Statistical approaches to explore clinical heterogeneity in psychosis

Islam, Atiqul

CHAPTER 2

A comparison of indices for identifying the number of clusters in hierarchical clustering: A study on cognition in schizophrenia patients

Md. Atiqul Islam

Behrooz Z. Alizadeh

Edwin R. van den Heuvel

GROUP Investigators

Communications in Statistics: Case Studies, Data Analysis and Applications, 1(2): 98-113

Abstract

Finding clusters in a complex dataset is not straightforward. Different indices were developed to quantify the number of clusters. Their performances were studied using unrealistic simulations, since they were considered at low dimensions. We investigated fourteen indices for eight dimensional data using simulations based on cognition measures. We focused on hierarchical clustering with Ward's agglomerative technique. Results indicated that Duda and Hart, Hartigan and Gap/pc were best performing. They estimated the number of clusters within ±1 with high probabilities. Duda and Hart index was most consistent while Gap/pc and WGap/pc together made a good distinction between single and multiple clusters.

Keywords: cluster indices; cluster analysis; hierarchical clustering; homogeneous subgroups; number of clusters

1. Introduction

1.1. Background

Clustering of data is used to classify groups of objects that are similar to one another within groups but different from each other between groups according to a defined mathematical criterion (e.g. Euclidean distance). In psychiatry, clustering is frequently used to group patients (e.g. suffering from schizophrenia) or their siblings to form homogeneous subtypes on the basis of variables or characteristics (e.g. cognition) that are related to the disease. Homogeneous subtypes may help explain or identify the disease severity (Cornblatt and Keilp, 1994; Chen and Faraone, 2000; Joyce and Roiser, 2007; Lin et al., 2009), in particular when the formed subtypes can be related to other clinical outcomes. In psychology, clustering is commonly used in family research to form family subtypes based on criteria such as parenting practices, church attendance, youth self-esteem, ethnic identity, personality, and patterns of change in marriage (Bray et al., 1995; Mandara, 2003; Henry et al., 2005).

Although there are highly complex clustering methods, common clustering techniques are hierarchical methods, which produce a nested sequence of clusters, and partitioning methods (e.g. K-means), which divide the whole set of objects into a pre-defined number of clusters (Hartigan and Wong, 1979; Borgen and Barnett, 1987; Gordon, 1999; Henry et al., 2005; Silver and Shmoish, 2008). Since partitioning methods are sensitive to the starting point (initial centroids) of formulating clusters, hierarchical clustering is typically used to obtain these initial centroids. Good reviews on these methods have been determined elsewhere (Clatworthy et al., 2005; Everitt et al., 2011).

In practice, a priori information about the actual groupings of subjects is typically missing, so it is necessary to identify the number of clusters on the basis of the observed data itself. Since hierarchical clustering is often used as initial step in formulating clusters, this technique can be helpful in defining the most likely number of clusters in combination with additional measures or criteria. Unfortunately, quantifying the true number of clusters has been a serious challenge in hierarchical clustering, since it produces an informative ordering of the objects and produce series of small clusters (Milligan and Cooper, 1985; Fernández and Gómez, 2008). Over the past decades, various measures or indices have been proposed for determining the number of clusters (Beale, 1969; Duda and Hart, 1973; Calinski and Harabasz, 1974; Hartigan, 1975; Sarle, 1983; Krzanowski and Lai, 1988; Kaufman and Rousseeuw, 1990; Tibshirani et al., 2001; Yan and Ye, 2007; Boone, 2011; Albalate et al., 2011; Albatineh and Niewiadomska-Bugaj, 2011), albeit not all indices originated from hierarchical clustering.

Milligan and Cooper (1985) applied the hierarchical clustering technique with a comprehensive Monte Carlo simulation to study thirty different procedures for determining the number of clusters. They suggested that the Calinski and Harabasz (CH) index (Calinski and Harabasz, 1974) and the Duda and Hart (DH) index (Duda and Hart, 1973) were the best performers followed by the C index (Dalrymple-Alford, 1970; Hubert and Levin, 1976), Beale (B) index (Beale, 1969), Cubic Clustering Criteria (CCC) (Sarle, 1983) and Hartigan (H) index (Hartigan, 1975). Despite the large number of indices already available in 1985, other indices have been developed since. Krzanowski and Lai (1988) modified the Marriott index (Marriott, 1971) using the within-group sum of squares as the objective function rather than the within-group determinant and demonstrated that the KL index

was superior to the Marriott index based on simulation studies. Kaufman and Rousseeuw (1990) introduced the *silhouette* (S) index which is based on the comparison of its "tightness" and "separation" using a dissimilarity matrix (e.g. Euclidean distance). They concluded that the highest average *silhouette* width determines the true number of clusters. Tibshirani et al. (2001) developed two Gap statistics (Gap/uniform and Gap/principal components) and compared them with the KL and the S index for K-means clustering. They concluded that the Gap/pc was better than the other indices. Yan and Ye (2007) proposed the weighted Gap (WGap) and the difference of difference-weighted Gap (DD-WGap) statistics using the weighted within-clusters sum of errors (a measure of the within-clusters homogeneity). They used K-means clustering and compared these two methods with the original Gap (uni/pc) statistics in simulation studies. They indicated that the WGap and the DD-WGap statistics were highly effective in determining the number of clusters.

Since the indices B, DH, CH, H and KL are all functions of the same within sums of squares, these indices can in principle be considered identical. However, they all used different functions or calculations with their own developed criteria to identify the true number of clusters. These criteria were based on different distributional characteristics. This makes it hard to compare them mathematically, thus we resorted to simulations. Furthermore, the vast literature on these indices also compared them with simulations, but mostly to relatively simple and nonrealistic settings, e.g. the dimension of the clustering variables were restricted to small numbers. In practice, the dimensions are typically larger (e.g. six to eight variables) and it is unclear if all variables truly contribute to the formation of homogeneous subtypes. This paper evaluates 14 indices for hierarchical clustering by simulation, implementing these practical settings on the basis of a real case study. We studied B, DH, CH, H, C-index, CCC, KL, S, Gap, WGap, and DD-WGap indices since these indices have been reported in literature as good performers. Some of the indices in this study are based on the (within and between clusters) sum of squares (B, DH, CH, H, CCC, KL, and Gap) and some others are calculated from a dissimilarity matrix (C and S). We will investigate how well they predict the simulated number of clusters, but we will also determine whether they can be used to decide if there would exist multiple clusters or not. Details on these indices are discussed in section 2, while the motivating example, which is used as input for the simulation study, is provided next. Details on the simulation study are provided in section 3 followed by the results of the simulation studies in section 4. Finally, Section 5 provides conclusions and final comments.

1.2. Motivating Example

Cognitive heterogeneity is an obstacle to clarify the neuropathological foundations of schizophrenia. Heterogeneity in cognition may possibly be addressed adequately using clustering techniques to form homogeneous cognitive subtypes (Jablensky, 2006; Joyce and Roiser, 2007; Dawes et al., 2011) to identify disease (schizophrenia) severity. In our study, data was extracted from the Genetic Risk and Outcome of Psychosis (GROUP), a longitudinal multi-center cohort study in the Netherlands and Belgium. A detailed description of the GROUP study has been published elsewhere (Korver et al., 2012). In brief, it comprised measures on cognitive functioning, clinical symptoms and genetic make-up of patients, their unaffected siblings and parents, and unrelated controls at multiple time points. Eight neurocognitive measures were obtained: Continuous Performance Test-HQ (CPT-HQ) and

Standard deviation of CPT-HQ (attention/vigilance) (Stinissen et al., 1970; Quee et al., 2011), Word Learning Task (WLT) Immediate and Delayed Recall (verbal learning and memory) (Brand and Jolles, 1985), Wechsler Adult Intelligence Scale-III (WAIS-III) Digit Symbol Coding (processing speed), WAIS-III Arithmetic (working memory), WAIS-III Block Design (reasoning and problem solving), and WAIS-III Information (verbal comprehension) (Blyler et al., 2000). These cognitive traits have been described in detail elsewhere (Meijer et al., 2012; Quee et al., 2014). We obtained data on 860 independent patients with 77.4% male and mean age 27.22 years (SD=7.46) and 439 independent controls with 48.1% male and mean age 30.46 years (SD=10.53) with complete cognitive measures and all from different families at baseline. The mean ± standard deviation of the raw scores of all cognitive variables were as follows: CPT performance index (220.95 ± 63.05), CPT standard deviation (92.90 ± 36.29), Block Design (40.28 ± 17.02), Digit Symbol (65.01 ± 15.98), Arithmetic (12.17 ± 4.79), Information (16.66 ± 5.51), WLT Immediate Recall (22.91 ± 6.11), and Delayed Recall (7.52 ± 2.85). The controls were used to obtain z-scores on patients for all eight cognitive measures that were age and gender specific, and these z-scores were used for clustering.

The number of clusters suggested by the fourteen selected indices using hierarchical (Ward's agglomerative) clustering are provided in Table 1. The dendrogram (Supplementary Figure 1) is presented in the Supplementary file. The indices do not agree on determining the number of clusters. The B index suggests that the patients represent one homogeneous group, while the CH, H, S, CCC, Gap/uni, Gap/pc, WGap/uni, and WGap/pc all suggest just two cognitive subtypes. Four subtypes are indicated by the KL, DD-WGap/uni and DD-WGap/pc, while the C-index indicates one subtype fewer and the DH even indicates five subtypes. Based on these results, it would be difficult to determine the correct number of clusters, unless we know which indices are most reliable for this type of data.

Table 1: The number of clusters determined by different indices from GROUP data

Number of clusters	Indices
1	B
2	CH, H, CCC, S, Gap/uni, Gap/pc, WGap/uni, WGap/pc
3	C
4	KL, DD-WGap/uni, DD-WGap/pc
5	DH

2. Methods

2.1. Overview of the selected indices

Before discussing the indices we need to introduce some notations. Assume *K* clusters have been identified in the set of patients, and let Y_{hcj} be the z-score for variable *h* (= 1, 2, …, *p*) on patient *j* (= 1, 2, …, n_c) in cluster *c* (= 1, 2, …, *K*). The total number of patients is $n\ (= n_1 + n_2 + \cdots + n_K)$. The within ($SSW_K$) and between ($SSB_K$) sums of squares of the clusters are given by $SSW_K = \sum_{h=1}^{p}\sum_{c=1}^{K}\sum_{j=1}^{n_c}\left(Y_{hcj} - \bar{Y}_{hc.}\right)^2$ and $SSB_K = \sum_{h=1}^{p}\sum_{c=1}^{K} n_c(\bar{Y}_{hc.} - \bar{Y}_{h..})^2$ with $\bar{Y}_{hc.}$ the average value of the variable *h* in cluster *c*, and $\bar{Y}_{h..}$ the average value of variable *h.* It should be noted that for hierarchical clustering, the clusters are nested, which means that an additional cluster would imply a split of one of the existing clusters $c = 1, 2, \ldots, K$. This means that SSW_1 is equal to the total sums of squares (*SST*) of all the data, with $SST = SSW_K + SSB_K$ for all numbers of clusters $K \geq 1$.

Furthermore, let d_{jk} be the distance between patient *j* and patient *k*, and let $x_{jk}^{c_1c_2}$ be the indicator function defined as follows:

$$x_{jk}^{c_1c_2} = \begin{cases} 1 \text{ if } j \in I_{c_1} \text{ and } k \in I_{c_2} \\ 0 \text{ otherwise,} \end{cases}$$

with I_c the set of patients that are contained in cluster *c*. For convenience, we define $d^{(r)}$ to be the r-th ranked distance among all $n(n-1)/2$ possible distances. Thus $d^{(1)}$ is the smallest distance between any two patients and $d^{\{n(n-1)/2\}}$ is the largest. In hierarchical clustering, clusters can be formed sequentially following the dendrogram from one homogeneous group to *n* clusters at the end. This procedure is followed for all indices and the indices will judge whether the observed set of *K* clusters should be changed to *K*+1 clusters. The way that we walk down the dendrogram is determined by the distance of two clusters (i.e. the length of the arms of the dendrogram), choosing larger distance first.

2.1.1. Beale (B) Index

Beale (1969) proposed an index which allows a significance test for choosing the number of clusters. The existence of an additional cluster is tested by the F-test, and the index is defined by

$$B_K = \frac{(n_c - 2)(SSW_K - SSW_{K+1})}{SSW_{K+1}\{(n_c - 1)2^{2/p} - (n_c - 2)\}}, K \geq 1,$$

where n_c is the number of observations of cluster *c* that was split up to form *K*+1 clusters in total. Note that $SSW_K - SSW_{K+1}$ depends on the data of cluster *c* alone. When B_K is smaller than or equal to the critical value of the F-distribution with *p* and $(n_c - 2)p$ degrees of freedom, cluster *c* is not considered heterogeneous and should not be divided into two clusters (Milligan and Cooper, 1985; Everitt et al., 2011).

2.1.2. Duda and Hart (DH) Index

Duda and Hart (1973) suggested the ratio of the two within sum of squares to decide whether a cluster can be divided into two clusters, i.e.

$$DH_K = \frac{SSW_{K+1}}{SSW_K}, K \geq 1.$$

For large number of observations in cluster *c*, SSW_K is approximately normally distributed with mean $n_c p\sigma^2$ and variance $2n_c p\sigma^4$ and SSW_{K+1} follows approximately a normal distribution with mean $n_c p(1 - 2/\pi p)\sigma^2$ and variance $2n_c p\{1 - 8/(\pi^2 p)\}\sigma^4$. The variance σ^2 represents the variance of the complete population. From these observations the following criterion was proposed to sub-divide whenever the following holds

$$DH_K < \left[1 - \frac{2}{\pi p} - z\sqrt{\frac{2\{1 - 8/(\pi^2 p)\}}{n_c p}}\right],$$

with *z* a standard normal quantile. Milligan and Cooper (1985) suggested to use the value 3.20 for *z* (Everitt et al., 2011; Milligan and Cooper, 1985; Duda and Hart, 1973).

2.1.3. Calinski and Harabasz (CH) Index

Calinski and Harabasz (1974) proposed the CH index for finding the number of clusters:

$$CH_K = \frac{SSB_K/(K-1)}{SSW_K/(n-K)}, K \geq 2.$$

The number of clusters was chosen to be the value of *K* that maximizes CH_K. Note that the CH index always assumes that there are at least two clusters and cannot be used to decide if clustering is needed.

2.1.4. Hartigan (H) Index

Hartigan (1975) originally proposed the *H* index for the number of clusters with K-means clustering. The index is defined by

$$H_K = (n-K-1)\left\{\frac{SSW_K - SSW_{K+1}}{SSW_{K+1}}\right\}, K \geq 2.$$

The term (*n*-*K*-1) is a penalty factor which avoids an increasing monotonicity with an increasing number of clusters. Hartigan (1975) recommended that the number of clusters is the smallest *K* for $H_K \leq 10$. Alternatively, Milligan and Cooper (1985) used $(H_{K+1} - H_K)$ and recommended the number of clusters *K* that maximizes this difference. In this study, we used the criterion of Milligan and Cooper (1985), but this implies that there should always exist at least two clusters even if no subtypes would be present.

2.1.5. C-index

Dalrymple-Alford (1970) proposed the C-index and later Hubert and Levin (1976) revised it. Let d_{jk} be the Euclidean distance, i.e.

$$d_{jk}^2 = \sum_{h=1}^{p} \left(Y_{hj} - Y_{hk}\right)^2,$$

with Y_{hj} the observation on variable *h* and patient *j* ignoring the possible cluster structure in which this patient belong. Within cluster *c*, there are $n_c(n_c-1)/2$ distances to be calculated and the total number of such pairs over all clusters is

$$n_{WK} = \sum_{c=1}^{K} (n_c^2 - n_c)/2.$$

As mentioned before, the total number of pairs in the data set is $n_T = n(n-1)/2$. This total number can be rewritten as $n_T = n_{WK} + n_{BK}$, with $n_{BK} = \sum_{c_1=1}^{K-1} \sum_{c_2=c_1+1}^{K} n_{c_1} n_{c_2}$ the number of pairs that do not belong to the same cluster. If we now define

$S_{WK} = \sum_{c=1}^{K} \sum_{j=1}^{n-1} \sum_{k=j+1}^{n} d_{jk} x_{jk}^{cc}$, $S_{min} = \sum_{r=1}^{n_{WK}} d^{(r)}$, and $S_{max} = \sum_{r=n_{BK}}^{n(n-1)/2} d^{(r)}$, the C-index is defined by

$$C_K = \frac{S_{WK} - S_{min}}{S_{max} - S_{min}}, K \geq 2,$$

under the assumption that $S_{min} \neq S_{max}$. Note that C_K is always an element of the interval [0, 1]. The number of clusters *K* is the number that minimizes C_K. It is always larger than one, which means that C_K cannot be used to decide if clustering should be proceeded.

2.1.6. Cubic Clustering Criteria (CCC)

Sarle (1983) developed a crude test for testing the null hypothesis that data have been sampled from a uniform distribution on a hyperbox against the alternative hypothesis that data have been sampled from a mixture of spherical multivariate normal distribution with equal variances and sampling probabilities. The author compared observed value $R_K^2 = 1 - SSW_K/SST$, the proportion of variance accounted for by the *K* clusters, with an approximation of its expected value $\mathbb{E}R_K^2$ under the assumption that the *K* clusters are generated by a *p*-dimensional uniform distribution. The Cubic Clustering Criteria (CCC) is defined as

$$CCC_K = ln\left[\frac{1 - E(R_K^2)}{1 - R_K^2}\right]\frac{\sqrt{np^*/2}}{[0.001 + E(R_K^2)]^{1.2}}, K \geq 1,$$

and $\mathbb{E}R_K^2$ is defined by

$$\mathbb{E}R_K^2 = 1 - \left[\frac{\sum_{h=1}^{p^*}(n + u_h)^{-1} + \sum_{h=p^*+1}^{p} u_h^2(n + u_h)^{-1}}{\sum_{h=1}^{p} u_h^2}\right][(n - K)^2/n](1 + 4/n),$$

where $u_h = s_h/m$, s_h is the square root of the *h*-th eigen value of *SST*/(*n*-1), $m = (v^*/K)^{1/p^*}$, with $v^* = \prod_{h=1}^{p^*} s_h$ and where p^* is chosen to be the largest integer less than *K* such that u_{p^*} is not less than one. A positive value of CCC_K means that the observed R_K^2 is greater than the expected R_K^2 under the uniform distribution and the cluster structure of the data is different from the uniform partition (i.e. reject the null hypothesis). The number of cluster *K* is determined by the number that maximizes CCC_K.

2.1.7. Krzanowski and Lai (KL) Index

Let $DIFF_K$ denote a scaled difference between the within sum of squares of two sequential clusterings, i.e.

$$DIFF_K = (K - 1)^{2/p} SSW_{K-1} - K^{2/p} SSW_K, K \geq 2.$$

Krzanowski and Lai (1988) argued that under independent uniformly distributed data the sum of squares $K^{2/p} SSW_K$ is constant and independent of the number of clusters *K*. They suggested to calculate the *KL* index by the ratio of two difference measures

$$KL_K = \left|\frac{DIFF_K}{DIFF_{K+1}}\right|, K \geq 2.$$

The proposed number of clusters is the number *K* that maximizes KL_K. The KL_K cannot be used to decide if clustering is needed or not.

2.1.8. Silhouette (S) Index

Kaufman and Rousseeuw (1990) proposed the *S*-index for assessing and estimating the true number of clusters. Let us define the within-cluster mean distance a_j^c as the mean distance of patient *j* to the other patients in cluster *c*, i.e.

$$a_j^c = \frac{1}{n_c - 1}\sum_{K=1} d_{jk} x_{jk}^{cc}.$$

The mean distance of patient *j* in cluster c_1 to the patients in another cluster c_2 is defined by

$$b_j^{c_1c_2} = \frac{1}{n_{c_2}} \sum_{K=1}^{n} d_{jk} x_{jk}^{c_1c_2}.$$

Then the smallest of these mean distances $b_j^{c_1c_2}$ over clusters $c_2 \in \{1, 2, \dots, K\} / \{c_1\}$ is defined by $b_j^c = min\left\{b_j^{c_1c_2} \middle| c_2 \in \{1, 2, \dots, K\} / \{c_1\}\right\}$.

The *silhouette* width of patient *j* in cluster *c* is now given by

$$S_j^c = \frac{b_j^c - a_j^c}{max\{a_j^c, b_j^c\}}.$$

Note that S_j^c is an element of the interval [-1, 1]. A value of S_j^c near one indicates that patient *j* is categorized within the right cluster while a value near minus one indicates that the patient could be better changed to another cluster. The average *silhouette* index for cluster *K* is defined by

$$S_K = \frac{1}{K} \sum_{c=1}^{K} \sum_{j=1}^{n_c} S_j^c / n_c \,, c = 1, 2, \dots, K.$$

The number of clusters is taken as the number that maximizes S_K across the hierarchical formulation of clusters.

2.1.9. Gap/uni and Gap/pc Statistic

Tibshirani, Walther, and Hastie (2001) proposed the Gap statistic as comparing the logarithm of within sums of squares with the expectation of this term under the reference distribution of the data. Therefore, the statistic is defined by

$$GAP_K = \mathbb{E}\{log(SSW_K)\} - log(SSW_K).$$

The expected value $\mathbb{E}log(SSW_K)$ is unknown and it is therefore determined using Monte Carlo simulation from a reference distribution and applying bootstrapping. On the basis of bootstrap sampling with *B* samples $\mathbb{E}log(SSW_K)$ is estimated with $\sum_{b=1}^{B} log(SSW_{K,b}^*) / B$, with $SSW_{K,b}^*$ be the within sum of squares for bootstrap sample *b*.

The number of clusters *K* is determined by the smallest number such that the following holds

$$GAP_K \geq GAP_{K+1} - S_{K+1}, K \geq 1,$$

with $S_K = sd_K\sqrt{(1 + 1/B)}$ is the total standard error and sd_K is given by

$$sd_K^2 = \frac{1}{B} \sum_{b=1}^{B} \left\{ log(SSW_{K,b}^*) - 1/B \sum_{b=1}^{B} log(SSW_{K,b}^*) \right\}^2.$$

There were two choices about the reference distribution for the Gap statistic (Tibshirani et al., 2001). In the first choice, each variable was generated from the uniformly distribution over the range of the observed values for the variable *p*. Determining the number of clusters via this choice was referred to as Gap/uniform or Gap/uni. In the second choice, the variables were sampled from a uniform distribution over a box aligned with the principal components of the centered design matrix. The new design matrix was then back transformed to obtain the reference dataset. This procedure of calculating the number of clusters was referred to as Gap/principal components or Gap/pc.

2.1.10. Weighted Gap (WGap)

Define the sum of pairwise distances between all patients within cluster *c* by

$$\overline{D}_c = \sum_{h=1}^{p}\sum_{j=1}^{n}\sum_{k=1}^{n}\left(Y_{hj} - Y_{hk}\right)^2 x_{jk}^{cc}/[2n_c(n_c-1)],$$

with Y_{hj} the value for patient *j* on variable *h* ignoring the cluster structure. The weighted within sum of squares is defined by $\overline{SSW}_K = \sum_{c=1}^{K}\overline{D}_c$. Yan and Ye (2007) proposed this WGap as an alternative to the original Gap statistic, but followed the exact same approach of Tibshirani *et al.* (2001) to compare it with its expectation. Therefore, the weighted Gap statistic is defined by

$$\overline{GAP}_K = \frac{1}{B}\sum_{b=1}^{B} log\left(\overline{SSW}_{K,b}^{*}\right) - log(\overline{SSW}_K), K \geq 1.$$

Both a WGap/uni and WGap/pc were obtained using the same reference distributions. The number of clusters *K* is determined to be the number that maximizes $\overline{GAP}_K$.

2.1.11. Difference of difference weighted Gap (DD-WGap)

The WGap method was used to test the null hypothesis of one homogeneous group against the alternative of multiple subtypes. If there is more than one clusters, the original Gap statistic and the WGap statistic may have a tendency to overestimate the number of clusters (Dudoit and Fridlyand, 2002; Yan and Ye, 2007). Therefore, the DD-WGap statistic has been proposed by Yan and Ye (2007) to find the best estimate of the number of clusters more efficiently. Let $D\overline{GAP}_K$ denote the difference in two sequential weighted Gap statistics

$$D\overline{GAP}_K = \overline{GAP}_K - \overline{GAP}_{K-1}, K \geq 2.$$

If the data are strongly grouped around *K* (*K* ≥ 2) modes based on $\overline{SSW}_K$, the function is defined by the difference of difference weighted Gap (DD-WGap) function

$$DD\overline{GAP}_K = D\overline{GAP}_K - D\overline{GAP}_{K+1}.$$

The number of clusters *K* is determined by the number that maximizes $DD\overline{GAP}_K$. Both the DD-WGap/uni and the DD-WGap/pc statistics can be computed like with the original Gap statistic.

2.2. Simulation Settings

Simulation designs in existing literature used almost the same approach and their designs were formulated based on hypothetical mean and covariance structures (Tibshirani et al., 2001; Albatineh and Niewiadomska-Bugaj, 2011; Albalate et al., 2011). The present study uses various means and covariance matrices derived from the GROUP study. Five different scenarios were chosen: (i) a single cluster structure, (ii) two clusters with a ratio of cluster sizes being 75% and 25%, (iii) three clusters with a ratio of cluster sizes being 40%, 35% and 25%, (iv) four clusters with a ratio of cluster sizes being 40%, 30%, 20% and 10%, and (v) five clusters with a ratio of cluster sizes set at 35%, 25%, 20%, 15% and 10%. The means and covariance matrices for the eight dimensional cognition variables for the clusters were determined from the GROUP study using K-means clustering with the pre-specified number of clusters described above. Based on these input and the associated ratios of sample sizes we simulated normally distributed data for the five settings described above. For each setting, 1000 datasets of 860 subjects with 8 dimensional cognitive z-scores were generated from the mixture of

multivariate normal distributions. The means and covariance matrices that we used for each setting are presented in Tables A1a to A5b in the Appendix file. We have also simulated the above mentioned five scenarios where we chose only six out of eight variables that contributed to the clusters. The two variables CPT performance index and CPT standard deviation were treated as nuisance variables to see if such setting would alter the performance of the indices. In practice we would not know which of these variables would or would not contribute to subtypes.

Two packages in R, clusterSim and NbClust were used to analyze the GROUP study and simulation studies. RStudio version 0.97.551 (R version 3.0.1) was used throughout the analysis.

3. Results of simulation study

Only seven indices were capable of identifying a single cluster. The other indices assume that clustering should be conducted. For these seven indices, the percentage of simulation runs that a single solution was selected is presented in Table 2.

Table 2: Percentage of identifying a single cluster solution from simulation study

	Simulated number of clusters (%)				
	1	2	3	4	5
B	97.1	91.4	28.3	1.4	0.6
DH	0	0	0	0	0
CCC	0	0	0	0	0
Gap/uni	0	0	5.5	0	0
Gap/pc	98.4	26.3	20.2	8.5	0
WGap/uni	0.6	1.8	0.5	0.1	0
WGap/pc	99.1	53.3	26.8	1.7	1.1

When just one single cluster was simulated, DH, CCC, Gap/uni, and WGap/uni were incapable of identifying just one cluster. They always seem to indicate incorrectly a multiple clusters solution. Contrary, B, Gap/pc, and WGap/pc correctly identified a single cluster solution with 97.1%, 98.4% and 99.1% respectively. However, when two and three clusters were simulated, B frequently incorrectly predicted a single cluster solution: 91.4% for 2 clusters and 28.3% for 3 clusters. Gap/pc seems to do better, because it incorrectly chooses 26.3% and 20.2% single cluster solutions when two and three clusters were simulated respectively. The Gap/pc seems to be a good index to answer the question if clustering should be conducted. The WGap/pc on the other hand, performs somewhere in between the performance of B and Gap/pc. Combining Gap/pc and WGap/pc improves the performance. If they both indicate that a single cluster solution is present, they correctly identify a single cluster with 97.5%. When two or three clusters are present they incorrectly predict a single cluster with 20.5% and 6.4%, respectively. For larger number of clusters these percentages vanish.

Table 3 shows the percentages of correctly identifying the exact number of simulated clusters from 1000 simulated datasets for each of the methods (columns (a)). Additionally, the percentage of datasets for which the identified number of clusters deviates no more than one cluster from the simulated number of clusters is also specified (columns (b)). This percentage informs us how frequent a method would predict within a range of K-1, K, K+1 clusters, when we intentionally

simulated *K* clusters. We believe that this is a measure of closeness or stability, which could be relevant too when the frequencies of predicting the correct number of clusters is relatively low. If the correct number is missed, how far away are the predictions from the correct number. For instance, compare CCC with KL on three simulated clusters. The CCC predicts three clusters with 12%, while KL predicts three clusters with 42.1%. Clearly KL seems to perform better than CCC for three clusters. This is only partly true, since KL predicts 2, 3 or 4 in 60.8% and predicts 1, 5, 6 or more clusters still with almost 40%. CCC never predicts 1, 5, 6 or more clusters and provides therefore a stable estimate on the number of clusters. So choosing between KL and CCC is not as simple.

The results demonstrate that the DH index predicts the simulated number of clusters within plus or minus one (*K*-1, *K*, *K*+1) quite well (82.3% - 95.4%) over the range of 2 to 5 clusters. When it comes to identifying the exact number, DH predicts this with about 50% when 3 to 5 clusters would be present.

Table 3: Percentage of identifying the exact number of simulated clusters and within a range of just one cluster from simulation study

Indices	Simulated number of clusters (%)							
	2		3		4		5	
	(a)*	(b)**	(a)	(b)	(a)	(b)	(a)	(b)
B	8.6	100.0	1.0	71.7	0	0.1	0	0
DH	28.1	82.3	43.4	95.4	50.0	89.9	52.6	89.7
CH	100.0	100.0	0.2	100.0	0	0.3	0	0
H	97.1	100.0	4.6	100.0	73.2	99.7	10.5	33.2
C	41.6	55.6	13.7	37.1	18.9	49.5	26.0	49.6
CCC	100.0	100.0	12.0	100.0	0.1	0.7	0.4	0.7
KL	40.3	54.2	42.1	60.8	46.4	61.2	14.3	34.9
S	100.0	100.0	0	100.0	0	0	0	0
Gap/uni	1.8	47.6	20.4	54.2	55.7	84.0	35.3	85.8
Gap/pc	55.5	99.8	30.7	56.1	70.0	87.3	18.4	45.8
WGap/uni	42.0	83.6	35.5	90.7	23.0	67.2	7.9	36.3
WGap/pc	45.0	99.9	13.5	73.0	4.7	26.0	1.3	8.9
DD-WGap/uni	82.7	91.3	5.3	94.4	0.3	1.8	0.2	1.3
DD-WGap/pc	36.4	46.9	13.2	71.0	2.1	7.4	0.8	3.7

*Predicting exact number of simulated clusters, **Predicting the exact number of simulated clusters within a range of just one cluster: {*K*-1, *K*, *K*+1}.

The H index performs better (99.0% to 100%) than DH in predicting the number of clusters within plus or minus one when the simulated number of clusters is less than 5, but it fails dramatically when 5 clusters are simulated (33.2%). Predicting the exact number of clusters with H is low when 3 and 5 underlying clusters are present (4.6% and 10.5%), while this percentage is quite good for 2 and 4 clusters (97.1% and 73.2%). On the other hand, Gap/pc predicts the number of clusters within the range of plus minus one better than DH at 2 clusters. However, DH and H predict the number of clusters better than Gap/pc when 3 and 4 underlying clusters are present. At 5 clusters, Gap/pc is better than H, but worse than DH. Some other indices, such as CH, CCC, and S have perfect performances when 2 clusters are present, but do not perform very well at other settings. They seem to have a preference to predict always 2 clusters, whatever the underlying cluster structure. Furthermore, the WGap/pc recovers quite well at 2 and 3 clusters, but failed at 4 and 5 clusters. It does not do much better than its originator Gap/pc index. B is not very good in predicting

the exact number of clusters for more than one cluster. Both C and KL indices are quite robust over the whole range of the number of clusters, but they have only a medium level performance. Interestingly enough, the Gap/uni seems to perform better when higher number of clusters is involved, although it never performs very well. Finally, the WGap/uni, DD-WGap/uni, and DD-WGap/pc, which represent the latest developments in indices, did not do better than earlier developed indices in our simulations. In addition, when we treated two out of eight variables as nuisance variables for all five simulation designs, the results did not change dramatically and our conclusions remained the same (data not shown). This may indicate that the performance of the indices is not that much affected by nuisance variables. When we would first decide if clustering is needed based on the Gap/pc and WGap/pc together, the results are not changed much either (data not shown).

4. Conclusions

Clustering is an explorative analysis and one of the major challenges is to determine the number of clusters in a complex heterogeneous dataset. Although complex statistical methods are available (using mixture models), relative simple and straightforward methods like K-means clustering are used most frequently. They are often supported with hierarchical clustering to help identifying the number of clusters and select good initial centroids for K-means clustering. Therefore, we investigated the most promising indices for detecting the correct number of clusters on the basis of hierarchical clustering (with Ward's agglomerative method). The indices were investigated on (i) how well they would decide between a single and multiple cluster solution and (ii) on how well they can predict the number of clusters in a multi cluster solution. We simulated clusters based on a real case study of patients with schizophrenia and their neurocognitive measured variables. This complements on the performances of the indices from literature, since the indices were mainly studied with artificial low dimensional simulated data. Although our results support earlier studies, we also found opposite results.

Milligan and Cooper (1985) demonstrated that the B index performed relatively well in identifying the number of clusters when three or more clusters were simulated and less for two clusters. They noted that the B index is an appropriate index when the clusters are well separated and spherical (Beale, 1969; Tibshirani et al., 2001). Although our clusters were also spherical, they were mostly not well separated and this explained the opposite result we found with the B index. In our study, the B index demonstrated a good performance for a single cluster and two clusters, but it failed to detect the correct number of clusters at four and five clusters. Milligan and Cooper (1985) already mentioned the good performance of DH. They also demonstrated that DH was particularly capable of identifying high numbers of clusters and they ranked DH as the second best index, among 30 indices. We demonstrated that DH was most consistent among all the indices we studied with high performances of recovering the number of clusters within a range of one. The CH index was considered the best in the simulation study of Milligan and Cooper (1985) and performed good in almost all investigations of Tibshirani et al. (2001). We confirmed that CH performed well up to 3 clusters, but it failed completely at higher number of clusters. This index seems to have a preference for choosing 2 clusters (in all settings). Tibshirani et al. (2001) showed that the H index did not

perform well at 3 and 4 clusters when the dimension of variables was low (2 or 3 variables). This was supported by the simulations of Albatineh and Niewiadomska-Bugaj (2011). Milligan and Cooper (1985) ranked the H index at place 16 among 30 indices. However, for our eight-dimensional cognition outcomes, the H index did particularly well at 2 to 4 clusters. The C index was originally developed for binary outcome data, but Milligan and Cooper (1985) demonstrated that it performed adequately for continuous outcome too. They showed that it performed well, except for 2 cluster solutions. They ranked this index as the third best index among 30 indices. In our simulation the index was consistent for all cluster solution, but the recovery of clusters was only moderate. It predicted the number of clusters within a range of one with approximately 50%. The CCC seems to have a preference for choosing 2 clusters in our simulation, similar as the CH index. It failed to predict the number of clusters for higher number of cluster solutions and underestimates the number of clusters. Milligan and Cooper (1985) and Boone (2011) demonstrated an opposite observation, namely that the CCC index would over-estimate the number of clusters. This could possibly be explained by the fact that our clusters are not clearly separated. We found that the KL index was consistent across different numbers of clusters, but it had only a medium performance. This supported the results of Albatineh and Niewiadomska-Bugaj (2011) and Marriott (1971), who showed that approximately 40 to 50% of the clusters were adequately identified with the KL index. It was however less than the results of Tibshirani et al. (2001), who demonstrated higher performances for 2 to 4 clusters. The S index performed quite well in the study of Albatineh and Niewiadomska-Bugaj (2011) and for 3 clusters in the study of Tibshirani et al. (2001). We demonstrated that the S index selected frequently two clusters under almost all settings, similar to the CH index. This implies that the S index is less suitable for higher number of clusters and explains its good performance for lower number of clusters. Our result on the Gap/pc index is in line with the results of the originator Tibshirani et al. (2001). The Gap/pc performs best at single clusters and lower number of clusters and less at higher number of clusters. This distinction in performance though was not observed by Albatineh and Niewiadomska-Bugaj (2011). The Gap/uni seemed to perform in our simulation reasonable well at 4 and 5 clusters, but not very well at lower number of clusters. This was contrary to the results of Tibshirani et al. (2001) and Albatineh and Niewiadomska-Bugaj (2011), who showed unsatisfactory results at higher number of clusters. Yan and Ye (2007) introduced several alternative indices that were based on the ideas of Gap, but their results could not be reproduced. We did not demonstrate that any of the WGap/uni, WGap/pc, DD-WGap/uni, and DD-WGap/pc performed systematically better than Gap/uni or Gap/pc. Only at three clusters did these indices performed better, but in other settings it was less than Gap/uni or Gap/pc.

Note that we also simulated settings where only six out of eight variables contributed to the subtypes, but these settings did not change our conclusions on the performance of the indices. They performed almost identical to the setting where all eight variables determined the subtypes. However, we focused only on one particular distance measure for identifying the number of clusters. It may be possible that some of the indices may provide different results if other distances or dissimilarity measures are used. Furthermore, our study investigated the proposed indices using sequential stopping criteria for the hierarchical formation of clusters, i.e. the contribution of a new cluster was evaluated with respect to the previous number of clusters, wherever the next cut in the

tree would happen. If the next set of clusters did not contribute, the number of clusters was determined. This means that the formation of clusters in the dendrogram cuts the branches horizontally. Alternatively, cutting branches can also be performed dynamically, which means that at certain trees in the dendrogram the splitting of clusters may continue, while at other parts of the dendrogram trees are not split up further. It would be of interest to find out whether dynamic cutting would improve the performance of certain indices. Furthermore, the criterion for comparing two sequential sets of hierarchical clusters is not changed with the number of previous comparisons (i.e. multiple testing issues). For instance, the z-value in the criterion for the DH index was selected at 3.20, which corresponds with a two-sided significance level of 0.0014. Instead, the criterion could be altered every time a next comparison is conducted, starting with a lower value for a single cluster and slowly rising to higher significance levels when the number of comparisons increases. A more strict criterion at the single cluster could improve the performance of the DH index for single clusters, but also for other indices. Finally, the indices may also be compared with alternative methods for identifying the number of clusters, such as the Bayesian Information Criterion that is applied to clustering methods that uses maximum likelihood. We studied only the indices, since they fit better with the relative simple but frequently used methods of clustering.

In conclusion, we found that the DH, H, and Gap/pc were the best performing indices in our simulation study based on eight-dimensional outcome variables taken from a real case study of schizophrenic patients. They predicted the simulated number of clusters within the range of one cluster with high probabilities. The DH index was most consistent, while Gap/pc in combination with WGap/pc is capable of answering the question if a multiple cluster solution is present.

References

Albalate, A., Suendermann, D., Minker, W. (2011). On cluster validation for detecting the number of clusters in a data set. *International Journal on Artificial Intelligence Tools* 20:941-953.

Albatineh, A.N., Niewiadomska-Bugaj, M. (2011). MCS: A Method for Finding the Number of Clusters. *Journal of classification* 28:184-209.

Beale, E.M.L. (1969). *Cluster analysis*. UK: London: Scientific Control Systems.

Blyler, C.R., Gold, J.M., Iannone, V.N., Buchanan, R.W. (2000). Short form of the WAIS-III for use with patients with schizophrenia. *Schizophrenia research* 46:209-215.

Boone, D.S. (2011). Determination of the number of clusters in a data set: A stopping rule X clustering algorithm comparison. *International Journal of Strategic Decision Sciences* 2(4):1-13.

Borgen, F.H., Barnett, D.C. (1987). Applying cluster-analysis in counseling psychology research. *Journal of counseling psychology* 34:456-468.

Brand, N., Jolles, J. (1985). Learning and retrieval rate of words presented auditorily and visually. *The Journal of general psychology* 112:201-210.

Bray, J.H., Maxwell, S.E., Cole, D. (1995). Multivariate statistics for family psychology research. *Journal of family psychology* 9:144-160.

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

Chen, W.J., Faraone, S.V. (2000). Sustained attention deficits as markers of genetic susceptibility to schizophrenia. *American Journal of Medical Genetics* 97:52-57.

Clatworthy, J., Buick, D., Hankins, M., Weinman, J., Horne, R. (2005). The use and reporting of cluster analysis in health psychology: a review. *British journal of health psychology* 10:329-358.

Cornblatt, B.A., Keilp, J.G. (1994). Impaired attention, genetics, and the pathophysiology of schizophrenia. *Schizophrenia bulletin* 20:31-46.

Dalrymple-Alford, E. (1970). Measurement of clustering in free recall. *Psychological bulletin* 74:32-34.

Dawes, S.E., Jeste, D.V., Palmer, B.W. (2011). Cognitive profiles in persons with chronic schizophrenia. *Journal of clinical and experimental neuropsychology* 33:929-936.

Duda, R.O., Hart, P.E. (1973). *Pattern classification and scene analysis*. New York, USA: New York: Wiley.

Dudoit, S., Fridlyand, J. (2002). A prediction-based resampling method for estimating the number of clusters in a dataset. *Genome biology* 3:RESEARCH0036.

Everitt, B.S., Landau, S., Leese, M., Stahl, D. (2011). *Cluster Analysis.* UK: Wiley Series in Probability and Statistics, A John Wiley and Sons, Ltd.

Fernández, A., Gómez, S. (2008). Solving Non-uniqueness in Agglomerative Hierarchical Clustering Using Multidendrograms. *Journal of Classification* 25:43-65.

Gordon, A.D. (1999). *Classification*. London, UK: Chapman & Hall/CRC Monographs on Statistics & Applied Probability.

Hartigan, J.A. (1975). *Clustering Algorithms. A Wiley Publication in Applied Statistics, John Wiley & Sons, New York.*

Hartigan, J.A., Wong, M.A. (1979). A K-Means Clustering Algorithm. *Applied Statistics* 28:100-108.

Henry, D.B., Tolan, P.H., Gorman-Smith, D. (2005). Cluster analysis in family psychology research. *Journal of family psychology : JFP : journal of the Division of Family Psychology of the American Psychological Association (Division 43)* 19:121-132.

Hubert, L.J., Levin, J.R. (1976). A general statistical framework for assessing categorical clustering in free recall. *Psychological Bulletin* 83:1072-1080.

Jablensky, A. (2006). Subtyping schizophrenia: implications for genetic research. *Molecular psychiatry* 11:815-836.

Joyce, E.M., Roiser, J.P. (2007). Cognitive heterogeneity in schizophrenia. *Current opinion in psychiatry* 20:268-272.

Kaufman, L., Rousseeuw, P.J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. USA: New York: Wiley-Interscience.

Korver, N., Quee, P.J., Boos, H.B.M., Simons, C.J.P., de Haan, L., GPOUPinvestigators, Investigators, G. (2012). Genetic Risk and Outcome of Psychosis (GROUP), a multi-site longitudinal cohort study focused on gene-environment interaction: objectives, sample characteristics, recruitment and assessment methods. *International Journal of Methods in Psychiatric Research* 21:205-221.

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* 44:23-34.

Lin, S.H., Liu, C.M., Liu, Y.L., Shen-Jang Fann, C., Hsiao, P.C., Wu, J.Y., Hung, S.I., Chen, C.H., Wu, H.M., Jou, Y.S., Liu, S.K., Hwang, T.J., Hsieh, M.H., Chang, C.C., Yang, W.C., Lin, J.J., Chou, F.H., Faraone, S.V., Tsuang, M.T., Hwu, H.G., Chen, W.J. (2009). Clustering by neurocognition for fine mapping of the schizophrenia susceptibility loci on chromosome 6p. *Genes, brain and behavior* 8:785-794.

Mandara, J. (2003). The typological approach in child and family psychology: a review of theory, methods, and research. *Clinical child and family psychology review* 6:129-146.

Marriott, F.H. (1971). Practical problems in a method of cluster analysis. *Biometrics* 27:501-14.

Meijer, J., Simons, C.J., Quee, P.J., Verweij, K., GROUP Investigators. (2012). Cognitive alterations in patients with non-affective psychotic disorder and their unaffected siblings and parents. *Acta Psychiatrica Scandinavica* 125:66-76.

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.

Quee, P.J., Alizadeh, B.Z., Aleman, A., van den Heuvel, E.R., GROUP Investigators. (2014). Cognitive subtypes in non-affected siblings of schizophrenia patients: characteristics and profile congruency with affected family members. *Psychological medicine* 44:395-405.

Quee, P.J., van der Meer, L., Bruggeman, R., de Haan, L., Krabbendam, L., Cahn, W., Mulder, N.C., Wiersma, D., Aleman, A. (2011). Insight in psychosis: relationship with neurocognition, social cognition and clinical symptoms depends on phase of illness. *Schizophrenia bulletin* 37:29-37.

Sarle, W.S. (1983). *Cubic clustering criterion (Tech. Rep. A-108)*. USA: Cary, N.C.: SAS Institute.

Silver, H., Shmoish, M. (2008). Analysis of cognitive performance in schizophrenia patients and healthy individuals with unsupervised clustering models. *Psychiatry research* 159:167-179.

Stinissen, J., Willems, P.J., Coetsier, O., Hulsman, W.L.L. (1970). *Manual of the Dutch Edition of the WAIS.* . The Netherlands: Lisse, The Netherlands: Swets & Zeitlinger.

Tibshirani, R., Walther, G., Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society.Series B, Statistical methodology* 63:411-423.

Yan, M., Ye, K. (2007). Determining the number of clusters using the weighted gap statistic. *Biometrics* 63:1031-7.

Appendix

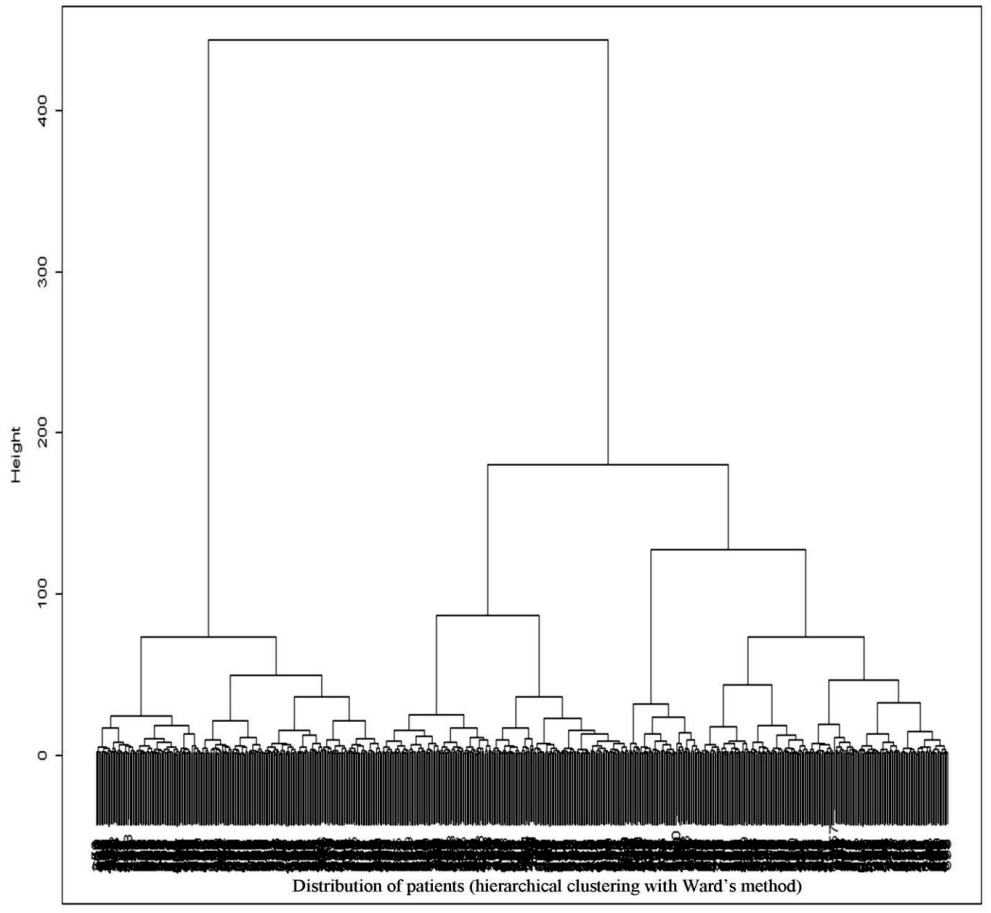

Figure A1: Dendrogram for schizophrenia patients using cognitive variables from GROUP

Table A1a: Means of cognition variables for the single cluster solution

Cluster	CPTpi	CPTsd	Block	Symbol	Calc	Info	WLT_I	WLT_D
1	-0.58	0.73	-0.49	-1.10	-0.80	-0.51	-0.92	-0.68

Table A1b. Variance-covariance matrix of cognition variables for the single cluster solution

Variables	CPTpi	CPTsd	Block	Symbol	Calc	Info	WLT_I	WLT_D
CPTpi	1.33	-0.78	0.27	0.42	0.28	0.20	0.31	0.23
CPTsd		1.62	-0.34	-0.48	-0.40	-0.27	-0.34	-0.26
Block			1.33	0.52	0.66	0.69	0.41	0.38
Symbol				1.05	0.54	0.50	0.41	0.34
Calc					1.19	0.76	0.48	0.40
Info						1.24	0.48	0.42
WLT_I							1.20	0.85
WLT_D								1.03

CPTpi: Continuous performance test index, CPTsd: Standard deviation of CPT, Block: Block design, Symbol: Digit symbol coding, Calc: Arithmetic/calculation, Info: Information, WLT_I: Word learning task immediate recall, and WLT_D: Word learning task delayed recall.

Table A2a: Means of cognition for the two clusters solution

Clusters	CPTpi	CPTsd	Block	Symbol	Calc	Info	WLT_I	WLT_D
1	-1.05	1.35	-1.25	-1.73	-1.6	-1.3	-1.59	-1.24
2	-0.19	0.22	0.12	-0.59	-0.15	0.14	-0.38	-0.23

Table A2b: Variance-covariance matrices of cognition variables for the two clusters solution

Cluster	Variables	CPTpi	CPTsd	Block	Symbol	Calc	Info	WLT_I	WLT_D
1	CPTpi	1.35	-0.63	0.08	0.24	0.06	-0.08	0.07	0.04
1	CPTsd		1.63	0.01	-0.21	-0.05	0.10	-0.04	0.01
1	Block			0.92	0.14	0.19	0.20	0.03	0.08
1	Symbol				0.64	0.17	0.10	0.09	0.08
1	Calc					0.66	0.24	0.07	0.06
1	Info						0.80	0.07	0.08
1	WLT_I							0.83	0.50
1	WLT_D								0.68
2	CPTpi	0.99	-0.48	-0.10	0.12	-0.11	-0.12	0.05	0.00
2	CPTsd		1.04	0.08	-0.13	0.06	0.15	0.02	0.03
2	Block			0.82	0.12	0.15	0.21	-0.03	0.01
2	Symbol				0.80	0.10	0.09	0.05	0.04
2	Calc					0.69	0.25	0.02	0.02
2	Info						0.68	0.04	0.04
2	WLT_I							0.83	0.59
2	WLT_D								0.86

CPTpi: Continuous performance test index, CPTsd: Standard deviation of CPT, Block: Block design, Symbol: Digit symbol coding, Calc: Arithmetic/calculation, Info: Information, WLT_I: Word learning task immediate recall, and WLT_D: Word learning task delayed recall.

Table A3a: Means of cognition variables for the three clusters solution

Clusters	CPTpi	CPTsd	Block	Symbol	Calc	Info	WLT_I	WLT_D
1	-0.25	0.30	0.35	-0.40	0.15	0.40	-0.15	0.00
2	-0.30	0.40	-0.65	-1.20	-1.0	-0.75	-1.15	-0.90
3	-1.60	0.20	-1.60	-2.10	-1.90	-1.55	-1.80	-1.40

Table A3b: Variance-covariance matrices of cognition variables for the three clusters solution

Clusters	Variables	CPTpi	CPTsd	Block	Symbol	Calc	Info	WLT_I	WLT_D
1	CPTpi	1.00	-0.55	0.00	0.18	0.00	0.00	0.09	0.00
1	CPTsd		1.21	0.00	-0.25	0.00	0.07	-0.05	-0.05
1	Block			0.56	0.07	0.05	0.07	-0.03	0.00
1	Symbol				0.81	0.06	0.00	0.00	-0.04
1	Calc					0.49	0.09	-0.06	-0.09
1	Info						0.42	0.00	0.00
1	WLT_I							0.72	0.47
1	WLT_D								0.72
2	CPTpi	0.81	-0.41	-0.14	0.09	-0.11	-0.16	0.00	0.00
2	CPTsd		0.81	0.18	-0.03	0.00	0.12	0.00	0.04
2	Block			1.00	0.00	0.04	0.14	-0.21	-0.16
2	Symbol				0.49	0.00	-0.03	-0.06	-0.06
2	Calc					0.64	0.14	-0.07	-0.06
2	Info						0.81	-0.11	-0.11
2	WLT_I							0.73	0.41
2	WLT_D								0.64
3	CPTpi	1.44	-0.23	-0.10	0.09	-0.05	-0.26	-0.06	0.00
3	CPTsd		1.56	0.21	0.00	0.09	0.32	0.12	0.11
3	Block			0.72	0.03	0.16	0.18	0.00	0.11
3	Symbol				0.56	0.14	0.16	0.00	0.10
3	Calc					0.56	0.26	0.07	0.10
3	Info						0.72	0.08	0.11
3	WLT_I							0.90	0.57
3	WLT_D								0.72

CPTpi: Continuous performance test index, CPTsd: Standard deviation of CPT, Block: Block design, Symbol: Digit symbol coding, Calc: Arithmetic/calculation, Info: Information, WLT_I: Word learning task immediate recall, and WLT_D: Word learning task delayed recall.

Table A4a: Means of cognition variables for the four clusters solution

Clusters	CPTpi	CPTsd	Block	Symbol	Calc	Info	WLT_I	WLT_D
1	-1.60	1.90	-1.70	-2.15	-2.00	-1.70	-1.95	-1.55
2	-1.05	1.42	-0.02	-1.15	-0.45	-0.10	-1.10	-0.80
3	0.00	-0.05	0.45	-0.30	0.20	0.45	0.05	0.15
4	0.10	-0.03	-1.05	-1.20	-1.30	-1.05	-1.00	-0.85

Table A4b: Variance-covariance matrices of cognition variables for the four clusters solution

Clusters	Variables	CPTpi	CPTsd	Block	Symbol	Calc	Info	WLT_I	WLT_D
1	CPTpi	1.50	-0.25	-0.10	0.10	0.10	-0.25	-0.10	0.00
1	CPTsd		1.65	0.05	-0.10	-0.10	0.20	0.15	0.10
1	Block			0.60	0.00	0.05	0.10	-0.10	0.05
1	Symbol				0.50	0.10	0.05	-0.05	0.10
1	Calc					0.35	0.15	0.05	0.00
1	Info						0.70	0.00	0.05
1	WLT_I							0.90	0.55
1	WLT_D								0.70
2	CPTpi	0.60	-0.10	0.10	0.05	-0.10	0.00	-0.10	-0.15
2	CPTsd		0.95	-0.10	-0.10	-0.10	-0.15	0.20	0.25
2	Block			0.70	0.05	0.00	0.00	-0.10	-0.15
2	Symbol				0.65	0.10	0.10	-0.10	-0.10
2	Calc					0.70	0.15	-0.05	-0.05
2	Info						0.70	-0.10	-0.05
2	WLT_I							0.60	0.35
2	WLT_D								0.65
3	CPTpi	0.80	-0.30	-0.05	0.10	-0.05	0.00	-0.10	-0.10
3	CPTsd		0.70	0.05	-0.10	0.05	0.10	0.20	0.15
3	Block			0.55	0.05	0.05	0.10	-0.05	0.05
3	Symbol				0.90	0.10	0.00	-0.10	-0.10
3	Calc					0.55	0.10	-0.10	-0.10
3	Info						0.50	0.05	0.05
3	WLT_I							0.65	0.50
3	WLT_D								0.80
4	CPTpi	0.85	-0.30	0.10	0.15	0.10	0.00	-0.15	-0.05
4	CPTsd		0.60	-0.15	-0.10	-0.15	-0.10	0.10	0.10
4	Block			1.00	0.15	-0.05	0.00	-0.15	-0.15
4	Symbol				0.50	0.05	0.05	-0.05	-0.05
4	Calc					0.65	0.20	0.00	-0.05
4	Info						0.70	-0.05	-0.05
4	WLT_I							0.85	0.50
4	WLT_D								0.70

CPTpi: Continuous performance test index, CPTsd: Standard deviation of CPT, Block: Block design, Symbol: Digit symbol coding, Calc: Arithmetic/calculation, Info: Information, WLT_I: Word learning task immediate recall, and WLT_D: Word learning task delayed recall.

Table A5a. Means of cognition variables for the five clusters solution

Clusters	CPTpi	CPTsd	Block	Symbol	Calc	Info	WLT_I	WLT_D
1	-1.85	2.05	-1.70	-2.25	-2.05	-1.65	-2.10	-1.70
2	-1.10	1.80	-0.20	-1.30	-0.65	-0.25	-1.00	-0.65
3	-0.25	0.20	0.40	-0.30	0.10	0.45	0.45	0.60
4	0.10	-0.25	0.30	-0.60	-0.10	0.10	-1.05	-1.00
5	-0.15	0.20	-1.35	-1.35	-1.55	-1.35	-1.15	-0.90

Table A5b: Variance-covariance matrices of cognition variables for the five clusters solution

Clusters	Variables	CPTpi	CPTsd	Block	Symbol	Calc	Info	WLT_I	WLT_D
1	CPTpi	1.50	-0.15	-0.10	0.10	0.05	-0.25	-0.15	-0.10
1	CPTsd		1.70	-0.05	-0.10	-0.05	0.15	0.25	0.10
1	Block			0.65	0.05	0.05	0.10	-0.10	-0.05
1	Symbol				0.50	0.10	0.10	-0.05	0.05
1	Calc					0.35	0.15	0.00	-0.05
1	Info						0.65	0.00	-0.05
1	WLT_I							0.85	0.50
1	WLT_D								0.60
2	CPTpi	0.55	-0.10	0.10	0.10	-0.15	-0.05	0.00	-0.15
2	CPTsd		0.85	0.00	0.10	0.05	-0.10	0.10	0.10
2	Block			0.70	0.05	0.05	0.00	-0.10	-0.10
2	Symbol				0.65	0.10	0.05	0.05	0.10
2	Calc					0.80	0.15	0.05	0.00
2	Info						0.70	0.00	0.00
2	WLT_I							0.55	0.30
2	WLT_D								0.60
3	CPTpi	0.90	-0.25	-0.15	0.15	-0.05	0.10	0.10	0.00
3	CPTsd		0.70	-0.05	-0.15	0.10	0.10	0.10	0.00
3	Block			0.65	0.10	0.15	0.15	-0.10	0.05
3	Symbol				0.90	0.10	0.15	-0.10	-0.10
3	Calc					0.55	0.15	-0.05	-0.05
3	Info						0.55	0.05	-0.10
3	WLT_I							0.40	0.20
3	WLT_D								0.45
4	CPTpi	0.95	-0.35	-0.05	0.05	-0.15	-0.15	0.15	0.25
4	CPTsd		0.50	0.10	0.00	0.05	0.15	-0.15	-0.20
4	Block			0.65	0.10	-0.05	-0.05	-0.05	-0.10
4	Symbol				0.75	0.05	0.05	-0.10	-0.10
4	Calc					0.60	0.20	-0.05	-0.10
4	Info						0.65	-0.05	0.10
4	WLT_I							0.50	0.20
4	WLT_D								0.50
5	CPTpi	0.75	-0.25	-0.05	0.10	0.05	-0.05	-0.05	0.05
5	CPTsd		0.60	-0.10	-0.10	-0.10	-0.10	-0.10	0.00
5	Block			0.70	0.15	-0.10	0.00	-0.15	0.00
5	Symbol				0.55	0.15	-0.05	0.15	0.10
5	Calc					0.60	0.10	0.00	-0.10
5	Info						0.65	0.00	0.05
5	WLT_I							0.90	0.55
5	WLT_D								0.70

CPTpi: Continuous performance test index, CPTsd: Standard deviation of CPT, Block: Block design, Symbol: Digit symbol coding, Calc: Arithmetic/calculation, Info: Information, WLT_I: Word learning task immediate recall, and WLT_D: Word learning task delayed recall.