Clustering Methods: A History of *k*-Means Algorithms

Hans-Hermann Bock

Institute of Statistics, RWTH Aachen University, D-52056 Aachen, Germany,
bock@stochastik.rwth-aachen.de

Abstract. This paper surveys some historical issues related to the well-known k-means algorithm in cluster analysis. It shows to which authors the different versions of this algorithm can be traced back, and which were the underlying applications. We sketch various generalizations (with references also to Diday's work) and thereby underline the usefulness of the k-means approach in data analysis.

1 Introduction

Cluster analysis was a main topic in the beginning of Edwin Diday's scientific career. In fact, the monograph 'Principles of numerical taxonomy' by Sokal and Sneath (1963) motivated world-wide research on clustering methods and initiated the publication of books such as 'Les bases de la classification automatique' (Lerman (1970)), 'Mathematical taxonomy' (Jardine and Sibson (1971)), 'Cluster analysis for applications' (Anderberg (1973)), 'Cluster analysis' (Bijnen (1973)), 'Automatische Klassifikation' (Bock (1974)), 'Empirische Verfahren zur Klassifikation' (Sodeur (1974)), 'Probleme und Verfahren der numerischen Klassifikation (Vogel (1975)), 'Cluster-Analyse-Algorithmen (Späth (1975, 1985)), and 'Clustering algorithms' (Hartigan (1975)). With the consequence that the basic problems and methods of clustering became well-known in a broad scientific community, in statistics, data analysis, and - in particular - in applications.

One of the major clustering approaches is based on the sum-of-squares criterion and on the algorithm that is today well-known under the name 'k-means'. When tracing back this algorithm to its origins, we see that it has been proposed by several scientists in different forms and under different assumptions. Later on, many researchers investigated theoretical and algorithmic aspects and modifications of the method, e.g., when considering 'continuous' analogues of the SSQ criterion (Cox (1957), Fisher (1958), Bock (1974)), by investigating the asymptotic behaviour under random sampling strategies (Hartigan (1975), Pollard (1982), Bock (1985)), and by extending its domain to new data types and probabilistic models. Certainly, Diday's monograph (Diday et al. 1979), written with 22 co-authors, marks a considerable level of generalization of the basic idea and established its usage for model-based clustering.

This article surveys the origins and some extensions of the k-means algorithm. In Section 2 we formulate the SSQ clustering problem and the *k-means algorithm*. Section 3 describes the most early papers proposing the SSQ criterion and the k-means algorithm. Section 4 concentrates on extensions of the SSQ criterion that lead to *generalized k-means algorithms*. Section 5 deals with one- and two-parameter criteria and shows how a 'convexity-based' clustering criterion can be minimized with a *k-tangent algorithm*.

2 k-means clustering for the SSQ criterion

There are two versions of the well-known SSQ clustering criterion: the 'discrete' and the 'continuous' case.

Discrete SSQ criterion for data clustering: Given n data points $x_1, ..., x_n$ in $\mathbb{R}^p$ and a k-partition $\mathcal{C} = (C_1, ..., C_k)$ of the set $\mathcal{O} = \{1, ..., n\}$ of underlying 'objects' with non-empty classes $C_i \subset \mathcal{O}$, the discrete SSQ criterion (also termed: variance criterion, inertia, or trace criterion) is given by

$$g_n(\mathcal{C}) := \sum_{i=1}^{k} \sum_{\ell \in C_i} ||x_\ell - \overline{x}_{C_i}||^2 \ \rightarrow \ \min_{\mathcal{C}} \tag{1}$$

where $\overline{x}_{C_i}$ denotes the centroid of the data points x_ℓ 'belonging' to class C_i (i.e. with $\ell \in C_i$). We look for a k-partition of $\mathcal{O}$ with minimum criterion value $g_n(\mathcal{C})$. The one-parameter optimization problem (1) is related, and even equivalent, to the two-parameter optimization problem

$$g_n(\mathcal{C}, \mathcal{Z}) := \sum_{i=1}^{k} \sum_{k \in C_i} ||x_\ell - z_i||^2 \ \rightarrow \ \min_{\mathcal{C}, \mathcal{Z}} \tag{2}$$

where minimization is also w.r.t. all systems $\mathcal{Z} = (z_1, ..., z_k)$ of k points $z_1, ..., z_k$ from $\mathbb{R}^p$ (class representatives, class prototypes). This results from part (i) of the following theorem:

Theorem 1:
(i) For any fixed k-partition $\mathcal{C}$ the criterion $g_n(\mathcal{C}, \mathcal{Z})$ is partially minimized w.r.t. $\mathcal{Z}$ by the system of class centroids $\mathcal{Z}^* = (\overline{x}_{C_1}, ..., \overline{x}_{C_k}) =: \mathcal{Z}(\mathcal{C})$:

$$g_n(\mathcal{C}, \mathcal{Z}) \geq g_n(\mathcal{C}, \mathcal{Z}^*) = \sum_{i=1}^{k} \sum_{k \in C_i} ||x_k - \overline{x}_{C_i}||^2 \ = g_n(\mathcal{C}) \quad \textit{for all } \mathcal{Z}. \tag{3}$$

(ii) For any fixed prototype system $\mathcal{Z}$ the criterion $g_n(\mathcal{C}, \mathcal{Z})$ is partially minimized w.r.t. $\mathcal{C}$ by any minimum-distance partition $\mathcal{C}^* := (C_1^*, ..., C_k^*) =: \mathcal{C}(\mathcal{Z})$ induced by $\mathcal{Z}$, i.e. with classes given by $C_i^* := \{\ell \in \mathcal{O} \mid d(x_\ell, z_i) =$

$\min_{j=1,...,k} d(x_\ell, z_j)\}$ *($i = 1, ..., n$) where* $d(x, z) = ||x - z||^2$ *is the squared Euclidean distance:*

$$g_n(\mathcal{C}, \mathcal{Z}) \geq g_n(\mathcal{C}^*, \mathcal{Z}) = \sum_{\ell=1}^{n} \min_{j=1,...,k} \{ \, ||x_\ell - z_j||^2 \, \} \qquad \textit{for all } \mathcal{C}. \tag{4}$$

A broad range of methods has been designed in order to minimize the discrete criteria (1) and (2), either exactly or approximately. They can be roughly grouped into enumeration methods, mathematical and combinatorial programming for exact minimization (Hansen and Jaumard (1997), Grötschel and Wakabayashi (1989)), integer, linear, and dynamic programming (Jensen (1969), Vinod (1969), Rao (1971)), heuristical and branch & bound methods (see also Anderberg (1973), Mulvey and Crowder (1979)).

The *k-means algorithm* tries to approximate an optimum k-partition by iterating the partial minimization steps (i) and (ii) from Theorem 1, in turn. It proceeds as follows[1]:

$t = 0$: Begin with an arbitrary prototype system $\mathcal{Z}^{(0)} = (z_1^{(0)}, ..., z_k^{(0)})$.

$t \to t + 1$:

(i) Minimize the criterion $g_n(\mathcal{C}, \mathcal{Z}^{(t)})$ w.r.t. the k-partition $\mathcal{C}$, i.e., determine a minimum-distance partition $\mathcal{C}^{(t+1)} := \mathcal{C}(\mathcal{Z}^{(t)})$.

(ii) Minimize the criterion $g_n(\mathcal{C}^{(t+1)}, \mathcal{Z})$ w.r.t. the prototype system $\mathcal{Z}$, i.e., calculate the system of class centroids $\mathcal{Z}^{(t+1)} := \mathcal{Z}(\mathcal{C}^{(t+1)})$.

Stopping: Iterate the steps (i) and (ii) until stationarity.

By construction, this algorithm yields a sequence $\mathcal{Z}^{(0)}, \mathcal{C}^{(1)}, \mathcal{Z}^{(1)}, \mathcal{C}^{(2)}, ...$ of prototypes and partitions with decreasing values of the criteria (1) and (2) that converge to a (typically local) minimum value.

Remark 1: In mathematical terms, the k-means algorithm is a *relaxation method* for minimizing a function of several parameters by iterative partial minimization steps (see also Mulvey and Crowder 1979), and also called an *alternating optimization* method.

Continuous SSQ criterion for space dissection: Considering $x_1, ..., x_n$ as realizations of a random vector X with distribution P in $\mathbb{R}^p$, we may formulate the following 'continuous' analogues of (1) and (2): We look for a k-partition $\mathcal{B} = (B_1, ..., B_k)$ of $\mathbb{R}^p$ with minimum value

$$g(\mathcal{B}) := \sum_{i=1}^{k} \int_{B_i} ||x - E[X|X \in B_i]||^2 \, dP(x) \;\to\; \min_{\mathcal{B}}. \tag{5}$$

As before we can relate (5) to a two-parameter optimization problem:

$$g(\mathcal{B}, \mathcal{Z}) := \sum_{i=1}^{k} \int_{B_i} ||x - z_i||^2 \, dP(x) \;\to\; \min_{\mathcal{B}, \mathcal{Z}} \tag{6}$$

[1] This is the *batch version* of the k-means algorithm; see *Remark 2*.

and formulate the analogue of Theorem 1:

Theorem 2:
(i) For any fixed k-partition $\mathcal{B}$ of $\mathbb{R}^p$ the criterion $g(\mathcal{B}, \mathcal{Z})$ is partially minimized w.r.t. $\mathcal{Z}$ by the prototype system $\mathcal{Z}^* = (z_1^*, ..., z_k^*) =: \mathcal{Z}(\mathcal{B})$ given by the conditional expectations $z_i^* := E[X|X \in B_i]$ of B_i:

$$g(\mathcal{B}, \mathcal{Z}) \geq g(\mathcal{B}, \mathcal{Z}^*) = \sum_{i=1}^{k} \int_{B_i} ||x - E[X|X \in B_i]||^2 \; = g(\mathcal{B}) \quad \textit{for all } \mathcal{Z}. \tag{7}$$

(ii) For any fixed prototype system $\mathcal{Z}$ the criterion $g(\mathcal{B}, \mathcal{Z})$ is partially minimized w.r.t. $\mathcal{B}$ by any minimum-distance partition $\mathcal{B}^* = (B_1^*, ..., B_k^*) =: \mathcal{B}(\mathcal{Z})$ generated by $\mathcal{Z}$, i.e. with classes given by $B_i^* := \{x \in \mathbb{R}^p \mid d(x, z_i) = \min_{j=1,...,k}\{d(x, z_j)\}\}$ $(i = 1, ..., n)$:

$$g(\mathcal{B}, \mathcal{Z}) \geq g(\mathcal{B}^*, \mathcal{Z}) = \int_{\mathcal{X}} \min_{j=1,...,k} \{||x - z_j||^2\} \, dP(x) \quad \textit{for all } \mathcal{B}. \tag{8}$$

It is obvious that Theorem 2 can be used to formulate, and justify, a continuous version of the k-means algorithm. However, in contrast to the discrete case, the calculation of the class centroids might be a computational problem.

3 First instances of SSQ clustering and *k*-means

The first formulation of the SSQ clustering problem I know has been provided by Dalenius (1950) and Dalenius and Gurney (1951) in the framework of optimum 'proportional' stratified sampling: For estimating the expectation $\mu = E[X]$ of a real-valued random variable X with distribution density $f(x)$ (e.g., the income of persons in a city), the domain $(-\infty, +\infty)$ of X is dissected into k contiguous intervals ('strata', 'classes') $B_i = (u_{i-1}, u_i]$ $(i = 1, ..., k+1$, with $u_0 = -\infty$ and $u_{k+1} = \infty)$ and from each stratum B_i a fixed number n_i of persons is sampled where $n_i = n \cdot P(B_i)$ is proportional to the probability mass of B_i. This yields n real data $x_1, ..., x_n$. The persons ℓ with income value x_ℓ in B_i build a class C_i with class average $z_i^* := \overline{x}_{C_i}$ $(i = 1, ..., k)$. The linear combination $\hat{\mu} := \sum_{i=1}^{k} (n_i/n) \cdot \overline{x}_{C_i}$ provides an unbiased estimator of μ with variance given by the SSQ criterion: $Var(\hat{\mu}) = g(\mathcal{B})/n$. Dalenius wants to determine a k-partition $\mathcal{B}$ with minimum variance, i.e., maximum accuracy for $\hat{\mu}$ – this means the continuous clustering problem (5).

Dalenius did not use a k-means algorithm for minimizing (5), but a 'shooting' algorithm that is based on the fact that for an optimum partition $\mathcal{B}$ of $\mathbb{R}^1$ the class boundaries u_i must necessarily lie midway between the neigbouring class centroids such that $u_i = (z_i^* + z_{i+1}^*)/2$ or $z_{i+1}^* = 2u_i - z_i^*$ must hold for $i = 1, ..., k-1$. Basically, he constructs a sequence $z_1 < u_1 < z_2 < u_2 < \cdots$ of centers and boundaries by

– choosing, for $i = 1$, an initial value $z_1 \in \mathbb{R}^1$

– determining, for $i = 1$, the upper boundary u_i of $B_i = (u_{i-1}, u_i]$ from the equation $E[X|X \in B_i] = [\int_{u_{i-1}}^{u_i} xf(x)dx]/[\int_{u_{i-1}}^{u_i} f(x)dx] \stackrel{!}{=} z_i$ (the expectation is an increasing function of u_i)
– then calculating the next centroid by $z_{i+1} = 2u_i - z_i$
– and iterating for $i = 2, 3, ..., k$.

By trial and error, the initial value z_1 is adapted such that the iteration stops with k classes and the k-th upper boundary $u_k = \infty$. A 'data version' of this approach for minimizing (1) has been described, e.g., by Strecker (1957), Stange (1960), and Schneeberger (1967).

Steinhaus (1956) was the first to propose explicitly the k-means algorithm in the multidimensional case. His motivation stems from mechanics (even if he refers also to examples from anthropology and industry): to partition a heterogeneous solid $\mathcal{X} \subset I\!R^p$ with internal mass distribution $f(x)$ into k subsets $B_1, ..., B_k$ and to minimize (6), i.e., the sum of the partial moments of inertia with respect to k points $z_1, ..., z_k \in I\!R^p$ by a suitable choice of the partition $\mathcal{B}$ and the z_i's. He does not only describe the (continuous version of the) k-means algorithm, but also discusses the existence of a solution for (6), its uniqueness ('minimum parfait', examples and counterexamples), and the behaviour of the sequence of minimum SSQ values for $k \to \infty$.

The first to propose the discrete k-means algorithm for clustering data, i.e., for solving (1), was Forgy (1965)[2], Jancey (1966a) was the first to mention it explicitly in a publication (see also Jancey (1966b)). The k-means method became a standard procedure in clustering and is known under quite different names such as *nuées dynamiques* (Diday 1971, 1972), *dynamic clusters method* (Diday 1973; Diday and Schroeder 1974a), *iterated minimum-distance partition method* (Bock 1974), *nearest centroid sorting* (Anderberg 1973), etc.

Remark 2: The name 'k-means algorithm' was first used by MacQueen (1967), but not for the 'batch algorithm' from Section 2. Instead he used it for his sequential, 'single-pass' algorithm for (asymptotically) minimizing the continuous SSQ criterion (5) on the basis of a sequence of data points $x_1, x_2, ... \in I\!R^p$ (sampled from P): The first k data (objects) defined k initial singleton classes $C_i^{(k)} = \{i\}$ with class centroids $z_i^{(k)} := \overline{x}_{C_i^{(k)}} = x_i$ $(i = 1, ..., k)$. Then, for $\ell = k+1, k+2, ...$, the data x_ℓ were sequentially observed and assigned to the class $C_i^{(\ell-1)}$ with closest class centroid $z_i^{(\ell-1)} := \overline{x}_{C_i^{(\ell-1)}}$ and (only) its class centroid was updated: $z_i^{(\ell)} := \overline{x}_{C_i^{(\ell)}} = z_i^{(\ell-1)} + (x_\ell - \overline{x}_{C_i^{(\ell-1)}})/|C_i^{(\ell)}|$. When stopping at some 'time' T, the minimium-distance partition $\mathcal{B}(\mathcal{Z}^{(T)})$ of $I\!R^p$ induced by the last centroid system $\mathcal{Z}^{(T)} = (\overline{x}_{C_1^{(T)}}, ..., \overline{x}_{C_k^{(T)}})$ approximates a (local) solution of (5) if T is large. This single-pass interpretation of 'k-means

[2] Forgy's abstract of his talk does not mention the k-means algorithm, however, details of his lecture were given by Anderberg (1973), p. 161 and MacQueen (1967) p. 294.

algorithm' is used in many monographs. – In Späth (1975) the batch-version of k-means is called HMEANS, whereas KMEANS denotes an algorithm that exchanges single objects between classes in order to decrease (1). Hartigan (1975) uses the term 'k-means' for various algorithms working with k class centroids, e.g. for Späth's exchange algorithm (on page 85/86), and k-means as described in our Section 2 is one of several options mentioned on page 102 of Hartigan (1975) (see also Hartigan and Wong (1979)).

In computer science and pattern recognition communities the k-means algorithm is often termed *Lloyd's algorithm I*. Lloyd (1957) considers the continuous SSQ clustering criterion (6) in $\mathbb{R}^1$ in the context of pulse-code modulation: 'Quantization' means replacing a random (voltage) signal X by a discretized approximate signal $\hat{X}$ that takes a constant value z_i ('quantum') if X belongs to the i-th class B_i of the partition $\mathcal{B} = (B_1, ..., B_k)$ of $\mathbb{R}^1$ such that $\hat{X} = z_i$ iff $X \in B_i$ $(i = 1, ..., k)$. Optimum quantification means minimization of the criterion (6). Lloyd reports the optimality of the class centroids $z_i^* = E[X|X \in B_i]$ for a fixed partition $\mathcal{B}$ and describes the one-dimensional version of the k-means algorithm as his 'Method I' whereas his 'Method II' is identical to the 'shooting method' of Dalenius.

4 Generalized k-means methods

The two-parameter SSQ clustering criteria (2) and (6) have been generalized in many ways in order to comply with special data types or cluster properties. In the discrete case, typical criteria have the two-parameter form

$$g_n(\mathcal{C}, \mathcal{Z}) := \sum_{i=1}^{k} \sum_{\ell \in C_i} d(\ell, z_i) \rightarrow \min_{\mathcal{C}, \mathcal{Z}} \tag{9}$$

where $d(\ell, z)$ measures the dissimilarity between an object ℓ and a class prototype z (sometimes written as $d(x_\ell, z)$ or $d_{\ell z}$ etc., depending on the context). There is much flexibility in this approach since

(1) there is almost no constraint on the type of underlying data (quantitative and/or categorical data, shapes, relations, weblogs, DNA strains, images)
(2) there are many ways to specify a family $\mathcal{P}$ of appropriate or admissible 'class prototypes' z to represent specific aspects of the clusters (points, hyperspaces in $\mathbb{R}^p$, subsets of $\mathcal{O}$, order relations),
(3) there exists a wealth of possibilities to choose the dissimilarity measure d, and we may, additionally, introduce weights w_ℓ for the objects $\ell \in \mathcal{O}$.

In all these cases, the following *generalized k-means algorithm* can be applied in order to attain a (locally or globally) optimum configuration $(\mathcal{C}, \mathcal{Z})$:

$t = 0$: Begin with an arbitrary prototype system $\mathcal{Z}^{(0)} = (z_1^{(0)}, ..., z_k^{(0)})$.

$t \rightarrow t + 1$:

(i) Minimize the criterion $g_n(\mathcal{C}, \mathcal{Z}^{(t)})$ w.r.t. the k-partition $\mathcal{C}$ from $\mathcal{P}$. Typically, this yields a minimum-distance partition $\mathcal{C}^{(t+1)} = \mathcal{C}(\mathcal{Z}^{(t)})$ with k classes $C_i^{(t+1)} := \{\ell \in \mathcal{O} \mid d(\ell, z_i^{(t)}) = \min_{j=1,...,k} d(\ell, z_j^{(t)})\ \}$.

(ii) Minimize the criterion $g_n(\mathcal{C}^{(t+1)}, \mathcal{Z})$ w.r.t. the prototype system $\mathcal{Z}$. Often, this amounts to determining, for each class $C_i = C_i^{(t+1)}$, a 'most typical configuration' $z_i^{(t+1)}$ in the sense:

$$Q(C_i, z) := \sum_{\ell \in C_i} d(\ell, z) \quad \rightarrow \quad \min_{z \in \mathcal{P}}. \tag{10}$$

Stopping: Iterate the steps (i) and (ii) until stationarity.

The first paper to propose the general criterion (9) and its generalized k-means method is Maranzana (1963): He starts from a $n \times n$ dissimilarity matrix $(d_{\ell t})$ for n factories $\ell = 1, ..., n$ in an industrial network where $d_{\ell t}$ are the minimum road transportation costs between ℓ and t. He wants to partition the set of factories into k classes $C_1, ..., C_k$ and to find a selection $\mathcal{Z} = (z_1, ..., z_k)$ of k factories as 'supply points' such that when supplying all factories of the class C_i from the supply point $z_i \in \mathcal{O}$, the overall transport costs are minimized in the sense of (9) where $d(\ell, z_i) = d_{\ell, z_i}$ means the dissimilarity between the factory (object) ℓ and the factory (supply point) $z_i \in \mathcal{O}$ (where we have omitted object-specific weights from Maranzana's formulation). So the family $\mathcal{P}$ of admissible prototypes consists of all singletons from $\mathcal{O}$ and (ii) means determining the 'most cheapest supply point' in C_i. Kaufman and Rousseeuw (1987, 1990) termed this method 'partitioning around medoids' (the *medoid* or *centrotype* of a class C_i is the most typical object in C_i in the sense of (10)).

Many authors, including Diday (1971, 1972, 1973) and Diday et al. (1979), have followed the generalized clustering approach via (9) in various settings and numerous variations and thereby obtained a plethora of generalized k-means algorithms, e.g., by

– using Mahalanobis or L_q distance in (1) instead of the Euclidean one, eventually including constraints (Diday and Govaert (1974, 1977): *méthode des distances adaptatives*)
– characterizing clusters by prototype hyperplanes, resulting in *principal component clustering* (Bock (1974) chap. 17, Diday and Schroeder (1974a)) and *clusterwise regression* (Bock (1969), Charles (1977), Späth (1979)).
– *projection pursuit clustering* where class centers are located on a low-dimensional hyperplane (Bock (1987, 1996c), Vichi (2005)),
– characterizing a class by the most typical subset (pair, triple,...) of objects from this class (Diday et al. (1979)).

A major step with new insight was provided by Diday and Schroeder (1974a, 1974b, 1976) and Sclove (1977) who detected that under a probabilistic 'fixed-partition' clustering model, maximum-likelihood estimation of an unknown k-partition $\mathcal{C}$ leads to a clustering criterion of the type (9) and can there-

fore be handled by a k-means algorithm[3]. The *fixed-partition model* considers the data $x_1, ..., x_n$ as realizations of n independent random vectors $X_1, ..., X_n$ with distributions from a density family $f(\cdot; \vartheta)$ (w.r.t. the Lebesgue or counting measure) with parameter ϑ (e.g., a normal, van Mises, loglinear,... distribution). It assumes the existence of a fixed, but unknown k-partition $\mathcal{C} = (C_1, ..., C_k)$ of $\mathcal{O}$ together with a system $\theta = (\vartheta_1, ..., \vartheta_k)$ of class-specific parameters such that the distribution of the data is class-specific in the sense that $X_\ell \sim f(\cdot; \vartheta_i)$ for all $\ell \in C_i$ $(i = 1, ..., n)$. Then maximizing the likelihood of $(x_1, ..., x_n)$ is equivalent to

$$g_n(\mathcal{C}, \theta) \;:=\; \sum_{i=1}^{k} \sum_{\ell \in C_i} [-\log f(x_\ell; \vartheta_i)] \;\rightarrow\; \min_{\mathcal{C}, \theta}, \tag{11}$$

this is the criterion (9) with $z_i \equiv \vartheta_i, \mathcal{Z} \equiv \theta$, and $d(\ell, z_i) = -\log f(x_\ell; \vartheta_i)$. The minimum-distance assignment of an object ℓ in (i) means maximum-likelihood assignment to a class C_i, and in (ii) optimum class prototypes are given by the maximum-likelihood estimate $\hat{\vartheta}_i$ of $z_i \equiv \vartheta_i$ in C_i. A major advantage of this approach resides in the fact that we can design meaningful clustering criteria also in the case of qualitative or binary data, yielding, *entropy clustering* and *logistic clustering* methods (Bock 1986), or models comprizing random noise or outliers (Gallegos (2002), Gallegos and Ritter (2005)). – A detailed account of these approaches is given, e.g., in Bock (1974, 1996a, 1996b, 1996c) and Diday et al. (1979).

5 Convexity-based criteria and the *k*-tangent method

The derivation of the k-means algorithm in Section 2 shows that it relies on the fact that the intuitive SSQ optimization problem (1) for *one* parameter $\mathcal{C}$ has an equivalent version (2) where optimization is w.r.t. *two* parameters $\mathcal{C}$ and $\mathcal{Z}$. In order to extend the domain of applicability of the k-means algorithm we may ask, more generally, if for an intuitively defined one-parameter clustering criterion we can find a two-parameter version such that both resulting optimization problems are equivalent and a k-means algorithm can be applied. A general investigation of this problem has been given by Windham (1986, 1987) and Bryant (1988). In the following we describe a situation where the answer is affirmative and leads to a new *k-tangent algorithm* (Bock (1983, 1992, 2003), Pötzelberger and Strasser (2001).

We consider the following 'convexity-based' clustering criterion for $x_1, ..., x_n \in \mathbb{R}^p$ that should be maximized w.r.t the k-partition $\mathcal{C}$:

$$k_n(\mathcal{C}) \;:=\; \sum_{i=1}^{k} (|C_i|/n) \cdot \phi(\overline{x}_{C_i}) \;\rightarrow\; \max_{\mathcal{C}} \tag{12}$$

[3] This fact was already known before, e.g., in the case of SSQ and the normal distribution, but these authors recognized its importance for more general cases.

Here $\phi(\cdot)$ is a smooth convex function, and (12) is a generalization of the SSQ clustering problem (1) since for $\phi(x) := ||x||^2$ (12) reduces to (1). Similarly, the continuous version

$$k(\mathcal{B}) \ := \ \sum_{i=1}^{k} \ P(B_i) \cdot \phi(E[X|X \in B_i]) \ \rightarrow \ \max_{\mathcal{B}} \tag{13}$$

is equivalent to (5), its generalization

$$K(\mathcal{B}) \ = \ \sum_{i=1}^{k} \ P_0(B_i) \cdot \phi(E_0[\lambda(X)|X \in B_i])) \ \rightarrow \ \max_{\mathcal{B}} \tag{14}$$

looks for an optimum dissection of $I\!R^p$ such that, for two equivalent alternative distributions P_0, P_1 on $I\!R^p$ with likelihood ratio $\lambda(x) = (dP_1/dP_0)(x) = f_1(x)/f_0(x)$, the discretized distributions $(P_0(B_1), ..., P_0(B_k))$ and $(P_1(B_1), ..., P_1(B_k))$ will be as different as possible. (Note that K is Cszizar's ϕ-divergence and reduces, e.g., to Kullback-Leibler and χ^2 distance for $\phi(u) = -\log u$ and $\phi(u) = (u-1)^2$, respectively; for other functions λ see Bock (2003).) Some analysis based on the convexity of ϕ shows that maximizing $K(\mathcal{C})$ is equivalent to the two-parameter minimization problem

$$G(\mathcal{B}, \mathcal{Z}) \ := \ \sum_{i=1}^{k} \ \int_{B_i} [\phi(\lambda(x)) - t(\lambda(x); z_i)] \ dP_0(x) \ \rightarrow \ \min_{\mathcal{B}, \mathcal{Z}} \tag{15}$$

where $\mathcal{Z} = (z_1, ..., z_k) \in I\!R_+^k$ and $t(\lambda; z) := \phi(z) + \phi'(z)(\lambda - z)$ is the tangent (support plane) of $y = \phi(\lambda)$ in the support point $z > 0$ ([....] is the weighted 'volume' between the curve and the corresponding segments of the tangents). Therefore we can apply the alternating partial minimization device. The resulting method is termed 'k-tangent algorithm' and comprizes the steps:

(i) For a given support point system $\mathcal{Z}$, determine the 'maximum-tangent partition' $\mathcal{B}$ with classes defined by maximum tangent values:

$$B_i \ := \ \{ \ x \in I\!R^p \mid t(\lambda(x); z_i) = \max_{j=1,...,k} t(\lambda(x); z_j) \ \} \tag{16}$$

(ii) For a given k-partition $\mathcal{B}$ determine the system $\mathcal{Z}$ of class-specific discrete likelihood ratios:

$$z_i \ := \ E_0[\ \lambda(X)] \mid X \in B_i \ ] \ = \ \frac{P_1(B_i)}{P_0(B_i)} \qquad i = 1, ..., k. \tag{17}$$

Iteration of (i) and (ii) yields a sequence of partitions with decreasing values in (14). – Pötzelberger and Strasser (2001) investigate the theoretical properties of the optimum partitions of (12) and (13), Bock (2003) shows, e.g., how the k-tangent method can be applied to the simultaneous classification of the rows and columns of a contingency table.

References

ANDERBERG, M.R. (1973): *Cluster analysis for applications.* Academic Press, New York.

BIJNEN, E.J. (1973): *Cluster analysis.* Tilburg University Press, Tilburg, Netherlands.

BOCK, H.-H. (1969): *The equivalence of two extremal problems and its application to the iterative classification of multivariate data.* Paper presented at the Workshop 'Medizinische Statistik', February 1969, Forschungsinstitut Oberwolfach.

BOCK, H.-H. (1974): *Automatische Klassifikation. Theoretische und praktische Methoden zur Strukturierung von Daten (Clusteranalyse).* Vandenhoeck & Ruprecht, Göttingen.

BOCK, H.-H. (1985): On some significance tests in cluster analysis. *Journal of Classification 2, 77-108.*

BOCK, H.-H. (1983): *A clustering algorithm for choosing optimal classes for the chi-square test.* Bull. 44th Session of the International Statistical institute, Madrid, Contributed Papers, Vol 2, 758-762.

BOCK, H.-H. (1986): Loglinear models and entropy clustering methods for qualitative data. In: W. Gaul, M. Schader (Eds.): *Classification as a tool of research.* North Holland, Amsterdam, 19-26.

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

BOCK, H.-H. (1992): A clustering technique for maximizing ϕ-divergence, noncentrality and discriminating power. In: M. Schader (Ed.): *Analyzing and modeling data and knowledge.* Springer, Heidelberg, 19-36.

BOCK, H.-H. (1996a): Probability models and hypotheses testing in partitioning cluster analysis. In: P. Arabie, L.J. Hubert, G. De Soete (Eds.): *Clustering and classification.* World Scientific, Singapore, 377-453.

BOCK, H.-H. (1996b): Probabilistic models in partitional cluster analysis. *Computational Statistics and Data Analysis 23, 5-28.*

BOCK, H.-H. (1996c): Probabilistic models in cluster analysis. In: A. Ferligoj, A. Kramberger (Eds.): *Developments in data analysis.* Proc. Intern. Conf. on 'Statistical data collection and analysis', Bled, 1994. FDV, Metodoloski zvezki, 12, Ljubljana, Slovenia, 3-25.

BOCK, H.-H. (2003): Convexity-based clustering criteria: theory, algorithms, and applications in statistics. *Statistical Methods & Applications 12, 293-317.*

BRYANT, P. (1988): On characterizing optimization-based clustering methods. *Journal of Classification 5, 81-84.*

CHARLES, C. (1977): *Regression typologique.* Rapport de Recherche no. 257. IRIA-LABORIA, Le Chesnay.

COX, D.R. (1957) Note on grouping. *J. Amer. Statist. Assoc. 52, 543-547.*

DALENIUS, T. (1950): The problem of optimum stratification I. *Skandinavisk Aktuarietidskrift 1950, 203-213.*

DALENIUS, T., GURNEY, M. (1951): The problem of optimum stratification. II. *Skandinavisk Aktuarietidskrift 1951, 133-148.*

DIDAY, E. (1971): Une nouvelle méthode de classification automatique et reconnaissance des formes: la méthode des nuées dynamiques. *Revue de Statistique Appliquée XIX (2), 1970, 19-33.*

DIDAY, E. (1972): Optimisation en classification automatique et reconnaissance des formes. *Revue Française d'Automatique, Informatique et Recherche Opérationelle (R.A.I.R.O.) VI, 61-96.*

DIDAY, E. (1973): The dynamic clusters method in nonhierarchical clustering. *Intern. Journal of Computer and Information Sciences 2 (1), 61-88.*

DIDAY, E. et al. (1979): *Optimisation en classification automatique. Vol. I, II.* Institut National der Recherche en Informatique et en Automatique (INRIA), Le Chesnay, France.

DIDAY, E., GOVAERT, G. (1974): Classification avec distance adaptative. *Comptes Rendus Acad. Sci. Paris 278 A, 993-995.*

DIDAY, E., GOVAERT, G. (1977): Classification automatique avec distances adaptatives. *R.A.I.R.O. Information/Computer Science 11 (4), 329-349.*

DIDAY, E., SCHROEDER, A. (1974a): The dynamic clusters method in pattern recognition. In: J.L. Rosenfeld (Ed.): *Information Processing 74.* Proc. IFIP Congress, Stockholm, August 1974. North Holland, Amsterdam, 691-697.

DIDAY, E., SCHROEDER, A. (1974b): *A new approach in mixed distribution detection.* Rapport de Recherche no. 52, Janvier 1974. INRIA, Le Chesnay.

DIDAY, E., SCHROEDER, A. (1976): A new approach in mixed distribution detection. *R.A.I.R.O. Recherche Opérationelle 10 (6), 75-1060.*

FISHER, W.D. (1958): On grouping for maximum heterogeneity. *J. Amer. Statist. Assoc. 53, 789-798.*

FORGY, E.W. (1965): Cluster analysis of multivariate data: efficiency versus interpretability of classifications. Biometric Society Meeting, Riverside, California, 1965. Abstract in *Biometrics 21 (1965) 768.*

GALLEGOS, M.T. (2002): Maximum likelihood clustering with outliers. In: K. Jajuga, A. Sokolowski, H.-H. Bock (Eds.): *Classification, clustering, and data analysis.* Springer, Heidelberg, 248-255.

GALLEGOS, M.T., RITTER, G. (2005): A robust method for cluster analysis. *Annals of Statistics 33, 347-380.*

GRÖTSCHEL, M., WAKABAYASHI, Y. (1989): A cutting plane algorithm for a clustering problem. *Mathematical Programming 45, 59-96.*

HANSEN, P., JAUMARD, B. (1997): Cluster analysis and mathematical programming. *Mathematical Programming 79, 191-215.*

HARTIGAN, J.A. (1975): *Clustering algorithms.* Wiley, New York.

HARTIGAN, J.A., WONG, M.A. (1979): A k-means clustering algorithm. *Applied Statistics 28, 100-108.*

JANCEY, R.C. (1966a): Multidimensional group analysis. *Australian J. Botany 14, 127-130.*

JANCEY, R. C. (1966b): The application of numerical methods of data analysis to the genus Phyllota Benth. in New South Wales. *Australian J. Botany 14, 131-149.*

JARDINE, N., SIBSON, R. (1971): *Mathematical taxonomy.* Wiley, New York.

JENSEN, R.E. (1969): A dynamic programming algorithm for cluster analysis. *Operations Research 17, 1034-1057.*

KAUFMAN, L., ROUSSEEUW, P.J. (1987): Clustering by means of medoids. In: Y. Dodge (Ed.): *Statistical data analysis based on the L_1-norm and related methods.* North Holland, Amsterdam, 405-416.

KAUFMAN, L., ROUSSEEUW, P.J. (1990): *Finding groups in data.* Wiley, New York.

LERMAN, I.C. (1970): *Les bases de la classification automatique.* Gauthier-Villars, Paris.

LLOYD, S.P. (1957): Least squares quantization in PCM. Bell Telephone Labs Memorandum, Murray Hill, NJ. Reprinted in: *IEEE Trans. Information Theory IT-28 (1982), vol. 2, 129-137.*

MacQUEEN, J. (1967): Some methods for classification and analysis of multivariate observations. In: L.M. LeCam, J. Neyman (eds.): *Proc. 5th Berkeley Symp. Math. Statist. Probab. 1965/66.* Univ. of California Press, Berkeley, vol. I, 281-297.

MARANZANA, F.E. (1963): On the location of supply points to minimize transportation costs. *IBM Systems Journal 2, 129-135.*

MULVEY, J.M., CROWDER, H.P. (1979): Cluster analysis: an application of Lagrangian relaxation. *Management Science 25, 329-340.*

PÖTZELBERGER, K., STRASSER, H. (2001): Clustering and quantization by MSP partitions. *Statistics and Decision 19, 331-371.*

POLLARD, D. (1982): A central limit theorem for k-means clustering. *Annals of Probability 10, 919-926.*

RAO, M.R. (1971): Cluster analysis and mathematical programming. *J. Amer. Statist. Assoc. 66, 622-626.*

SCHNEEBERGER, H. (1967): Optimale Schichtung bei proportionaler Aufteilung mit Hilfe eines iterativen Analogrechners. *Unternehmensforschung 11, 21-32.*

SCLOVE, S.L. (1977): Population mixture models and clustering algorithms. *Commun. in Statistics, Theory and Methods, A6, 417-434.*

SODEUR, W. (1974): *Empirische Verfahren zur Klassifikation.* Teubner, Stuttgart.

SOKAL, R.R., SNEATH, P. H. (1963): *Principles of numerical taxonomy.* Freeman, San Francisco - London.

SPÄTH, H. (1975): *Cluster-Analyse-Algorithmen zur Objektklassifizierung und Datenreduktion.* Oldenbourg Verlag, München - Wien.

SPÄTH, H. (1979): Algorithm 39: Clusterwise linear regression. *Computing 22, 367-373.* Correction in *Computing 26 (1981), 275.*

SPÄTH, H. (1985): *Cluster dissection and analysis.* Wiley, Chichester.

STANGE, K. (1960): Die zeichnerische Ermittlung der besten Schätzung bei proportionaler Aufteilung der Stichprobe. *Zeitschrift für Unternehmensforschung 4, 156-163.*

STEINHAUS, H. (1956): Sur la division des corps matériels en parties. *Bulletin de l'Académie Polonaise des Sciences, Classe III, vol. IV, no. 12, 801-804.*

STRECKER, H. (1957): *Moderne Methoden in der Agrarstatistik.* Physica, Würzburg, p. 80 etc.

VICHI, M. (2005): Clustering including dimensionality reduction. In: D. Baier, R. Decker, L. Schmidt-Thieme (Eds.): *Data analysis and decision support.* Springer, Heidelberg, 149-156.

VINOD, H.D. (1969): Integer programming and the theory of grouping. *J. Amer. Statist. Assoc. 64, 506-519.*

VOGEL, F. (1975): *Probleme und Verfahren der Numerischen Klassifikation.* Vandenhoeck & Ruprecht, Göttingen.

WINDHAM, M.P. (1986): A unification of optimization-based clustering algorithms. In: W. Gaul, M. Schader (Eds.): *Classification as a tool of research.* North Holland, Amsterdam, 447-451.

WINDHAM, M.P. (1987): Parameter modification for clustering criteria. *Journal of Classification 4, 191-214.*