Solving the Minimum Sum-of-Squares Clustering Problem by Hyperbolic Smoothing and Partition into Boundary and Gravitational Regions

Adilson Elias Xavier
Vinicius Layter Xavier
Dept. of Systems Engineering and Computer Science
Graduate School of Engineering (COPPE)
Federal University of Rio de Janeiro
P.O. Box 68511
Rio de Janeiro,RJ 21941-972, BRAZIL
e-mail: {adilson,vinicius}@cos.ufrj.br

Abstract

This article considers the minimum sum-of-squares clustering (MSSC)problem. The mathematical modeling of this problem leads to a $min - sum - min$ formulation which, in addition to its intrinsic bi-level nature, has the significant characteristic of being strongly nondifferentiable. To overcome these difficulties, the proposed resolution method, called Hyperbolic Smoothing, adopts a smoothing strategy using a special C^∞ differentiable class function. The final solution is obtained by solving a sequence of low dimension differentiable unconstrained optimization subproblems which gradually approach the original problem. This paper introduces the method of partition of the set of observations into two non overlapping groups: "data in frontier" and "data in gravitational regions". The resulting combination of the two methotologies for the MSSC problem has interesting properties, which drastically simplify the computational tasks.

Keywords: Cluster Analysis, Pattern Recognition, Min-Sum-Min Problems, Nondifferentiable Programming, Smoothing

1 Introduction

Cluster analysis deals with the problems of classification of a set of patterns or observations, in general represented as points in a multidimensional space, into clusters, following two basic and simultaneous objectives: patterns in the same clusters must be similar to each other (homogeneity objective) and different from patterns in other clusters (separation objective) [Anderberg (1973), Hartingan (1975) and Späth (1980)].

Clustering is an important problem that appears in a broad spectrum of applications, whose intrinsic characteristics engender many approaches to this problem, as described by Dubes and Jain (1976), Jain and Dubes (1988) and Hansen and Jaumard (1997).

Clustering analysis has been used traditionally in disciplines such as: biology, biometry, psychology, psychiatry, medicine, geology, marketing and finance. Clustering is also a fundamental tool in modern technology applications, such as: pattern recognition, data mining, web mining, image processing, machine learning and knowledge discovering.

In this paper, a particular clustering problem formulation is considered. Among many criteria used in cluster analysis, the most natural, intuitive and frequently adopted criterion is the minimum sum-of-squares clustering (MSSC). This criterion corresponds to the minimization of the sum-of-squares of distances of observations to their cluster means, or equivalently, to the minimization of within-group sum-of-squares. It is a criterion for both the homogeneity and the separation objectives. According to the Huygens Theorem, minimizing the within-cluster inertia of a partition (homogeneity within the cluster) is equivalent to maximizing the between-cluster inertia (separation between clusters).

The minimum sum-of-squares clustering (MSSC) formulation produces a mathematical problem of global optimization. It is both a nondifferentiable and a nonconvex mathematical problem, with a large number of local minimizers.

There are two main strategies for solving clustering problems: hierarchical clustering methods and partition clustering methods. Hierarchical methods produce a hierarchy of partitions of a set of observations. Partition methods, in general, assume a given number of clusters and, essentially, seek the

optimization of an objective function measuring the homogeneity within the clusters and/or the separation between the clusters.

For the sake of completeness, we present first the Hyperbolic Smothing Clustering Method (HSCM), Xavier (2010). Basically the method performs the smoothing of the nondifferentiable $min - sum - min$ problem engendered by the modeling of the clustering problem. This technique was developed through an adaptation of the hyperbolic penalty method originally introduced by Xavier (1982). By smoothing, we fundamentally mean the substitution of an intrinsically nondifferentiable two-level problem by a C^{∞} unconstrained differentiable single-level alternative.

Additionally, the paper presents a new, faster, methodology. The basic idea is the partition of the set of observations into two non overlapping parts. By using a conceptual presentation, the first set corresponds to the observation points relatively close to two or more centroids. This set of observations, named boundary band points, can be managed by using the previously presented smoothing approach. The second set corresponds to observation points significantly closer to a single centroid in comparison with others. This set of observations, named gravitational points, is managed in a direct and simple way, offering much faster performance.

This work is organized in the following way. A step-by-step definition of the minimum sum-of-squares clustering problem is presented in the next section. The original smoothing hyperbolic smoothing approach and the derived algorithm are presented in section 3. The boundary and gravitational regions partition scheme and the new derived algorithm are presented in section 4. Computational results are presented in section 5. Brief conclusions are drawn in section 6.

2 The Minimum Sum-of-Squares Clustering Problem

Let $S = \{s_1, \ldots, s_m\}$ denote a set of m patterns or observations from an Euclidean n-space, to be clustered into a given number q of disjoint clusters. To formulate the original clustering problem as a $min-sum-min$ problem, we proceed as follows. Let $x_i, i = 1, \ldots, q$ be the centroids of the

clusters, where each $x_i \in \mathbb{R}^n$. The set of these centroid coordinates will be represented by $X \in \mathbb{R}^{nq}$. Given a point s_j of S, we initially calculate the Euclidian distance from s_j to the center in X that is nearest. This is given by

$$z_j = \min_{x_i \in X} \|s_j - x_i\|_2. \tag{1}$$

The most frequent measurement of the quality of a clustering associated to a specific position of q centroids is provided by the sum of the squares of these distances, which determines the MSSC problem:

$$\begin{aligned} & minimize \sum_{j=1}^{m} z_j^2 \\ \text{subject to} \quad & z_j = \min_{i=1,\ldots,q} \|s_j - x_i\|_2, \quad j = 1, \ldots, m \end{aligned} \tag{2}$$

3 The Hyperbolic Smoothing Clustering Method

Considering its definition, each z_j must necessarily satisfy the following set of inequalities:

$$z_j - \|s_j - x_i\|_2 \leq 0, \quad i = 1, \ldots, q. \tag{3}$$

Substituting these inequalities for the equality constraints of problem (2), it is produced the relaxed problem:

$$\begin{aligned} & minimize \sum_{j=1}^{m} z_j^2 \\ \text{subject to} \quad & z_j - \|s_j - x_i\|_2 \leq 0, \quad j = 1, \ldots, m, \;\; i = 1, \ldots, q. \end{aligned} \tag{4}$$

Since the variables z_j are not bounded from below, the optimum solution of the relaxed problem will be $z_j = 0,\ j = 1, \ldots, m$. In order to obtain the desired equivalence, we must, therefore, modify problem (4). We do so by first letting $\varphi(y)$ denote $max\{0, y\}$ and then observing that, from the set of inequalities in (4), it follows that

$$\sum_{i=1}^{q} \varphi(z_j - \|s_j - x_i\|_2\ ) = 0, \quad j = 1, \ldots, m. \tag{5}$$

Using (5) in place of the set of inequality constraints in (4), we would obtain an equivalent problem maintaining the undesirable property that $z_j,\ j = 1, \ldots, m$ still has no lower bound. Considering, however, that the objective function of problem (4) will force each $z_j,\ j = 1, \ldots, m,$ downward, we can think of bounding the latter variables from below by including an ε perturbation in (5). So, it is obtained the following modified problem:

$$\begin{aligned} & minimize \ \sum_{j=1}^{m} z_j^2 \\ \text{subject to} \quad & \sum_{i=1}^{q} \varphi(z_j - \|s_j - x_i\|_2\ ) \ \geq \ \varepsilon\ , \quad j = 1, \ldots, m \end{aligned} \tag{6}$$

for $\varepsilon > 0$. Since the feasible set of problem (2) is the limit of that of (6) when $\varepsilon \to 0_+$, we can then consider solving (2) by solving a sequence of problems like (6) for a sequence of decreasing values for ε that approaches 0.

Analyzing the problem (6), the definition of function φ endows it with an extremely rigid nondifferentiable structure, which makes its computational solution very hard. In view of this, the numerical method we adopt for solving problem (1), takes a smoothing approach. From this perspective, let us define the function:

$$\phi(y, \tau) = \left(y + \sqrt{y^2 + \tau^2}\ \right)/2 \tag{7}$$

for $y \in \mathbb{R}$ and $\tau > 0$.

Function ϕ has the following properties:

(*a*) $\phi(y,\tau) > \varphi(y), \quad \forall\, \tau > 0$;

(*b*) $\lim\limits_{\tau \to 0} \phi(y,\tau) = \varphi(y)$;

(*c*) $\phi(y,\tau)$ is an increasing convex C^∞ function in variable y.

By using function ϕ in the place of function φ, the problem

$$\begin{aligned} & minimize \ \sum_{j=1}^{m} z_j^2 \\ \text{subject to} \quad & \sum_{i=1}^{q} \phi(z_j - \|s_j - x_i\|_2, \tau) \geq \varepsilon, \quad j = 1, \ldots, m. \end{aligned} \tag{8}$$

is produced.

To obtain a differentiable problem, it is still necessary to smooth the Euclidean distance $\|s_j - x_i\|_2$. For this purpose, let us define the function

$$\theta(\, s_j \,,\, x_i \,,\, \gamma \,) \;=\; \sqrt{\sum_{l=1}^{n} (s_j^l - x_i^l)^2 \,+\, \gamma^2} \tag{9}$$

for $\gamma > 0$.

Function θ has the following properties:

(*a*) $\lim\limits_{\gamma \to 0} \theta(\, s_j \,,\, x_i \,,\, \gamma \,) \;=\; \|s_j - x_i\|_2$;

(*b*) θ is a C^∞ function.

By using function θ in place of the distance $\|s_j - x_i\|_2$, the following completely differentiable problem is now obtained:

$$minimize \ \sum_{j=1}^{m} z_j^2 \tag{10}$$

$$\text{subject to} \quad \sum_{i=1}^{q} \phi(z_j - \theta(s_j, x_i, \gamma), \tau) \geq \varepsilon, \quad j = 1, \ldots, m.$$

So, the properties of functions ϕ and θ allow us to seek a solution to problem (6) by solving a sequence of subproblems like problem (10), produced by the decreasing of the parameters $\gamma \to 0$, $\tau \to 0$, and $\varepsilon \to 0$.

Since $z_j \geq 0$, $j = 1, \ldots, m$, the objective function minimization process will work for reducing these values to the utmost. On the other hand, given any set of centroids x_i, $i = 1, \ldots, q$, due to property (c) of the hyperbolic smoothing function ϕ, the constraints of problem (10) are a monotonically increasing function in z_j. So, these constraints will certainly be active and problem (10) will at last be equivalent to problem:

$$minimize \ \sum_{j=1}^{m} z_j^2 \tag{11}$$

$$\text{subject to} \quad h_j(z_j, x) = \sum_{i=1}^{q} \phi(z_j - \theta(s_j, x_i, \gamma), \tau) - \varepsilon = 0, \quad j = 1, \ldots, m.$$

The dimension of the variable domain space of problem (11) is $(nq+m)$. As, in general, the value of the parameter m, the cardinality of the set S of the observations s_j, is large, problem (11) has a large number of variables. However, it has a separable structure, because each variable z_j appears only in one equality constraint. Therefore, as the partial derivative of $h(z_j, x)$ with respect to z_j, $j = 1, \ldots, m$ is not equal to zero, it is possible to use the Implicit Function Theorem to calculate each component z_j, $j = 1, \ldots, m$ as a function of the centroid variables x_i, $i = 1, \ldots, q$. In this way, the unconstrained problem

$$minimize \ f(x) = \sum_{j=1}^{m} z_j(x)^2 \tag{12}$$

is obtained, where each $z_j(x)$ results from the calculation of a zero of each equation

$$h_j(z_j, x) \;=\; \sum_{i=1}^{q} \phi(z_j - \theta(s_j, x_i, \gamma), \tau) \;-\; \varepsilon \;=\; 0, \qquad j = 1, \ldots, m. \tag{13}$$

Due to property (c) of the hyperbolic smoothing function, each term ϕ above is strictly increasing with variable z_j and therefore the equation has a single zero.

Again, due to the Implicit Function Theorem, the functions $z_j(x)$ have all derivatives with respect to the variables x_i, $i = 1, \ldots, q$, and therefore it is possible to calculate the gradient of the objective function of problem (12),

$$\nabla f(x) \;=\; \sum_{j=1}^{m} 2\, z_j(x)\, \nabla z_j(x) \tag{14}$$

where

$$\nabla z_j(x) \;=\; -\, \nabla h_j(z_j, x) \;/\; \frac{\partial\, h_j(z_j, x)}{\partial\, z_j}, \tag{15}$$

while $\nabla h_j(z_j, x)$ and $\partial\, h_j(z_j, x)/\partial\, z_j$ are obtained from equations (7), (9) and (13).

In this way, it is easy to solve problem (12) by making use of any method based on first order derivative information. At last, it must be emphasized that problem (12) is defined on a $(nq)-$dimensional space, so it is a small problem, since the number of clusters, q, is, in general, very small for real applications.

The solution of the original clustering problem can be obtained by using the Hyperbolic Smoothing Clustering Algorithm, described below in a simplified form.

The Simplified HSC Algorithm

Initialization Step: Choose initial values: $x^0,\ \gamma^1\ ,\ \tau^1\ ,\ \varepsilon^1$.

Choose values $0 < \rho_1 < 1,\ 0 < \rho_2 < 1,\ 0 < \rho_3 < 1;$ let $k = 1$.

Main Step: Repeat until a stopping rule is attained

Solve problem (12) with $\gamma = \gamma^k,\ \tau = \tau^k$ and $\varepsilon = \varepsilon^k,$ starting at the initial point x^{k-1} and let x^k be the solution obtained.

Let $\gamma^{k+1} = \rho_1\,\gamma^k\ ,\ \tau^{k+1} = \rho_2\,\tau^k\ ,\ \varepsilon^{k+1} = \rho_3\,\varepsilon^k\ ,\ k := k+1.$ ∎

Just as in other smoothing methods, the solution to the clustering problem is obtained, in theory, by solving an infinite sequence of optimization problems. In the HSC algorithm, each problem to be minimized is unconstrained and of low dimension.

Notice that the algorithm causes τ and γ to approach $0,$ so the constraints of the subproblems as given in (10) tend to those of (6). In addition, the algorithm causes ε to approach $0,$ so, in a simultaneous movement, the solved problem (6) gradually approaches the original MSSC problem (2).

4 The Accelerated Hyperbolic Smoothing Clustering Method

The calculation of the objective function of the problem (12) demands the determination of the zeros of m equations (13), one equation for each observation point. This is a relevant computational task associated to HSC Algorithm.

In this section, it is presented a faster procedure. The basic idea is the partition of the set of observations into two non overlapping regions. By using a conceptual presentation, the first region corresponds to the observation points that are relatively close to two or more centroids. The second region corresponds to the observation points that are significantly close to a unique centroid in comparison with the other ones.

So, the first part J_B is the set of boundary observations and the second is the set J_G of gravitational observations. Considering this partition, equation (12) can be expressed in the following way:

$$minimize\ f(x) = \sum_{j=1}^{m} z_j(x)^2 = \sum_{j \in J_B} z_j(x)^2 + \sum_{j \in J_G} z_j(x)^2, \tag{16}$$

so, the objective function can be presented in the form:

$$minimize\ f(x) = f_B(x) + f_G(x), \tag{17}$$

where the two components are completely independent.

The first part of expression (17), associated with the boundary observations, can be calculated by using the previous presented smoothing approach, see (12) and (13):

$$minimize\ f_B(x) = \sum_{j \in J_B} z_j(x)^2, \tag{18}$$

where each $z_j(x)$ results from the calculation of a zero of each equation

$$h_j(z_j, x) = \sum_{i=1}^{q} \phi(z_j - \theta(s_j, x_i, \gamma), \tau) - \varepsilon = 0, \quad j \in J_B. \tag{19}$$

The second part of expression (17) can be calculated by using a faster procedure, as we will show right away.

Let us define the two parts in a more rigorous form. Let be $\overline{x}_i$, $i = 1, \ldots, q$ be a referential position of centroids of the clusters taken in the iterative process.

The boundary concept in relation to the referential point $\overline{x}$ can be easily specified by defining a δ band zone between neighboring centroids. For a generic point $s \in \mathbb{R}^n$, we define the first and second nearest distances from s to the centroids:

$$d_1(s, \overline{x}) = \| s - \overline{x}_{i_1} \| = \min_i \| s - \overline{x}_i \| \tag{20}$$

$$d_2(s,\overline{x}) \;=\; \| \, s - \overline{x}_{i_2} \, \| \;=\; \min_{i \neq i_1} \| \, s - \overline{x}_i \, \| \;, \tag{21}$$

where i_1 and i_2 are the labeling indexes of these two nearest centroids.

By using the above definitions, let us define precisely the δ boundary band zone:

$$Z_\delta(\overline{x}) \;=\; \{ s \in \mathbb{R}^n \mid d_2(s,\overline{x}) \;-\; d_1(s,\overline{x}) \;<\; 2\,\delta \,\} \tag{22}$$

and the gravity region, this is the complementary space:

$$G_\delta(\overline{x}) \;=\; \{ s \in \mathbb{R}^n \;-\; Z_\delta(\overline{x}) \,\} \,. \tag{23}$$

Figure 1 illustrates in $\mathbb{R}^2$ the $Z_\delta(\overline{x})$ and $G_\delta(\overline{x})$ partitions. The central lines form the Voronoy polygon associated with the referential centroids $\overline{x}_i$, $i = 1, \ldots, q$. The region between two parallel lines to Voronoy lines constitutes the boundary band zone $Z_\delta(\overline{x})$.

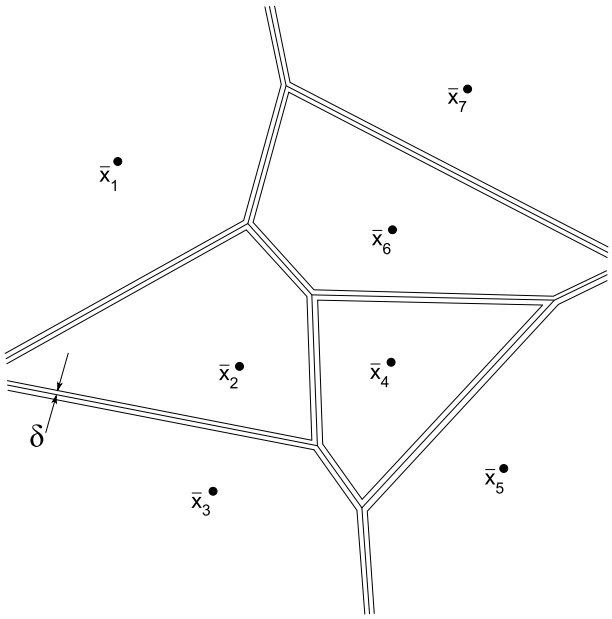

Figure 1: The $Z_\delta(\overline{x})$ and $G_\delta(\overline{x})$ partitions.

Now, the sets J_B and J_G can be defined in a precise form:

$$J_B(\overline{x}) = \{j = 1, \ldots, m \,|\, s_j \in Z_\delta(\overline{x})\}\,, \tag{24}$$

$$J_G(\overline{x}) = \{j = 1, \ldots, m \,|\, s_j \in G_\delta(\overline{x})\}\,. \tag{25}$$

Proposition 1:

Let s be a generic point belonging to the gravity region $G_\delta(\overline{x})$, with nearest centroid i_1. Let x be the current position of the centroids. Let $\Delta x = \max_i \|x_i - \overline{x}_i\|$ be the maximum displacement of the centroids.

If $\Delta x < \delta$ then s will continue to be nearest to centroid x_{i_1} than to any other one,so

$$\min_{i \neq i_1} \| s - x_i \| - \| s - x_{i_1} \| \geq 0. \tag{26}$$

Proof.

$$\min_{i \neq i_1} \| s - x_i \| - \| s - x_{i_1} \| = \min_{i \neq i_1} \| s - \overline{x}_i + \overline{x}_i - x_i \| - \| s - \overline{x}_{i_1} + \overline{x}_{i_1} - x_{i_1} \| \geq \tag{27}$$

$$\min_{i \neq i_1} \| s - \overline{x}_i \| - \| \overline{x}_i - x_i \| - \| s - \overline{x}_{i_1} \| - \| \overline{x}_{i_1} - x_{i_1} \| \geq \tag{28}$$

$$2\,\delta - 2\,\Delta x \geq 0 \qquad \blacksquare \tag{29}$$

Since $\delta \geq \Delta x$, Proposition 1 makes it possible to calculate exactly expression (16) in a very fast way. First, let us define the subsets of gravity observations associated with each referential centroid:

$$J_i(\overline{x}) = \left\{ j \in J_G \,|\, \min_{l=1,\ldots,q} \| s_j - \overline{x}_l \| = \| s_j - \overline{x}_i \| \right\} \tag{30}$$

The center of the observations in each non-empty subset is given by

$$v_i = \frac{1}{|J_i|} \sum_{s_j \in J_i} s_j, \quad \forall\, i = 1, \ldots, q. \tag{31}$$

Let us consider the second sum in expression (16). It will be computed by taking into account the centers defined above.

$$minimize\ f_G(x) = \sum_{j \in J_G} z_j(x)^2 = \sum_{i=1}^{q} \sum_{j \in J_i} \| s_j - x_i \|^2 = \tag{32}$$

$$\sum_{i=1}^{q} \sum_{j \in J_i} \| s_j - v_i + v_i - x_i \|^2 = \tag{33}$$

$$\sum_{i=1}^{q} \left[\sum_{j \in J_i} \| s_j - v_i \|^2 + 2 (v_i - x_i) \sum_{j \in J_i} (s_j - v_i) + \sum_{j \in J_i} \| x_i - v_i \|^2 \right] \tag{34}$$

Equation (31) implies that

$$\sum_{j \in J_i} (s_j - v_i) = 0, \tag{35}$$

and then:

$$minimize\ f_G(x) = \sum_{i=1}^{q} \sum_{j \in J_i} \| s_j - v_i \|^2 + \sum_{i=1}^{q} |J_i| \, \| x_i - v_i \|^2 \tag{36}$$

When the position of centroids $x_i , i = 1, \ldots, q$ moves during the iterative process, the value of the first sum in (36) assumes a constant value, since the vectors s and v are fixed. On the other hand, for the calculation of the second sum, it is only necessary to calculate q distances, $\| v_i - x_i \|$, $i = 1, \ldots, q$.

The gradient of the second part of the objective function is easily calculated by:

$$\nabla f_G(x) = \sum_{i=1}^{q} 2 |J_i| (x_i - v_i), \tag{37}$$

where the vector $(v_i - x_i)$ must be in $\mathbb{R}^{nq}$, so it has the first $(i-1)q$ components and the last $l = iq + 1, \ldots, nq$ components equal zero.

Therefore, if it is observed that $\delta \geq \Delta x$ within the iterative process, the calculation of the expression $\sum_{j \in J_G} z_j(x)^2$ and its gradient can be computed exactly by very fast procedures, equations (36) and (37) .

By using the above results, it is possible to construct a procedure, the Accelerated Hyperbolic Smoothing Clustering Algorithm, which has conceptual properties that offer a faster computational performance for solving the clustering problem given by formulation (16).

A fundamental question is the proper choice of the boundary parameter δ. Moreover, there are two main options for updating the boundary parameter δ, inside the internal minimization procedure or after it. For simplicity sake, the AHSC-L2 algorithm, the hyperbolic smoothing approach connected with the partition scheme, presented below adopts the second option, which offers a better computational performance, in spite of an eventual violation of the $\delta \geq \Delta x$ condition, which is corrected in the next partition update.

The Simplified AHSC-L2 Algorithm

Initialization Step:

Choose initial start point: x^0;

Choose parameter values: γ^1 , τ^1 , ε^1;

Choose reduction factors: $0 < \rho_1 < 1$, $0 < \rho_2 < 1$, $0 < \rho_3 < 1$;

Specify the boundary band width: δ^1;

Let $k = 1$.

Main Step: Repeat until an arbitrary stopping rule is attained

For determining the $Z_\delta(\overline{x})$ and $G_\delta(\overline{x})$ partitions, given by (22) and (23), use $\overline{x} = x^{k-1}$ and $\delta = \delta^k$.

Calculate the centers $v_i, i = 1, \ldots, q$ of gravitational regions by using (31).

Solve problem (17) starting at the initial point x^{k-1} and let x^k be the solution obtained:

For solving the equations (19), associated to the first part given

by (19), take the smoothing parameters: $\gamma = \gamma^k$, $\tau = \tau^k$ and $\varepsilon = \varepsilon^k$;

For solving the second part, given by (36), use the above calculated centers of the observations.

Updating procedure:

Let $\gamma^{k+1} = \rho_1 \gamma^k$, $\tau^{k+1} = \rho_2 \tau^k$, $\varepsilon^{k+1} = \rho_3 \varepsilon^k$

Redefine the boundary value: δ^{k+1}

Let $k := k + 1$. ∎

The above algorithm does not include any procedure for considering the occurrence of empty gravitational regions. This possibility can be overcome by simply moving the centroids.

The efficiency of the AHSC-L2 (HSC Method Connected with the Boundary and Gravitational Regions Partition Scheme) depends strongly on the parameter δ. A choice of a small value for it will imply an improper definition of the set $G_\delta(\overline{x})$ and frequent violation of the basic condition $\Delta x < \delta$, for the validity of Proposition 1. Otherwise, a choice of a large value will imply a decrease in the number of gravitational observation points and, therefore, the computational advantages given by formulation (36) will be reduced.

As a general strategy, within first iterations, larger δ values must be used, because the centroid displacements are more expressive. The δ values must be gradually decreased in the same proportion of the decrease of these displacements.

5 Computational Results

The computational results presented below were obtained from a preliminary implementation of the AHSC-L2 algorithm. The numerical experiments have been carried out on a PC Intel Celeron with 2.7GHz CPU and 512MB RAM. The programs are coded with Compac Visual FORTRAN, Version 6.1. The unconstrained minimization tasks were carried out by means of a Quasi-Newton algorithm employing the BFGS updating formula from the Harwell Library, obtained in the site: (http://www.cse.scitech.ac.uk/nag/hsl/).

In order to show the distinct performance of the AHSC-L2 algorithm, results obtained by solving a set of the largest problems of the TSP collection, Reinelt(1991) (http://www.iwr.uni-heidelberg.de/groups/comopt/software), are shown below.

Table 1 presents the results for the TSPLIB-3038 data set. It exhibits the results produced by the AHSC-L2 algorithm and, for comparison, those of two algorithms presented by Bagirov (2008). The first two columns show the number of clusters (q), and the best known value for the global optimum (f_{opt}) taken from Bagirov (2008). The next columns show the error (E) for the best solution produced (f_{Best}), and the mean CPU time $(Time)$ given in seconds associated to three algorithms: multi-start k-means (MS k-means), modified global k-means (MGKM) and the proposed AHSM-L2. The errors are calculated in the following way: $E = 100\,(f_{Best} - f_{opt})\,/\,f_{opt}$.

The multi-start k-means algorithm is the traditional k-means algorithm with multiple initial starting points. In this experiment, to find q clusters, 100 times q starting points were randomly chosen in the MS k-means algorithm. The global k-means algorithm, introduced by Likas et alli (2003), is a significant improvement of the k-means algorithm. The MGKS is an improved version of the Likas algorithm proposed by Bagirov(2008). The AHSC-L2 solutions were produced by using 10 starting points in all cases, except $q = 40$ and $q = 50$, where 20 and 40 starting points were taken, respectively.

It is possible to observe in each row of Table 1 that the best solution produced by the new AHSC-L2 algorithm becomes significantly smaller than that by MS k-means when the number of clusters q increases. In fact, this algorithm does not perform well for big instances, despite being one of most used algorithms. Wu et alli (2008) present the top 10 data mining algorithms identified by the IEEE International Conference on Data Mining in December 2006. The k-means assumes the second place in this list. The comparison between AHSC-L2 and MGKM solutions demonstrates similar superiority of proposed algorithm. In the same way, the comparison of time columns shows a consistent speed advantage of the proposed algorithm over the older ones.

On the other hand, the best solution produced by the AHSC-L2 algorithm

q	f_{opt}	MS k-means		MGKM		AHSC-L2 Algorithm	
		E	$Time$	E	$Time$	E	$Time$
2	0.31688E10	0.00	12.97	0.00	0.86	0.05	0.07
10	0.56025E09	0.00	11.52	0.58	3.30	0.01	0.28
20	0.26681E09	0.42	14.53	0.48	5.77	0.05	0.59
30	0.17557E09	1.16	19.09	0.67	8.25	0.31	0.86
40	0.12548E09	2.24	22.28	1.35	10.70	**-0.11**	1.09
50	0.98400E08	2.60	23.55	1.41	13.23	0.44	1.36
60	0.82006E08	5.56	27.64	0.98	15.75	**-0.80**	1.91
80	0.61217E08	4.84	30.02	0.63	20.94	**-0.73**	6.72
100	0.48912E08	5.99	33.59	0.00	26.11	**-0.60**	9.79

Table 1: Results for the TSPLIB-3038 Instance

is very close to the putative global minimum, the best known solution of the TSPLIB-3038 instance. Moreover, in this preliminary experiment, by using a relatively small number of initial starting points, four new putative global minimum results ($q = 40$, $q = 60$, $q = 80$ and $q = 100$) have been established.

q	$f_{Calculated}$	Algorithm HSC			Algorithm AHSC-L2		
		Occur.	E_{Mean}	$Time_{Mean}$	$Occur.$	E_{Mean}	$Time_{Mean}$
2	0.37491E16	4	0.86	23.07	5	0.58	3.65
3	0.22806E16	10	0.00	47.41	7	0.04	4.92
4	0.15931E16	10	0.00	76.34	10	0.00	5.76
5	0.13397E16	1	0.80	124.32	1	1.35	7.78
6	0.11366E16	8	0.12	173.44	2	1.25	7.87
7	0.97110E15	4	0.42	254.37	1	0.87	9.33
8	0.83774E15	8	0.55	353.61	4	0.37	12.96
9	0.74660E15	3	0.68	438.71	1	0.25	13.00
10	0.68294E15	4	0.29	551.98	3	0.46	14.75

Table 2: Results for the Pla85900 Instance

Table 2 presents the results for the Pla85900 data set. Ten different randomly chosen starting points were used. The first column presents the

specified number of clusters (q). The second column presents the best objective function value $(f_{Calculated})$ produced by the HSC algorithm and by the AHSC-L2 algorithm, both alternatives obtained the same results within a 5 decimal digits precision. The following three columns give the particular data associated to the original HSC algorithm: the number of occurrences of the best solution, the average error of the 10 solutions (E_{Mean}) in relation to the best solution obtained and CPU mean time given in seconds. The last three columns give the same data associated to the new AHSC-L2 algorithm

The results presented in Table 2 show a coherent performance of both algorithms. It was impossible to find any record of solutions of this instance. Indeed, the clustering literature seldom considers instances with such number of observations. The high number of occurrences of the best solution $(Occur.)$ and the low values presented in columns (E_{Mean}) show a consistent performance of both algorithms. The principal issue, the comparison between the mean CPU time values, shows clearly the extra performance of the new proposed AHSC-L2 algorithm resulting from the very fast procedures associated to equations (36) and (37).

Instance	$q = 5$			$q = 10$		
	$f_{AHSC-L2_{Best}}$	E_{Mean}	$Time_{Mean}$	$f_{AHSC-L2_{Best}}$	E_{Mean}	$Time_{Mean}$
FL3795	0.368283E09	6.18	0.18	0.106394E09	2.30	0.26
FNL4461	0.181667E10	0.43	0.31	0.853304E09	0.36	0.52
RL5915	0.379585E11	1.01	0.45	0.187794E11	0.41	0.74
RL5934	0.393650E11	1.69	0.39	0.191761E11	2.35	0.76
Pla7397	0.506247E14	1.94	0.34	0.243486E14	2.10	0.80
RL11849	0.809552E11	1.11	0.83	0.369192E11	0.53	1.55
USA13509	0.329511E14	0.01	1.01	0.149816E14	1.39	1.69
BRD14051	0.122288E11	1.20	0.82	0.593928E10	1.17	2.00
D15112	0.132707E12	0.00	0.88	0.644901E11	0.71	2.27
BRD18512	0.233416E11	1.30	1.25	0.105912E11	1.05	2.24
Pla33810	0.335680E15	0.22	3.54	0.164824E15	0.68	5.14

Table 3: Results for larger instances of the TSPLIB collection

Table 3 presents the computational results produced by the AHSC-L2 algorithm for the largest instances of the Symmetric Traveling Salesman

Problem (TSP) collection: FL3795, FNL4461, RL5915, RL5934, Pla7397, RL11849, USA13509, BRD14051, D15112, BRD18512 and Pla33810. For each instance, two cases are presented : $q = 5$ and $q = 10$. Ten different randomly chosen starting points were used. For each case, the table presents: the best objective function value produced by the AHSC-L2 algorithm ($f_{AHSC-L2_{Best}}$), the average error of the 10 solutions in relation to the best solution obtained (E_{Mean}) and CPU mean time given in seconds ($Time_{Mean}$).

It was impossible to perform a comparison, given the lack of records of solutions of these instances . Indeed, the clustering literature seldom considers instances with such number of observations. Only a possible remark: the low values presented in columns (E_{Mean}) show a consistent performance of the proposed algorithm.

6 Conclusions

In this paper, a new method for the solution of the minimum sum-of-squares clustering problem is proposed. It is a natural development that improves the global performance of the original HSC method presented by Xavier (2010). The robustness of the performance of the AHSC-L2 algorithm can be attributed to the complete differentiability of the approach. The high speed of the AHSC-L2 algorithm can be attributed to the partition of the set of observations into two non overlapping parts. This approach offers a drastic simplification of computational tasks.

It must be observed that the AHSC-L2 algorithm, as here presented, is firmly linked to the MSSC problem formulation. Thus, each different problem formulation requires a specific methodology to be developed, in order to apply the partition into boundary and gravitational regions.

Finally, it must be remembered that the MSSC problem is a global optimization problem with several local minima, so that both algorithms can only produce local minima. The obtained computational results exhibit a deep local minima property, which is well suited to the requirements of practical applications.

Acknowledgments

The author would like to thank Cláudio Joaquim Martagão Gesteira and Geraldo Veiga of Federal University of Rio de Janeiro for the helpful review of the work and constructive comments.

References

ANDERBERG, M. R. (1973) "Cluster Analysis for Applications", Academic Press Inc., New York.

BAGIROV, A. M. (2008) "Modified Global k-means Algorithm for Minimum Sum-of-Squares Clustering Problems", Pattern Recognition, Vol 41 Issue 10 pp. 3192-3199.

DUBES, R. C. and JAIN, A. K. (1976) "Cluster Techniques: The User's Dilemma", Pattern Recognition No. 8 pp. 247-260.

HANSEN, P. and JAUMARD, B. (1997) "Cluster Analysis and Mathematical Programming", Mathematical Programming No. 79 pp. 191-215.

HARTINGAN, J. A. (1975) "Clustering Algorithms", John Wiley and Sons, Inc., New York, NY.

JAIN, A. K. and DUBES, R. C. (1988) "Algorithms for Clustering Data", Prentice-Hall Inc., Upper Saddle River, NJ.

LIKAS, A., VLASSIS, M., and VERBEEK, J. (2003) "The Global k-means Clustering Algorithm", Pattern Recognition, Vol 36 pp. 451-461.

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

SPÄTH, H. (1980) "Cluster Analysis Algorithms for Data Reduction and Classification", Ellis Horwood, Upper Saddle River, NJ.

Wu, X. et alli (2008) "Top 10 algorithms in data mining", Knowledge and Information Systems, Springer, 14, pp. 1-37.

XAVIER, A.E. (1982) "Penalização Hiperbólica: Um Novo Método para Resolução de Problemas de Otimização", M.Sc. Thesis - COPPE - UFRJ, Rio de Janeiro.

XAVIER, A.E. (2010) "The Hyperbolic Smothing Clustering Method", Pattern Recognition, Vol 43, pp. 731-737.