Journal of Classification 22:287-310 (2005)
DOI: 10.1007/s00357-005-0018-3

Analysis of Global k-Means, an Incremental Heuristic for Minimum Sum-of-Squares Clustering

Pierre Hansen

GERAD and HEC Montréal
Montréal, Québec, Canada

Eric Ngai

The Hong Kong Polytechnic University
Hong Kong

Bernard K. Cheung

GERAD and École Polytechnique
Montréal, Québec, Canada

Nenad Mladenović

Mathematical Institute and GERAD
SANU, Belgrade, Yugoslavia

Abstract: The global k-means heuristic is a recently proposed (Likas, Vlassis and Verbeek, 2003) incremental approach for minimum sum-of-squares clustering of a set X of N points of $\mathbf{R}^d$ into M clusters. For $k = 2, 3, \ldots, M - 1$ it considers the best-known set of $k - 1$ centroids previously obtained, adds a new cluster center at each point of X in turn and applies k-means to each set of k centroids so-obtained, keeping the best k-partition found. We show that global k-means cannot be guaranteed to find the optimum partition for any $M \geq 2$ and $d \geq 1$; moreover, the same holds for all $M \geq 3$ if the new cluster center is chosen anywhere in $\mathbf{R}^d$ instead of belonging to X. The empirical performance of global k-means is also evaluated by comparing the values it obtains with those obtained for three data sets with $N \leq 150$ which are solved optimally, as well as with values obtained by the recent j-means heuristic and extensions thereof for three larger data sets with $N \leq 3038$.

Keywords: Clustering; k-means; j-means; Global k-means; Minimum sum-of-squares.

Author's Address: Pierre Hansen, GERAD and HEC Montréal, 3000 chemin de la Côte-Sainte-Catherine, Montréal, Québec, Canada H3T 2A7, tel: 514-340-6053-5675, fax: 514-340-5665, e-mail: Pierre.Hansen@GERAD.ca

1. Introduction

Cluster analysis aims at solving the following very general problem, which has numerous applications in the natural and social sciences as well as in medicine and engineering: given a set X of N entities, often described by measurements as points of the real d-dimensional space $\mathbf{R}^d$, find subsets of X which are *homogeneous* and/or *well-separated.* Homogeneity means that entities in the same cluster must be similar and separation that entities in different clusters must differ one from the other. These concepts can be made precise in a variety of ways, which lead to as many clustering problems and even more heuristic or exact algorithms. So, clustering is a vast subject. Good introductory texts are Kaufman and Rousseeuw (1990), Gordon (1981) and the survey Jain, Murty and Flinn (1999); another more mathematical survey is Hansen and Jaumard (1997).

In some rare cases, the criterion adopted expresses both homogeneity and separation. This is so for minimization of the sum-of-squared distances from all entities to the centroid of the cluster to which they belong. Indeed, minimizing the within clusters sum-of-squares, a criterion of homogeneity, is tantamount to maximizing the between clusters sum-of-squares, a criterion of separation. For short, we will refer to this criterion as "minimum sum-of-squares".

A mathematical programming formulation of the minimum sum-of-squares clustering problem is as follows:

$$\min f(\mathcal{M}, Z) = \sum_{i=1}^{M} \sum_{j=1}^{N} z_{ij} \cdot \|x_j - m_i\|^2$$

subject to

$$\sum_{i=1}^{M} z_{ij} = 1, \quad j = 1, 2, \dots, N,$$

$$z_{ij} \in \{0, 1\} \quad i = 1, 2, \dots, M; \; j = 1, 2, \dots, N,$$

where

$$m_i = \frac{\sum_{j=1}^{N} z_{ij} x_j}{\sum_{j=1}^{N} z_{ij}}, \quad i = 1, 2, \dots, M.$$

The N entities to be clustered are at given points $x_j = (x_{j1}, x_{j2}, \dots, x_{jd})$ of $\mathbf{R}^d$ for $j = 1, \dots, N$; M cluster centroids must be located at unknown points $m_i \in \mathbf{R}^d$ for $i = 1, \dots, M$. The decision variable z_{ij} is equal to 1 if point j is assigned to cluster i, at a squared Euclidean distance $\|x_j - m_i\|^2$ from its centroid. It is well-known that condition $z_{ij} \in \{0, 1\}$ may be replaced by

$z_{ij} \in [0,1]$, since in an optimal solution, each entity belongs to the cluster with the nearest centroid (ties being broken arbitrarily).

A combinatorial optimization formulation is as follows:

Let $P_M = \{C_1, C_2, \ldots, C_M\}$ denote a partition of X into M clusters (or classes):

$$C_i \neq \emptyset \quad \forall i = 1, \ldots, M, \quad C_i \cap C_j = \emptyset \quad \forall i,j = 1, \ldots, M,\ i \neq j$$

and

$$\cup_{i=1}^{M} C_i = X.$$

Let $\mathcal{P}_M$ denote the set of all M-partitions of X. Then find P_M^* such that

$$P_M^* = \min_{P_M \in \mathcal{P}_M} \sum_{i=1}^{M} \sigma^2(C_i),$$

where

$$\sigma^2(C_i) = \sum_{j | x_j \in C_i} \|x_i - m_i\|^2,$$

the sum of squared distances from entities of cluster C_i to its centroid m_i. Note that $P_M^* = (C_1^*, C_2^*, \ldots, C_M^*\}$ is such that

$$C_i^* = \{x_j \in X \mid z_{ij}^* = 1\},$$

where (z_{ij}^*) is the assignment matrix of the optimal solution of the mathematical programming formulation.

Minimum sum-of-squares clustering is among the most central problems of cluster analysis. It has been extensively studied since the sixties of last century, leading to several hundred papers on exact or approximate algorithms and their properties, as well as several thousand papers on generalizations and applications in various fields. A complete survey of this literature would require a long paper in itself.

The best-known heuristic for minimum sum-of-squares clustering is Mac Queen's (1967) *k-means*. It proceeds by selecting a first set of M points as candidate centroid set, then alternately (i) assigning points of X to their closest centroid and (ii) recomputing centroids of the clusters so-obtained, until stability is attained.

Many variants of this scheme have been proposed. We find of particular interest, that one recently proposed by Likas, Vlassis and Verbeek (2003). It advocates an incremental approach in which an initial solution for a partition in M clusters is obtained by adding one point to the set of centroids of the best points into $M-1$ clusters obtained. This point is usually chosen among those of X but the possibility of choosing a point of R^d in general is also mentioned.

Likas et al (2003) make a conjecture about the optimality of such a procedure. Their method has already attracted much interest (Godin, Huguet, Gaertner and Salmon 2004; Marques, Carvalho, Costa and Medeiros 2004; Pham, Dimov and Nguyen 2004; Schenken, Last, Bunke and Kandel 2004; Tsingos, Gallo and Drettakis 2004; Whitfield, Hall and Cannon 2004).

The purpose of the present paper is two-fold:

(i) to study the optimality of global k-means heuristics, first when a point of X is added at each iteration, then when a point of R^d is chosen, thus answering (in the negative) the conjecture of Likas et al.(2003).

(ii) to study empirically the version of Global k-means proposed by Likas et al. (2003) and compare it with the recent *j-means* heuristic (Hansen and Mladenović 2001a).

The paper is organized as follows: complexity of minimum sum-of-squares clustering is examined in the next section. A brief and selective survey of heuristics and exact algorithms for that problem is given in Section 3. The Global k-means and j-means heuristics are described in more details than other methods as they form the subject matter of this paper. In Section 4 we discuss for which values of M and d does global k-means always lead to an optimal partition. An empirical comparison of Global k-means and j-means is reported on in Section 5. Conclusions are drawn in Section 6. The rather technical proof of our main result is given in the Appendix.

2. Complexity of Minimum Sum-of-squares Clustering

To the best of our knowledge the computational complexity of minimum sum-of-squares clustering for general values of M and d is unknown. However, several incorrect statements have been made about this problem being known to be NP-hard (including one by two of the present authors in Hansen and Mladenović (2001a)). Reasons of these confusions are worth discussing.

First, in their classical book on *Computers and Intractability*, Garey and Johnson (1979) mention in their list of NP-hard problems:

MINIMUM SUM-OF-SQUARES

INSTANCE: Finite set A, a size $s(a) \in Z^+$ for each $a \in A$, positive integers $K \leq |A|$ and J.

QUESTION: Can A be partitioned into K disjoint sets $A_1, A_2, \ldots, A_K$ such that

$$\sum_{i=1}^{K} \left(\sum_{a \in A_i} s(a) \right)^2 \leq J.$$

A close look at this statement shows it is not the same as, nor a particular case of, the minimum sum-of-squares problem stated in the introduction.

Second, Brücker (1978) and Hansen and Delattre (1978) independently proposed a graph-theoretical proof that *Minimum diameter partitioning* is NP-hard (given dissimilarities between all pairs of entities, the diameter of a partition is the largest dissimilarity between a pair of entities in the same cluster). The reduction is to chromatic number: given a graph $G(V, E)$ with vertex set V and edge set E and an integer M is G $M-$colorable (i.e., colorable in M colors such that no pair of adjacent vertices receive the same color)? Taking $d_{kl} = 1$ if $\{v_k, v_l\} \in E$ and $d_{kl} = 0$ otherwise expressed this problem as a minimum diameter partitioning one: G is M-colorable is the optimum diameter is 0 and not if it is 1.

Welch (1982) examined if this proof technique could be extended to show NP-hardness of other clustering problems including minimum sum-of-squares clustering. However, this neglects the fact that points belonging to R^d severely restricts the possible values of the distances between them (e.g. one cannot have $d_{ij} = d_{jl} = 0$ and $d_{ik} = 1$, due to the triangle inequality).

If M and d are fixed, minimum sum-of-squares clustering can, in principle, be solved in polynomial time. Indeed, the number of bipartitions of X is polynomial so X can be bipartitioned in all possible ways, then the operation iterated $M - 1$ times on the clusters obtained. This is not practical for $d > 4$ (see Hansen, Jaumard and Mladenović 1998).

If $d = 1$ the minimum sum-of-squares clustering problem can be solved in $O(N^2M)$ time using dynamic programming as observed by several authors, e.g. Bellman and Dreyfus (1962) and Rao (1971).

A hierarchical agglomerative clustering method for minimum sum of squares clustering, known as Ward's method, has been proposed by Ward (1963) long ago and extensively applied (this paper is one of the most cited of the scientific literature: over 2000 times). A hierarchical divisive method has been developed more recently Hansen et al. (1998) and works well for small d, i.e., it can solve instances with $N \leq 20,000$ for $d = 2$, $N \leq 1,000$ for $d = 3$ and $N \leq 150$ for $d = 4$. As all other hierarchical methods, these two suffer from the defect that a non-optimal merging or splitting is never corrected. Therefore, partitioning (i.e., non-hierarchical) algorithms and heuristics have been proposed. A first branch-and-bound algorithm is presented in Koontz, Narendra and Fukunaga (1975) and elaborated on in Diehr (1985) Recently, jointly using several tools from mathematical programming (column generation, interior point, hyperbolic and quadratic 0-1 algorithms, together with Variable Neighborhood Search heuristics and branch-and-bound) led to an exact solution of problems with $N \leq 150$ (du Merle, Hansen, Jaumard and Mladenović (2000), including Fisher's Iris (Anderson 1935, Fisher 1936). However, numerous data sets have more than 150 entities.

Many heuristics have been proposed for the minimum some-of-squares clustering problem. The best known one is *k-means* (MacQueen 1967). It proceeds from a seed solution consisting of M points $m_1, m_2, \ldots, m_M$ (not necessarily belonging to X) considered as tentative centroids, by alternatively:

(i) allocating each point to its closest centroid, thus obtaining an M-partition $C_1, C_2, \ldots, C_M$ of X (where C_j denotes the set of points closer to m_j than to any other m_i with $i \in \{1, \ldots, M\}$;

(ii) recomputing centroids $m_1, m_2, \ldots, m_M$ for the clusters $C_1, C_2, \ldots, C_M$ so-obtained and returning to (i) until no more points change cluster.

Observe that in k-means, a cluster may, at some iteration, become empty, a phenomenon known as *degeneracy*. It is then worthwhile to choose a point, which does not yet coincide with a centroid, to replace the centroid of the cluster which has become empty (Hansen and Mladenović 2001a).

Another heuristic for minimum sum-of-squares clustering (Späth 1985, Hansen and Mladenović 2001a) consists at each iteration in moving an entity from its cluster to another one in such a way that the objective function value is most reduced. This is repeated until no more such move improves that value. In this paper, we shall refer to that heuristic as h-means.

Usually, one allocation step (i) of k-means implies several reassignments of entities. As noted in Hansen and Mladenović (2001), local minima obtained by k-means may be improved by h-means, while the converse is not true.

The k-means heuristic, despite being much used, has several defects: (a) it stops in a local optimum which can be far from the optimum; examples show the value obtained for large N and M can be several times the optimal one (Hansen and Mladenović 2001a); (b) the solution obtained depends largely on the seed solution used (see e.g. Peña, Lozano and Larañaga 1999); (c) computing time may be substantial for very large data sets, such as those considered in data mining (other methods, however are even more time-consuming).

To alleviate, as far as possible, the defects (a) and (b) two new and rather similar heuristics for minimum sum-of-squares clustering have recently been proposed.

On the one hand, the heuristic *j-means* (for *jump*-means, Hansen and Mladenović 2001a) extends k-means by adding a jump move, i.e., a point of X where there is no centroid is chosen, considered as a new centroid, and replaces that centroid among the M previous ones whose deletion augments least the objective function value, all other centroids remaining fixed. All such possible centroid-to-entity relocation moves define the *jump neighborhood* of the current solution, which is systematically explored. Once the best (*add,drop*) pair is found, the corresponding reassignments are performed, and possibly improved by k-means. The procedure is iterated until a local optimum for the jump neighborhood is reached. While the h-means heuristic improves the search by refining the allocation step (i) of k-means, j-means concentrates on improving k-means's location step (ii).

It is also possible to find the best add/drop pair in a more precise way by using k-means (or at least its first iteration) after each add step, but, as explained in Hansen and Mladenović (2001), in an efficient implementation, all the jump neighborhoods can be explored in $O(MN)$ operations, which is much less than $O(N)$ applications of k-means. Two variants of j-means are tested in Hansen and Mladenović (2001): one that only reassigns entities after finding the best add/drop pair (called j-means) and another that use k-means and then h-means to improve that best add/drop solution (j-means+).

A more powerful extension of j-means is to embed it in the framework of the Variable Neighborhood Search (VNS) metaheuristic (Mladenović and Hansen 1997, Hansen and Mladenović 2001b). That metaheuristic exploits a descent method such as j-means or j-means+ together with a systematic change of neighborhoods within the search space. To that effect, given any solution $P_M = \{C_1, C_2, \ldots, C_M\}$, a set of neighborhoods $N_1(P_M)$, $N_2(P_M), \ldots,$ $N_{k_{max}}(P_M)$ is defined. For minimum sum-of-squares clustering it correspond to all possible sequences of $1, 2, \ldots, k_{max}$ centroid-to-entity relocation moves. Once a local optimum P_M is found by j-means, a solution P'_M is chosen randomly in its first neighborhood $N_1(P_M)$ (or in other words one proceeds to a jump move) and used as initial solution for a descent, using again j-means (or j-means+). If the value of the locally optimal solution P''_M obtained at the end of this descent is worse than the best-known one, or incumbent, one proceeds to the random choice of a solution within the next neighborhood $N_{k+1}(P_M)$. If this value is better, it is stored, together with P''_M and the search is recentered around P''_M. If neighborhood k_{max} is reached without any improvement one begins again at the first neighborhood, until a stopping condition (e.g. maximum computing time, or number of iterations, or number of iterations since the last improvement of the best known solution) is met.

Solutions obtained by j-means may be substantially better than those obtained by *multistart k-means*, in which k-means is repeated from randomly generated initial partitions ans the best result high, and (due to the fact that descents from a perturbed local optimum are shorter than from a random solution) obtained in less computing time Hansen and Mladenović 2001a).

On the other hand, the *global* $k - means$ heuristic (Likas *et al.* 2003), is an incremental approach which builds partitions of X into $k = 1, 2, \ldots, M$ clusters successively. At a current iteration, for $k \geq 2$, its steps are as follows:

(i) consider the centroids $m_1(k-1), m_2(k-1), \ldots, m_{k-1}(k-1)$ of the best partition obtained at the previous iteration (into $k-1$ clusters);

(ii) add in turn, each point of X to this set of centroids, thus obtaining N initial solutions with k points; apply k-means to each of them; keep the best k-partition so-obtained and its centroids $m_1(k), m_2(k), \ldots, m_k(k)$;

(iii) augment k by 1 and return to (i) as long as $k \leq M$.

Note that, as the locally optimal partitions with $k = 2, 3, \ldots, M$ are kept, one can apply tests (Milligan and Cooper 1985, Likas 2003) to determine the best number of clusters without further effort.

Computing time of global k-means is fairly large, as $M \cdot N$ applications of k-means are made; two procedures are proposed to reduce it (possibly at the cost of obtaining less good solutions). In the first one, the effect of the first iteration of k-means is evaluated for all possible additions of a new point; then k-means is applied to the solution corresponding to the greatest first-step reduction in objective function value (which is clearly a lower bound on the total reduction obtained with k-means). This short-cut does not seem to affect much the value of the best solution obtained. In the second one, which applies to low-dimensional data, an efficient data structure for handling points known as a $k - d$ tree (Bentley 1991, Sproull 1991) is used to partition X into $N' << N$ subsets; their centroids are used as initial points instead of the more numerous points of X in the global k-means scheme.

Observe that the jump move of j-means and the addition of a point of X in global k-means are fairly similar. This is even more so when one considers a particular version of j-means introduced to study its relationship with the greedy heuristic. In Hansen and Mladenović (2001a) it is noted that:

> "…j-means can be viewed as an extended *Greedy* heuristic. [i.e., a heuristic performing the best move at every iteration, in a myopic way]. Indeed, assume that all points are initially assigned to the same cluster, i.e., all M centroids are located at the same far away point (for example at origin). Then, in each iteration a new centroid is added, and one deleted from the origin. However, the *Greedy* heuristic stops when the number of origin centroids becomes zero, while *j-means* could continue the search".

Global k-means thus appears to be a version of the previous scheme in which a larger effort is made in the choice of the new centroids, and none to improve further the M-partition obtained at the last iteration.

Recently, Kanungo,Mount, Netanyahu, Piatko, Silverman and WU (2002) proposed a single and practical approximation algorithm based on swapping centers in and out which has $9 + \varepsilon$ approximation factor. Further results, involving the spread of X, i.e., the ratio of its diameter to distance between its two closest points are given in Har-Peled and Mazumdar (2004), and Har-Peled and Sadri (2005). Of particular interest are lower and upper bounds on the number of iterations k-means.

3. Some Heuristics and Exact Algorithms for Minimum Sum-of-squares Clustering

The minimum sum-of-squares clustering problem as well as the k-means heuristic have been generalized in many ways. We mention a few:

(i) *Generalized centroid:* in his important early work on the method of "dynamic clusters" Diday (1972, 1974) proposed to replace the centroid of a cluster by a more general set of points or a surface which would better represent it. This can be done in many ways while keeping the essence of the k-means heuristic at the center of the resolution method. Diday also proposed to study "strong forms", i.e., common subsets of entities within the classes of the partitions obtained in several runs.

(ii) *Fuzzy clustering:* instead of imposing that each point belong to one and only one cluster, it may be allowed that it belong to several, with different degrees of membership, summing to 1. The mathematical programming model of the introduction is then modified by giving an exponent $\alpha_{ij} \in (0, 1)$ to each z_{ij} in the objective function Bezdek (1980). Several variants of k-means and j-means (Belacel, Hansen and Mladenović, 2003) have been proposed to solve this problem.

(iii) *Expectation-maximization:* the k-means heuristic can be generalized to the extension of missing data, see Estivill-Castro and Yang (2004), for thorough discussion.

(iv) *Categorical and mixed data sets:* when the data consist in categorical observations instead of measurements of real values, the k-modes heuristic can be modified into a k-median heuristic e.g. Huang (1998): when both categorical and real data are present one can compute a mode/centroid by doing computation as above on real and categorical components of points according to the centroid and median calculations.

(vi) *On-line scheduling:* large data set arriving in a continuous stream can be classified by on-line extensions of k-means (Bermejo and Cabestany 2002; Bougeuettaya 1996; El-Sonbaty and Ismail 1998).

(v) *Multicriteria clustering:* (De Smet and Montano-Guzman 2004).

4. Some Small Hard to Solve Examples

The authors of the global k-means method comment on its ability to find the optimal solution as follows (Likas et al., 2003):
"The rationale behind the proposed method is based on the following assumption: an optimal clustering solution with k clusters can be obtained through local search (using k-means) starting from an initial state with

- the $k-1$ centers placed at optimal positions for the $(k-1)$ - clustering problem and
- the remaining k^{th} center placed at an appropriate position to be discovered."

Choice of position for the k^{th} center leads to a further assumption:
"It is also reasonable to restrict the set of possible initial positions to the set X of available data points".

Then favorable experimental results, as compared with using numerous random restarts of k-means, suggest to
"cautiously state that the proposed method is *experimentally optimal* (although it is difficult to prove theoretically)".

In this section we examine, through examples and propositions, for which values of the number M of clusters and the dimension d of the space considered, do the assumptions necessarily hold. We first consider both assumptions together; the more difficult case where only the first assumption is made, i.e., the initial position of the k^{th} center is not restricted to X, will be examined afterwards.

Example 1 *Let* $N = 4$, $d = 2$, $x_1 = (0,1)$, $x_2 = (1,0)$, $x_3 = (0,-1)$, $x_4 = (-1,0)$ *and* $M = 2$ *(see Figure ?? (a)). For* $k = 1$, $C_1 = \{x_1, x_2, x_3, x_4\}$ *and* $m_1 = (0,0)$ *is the centroid.*

For $k = 2$, by symmetry, only one of the four points, say x_1, needs to be considered as initial position for m_2. Then $C_1 = \{x_2, x_3, x_4\}$, $C_2 = \{x_1\}$, (see Figure ?? (b)). Computing centroids one gets $m_1^* = (-1/3, 0)$ and $m_2 = (1,0)$ (see Figure ?? (c)). The error is $2[1^2 + (\frac{1}{3})^2] + (\frac{2}{3})^2 = \frac{8}{3}$.

However, consider the partition $C_1' = \{x_2, x_3\}$, $C_2' = \{x_1, x_4\}$, which has centroids $m_1' = (\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}})$ and $m_2' = (-\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$ (see Figure ?? (d)). The error is then $4(\frac{1}{\sqrt{2}})^2 = 2$ and it is easily seen that this partition is optimal. So in this case global k-means stops with an error exceeding the optimal value by $(8/3-2)/2 = 33.3\%$.

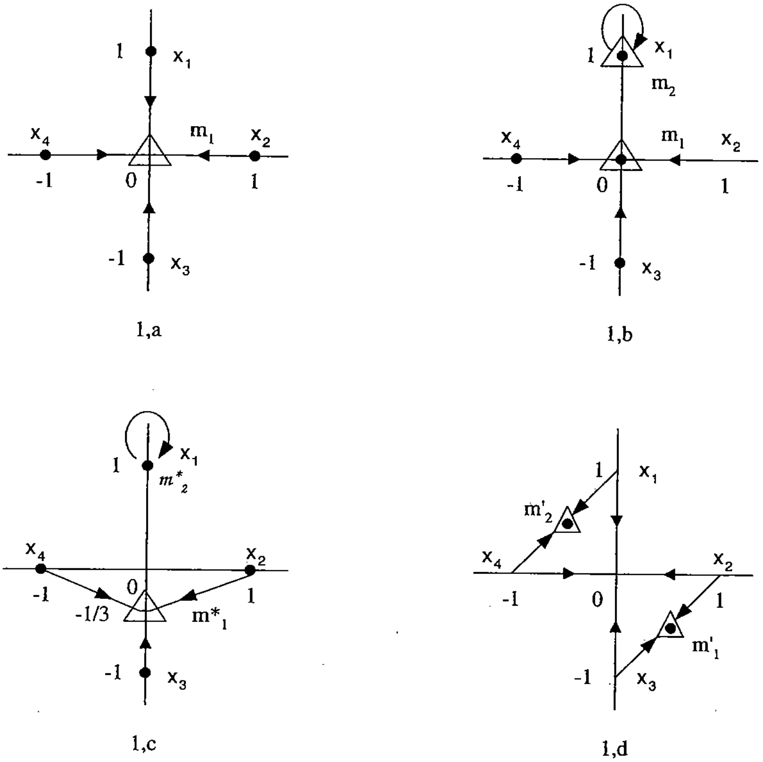


Figure 1. Example 1.

Similar results hold for $M > 2$. Indeed, modifying Example ?? by adding $M-2$ points $x_5 = (4,0), x_6 = (7,0), \ldots, x_{M+2} = (1+3(M-2), 0)$ (see Figure ??) one gets an example for which the optimal $(M-1)$ - clustering is $C_1 = \{x_1, x_2, x_3, x_4\}$, $C_2 = \{x_5\}$, $C_3 = \{x_6\}, \ldots$, $C_M = \{x_{M+2}\}$, $m_1 = (0,0), m_2 = (4,0), m_3 = (7,0), \ldots, m_{M-1} = (1+3(M-2), 0)$, as the additional points are sufficiently far from the others and between themselves to form 1 entity clusters. The analysis done for Example ?? then carries over.

Moreover, as the d-dimensional Euclidean space $R^d \subset R^{d'}$ for $d' > d$ the above example can be viewed as belonging to d'-space for any $d' > 2$. If, however, one wishes to have points in general position in d'-space, this can be achieved by (i) duplicating all points of Example ?? or its extension a sufficient number of times to have at least d' points, and (ii) perturbing very slightly the first d' points in turn along the 1^{st}, 2^{nd}, $\ldots$, d'^{th} axis.

There remains an open case, i.e., $d = 1$. Note that in this case an $O(N^3)$ dynamic programming algorithm is available (Späth 1980). Nevertheless, let us examine if the two assumptions hold.

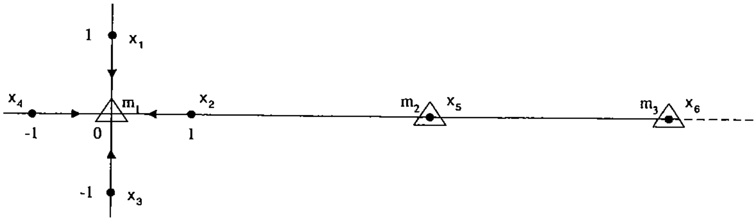


Figure 2. Example 1 modified for $M > 2$.

Example 2 *Let $N = 6$, $d = 1$, $x_1 = 0$, $x_2 = 2$, $x_3 = 4$, $x_4 = 7$, $x_5 = 9$, $x_6 = 11$ and $M = 3$ (see Figure* **??** *(a)). For $k = 2$ it is easily seen that the best partition is $C_1 = \{x_1, x_2, x_3\}$ and $C_2 = \{x_4, x_5, x_6\}$ with $m_1^* = 2$ and $m_2^* = 9$. The error is $2 \cdot 2 \cdot (2^2) = 16$ (see Figure* **??** *(b)). As there are already centroids at x_2 and x_5 the initial position for m_3 may be x_1, x_3, x_4 or x_6. Consider x_1. Then $C_3 = \{x_1\}$, $C_1 = \{x_2, x_3\}$, $C_2 = \{x_4, x_5, x_6\}$, $m_3^* = 0$, $m_1^* = 3$, $m_2^* = 9$ and the error is 0+2 $\cdot 1^2$ + 2 $\cdot$ 2^2 = 9 (see Figure* **??** *(c)). Taking x_3, x_4 or x_6 as initial positions for the new center yields partitions with classes containing 1 point, 2 points at distance 2 and 3 points with the central one at distance 2 from each of the outer ones. So the error is the same as in the first case.*

However, consider the partition $C_1' = \{x_1, x_2\}$, $C_2' = \{x_3, x_4\}$, $C_3' = \{x_5, x_6\}$ with centroids $m_1' = 1$, $m_2' = 5.5$ and $m_3' = 10$. The error is then $2 \cdot 1^2 + 2 \cdot (2.25)^2 + 2 \cdot 1^2 = 8.5$ (see Figure **??** (d)). So, in this case, global k-means stops with an error exceeding the optimal value by (9-8.5)/8.5 = 5.98 %. Again, this example can be modified by adding points at sufficient distance from those already chosen and one from another to obtain a counter-example with $M > 3$.

Finally, a single case remains when initial points for the k^{th} centroid are chosen in X: $d = 1$ and $M = 2$. For that case we have only been able to build a counter-example with a very large number of points to cluster; all of these are located at four distinct points of the real line.

Example 3 *Let $N = 10010$, $d = 1$, $x_1 = x_2 = x_3 = 0$, $x_4 = 604$, $x_5 = \ldots, x_{10}$=902, $x_{11} = \ldots x_{10010} = 1202$ and $M = 2$ (see Figure* **??** *(a)). Then $m_1 = (3.0 + 1.604 + 6.902 + 10000.1202)/10010 = 1201.40$ and the error with a single cluster is*

$$3 \cdot (1201.40)^2 + (597.40)^2 + 6 \cdot (299.40)^2 + 10000(0.60)^2 = 5,228,414.8.$$

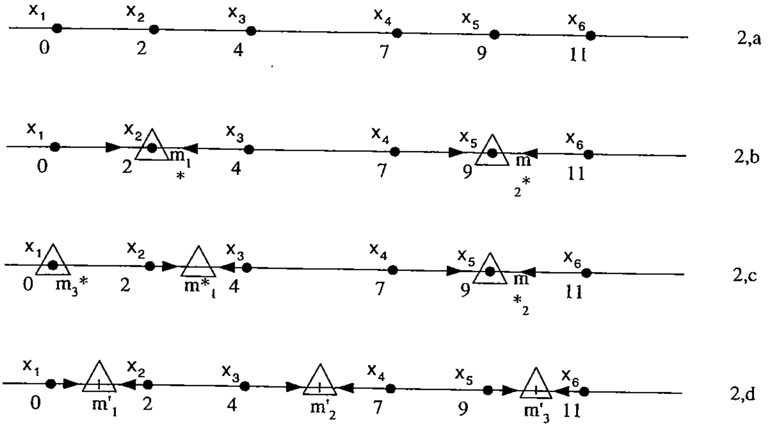


Figure 3. Example 2.

Taking $x_1 = 0$ as initial position for m_2, a partition $C'_1 = \{x_4, \ldots, x_{10010}\}$, $C'_2 = \{x_1, x_2, x_3\}$ is obtained as the distance between x_1 and x_4 is 604 > 597.40, i.e. the distance between x_4 and m_1. By an easy computation $m_2^* = 1201.76$ and no point changes cluster anymore. The error is 897,029 (see Figure ?? (b)). Taking $x_4 = 604$, a partition $C_1 = \{x_{11}, \ldots, x_{10010}\}$ and $C'_2 = \{x_1, x_2, \ldots, x_{10}\}$ is obtained as the distance between x_4 and $x_5, x_6, \ldots, x_{10}$ is 902 - 604 =298 < 1201.40 - 902 = 299.40, i.e., the distance between x_{10} and m_1. By an easy computation $m_1^* = 1202$, $m_2^* = 601.6$ and the error is 1,627,214 (see Figure ?? (c)).

Taking $x_5 = 902$ or $x_{11} = 1202$ as initial position for m_2 yields in both cases the same partition as when taking $x_1 = 0$ (up to a permutation of indices in one case). That partition is the best one for global k-means. However, consider the partition $C'_1 = \{x_1, x_2, x_3, x_4\}$ and $C'_2 = \{x_5, x_6, \ldots, x_{10010}\}$ with centroids $m'_1 = 151$ and $m'_2 = 1201.82$. Then error is 813,288. So in this case global k-means stops with an error exceeding the optimal value by (897029-813288)/813288 = 10.30%.

We summarize results obtained up to now as follows:

Proposition 1 *For all real spaces $\mathbf{R}^d$, with $d \geq 1$ and all numbers of clusters $M \geq 2$, there are instances of minimum sum-of-squares clustering such that beginning with the $M - 1$ centroids of an optimal $(M - 1)$-partition and any point $x_j \in X$, k-means never gets an optimal M-partition.*

Let us now turn to the more general case in which the k^{th} initial point can be chosen anywhere in $\mathbf{R}^d$ and not only as a point of X. We then get a positive result for $M = 2$.

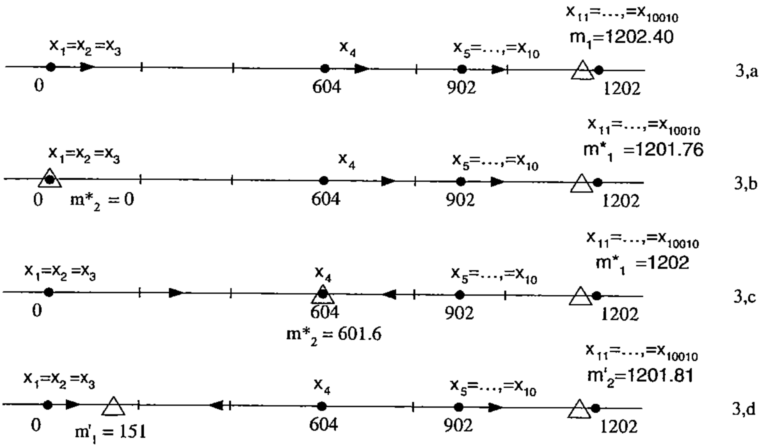


Figure 4. Example 3.

Proposition 2 *For any minimum sum-of-squares clustering problem in R^d with $M = 2$, there exists a point $y \in \boldsymbol{R}^d$ such that choosing y and the centroid of X (with small perturbation in case of coincidence) as initial points k-means gives an optimal clustering.*

Proof. Assume not all points of X coincide (otherwise, partitioning is arbitrary). Consider an optimal solution with clusters C_1, C_2 and centroids m_1 and m_2. Then the hyperplane H perpendicular to the line segment joining m_1 and m_2 and passing through its middle separates the points from C_1 from those of C_2 (Gower 1967); points on the hyperplane, if any, could be moved very slightly towards their centroids. Then consider the centroid $m_1(1)$ of X and its symmetric y with respect to H. Assume first that $m_1(1)$ is not on H. Then choosing $m_i(1)$ and y as initial centers gives the partition (C_1, C_2) (up to permutation of indices). So k-means stops after one iteration with the optimal partition. Then consider the case where $m_1(1)$ is on H; take initial centers m_1' and m_2' on the line-segment through $m_1(1)$ parallel to the line-segment joining m_1 and m_2 at a small distance ε from H on both sides. Again the optimal partition is obtained after one iteration of k-means. □

Unfortunately, the proof of Proposition 2 is not constructive. Moreover, if $M \geq 3$, again there are small hard to solve cases, which leads us to our main result. Its proof, being rather technical, is relegated to the Appendix.

Theorem 1 *For all real spaces $\boldsymbol{R}^d$, with $d \geq 1$ and all numbers of clusters $M \geq 3$, there are instances of minimum sum-of-squares clustering such that* *beginning with the $M - 1$ centroids of an optimal ($M - 1$)-partition and any point $y \in \boldsymbol{R}^d$, k-means never gets an optimal M-partition.*

This impossibility result can also be expressed in an alternative way: **Theorem 1'** *There is no global k-means algorithm which is optimal for $M \geq 3$ and $d \geq 1$.*

5. Computational Results

Global k-means has been shown to give better results than multistart k-means (Likas *et al.* 2003) in which k-means is repeated from randomly chosen initial solutions until a time-limit is reached and the best solution kept. It is worthwhile, however, to evaluate how close the solutions obtained are from the optimal ones. This is done in the present section, in two ways. First, global k-means, multistart k-means, j-means and j-means combined with Variable Neighborhood Search are applied to three data sets with $N \leq 150$ and $M \leq 20$ for which optimal solutions are known (du Merle et al. 2000). Second, three larger data sets from the Irvine repository (Blake and Merz 1998) and TSP-Lib (Reinelt 1991) are considered and the same heuristics applied. All computations are done on a Sun Ultra 2, 450 Mhz workstation.

The three data sets with known optimal solutions are:

- 4-dimensional data on 150 iris from the Gaspé peninsula (Anderson 1935; Fisher 1936);
- 3-dimensional data on 89 postal districts in Bavaria, Germany (Späth 1985)
- 2-dimensional data on 75 points in $\mathbf{R}^2$, from an artificial, but much studied, example of Ruspini (1970).

Results of the first series of experiments are presented in Table **??**. The first two columns give the number N of points and M of clusters of the partitions. Optimal values of the sum-of-squares, determined by the algorithm of du Merle et al. (2000) are recalled in column 3. The next four columns give % error for global k-means (G-1 for short) and fast global k-means (G-2 for short, i.e., the version evaluating the effect of adding a center at a point of X, applying only a first iteration of k-means then full k-means after the best addition is found) as well as computing times for both versions. The same computing time as taken by global k-means is given to k-means, j-means+ and VNS with in the last case two variants: 10 restarts in 1/10 of that time for column VNS, the same time without restarts for column VNS-1. Average results for all restarts are given in columns 8 to 11 and best results in columns 12 to 15.

Table 1. Comparative results for problems with known optimal solution

		Optimal	Error		Time		Average error			Error of the best			Error
N	*M*	value	G-1	G-2	G-1	G-2	K-M	J-M+	VNS	K-M	J-M+	VNS	VNS-1
150	2	152.35	0.00	0.00	0.06	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	3	78.851	0.00	0.01	0.21	0.04	13.35	0.00	0.00	0.00	0.00	0.00	0.00
	4	57.228	0.00	0.00	0.37	0.06	11.26	2.40	0.00	0.00	0.00	0.00	0.00
	5	46.446	0.00	0.06	0.51	0.09	13.83	3.97	1.46	0.00	0.00	0.00	0.00
	6	39.040	0.00	0.07	0.66	0.12	16.21	4.26	0.01	0.00	0.00	0.00	0.00
	7	34.298	0.02	1.55	0.85	0.15	17.43	2.86	0.00	0.00	0.00	0.00	0.00
	8	29.989	0.01	0.25	1.05	0.19	20.81	2.76	0.02	0.00	0.00	0.00	0.00
	9	27.786	0.01	0.82	1.25	0.23	18.80	1.62	0.00	0.00	0.00	0.00	0.00
	10	25.834	0.51	1.00	1.42	0.27	17.43	2.84	0.42	0.16	0.00	0.00	0.00
	Average		0.06	0.42	0.71	0.13	14.35	2.30	0.21	0.02	0.00	0.00	0.00
89	2	$0.60255 \cdot 10^{12}$	0.00	0.00	0.02	0.00	7.75	0.00	0.00	7.75	0.00	0.00	0.00
	3	$0.29451 \cdot 10^{12}$	0.00	0.00	0.05	0.01	23.40	0.00	7.04	20.02	0.00	0.00	0.00
	4	$0.10447 \cdot 10^{12}$	0.00	0.00	0.09	0.02	156.17	0.00	0.00	0.08	0.00	0.00	0.00
	5	$0.59762 \cdot 10^{11}$	0.00	0.00	0.14	0.02	315.28	0.00	0.00	23.58	0.00	0.00	0.00
	6	$0.35909 \cdot 10^{11}$	0.00	0.00	0.19	0.03	531.44	27.70	11.06	27.79	27.65	0.00	0.00
	7	$0.21983 \cdot 10^{11}$	0.00	0.00	0.25	0.04	832.60	44.00	7.07	69.39	0.00	0.00	0.00
	8	$0.13385 \cdot 10^{11}$	0.00	0.00	0.33	0.05	1239.64	0.24	0.00	0.00	0.00	0.00	0.00
	9	$0.84237 \cdot 10^{10}$	0.00	0.00	0.38	0.06	1697.17	28.59	0.00	35.81	0.00	0.00	0.00
	10	$0.64465 \cdot 10^{10}$	0.00	0.00	0.44	0.07	1638.30	0.16	0.00	57.81	0.00	0.00	0.00
	14	$0.21155 \cdot 10^{10}$	1.48	0.11	0.65	0.13	1922.39	11.48	0.00	67.10	0.00	0.00	0.00
	18	$0.98069 \cdot 10^{9}$	0.00	0.00	0.86	0.20	2703.85	5.62	1.11	244.13	0.00	0.00	0.00
	22	$0.54214 \cdot 10^{9}$	0.00	0.00	1.12	0.29	4735.31	18.55	5.56	228.89	4.71	0.00	0.00
	26	$0.28223 \cdot 10^{9}$	6.44	9.14	1.42	0.38	8835.90	6.19	0.00	105.62	0.00	0.00	0.00
	30	$0.17138 \cdot 10^{9}$	0.00	3.80	1.70	0.49	14032.88	8.17	0.53	171.98	0.00	0.00	0.00
	Average		0.57	0.93	0.55	0.13	2762.2 9	10.76	2.31	75.71	2.31	0.00	0.00
75	2	89337.83	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	3	51063.47	0.00	0.00	0.02	0.01	0.09	0.05	0.07	0.00	0.00	0.00	0.00
	4	12881.05	0.00	0.00	0.03	0.01	141.57	0.00	0.00	0.00	0.00	0.00	0.00
	5	10126.72	0.00	0.22	0.05	0.01	109.22	6.04	2.38	0.00	0.00	0.00	0.00
	6	8575.41	0.00	0.98	0.06	0.02	81.06	2.09	0.90	0.00	0.00	0.00	0.00
	7	7126.20	0.00	1.69	0.08	0.03	63.96	5.71	0.68	0.00	0.00	0.00	0.00
	8	6149.64	0.14	1.73	0.10	0.03	48.94	4.32	0.81	0.27	0.00	0.14	0.00
	9	5181.65	0.32	1.78	0.12	0.04	50.62	7.76	2.81	0.25	0.00	0.00	0.00
	10	4446.28	0.38	2.73	0.15	0.05	52.83	7.22	0.08	0.29	0.00	0.00	0.00
	15	2559.35	0.19	2.73	0.25	0.09	58.81	6.06	2.72	5.02	0.00	0.00	0.00
	20	1721.22	0.65	0.04	0.38	0.15	51.97	6.93	2.65	10.27	1.75	0.91	0.00
	25	1162.67	1.24	1.24	0.53	0.21	57.72	10.45	1.74	18.81	0.47	0.00	0.00
	30	741.83	1.46	3.27	0.69	0.29	81.97	9.97	1.67	20.51	0.00	0.00	0.00
	Average		0.34	1.26	0.19	0.07	61.44	5.12	1.27	4.26	0.17	0.08	0.00

It appears that:

(i) global *k*-means (G-1) obtains optimal partitions for small *M* (up to 6 for iris, 10 for postal districts, 7 for points in the plane) but usually not for larger values;

(ii) fast global k-means (G-2) divides the computing time of global k-means by a factor of about 3 but at the price of a notable increase in both the number of non-optimal partitions and the size of the errors;

(iii) results of multistart k-means are problem dependent (but always bad in average for one descent): the iris problem is solved optimally except for $M = 10$, best partitions with an even number of clusters for the postal district problem have very large errors (up to more than 200 % !) and partition into 7 clusters at most for the points in the plane are optimal but errors rapidly increase for larger values of M;

(iv) j-means+ gives optimal partitions in all cases but four (although average results for its use until a restart takes place show substantial errors);

(v) VNS always attains optimal partitions (with errors in 2 cases only if 10 runs are made with 1/10 of the time of global k-means).

The three larger data sets are:

- 9-dimensional data on 214 glass specimens for forensic identification (Blake and Merz 1998);
- 19-dimensional data on 2310 instances drawn randomly as parts of 7 outdoors images (Blake and Merz 1998);
- 2-dimensional data on 3038 points in the plane from a traveling salesman problem of Reinelt (1991).

Results of this second series are presented in Table **??**. Columns are similar to those of Table **??** except that the third one contains values of the best solutions found during our experiments instead of optimal values, which are unknown.

It appears that, in addition to conclusions similar to those given for Table **??**:

(vi) for large instances, with $N > 2000$, global k-means gives good results, i.e., the best known ones when $M \leq 20$, and with moderate error otherwise;

(vii) computing time of global k-means becomes high (up to about 6 hours per instance) for $N > 2000$; reduction in this time when using fast global k-means is now much larger and again at the price of a notable increase in the number of suboptimal partitions and size of the errors;

(viii) j-means combined with Variable Neighborhood Search provides best known solutions in all but a few cases in 10 times less computing time than the global k-means.

Table 2. Comparative results on large test instances

		Best kn.	Error		Time		Average error			Error of the best			Error	Time
N	M	value	G-1	G-2	G-1	G-2	K-M	J-M+	VNS	K-M	J-M+	VNS	VNS-1	VNS-1
214	2	819.63	0.00	0.00	0.37	0.06	3.86	0.16	0.06	0.00	0.00	0.00	0.00	0.37
	3	589.03	0.00	0.00	0.90	0.13	12.13	0.00	0.00	0.00	0.00	0.00	0.00	0.90
	4	489.04	0.00	0.00	1.53	0.22	10.67	1.41	0.00	0.00	0.00	0.00	0.00	1.53
	5	400.26	0.01	0.07	2.24	0.32	12.72	0.63	0.00	0.01	0.00	0.00	0.00	2.24
	6	336.06	0.00	0.07	3.16	0.43	19.73	3.38	0.01	0.00	0.00	0.00	0.00	3.16
	7	292.25	0.00	0.13	4.09	0.55	21.46	2.38	0.00	0.00	0.00	0.00	0.00	4.09
	8	266.50	0.09	0.00	5.18	0.69	19.32	1.69	0.00	0.09	0.00	0.00	0.00	5.18
	9	245.35	0.00	0.40	6.54	0.84	17.80	1.75	0.20	0.00	0.00	0.00	0.00	6.54
	10	225.19	0.00	0.00	7.63	1.00	18.34	2.13	0.10	0.01	0.00	0.00	0.00	7.63
	20	114.65	0.78	0.00	22.14	3.26	50.06	5.33	1.03	16.97	1.56	0.00	0.00	22.14
	30	63.25	4.11	1.42	41.39	6.70	111.28	5.94	0.09	43.00	0.72	0.00	0.00	41.39
	40	39.50	2.27	2.97	60.29	11.32	172.56	9.01	0.72	63.03	3.39	0.00	0.00	60.29
	50	26.78	0.31	1.30	79.52	16.83	231.78	7.69	1.22	67.10	3.01	0.27	0.00	79.52
	Average		0.58	0.49	18.08	3.26	53.98	3.19	0.26	14.63	0.67	0.02	0.00	18.08
2310	2	$0.35606 \cdot 10^8$	0.00	0.00	97.58	7.92	0.17	0.00	0.00	0.00	0.00	0.00	0.00	9.76
	3	$0.27416 \cdot 10^8$	0.00	0.00	198.96	17.38	1.38	0.37	0.46	0.00	0.00	0.00	0.00	19.90
	4	$0.19456 \cdot 10^8$	0.00	0.00	313.01	28.46	25.32	0.00	0.00	0.00	0.00	0.00	0.00	31.30
	5	$0.17143 \cdot 10^8$	0.00	0.00	497.86	41.05	23.48	0.81	0.10	0.00	0.00	0.00	0.00	49.79
	6	$0.15209 \cdot 10^8$	0.00	0.47	707.27	54.82	16.41	2.47	2.87	0.60	0.81	0.81	0.00	70.73
	7	$0.13404 \cdot 10^8$	0.00	0.51	981.72	70.01	12.49	5.39	2.53	0.00	1.80	0.00	0.00	98.17
	8	$0.12030 \cdot 10^8$	0.00	0.58	1249.57	86.59	10.94	6.29	0.16	0.17	0.00	0.00	0.00	124.96
	9	$0.10784 \cdot 10^8$	0.00	0.64	1509.90	104.85	13.58	6.92	0.32	0.00	0.00	0.00	0.00	150.99
	10	$0.97952 \cdot 10^7$	0.00	1.75	1822.81	124.43	15.83	3.65	0.18	0.63	0.00	0.00	0.00	182.28
	20	$0.51283 \cdot 10^7$	0.03	0.46	6033.26	401.95	34.63	3.46	0.97	7.29	0.78	0.00	0.00	603.33
	30	$0.35076 \cdot 10^7$	0.00	0.16	11270.80	825.71	44.68	5.62	0.40	10.80	0.95	0.51	0.23	1127.08
	40	$0.27398 \cdot 10^7$	0.15	0.31	17052.50	1396.75	53.43	4.90	0.43	14.33	1.14	0.00	0.00	1705.25
	50	$0.22249 \cdot 10^7$	0.21	1.06	24174.30	2112.44	58.51	4.32	0.35	19.65	0.65	0.00	0.00	2417.43
	Average		0.03	0.46	5069.96	405.57	23.91	3.40	0.67	4.11	0.47	0.10	0.02	507.00
3038	2	$0.31688 \cdot 10^{10}$	0.00	0.00	71.93	4.73	0.00	0.00	0.00	0.00	0.00	0.00	0.00	7.19
	3	$0.21763 \cdot 10^{10}$	0.00	0.00	196.16	10.78	1.55	1.29	1.37	0.00	0.00	0.00	0.00	19.62
	4	$0.14790 \cdot 10^{10}$	0.00	0.00	328.89	17.63	0.03	0.00	0.00	0.00	0.00	0.00	0.00	32.89
	5	$0.11982 \cdot 10^{10}$	0.00	0.00	539.29	25.67	0.12	0.11	0.10	0.00	0.00	0.00	0.00	53.93
	6	$0.96918 \cdot 10^9$	0.00	0.02	730.64	34.76	1.22	1.98	0.01	0.00	0.00	0.00	0.00	73.06
	7	$0.83966 \cdot 10^9$	0.00	0.84	1012.89	44.91	1.65	1.48	0.73	0.00	0.00	0.00	0.00	101.29
	8	$0.73475 \cdot 10^9$	0.00	1.28	1226.30	56.05	1.90	1.48	0.62	0.00	0.00	0.00	0.00	122.63
	9	$0.64477 \cdot 10^9$	0.00	1.41	1589.68	68.15	1.47	0.99	0.11	0.00	0.01	0.00	0.00	158.97
	10	$0.56025 \cdot 10^9$	0.00	0.00	1890.17	81.37	2.44	1.81	0.06	0.00	0.00	0.00	0.00	189.02
	20	$0.26681 \cdot 10^9$	0.01	0.42	6412.25	258.94	3.16	2.60	0.09	0.02	0.05	0.01	0.00	641.22
	30	$0.17557 \cdot 10^9$	0.00	1.48	11155.20	528.63	4.04	2.89	0.91	0.60	0.42	0.01	0.00	1115.52
	40	$0.12548 \cdot 10^9$	0.42	1.42	16211.00	894.52	6.21	3.49	0.93	0.67	0.00	0.00	0.00	1621.10
	50	$0.98400 \cdot 10^8$	0.07	1.18	21506.40	1335.97	6.79	3.51	0.33	0.79	0.74	0.00	0.00	2150.64
	Average		0.04	0.62	4836.22	258.62	2.35	1.66	0.40	0.16	0.09	0.00	0.00	483.62

An anonymous referee suggested that G_1 and G_2 could be improved by small perturbation of points of X. While it follows from Theorem 1 that this would not guarantee optimality, it might give better solution in some cases. For further empirical comparison of heuristics for minimum sum-of-squares clustering see e.g. Taristano (2003) and Hansen and Mladenović (2001a).

Table 3. Summary of cases for Example ??

Case	Interval	Cardinality			Centroids			Errors			Total
	for y				m_1	m_2	m_3	C_1	C_2	C_3	Error
1	$(-\infty,-150]$	0	4	5	<-150	150	654	0	50000	100000	150000
2	(150,150]	1	3	5	0	200	654	0	20000	100000	120000
3	(50,250]	2	2	5	50	250	654	5000	5000	100000	110000
4	(250,300]	3	1	5	100	300	654	20000	0	100000	120000
5	(300,450]	3	2	4	100	377	704	20000	11858	50000	81858
6	(450,454]	4	1	4	150	454	704	50000	0	50000	100000
7	(454,654]	4	2	3	150	504	754	50000	50000	20000	75000
8	(654,854]	4	3	2	150	554	804	50000	20000	5000	75000
9	(854,1054]	4	4	1	150	604	854	50000	50000	0	100000
10	$(1054,\infty)$	4	5	0	150	654	>1054	50000	100000	0	150000
Optimal		3	3	3	100	434	754	20000	32744	20000	72744

6. Conclusions

From the theoretical point of view, a complete analysis has been made of the conditions in which the assumption of Likas *et al.* (2003) holds, i.e., that beginning with the $M - 1$ centroids of an optimal $(M - 1)$ partition there is always an additional point $y \in X$ or $y \in \mathbf{R}^d$ to be chosen as M^{th} cluster center such that k-means then yields an optimal M-partition. Results are rather negative: if $y \in X$ there are cases for all $M \geq 2$ and $d \geq 1$ where k-means stops with a suboptimal M-partition. If $y \in \mathbf{R}^d$ and $M = 2$ an adequate cluster center exists, but the proof given is not constructive. If $y \in \mathbf{R}^d$, $M \geq 3$ and $d \geq 1$ again there are cases where k-means stops with a suboptimal partition. This proves the impossibility of deriving an optimal incremental k-means algorithm.

From the empirical point of view, it appears that for moderate size problems, i.e., with $N \leq 150$, for which the optimal solution is known, global k-means attains this solution for small M, i.e., $M \leq 7$. For larger values of M global k-means makes small errors.

Considering a bit larger problems with $N = 214$, 2310 and 3038 leads to similar conclusions. Moreover, the recent j-means heuristic extended by embedding it in a Variable Neighborhood Search framework gives better results than global k-means in equal computing time and for large M and N in 10 times less computing time.

These properties and experiments illustrate the difficulty of obtaining an exact solution to large minimum sum-of-squares clustering problem: when M grows, an incremental approach does not suffice for reorganizing the current partition sufficiently. Jump moves then appear to be necessary in order to avoid that the effect of the new center be only local.

Appendix

Proof of Theorem 1.

The proof is constructive. It consists in providing an example which satisfies the conditions of Theorem 1. As this theorem is an existential statement, the result follows.

Example 4 *Let $N = 9$, $d = 1$, $x_1 = 0$, $x_2 = 100$, $x_3 = 200$, $x_4 = 300$, $x_5 = 454$, $x_6 = 554$, $x_7 = 654$, $x_8 = 754$, $x_9 = 854$ and $M = 3$. It is easily seen that the optimal clustering into $k = 2$ clusters is $C_1 = \{x_1, x_2, x_3, x_4\}$, $C_2 = \{x_5, x_6, x_7, x_8, x_9\}$ with centroids $m_1 = 150$, $m_2 = 654$ and an error of 150000 (see Figure ?? (a)).*

Let us now examine which partitions are obtained for all possible choices of the additional initial point y. There are 10 cases, summarized in Table ?? and illustrated on Figure ??(b) to ??(k). We only discuss the three first ones, the others being similar. We assume that points do not change cluster in case of ties in their distances to centroids (or initial points).

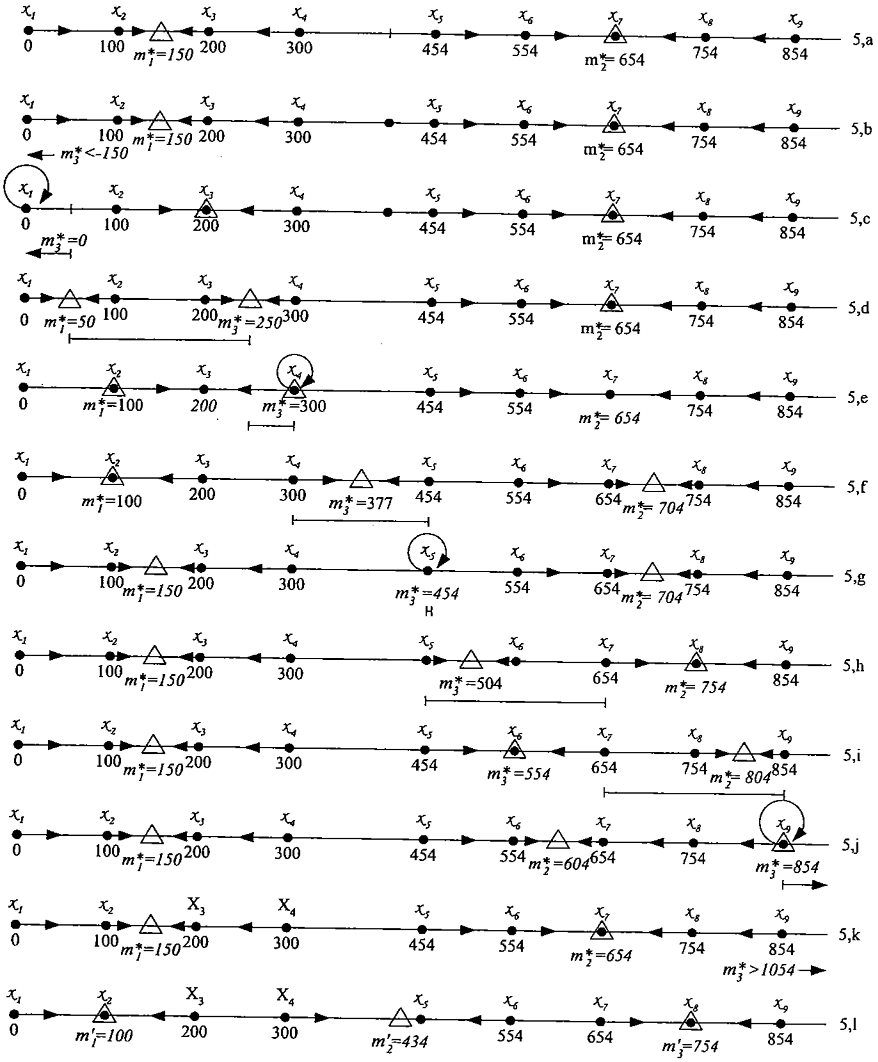
If $y \in (-\infty, -150]$, the new initial point is so far to the left that it captures none of the points of X. An empty cluster is thus added to the optimal cluster for $k = 2$. The centroids are at < -150; 150 and 654, the errors of the clusters are 0, 50000 and 100000 and the total error 150000 (see Figure ??(b)).

If $y \in (-150, 50]$ the new initial point captures x_1 only. Clusters with 1, 3 and 5 points, from left to right, are obtained. Their centroids are at 0, 200 and 654. Due to the tie-breaking rule, k-means stops. The errors of the clusters are 0, 20000 and 100000 and the total error 120000 (see Figure ??(c)).

If $y \in (50, 250]$ then either $y \in (50, 150)$, it captures the first 2 points, the leftmost cluster keeps the next 2 and the rightmost cluster remains as before; $y = 150$ it coincides with m_1 and is useless, or $y \in (150, 250]$ and it captures points x_3 and x_4, the leftmost cluster retaining points x_1 and x_2. In both cases (i.e., excluding y=150) the obtained clusters have, from left to right 2, 2 and 5 points. Their centroids are at 50, 250 and 654 respectively, their errors are 5000, 5000 and 100000. The total error is 110000 (see Figure ??(d)).

The seven other cases yield partitions with a total error greater than or equal to 75000. Yet, consider the partition $C_1 = \{x_1, x_2, x_3\}$, $C_2 = \{x_4, x_5, x_6\}$, $C_3 = \{x_7, x_8, x_9\}$ which can easily be shown to be optimal. Then the centroids of the clusters are at 100, 434 and 754, the errors of the clusters are 20000, 32724 and 20000 respectively. The total error is 72744 (see Figure ??(l)).

So, regardless of the initial point y chosen in addition to the two centroids m_1 and m_2, k-means stops with an error of at least $(75000 - 72744)/72744 = 3.10\%$ of the optimum value.



x1 x2 x3 x4 x5 x6 x7 x8 x9
0 100 m*1=150 200 300 454 554 m*2=654 754 854 5,a
0 100 m*1=150 200 300 454 554 m*2=654 754 854 5,b
m*3<-150
0 100 200 300 454 554 m*2=654 754 854 5,c
m*3=0
0 m*1=50 100 200 m*3=250 300 454 554 m*2=654 754 854 5,d
0 m*1=100 200 m*3=300 454 554 m*2=654 754 854 5,e
0 m*1=100 200 300 m*3=377 454 554 654 m*2=704 754 854 5,f
0 100 m*1=150 200 300 m*3=454 554 654 m*2=704 754 854 5,g
H
0 100 m*1=150 200 300 m*3=504 654 m*2=754 854 5,h
0 100 m*1=150 200 300 454 m*3=554 654 754 m*2=804 854 5,i
0 100 m*1=150 200 300 454 554 m*2=604 654 754 m*3=854 5,j
0 100 m*1=150 200 300 454 554 m*2=654 754 854 m*3>1054 5,k
0 m'1=100 200 300 m'2=434 554 654 m'3=754 854 5,l

Figure 5. Example 4.

Note that Example ?? can be modified, as was done for Example ??, by adding points sufficiently far apart, then duplicating them and giving them small perturbations, to build examples where global k-means does not give the optimal partition with $M > 3$ and/or $d > 1$.

References

ANDERSON, E. (1935), "The Irises of Gaspe Peninsula", *Bull. Amer. Iris Soc., 59*, 2–5.

BELACEL, N., HANSEN, P. and MLADENOVIĆ, N. (2003) "Fuzzy J-Means: A New Heuristic for Fuzzy Clustering", *Pattern Recognition, 35*, 2193–2200.

BELLMAN, R. and DREYFUS, S. (1962), *Applied Dynamic Programming*, Princeton University Press.

BENTLEY, J.L. (1991), "Multidimensional Binary Search Trees Used for Associative Searching", *Commun. ACM, 18, 9*, 509–517.

BERMEJO, S. and CABESTANY, J. (2002), "The Effect of Finite Sample Size on On-line K-means", *Neurocomputing, 48*, 511–539.

BEZDEK, J.C. (1980), "Convergence Theorem for the Fuzzy Isodata Clustering Algorithms", *IEE Transactions on Pattern Analysis and Machine Intelligence 2*, 1–8.

BLAKE, C.L. and MERZ, C.J. (1998), "CI Repository of Machine Learning Databases [http://www.ics.uci.edu/ mlearn/MLRepository.html]", Irvine, CA: University of California, Department of Information and Computer Science.

BOUGUETTAYA, A. (1996), "On-line Clustering", *IEEE Transactions on Knowledge and Data Engineering, 8*, 333–339.

BRÜCKER, P. (1978), "On the Complexity of Clustering Problem", *Lecture Notes in Economic and Mathematical Systems, 157*, 45–54.

DE SMET, Y. and MONTANO-GUZMÁN L. (2004), "Towards Multicriteria Clustering: An Extension of the k-means Algorithm", *European Journal of Oper. Res. 158*, 390–398.

DIDAY, E. (1972), "Introduction to Dynamic Clusters Method (DYC Program)", *METRA, 11*, 505–519.

DIDAY, E. (1974), "Optimization in Non-hierarhical Clustering", *Pattern Recognition, 6*.

DIEHR, G. (1985), "Evaluation of a Branch and Bound Algorithm for Clustering", *SIAM J. Sci. and Statist. Computing, 6*, 268–284.

DU MERLE, O., HANSEN, P., JAUMARD, B. and MLADENOVIĆ, N. (2000), "An Interior Point Algorithm for Minimum Sum-of-Squares Clustering", *SIAM J. Sci. Comput., 21*, 1485–1505.

EL-SONBATY, Y. and ISMAIL, M.A. (1998), "On-line Hierarchical Clustering", *Pattern Recognition Letters, 19*, 1285–1291.

ESTIVILL-CASTRO, V. and YANG, J. (2004), "Fast and Robust General Purpose Clustering Algorithms", *Data Mining and Knowledge Discovery, 8*, 127–150.

FISHER, R.A. (1936), "The Use of Multiple Measurements in Taxonomic Problems", *Ann. Eugenics, VII part II*, 179–188. Reprinted in R. A. Fisher, Contributions to Mathematical Statistics, Wiley, 1950.

GAREY, M. and JOHNSON, D. (1979), *Computers and Intractability*, W.H. Freeman and company, New York.

GODIN, N., HUGUET, S., GAERTNER, R. and SALMON, L. (2004), "Clustering of Acoustic Emission Signals Collected during Tensile Tests on Unidirectional Glass/polyester Composite Using Supervised and Unsupervised Classifiers", *NDT & E International 37*, 253–264.

GORDON, A.D. (1981), *Classification*, Chapman and Hall: London.

GOWER, J.C. (1967), "A Comparisom of Some Methods of Cluster Analysis", *Biometrics, 23*, 623–637.

HANSEN, P. and DELATTRE, M. (1978), "Complete-link Cluster Analysis by Graph Coloring", *Journal of the American Statistical Association, 73*, 397–403.

HANSEN, P. and JAUMARD, B. (1997), "Cluster Analysis and Mathematical Programming", *Mathematical Programming, 79*, 191–215.

HANSEN, P., JAUMARD, B. and MLADENOVIĆ, N. (1998), "Minimum Sum-of-squares Clustering in a Low Dimensional Space", *Journal of Classification, 15*, 37–56.

HANSEN, P. and MLADENOVIĆ, N. (2001a), "J-Means: A New Local Search Heuristic for Minimum Sum-of-squares Clustering", *Pattern Recognition, 34, 2*, 405–413.

HANSEN, P. and MLADENOVIĆ, N. (2001b), "Variable Neighborhood Search: Principles and Applications", *European Journal of Oprnl. Res. 130*, 449–467.

HAR-PELED, S. and MAZUMDAR, S. (2004), "Coresets for k-means and k-median Clustering and Their Applications", *STOC.*

HAR-PELED, S. and SADRI, B., "On Lloyd's k-means Method", SODA 2005, *Algorithmica* (to appear).

HUANG, Z. (1998), "Extensions to the k-means Algorithm for Clustering Large Data Sets with Categorical Values", *Data Mining and Knowledge Discovery, 2*, 283–304.

JAIN, A.K., MURTY, M.N. and FLINN, P.J. (1999), "Data Clustering: A Review", *ACM Computing Surveys, 31, 3*, 264–323.

KANUNGO, T., MOUNT, D.M., NETANYAHU, N.S., PIATKO, C.D., SILVERMAN, R. and WU, A.Y. (2002), "An Efficient k-means Clustering Algorithm: Analysis and Implementation", *IEEE Transactions on Pattern Analysis and Machine Intelligence, 24*, 881–892.

KAUFMAN, L. and ROUSSEEUW, P.J. (1990), *Finding Groups in Data: An Introduction to Cluster Analysis*, New York, Wiley.

KOONTZ, W.L.G., NARENDRA, P.M. and FUKUNAGA, K. (1975), "A Branch and Bound Clustering Algorithm", *IEEE Trans. Computers C, 24*, 908–915.

LIKAS, A., VLASSIS, N. and VERBEEK, J. (2003), "The Global k-means Clustering Algorithm", *Pattern Recognition, 36, 2*, 451–461.

MACQUEEN, J.B. (1967), "Some Methods for Classification and Analysis of Multivariate Observations", *Proceedings of the* 5^{th} *Berkeley Symposium on Mathematical Statistics and Probability, 1*, 281–297.

MAKARENKOV, V. and LEGENDRE, P. (2001), "Optimal Variable Weighting for Ultrametric and Additive Trees and K-means Partitioning: Methods and Software", *Journal of Classification, 18*, 245–271.

MARQUES, W.C.P, CARVALHO, E.A., COSTA, R.C.S. and MEDEIROS, F.N.S. (2004), "Filtering Effects on SAR Images Segmentation", *Lecture Notes in Computer Science, 3124*, 1041–1046.

MILLIGAN, G.W. and COOPER, M.C. (1985), "An Examination of Procedures for Determinating the Number of Clusters in Data Set", *Psychometrika, 50*, 159–179.

MLADENOVIĆ, N. and HANSEN, P. (1997), "Variable Neighborhood Search", *Comps. and Opns. Res. 24*, 1097–1100.

PEÑA, J.M., LOZANO, J.A. and LARRAÑAGA, P. (1999), "An Empirical Comparison of Four Initialization Methods for the k-means Algorithm", *Pattern Recognition Letters, 20*, 1027–1040.

PHAM, D.T., DIMOV, S.S. and NGUYEN, C.D. (2004), "An Incremental k-means Algorithm", *Journal of Mechanical Engineering Science, 218*, 783–795.

RAO, M.R. (1971), "Cluster Analysis and Mathematical Programming", *Journal of the American Statistical Association, 66*, 622–645.

REINELT, G. (1991), "TSP-LIB-A Traveling Salesman Library", *ORSA J. Comput., 3*, 376–384.

RUSPINI, E. (1970), "Numerical Methods for Fuzzy Clustering", *Information Sci., 2*, 319-350.

SAN, O.M., HUYNH, V-N. and NAKAMORI, Y. (2004), "An Alternative Extension of the k-means Algorithm for Clustering Categorical Data", *Int. J. Appl. Math. Comput. Sci., 14*, 241–247.

SCHENKER, A., LAST, M., BUNKE, H. and KANDEL, A. (2004), "Comparison of Algorithms for Web Document Clustering Using Graph Representation of Data", *Lecture Notes in Computer Sciences, 3138*, 190–197.

SPÄTH, H. (1980), *Cluster Analysis Algorithms for Data Reduction and Classification of Objects*, Ellis Horwood: Chichester.

SPÄTH, H. (1985), *Cluster Dissection and Analysis (Theory, Fortran Programs, Examples)*, Ellis Horwood: Chichester.

SPROULL, R.F. (1991), "Refinements to Nearest-neighbor Searching in k-dimensional Trees", *Algoritmica, 6*, 579–589.

TARSITANO, A. (2003), "A Computational Study of Several Relocation Methods for k-means Algorithms", *Pattern Recognition, 36*, 2955–2966.

TSINGOS, N., GALLO, E. and DRETTAKIS, G. (2004), "Perceptual Audio Rendering of Complex Virtual Environments", *ACM Transactions on Graphics, 23*, 249–258.

WARD, J.H. (1963), "Hierarchical Grouping to Optimize an Objective Function", *Journal of the American Statistical Association, 58*, 236–244.

WELCH, W.J. (1982), "Algorithmic Complexity Three NP-hard Problems in Computational Statistics", *J Stat. Comp. and Sim., 15*, 17–25.

WHITFIELD, P.H., HALL, A.W. and CANNON, A.J. (2004), "Changes in the Seasonal Cycle in the Circumpolar Arctic, 1976-95: Temperature and Orecipitation", *Arctic, 57*, 80-93.