is minimal. Note that this strategy can be used with any user-defined number of data folds n. In practice, often $n = 5$ or $n = 10$ is used. As the data set in this strategy is split only once into n folds, the particular split that is used may strongly influence the selection of the optimal K as the $s_{av}(K)$-value for each K may be dependent on this particular split. To reduce the uncertainty in the $s_{av}(K)$-values, a procedure may be implemented in which the "complex" cross-validation procedure is applied multiple times to $\mathbf{X}$, each time with a different split of the objects in $\mathbf{X}$ into n folds. The final $s_{av}(K)$-value is then obtained by averaging over the $s_{av}(K)$-values obtained per split. In this paper, however, we will only study the cross-validation procedure with a single split.

4 Simulation Study

In this section, the results of a simulation study are presented to compare the effectiveness of thirteen strategies for determining the optimal number of overlapping clusters in ADPROCLUS. To determine their model selection performance, the model selection strategies are evaluated in terms of the *accuracy* and *precision* to which they identify the optimal number of clusters underlying the data (Steinley, 2003). Besides overall (i.e., across all generated data sets) differences between the proposed model selection strategies, we are also interested in how both performance aspects depend on the following data characteristics: the size of the data, the amount of noise in the data, the number of overlapping clusters underlying the data and the amount and type of cluster overlap.

4.1 Design and Procedure

Design To generate data sets with varying data characteristics, the following factors were systematically manipulated:

- The *size*, $I \times J$, of data matrix $\mathbf{X}$, at two levels: 200×15 and 400×15; note that the number of variables J is fixed to 15;
- The *number of overlapping clusters*, K, at two levels: 3 and 5;

- The *amount of cluster overlap*, which is defined as the percentage of objects belonging to more than a single cluster, at three levels: 0%, 35% and 75%.
- The *degree of non-occurring overlapping membership patterns*, which is defined as the number of distinct binary membership patterns indicating cluster overlap (i.e., multiple 1's in the membership pattern) that do not occur in **A**, at three levels: none, medium and high. Table 1 specifies the number of distinct non-occurring overlapping membership patterns for each condition;
- The *noise level*, ϵ, which is defined as the percentage of the total variance in **X** that is accounted for by **E**, at three levels: 0.1, 0.4 and 0.7. Note that this implies that 10%, 40% or 70%, respectively, of the data is noise.

These data characteristics—and their manipulated levels—were chosen in order to generate data sets that are representative for typical data from psychology and the social sciences. In particular, the considered number of cases, variables and clusters, along with the adopted noise levels (except maybe for 70% of noise, which is included to seriously challenge the model selection strategies under study) are commonly encountered in social sciences data. The other two factors (i.e., the amount of cluster overlap and the degree of non-occurring overlapping membership patterns), which are hard to determine for empirical data, were chosen in order to investigate the performance of the model selection strategies considered under widely varying cluster overlap patterns.

Procedure Each data set **X** is obtained through independent generation of a cluster membership matrix **A**, a profile matrix **P** and noise **E**, and subsequently calculating $\mathbf{X} = \mathbf{AP}+\mathbf{E}$. The values in the real-valued cluster profile matrix **P** are independently drawn from a normal distribution $N(0, 10)$. The entries of E are independently drawn from a normal distribution $N(0, \sigma_\epsilon^2)$, where the variance σ_ϵ^2 is chosen such that the data contain the desired proportion of noise ϵ. To generate the binary cluster membership matrix **A**, first, for each possible membership pattern, it is determined exactly how often it should occur in the rows of **A** according to the levels of the data characteristics *number of overlapping clusters*, *amount of cluster overlap* and *degree of non-occurring overlapping membership patterns*, with the all-zero pattern always occurring for one out of 20 objects. To control the desired *amount of cluster overlap*, it is calculated how many membership patterns should show overlap (i.e., 0%, 35% or 75%) and, as a consequence, how many patterns should not show overlap (i.e., 95%, 60% or 20%, respectively). For the patterns showing no overlap, it is taken care of that each such pattern occurs the same amount of times (e.g., $\frac{60\%}{3} = 20\%$ of the cases for the conditions with $k = 3$ clusters and an overlap of 35%) in the rows of **A**. Subsequently, to account for the desired *degree of non-occurring overlapping membership patterns*, from

Table 1 The number of distinct non-occurring overlapping membership patterns as a function of the amount of underlying clusters K and the degree of non-occurring overlapping membership patterns

		Degree of non-occurring membership patterns		
K	$P_{OL}{}^*$	None	Medium	High
3 Clusters	4	0	1	3
5 Clusters	26	0	9	17

* $P_{OL} = P_{\text{MAX}} - K - 1$ indicates, for a given number of clusters K, the number of distinct membership patterns that contain more than a single 1 (i.e., denoting cluster overlap), with $\text{P}_{\text{MAX}} = 2^K$ indicating the total number of possible (distinct) membership patterns for K clusters

all possible membership patterns indicating cluster overlap, a certain number, as specified in Table 1, of patterns is independently sampled and removed from the set of overlapping membership patterns occurring in the rows of $\mathbf{A}$. Note that for K clusters, the number of membership patterns showing overlap equals $2^K - (K + 1)$ as there are 2^K possible patterns in total of which there are $K + 1$ showing no overlap (i.e., the all-zero pattern and K patterns, one for each cluster, with a single one). It is guaranteed that each remaining overlapping membership pattern appears the same number of times in the rows of $\mathbf{A}$. Finally, a matrix $\mathbf{A}$ is filled according to how often each membership pattern should be encountered (as calculated above) and the rows of $\mathbf{A}$ are permuted randomly.

A full crossing of all levels of all manipulated factors is not possible as in the 0%-cluster-overlap condition the degree of non-occurring overlapping membership patterns cannot be varied. Therefore, data sets without any cluster overlap (0% condition; referred to as NOL data sets) and data sets with cluster overlap (35% and 75% conditions; referred to as OL data sets) will be analyzed and discussed separately. For NOL data sets, the design yields a total of 2 (*number of clusters*) × 2 (*data size*) × 1 (*amount of cluster overlap*) × 1 (*degree of non-occurring overlapping membership patterns*) × 3 (*noise level*) = 12 conditions. For OL data sets, the design yields 2 (*number of clusters*) × 2 (*data size*) × 2 (*amount of cluster overlap*) × 3 (*degree of non-occurring overlapping membership patterns*) × 3 (*noise level*) = 72 conditions. For each condition, 10 replicate data sets are generated, resulting in a total of 840 simulated data sets.

For each generated data set, ADPROCLUS models with $K = 1$ up to $K = 8$ clusters are obtained. For each analysis, a multi-start procedure is performed consisting of the following (see Section 2.3): one SEFIT-rational start, nine pseudo-rational starts based on the SEFIT-rational start, five $\mathbf{A}^{(\mathbf{P})}$-rational starts and 15 random starts. Further, to lower the risk of the algorithm retaining a local optimal solution, also an alternative rational start is included that makes use of the fact that the cluster analysis is performed sequentially with an increasing number of clusters K. In particular, an alternative rational start for $\mathbf{A}$ is obtained by taking the optimal solution $\mathbf{A}_{(K-1)}$ encountered for $K - 1$ clusters and adding one extra column filled with zeros and ones drawn independently from a Bernoulli distribution with probability 0.50; we call this the $\mathbf{A}_{(K-1)}$-rational start. Note that for $K = 1$, this type of rational start boils down to a random start. Subsequently, also nine pseudo-rational starts based on this alternative rational start were included. Finally, ten pseudo-rational starts were included that were obtained by perturbating (with each entry of $\mathbf{A}$ having a chance of 0.20 to be changed, see Section 2.3) the best solution encountered across the first 40 starts (see above); we call this the best-of-40-pseudo-rational start. Note that, in total, the multi-start procedure was run with 50 starts. Both cross-validation based procedures (see Section 3.2) were applied to each data set (also with 50 multi-starts). Regarding the split of the data into training and validation sets for performing the simplified (cross-)validation, we set the number of objects m in the training sets $\mathbf{X}_1$ and $\mathbf{X}_2$ equal to $m = \frac{2I}{5}$ (i.e., each training set contains 40% of the observations and the validation set has 20% of the observations). For the complex cross-validaton, we used fivefold cross-validation. The simulation study was programmed in R version 3.2.3 and performed on a supercomputer consisting of Intel Xeon E5-2660 CPUs with a clock frequency of 2.6 GHz.

4.2 Evaluation Criteria and Analysis Strategy

Evaluation Criteria Two measures were taken into account when assessing the performance of the model selection strategies: *accuracy* and *precision* (Steinley, 2003). The *accuracy* Ω of a strategy pertains to whether or not the strategy in question correctly identifies the true

number of clusters underlying the data. The *precision* Φ of a strategy equals the absolute difference between the by the strategy estimated number of clusters K^{est} and the true number of clusters K^{true}; for a particular data set d, the precision Φ_d equals $\Phi_d = |K_d^{est} - K_d^{true}|$. Note that this measure is an adapted version (i.e., absolute instead of squared differences) of the precision measure proposed in Steinley and Brusco (2011).

Analysis Strategy A mixed-design (repeated measures-) analysis of variance (ANOVA) is conducted to investigate how the precision Φ of a model selection strategy varies as a function of the manipulated data characteristics and the adopted model selection strategy. This mixed-ANOVA is performed for OL and NOL data sets separately, with the manipulated data characteristics acting as between-subject variables and the model selection strategy as a within-subject variable. As effect size measure for the mixed-design ANOVA, we use generalized eta squared (η_G^2), as proposed by Bakeman (2005). Only significant main and interaction effects with a generalized eta squared larger than or equal to 0.15 ($\eta_G^2 \geq 0.15$), indicating a medium effect (Cohen, 1988; Bakeman, 2005), are reported. Note that no ANOVA is performed on the accuracy measure as this is a binary outcome variable. To investigate how accuracy varies in function of the data characteristics and the model selection strategies, a table with the mean accuracy for each level of each manipulated factor will be displayed for each strategy.

4.3 Results

First, the overall performance of the model selection strategies in terms of accuracy and precision is discussed and the pattern of over- and underestimation of the model selection strategies is investigated. Next, the influence of data characteristics on accuracy and precision is examined. Subsequently, the best performing model selection strategies are compared to each other in more detail. Finally, the computation time of the ADPROCLUS algorithm is discussed.

4.3.1 Accuracy Ω, Precision Φ and Degree of Over- and Underestimation of the Model Selection Strategies

Accuracy In Table 2, the mean accuracy Ω of all model selection strategies is displayed, computed globally across all simulated data sets and as a function of the noise level ϵ. From this table it appears that, in data sets with overlap (OL), CHull using the (negative) log-likelihood as a measure of (mis-)fit (CHull NLL) performed best, selecting the correct model 460 out of 720 times (63.8%). CHull LSQ, using the loss function as (mis-)fit measure, performed slightly worse, with 434 out of 720 correctly identified models (60.3%). The procedure with the second best accuracy in OL data sets was Akaike Information Criterion (AIC; 48.2%), followed by the corrected AIC (AICc; 37.8%), both of which performed notably worse than CHull. The Dunn index (DI; 31.5%), the Lower Bound Technique (LBT; 30.8%), the Hannan-Quinn Measure (HQM; 30.6%) and complex cross-validation (CVc; 30.4%) all retrieved the correct number of clusters in about one third of all OL cases. Performance of the Average Silhouette Width (ASW; 19.9%), simplified cross-validation (CVs; 19.4%), the Krzanowski-Lai Index (KL; 16.3%), the Caliński-Harabasz Index (CH; 13.2%) and the Bayesian Information Criterion (BIC; 11.3%) was disappointing. A similar performance pattern was observed for data sets without overlap (NOL; see the values between brackets in Table 2): CHull NLL (75.8%) showed the highest accuracy, followed by CHull LSQ (70.0%). AIC (57.5%) and AICc (49.2%) had the second and third highest accuracy

levels, respectively, followed by HQM (39.2%), CVc (32.5%), LBT (24.2%), DI (20.8%) and KL (20.0%). Again, CH, BIC and CVs performed very poorly, also with low noise, with only 6.7%, 16.7% and 19.2% correctly retrieved models, respectively. Finally, the ASW completely fails for data sets without overlap (1.7%).

Table 2 Mean accuracy Ω (expressed as a percentage) and precision Φ of thirteen model selection strategies for selecting the optimal number of overlapping clusters in ADPROCLUS for OL (n = 720) and, between brackets, NOL data sets (n = 120), computed globally and as a function of the noise level ϵ

	Noise level ϵ			
	0.1	0.4	0.7	Overall
CHull - Negative Log-Likelihood (CHull NLL)				
$\Omega_{Correct}$	**99.6 (100.0)%**	**70.8 (85.0)%**	**21.2 (42.5)%**	**63.8 (75.8)%**
Φ	< 0.01 (0.00)	0.47 (0.18)	1.49 (0.98)	0.67 (0.38)
$\Phi_{underestimation}$	1.00 (-)	1.63 (1.17)	1.93 (1.70)	1.84 (1.59)
$\Phi_{overestimation}$	− (-)	1.00 (-)	2.00 (-)	1.87 (-)
CHull - Least Squares Loss Function (CHull LSQ)				
$\Omega_{Correct}$	**99.6 (100.0)%**	**63.3 (75.0)%**	**17.9 (35.0)%**	**60.3 (70.0)%**
Φ	< 0.01 (0.00)	0.61 (0.38)	1.55 (1.07)	0.73 (0.48)
$\Phi_{underestimation}$	1.00 (-)	1.69 (1.50)	1.93 (1.65)	1.85 (1.61)
$\Phi_{overestimation}$	− (-)	1.00 (-)	1.91 (-)	1.77 (-)
Akaike Information Criterion (AIC)				
$\Omega_{Correct}$	**100.0 (100.0)%**	**44.2 (70.0)%**	**0.4 (2.5)%**	**48.2 (57.5)%**
Φ	0.00 (0.00)	0.95 (0.32)	2.94 (2.52)	1.30 (0.95)
$\Phi_{underestimation}$	− (-)	1.70 (1.08)	2.95 (2.59)	2.50 (2.24)
$\Phi_{overestimation}$	− (-)	− (-)	− (-)	− (-)
Corrected AIC (AICc)				
$\Omega_{Correct}$	**94.2 (100.0)%**	**19.2 (47.5)%**	**0.0 (0.0)%**	**37.8 (49.2)%**
Φ	0.06 (0.00)	2.04 (1.27)	3.00 (2.92)	1.70 (1.40)
$\Phi_{underestimation}$	1.07 (-)	2.53 (2.43)	3.00 (2.93)	2.73 (2.75)
$\Phi_{overestimation}$	− (-)	− (-)	− (-)	− (-)
Bayesian Information Criterion (BIC)				
$\Omega_{Correct}$	**33.8 (50.0)%**	**0.0 (0.0)%**	**0.0 (0.0)%**	**11.3 (16.7)%**
Φ	2.27 (1.92)	3.00 (3.00)	3.00 (3.00)	2.76 (2.64)
$\Phi_{underestimation}$	3.43 (3.85)	3.00 (3.00)	3.00 (3.00)	3.11 (3.17)
$\Phi_{overestimation}$	− (-)	− (-)	− (-)	− (-)
Hannan-Quinn Measure (HQM)				
$\Omega_{Correct}$	**90.4 (100.0)%**	**1.2 (17.5)%**	**0.0 (0.0)%**	**30.6 (39.2)%**
Φ	0.16 (0.00)	2.90 (2.33)	3.00 (3.00)	2.02 (1.77)
$\Phi_{underestimation}$	1.65 (-)	2.93 (2.82)	3.00 (3.00)	2.91 (2.92)
$\Phi_{overestimation}$	− (-)	− (-)	− (-)	− (-)
Lower Bound of SSE Technique (LBT)				
$\Omega_{Correct}$	**89.6 (65.5)%**	**2.9 (7.5)%**	**0.0 (0.0)%**	**30.8 (24.2)%**
Φ	0.19 (0.82)	2.87 (2.62)	3.00 (2.95)	2.02 (2.13)
$\Phi_{underestimation}$	1.80 (2.36)	2.96 (2.84)	3.00 (2.95)	2.92 (2.81)
$\Phi_{overestimation}$	− (-)	− (-)	− (-)	− (-)

Table 2 (continued)

	Noise level ϵ			
	0.1	0.4	0.7	Overall
Simplified Cross-validation (CVs)				
$\Omega_{Correct}$	**29.6 (22.5)%**	**14.6 (15.0)%**	**14.2 (20.0)%**	**19.4 (19.2)%**
Φ	1.32 (1.35)	1.96 (1.82)	1.73 (1.77)	1.67 (1.65)
$\Phi_{underestimation}$	2.12 (2.31)	2.46 (2.45)	2.36 (2.32)	2.34 (2.36)
$\Phi_{overestimation}$	1.64 (1.13)	1.97 (1.71)	1.45 (2.00)	1.68 (1.56)
Complex Cross-validation (CVc)				
$\Omega_{Correct}$	**58.8 (57.5)%**	**22.9 (20.0)%**	**9.6 (20.0)%**	**30.4 (32.5)%**
Φ	0.71 (0.80)	1.73 (1.60)	2.25 (1.77)	1.56 (1.39)
$\Phi_{underestimation}$	2.02 (2.27)	2.53 (2.37)	2.65 (2.50)	2.53 (2.40)
$\Phi_{overestimation}$	1.47 (1.17)	1.48 (1.46)	1.53 (1.75)	1.49 (1.52)
Calińsky-Harabasz Index (CH)				
$\Omega_{Correct}$	**13.2 (6.7)%**	**0.0 (0.0)%**	**0.0 (0.0)%**	**13.2 (6.7)%**
Φ	1.51 (.93)	2.68 (2.53)	2.92 (2.88)	2.37 (2.11)
$\Phi_{underestimation}$	2.50 (1.16)	2.68 (2.53)	2.92 (2.88)	2.73 (2.26)
$\Phi_{overestimation}$	− (-)	− (-)	− (-)	− (-)
Dunn Index (DI)				
$\Omega_{Correct}$	**83.3 (37.5)%**	**8.3 (20.0)%**	**2.9 (5.0)%**	**31.5 (20.8)%**
Φ	0.23 (0.83)	3.08 (1.98)	3.46 (3.25)	2.26 (1.98)
$\Phi_{underestimation}$	1.20 (1.16)	1.54 (1.58)	2.00 (1.00)	1.37 (1.29)
$\Phi_{overestimation}$	2.40 (-)	3.48 (3.00)	3.60 (3.49)	3.53 (3.32)
Krzanowski-Lai Index (KL)				
$\Omega_{Correct}$	**17.9 (27.5)%**	**15.0 (15.0)%**	**15.83 (17.5)%**	**16.3 (20.0)%**
Φ	1.65 (1.55)	1.68 (1.75)	1.73 (1.73)	1.69 (1.68)
$\Phi_{underestimation}$	1.75 (1.93)	1.90 (2.00)	1.89 (2.15)	1.85 (2.03)
$\Phi_{overestimation}$	2.31 (2.36)	2.06 (2.09)	2.23 (2.05)	2.20 (2.14)
Average Silhouette Width (ASW)				
$\Omega_{Correct}$	**56.7 (5.0)%**	**2.9 (0.0)%**	**0.0 (0.0)%**	**19.9 (1.7)%**
Φ	0.62 (1.60)	2.68 (2.65)	2.99 (2.95)	2.10 (2.40)
$\Phi_{underestimation}$	1.43 (1.68)	2.76 (2.65)	2.99 (2.95)	2.62 (2.44)
$\Phi_{overestimation}$	− (-)	− (-)	− (-)	− (-)

The accuracy Ω is displayed for the true number of clusters ($\Omega_{Correct}$; in bold). $\Phi_{underestimation}$ and $\Phi_{overestimation}$ pertain to the precision Φ for data sets that respectively under- and overestimate the true number of overlapping clusters

Precision An examination of the precision Φ of the model selection strategies, which is presented in Table 2 with smaller Φ-values indicating the strategy being more precise, further confirmed CHull and AIC being the best and second best performing strategies, respectively, with all other strategies performing substantially worse. In OL data sets, CHull NLL ($\Phi = 0.67$) and CHull LSQ ($\Phi = 0.73$) performed best in terms of precision, followed by AIC ($\Phi = 1.30$) which performed worse by a factor of 2 compared to CHull NLL. The fourth to seventh best strategies were complex ($\Phi = 1.56$) and simplified ($\Phi = 1.67$) CV,

KL ($\Phi = 1.69$), and AICc ($\Phi = 1.70$), respectively. HQM ($\Phi = 2.02$), LBT ($\Phi = 2.02$), ASW ($\Phi = 2.10$), DI ($\Phi = 2.26$), CH ($\Phi = 2.37$) and BIC ($\Phi = 2.76$) performed worse by multiple orders of magnitude compared to CHull, with the average distance from the estimated to the true number of clusters being larger than two for all these five strategies (i.e., selecting a solution that has two clusters more or less than the optimal solution with the true number of clusters). A similar pattern was encountered for NOL data sets in which AIC ($\Phi = 0.95$) performed worse than CHull NLL ($\Phi = 0.38$) and CHull LSQ ($\Phi = 0.48$) by a factor of 2 to 2.5 orders of magnitude. Similar to OL data sets, complex CV ($\Phi = 1.39$), AICc ($\Phi = 1.40$), simplified CV ($\Phi = 1.65$) and KL ($\Phi = 1.68$) come in fourth to seventh place, respectively, followed by HQM ($\Phi = 1.77$), DI ($\Phi = 1.98$), CH ($\Phi = 2.11$), LBT ($\Phi = 2.13$), ASW ($\Phi = 2.40$) and BIC ($\Phi = 2.64$).

Degree of Over- and Underestimation To study the degree to which each model selection strategy over- or under-estimates the true number of clusters in ADPROCLUS, we computed the Φ-measure, taking for each strategy the data sets in which respectively overestimation ($\Phi_{overestimation}$) and underestimation ($\Phi_{underestimation}$) occurred separately. The mean $\Phi_{underestimation}$ and $\Phi_{overestimation}$ for each model selection strategy, computed globally and per noise level, is presented in Table 2. In this table, one can see that, when not identifying the true number of clusters, the CH Index, LBT, ASW and all information-theoretic approaches (i.e., AIC, AICc, BIC, HQM) have a strong tendency to underestimate the number of clusters in ADPROCLUS. In particular, for these methods, underestimation occurs for all data sets (in both OL and NOL data sets) in which the estimated number of clusters does not equal the true number of clusters. As such, overestimation was never observed for these strategies. For AIC, this may be a rather surprising result as this strategy is known—as demonstrated in several studies—for its tendency to select too complex models (i.e., overestimation) in regression, partitioning methods and mixtures of factor analyzers (Steinley & Brusco, 2011; Vrieze, 2012; Bulteel et al., 2013). These studies, however, were performed under different conditions than the current study, which may—partly—explain this rather surprising result. Nevertheless, in general, which is in line with previous studies, AIC selects a more complex model (i.e., a smaller amount of underestimation) than AICc and BIC. For the Dunn index, overestimation occurred to a stronger extent than underestimation, whereas the opposite is true for the cross-validation based strategies. For the CHull strategies and the KL Index, over- and underestimation was encountered to the same extent. In general, under- and overestimation becomes more severe with increasing noise levels, except for the simplified cross-validation (CVs) strategy and the KL Index in which the effect of the noise level is less univocal. For the information-theoretic strategies (and also for LBT and CH), underestimation was (almost) maximal for the largest noise level, implying that for these data sets, the information-theoretic measures (almost) always erroneously select the most simple model with a single cluster. This pattern of results seems to suggest that for ADPROCLUS information-theoretic strategies punish the model fit too hard by the complexity of the model, indicating that the definition of model complexity fp used may not be optimal for these strategies (for a further discussion of this issue, see Section 4.4).

4.3.2 Effect of Data Characteristics on Accuracy and Precision

Accuracy The percentage of data sets for which the true number of clusters is identified (i.e., the mean accuracy $\Omega_{Correct}$ across data sets), overall and per model selection strategy, is displayed in Table 3 as a function of the levels of the manipulated data characteristics. From this table, it can be concluded that, overall, accuracy was more severely impaired for

Table 3 Mean accuracy Ω (displayed as a percentage and rounded to one decimal place) computed for each model selection strategy separately and across all model selection strategies, as a function of the various levels of the manipulated data characteristics

		CHull	CHull											
Factor	Level	NLL	LSQ	AIC	AICc	BIC	HQM	LBT	CVs	CVc	CH	DI	KL	ASW	Overall
Number	3	73.3	70.7	58.8	47.6	24	32.6	30	15.5	31.4	16	30	16.2	17.6	35.7
of clusters	5	58.1	53.1	40.2	31.2	0	31	29.8	23.3	30	8.6	30	17.4	16.9	28.3
Number	200	62.9	60	50.2	38.6	11.4	30.7	29.3	20	29	11	29.3	16.9	16.7	31.2
of objects	400	68.6	63.3	48.8	40.2	12.6	32.9	30.5	18.8	32.4	13.6	30.7	16.7	17.9	32.8
Cluster	0%*	75.8	70	57.5	49.2	16.7	39.2	24.2	19.2	32.5	6.7	20.8	20	1.6	33.3
overlap	35%	72.5	68.3	54.2	43.6	15.3	33.9	30.6	22.5	33.9	17.2	34.7	21.9	23.6	36.3
	75%	55.2	52.2	42.2	31.9	7.2	27.2	31.1	16.4	26.9	9.2	28.3	10.6	16.1	27.3
Degree of	none	74.2	69.4	55.3	43.6	15.3	34.7	29.7	18.9	32.5	8.9	27.8	16.4	17.2	34.1
non-occurring	medium	61.7	58.8	49.2	40	12.5	33.8	33.3	24.2	30.8	12.9	32.9	15.4	20.8	32.8
patterns	high	56.7	52.9	41.3	32.5	6.7	25.4	26.7	15.4	27.9	16.7	30.4	18.8	13.8	28.1
Noise	0.1	99.6	99.6	100	95.0	36.1	91.8	86.1	28.6	58.6	36.8	76.8	19.3	49.3	67.5
level	0.4	72.9	65.0	47.9	23.2	0	3.6	3.6	14.6	22.5	0	10	15	2.5	21.6
	0.7	24.3	20.4	0.7	0	0	0	0	15.0	11.1	0	3.2	16.1	0	7

*NOL data sets ($N = 120$)

data sets that contain more noise (i.e., $\Omega = 67.5\%$, 21.6% and 7.0% for $\epsilon = 0.1$, 0.4 and 0.7, respectively), for data sets with more underlying clusters (i.e., $\Omega = 35.7\%$ for three- and 28.3% for five-cluster data sets), for data with a larger amount of cluster overlap (i.e., $\Omega = 27.3\%$ for large versus 36.3% and 33.3% for medium and no overlap, respectively) and for data with a larger degree of non-occurring overlapping membership patterns (i.e., for a high and medium degree $\Omega = 28.1\%$ and 32.8%, respectively, versus 34.1% for no overlap). For the number of objects in the data, only a small effect was observed (i.e., $\Omega = 31.2\%$ versus 32.8% for 200 and 400 objects, respectively).

Precision A mixed-ANOVA with Φ as dependent variable suggests that for OL data sets performance deteriorates the strongest when the noise level increases ($\eta^2_G = 0.60$; $\Phi = 0.67$, 2.05 and 2.55 for $\epsilon = 0.1$, 0.4 and 0.7, respectively) and the true number of overlapping clusters underlying the data increases ($\eta^2_G = 0.31$; $\Phi = 1.28$ and 2.23 for $k = 3$ and 5, respectively). Similarly, for NOL data sets, performance also deteriorates the strongest with increasing noise level ($\eta^2_G = 0.44$; $\Phi = 0.75$, 1.73 and 2.37 for $\epsilon = 0.1$, 0.4 and 0.7, respectively) and an increasing true number of overlapping clusters ($\eta^2_G = 0.31$; $\Phi = 1.11$ and 2.12 for $k = 3$ and 5, respectively).

Further, the choice of model selection strategy had a strong effect ($\eta^2_G = 0.40$ and 0.43 for OL and NOL data sets, respectively) on model selection performance, suggesting large performance differences between strategies. In particular, overall, both CHull versions are clearly more precise than AIC and the cross-validation based strategies, with BIC performing the worst (see Table 2). The main effect of the model selection strategies, however, is qualified by a model selection strategy by number of clusters interaction ($\eta^2_G = 0.28$ and 0.26 in OL and NOL data sets, respectively) and a model selection strategy by noise level interaction ($\eta^2_G = 0.33$ and 0.24 in OL and NOL data sets, respectively), both of which are visualized in Fig. 1 (for OL data sets). Regarding the interaction with the number of clusters (see the upper left panel in Fig. 1), in general, for both OL and NOL (not shown in figure) data sets, the performance deterioration with increasing numbers of underlying clusters is stronger for strategies that perform overall worse than for strategies that perform overall better. An exception to this are the cross-validation based strategies for which precision is nearly influenced by the number of clusters and AIC for which precision increased with increasing number of clusters. With respect to the interaction between model selection strategies and noise level (see the upper right panel in Fig. 1), in both OL and NOL (not shown in figure) data sets, the decrease in precision differs between the best and worst performing model selection strategies. In particular, for the overall best performing strategies (i.e., CHull NLL, CHull LSQ, and AIC), the deterioration in precision becomes stronger when the noise increases (i.e., a larger deterioration when going from $\epsilon = 0.4$ to 0.7 compared to ϵ going from 0.1 to 0.4); for the overall worse performing strategies (i.e., BIC, ASW, CH, DI, HQM and the cross-validation based strategies), the strongest deterioration is observed for ϵ going from 0.1 to 0.4 instead of from 0.4 to 0.7. For KL, the precision is almost not influenced by the amount of noise in the data. Both interaction effects are qualified by a three-way model selection strategy by number of clusters by noise level interaction ($\eta^2_G = 0.16$ and 0.19 in OL and NOL data sets, respectively). In the lower panels of Fig. 1, in which this three-way interaction is displayed, one can see that for all model selection strategies—except the DI Index—the deterioration in precision with increasing noise level is stronger for data sets with five underlying clusters (bottom right panel) than for data sets with three underlying clusters (bottom left panel). Again, for KL, the precision is almost constant across noise levels.

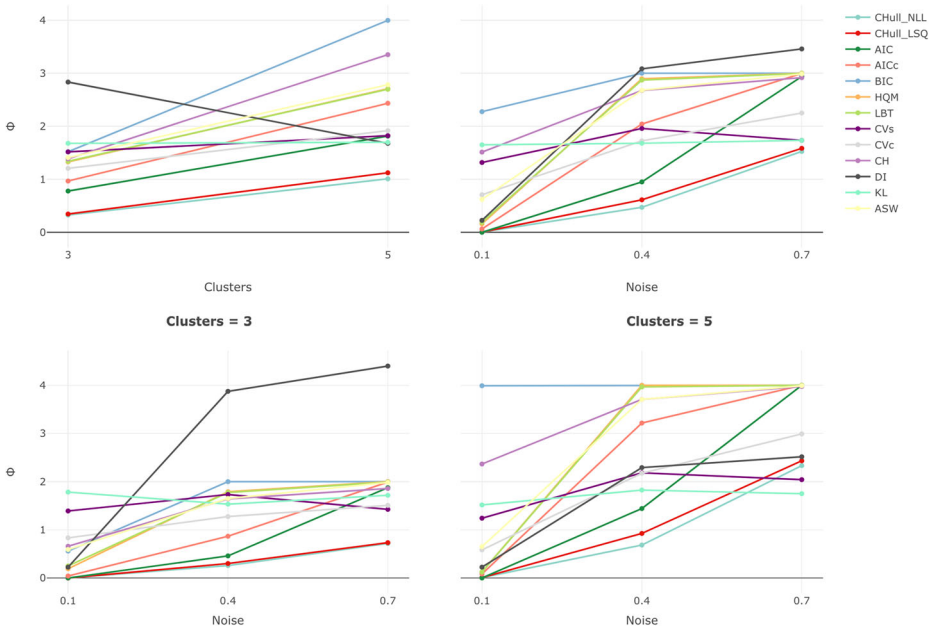

Fig. 1 Two-way interactions (upper panels) and three-way interaction (lower panels) for the mixed-ANOVA on OL data sets: (upper left) Precision Φ as a function of the true number of clusters K and the model selection strategy used; (upper right) Precision Φ as a function of the noise level ϵ and the model selection strategy used; (lower panels) Precision Φ as a function of the noise level ϵ, the model selection strategy used and the true number of clusters K for $K = 3$ (left panel) and $K = 5$ (right panel)

4.3.3 Comparison between the Best Performing Model Selection Strategies: CHull NLL and AIC

A cross-tabulation of the accuracy of CHull NLL and AIC, which is depicted in Table 4, shows that for 48.9% of the data sets both strategies correctly identify the true underlying number of clusters, whereas both strategies fail to do so for 33.8% of the data sets. The 284 data sets for which both methods fail mainly belong to the very hard conditions that are characterized by a large noise level of $\epsilon = 0.7$ ($N = 212$), a large amount of cluster overlap ($N = 160$) and five clusters underlying the data ($N = 160$). When looking at the data sets where AIC retrieves the correct model (i.e., for 49.5% of the data sets), CHull almost always also identifies the correct model (i.e., 99% of the cases); on the contrary, AIC only selects the correct model in 74.6% of the data sets for which CHull retrieves the correct model (i.e., for 65.6% of the data sets). When investigating the cross-tabulation across noise levels, CHull appears to consistently identify the correct model for the data sets for which AIC is correct (99.6%, 97% and 100% of the cases for $\epsilon = 0.1$, 0.4 and 0.7, respectively), while in data sets where CHull retrieves the correct model, AIC is only correct at low noise levels and AIC's performance deteriorates drastically with increasing noise (100%, 63.7% and 2.9% of the cases for $\epsilon = 0.1$, 0.4 and 0.7, respectively). It can be concluded that AIC does not have a substantial advantage over CHull and that the advantage of CHull over AIC becomes more prominent when the data contain more noise. In particular, the 140 data sets for which CHull—but not AIC—retrieved the correct model belong mainly to more difficult

Table 4 Cross-tabulation of the number of correctly retrieved models for CHull NLL and AIC, computed overall and per noise level ϵ

		AIC	
		Correct	Incorrect
$\epsilon = 0.1$ ($N = 280$)			
CHull	Correct	279	0
NLL	Incorrect	1	0
$\epsilon = 0.4$ ($N = 280$)			
CHull	Correct	130	74
NLL	Incorrect	4	72
$\epsilon = 0.7$ ($N = 280$)			
CHull	Correct	2	68
NLL	Incorrect	0	212
Overall ($N = 840$)			
CHull	Correct	411	140
NLL	Incorrect	5	284

conditions with a larger noise level of $\epsilon = 0.4$ ($N = 74$) and $\epsilon = 0.7$ ($N = 66$) and 400 observations ($N = 83$).

4.3.4 Computation Time of the ADPROCLUS Algorithm

Table 5 displays the computation time for an average multi-start of the ADPROCLUS algorithm as a function of the number of clusters K and the number of objects I, the two factors that have the largest impact on the computation time. It appears that the computation time increased drastically both with increasing number of clusters K and increasing number of objects I. On average, the computation time for a solution with K clusters was about two

Table 5 Mean computation time (in minutes, rounded to two decimal places) and, between brackets, standard deviations, for a single multi-start of the ADPROCLUS algorithm, averaged across all simulated data sets, as a function of the number of objects I and the number of clusters K

	Observations I	
Number of clusters K	200	400
1	0.44 (0.15)	1.28 (0.44)
2	1.99 (0.70)	5.49 (1.95)
3	4.86 (1.81)	14.25 (5.96)
4	13.28 (3.60)	43.24 (13.48)
5	31.91 (8.73)	110.29 (34.96)
6	80.20 (13.32)	286.37 (55.15)
7	97.44 (16.04)	393.50 (61.50)
8	230.73 (34.71)	942.99 (126.38)
Sum	460.84 (58.50)	1797.42 (234.77)

to three times longer than the computation time for a solution with $K - 1$ clusters. Likewise, computation of ADPROCLUS solutions for data sets with $I = 400$ objects took, on average, 3.90 times longer compared to data sets with $I = 200$ objects.

4.4 Discussion of the Results

4.4.1 Model Complexity: Post Hoc Analysis

As suggested by previous research on model selection in other statistical models (e.g., Hamerly & Elkan, 2004; Steinley & Brusco, 2011; Preacher et al., 2013), model complexity in this study was defined as the number of estimated model parameters: $fp = ((I+J)\times K)+1$ (see Section 3). Defining complexity in this way, however, is based on the assumption that each parameter, regardless of its type (i.e., binary membership versus real-valued profile), should be weighed equally when determining the complexity of a model. For ADPROCLUS, one may doubt whether it makes sense to consider a binary membership parameter (from **A**) as important for the definition of model complexity as a real-valued profile parameter (from **P**). This implies that our definition of model complexity for ADPROCLUS may not be optimal.

The strong tendency of information-theoretic strategies to underestimate the number of clusters, as demonstrated in our simulation study, suggests that the associated test statistics penalize complex models too hard. In particular, the information criteria used in this study all select the model that has the lowest value on a test statistic that takes the following general form:

$$\text{IC} = (\text{Model likelihood}) + \lambda(\text{Model complexity}), \tag{31}$$

where the first term decreases and the second term increases for more complex models, and λ is a weight that determines the penalty for increasing model complexity. As such, the number of clusters in ADPROCLUS may be underestimated when complex models are penalized too much. One reason for this over-penalization may be that the complexity of a model is not well defined (i.e., *fp* does not represents the true complexity of the model). Applied to ADPROCLUS, it seems that the definition of *fp* in (8) yields a too large estimate of the complexity of the model.

Relevant in this regard is the work of Lee (2001) on model complexity of additive clustering models. Bear in mind that additive clustering is similar—but not identical—to additive profile clustering. In particular, additive clustering models, like the ADCLUS model (Shepard & Arabie, 1979), operate on object by object (dis)similarity data instead of on object by variable data. Moreover, additive clustering tries to explain the observed (dis)similarities in terms of a set of overlapping clusters and associated weights that are summed across all clusters that are shared by the members of an object pair. Additive overlapping clustering, however, tries to reconstruct object by variable data by means of an overlapping object clustering and a set of associated cluster profiles that are summed across the clusters an object belongs to. In an attempt to derive a BIC statistic for additive clustering models, Lee (2001) used the number of clusters K as an estimate for model complexity (i.e., $fp = K$). Using this very conservative estimate of model complexity in the AIC and BIC measure (see (14) and (16)) for ADPROCLUS resulted in an overestimation of the number of clusters for all 840 data sets. In particular, for each data set, the most complex model (i.e., $K = 8$) was erroneously identified as the correct model by both AIC and BIC. It can be concluded that a sensible estimate for model complexity in an ADPROCLUS context should be substantially larger than the number of clusters K and substantially smaller than the estimate given in (8).

A post hoc analysis was conducted to investigate how the performance of AIC changes as a function of the definition of model complexity. Note that the AIC test statistic (14) can be rewritten as

$$\text{AIC} = 2(\text{-}l) + 2(w \times fp), \tag{32}$$

where $w = 1$ for AIC and $w = \frac{n}{n-fp-1}$ (AICc), $w = \frac{\log(n)}{2}$ (BIC) and $w = \log(\log(n))$ (HQM) for the other information-theoretic measures. To investigate the effect of using other estimates for model complexity (i.e., $w \times fp$) on the model selection performance of information-theoretic strategies in ADPROCLUS, the parameter w was manipulated to increase in steps of 0.025 over the interval between 0.5 and 1 (i.e., 0.500, 0.525, 0.550, ... , 1) and this post hoc AIC-based strategy with all different w-values was applied to all 840 simulated data sets. In Fig. 2, mean accuracy (upper panel) and precision (lower panel) is displayed for each value of w, with the dashed black line indicating the performance of CHull NLL, which was the best performing model selection strategy in the simulation study. Both accuracy and precision varied dramatically for different values of w. Accuracy ($\Omega = 77\%$) and precision ($\Phi = 0.53$) were both maximal for $w = 0.625$. Note that when using the optimal w, the post hoc AIC-based strategy clearly outperforms both CHull versions in terms of accuracy and precision. Even for w-values close to the optimal w, the post hoc strategy outperforms CHull NLL. When investigating the proportion of correctly retrieved models by the post hoc strategy for $w = 0.625$ as a function of noise level, we see that for data sets with noise level $\epsilon = 0.1$ ($\Omega = 100\%$) and $\epsilon = 0.4$ ($\Omega = 90.0\%$), the post hoc AIC-based strategy could reliably estimate the number of clusters, whereas performance decreased to 31.4% for data sets with a noise level of $\epsilon = 0.7$.

It is not clear to which extent the optimal value of w and associated optimal definition of model complexity for ADPROCLUS as $fp = 0.625(IK + JK + 1)$ is applicable to data sets with other data characteristics than the data sets generated in the simulation study. In particular, the optimal w-value may strongly depend on the number of objects I and/or on the number of variables J that was used for the simulated data. Further research on this issue may show how generalizable the obtained results are and may result in a modification of the complexity value fp that appropriately captures the true complexity of the ADPROCLUS model. However, the results demonstrate the usefulness of information criteria to tackle the non-trivial model selection problem in ADPROCLUS, given an appropriate definition of model complexity is used.

4.4.2 Omega: Alternative Instability Measure for the Cross-validation Based Strategies

Both cross-validation based strategies, CVs and CVc, performed very poor in the simulation study. One reason for this may be the particular choice of instability measure s in (30). Instead of looking at the misfit for $\mathbf{X}_3$ when using $\mathbf{P}_1$ and $\mathbf{P}_2$ as profile matrices, we can alternatively look at the dissimilarity between the obtained clusterings for $\mathbf{X}_3$ based on $\mathbf{P}_1$ and $\mathbf{P}_2$. To this end, the Omega index (Collins & Dent, 1988), which generalizes Hubert and Arabie's Adjusted Rand Index (ARI; Hubert & Arabie, 1985) to the case of overlapping clustering, may be used. In particular, the Omega index between $\mathbf{A}_3^{(\mathbf{P}_1)}$ and $\mathbf{A}_3^{(\mathbf{P}_2)}$ may be adopted as a stability measure. Applied to the 840 data sets at hand, the accuracy of simplified CV using the Omega index improved to 29.4% (compared to 19.4% using the original test statistic (30)) and accuracy of complex CV improved to 34.8% (compared to 30.4% using (30)). Although the Omega index seems to somewhat improve the performance of the proposed cross-validation strategies, the performance of the CV strategies is not at the level of computationally less demanding methods like CHull and AIC.

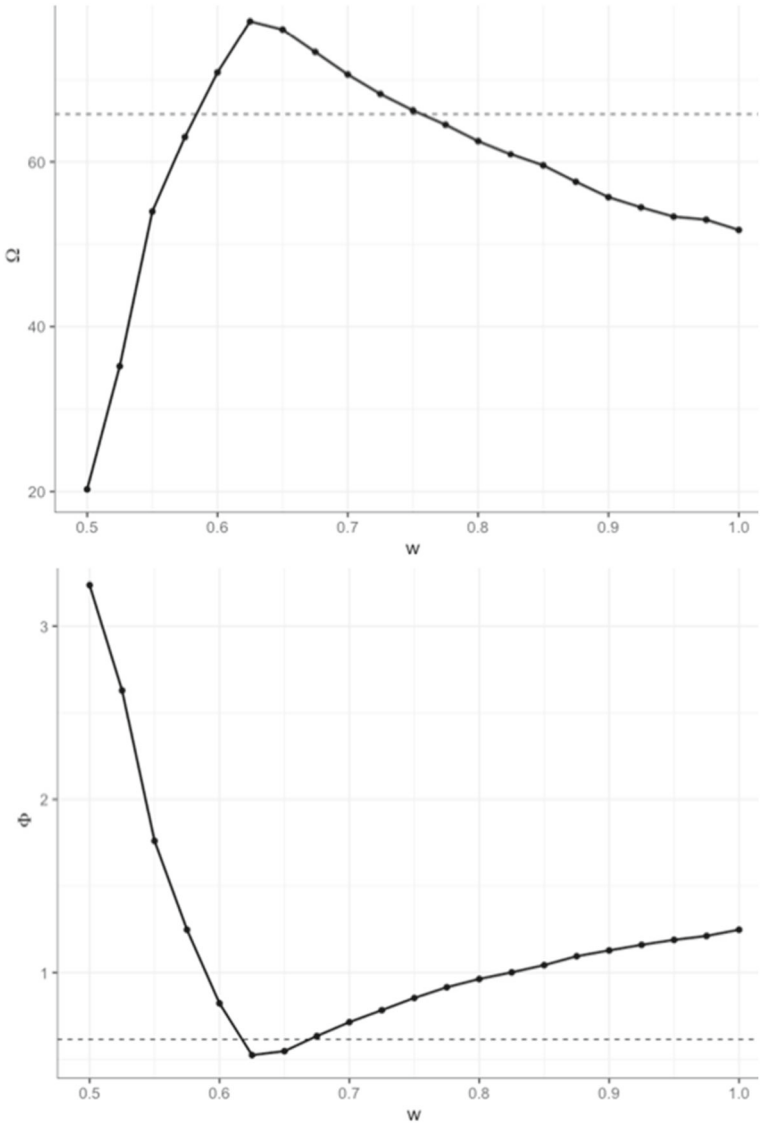

Fig. 2 Accuracy Ω (upper panel) and precision Φ (lower panel), averaged across all data sets, of selecting the optimal number of clusters in ADPROCLUS for different values of the weighting parameter w in a post hoc analysis of AIC performance

4.4.3 Generalizability of the Results to Data Sets with a Different Number of Variables

In the simulation study, the number of variables in the data set was kept constant at $J = 15$. In order to show that the performance results of the different model selection strategies that were reported above can also be generalized to data sets with a different number of variables, we ran a second simulation study in which we fixed the number of variables at $J = 8$. As the ANOVA results (see Section 4.3.2) pointed out that the number of objects had a negligible

effect on the performance of the different model selection strategies, in order to reduce the computation time, we fixed the number of objects to 200. As such, the *size*, $I \times J$, of data matrix $\mathbf{X}$ was kept constant at 200×8. All other factors and manipulated factor levels stayed the same as described in Section 4.1, resulting in the generation of 420 (60 NOL and 360 OL) data sets. For each generated data set, we fitted ADPROCLUS models with $K = 1$ up to $K = 6$ clusters. Again, in order to reduce computation time and because increasing the number of clusters from $K = 6$ to $K = 8$ for the simulation study reported above did not change the pattern of the performance results, we decided to run ADPROCLUS only with $K = 1$ up to $K = 6$ clusters. For each analysis, we applied a multi-start procedure with 50 starts (also for the analyses for both cross-validation based procedures with again using fivefold cross-validation for the complex cross-validation procedure).

The results show that, for OL data sets, both CHull strategies perform best (overall mean accuracy $\Omega = 41.94\%$ and 40.83% for CHull NLL and CHull LSQ, respectively). The next best procedure for OL data sets is AIC (32.22%), followed by the Dunn (DI; 30.8%) and Krzanowski-Lai Index (KL; 28.06%). The LBT (22.5%), CVc (22.22%) and CVs (21.11%) procedure perform equally worse. The worst performing strategies are HQM (10.83%), AICc (10.28%), ASW (9.44%), CH (6.94%) and BIC (0%). For NOL data sets, both CHull strategies performed best, with an overall mean $\Omega = 55\%$ and 48.33% for CHull NLL and CHull LSQ, respectively. Again, the third best performing method is the AIC procedure (41.67%). The CVs(25%), CVc (25%), KL (23.33%), DI (18.33%), HQM (16.67%), AICc (16.67%) and LBT(15%) showed similar results for the NOL data sets. The CH (6.67%), ASW (0%) and BIC (0%) Index failed completely. In general, it appears that the accuracy of the considered model selection strategies is smaller for data sets with $J = 8$ compared to $J = 15$ variables, with this effect being more pronounced for the best performing (i.e., CHull) strategies. A possible reason for this decrease in performance could be the availability of less data (i.e., only 8 instead of 15 variables) to estimate the ADPROCUS models (with varying k), which may result in clustering solutions of a lower quality, implying a more difficult model selection task.

5 Discussion

The problem of model selection in ADPROCLUS, specifically, and in overlapping clustering, in general, has not yet been tackled in a satisfactory way nor has it been studied extensively. Therefore, in this paper, being one of the first to do so, the capacity of different model selection strategies to determine the correct number of overlapping clusters in ADPROCLUS is investigated in a systematic way by means of a large-scale simulation study. The results demonstrated that for the more easy data sets (i.e., low amounts of noise and a low number of underlying clusters), model selection strategies based on CHull and AIC reliably selected the model with the correct number of clusters. The performance of CHull proved to be more stable than AIC in more difficult, but also more realistic, data conditions (i.e., more noise in the data, a larger number of underlying clusters that show more overlap, and the non-occurrence of certain membership patterns that indicate overlap). The amount of noise in the data had the largest effect on the overall performance and data sets with a large noise level posed a major difficulty for all model selection strategies. In the more realistic data conditions, AIC, as well as the three other information criteria (i.e., AICc, BIC and HQM), tended to underestimate the number of clusters, suggesting that for more complex models (i.e., with more clusters) these criteria penalize the model fit too

strongly. In general, when not identifying the correct number of clusters, CHull tends to under/over-estimates the number of clusters to a lesser degree than AIC and the other model selection strategies. The results for the Caliński-Harabasz and Dunn Index, the LBT measure (Steinley & Brusco, 2011), the KL and ASW Index and the proposed cross-validation (CV) based strategies were disappointing, implying that these methods should not be used for model selection in ADPROCLUS. It can be concluded that CHull (Ceulemans et al., 2011; Wilderjans et al., 2012), especially the version in which the likelihood (11) is used as fit measure, is the best strategy for determining the optimal ADPROCLUS model out of a range of models with an increasing number of clusters.

In the remainder of this section, several issues related to determining the number of clusters in ADPROCLUS are discussed that deserve further investigation. First, the presence of—large amounts of—noise in the data that may contain some cluster structure is discussed. Second, some thoughts about the local minima problem in ADPROCLUS are presented. Third, some considerations on the scalability of ADPROCLUS algorithms is presented and, finally, determining whether data need to be modelled with an overlapping versus non-overlapping cluster structure is discussed.

Noise in the Data That Has a Cluster Structure For all design cells in the simulation study with large amounts of noise (i.e., 70% of the data is noise), it is observed that for none of the model selection strategies the performance is at a satisfactory level. In particular, the best—although not very encouraging—results are obtained for both CHull methods (i.e., an accuracy Ω around 20%) and both cross-validation methods (i.e., Ω between 10% and 15%), whereas for the other strategies the optimal number of cluster K is (almost) never identified (i.e., Ω equals or is close to zero). A possible reason for this poor performance of the model selection strategies when the data are very noisy is that the noise may contain some—overlapping—cluster structure which misguides the algorithm and, as a consequence, obstructs the algorithm from uncovering the true cluster structure and true number of clusters underlying the data. This may especially be the case when the amount of noise in the data is large and when clusters overlap to a large extent as in these situations the true clusters are hard to disclose and easily can be obscured by cluster structure present in the noise. An indication in this direction are the results of a study of Brusco (2004) in which, in the context of clustering binary data, the effect of structured noise on the recovery of the true cluster structure is investigated. In particular, this study shows that, when generating data with two different true underlying cluster structures, the clustering algorithm tends to only uncover one of both true structures. Moreover, the recovery of this true cluster structure is hampered by the other true cluster structure, which in this case could be considered as structured noise. Indeed, the structure in the noise is picked up by the algorithm, herewith obscuring the simulated true cluster structure, which can be expected to have a serious deteriorating effect on the performance of any model selection strategy. This may especially be the case when the true structure is rather weak compared to the—strong—structure present in the noise (Ceulemans & Kiers, 2009).

Local Minima in ADPROCLUS All model selection strategies proposed in this paper rely in some way on the fit of the ADPROCLUS model (i.e., sum of squared residuals or negative log-likelihood) to the data. As such, when comparing model selection strategies, it is of utmost importance that for each ADPROCLUS model included in the model comparison (e.g., a set of models with increasing numbers of clusters), the global optimal solution is found and its (optimal) fit value is used in the calculations for the model selection strategies; otherwise, the local minima problem in ADPROCLUS may negatively impact—as it

may get intertwined with—the ADPROCLUS model selection problem. The ADPROCLUS loss function, however, is known to suffer from a severe local minima problem, implying that the algorithm sometimes retains a local optimal solution that fits the data worse than the global optimal solution (Depril et al., 2008). Moreover, when the number of clusters of the ADPROCLUS model and the amount of noise in the data increases, it may be expected that the local minima problem enlarges as the former implies a larger solution space and the latter an increase in the number of local optimal solutions in which the algorithm can get stuck. To mitigate the local minima problem, often an extensive multi-start procedure is used in which the algorithm is run multiple times and the best solution encountered across all runs is retained. However, the number of multi-starts used is often kept the same for each K, implying that for smaller K, where the problem of local minima is expected to be less of an issue, the algorithm may mitigate the local minima problem to a larger extent than for a larger K. It is not clear yet how this problem of local minima and the dependence of its severity on K affects the proposed model selection strategies. It is also not yet thoroughly investigated how the number of starts in a multi-start procedure should be tuned to account for the differences in severity of the local minima problem for increasing values of K. To deal with the larger local minima problem caused by larger amounts of noise in the data, an interesting proposal has been launched by Steinley (2006). The author observes that, when the data contain noise and clusters are hard to disentangle from each other (because, for example, clusters overlap to a larger extent), the global optimum solution is not necessarily the solution that captures the true underlying cluster structure the best. As a consequence, in such a situation, one should not only look at the global optimum solution but also take local optimal solutions into account. As a way out, Steinley (2006) proposes to profile the local optima in order to aid the model selection process. Further research regarding these issues is highly recommended and may also have important ramifications for other problems suffering from local minima, like K-means, model-based clustering and biclustering.

Limited Size of the Data Sets Used and Scalability of the ADPROCLUS Algorithm The data sets used in this simulation study were limited to only include 200 or 400 objects and a small number of variables (i.e., 8 or 15). Data sets in scientific practice, however, easily have a (much) larger number of objects and variables. It is not clear whether the (slight) trend of model selection strategies performing better when the data size increases, as observed in the simulation, also holds when the number of objects and variables becomes (really) large. Fitting ADPROCLUS models to large data sets, however, may be problematic for two reasons. First, the increasing severity of the local minima problem when the data size and thus the number of parameters to be estimated (i.e., solution space) increases (see earlier comment). And secondly, the enormous increase in computational burden for the ADPROCLUS algorithm, which may become prohibitive for large numbers of objects I and/or variables J, even when parallel computing infrastructure can be used. Note that the sometimes large computation time for fitting the ADPROCLUS model—even when applied to data of a moderate size—and this especially for larger number of clusters, limits to some extent the real world application of the proposed model selection strategies as these strategies require the fitting of ADPROCLUS models with increasing number of clusters. Regarding the latter, as a way out, it may be worthwhile to develop a faster and more scalable algorithm for ADPROCLUS. A good starting point here may be the work of France and Abbassi (2011) in which additive clustering algorithms are boosted by combining them with clusterwise optimization and multi-label learning strategies. The rationale of this approach consists of training an overlapping clustering model, like ADPROCLUS, on a manageable subset of the data

and using the resulting cluster solution subsequently in a multi-label learning method to also find the cluster memberships of the objects that were omitted in first instance. Increasing the algorithmic speed of the ADPROCLUS algorithm will improve the usability of ADPROCLUS and the proposed model selection strategies in empirical practice.

Selecting between Overlapping and Non-overlapping Underlying Cluster Structures Before identifying the optimal number of clusters in ADPROCLUS, it first should be determined whether an overlapping clustering structure is underlying the data at hand, warranting the use of ADPROCLUS as such. When clusters do not overlap, and thus the true cluster structure is a partition, K-means or another partitioning method (e.g., model-based clustering) should be preferred over ADPROCLUS. Just performing ADPROCLUS on the data and checking whether or not objects are assigned to multiple clusters is not a good strategy to determine the nature of the cluster structure underlying a data set at hand as the data contain noise which may blur the underlying cluster structure. In particular, in our simulation study, an ADPROCLUS model often yielded a solution with overlapping clusters, even when the true underlying cluster structure was a partition. In particular, for data sets without cluster overlap, ADPROCLUS assigned, on average, 59.5% of the objects to multiple clusters; in general, this tendency became larger when the data contained more noise (i.e., 31.3%, 79.4% and 67.8% for $\epsilon = 0.1$, 0.4 and 0.7, respectively).

An alternative, and probably better, option to determine the nature of the underlying cluster structure consists of constructing a test to determine whether clusters overlap. One way to obtain this is to derive a similar minimal stochastic extension for K-means as was proposed for ADPROCLUS. In particular, such an extension of K-means would lead to the same formulas (9)-(13) as presented in Section 3.1.1, with only m_{ij} now being the predicted data value under K-means instead of under ADPROCLUS. As the stochastic extension of the K-means model is nested within the similar extension of the ADPROCLUS model (i.e., the restriction in K-means that rows of $\mathbf{A}$ should sum to one is omitted in ADPROCLUS), a likelihood ratio (LR) test may be constructed that tests the hypothesis whether the more simple K-means model fits the data equally well as the more complex ADPROCLUS model. When this hypothesis is rejected, it can be concluded that the underlying cluster structure indeed shows overlap. In such a LR test, the change in likelihood between both models, when applied to the same data set with the same number of clusters K, is weighted against the difference in model complexity. In other words, it is determined whether the extra freedom—and complexity—in the ADPROCLUS model (i.e., allowing clusters to overlap) is compensated by a substantial increase in model fit/likelihood.

Applying this LR test (using as model complexity for ADPROCLUS (8) and for K-means $fp = I + JK + 1$) to the data sets from our simulation study resulted in 49.0% of the OL data sets being correctly identified as data sets with an underlying overlapping cluster structure, whereas only 14.2% of NOL data sets were erroneously identified as data sets containing overlapping clusters. It should be noted that for NOL data sets, the true underlying model did not fully comply with the K-means model (i.e., the underlying cluster structure being a partition) as always 5% of the objects belonged to no cluster at all (i.e., have an all-zero membership pattern). For data sets with a larger noise level ϵ, the LR test, in general, less often indicated an overlapping cluster structure, resulting in the LR test being less good in detecting an overlapping cluster structure when the noise in the data increased. In particular, when $\epsilon = 0.1$, for OL data sets, 97.1% of them were correctly identified as containing an overlapping cluster structure. For $\epsilon = 0.4$ and $\epsilon = 0.7$, this percentage shrank to 50% and 0%, respectively. Moreover, all NOL data sets that were erroneously marked as showing cluster overlap belonged to the condition with almost no noise ($\epsilon = 0.1$).

Some caution regarding the use of such a LR test is needed as its performance critically depends on an adequate definition of model complexity for the models compared (i.e., the distribution of the test statistic depends on the difference in model complexity between the models compared). However, both for ADPROCLUS and *K*-means, the definition of model complexity is still an issue that is not yet resolved in a satisfactory way.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723.

Azaouzi, M., Rhouma, D., & Bem Romdhane, L. (2019). Community Detection in Large-Scale Social Networks: State-of-the-art and Future Directions. *Social Network Analysis and Mining, 9*(23).

Baadel, S., Thabtah, F., & Lu, J. (2015). Mcoke: Multi-cluster Overlapping *K*-means Extension Algorithm. *International Journal of Computer, Electrical, Automation, Control and Information Engineering*, *9*(2), 427–430.

Baadel, S., Thabtah, F., & LU, J. (2016). Overlapping clustering: a review. In *Proceedings of the 2016 SAI Computing Conference* (pp. 233–237). London: IEEE.

Bakeman, R. (2005). Recommended effect size statistics for repeated measures designs. *Behavior Research Methods*, *37*(3), 379–384.

Banerjee, A., Krumpelman, C., Basu, S., Mooney, R. J., & Ghosh, J. (2005). Model-Based Overlapping clustering. In R. L. Grossman, R. J. Bayardo, & K. Bennett (Eds.) *Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD05)* (pp. 532–537). New York: Association for Computing Machinery.

Battle, A., Segal, E., & Koller, D. (2004). Probabilistic Discovery of Overlapping Cellular Processes and Their Regulation Using Gene Expression Data. In *Proceedings of the Eigth International Conference on Research in Computational Molecular Biology (RECOMB-2004)* (pp. 909–927).

Ben N'Cir, C., Cleuziou, G., & Essoussi, N. (2013). Identification of Non-Disjoint Clusters With Small and Parameterizable Overlaps. In *IEEE International Conference on Computer Applications Technology (ICCAT)* (pp. 1–6).

Ben N'Cir, C.-E., Cleuziou, G., & Nadia, E. (2015). Overview of overlapping partitional clustering methods. In M. E. Celebi (Ed.) *Partitional Clustering Algorithms* (pp. 245–275). Switzerland: Springer International Publishing.

Ben N'Cir, C., & Essoussi, N. (2012). Overlapping Patterns Recognition With Linear and Non-Linear Separations Using Positive Definite Kernels. In *International Journal of Computer Applications (IJCA)* (pp. 1–8).

Ben N'Cir, C., Essoussi, N., & Bertrand, P. (2010). Kernel Overlapping *K*-means for Clustering in Feature Space. In *International Conference on Knowledge Discovery and Information Retrieval (KDIR)* (pp. 250–256).

Bertrand, P., & Janowitz, M. (2003). The k-weak Hierarchical Representations: An Extension of the Indexed Closed Weak Hierarchies. *Discrete Applied Mathematics*, *127*(2), 199–220.

Bezdek, J. C. (1981). *Pattern Recognition With Fuzzy Objective Function Algorithm*. Plenum Press.

Bonchi, F., Gionis, A., & Ukkonen, A. (2013). Overlapping correlation clustering. *Knowledge and Information Systems*, *35*(1), 1–32.

Bozdogan, H. (2000). Akaike's information criterion and recent developments in information complexity. *Journal of Mathematical Psychology*, *44*, 62–91.

Bozdogan, H. (1987). Model selection and akaike information criterion (AIC) - the general theory and its analytical extensions. *Psychometrika*, *52*(3), 345–370.

Browne, M. W. (2000). Cross-Validation Methods. *Journal of Mathematical Psychology*, *44*(1), 108–132.
Brusco, M. J. (2004). Clustering binary data in the presence of masking variables. *Psychological Methods*, *9*(4), 510–523.
Bulteel, K., Wilderjans, T. F., Tuerlinckx, F., & Ceulemans, E. (2013). Chull as an alternative to AIC and BIC in the context of mixtures of factor analyzers. *Behavior Research Methods*, *45*, 782–791.
Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods and Research*, *33*(2), 261–304.
Caliński, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics - Theory and Methods*, *3*(1), 1–27.
Cattell, R. B. (1966). Scree Test for Number of Factors. *Multivariate Behavioral Research*, *1*(2), 245–276.
Ceulemans, E., & Kiers, H. A. L. (2009). Discriminating between strong and weak structures in Three-Mode principal component analysis. *British Journal of Mathematical & Statistical Psychology*, *62*, 601–620.
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.
Ceulemans, E., Van Mechelen, I., & Leenen, I. (2007). The Local Minima Problem in Hierarchical Classes Analysis: an Evaluation of a Simulated Annealing Algorithm and Various Multistart Procedures. *Psychometrika*, *72*, 377–391.
Chaturvedi, A., & Carroll, J. D. (1994). An alternating combinatorial optimization approach to fitting the INDCLUS and generalized INDCLUS models. *Journal of Classification*, *11*, 155–170.
Chen, Y., & Hu, H. (2006). An overlapping cluster algorithm to provide Non-Exhaustive clustering. *European Journal of Operational Research*, *173*, 762–780.
Claeskens, G., & Hjort, N. L. (2008). *Model Selection and Model Averaging*. Cambridge: Cambridge University Press.
Cleuziou, G. (2008). An Extended Version of the *K*-means Method for Overlapping Clustering. In *IEEE International Conference on Pattern Recognition (ICPR)* (pp. 1–4).
Cleuziou, G. (2013). Osom: a method for building overlapping topological maps. *Pattern Recognition Letters*, *34*(3), 239–246.
Cleuziou, G. (2009). Two Variants of the OKM for Overlapping Clustering. In *Advances in Knowledge Discovery and Management* (pp. 149–166).
Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*, 2nd edn. Hillsdale: Lawrence Earlbaum Associates.
Collins, L. M., & Dent, C. W. (1988). Omega: a general formulation of the rand index of cluster recovery suitable for Non-Disjoint solutions. *Multivariate Behavioral Research*, *23*(2), 231–242.
Cormack, R. M. (1971). A review of classification. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, *134*(3), 321–367.
Davis, G. B., & Carley, K. M. (2008). Clearing the fog: fuzzy, Overlapping Groups for Social Networks. *Social Networks*, *30*(3), 201–212.
Depril, D., Van Mechelen, I., & Mirkin, B. (2008). Algorithms for additive clustering of rectangular data tables. *Computational Statistics and Data Analysis*, *52*, 4923–4938.
Depril, D., Van Mechelen, I., & Wilderjans, T.F. (2012). Lowdimensional Additive Overlapping Clustering. *Journal of Classification*, *29*(3), 297–320.
Ding, Z., Zhang, X., Sun, D., & Luo, B. (2016). Overlapping Community Detection based on Network Decomposition. *Scientific Reports*, *6*, 24115.
Duda, R. O., Hart, P. E., & Stork, D.G. (2001). *Pattern Classification*. Hoboken: Wiley-Interscience.
Dunn, J. C. (1973). A fuzzy relative of the ISODATA process and its use in detecting compact Well-Separated clusters. *Journal of Cybernetics*, *3*(3), 32–57.
Dunn, J. C. (1974). Well-Separated Clusters and optimal fuzzy partitions. *Journal of Cybernetics*, *4*(1), 95–104.
Fellows, M. R., Guo, J., Komusiewicz, C., Niedermeier, R., & Uhlmann, J. (2011). Graph-Based Data clustering with overlaps. *Discrete Optimization*, *8*(1), 2–17.
Forgy, E. W. (1965). Cluster analyses of multivariate data: Efficiency versus interpretability of classifications. *Biometrics*, *21*, 768–769.
Fraley, C., & Raftery, A. E. (2002). Model-Based Clustering, discriminant analysis and density estimation. *Journal of the American Statistical Association*, *97*, 611–631.
France, S. L., & Abbassi, A. (2011). Boosting Unsupervised Additive Clustering using Cluster-Wise Optimization and Multi-Label Learning. In *Proceedings of the 2011 IEEE 11th International Conference on Data Mining Workshops (ICDM2011)* (pp. 236–243). Vancouver.
Fu, Q., & Banerjee, A. (2008). Multiplicative Mixture Models for Overlapping Clustering. In *Eight IEEE International Conference on Data Mining* (pp. 791–796).
Gil-Garcİa, R., & Pons-Porrata, A. (2010). Dynamic hierarchical algorithms for document clustering. *Pattern Recognition Letters*, *31*(6), 469–477.

Girvan, M., & Newman, M. E. J. (2002). Community structure in social and biological networks. *Proceedings of the National Academic Society of the United States of America*, *99*, 7821–7826.
Gordon, A. D. (1987). A review of hierarchical classification. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, *150*(2), 119–137.
Gordon, A. D., & De Soete, G. (1996). new york hierarchical classification. In P. Arabie, & L. J. Hubert (Eds.) *Clustering and classification* (pp. 119–137). NY: World Publishing.
Hamerly, G., & Elkan, C. (2004). Learning the k in *K*-means. In S. Thrun, L. Saul, & B. Schölkopf (Eds.) *Advances in neural information processing systems*, (Vol. 16 pp. 281–288). Cambridge: MIT Press.
Hannan, E. J., & Quinn, B. G. (1979). Determination of the Order of an Autoregression. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, *41*(2), 190–195.
Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS136: A *K*-means Clustering Algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *28*(1), 100–108.
Hastie, T., Tibshirani, R., Eisen, M. B., Alizadeh, A., Levy, R., Staudt, L., Chan, W. C., Botstein, D., & Brown, P. (2000). 'Gene Shaving' as a Method for Identifying Distinct Sets of Genes With Similar Expression Patterns. *Genome Biology*, *1*(2), 1–21.
Heller, K., & Ghahramani, Z. (2007). A Nonparametric Bayesian Approach to Modeling Overlapping Clusters. In *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics (PMLR)* (pp. 187–194).
Höppner, F., Klawonn, F., Kruse, R., & Runkler, T. (1999). *Fuzzy cluster analysis: Methods for classification*. Wiley: Data Analysis and Image Recognition.
Hruschka, E. R., Campello, R. J. G. B., Freitas, A. A., & De Carvalho, A.C.P.L.F (2009). A survey of evolutionary algorithms for clustering. In *IEEE Transactions on systems, man, and cybernetics, Part C (applications and reviews)*, (Vol. 39 pp. 133-155).
Huang, J. Z., Ng, M. K., Rong, H., & Li, Z. (2005). Automated Variable Weighting in *K*-means Type Clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27*(5), 657–688.
Hubert, L. J., & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, *2*(1), 193–218.
Hurvich, C. M., & Tsai, C. L. (1989). Regression and time series model selection in small samples. *Biometrika*, *76*(2), 297–307.
Jancey, R. C. (1966). Multidimensional group analysis. *Australian Journal of Botany*, *14*, 127–130.
Kanaya, S., Altaf-Ui-Amin, M., Kiboi, S. K., & Afendi, F.M. (2014). Big Data and Network Biology. *BioMed Research International, Article ID 836708*.
Khandekar, R., Kortsarz, G., & Mirrokni, V. (2012). On the advantage of overlapping clusters for minimizing conductance. *Algorithmica*, *69*, 844–863.
Khanmohammadi, S., Adibeig, N., & Shanehbandy, S. (2017). An Improved Overlapping *K*-means Clustering Method for Medical Applications. *Expert Systems with Applications*, *67*, 12–18.
Kohonen, T. (1995). *Self-Organizing Maps* Vol. 30. Berlin: Springer.
Krishnapuram, R., & Keller, J. M. (1993). A possibilistic approach to clustering. *IEEE Transactions on Fuzzy Systems*, *1*(2), 98–110.
Krishnapuram, R., & Keller, J. M. (1996). The Possibilistic C-means Algorithm: Insights and Recommendations. *IEEE Transactions on Fuzzy Systems*, *4*(3), 385–393.
Krzanowski, W. J., & Lai, Y. T. (1988). A criterion for determining the number of groups in a data set using sum of squares clustering. *Biometrics*, *44*(1), 23–34.
Lang, K. (1995). NewsWeeder: Learning To Filter Netnews. In *Proceedings of the Twelveth International Conference on Machine Learning (ICML-95)* (pp. 331–339). San Francisco.
Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent Structure Analysis*. Boston: Houghton Mill.
Lee, M. D. (2001). On the complexity of additive clustering models. *Journal of Mathematical Psychology*, *45*, 131–148.
Lingras, P., & West, C. (2004). Interval Set Clustering of Web Users With Rough *K*-means. *Journal of Intelligent Information System*, *23*(1), 5–16.
Liu, Z.-G., Dezert, J., Mercier, G., & Pan, Q. (2012). Belief c-means: An Extension of Fuzzy c-means Algorithm in Belief Functions Framework. *Pattern Recognition Letters*, *33*(3), 291–300.
Macqueen, J. B. (1967). Some methods for classification and analysis of multivariate observation. In L. M. Le Cam, & J. Neyman (Eds.) *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, (Vol. 1 pp. 281–297). Berkeley: University of California Press.
Manning, C. D., Raghavan, P., & SchüTze, H. (2008). *Introduction to Information Retrieval*. Cambridge: Cambridge University Press.
Masson, M.-H., & Denoeux, T. (2008). ECM: An Evidential Version Of the Fuzzy c-means Algorithm. *Pattern Recognition*, *41*(4), 1384–1397.
Mclachlan, G. J., & Chang, S. U. (2004). Mixture modelling for cluster analysis. *Statistical Methods in Medical Research*, *13*, 347–361.

Mclachlan, G. J., & Peel, D. (2000). *Finite Mixture Models*. New York: Wiley.
Milligan, G. W., & Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, *50*(2), 159–179.
Milligan, G. W., & Cooper, M. C. (1988). A study of standardization of variables in cluster analysis. *Journal of Classification*, *5*, 181–204.
Mirkin, B. G. (1987). The method of principal clusters. *Automation and Remote Control*, *10*, 131–143.
Mirkin, B. G. (1990). A sequential fitting procedure for linear data analysis models. *Journal of Classification*, *7*(2), 167–195.
Moore, E. H. (1920). On the reciprocal of the general algebraic matrix. *Bulletin of the American Mathematical Society*, *26*(9), 394–395.
Müller, H. G., & Stadtmüller, U. (2005). Generalized functional linear models. *Annals of Statistics*, *33*(2), 774–805.
Pal, K., Keller, J. M., & Bezdek, J. C. (2005). A Possibilistic Fuzzy c-means Clustering Algorithm. *IEEE Transactions of Fuzzy Systems*, *13*(4), 517–530.
Palla, G., DerÉnyi, I., Farkas, I., & Vicsek, T. (2008). Uncovering the Overlapping Community Structure of Complex Networks in Nature and Society. *Nature*, *435*(7043), 814–818.
PėRez-Suárez, A., MartİNez-Trinidad, J. F., Carrasco-Ochoa, J. A., & Medina-Pagola, J.E. (2013b). An algorithm based on density and compactness for dynamic overlapping clustering. *Pattern Recognition*, *46*(11), 3040–3055.
PėRez-Suárez, A., MartİNez-Trinidad, J. F., Carrasco-Ochoa, J. A., & Medina-Pagola, J.E. (2013a). Oclustr: A new Graph-Based Algorithm for Overlapping Clustering. *Neurocomputing*, *109*, 1–14.
Preacher, K. J., Zhang, G., Kim, C., & Mels, G. (2013). Choosing the optimal number of factors in exploratory factor analysis: a model selection perspective. *Multivariate Behavioral Research*, *48*(1), 28–56.
Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Computational and Applied Mathematics*, *20*, 53–65.
Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464.
Scott, J. (2000). *Social network analysis: a handbook*, 2nd edn. London: Sage.
Segal, E., Battle, A., & Koller, D. (2003). Decomposing Gene Expression into Cellular Processes. In *Proceedings of the Eigth Pacific Symposium on Biocomputing (PSB)* (pp. 89–100).
Shepard, R. N., & Arabie, P. (1979). Additive clustering: Representation of similarities as combinations of discrete overlapping properties. *Psychological Review*, *86*(2), 87–123.
Sneath, P. H., & Sokal, R. R. (1973). *Numerical taxonomy. The principles and practice of numerical classification*. San Francisco: W. H. Freeman Publishers.
Snoek, C. G. M., Worring, M., Van Gemert, J. C., Geusebroek, J.-M., & Smeulders, A.W.M. (2006). The Challenge Problem for Automated Detection of 101 Semantic Concepts in Multimedia. In *Fourteenth Annual ACM International Conference on Multimedia* (pp. 421–430).
Solka, J. L., Wegman, E. J., Priebe, C. E., & Poston, W. L. (1998). Mixture structure analysis using the akaike information criterion and the bootstrap. *Statistics and Computing*, *8*, 177–188.
Steele, R. J., & Raftery, A. E. (2009). *Performance of bayesian model selection criteria for gaussian mixture models, technical report 559, department of statistics*. WA: University of Washington Seattle.
Steinheus, H. (1956). Sur la division des corps matériels en parties. *Bulletin de l'Academie Polonaise des Sciences, Classe III*, *IV*(12), 801–804.
Steinley, D. (2003). Local optima in *K*-means clustering: what you don't know may hurt you. *Psychological Methods*, *8*(3), 294–304.
Steinley, D. (2006). Profiling local optima in *K*-means clustering: Developing a diagnostic technique. *Psychological Methods*, *11*(2), 178–192.
Steinley, D., & Brusco, M. J. (2011). Choosing the Number of Clusters in *K*-means Clustering. *Psychological Methods*, *16*(3), 285–297.
Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions (with discussion). *Journal of the Royal Statistical Society. Series B (Methodological)*, *36*(2), 111–147.
Sugar, C. A., & James, G. M. (2003). Finding the number of clusters in a dataset: an Information-Theoretic approach. *Journal of the American Statistical Association*, *98*(463), 750–763.
Tang, L., & Liu, H. (2009). Scalable learning of collective behavior based on sparse social dimensions. In *ACM Conference on Information and Knowledge Management* (pp. 1107–1116).
Tibshirani, R., Walther, G., Botstein, D., & Brown, P. (2001a). Cluster validation by prediction strength. *Journal of Computational and Graphical Statistics*, *14*(3), 511–528.
Tibshirani, R., Walther, G., & Hastie, T. (2001b). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, *63*(2), 411–423.
Van Mechelen, I., Bock, H.-H., & De Boeck, P. (2004). P Two-Mode clustering methods: a structured overview. *Statistical Methods in Medical Research*, *13*, 363–394.

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*(2), 228–243.

Wang, J. (2010). Consistent selection of the number of clusters via crossvalidation. *Biometrika*, *97*(4), 893–904.

Wang, Q., & Fleury, E. (2011). Uncovering overlapping community structure. *Complex Networks*, *116*, 176–186.

Wang, X., Tang, L., Gao, H., & Liu, H. (2010). Discovering Overlapping Groups in Social Media. In *IEEE International Conference on Data Mining* (pp. 569–578).

Wieczorkowska, A., Synak, P., & Ras, Z. (2006). Multi-label classification of emotions in music. In *Intelligent Information Processing and Web Mining, volume 35 of Advances in Soft Computing* (pp. 307–315).

Wilderjans, T. F., Ceulemans, E., & Meers, K. (2012). CHUll: A generic convex-hull-based model selection method. *Behavior Research Methods*, *45*(1), 1–15.

Wilderjans, T. F., Ceulemans, E., Van Mechelen, I., & Depril, D. (2011). ADPROCLUS: A graphical user interface for fitting additive profile clustering models to object by variable data matrices. *Behavior Research Methods*, *43*(1), 56–65.

Wilderjans, T. F., Depril, D., & Van Mechelen, I. (2013). Additive biclustering: a comparison of one new and two existing ALS algorithms. *Journal of Classification*, *30*(1), 56–74.

Xie, J., Kelley, S., & Szymanski, B. K. (2013). Overlapping community detection in networks: The state of the art and comparative study. *ACM computing surveys*, *45*(4), Article 43.

Zhang, S., Wang, R.-S., & Zhang, X. -S. (2007). Identification of overlapping community structure in complex networks using fuzzy c-means clustering. *Physica A: Statistical Mechanics and Its Applications*, *374*(1), 483–490.

Zhong, S., & Ghosh, J. (2003). A unified framework for model-based clustering. *Journal of Machine Learning Research*, *4*, 1001–1037.