Behav Res (2013) 45:1–15
DOI 10.3758/s13428-012-0238-5

CHull: A generic convex-hull-based model selection method

Tom F. Wilderjans · Eva Ceulemans · Kristof Meers

Published online: 6 October 2012

Abstract When analyzing data, researchers are often confronted with a model selection problem (e.g., determining the number of components/factors in principal components analysis [PCA]/factor analysis or identifying the most important predictors in a regression analysis). To tackle such a problem, researchers may apply some objective procedure, like parallel analysis in PCA/factor analysis or stepwise selection methods in regression analysis. A drawback of these procedures is that they can only be applied to the model selection problem at hand. An interesting alternative is the CHull model selection procedure, which was originally developed for multiway analysis (e.g., multimode partitioning). However, the key idea behind the CHull procedure—identifying a model that optimally balances model goodness of fit/misfit and model complexity—is quite generic. Therefore, the procedure may also be used when applying many other analysis techniques. The aim of this article is twofold. First, we demonstrate the wide applicability of the CHull method by showing how it can be used to solve various model selection problems in the context of PCA, reduced *K*-means, best-subset regression, and partial least squares regression. Moreover, a comparison of CHull with standard model selection methods for these problems is performed. Second, we present the CHULL software, which may be downloaded from http://ppw.kuleuven.be/okp/software/CHULL/, to assist the user in applying the CHull procedure.

Keywords Model selection · CHULL · Graphical user interface · PCA · Regression · PLS

T. F. Wilderjans (✉) · E. Ceulemans
Methodology of Educational Sciences Research Group,
Faculty of Psychology and Educational Sciences, KU Leuven,
Andreas Vesaliusstraat 2, Box 3762, 3000 Leuven, Belgium
e-mail: tom.wilderjans@ppw.kuleuven.be

E. Ceulemans
e-mail: eva.ceulemans@ppw.kuleuven.be

K. Meers
Research Group of Quantitative Psychology
and Individual Differences, Faculty of Psychology
and Educational Sciences, KU Leuven,
Tiensestraat 102,
Leuven 3000, Belgium
e-mail: kristof.meers@ppw.kuleuven.be

When analyzing data, researchers very often face a (complex) model selection problem. Take, as a first example, a clinical psychologist who wants to study the dimensionality of a particular psychological construct, such as alexithymia (i.e., having difficulties distinguishing between and expressing emotions). To this end, the researcher administers to a sample of subjects a questionnaire that measures the construct in question. Next, the researcher may examine the internal structure of the questionnaire and the dimensionality of the underlying construct by performing a principal components analysis (PCA) or an exploratory factor analysis (EFA) on the collected data. In doing so, the researcher needs to determine the optimal number of components or factors, and thus has to solve a model selection problem. As a second example, consider an economist who wants to assess which "factors" affect the selling price of a house in a particular neighborhood. To study this, the researcher may collect data (e.g., price and size of the house or the number of rooms) about a sample of houses from the neighborhood under study and may perform a regression analysis. In this case, again a model selection problem arises, which consists of selecting the (sub)set of predictors (optionally also including interactions between the predictors) that optimally predicts the selling price of a house.

In general, model selection boils down to selecting, out of a set of models, one that yields a good description of the data, in that it fits the data well, without being overly

complex (Pitt, Myung, & Zhang, 2002).[1] Sometimes, the different models are nested, in that one model may be obtained from another model by setting some of the parameters of the latter model to zero (e.g., when comparing a PCA model with two or three components), but often this is not the case (e.g., determining whether predictor A or B should be included in a regression model). One way to identify the most appropriate model is to visually inspect a scree-like plot that displays, for each model under consideration, a measure of model complexity against a measure of model fit (see e.g., Kroonenberg & Oort, 2003; Kroonenberg & Van der Voort, 1987)[2]; a well-known example of a model selection technique that is based on visual inspection is the scree test of Cattell (1966). Because the results of visual inspection are often very subjective, it may be wise to also use a more objective model selection procedure.

A few objective model selection procedures have already been proposed for several data-analytic problems. For example, in order to determine the number of components in PCA or (common) factors in EFA, different procedures exist, including the Kaiser–Guttman rule (Cliff, 1988; Guttman, 1954; Kaiser, 1960), parallel analysis (Horn, 1965; Peres-Neto, Jackson, & Somers, 2005), the scree test (Cattell, 1966), and the minimum average partial (MAP) test (Velicer, 1976). To select the optimal (sub)set of predictors in regression analysis, one may use, for instance, the *Cp* statistic (Mallows, 1973), forward selection, stepwise regression, and backward elimination (Draper & Smith, 1981; Hocking, 1976). A drawback of all of these objective procedures is that they must be tailor-made for the problem at hand and cannot be used for (or adapted to, in a straightforward way) other analysis problems (e.g., there is no equivalent for the Kaiser–Guttman rule in regression analysis). Therefore, it would be interesting to identify a more generic model selection strategy that is not limited to a specific technique.

In the case that the model selection problem only pertains to stochastic models (i.e., models with distributional assumptions about the noise in the data), generic methods based on information theory, such as the Akaike information criterion (AIC; Akaike, 1973, 1987; Bozdogan, 1987), the Bayesian information criterion (BIC; Schwarz, 1978), and minimum description length (MDL; Grunwald, 2000; Rissanen, 1983, 1996), may be adopted. However, these information-theoretic measures cannot be used in the case of deterministic models (i.e., models with no distributional assumptions regarding the noise in the data), such as dimension reduction methods (e.g., PCA, PARAFAC), scaling methods (e.g., multidimensional scaling), and clustering methods (e.g., *K*-means).

In this article, we claim that the CHull procedure, proposed by Ceulemans and Kiers (2006), is a very promising generic model selection heuristic, because it automates the very general idea behind the visual-inspection-based method that was described above. The CHull method has already been successfully applied for many techniques that imply a dimensional or categorical reduction of the data, including three-mode component analysis (Ceulemans & Kiers, 2006, 2009), multimode partitioning models (Schepers, Ceulemans, & Van Mechelen, 2008), hierarchical classes analysis (Ceulemans & Van Mechelen, 2005), multilevel component analysis (Ceulemans, Timmerman, & Kiers, 2011; Timmerman, 2006), and common-factor analysis (Lorenzo-Seva, Timmerman, & Kiers, 2011); extensive simulation results have shown that CHull outperforms other model selection methods that have been proposed for these techniques. The CHull procedure consists of (1) determining the convex hull of the fit-measure-by-complexity-measure plot of the models under consideration and (2) identifying the model on the boundary of the convex hull for which it is true that increasing the complexity (i.e., adding more parameters) has only a small effect on the fit measure, whereas lowering complexity (e.g., dropping parameters from the model) changes the goodness of fit (or, respectively, the misfit) substantially. CHull is very powerful because model goodness of fit/misfit and model complexity may be defined according to the analysis problem at hand. Besides being a generic strategy to identify one "optimal" model, the CHull method may also be used to determine a number of "good-quality" models that need to be investigated further. The goal of this article is twofold. First, we will show, by means of various examples that imply both rather easy and more complex model selection problems, the broad applicability of the CHull method. In particular, two component analysis illustrations (i.e., PCA and reduced *K*-means) and two regression examples (i.e., best-subset regression and partial least squares regression) will be discussed. Second, because up to now no software program has been publicly available to apply the CHull method in data-analytical practice, an implementation of the CHull method will be presented, together with a user-friendly program, called CHULL, that has been developed in MATLAB. For users not experienced with or without access to MATLAB, a standalone version of the CHULL program is provided.

[1] It should be noted that the complexity of a model not only depends on the number of parameters in the model (e.g., the number of components/factors or the number of regression coefficients), but also is influenced by the functional form (i.e., the way in which the model equation combines the model parameters with the data) of the model (see e.g., Pitt et al., 2002). However, when only comparing models that have the same functional form (i.e., determining the number of components/factors in PCA/EFA or best-subset regression), as is the case in this article, model complexity is a function of the number of parameters only.

[2] In the regression example, the fit measure may indicate the goodness of fit of the model (e.g., the amount of variance explained in the dependent variable by the predictors) or the badness of fit/misfit (e.g., the sum of squared residuals) of the model.

The remainder of this article is organized in two main sections. In the first section, the CHull method is presented and illustrated by applying it to component analysis (PCA), reduced *K*-means (RKM), best-subset regression, and partial least squares regression (PLS), which are four analysis techniques with which CHull had never been used before. In Section 2, we demonstrate how the CHULL program can be used to perform model selection by means of the CHull method. In Section 3, we provide some concluding remarks.

The CHull method

Method

To apply the CHull method, one needs to fit a set of models (i.e., the models one wants to incorporate in the comparison) and to compute for each model a corresponding goodness of fit/misfit value *f* (e.g., the sum of squared residuals or the amount of explained variance) and a complexity value *c*. Regarding the latter measure, different possibilities exist that may affect the solution that will be selected, including the number of (effective) parameters, the number of factors/components (in the context of factor analysis and component analysis), or the number of clusters (in the context of cluster analysis). Starting from a *c* and a *f* value for each of the models under consideration, the CHull procedure consists of seven steps. The first four steps serve to find the models on the upper (in the case of goodness of fit) or lower (in case of misfit) boundary of the convex hull, whereas in the last three steps the optimal model is identified by looking for solutions at which the boundary of the convex hull levels off. To this end, for each solution on the boundary of the convex hull (i.e., hull solutions), a scree-test value *st* is computed that indicates how much better a solution is in comparison to a less complex one, relative to how much worse a solution is in comparison with a more complex one; the solution with the largest *st* value should be selected. The seven steps of the CHull method are detailed below (for a more detailed description of the procedure, see Ceulemans & Kiers, 2006; Ceulemans & Van Mechelen, 2005):

1. For each level of complexity *c*, retain only the best model (i.e., the model with the largest goodness of fit or smallest misfit). When two (or more) models yield the same optimal goodness of fit/misfit, select one of these models at random.
2. Order the models m_i ($i = 1, \ldots, n$) that have been retained from the first step on the basis of their complexity value c_i, going from the most simple ($i = 1$) to the most complex ($i = n$).
3. Consider all pairs of adjacent models, and exclude a model m_i if $f_j \geq f_i$ (in the case of goodness of fit) or $f_j \leq f_i$ (in case of misfit) with $j < i$. Repeat this step until no more models can be excluded. The resulting f_i values increase (in the case of goodness of fit) or decrease (in the case of misfit) monotonically (in a strict sense).
4. For each triplet of adjacent models (m_i, m_j, m_k), exclude m_j if $f_j \leq f_i + \left(c_j - c_i\right)\frac{(f_k - f_i)}{(c_k - c_i)}$ (in the case of goodness of fit) or $f_j \geq f_i + \left(c_j - c_i\right)\frac{f_k - f_i}{c_k - c_i}$ (in the case of misfit), which implies that a model m_j is excluded when it is located on or under (in the case of goodness of fit) or above (in the case of misfit) the line that connects the other two models (i.e., m_i and m_k). Repeat this step until no more models can be excluded. The resulting models $\tilde{m}_i\left(i = 1, \ldots, \tilde{n}\right)$ are all located on the upper (in the case of goodness of fit) or the lower (in the case of misfit) boundary of the convex hull.
5. Compute for each model $\tilde{m}_i$ the following *st* value: $st_i = \frac{\frac{f_i - f_{i-1}}{c_i - c_{i-1}}}{\frac{f_{i+1} - f_i}{(c_{i+1} - c_i)}}$.
6. Retain the model $\tilde{m}_i$ that has the highest *st* value. In the case that two models have a maximal *st* value, retain the simplest one (i.e., smallest c_i value). As such, a model is obtained after which the increase in goodness of fit (or decrease in misfit) levels off. To see this, note that the numerator and denominator of the *st* ratio equal the slope of the line connecting two adjacent models and, consequently, that a large *st* value for $\tilde{m}_i$ implies that model $\tilde{m}_i$ fits the data considerably better than $\tilde{m}_{i-1}$ (i.e., a large numerator), whereas only a small improvement in goodness of fit/misfit is encountered when going from $\tilde{m}_i$ to $\tilde{m}_{i+1}$ (i.e., a small denominator).
7. If, in the first step of the CHull procedure, a model or models have been excluded that have the same c_i and f_i values as the retained model, then also retain this model/these models.

It should be noted that adding (very) complex models to the comparison may cause CHull to select a too-complex model. To see this, one should know that in most cases there is a degree of complexity, which often exceeds the complexity of the optimal model, after which adding more complexity does not result in large differences in fit. Consequently, for these models, both the numerator and the denominator of the *st* statistic become very small, so that the *st* value is artificially inflated, resulting in a too-complex model being selected by CHull. Two strategies exist to circumvent this problem. First, the researcher may decide in advance about an appropriate range of models to be included in the model comparison (for an example using common-factor analysis, see Lorenzo-Seva et al., 2011). Second, one may refine the CHull procedure by discarding models for which the numerator of the *st* ratio is "small" (for

an illustration in the context of the Diffit method, which is a scree-test-like model selection procedure for Tucker3 analysis, see Timmerman & Kiers, 2000). However, determining what is "small" is not a trivial problem. In this regard, we recommend adding an extra step, which should be included between the fourth and fifth step of the CHull procedure: Exclude models for which the fit (almost) equals the fit of a model that is less complex. In particular, a complex model is discarded from the boundary of the convex hull when its fit is less than 1 % better than the fit of a less complex model. Three further remarks about the CHull procedure are in order. First, in practice it may be wise not to focus exclusively on the model with the largest *st* value, but also to study models that yield a relatively large *st* value and/or models that are located close to the optimal one in the goodness of fit/misfit-versus-complexity plot (e.g., models with the same c_i value as the optimal one, but with a slightly different f_i value). Indeed, the f_i and c_i values are often calculated on the basis of a specific sample (drawn from a broader population), and consequently, these values may show some sampling variability (i.e., when analyzing data from another sample from the same population, slightly different f_i and c_i values may be obtained). Second, one should never select a final model on the basis of the CHull results only; instead, one should always take the interpretability of the model into consideration. Third, note that no *st* value is obtained for the first (most simple) and last (most complex) model on the boundary of the convex hull, which implies that these models cannot be selected by the CHull procedure (see the concluding remarks). Therefore, it is of utmost importance to carefully determine the set of models that one wants to choose among (see Lorenzo-Seva et al., 2011).

Principal components analysis

In this first example, we will illustrate the use of the CHull procedure in PCA by analyzing a data set consisting of the standardized (i.e., mean of zero and variance of one) scores of 133 subjects on the TAS-20 questionnaire (i.e., 20 items), which has been developed for measuring alexithymia. In the literature, ongoing debate has concerned whether alexithymia (i.e., difficulties in expressing or distinguishing emotions) is a two- or three-dimensional construct (see e.g., Erni, Lötscher, & Modestin, 1997; Haviland & Reise, 1996; Pandey, Mandal, Taylor, & Parker, 1996; Parker, Bagby, Taylor, Endler, & Schmitz, 1993). To resolve this issue, one may study the internal structure of the TAS-20 questionnaire (Bagby, Parker, & Taylor, 1994). More specifically, one may perform PCA and determine the number of components that are needed to adequately describe the TAS-20 data.

To this end, PCA was performed on the data, with the number of components ranging from one to 20. The number of extracted components was considered as the complexity measure, whereas the fit measure equaled the total amount of variance explained by the extracted components (i.e., a measure of goodness of fit, with a higher amount of explained variance denoting a better model). In Table 1, for each model (first column), the number of extracted components (second column), the amount of explained variance (third column), and the *st* value (fourth column) are displayed. In the upper panel of Fig. 1, which displays the CHull plot, one can see that all models are located on the upper boundary of the convex hull. This will always be the case in PCA (and also in common—principal axis—factor analysis), because components/factors are extracted in such a way that they explain a decreasing amount of variance (i.e., the first component/factor always explains the most variance, the second one the second most, etc.). Therefore, when the total amount of variance explained is plotted against the number of components/factors (i.e., the CHull plot), all models will be located on a strictly increasing and decelerating function. As a consequence, for each triplet of adjacent models (m_i, m_j, m_k), the middle model m_j will always be located above the line connecting m_i and m_k, which implies that all models will be located on the upper boundary of the convex hull.[3] In the upper panel of Fig. 1, one can further see that the CHull procedure (adopting the 1 % rule; see section Method) suggests retaining a model with two components. Note that the selected solution is indicated with a green circle and that the associated (maximal) *st* value (i.e., 1.574) can be found in Table 1. This result supports the researchers who defend the bidimensionality of the alexithymia construct, as measured by the TAS-20 questionnaire (see Erni et al., 1997; Haviland & Reise, 1996).

When comparing the results of the CHull method with standard methods for determining the number of components, it appears that mixed results are obtained. In particular, the Kaiser–Guttman rule (Cliff, 1988; Guttman, 1954; Kaiser, 1960), which selects all components with an eigenvalue greater than one, retains a model with six components, which is a too-complex model. When using a permutation method, like parallel analysis (Horn, 1965; Peres-Neto, Jackson, & Somers, 2005), the number of selected components depends on the cutoff values used: For example, when comparing the observed eigenvalues with the median or the last quartile of the distribution of eigenvalues for the permuted data (see Lorenzo-Seva et al., 2011), a model with four components is retained; when the 95th or 99th percentile is taken, a model with two components is selected. The scree test (Cattell, 1966), which boils down to identifying

[3] Note that in only a few, very exceptional cases, which will almost never be encountered in practice, this may not be true (i.e., when two components/factors explain exactly equal amounts of variance and/or when one component/factor explains no variance at all).

Table 1 Numbers of components extracted, amounts of explained variance, and *st* values of the 20 models on the upper boundary of the convex hull for the TAS-20 data

Model	Complexity (Number of Components)	Fit (Amount of Explained Variance)	*st* Value
M_1	1	.2363	—*
M_2	2	.3606	1.574
M_3	3	.4396	1.092
M_4	4	.5119	1.153
M_5	5	.5747	1.179
M_6	6	.6279	1.187
M_7	7	.6728	1.062
M_8	8	.7150	1.125
M_9	9	.7525	1.082
M_{10}	10	.7873	1.047
M_{11}	11	.8204	1.236
M_{12}	12	.8472	1.039
M_{13}	13	.8731	1.074
M_{14}	14	.8971	1.069
M_{15}	15	.9196	1.067
M_{16}	16	.9407	1.071
M_{17}	17	.9604	1.261
M_{18}	18	.9759	1.135
M_{19}	19	.9897	1.333
M_{20}	20	1	—*

M_1 = Model 1, M_2 = Model 2, etc. As noted, all models are located on the (upper) boundary of the convex hull (adopting the 1 % rule); the model retained by CHull is underlined. * For the first and last hull models, the *st* value is not defined

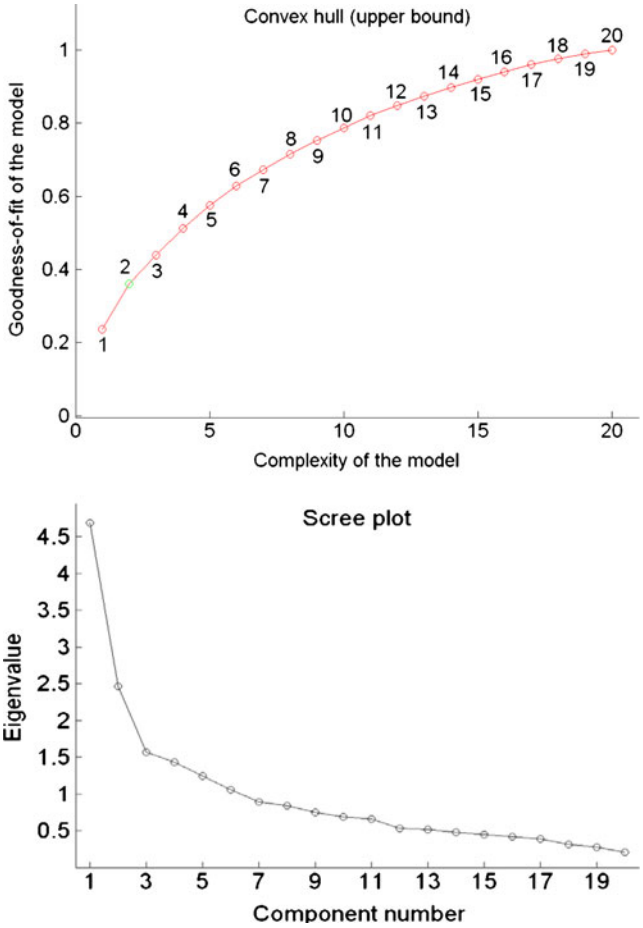

Fig. 1 (Upper panel) Graphical representation of the numbers of extracted components (i.e., complexity measure) versus the amount of explained variance (i.e., goodness of fit measure) for the considered component models for the TAS-20 data. (Lower panel) Scree plot (i.e., successive eigenvalues against their rank order) for the TAS-20 data

the elbow in the scree plot (i.e., the plot of the successive eigenvalues against their rank order; see the lower panel of Fig. 1) and retaining the model before the elbow, suggests a model with two components. When adopting the MAP test (Velicer, 1976), which takes into account the partial correlation matrices, a model with two components is selected. Finally, the broken-stick method (Frontier, 1976; Legendre & Legendre, 1983) retains a model with only one component.

Reduced K-means

A second illustration pertains to RKM (De Soete & Carroll, 1994; Timmerman, Ceulemans, Kiers, & Vichi, 2010), also called *projection pursuit clustering* (Bock, 1987). In this model, the variables are reduced to a limited set of components, and simultaneously the objects are clustered on the basis of the scores of the objects on these components. As such, each obtained cluster has an associated component profile (i.e., cluster component scores)—which, when there are many variables, is much easier to interpret than the associated variable profile. The model selection problem for RKM is more complex than that for PCA because one has (simultaneously) to decide on both the number of clusters K and the number of components Q.

To illustrate the usefulness of CHull, we will analyze the well-known iris data set. In order to study the morphologic variation among three (related) species of iris flowers (i.e., *setosa*, *versicolor*, and *virginica*), Anderson (1935, 1936) measured 150 iris flowers (i.e., 50 samples from each of the three species of iris flowers) on four variables (i.e., the length and the width of the sepal and the petal). RKM models were fitted to the centered data, with K ranging from two to six and with Q varying, for each particular K value, from one to K minus one. The latter restriction was imposed because, due to the centering, the K cluster-specific variable profiles can be perfectly modeled with $K - 1$ components, implying that adding more components would not increase the fit. As a fit measure, the sum of squared residuals, which was minimized by the algorithm, was adopted. The complexity measure equaled the number of fitted parameters,

where the parameter set consisted of the 150 × K entries of the partition matrix that indicated to which cluster each flower belonged, the 4 × Q component loadings, and the K × Q cluster component scores. To account for the rotational freedom of the components, Q^2 parameters were subtracted from this total number of parameters (see Ceulemans, Timmerman, & Kiers, 2011).

Table 2 displays the numbers of clusters K, the numbers of components Q, and the complexity and fit values for all of the RKM models under consideration; moreover, it indicates which of these models are located on the lower boundary of the convex hull (i.e., the hull models are indicated in italic), and the associated *st* values are given. From Table 2, one can conclude that CHull (adopting the 1 % rule; see section Method) selects the model with three clusters and two components (i.e., Model 3 has the greatest *st* value in Table 2).

The clusters of the retained model contain 50, 62, and 38 flower samples, respectively. When comparing these clusters to the true clustering (i.e., the three types of species), it appears that the first cluster perfectly corresponds with the setosa type, the second cluster consists of 48 flowers of the versicolor type and 16 of the virginica type, and the third cluster encompasses 36 flowers of the virginica type and 2 of the versicolor type. In sum, the true clustering is recovered quite well, except that some virginica flowers are erroneously classified as versicolor flowers. From Table 3, which presents the component loadings (left part) and the cluster component scores (right part), one can see that the first component has a very high loading for petal length and relatively high loadings for sepal length and petal width, whereas the second component pertains to the sepal characteristics. When inspecting the cluster component scores, it appears that the second cluster (i.e., versicolor and a bit of virginica) clearly differs from the other two clusters on both components, whereas the first (i.e., setosa) and third (i.e., virginica) clusters only can be discriminated on the basis of the first component. Note that these results nicely resonate with the discriminant analysis that Fisher (1936) applied to these data.

For the RKM model, no standard model selection procedure exists to simultaneously determine K and Q. However, the Calinski–Harabasz (CH) index (Calinski & Harabasz, 1974), also called the *pseudo-F* statistic, has already been successfully applied in other clustering techniques (e.g., K-means analysis; Milligan & Cooper, 1985) and can easily be adapted to the RKM case. Specifically, for each model under consideration, one may compute for each cluster the predicted variable profile x_k ($k = 1..K$) by multiplying the cluster component scores by the component loadings. Subsequently, the CH index is computed as follows:

$$CH = \frac{\frac{trace(X_B)}{K-1}}{\frac{trace(X_W)}{N-K}}, \quad \text{with} \quad X_B = \sum_{k=1}^{K} n_k (x_k - \bar{x})(x_k - \bar{x})',$$

$$X_W = \sum_{k=1}^{K} \sum_{j=1}^{n_k} (x_{kj} - x_k)(x_{kj} - x_k)',$$

$\bar{x}$ as the mean variable profile computed over all objects (i.e., a zero vector, in the case of centered/standardized data),

Table 2 Numbers of clusters, numbers of components, complexity, (mis)fit (i.e., sum of squared residuals) values, *st* values (only for hull models), and Calinski–Harabasz (CH) values for all fitted reduced K-means models for the iris data

Model	Number of Clusters (K)	Number of Components (Q)	Complexity	Misfit (Sum of Squared Residuals)	*st* Value	CH Value
M_1	*2*	*1*	*305*	*152.348*	—*	*513.92*
M_2	3	1	456	88.742		490.84
M_3	*3*	*2*	*460*	*78.851*	*3.34*	*561.63*
M_4	4	1	607	70.705		420.31
M_5	*4*	*2*	*612*	*57.266*	*2.05*	*530.38*
M_6	4	3	615	57.229		530.77
M_7	5	1	758	63.656		351.77
M_8	5	2	764	48.683		471.11
M_9	*5*	*3*	*768*	*46.446*	*1.43*	*495.54*
M_{10}	5	4	770	46.446		495.54
M_{11}	6	1	909	60.005		298.23
M_{12}	6	2	916	41.307		446.26
M_{13}	*6*	*3*	*921*	*39.041*	—*	*473.83*
M_{14}	6	4	924	39.040		473.85

M_1 = Model 1, M_2 = Model 2, etc. The models that are located on the (lower) boundary of the convex hull (adopting the 1 % rule) are indicated in italics; the model retained by CHull is underlined. * For the first and last hull models, the *st* value is not defined

Table 3 Component loadings and cluster component scores for the retained reduced *K*-means model with three clusters and two components for the iris data

	Component Loadings			Cluster Component Scores	
	Component 1	Component 2		Component 1	Component 2
Sepal length	.34	.59	Cluster 1	–2.65	0.13
Sepal width	–.10	.79	Cluster 2	0.68	–0.30
Petal length	.86	–.16	Cluster 3	2.37	0.33
Petal width	.36	.04			

x_{kj} as the observed variable profile for the *j*th object in the *k*th cluster, N as the (total) number of objects (i.e., the number of iris flowers), and n_k as the number of objects for the *k*th cluster. On the basis of the CH values, which are listed in Table 2, a model with three clusters and two components would be retained, as this model has the maximal CH value.

Best-subset regression

This example will demonstrate how the CHull procedure can be used to select an appropriate regression model—that is, to determine which predictors should be retained in a regression analysis. To this end, different multiple regression analyses were performed on a data set (Agresti & Finlay, 2008) containing the prices (in thousands of dollars) of 93 houses in Gainesville, Florida, in January 1996, together with information regarding the sizes of the houses (in thousands of square feet), the numbers of bedrooms, the numbers of bathrooms, the amounts of annual taxes, and whether a house was newly built.

After standardizing the data, all possible regression models (without an intercept, because of the standardization) with one (five models in total), two (ten models), three (ten models), four (five models), and five (one model) predictors were estimated. The number of estimated parameters (i.e., the number of estimated regression coefficients) was adopted as a complexity measure, and the sum of the squared residuals was taken as a badness of fit measure. In Table 4, for all 31 models under consideration, the included predictors, the complexity (i.e., the number of predictors), and the (mis)fit (i.e., sum of squared residuals) values are displayed. When applying the CHull procedure (adopting the 1 % rule; see section Method) to these 31 models, four models appeared to be located on the lower boundary of the convex hull; note that the hull plot is displayed in Fig. 2 as part of a screenshot using the CHULL software that will be discussed later (see section Program Handling). From Table 4, which also indicates the hull models (in italics) and the associated *st* values, one can read that the CHull procedure selected the model with three predictors.

On the basis of the estimated regression coefficients, one may conclude that the variability in house-selling prices in Gainesville is mainly explained by differences in the amount of annual taxation ($\beta = .45$) and in the size of the house ($\beta = .41$), and to a lesser extent by differences in the age (i.e., new or not) of the house ($\beta = .14$). The number of bedrooms and the number of bathrooms, which are not included in the model, seem to be of no importance at all.

When analyzing these data, Agresti and Finlay (2008) came to the same conclusion regarding the importance of the different factors affecting the house-selling price. Moreover, the model with three predictors was selected by various standard model selection procedures in the context of regression analysis, including forward selection, backward elimination, and stepwise methods. Furthermore, as one can see in Table 5, this model with three predictors has the lowest values for the AIC, BIC, and *Cp* statistics. However, as one can also see in Table 5, a model with two predictors (i.e., the amount of annual taxation and the size of the house) has the lowest (tenfold) cross-validation error, which is also true when adopting the 10 % rule[4] (see Table 5). Note that these two predictors have the largest regression weights in the model retained by CHull, and that the two-predictor model has the second largest *st* value (see Table 4).

Partial least squares regression

In this final example, the applicability of CHull to PLS regression (Wold, 1966) will be illustrated by analyzing a data set in which the level of prostate-specific antigen, which is related to prostate cancer, is predicted by a set of clinical measures, such as the weight of the prostate (Stamey et al., 1989; Hastie, Tibshirani, & Friedman, 2009, used the same data set for illustrating different regression techniques). PLS, which is often used to overcome collinearity problems in regression analysis, bears characteristics from

[4] Because the observed cross-validation error for a model is influenced by sampling error, Hastie et al. (2009) advised researchers to select the least complex model that has a cross-validation error that is less than 10 % larger than the lowest cross-validation error.

Table 4 Included predictors, complexity (i.e., number of predictors), (mis)fit (i.e., sum of squared residuals) values, and *st* values (only for hull models) for all best-subset regression models under consideration for the house-selling price data

Model	Included Predictors	Complexity (Number of Predictors)	Misfit (Sum of Squared Residuals)	*st* Value
M_1	*P_1*	*1*	*28.8159*	—*
M_2	P_2	1	83.6350	
M_3	P_3	1	68.1470	
M_4	P_4	1	76.8264	
M_5	P_5	1	30.1755	
M_6	P_4, P_5	2	27.4630	
M_7	P_3, P_5	2	30.1599	
M_8	P_3, P_4	2	56.4369	
M_9	P_2, P_5	2	29.6640	
M_{10}	P_2, P_4	2	63.1948	
M_{11}	P_2, P_3	2	66.2914	
M_{12}	*P_1, P_5*	*2*	*22.5545*	*3.63*
M_{13}	P_1, P_4	2	26.1203	
M_{14}	P_1, P_3	2	28.3111	
M_{15}	P_1, P_2	2	28.8126	
M_{16}	P_3, P_4, P_5	3	27.4465	
M_{17}	P_2, P_4, P_5	3	27.3163	
M_{18}	P_2, P_3, P_5	3	29.5842	
M_{19}	P_2, P_3, P_4	3	53.6440	
M_{20}	*P_1, P_4, P_5*	*3*	*20.8319*	*4.67*
M_{21}	P_1, P_3, P_5	3	22.4996	
M_{22}	P_1, P_3, P_4	3	25.6902	
M_{23}	P_1, P_2, P_5	3	21.7864	
M_{24}	P_1, P_2, P_4	3	26.0762	
M_{25}	P_1, P_2, P_3	3	28.2320	
M_{26}	P_1, P_2, P_3, P_4	4	25.6901	
M_{27}	P_1, P_2, P_3, P_5	4	21.7834	
M_{28}	*P_1, P_2, P_4, P_5*	*4*	*20.4632*	—*
M_{29}	P_1, P_3, P_4, P_5	4	20.7878	
M_{30}	P_2, P_3, P_4, P_5	4	27.2697	
M_{31}	P_1, P_2, P_3, P_4, P_5	5	20.4558	

M_1 = Model 1, M_2 = Model 2, etc.; P_1 = annual taxes, P_2 = number of bedrooms, P_3 = number of bathrooms, P_4 = new or old house, P_5 = size of the house. The models that are located on the (lower) boundary of the convex hull (adopting the 1 % rule) are indicated in italics; the model retained by CHull is underlined. * For the first and last hull models, the *st* value is not defined

PCA and regression, as it jointly summarizes the predictor variables in a few components and regresses the criterion variable on these components. The model selection issue in PLS thus pertains to determining the optimal number of PLS components.

We analyzed the prostate cancer data, with the number of PLS components varying from one to eight. Note that, like Hastie et al. (2009), we only used the training data set (i.e., 67 observations were picked at random from the full, autoscaled data set). The number of PLS components was adopted as the complexity measure, and the sum of squared residuals for the criterion as the fit measure. From Table 6, which presents the complexity (i.e., number of PLS components) and fit (i.e., sum of squared residuals) values for all considered models and the *st* values of all hull models (indicated in italic), we conclude that the CHull procedure (adopting the 1 % rule; see section Method) suggests retaining the solution with two PLS components. This model indicates that the volume of the cancer ($\beta = .44$), the weight of the prostate ($\beta = .35$), the size of the seminal vesicle invasion ($\beta = .25$), and the amount of benign prostatic hyperplasia ($\beta = .25$) are the most important predictors for prostate-specific antigen. Note that, on the basis of tenfold cross-validation (adopting the 10 % rule; see section Best-subset regression), Hastie et al. (2009) also selected the model with two PLS components.

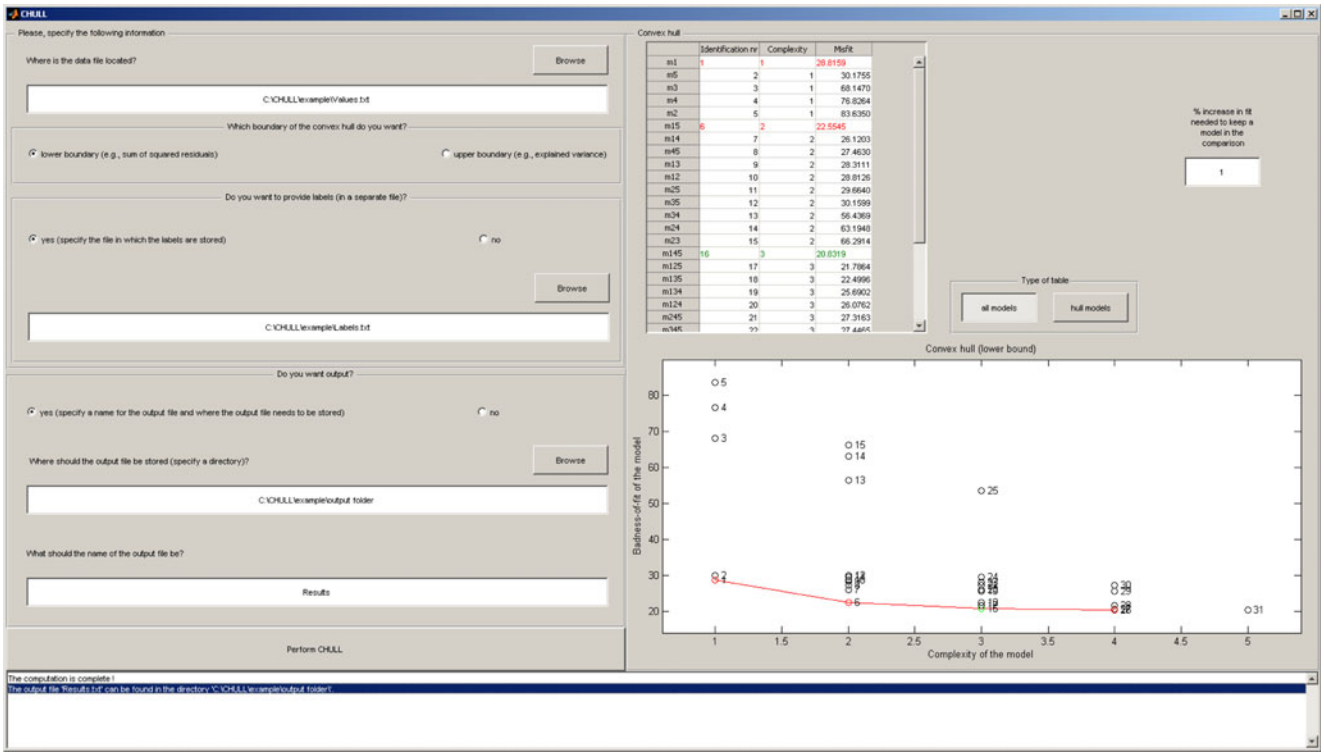

Fig. 2 Screenshot of the graphical user interface of the CHULL program

We also performed a best-subset regression analysis to the (training) data. From Table 7, in which for all hull solutions the included predictors, associated complexity, fit, and *st* values are presented, it appears that CHull (again adopting the 1 % rule) retains a model with two predictors (i.e., cancer volume and prostate weight). When applying best-subset regression to these data, Hastie et al. (2009) selected the same model on the basis of tenfold cross-validation (adopting the 10 % rule; see section Best-subset regression). Note that the two selected predictors are the ones with the largest regression weights in the PLS model with two PLS components that was retained by CHull. When considering the standard model selection methods for regression, it appears that only BIC, forward selection, and stepwise selection retain the model with two predictors, whereas the other methods select a more complex model. In particular, AIC, backward elimination, and *Cp* retain models with two, four, and five extra predictors, respectively.

The CHULL software

Program handling

Two versions of the CHULL software are available. One is a standalone application that can be run on any Windows computer and does not require any MATLAB knowledge or license. After downloading the CHULL application from http://ppw.kuleuven.be/okp/software/CHULL/ and installing it (see the instructions file "ReadMe_Standalone.txt" on the same website), this standalone version can be launched from the Start menu. The other version consists of a graphical user interface (GUI) that is built around the MATLAB functions (m-files) for computing the boundary of the convex hull and locating the elbow in this boundary, and which is used within the MATLAB environment (version R2011b). Specifically, after downloading and storing all necessary files in the same folder and ensuring that the current MATLAB directory is set to this folder, the software can be launched in MATLAB by typing CHULL at the command prompt[5]:

>>CHULL <ENTER>

The GUI of the CHULL software, which is displayed in Fig. 2, consists of three main compartments (i.e., one compartment for providing information regarding the file for the data/labels, one for providing information regarding the output file, and one in which the data and the analysis results will be plotted and displayed in a table). Note that in Fig. 2, the contents of the boxes of the GUI pertain to the guiding example, which is the best-subset regression model selection problem discussed in section Best-subset regression. To

[5] While performing the CHULL analysis with the CHULL software, it is important that the current MATLAB directory is not be changed (and thus that it be the folder where the .m-files are stored).

Table 5 Included predictors, Akaike information criterion (AIC) values, Bayesian information criterion (BIC) values, *Cp* values, and cross-validation error for all best-subset regression models under consideration for the house-selling price data

Model	Included Predictors	AIC	BIC	*Cp* Value	Cross-Validation Error*
M_1	P_1	–121.42	–118.81	35.83	3.0336
M_2	P_2	–14.87	–12.26	290.41	8.8647
M_3	P_3	–35.35	–32.74	218.49	7.1321
M_4	P_4	–23.36	–20.75	258.79	8.0471
M_5	P_5	–116.81	–114.20	42.14	3.3878
M_6	P_4, P_5	–123.21	–118.00	31.54	3.1928
M_7	P_3, P_5	–113.85	–108.64	44.07	3.4675
M_8	P_3, P_4	–51.18	–45.97	166.10	6.6781
M_9	P_2, P_5	–115.50	–110.29	41.76	3.4260
M_{10}	P_2, P_4	–39.87	–34.66	197.49	6.9341
M_{11}	P_2, P_3	–35.09	–29.88	211.87	7.0879
M_{12}	P_1, P_5	–142.90	–137.69	8.75	2.3704
M_{13}	P_1, P_4	–128.23	–123.02	25.31	2.9224
M_{14}	P_1, P_3	–120.17	–114.96	35.48	2.9942
M_{15}	P_1, P_2	–118.42	–113.21	37.81	3.1561
M_{16}	P_3, P_4, P_5	–120.25	–112.43	33.47	3.1254
M_{17}	P_2, P_4, P_5	–120.72	–112.91	32.86	3.6369
M_{18}	P_2, P_3, P_5	–112.75	–104.93	43.39	3.8604
M_{19}	P_2, P_3, P_4	–53.23	–45.42	155.13	6.1137
M_{20}	P_1, P_4, P_5	–147.82	–140.01	2.75	2.5375
M_{21}	P_1, P_3, P_5	–140.12	–132.31	10.49	2.6978
M_{22}	P_1, P_3, P_4	–126.86	–119.04	25.31	3.1943
M_{23}	P_1, P_2, P_5	–143.34	–135.53	7.18	2.9720
M_{24}	P_1, P_2, P_4	–125.37	–117.55	27.10	3.2000
M_{25}	P_1, P_2, P_3	–117.43	–109.61	37.11	3.0528
M_{26}	P_1, P_2, P_3, P_4	–123.82	–113.40	27.31	3.1254
M_{27}	P_1, P_2, P_3, P_5	–140.32	–129.90	9.17	2.7464
M_{28}	P_1, P_2, P_4, P_5	–146.57	–136.15	3.03	2.6671
M_{29}	P_1, P_3, P_4, P_5	–145.00	–134.58	4.54	2.8725
M_{30}	P_2, P_3, P_4, P_5	–117.86	–107.44	34.64	3.2479
M_{31}	P_1, P_2, P_3, P_4, P_5	–143.56	–130.54	5.00	2.9430

M_1 = Model 1, M_2 = Model 2, etc.; P_1 = annual taxes, P_2 = number of bedrooms, P_3 = number of bathrooms, P_4 = new or old house, P_5 = size of the house. The models retained by the different model selection strategies are underlined. * Adopting the 10 % rule

apply the CHull procedure to a model selection problem at hand, the user needs to specify the necessary information in the different compartments and to click the "Perform CHULL" button. In the following sections, the functionality of the GUI (i.e., how to input data, apply the procedure, ask for an output file, etc.) will be discussed.

Information regarding the data, labels, output file, and analysis options

File with complexity and (mis)fit values The user has to specify the file that contains the complexity and (mis)fit measures for all considered models by clicking the "Browse" button at the top of the "Please, specify the following information" compartment and selecting the file in question (e.g., in Fig. 2, "Values.txt"). This file should be an ASCII file (i.e., a .txt file) that contains as many rows as there are models under consideration, with each row having a complexity (first column) and a goodness of fit/misfit (second column) value for the corresponding model (the file "Values.txt" is displayed in the left-hand panel of Fig. 3); in the ASCII file, empty lines are allowed. Within each row, the complexity and fit values may be separated by one or more spaces, commas, semicolons, tabs, or any combination of these. The complexity and fit values should be integers or real numbers, with decimal separators being denoted by a

Table 6 Complexity (i.e., number of partial least squares [PLS] components), (mis)fit (i.e., sum of squared residuals) values, and *st* values (only for hull models) for all fitted PLS models for the prostate cancer data

Model	Complexity (Number of PLS Components)	Misfit (Sum of Squared Residuals)	*st* Value
M_1	*1*	*.6346*	—*
M_2	*2*	*.5030*	*3.43*
M_3	*3*	*.4645*	*1.80*
M_4	*4*	*.4433*	—*
M_5	5	.4400	
M_6	6	.4392	
M_7	7	.4392	
M_8	8	.4392	

M_1 = Model 1, M_2 = Model 2, etc. The models that are located on the (lower) boundary of the convex hull (adopting the 1 % rule) are indicated in italics; the model retained by the CHull procedure is underlined. * For the first and last hull models, the *st* value is not defined

period and not by a comma. Obviously, the CHULL program cannot handle missing values (i.e., models for which no complexity or goodness of fit/misfit value is provided cannot be considered).

Goodness of fit or misfit The user needs to indicate whether the fit measure pertains to goodness of fit (e.g., the amount of explained variance) or badness of fit/misfit (e.g., sum of the squared residuals) by checking the appropriate option in the "Which boundary of the convex hull do you want?" part of the "Please, specify the following information" compartment: in the case of badness of fit, "lower boundary", and in the case of goodness of fit, "upper boundary". Because in the regression example the loss function pertains to badness of fit (i.e., sum of squared residuals), the "lower boundary" option is selected in Fig. 2. Furthermore, the user also needs to specify the percentage of increase in fit that a more complex model should have, as compared to a less complex model, so as to keep the complex model in the comparison (see section Method). To this end, the user needs to specify the required percentage in the box below "% increase in fit needed to keep a model in the comparison", to the right in the "Convex hull" compartment. Because for the best-subset regression example the 1 % rule is adopted, a "1" is entered in this box (see Fig. 2).

Label file Optionally, the user may also provide labels for all models under study. This can be achieved by checking the "yes" box in the "Do you want to provide labels (in a separate file)?" part of the "Please, specify the following information" compartment and by browsing for the file that contains the labels (e.g., in Fig. 2, "Labels.txt"). This file should also be an ASCII file (.txt format). Each line of text (and/or symbols) will be considered as a separate label, where the different text lines may be separated by one or more empty lines. Obviously, the number and order of the labels has to correspond to the number and order of the models in the file with the complexity and (mis)fit values. In the right-hand panel of Fig. 3, the label file ("Labels.txt") for the guiding example is shown. If the "no" option is selected (no labels), the CHULL program generates default labels (e.g., "model1", "model2", etc.).

Output file Optionally, the user may also want to store the results of the CHull procedure in an output file (.mht). To this end, one may select the "yes" option in the "Do you want output?" panel, specify the directory where this output file should be stored by means of the "Browse" button, and provide a name for this output file, which should be typed into the corresponding box (without the extension, such as .mht). For the guiding example, one can see in Fig. 2 that the output file will be stored in the directory "C:\CHULL \example\output folder" and that this file will be named "Results".

Table 7 Included predictors, complexity (i.e., number of predictors), (mis)fit (i.e., sum of squared residuals) values, and *st* values for all best-subset regression models that are located on the lower boundary of the convex hull (adopting the 1 % rule) for the prostate cancer data

Model	Included Predictors	Complexity (Number of Predictors)	Misfit (Sum of Squared Residuals)	*st* Value
M_1	P_1	1	30.524	—*
M_9	P_1, P_2	2	25.426	3.40
M_{39}	P_1, P_2, P_3	3	23.929	1.04
M_{57}	P_5, P_6, P_7, P_8	4	22.494	1.84
M_{235}	P_3, P_4, P_5, P_6, P_7, P_8	6	20.935	1.03
M_{248}	P_1, P_2, P_3, P_4, P_5, P_6, P_7	7	20.179	—*

M_1 = Model 1, M_9 = Model 9, etc.; P_1 = cancer volume, P_2 = prostate weight, P_3 = age, P_4 = amount of benign prostatic hyperplasia, P_5 = seminal vesicle invasion, P_6 = capsular penetration, P_7 = Gleason score, P_8 = percentage of Gleason scores 4 or 5. The model retained by the CHull procedure is underlined. * For the first and last hull models, the *st* value is not defined

Checking the analysis results in the CHull plot and in a table

After providing the necessary information (see the previous section), the user clicks the "Perform CHULL" button. The program will immediately generate a scree-like plot of the complexity and fit values of the different models. The models that are not located on the appropriate boundary of the convex hull are indicated by black circles, and the "hull" models by red circles. The boundary itself is indicated by a red line, and the model retained by CHull is marked with a green circle. The models, which are indicated in the plot by "1", "2", and so forth (i.e., the identification number), are numbered as follows: A model receives a higher number when it is more complex than another model; when some of the models are equally complex, increasing numbers indicate increasing fit values. Note that this numbering of the models does not necessarily follow the order of the models provided by the user in the file with the complexity and (mis)fit values (i.e., compare with the left-hand panel of Fig. 3). Furthermore, a table is displayed in which all of the models under study are listed by their specified label, together with their identification number, which is also used in the CHull plot (i.e., "1", "2", etc.), and their complexity and (mis)fit values; in this table, the different models are indicated by the same colors as in the plot (i.e., black, red, and green). When the "hull models" button is clicked, the former table is replaced by another table in which only the "hull" models are listed (again, using the same coloring of the different models as before and the specified labels), along with their identification numbers and complexity, (mis)fit, and *st* values. The user may return to the original table by clicking the "all models" button.

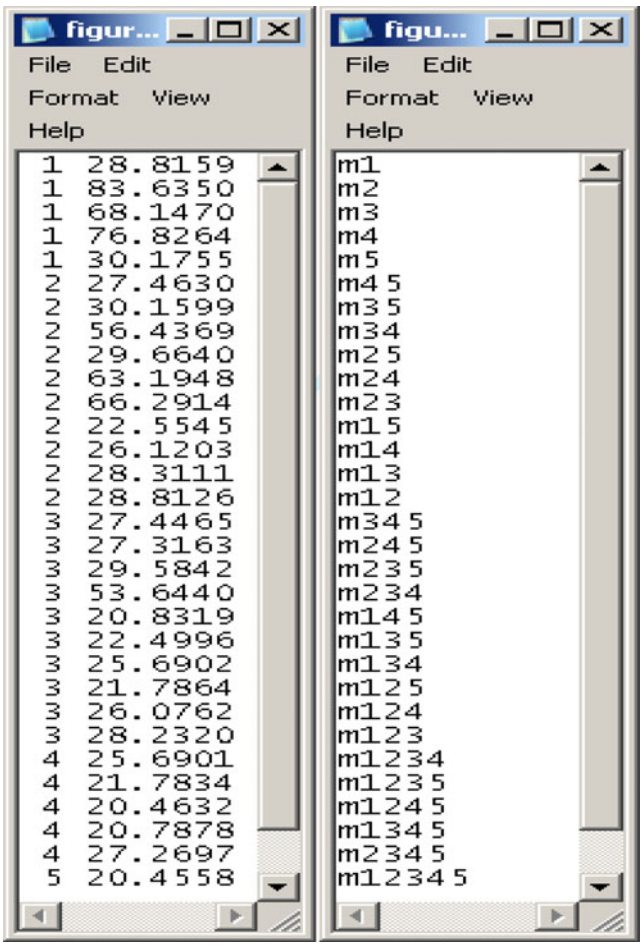

Fig. 3 (Left) Screenshot of the file ("Values.txt") with the complexity and (mis)fit measures for all models under consideration. (Right) Screenshot of the label file ("Labels.txt")

In the plot, the user may highlight a particular model (by dragging the white cross to the circle corresponding to the desired model) and click the left mouse button. As a result, the selected model will also be highlighted in the table displaying the different models. Conversely, it is also possible to highlight a model by selecting in the table the row pertaining to the desired model and clicking the left mouse button, resulting in the selected model being highlighted in both the table and the plot. In both cases, the specified label, identification number, complexity, and (mis)fit measure of the selected model will be displayed above the plot (at the right of the table). As such, the user may interactively inspect the different models and their quality (i.e., whether or not the model is located on the boundary of the convex hull, how far away from the boundary the model is, the *st* value if the model is on the boundary, etc.).

Consulting the analysis results in the output file

When the user requests an output file, the CHULL program will generate a .mht file with the specified name (i.e., "Results.mht") and store the file in the specified folder "(i.e., "C:\CHULL\example\output folder)". This output file contains the following information: (1) a table that displays the identification number, specified label, complexity, and (mis)fit measure of each model under study; (2) a second table that shows, for each model located on the upper (in the case of goodness of fit) or lower (in the case of misfit) boundary of the convex hull, the identification number, specified label, complexity, goodness of fit/misfit, and *st* value; and (3) the complexity-versus-fit plot, including the appropriate boundary of the convex hull (see the description of the plot in section Checking the analysis results in the CHull plot and in a table). When using the MATLAB version of CHULL, a .mat file (in our example, "Results.mat") containing the same information is also stored in the specified folder, and an object called "CHULL_result" appears in the current workspace.

Error handling

After clicking the "Perform CHULL" button, the CHull procedure will be applied to the models at hand (make sure

that the current MATLAB directory is always the folder where the .m-files are stored). When the data or label file is incorrectly specified or does not comply with the required format, or when no analysis options are checked, one or more error screens will appear with information regarding the problem(s) encountered. After clicking the OK button (s), the user is given the opportunity to correct the errors. To assist the user, the content of the error messages is displayed in the box at the bottom of the GUI screen. Once all errors have been corrected, the user should again click the "Perform CHULL" button.

Concluding remarks

In this article, we have demonstrated by means of several illustrations that the CHull procedure is a promising and generic model selection technique that may be useful for solving various model selection problems. In particular, CHull can be applied in all cases in which a model is sought that balances model fit (e.g., the sum of squared residuals or the likelihood) and model complexity (e.g., the number of—free, effective—model parameters). Some examples are: (1) determining the optimal causal model in a structural equation analysis or a log-linear analysis, (2) identifying the best subset of predictors to include in a general/generalized linear (mixed) model (e.g., logistic or multilevel regression), (3) deciding on the value for the penalty parameter in penalized regression problems (e.g., ridge regression, LASSO, or penalized logistic regression), (4) deciding on the optimal number of underlying clusters in a cluster analysis, and (5) obtaining the optimal number of (mixture) components in a mixture analysis (e.g., model-based clustering or mixture of factor analyzers). Furthermore, we presented the CHULL software to assist the user in applying the CHull procedure.

A nice feature of the CHull method, which is part of its generic nature, is that a researcher can decide which fit measure is to be used. In the regression example, for instance, one may use the amount of explained variance or the adjusted *R*-square value as a goodness of fit measure, or the sum of squared residuals or the mean squared error as a badness of fit measure. The choice of a particular fit measure has important implications for which model(s) and which model characteristics will be favored, and the choice should therefore be based on theoretical grounds and/or on the research question at hand. In the regression example, for instance, the mean squared error may be preferred when the generalizability of the model (i.e., the accuracy in predicting the criterion value of new observations) is of utmost importance. Conversely, when one is only interested in the particular data sample that has been collected, the (adjusted) *R*-square value may be recommended. Note that the researcher needs to make a similar choice regarding the complexity measure. The complexity value of a Tucker3 analysis with rank (P, Q, R) of an $I \times J \times K$ data block, for instance, may be expressed by the total number of components (i.e., $P + Q + R$) or by the number of free parameters [i.e., $(I \times P) + (J \times Q) + (K \times R) + (P \times Q \times R) - P^2 - Q^2 - R^2$], with the former measure being more robust (i.e., less influenced by the data characteristics/size) than the latter one (see Ceulemans & Kiers, 2006). Choosing a useful complexity measure may become not trivial at all when models are compared that differ in functional form, because in that case model complexity does not simply depend on the number of free parameters (see Cutting, 2000; Myung, 2000; Pitt et al., 2002). In three-mode component analysis, for instance, a two-component PARAFAC model and a Tucker3 model of rank (2, 2, 2) have the same number of parameters but may not fit the data equally well, and consequently they differ in complexity (see e.g., Kroonenberg & ten Berge, 2011).

When applying the CHull method, one should be cautious when including a set of (very) complex models in the comparison (see section Method). However, because the CHull procedure aims at selecting the simplest model that describes the data well and cannot select the model with the lowest complexity value, researchers should also think carefully about which simple models to include in the comparison. In the PCA (subset regression) example, for instance, the model with one component (predictor) was the simplest model under consideration, implying that this model could not be selected by CHull. However, to circumvent this, one may include in the comparison an even simpler model. In particular, for PCA, a model with no components may be included, and for regression, a model without predictors (i.e., including an intercept only). Yet, when including a model with zero complexity, CHull may be tempted to select the simplest of the other models, because the zero-complexity model often will fit very badly, resulting in an artificially inflated *st* value for the next model (because of the large numerator of the *st* ratio). This model, however, may be too simple.

Finally, we would emphasize again that an automated model selection procedure such as CHull, however tempting it may sound, should never be applied in a rigid way. Indeed, in empirical practice, one should always combine CHull with substantive information regarding the models under study. In other words, we advise researchers to use the CHull procedure as a helpful tool for making a first selection of (potentially) interesting models (by considering models with large *st* values and by also checking models that are located close to the models with large *st* values). When selecting the final model, however, substantive information and the interpretability and stability of the parameter estimates should also be taken into account.

Author note The research reported in this article was partially supported by the Research Fund of KU Leuven (PDM-Kort; T.F.W.); by the Fund for Scientific Research (FWO)–Flanders (Belgium), Project No. G.0477.09 N awarded to E.C., Marieke Timmerman, and Patrick Onghena; and by the Research Council of KU Leuven (GOA/10/02). Requests for reprints should be sent to T.F.W.

References

Agresti, A., & Finlay, B. (2008). *Statistical methods for the social sciences* (4th ed.). Upper Saddle River, NJ: Prentice Hall.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csáki (Eds.), *Second International Symposium on Information Theory* (pp. 267–281). Budapest, Hungary: Akadémiai Kiadó.

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika, 52,* 317–332.

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

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

Bagby, R. M., Parker, J. D. A., & Taylor, G. J. (1994). The twenty-item Toronto Alexithymia Scale: I. Item selection and cross-validation of the factor structure. *Journal of Psychosomatic Research, 38,* 23–32.

Bock, H.-H. (1987). On the interface between cluster analysis, principal component analysis, and multidimensional scaling. In H. Bozdogan & A. K. Gupta (Eds.), *Multivariate statistical modeling and data analysis* (pp. 17–34). Dordrecht, The Netherlands: Reidel.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika, 52,* 345–370.

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

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

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

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

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

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

Cliff, N. (1988). The eigenvalues-greater-than-one rule and the reliability of components. *Psychological Bulletin, 103,* 276–279.

Cutting, J. E. (2000). Accuracy, scope, and flexibility of models. *Journal of Mathematical Psychology, 44,* 3–19. doi:10.1006/jmps.1999.1274

De Soete, G., & Carroll, J. D. (1994). K-means clustering in a low-dimensional Euclidean space. In E. Diday, Y. Lechevallier, M. Schader, P. Bertrand, & B. Burtschy (Eds.), *New approaches in classification and data analysis* (pp. 212–219). Heidelberg, Germany: Springer.

Draper, N., & Smith, H. (1981). *Applied regression analysis* (2nd ed.). New York, NY: Wiley.

Erni, T., Lötscher, K., & Modestin, J. (1997). Two-factor solution of the 20-item Toronto Alexithymia Scale confirmed. *Psychopathology, 30,* 335–340.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics, 7,* 179–188. doi:10.1111/j.1469-1809.1936.tb02137.x

Frontier, S. (1976). Étude de la décroissance des valeurs propres dans une analyse en composantes principales: Comparaison avec le modèle de baton brisé. *Journal of Experimental Marine Biology and Ecology, 25,* 67–75.

Grünwald, P. (2000). Model selection based on minimum description length. *Journal of Mathematical Psychology, 44,* 133–152. doi:10.1006/jmps.1999.1280

Guttman, L. (1954). Some necessary conditions for common factor analysis. *Psychometrika, 19,* 149–161.

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). New York, NY: Springer.

Haviland, M. G., & Reise, S. P. (1996). Structure of the twenty-item Toronto Alexithymia Scale. *Journal of Personality Assessment, 66,* 116–125.

Hocking, R. R. (1976). The analysis and selection of variables in linear regression. *Biometrics, 32,* 1–49.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30,* 179–185.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement, 20,* 141–151.

Kroonenberg, P. M., & Oort, F. J. (2003). Three-mode analysis of multimode covariance matrices. *British Journal of Mathematical and Statistical Psychology, 56,* 305–336.

Kroonenberg, P. M., & ten Berge, J. M. F. (2011). The equivalence of Tucker3 and Parafac models with two components. *Chemometrics and Intelligent Laboratory Systems, 106,* 21–26.

Kroonenberg, P. M., & Van der Voort, T. H. A. (1987). Multiplicatieve decompositie van interacties bij oordelen over de werkelijkheidswaarde van televisiefilms [Multiplicative decomposition of interactions of judgements of realism of television films]. *Kwantitatieve Methoden, 8,* 117–144.

Legendre, L., & Legendre, P. (1983). *Numerical ecology*. Amsterdam, The Netherlands: Elsevier Scientific.

Lorenzo-Seva, U., Timmerman, M. E., & Kiers, H. A. L. (2011). *The Hull method for selecting the number of common factors*. Manuscript submitted for publication.

Mallows, C. L. (1973). Some comments on *Cp*. *Technometrics, 15,* 661–675. doi:10.2307/1267380

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

Myung, I. J. (2000). The importance of complexity in model selection. *Journal of Mathematical Psychology, 44,* 190–204. doi:10.1006/jmps.1999.1283

Pandey, R., Mandal, M. K., Taylor, G. J., & Parker, J. D. A. (1996). Cross-cultural alexithymia: Development and validation of a Hindi translation of the 20-item Toronto Alexithymia Scale. *Journal of Clinical Psychology, 52,* 173–176.

Parker, J. D. A., Bagby, R. M., Taylor, G. J., Endler, N. S., & Schmitz, P. (1993). Factorial validity of the 20-item Toronto Alexithymia Scale. *European Journal of Personality, 7,* 221–232.

Peres-Neto, P. R., Jackson, D. A., & Somers, K. M. (2005). How many principal components? Stopping rules for determining the number of non-trivial axes revisited. *Computational Statistics and Data Analysis, 49,* 974–997.

Pitt, M. A., Myung, I. J., & Zhang, S. (2002). Toward a method of selecting among computational models of cognition.

Psychological Review, 109, 472–491. doi:10.1037/0033-295X.109.3.472

Rissanen, J. (1983). A universal prior for integers and estimation by minimum description length. *The Annals of Statistics, 11,* 416–431.

Rissanen, J. (1996). Fisher information and stochastic complexity. *IEEE Transactions on Information Theory, 42,* 40–47.

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

Schwarz, R. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6,* 461–464.

Stamey, T., Kabalin, J., McNeal, J., Johnstone, I., Freiha, F., Redwine, E., & Yang, N. (1989). Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate: II. radical prostatectomy treated patients. *The Journal of Urology, 16,* 1076–1083.

Timmerman, M. E. (2006). Multilevel component analysis. *British Journal of Mathematical and Statistical Psychology, 59,* 301–320.

Timmerman, M. E., Ceulemans, E., Kiers, H. A. L., & Vichi, M. (2010). Factorial and reduced *k*-means reconsidered. *Computational Statistics and Data Analysis, 54,* 1858–1871. doi:10.1016/j.csda.2010.02.009

Timmerman, M. E., & Kiers, H. A. L. (2000). Three-mode principal component analysis: Choosing the numbers of components and sensitivity to local optima. *British Journal of Mathematical and Statistical Psychology, 53,* 1–16.

Velicer, W. F. (1976). Determining the number of components from the matrix of partial correlations. *Psychometrika, 41,* 321–327.

Wold, H. (1966). Estimation of principal components and related models by iterative least squares. In P. R. Krishnaiaah (Ed.), *Multivariate analysis* (pp. 391–420). New York, NY: Academic Press.