An *a contrario* approach to hierarchical clustering validity assessment

Frédéric CAO*, Julie DELON†, Agnès DESOLNEUX‡, Pablo MUSÉ†, Frédéric SUR†

28th September 2004

Abstract

In this paper we present a method to detect natural groups in a data set, based on hierarchical clustering. A measure of the meaningfulness of clusters, derived from a *background model* assuming no class structure in the data, provides a way to compare clusters, and leads to a cluster validity criterion. This criterion is applied to every cluster in the nested structure. While all clusters passing the validity test are meaningful in themselves, the set of all of them will probably provide a redundant data representation. By selecting a subset of the meaningful clusters, a good data representation, which also discards outliers, can be achieved. The strategy we propose combines a new merging criterion (also derived from the *background model*) with a selection of local maxima of the meaningfulness with respect to inclusion, in the nested hierarchical structure.

1 Introduction

The unsupervised classification of data into groups is commonly referred as clustering. The aim of clustering is to discover structure in a data set, by dividing it into its "natural" groups, so that data items within each cluster are more closely related to one another than to data items assigned to different clusters. This paper considers clustering problems in which data are represented as patterns (D-dimensional feature vectors) in a bounded subset E of $\mathbb{R}^D$. Hundreds of methods for finding clusters in data have been reported in the literature. Most of the clustering methods are either partitional, either hierarchical methods [10]. Partitional methods identify the partition that optimizes a clustering criterion (*e.g.* minimum variance partition). Hierarchical methods produce a hierarchical representation, in which each level of the hierarchy is itself a partition of the data. Hierarchial algorithms can be agglomerative or divisive; in the agglomerative case, the hierarchy is built by recursively merging the two closest clusters in the sense of a predefined proximity measure.

In general, the actual number of clusters which are present in the data is unknown. There may be possibly none. Even if this number is known in advance, the clusters produced by a particular clustering algorithm may not reveal the underlying structure of data. Consequently, cluster validation techniques have to be applied. There are different validation approaches [5, 8], depending on the amount of prior information on the data. Internal validation tests consist in determining if the structure is intrinsically adapted to the data with no other information than the data themselves. In this work we will focus on the assessment and internal validation of clusters obtained by agglomerative hierarchical procedures. We will not discuss the choice of the clustering algorithm, since no method performs universally better than the other ones; depending on the cluster proximity measure and the clusters' shape, different methods of clustering can be more or less successful [6, 9, 10].

*IRISA, INRIA Rennes, Campus Universitaire de Beaulieu, 35042 Rennes Cedex, France.
E-mail: fcao@irisa.fr
†CMLA, ENS de Cachan, 61 avenue du Président Wilson, 94235 Cachan Cedex, France.
E-mail: {delon,muse,sur}@cmla.ens-cachan.fr
‡MAP5 - UFR Maths-Info Université Paris 5, 45 rue des Saints-Pères, 75270 Paris cedex 06, France.
E-mail: desolneux@math-info.univ-paris5.fr

Hierarchical clustering can be represented by a dendrogram, where each node is a cluster in the hierarchical structure. Determining the partition that best fits the data amounts to cutting the dendrogram at a certain level. Rules for deciding the appropriate level are known as stopping rules, because they can be seen as stopping the merging process. Typically, an index based on the cohesion and/or the separation of clusters at each level is computed, and the level optimizing this index is chosen. In general, stopping rules proposed in the literature are heuristic, *ad hoc* procedures, and lack of theoretical foundation [13]. This common approach for hierarchical clustering validation suffers from two main problems. First, since these indices have no absolute meaning, nothing ensures that clusters at the optimum level actually correspond to the "natural" clusters. Second, noise or outliers cannot be rejected since every pattern is assigned to a cluster.

These considerations motivate the work which is presented in this paper. We propose a fully unsupervised method to determine a set of relevant clusters within the hierarchical clustering structure, that suitably represents the data set. The approach is as follows:

1. Following [4], we define a measure of meaningfulness of a group of patterns: the "number of false alarms" (NFA). The lower the NFA is, the more the group is meaningful. This measure not only enables to rank the clusters in the hierarchical clustering according to their relevance, but also to decide if a cluster is valid (*i.e.* represents a "natural" group in which outliers have been discarded) or not.

   The definition of the NFA is based on the Helmholtz principle, which states that if an observed arrangement of objects is highly unlikely, the occurrence of such arrangement is significant and the objects should be grouped together into a single structure. This perceptual organization principle, also known as the principle of common cause or of the coincidental explanation, was first stated in computer vision by Lowe [12]. Roughly speaking, a cluster or group of patterns is significant if its density is so high that such an arrangement is unlikely to be due to randomness. In other words, there must be a better explanation for the observed cluster than randomness: the formation of causal relations. This qualitative definition can be made more precise in a hypothesis testing framework. In his founding paper, Bock [1] presents several significance tests for distinguishing between a null hypothesis modeling the lack of structure in the data, and an alternative hypothesis involving clustering. Hence, clusters are detected *a contrario* to a null hypothesis or background model. The problem with these significance tests is that there is no intuitive rule to fix their significance level. Desolneux *et al.* [4] propose to bypass this difficulty by controlling the expected number of false alarms (the NFA).

2. All clusters showing a small enough NFA are meaningful in themselves, when considered as isolated entities. However, since they are embedded in a hierarchical structure, the set of all of them will probably provide a very redundant representation. In order to achieve a good data representation, a disjoint subset of meaningful clusters has to be selected. We propose a selection strategy which combines the ordering established by the NFA, the inclusion relations in the hierarchical structure, and a merging criterion for deciding whether two clusters should be merged or not. The proposed merging criterion is based on the same *a contrario* approach: two clusters G_1 and G_2 should be merged if, under the background model, their union is less likely to be observed than G_1 and G_2 separately.

The plan of this article is as follows. In section 2 we recall classical issues in clustering validation and stopping rules for hierarchical classifications. In section 3 we present the notion of meaningful group of patterns, based on the work by Desolneux *et al.* [4]. We then propose a new merging criterion. This criterion, combined with the selection of local maxima of the NFA with respect to inclusion, leads to a reasonable set of clusters that well represents the data. This is confirmed by the experiments on synthetic and real data that we present in section 4. We conclude in section 5.

2 Cluster validity and stopping rules

The large variety of clustering methods that have been proposed in the recent past has been followed by an increasing interest in clustering validation methods. In [8], a comprehensive study of these techniques is presented. Classical issues in cluster validity analysis are the assessment of individual cluster validity, and the assessment of a whole partition. In what follows we briefly summarize these two issues.

2.1 Partition validity assessment

A relevant question to address in order to assess the validity of a partition, is deriving the number of clusters [5], that we denote by c. Notice however that by solving this problem, it cannot be ensured that the c clusters are valid clusters. The most common approach to decide how many clusters are best consists in finding partitions for $c = 1, \ldots, c_{max}$ and optimizing a measure $A(c)$ of partition adequacy, which is usually based on the within-cluster and between-cluster variability. When applied to hierarchical clustering methods, these techniques are known as *global stopping rules*, because the choice of c can be seen as stopping the merging process (in the agglomerative case) at a certain level of the dendrogram. A popular stopping rule for assessing partitions was proposed by Calinski and Harabasz in [2]. They define $A(c)$ as the ratio between the total within-cluster sum of squared distances to the centroids, and the total between-cluster sum of squared distances. Since the index is based on the sum of squares criterion, it has a tendency to partition the data into hyperspherical shaped clusters, having roughly equal numbers of patterns [8]. We can summarize the main drawback of global stopping rules by quoting Bock [1]: *"Some care is needed when applying any test for clustering, bearing in mind that different types of clusters may be present simultaneously in the data, and that the number of clusters is, in some sense, dependent on the intended level of information compression. Thus, a global application of a cluster test to a large or high-dimensional data set will not be advisable in most cases. However, a "local" application (...) to a specific part of the data will often be useful for providing evidence for or against a prospective clustering tendency".*

In agglomerative hierarchical classifications, local approaches can be conceived by deciding if two clusters should be merged or not. Usually, the merging process is continued until it is decided, for the first time, that two clusters should not be aggregated (see for instance Duda and Hart's rule, [6] chapter 10.10). These strategies, referred to as *local stopping rules*, present an inherent drawback: a global decision is inferred from a too much local information.

Comparative studies of stopping rules are presented by Milligan and Cooper in [13], or Dubes in [5]. A major conclusion of their works is that the majority of the described stopping rules are based on heuristics and lack of theoretical foundation. Those derived from rigorous statistical techniques, assume in general hypotheses on the data which are unrealistic in most real applications (*e.g.* multivariate normal distribution for the patterns).

2.2 Validity assessment of individual clusters

Now we are concerned with the problem of deciding, among the candidate clusters furnished by the clustering procedure, which are the ones that correspond to "natural" clusters. But what does a "natural" cluster look like? As pointed out by Gordon [8], it may be difficult to specify a relevant definition of ideal cluster for a particular data set. However, we can think of clusters as some structure in the data. Clustered data reveal then structure, that is perceived as opposite to a complete absence of structure. Thus, in order to decide whether the clusters we have found are significant, we can proceed by comparing our actual data with some appropriate random distribution. This leads to a general methodology for cluster validity analysis, based on the statistical approach of hypothesis testing [1, 7, 8], that we call *a contrario* strategy. Following Bock [1], this framework consists in:

1. Designing a null hypothesis $\mathcal{H}$ for the absence of class structure in the data (a *background model*, or *null model*), meaning that patterns are sampled from a "homogeneous" population. Then, "heterogeneity" or "clustering structure" are involved in the alternative hypothesis $\mathcal{A}$.

2. Defining a test statistic, which will be used as validity index to discriminate between $\mathcal{H}$ and $\mathcal{A}$.

3. Rejecting the null hypothesis $\mathcal{H}$ in favor of $\mathcal{A}$ if, for a given significance level (error probability) α, the test statistic of the observed data exceeds the corresponding critical value c_α.

This general framework can be adapted for assessing the validity of individual clusters. A popular approach within this framework is Monte-Carlo validation, which is described in [8]. Assume one wants to assess the validity of an observed cluster G having n patterns, in a data set having M patterns. In the Monte-Carlo validation method, data sets of M patterns are simulated under the background model, and classified using the same clustering procedure that was used to classify the original data. The test statistic is computed for those clusters having n patterns, and the distribution of the test statistic is estimated. Then, using the value of the test statistic of G, one can compute the significance level of rejecting $\mathcal{H}$. Two popular test statistics are the maximum F test and the U statistic (see Bock [1] and Gordon [8]).

We have not addressed the choice of the null model yet. The specification of appropriate null models for data is the subject of the study presented in [7]. These models, which specify the distribution of patterns in the absence of structure in the data, can be of two types:

– *Standard (data-independent) null models*. Two well known standard null models are the *Poisson model* and the *Unimodal model* [1]. The main problem with the Poisson model is the choice of the region R within which patterns are uniformly distributed (standard choices for normalized data are the unit hypercube and the unit hypersphere). The Unimodal model assumes that the joint distribution of the variables describing the patterns is unimodal, but the choice of the distribution may not be easy.

– *Data-influenced null models*. Here the data are used to influence the specification of the null model. Examples of these null models are the Poisson model where R is chosen to be the convex hull of the data set, or the *Ellipsoidal model*, which is a multivariate normal distribution, whose mean and covariance matrix are given by the data set.

In [7], Gordon concludes that the results of the tests considerably depend on the choice of the null model, and that, in general, the results based on data-influenced null models are more relevant than those obtained using a standard null model.

In the following section we propose a method to detect valid clusters from an agglomerative hierarchical classification, that combines an individual cluster validity method and a merging criterion. The first step consists in deciding, *a contrario* to a data-influenced background model, whether a cluster is valid or not. All clusters in the hierarchical structure are examined. While all clusters passing the validity test are meaningful in themselves, the set of all of them does not necessarily reveal the structure of the data set. However, by selecting a subset of the meaningful clusters, a good data representation can be achieved. Hence, in the second step such a selection is performed, by means of a new merging criterion, also derived from the *background model*. Unlike the classical hypothesis testing methods presented in this section, the proposed method does not require to *a priori* fix an arbitrary significance level for deciding the validity of clusters.

3 Meaningful clusters

3.1 *A contrario* definition of meaningful groups

Let us suppose that a distribution function p (which we call *background law*) is defined over the pattern space E. In the following, for every subset R of E, we will denote by $p(R)$ the probability $\int_E \mathbf{1}_R dp$ (where dp is the probability measure associated to p). This distribution function enables us to define a *background process*.

Definition 1 *We call* background process *the stochastic process $(X_i)_{i\in\{1\ldots M\}}$, where the M random vectors X_i (whose values are in E) are independent and identically distributed, with distribution function p.*

Now, consider a set of M patterns. The cornerstone of the proposed methodology consists in the following assumption:

(A) *the M-tuple $(x_i)_{i\in\{1\ldots M\}}$ made of the M observed patterns is a realization of the background process.*

The definition of the background law is problem specific. In general, it is given *a priori*, or can be empirically estimated over the data.

Having a background model, we are in position to evaluate the probability of a given cluster of patterns as a realization of the background process. Hence, we are able to detect relevant clusters by Helmholtz principle: those clusters being unlikely to be observed by chance will be considered as meaningful groups. Let us give an example to illustrate this idea. In Figure 1, we display two 2-D representations of a group of objects, as points in the (diameter, elongation) space, and in the space of coordinates, respectively. These two features have no reason to show such a special arrangement, one could therefore expect that, in general, they are uniformly distributed. Consequently, the "high density" cluster we observe at the top left corner in the first 2-D representation reveals a conspicuous coincidence. Indeed, the probability of it being a realization of the background process should be very low, and one would expect it to be an exception to randomness. This cluster corresponds to the set of all the entire circles in the image. On the contrary, no cluster can be seen in the second representation, what means that we should not find groups with respect to these two features.

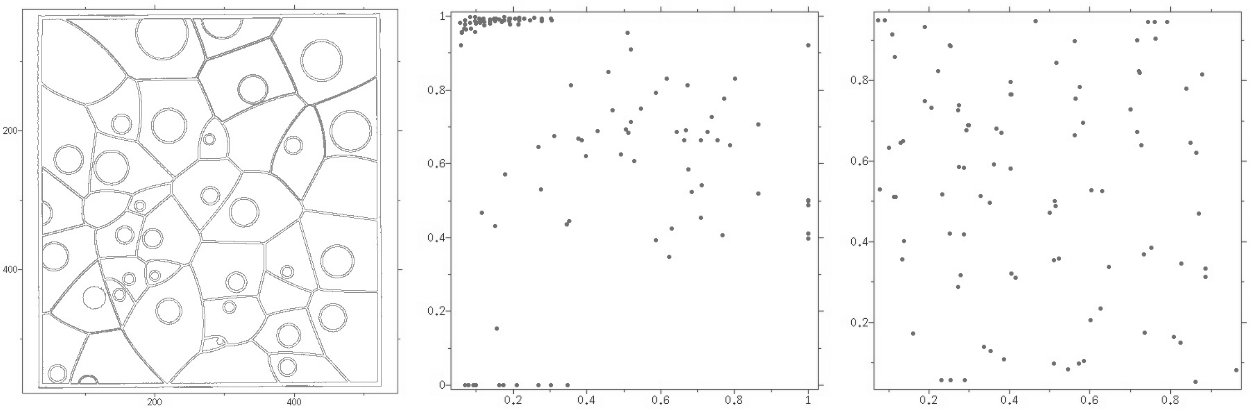


Figure 1: Left: image of geometrical objects. Center: 2-D representation of these objects in the space (diameter, elongation). Right: 2-D representation of these objects in the space of objects' barycenter coordinates.

Let us make things more precise. For any region R in the feature space, we know how to compute the "region probability" $p(R)$, the probability that a pattern generated by the background process falls in R. Then, since patterns are mutually independent, the probability that R contains at least k patterns out of M under the background model is given by the tail of the binomial probability distribution

$$\mathcal{B}(M,k,p(R)) = \sum_{i=k}^{M} b(M,i,p(R)),$$

where

$$\forall\, i \in \{0,\dots,M\},\ \ b(M,i,p(R)) = \binom{M}{i} p(R)^i (1-p(R))^{M-i}.$$

For obvious algorithmic reasons, we cannot explicitly test all regions R in E. Hence, the next step prior to detection is to define a set $\mathcal{R}$ of reasonable region candidates. For the sake of simplicity, $\mathcal{R}$ will consists in a finite set of hyper-rectangles. We may consider one of the most basic possibilities, which is the set of all hyper-rectangles of different sizes within a given quantization grid of the (bounded) feature space E. Let us denote by $\#\mathcal{R}$ the cardinality of $\mathcal{R}$. If each dimension of E is divided into L bins,

$$\#\mathcal{R} = \left(\frac{L(L+1)}{2}\right)^D.$$

Definition 2 (ε-meaningful region) *We say that a region R of $\mathcal{R}$, containing k patterns, is ε-meaningful if*

$$\#\mathcal{R} \cdot \mathcal{B}(M,k,p(R)) \leqslant \varepsilon.$$

Proposition 1 *The expected number of ε-meaningful regions in $\mathcal{R}$ is less than ε.*

Proof: Let us denote by χ_R the binary random variable equal to 1 if the region R in $\mathcal{R}$ is ε-meaningful and 0 otherwise. Let $S = \sum_{R\in\mathcal{R}} \chi_R$ be the random variable representing the number of ε-meaningful regions in $\mathcal{R}$.
By linearity, the expectation of S is $\mathbb{E}[S] = \sum_{R\in\mathcal{R}} \mathbb{E}[\chi_R]$. Hence, since χ_R is a Bernoulli variable,

$$\mathbb{E}[S] = \sum_{R\in\mathcal{R}} \Pr(\chi_R = 1).$$

Let us denote by $k^*(\varepsilon)$ the minimum number of points in R such that R is ε-meaningful:

$$k^*(\varepsilon) = \min\left\{k \in \mathbb{N},\ \mathcal{B}(M, k, p(R))\right\} \leqslant \frac{\varepsilon}{\#\mathcal{R}}.$$

This number is well defined because $\mathcal{B}(M, k, p(R))$ is a decreasing function of k. It follows that

$$\Pr(\chi_R = 1) = \Pr\left(k \geqslant k^*(\varepsilon)\right) = \mathcal{B}(M, k^*(\varepsilon), p(R)) \leqslant \frac{\varepsilon}{\#\mathcal{R}}.$$

Thus,

$$\mathbb{E}[S] \leqslant \sum_{R\in\mathcal{R}} \frac{\varepsilon}{\#\mathcal{R}},$$

yielding $\mathbb{E}[S] \leqslant \varepsilon$. ■

Remark 1 The key point is that we control the expectation of S. Since dependencies between random variables χ_R are unknown, we are not able to compute the probability law of S. Nevertheless, linearity still allows to compute the expectation.

The following definition provides a quality measure for a cluster or group of patterns in the feature space.

Definition 3 (Number of False Alarms) *Given a group G of k meaningful patterns among M, we call number of false alarms of G the number*

$$NFA_g(G) = \#\mathcal{R} \cdot \mathcal{B}(M, k, p(R)),$$

where R is the smallest region of $\mathcal{R}$ containing all k patterns.

The number of false alarms of G is a measure of how likely it is that a group having at least k meaningful patterns and a region probability $p(R)$, was generated "by chance", as a realization of the background process. The lower is $NFA_g(G)$, the less likely G is generated by the background, and hence, the more meaningful is G. In other words, if $NFA_g(G)$ is very small, elements in G certainly violate assumptions in (A), leading to an *a contrario* detection. Notice also, from Proposition 1, that the only parameter that controls detections is ε. This provides a handy way to control false detections: if we want to detect on the average at most one "non relevant cluster" among all ε-meaningful clusters, we just set $\varepsilon = 1$. From now on, we refer to 1-meaningful groups as "meaningful groups".

3.2 Finding a suitable data representation: merging condition and maximality criterion

In section 3.1 we have defined the notion of *meaningful groups of patterns*, and proposed to restrict the space of tests to the smallest regions of $\mathcal{R}$ containing clusters from the dendrogram. While each meaningful group we detect will be relevant by itself, the whole set of meaningful groups will probably exhibit high redundancy in the sense that we will get many nested meaningful groups. In this section we describe a strategy to reduce this redundancy by combining the inclusion tree given by the hierarchical clustering procedure, and the measure of meaningfulness given by NFA_g.

Let us start this discussion by the following issue. At each step of the hierarchical clustering procedure, two clusters are merged. This merging is not necessarily a better data representation than the two separate clusters. By using the complete dendrogram (that we denote by $\mathcal{D}$) of $2M-1$ clusters, we can decide *a posteriori* whether pairs of clusters should be merged or not. Let us denote by G, G_1 and G_2 the groups of patterns corresponding respectively to a node and its two children nodes in $\mathcal{D}$. Roughly speaking, we will accept merging if, under the *a contrario* model, the expected number of groups like G we would observe is smaller than the one of observing groups like G_1, G_2, or the pair G_1 and G_2. In this case, we will say that G satisfies the merging criterion. Before giving a precise merging criterion definition, let us define $NFA_{gg}(G_1, G_2)$, the number of false alarms of the pair (G_1, G_2):

$$NFA_{gg}(G_1, G_2) = \frac{\#\mathcal{R}\,(\#\mathcal{R}-1)}{2} \sum_{i=k_1}^{M} \sum_{j=k_2}^{M-i} \binom{M}{i,j} {p_1}^i {p_2}^j (1-p_1-p_2)^{M-i-j}, \tag{1}$$

where k_1 and k_2 are the number of elements in G_1 and G_2, and p_1 and p_2 their associated region probabilities. $\binom{M}{i,j}$ denotes the trinomial coefficient. $NFA_{gg}(G_1, G_2)$ is an estimate of the number of occurrences, under the *a contrario* model, of the event $\mathcal{E}$: "there are two disjoint groups A and B, with region probabilities p_1 and p_2 (resp.), containing at least k_1 and k_2 patterns (resp.) among M". Indeed, $\#\mathcal{R}\,(\#\mathcal{R}-1)/2$ is the number of pairs of clusters of $\mathcal{R}$, and the probability of event $\mathcal{E}$ is given by the joint tail of the trinomial probability distribution.

Definition 4 (Merging condition) *Let G, G_1 and G_2 be the groups of patterns corresponding respectively to a node and its two children nodes in $\mathcal{D}$. We say that G satisfies the merging condition if both following inequalities hold:*

$$NFA_g(G) < \min\{NFA_g(G_1), NFA_g(G_2)\}, \tag{2}$$

$$NFA_g(G) \leqslant NFA_{gg}(G_1, G_2). \tag{3}$$

Eq. (2) corresponds to the condition that merging cannot be suitable if one of the child nodes is more meaningful than the father. Eq. (3) means that for G to be valid, it is necessary that its number of false alarms is lower than the number of false alarms of the pair (G_1, G_2). The following lemma leads to a necessary condition for merging.

Lemma 1 *For every k_1 and k_2 in $\{0, \dots, M\}$, for every p_1 and p_2 in $[0, 1]$, the following inequality stands:*

$$\sum_{i=k_1}^{M} \sum_{j=k_2}^{M-i} \binom{M}{i,j} {p_1}^i {p_2}^j (1-p_1-p_2)^{M-i-j} \leqslant \mathcal{B}(M, k_1, p_1) \cdot \mathcal{B}(M, k_2, p_2). \tag{4}$$

Proving lemma 1 directly by calculation is not trivial. Inequality (4) is a consequence of the *negative dependence* amongst random variables $\#A$ and $\#B$, the number of patterns in two random clusters A and B. Intuitively, this dependence (which is obvious because of the condition $\#A + \#B \leqslant M$) is negative in the sense that, if $\#A$ is "large", $\#B$ is less likely to be "large". The notion of negative dependence was introduced in [11], where the authors also prove that multinomial distributions are negatively associated.

Proposition 2 *If G satisfies the merging condition, then $NFA_g(G) < \frac{1}{2} \cdot NFA_g(G_1) \cdot NFA_g(G_2)$.*

Proof: The result follows immediately from (1), definition 4 and lemma 1. ■

Proposition 2 is useful from the computational viewpoint, since in many cases one can avoid computing the tail of the trinomial distribution, by "filtering" those clusters that do not pass the necessary condition.

The union of all meaningful groups which satisfy the merging condition is not disjoint, and consequently, it does not provide a compact representation of data. This consideration motivates the following definition.

Definition 5 (Maximal ε-meaningful group) *We say that a group of patterns G is a maximal ε-meaningful group if and only if:*

1. $NFA_g(G) \leqslant \varepsilon$,
2. G *satisfies the merging condition (*cf *definition 4),*
3. *for all descendant F that satisfies the merging condition,* $NFA_g(F) > NFA_g(G)$,
4. *for all ancestor F that satisfies the merging condition,* $NFA_g(F) \geqslant NFA_g(G)$.

Imposing items 3 and 4 ensures that two different maximal meaningful groups are disjoint. Indeed, since they are nodes of the dendrogram, they cannot overlap, and because of points 3 and 4, they cannot be nested. Hence, maximal meaningful groups define a set of groups on the data, which is optimal in the sense that these groups are maxima of meaningfulness with respect to inclusion, and where outliers have been automatically rejected.

3.3 Influence of the merging condition

Let us illustrate the critical importance of the merging condition with two simple examples. The top of Figure 2 shows a configuration of 150 points, distributed on $[0, 1]^2$, and naturally grouped in two clusters G_1 and G_2. In the hierarchical structure, G_1 and G_2 are the children of $G = G_1 \cup G_2$. All three nodes are obviously meaningful, since their corresponding NFA_g are less than 1 (and are all the lower since they represent very concentrated clusters of points). Their NFA_g are also lower than the other groups in the dendogram. Now, the following situation occurs:

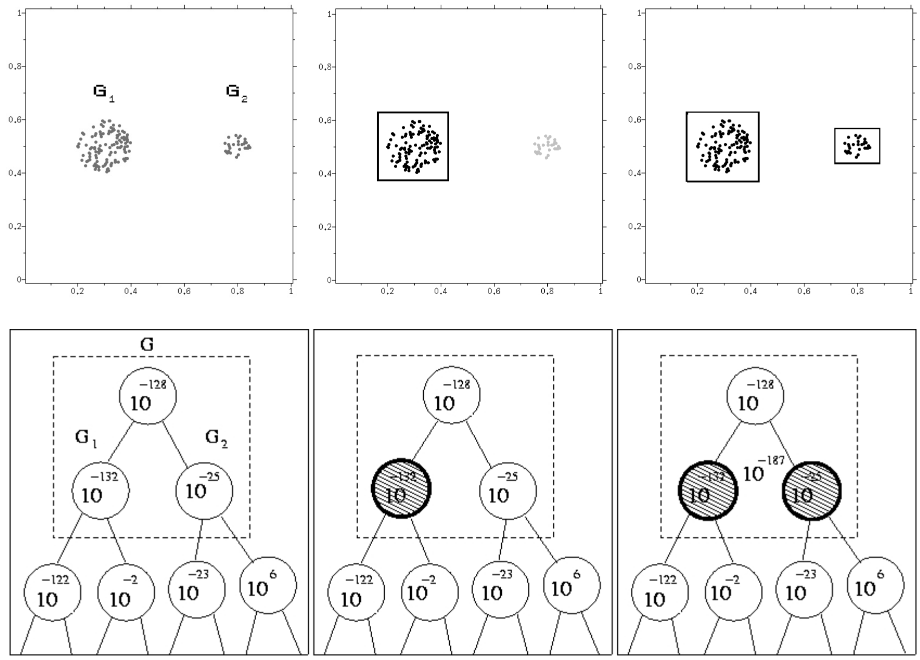
$$NFA_g(G_1) \; < \; NFA_g(G) \; < \; NFA_g(G_2)$$

If we define the maximality in the tree without taking into account the merging condition (in this case, a group is maximal meaningful if it is more meaningful than all the other groups in its branch), then G_1 will be considered as only maximal meaningful cluster of the tree. Neither G nor G_2 will be selected, since $NFA_g(G_1) < NFA_g(G)$ and $NFA_g(G) < NFA_g(G_2)$. Now, it is clear that G_2 represents an informative part of the data that should be kept. It happens that $NFA_{gg}(G_1, G_2) < NFA_g(G)$, which means that G does not satisfy the merging condition. Thus, if we add the merging condition in the maximality definition, G will not be taken into account in the maximality search. Consequently, G_2 is considered as a maximal group in its branch, and is also selected as a maximal meaningful group.

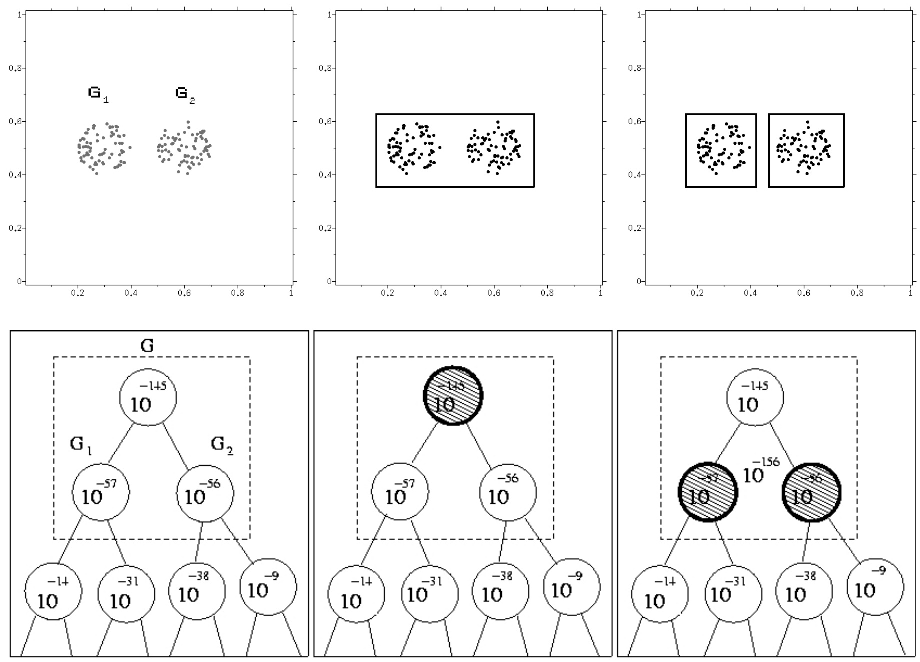
The second part of Figure 2 shows another typical configuration where the merging criterion is needed. In this example, the union G of two clusters G_1 and G_2 is more meaningful than each separate cluster. This implies that if we do not take into account the merging condition in the maximality definition, G will be the only maximal meaningful group selected. It would be coherent to represent the data by G if G_1 and G_2 were mixed enough. However, we see on the figure that these two clusters are clearly separated by an empty space. It seems much more relevant here to represent the set of points by two separate clusters than by G. The merging condition confirms this expectation, since $NFA_{gg}(G_1, G_2) < NFA_g(G)$.

4 Some experiments

In this section, in order to illustrate the proposed method, we present some experiments based on synthetic data. In all the experiments, we use the single-linkage algorithm [10, 6] (the nearest-neighbor points determine the nearest subsets) to build the hierarchical clustering structure. The set $\mathcal{R}$ of region candidates is the set of all hyper-rectangles supported by the regular grid dividing each dimension of the feature space in $L = 100$ bins. Thus, in D-dimensions, the number of tests is $(L(L + 1)/2)^D$.



(a) *Left: original configuration. Middle: the node selected without merging criterion; this maximality criterion yields some relevant misses, such as the cluster having* $NFA_g = 10^{-25}$*. Right: by combining merging and maximality criteria, both clusters are selected.*



(b) *Left: original configuration. Middle: the node selected without merging condition. Right: selected nodes obtained by combining merging and maximality criteria. The merging criterion implies that the selected pair of nodes is more significant than its ancestor. Indeed,* $NFA_{gg}(G_1, G_2) = 10^{-156} < 10^{-145} = NFA_g(G)$*.*

Figure 2: Influence of the merging condition. Each subfigure shows a configuration of points, and a piece of the corresponding dendrogram, with the selection of maximal meaningful groups, depicted in gray. The numbers in each node correspond to the NFA_g of its associated clusters. The one between two nodes is the corresponding pair NFA_{gg}.

4.1 Synthetic data

2-D experiments. Figure 3 shows a realization of a process of 950 independent, uniformly distributed points in $[0,1]^2$. Then, 25 points are added "by hand" in the neighborhood of $(0.4, 0.4)$, and 25 other points in the neighborhood of $(0.7, 0.7)$. If we use as background model the uniform distribution on $[0,1]^2$, two maximal meaningful groups are found, corresponding approximately to the handmade clusters. They contain the added points and the points of the uniform background lying in the same area. The left group is composed of 32 points and its NFA is 10^{-8}. The second one is composed of 36 points and its NFA is 10^{-12}. As expected, no maximal meaningful group is found in the uniform bakground.

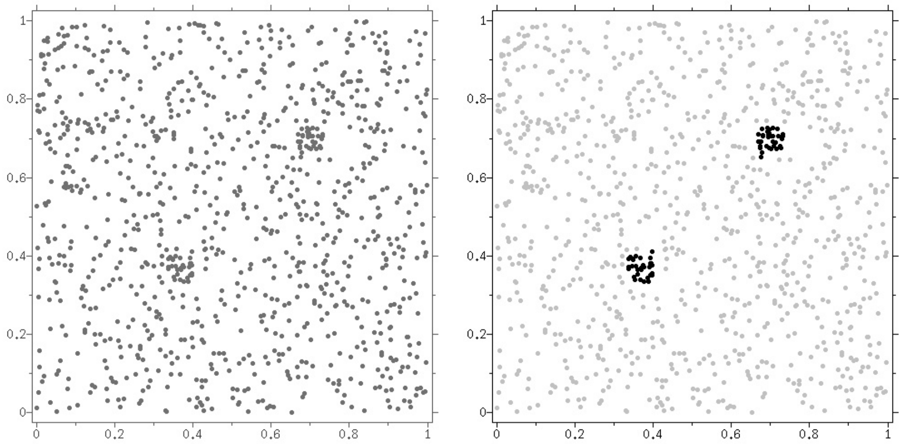


Figure 3: Left: distribution of 1000 points on $[0,1]^2$. Among them, 950 are uniformly distributed, and 50 are put "by hand" around $(0.4, 0.4)$ and $(0.7, 0.7)$. Right: two maximal meaningful groups are detected. The NFA of the right one is 10^{-11}, and 10^{-8} for the other one.

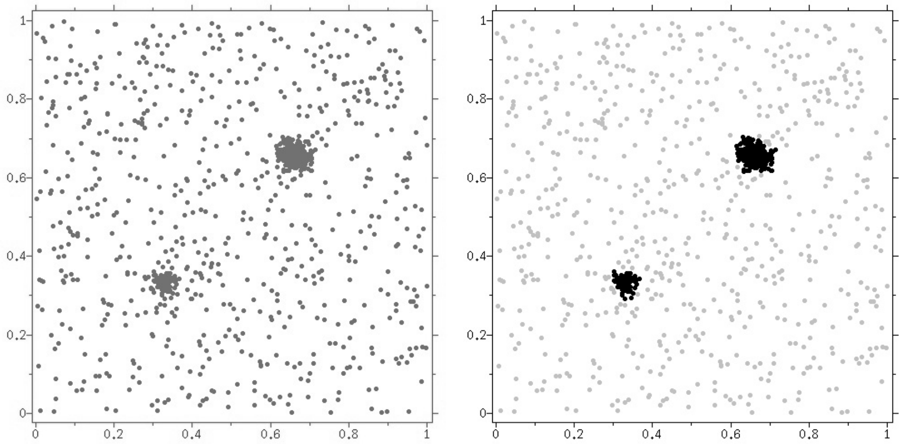


Figure 4: Left: distribution of 1000 points on $[0,1]^2$. The distribution law used here is a linear combination of the uniform distribution and of two gaussian laws. Right: two maximal meaningful groups are detected. Their NFA are 10^{-270}, for the largest one, and 10^{-51} for the other one.

Figure 4 shows another example, where 1000 points are distributed on $[0,1]^2$, according to a distribution law, which is a linear combination of the uniform distribution and two gaussian laws (with different means and standard deviations). If we use as background model the uniform distribution on $[0,1]^2$, two maximal meaningful groups are found, corresponding approximately to the modes of the gaussians. The first one, corresponding to the gaussian of largest coefficient, contains 251 points, and its NFA is 10^{-270}. The second one contains 71 points and its NFA is 10^{-51}. As we could expect, the groups are much more meaningful in this experiment. Notice that this example shows that we can detect gaussian modes without any *a priori* parametric model.

4.2 Object grouping based on elementary features.

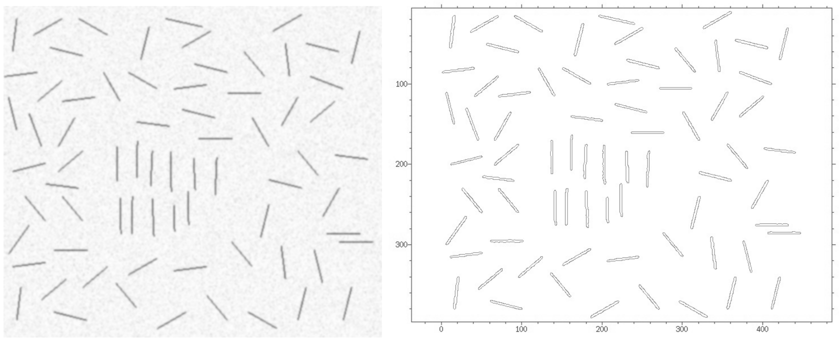
Grouping phenomena are essential in human perception, since they are responsible for the organization of information. In vision, grouping has been especially studied by Gestalt psychologists like Wertheimer [14]. In these experiments, we aim at detecting the groups of objects in an image, that share some elementary features. The objects boundaries are extracted as some contrasted level lines in the image (see [3] for a full description of this extraction process). Once these objects are detected, say $O_1, \dots O_M$, we can compute for each of them a list of D features (grey level, position , orientation, etc...). If k objects among M have one or several features in common, we wonder if it is happening by chance or if it is enough to group them. In order to answer this question, we choose to normalize each feature in $[0,1]$, and to represent the objects as points in the feature space $[0,1]^D$. We can then look for maximal meaningful groups in this space.

Segments. In this first example, groups are perceived as a result of the collaboration between two different features. Figure 5 shows 71 straight segments with different orientations, almost uniformly distributed in position. As expected, no maximal meaningful cluster is detected in the space of position coordinates. If we choose the orientation as only feature ($D = 1$), 3 maximal meaningful groups are detected, corresponding to the most represented orientations. None of these clusters exhibits a very low NFA, and we can hardly perceive them. The other orientations do not create meaningful groups because they are not enough represented.

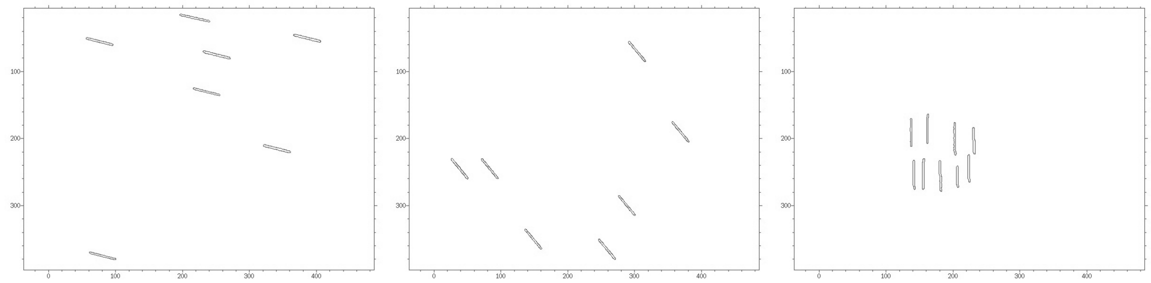
Now, let us see what happens when considering two features ($D = 2$). In the space (x-coordinate, orientation), only one maximal meaningful cluster is found. This cluster corresponds to the group G of 11 central vertical segments. Its NFA is equal to 10^{-4}. Here, the conjonction of qualities reveals a much more meaningful coincidence than separated features. In the space (y-coordinate, orientation), two maximal meaningful clusters are found. They correspond to the two lines of segments composing G. The first one, say G_1, contains 6 segments and its NFA is 0.018. The second one, say G_2 contains 5 segments and its NFA is 0.0047. The role of the merging criterion is decisive here. In the space (y-coordinate, orientation), the combination of the maximality and the merging criterion yields that it is more meaningful to observe at the same time G_1 and G_2 than the whole G. This is coherent with the visual perception, since we actually see two lines of segments here. On the contrary, in the (x-coordinate, orientation) space, the merging criterion indicates that observing G is more meaningful than observing simultaneously its children in the dendrogram. This decision is still conform with observation: no particular group within G can be distinguished with regards to the x-coordinate. The same group is obtained in the space (x-coordinate, y-coordinate, orientation), with an obviously lower NFA.

DNA image. The 80 objects in Figure 6 are more complex, in the sense that more features are needed in order to represent them (diameter, elongation, orientation, etc.). It is clear that a projection on a single feature is not really enough to differentiate the objects. Globally, we see three groups of objects: the DNA marks, which share the same form, size and orientation; the numbers, all on the same line, almost of the same size; finally the elements of the ruler, also on the same line and of similar diameters. The position appears to be decisive in the perceptive formation of these groups. In the space (diameter, y-coordinate), we find 6 maximal meaningful groups. Four correponds to the 4 lines of DNA marks (First line: 5 objects, $NFA = 10^{-4.5}$. Second line: 6 objects, $NFA = 10^{-7}$. Third line: 6 objects, $NFA = 10^{-7.5}$. Fourth line: 6 objects, $NFA = 10^{-5.4}$), one to the group of numbers (22 objects, $NFA = 10^{-40}$), and one to the group of objects in the ruler (27 objects, 10^{-45}).

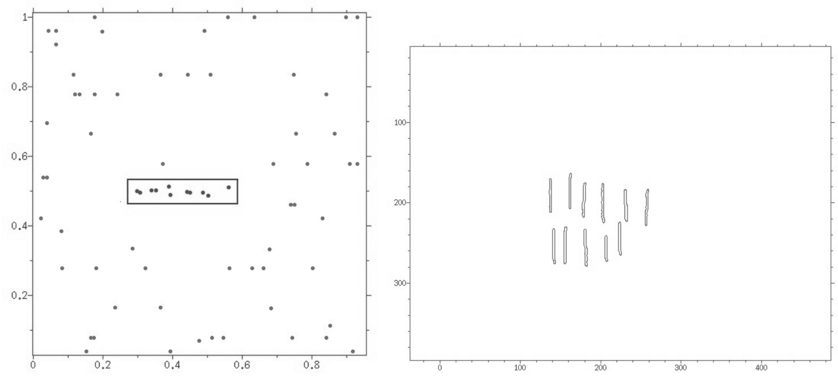
Now, assume that we do not consider the position information. Do we still see the DNA marks as a coherent group? By taking several other features into account, the DNA marks form an isolated and very meaningful group: the combination of features (orientation, diameter, elongation, convexity coefficient) reveals the DNA marks as the only one maximal meaningful cluster ($NFA = 10^{-88}$). This means that an



(a)



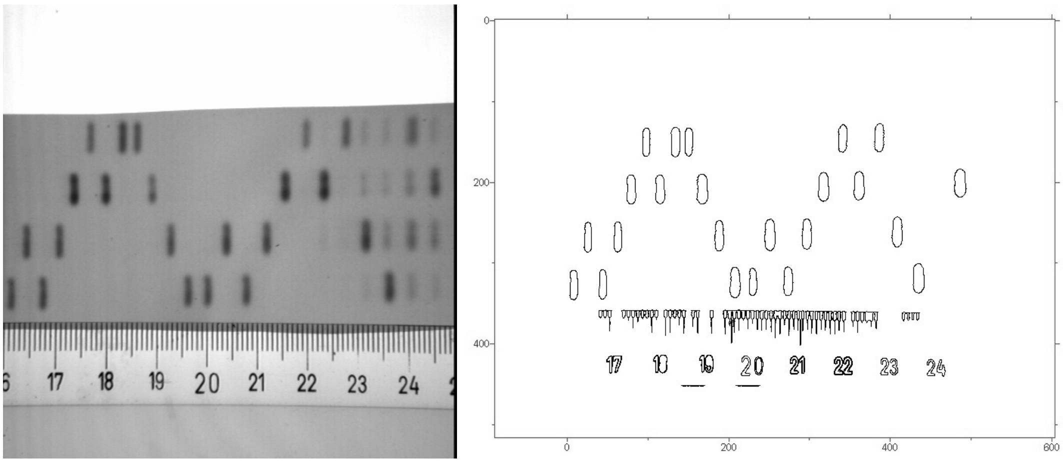
(b)



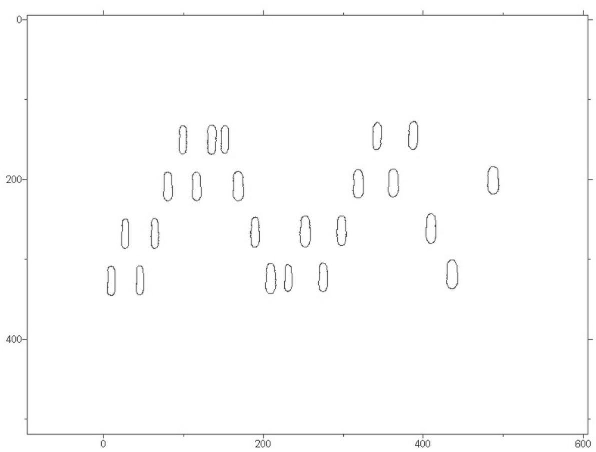
(c)

Figure 5: (a) Image of segments and its maximal meaningful level lines ($\varepsilon = 10^{-20}$, 71 objects). (b) Maximal meaningful groups for orientation. The more meaningful one is the group of vertical lines ($NFA = 0.015$, 9 objects). (c) Left: 2D-representation of the objects in the space (x-coordinate,orientation). There is only one maximal meaningful group, indicated by the rectangle. There is no maximal meaningful group for position only. Right: corresponding group of objects ($NFA = 10^{-4}$, 11 objects).

elementary form description is enough to perceptively distinguish the DNA group among all these objects, whatever their position.



(a)



(b)

Figure 6: (a) Left: DNA mage. Right: its maximal meaningful level lines ($\varepsilon = 10^{-20}$, 80 objects). (b) Maximal meaningful group ($NFA = 10^{-88}$, 23 objects) for the following combination of qualities: diameter, orientation, elongation, convexity coefficient.

5 Conclusion

Finding groups in data sets is a major problem in many fields of knowledge such as statistical pattern recognition, image processing, or data mining. In this paper, we proposed a method to select the clusters in a hierarchical structure, which suitably represent the dataset. We introduced a quantitative measure of validity (the NFA), inspired by Desolneux *et al.*'s work [4], which aims at controlling the number of false alarms. This measure enables to detect the groups of patterns which are meaningful when considered as isolated entities. In order to obtain a non-redundant description of the data, and to discard outliers, we propose a selection strategy which combines the natural ordering induced by the NFA in the tree, and a new merging criterion, which is consistent with the NFA measure. A generalization of this merging rule, using multinomial distributions, can be applied to non-binary trees. Experiments in section 4 illustrate the good performances of the method concerning the previous claims. Moreover, as shown in the last experiment, the clustering validation method presented here, provides an appealing framework to the analysis of the collaboration between perceptive cues. An application of this method to shape recognition is being investigated by the authors.

References

[1] H.H. Bock. On some significance tests in cluster analysis. *Journal of Classification*, 2:77–108, 1985.

[2] T. Calinski and J. Harabasz. A dendrite method for cluster analysis. *Communications in statistics*, 3(1):1–27, 1974.

[3] A. Desolneux, L. Moisan, and J.-M. Morel. Edge detection by Helmholtz principle. *Journal of Mathematical Imaging and Vision*, 14(3):271–284, 2001.

[4] A. Desolneux, L. Moisan, and J.-M. Morel. A grouping principle and four applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(4):508–513, 2003.

[5] R . C. Dubes. How many clusters are best? – an experiment. *Pattern Recognition*, 20(6):645–663, 1987.

[6] R.O. Duda, P.E. Hart, and D.G. Stork. *Pattern Classification*. John Wiley and Sons, 2000.

[7] A.D. Gordon. Null models in cluster validation. In W. Gaul and D. Pfeifer, editors, *From Data to Knowledge: Theoretical and Practical Aspects of Classification, Data Analysis, and Knowledge Organization*, pages 32–44. Springer Verlag, 1996.

[8] A.D. Gordon. *Classification*. Monographs on Statistics and Applied Probability 82, Chapman & Hall, 1999.

[9] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning Data Mining, Inference, and Prediction*. Springer Series in Statistics, 2001.

[10] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: a review. *ACM Computing Surveys*, 31(3):264–323, 1999.

[11] K. Joag-Dev and F. Proschan. Negative association of random variables, with applications. *Annals of Statistics*, 11(1):286–295, 1983.

[12] D.G. Lowe. *Perceptual Organization and Visual Recognition*. Kluwer Academic Publisher, 1985.

[13] G.W. Milligan and M.C. Cooper. An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 50(2):159–179, 1985.

[14] M. Wertheimer. Untersuchungen zur Lehre der Gestalt, II. *Psychologische Forschung*, (4):301–350, 1923. Translation published as Laws of Organization in Perceptual Forms, in Ellis, W. (1938). A source book of Gestalt psychology (pp. 71-88). Routledge & Kegan Paul.