Bregman divergences in the \((m \times k)\)-partitioning problem

G. Kokolakis\(^a,\)\(^*\), Ph. Nanopoulos\(^b\), D. Fouskakis\(^c\)

\(^a\)Department of Mathematics, National Technical University of Athens, Zografou Campus, Athens 15780, Greece
\(^b\)European Commission, DG Administration & Personnel, JMO, C2/28, L-2720, Luxembourg
\(^c\)Department of Mathematics, National Technical University of Athens, Zografou Campus, Athens 15780, Greece

Received 22 July 2005; received in revised form 13 December 2005; accepted 16 February 2006
Available online 20 March 2006

Abstract

A method of fixed cardinality partition is examined. This methodology can be applied on many problems, such as the confidentiality protection, in which the protection of confidential information has to be ensured, while preserving the information content of the data. The basic feature of the technique is to aggregate the data into \(m\) groups of small fixed size \(k\), by minimizing Bregman divergences. It is shown that, in the case of non-uniform probability measures the groups of the optimal solution are not necessarily separated by hyperplanes, while with uniform they are. After the creation of an initial partition on a real data-set, an algorithm, based on two different Bregman divergences, is proposed and applied. This methodology provides us with a very fast and efficient tool to construct a near-optimum partition for the \((m \times k)\)-partitioning problem.

© 2006 Elsevier B.V. All rights reserved.

Keywords: Confidentiality; Data masking; Fixed cardinality partitioning; Fixed size micro-aggregation; Bregman divergences; Pythagorean property; Convex partition

1. Introduction

This paper has its origins in the treatment of the problem of confidentiality encountered in Official Statistics. Ethics and legal provisions prevent National Statistical Services from disseminating individual records, even anonymized, if there is any possibility of direct or indirect identification of the individual behind a record. This basic principle frequently has negative implications when performing social and economic analysis. An account of concerns about the tension between confidentiality protection and data accessibility is given by Duncan and Pearson (1991) and Fienberg (1994).

It is obvious that methods aiming to control the risk of identity disclosure could suffer from bias introduction, correlation structure modification and data distortion. Thus, such methods could have, to a lesser or higher degree, negative implications in pursuing accurate statistical analysis or research studies for decision support purposes, see e.g., Adam and Wortmann (1989). It is essential therefore the development of a methodology that could soften the problem between data accessibility and confidentiality.

Many of the methods applied on the identity disclosure control problem belong to the matrix masking model, introduced by Duncan and Pearson (1991), where a triplet of matrices \((A, B, C)\) is used to transform the original...
microdata file $M_0$ into the masked microdata file $M_1 = AM_0B + C$. The problem of course is how to choose the mask $(A, B, C)$ in order to preserve in $M_1$ the maximum possible statistical information content about $M_0$ while protecting confidentiality.

One of the methods for the protection of individual information, that can be seen as a matrix masking technique, is that of micro-aggregation (Defays and Nanopoulos, 1993), where the individual records of the original microdata file are replaced by averages of small homogenous clusters of similar microdata from the original set. The group homogeneity is usually based on the squared Euclidean distance. This methodology was introduced in order to cope with a rule, formally applied in Official Statistics, according to which a minimum number of individual data have to be averaged before publishing. For this kind of approach cf. Domingo-Ferrer and Mateo-Sanz (2002).

In this paper, we are focusing on the concept of fixed cardinality partitioning, i.e., fixed size micro-aggregation. Instead of the Euclidean distance, a general family of loss functions is considered here, namely the Bregman divergence (Bregman, 1967). The homogeneity of the groups is achieved by minimizing the within-group divergence. The Bregman divergences between two $d$-dimensional random variables, defined as expectations of Bregman loss functions, have received an increased attention in the areas of decision theory and statistical inference, cf. Csiszár (1975), Jones and Byrne (1990), Csiszár (1991), Lafferty (1999), Grünwald and Dawid (2004) and Banerjee et al. (2005a).

In Section 2, basic micro-aggregation concepts together with the necessary terminology are introduced. In Section 3, mathematical results that refer to the Bregman family are presented. In Section 4, we are considering the problem of optimal partition with only two groups of $k$ and $(n - k)$ elements, respectively, while in Section 5 we are dealing with the problem of the $(m \times k)$-partitioning. Section 6 presents experimental results for two different divergences and in Section 7, conclusions are drawn.

2. Micro-aggregation and formalization of the general problem

Regarding the group size in micro-aggregation, as mentioned above, the rules in Official Statistics do not permit the transmission of aggregated data when the number of individuals aggregated is less than a threshold value $k$ (in most countries $k = 3$). We propose to apply this rule strictly and replace individual data by averages of small aggregates, which can play the role of fictitious individuals on which statistical data analysis could be performed. The problem is to partition the whole population in clusters of fixed size in an optimal way. The means of the “optimal” clusters would define the fictitious individuals whose data would be transmitted, in other words group-means would be the representatives of all data within groups.

Suppose that the total population $\Omega$ is formed by $n$ units, and $P$ is a probability measure over the power set $\mathcal{A}$ of $\Omega$. To each unit $\omega$ it corresponds a vector $X$ of $d$ variables, i.e. $X = X(\omega) = (X_1(\omega), \ldots, X_d(\omega))^T$. The objective is the partitioning of the set $\Omega$ into $m$ (=: $n/k$), groups of $k$ units, where $[\cdot]$ denotes the integral part, with the remaining ($n - mk$) data allocated to the most “appropriate” groups according to the methodology that follows. The $m$ groups, say $G_1, \ldots, G_m$, have to be formed in such a way that they would be as homogeneous as possible.

To define homogeneity of groups we need a notion of proximity or divergence $d(\omega, \omega')$ of the points in $\Omega$, which has to depend on the observed variables. Specifically we take:

$$d(\omega, \omega') = D(X(\omega), X(\omega')),$$

where $D$ is an appropriate divergence between two points in $\mathbb{R}^d$. Since a group-mean is the natural representative of all data within a group, the group homogeneity can be defined by:

$$\psi(G_i) = \sum_{\omega \in G_i} P(\omega | G_i) D(X(\omega), m_i) = E[D(X, m_i) | G_i],$$

where $m_i = E[X | G_i], i = 1, \ldots, m$. 
The quality of a partition can be defined by a kind of within-group divergence:

$$\Psi(\mathcal{G}) = \Psi(G_1, \ldots, G_m) = \sum_{i=1}^{m} P(G_i) \psi(G_i) = E[D(X, E[X|\mathcal{G}]])$$

(3)

where

$$E[X|\mathcal{G}] = \sum_{i} m_{i} I_{G_i}$$

(4)

with $I_{G_i}$ the indicator function of the set $G_i, i = 1, \ldots, m$.

The quantity in (3) has to be minimized with respect to $\mathcal{G}$ under the constraints:

$$|G_i| = k, \quad i = 1, \ldots, m,$$

where $|G_i|$ denotes the cardinality of the group $G_i$. These constraints will be violated only after the formation of the $m$ groups, when allocating the remaining $(n - mk)$ data to the most “appropriate” groups. This is what we call the last step augmentation.

With the above constraints we have the $(m \times k)$-partitioning problem. This differs from the classical well known clustering problem (Hartigan, 1975), where one wants to divide the whole population into a fixed number of groups without cardinality constraints.

3. The Bregman divergences in the $(m \times k)$-partitioning problem

The divergence functions we consider here are derived from the Bregman family of loss functions (Bregman, 1967). Such a loss function, named Bregman divergence, is defined by the quantity $D_\phi(x, y) = \phi(x) - \phi(y) - (x - y, \nabla \phi(y))$, where $\phi : \mathbb{R}^d \supseteq S \rightarrow \mathbb{R}$ is a strictly convex function, differentiable on the interior of $S$, and $\nabla \phi$ is the gradient of $\phi$. Under very mild conditions, the set of points where a convex function $\phi$ is not differentiable is of Lebesgue measure zero (Rockafellar, 1997, Theorem 25.2). Since $\phi$ is strictly convex, $D_\phi(x, y) > 0$, with equality if and only if $x = y$.

The Bregman divergence between two $d$-dimensional random variables $X$ and $Y$ is expressed by:

$$B(X, Y) = E[D_\phi(X, Y)] = E[\phi(X) - \phi(Y) - (X - Y, \nabla \phi(Y))]$$

(5)

with $B(X, Y) > 0$. The equality holds, if and only if, $X = Y$ a.e. $P$.

The most common case is when $\phi(x) = \|x\|^2$, the squared Euclidean distance. Then, $B(X, Y) = E[\|X - Y\|^2]$, i.e., the usual mean squared error. Another quite common case is the Mahalanobis distance, with $\phi(x) = x^T \Sigma^{-1} x$ and $\Sigma$ an $(d \times d)$ positive definite symmetric matrix. Then, $B(X, Y) = E[(X - Y)^T \Sigma^{-1} (X - Y)]$.

One of the most important properties of the above family of divergences, cf. Banerjee et al. (2005a,b), is that for any $\sigma$-field $\mathcal{F} \subset \mathcal{G}$ and for any $\mathcal{F}$-measurable function $Y$ we have:

$$\arg \min_{Y \in \mathcal{F}} E[D_\phi(X, Y)] = E[X|\mathcal{F}].$$

(6)

This is based on the fact that if $Y \in \mathcal{F}$ and $Y^* = E[X|\mathcal{F}]$, then

$$E[E[\{Y^* - X, \nabla \{\phi(Y^*) - \phi(Y)\}\]|\mathcal{F}]] = 0,$$

(7)

which implies

$$E[D_\phi(X, Y)] - E[D_\phi(X, Y^*)] = E[D_\phi(Y^*, Y)].$$

(8)
This is the Pythagorean property (Csiszár, 1991) of the Bregman divergences. The right-hand side of (8) is non-negative, and is equal to zero if and only if $Y = Y^*$, with probability one. This establishes (6).

Conversely, it has been proved by Banerjee et al. (2005a), that any divergence function with the property (6) belongs to the Bregman family.

By setting $Y = E[X]$ in (8) we have:

$$E[D_\phi(X, E[X])] - E[D_\phi(X, E[X|G])] = E[D_\phi(E[X|G], E[X])].$$

(9)

From (6) and (9) we realize that the within-group divergence $\Psi(G)$ in (3) can be written as follows:

$$\Psi(G) = E[D_\phi(X, E[X|G])] = E[D_\phi(X, E[X])] - E[D_\phi(E[X|G], E[X])],$$

(10)

and introducing (5), we can easily obtain

$$\Psi(G) = E[\phi(X)] - \Pi(G),$$

(11)

where

$$\Pi(G) = E[\phi(E[X|G])].$$

(12)

Thus, in the $(m \times k)$-partitioning problem, the minimization of $\Psi(G)$ is equivalent to the maximization of $E[D_\phi(E[X|G], E[X])]$, i.e. the between-group divergence, or equivalently the maximization of $\Pi(G)$.

Finally, the information loss, standardized between 0 and 1, can be provided by the quantity

$$L(G) = \frac{B(X, E[X|G])}{B(X, E[X])} = \frac{E[D_\phi(X, E[X|G])]}{E[D_\phi(X, E[X])]}.$$

(13)

and the optimal $(m \times k)$ partition is the one that minimizes $L(G)$.

4. The $(k, n - k)$-partitioning problem

Before we consider the general $(m \times k)$ case, let us focus on the problem with only two groups, namely $A$ and $A^c$ with sizes $k$ and $(n - k)$, respectively, i.e. the $(k, n - k)$-partitioning problem. Then we have to minimize the quantity

$$\Psi(G) = E[D_\phi(X, E[X|G])],$$

(14)

where $\phi$ is a strictly convex function in $\mathbb{R}^d$ and $G = \{A, A^c\}$, under the constraint $|A| = k$.

**Definition 1.** Let $G_1$ and $G_2$ be two different $(k, n - k)$ partitions of $\Omega$. We will call $G_1$ superior to $G_2$, if and only if $\Psi(G_1) < \Psi(G_2)$.

**Remark 1.** From (10) we have equivalently to maximize the quantity

$$E[D_\phi(E[X|G], E[X])] = P(A)D_\phi(m_A, m) + P(A^c)D_\phi(m_{A^c}, m),$$

(15)

where $m_A$ is the mean of the $A$ group, $m_{A^c}$ is the mean of the $A^c$ group and $m$ is the total mean. It is straightforward that in order to find the optimal $(k, n - k)$ partition, it is sensible to create the group $A$ starting from an extreme point $x_e$, i.e. a point whose divergence $D_\phi(x_e, m)$ from the total mean $m$ is high.
The following proposition provides us with a characterization of the optimal two-group partition $\mathcal{G} = \{A, A^c\}$ without cardinality constraints. In our case with cardinality constraints this proposition will be used as a guide for the creation of an initial partition.

**Proposition 1.** Let $\mathcal{G} = \{A, A^c\}$ be an unconstrained two-group partition of $\Omega$. Let also $x^* \in A$ such that:

$$D_\phi(x^*, m_A) > D_\phi(x^*, m_{A^c}),$$

(16)

Then, the partition $\mathcal{F} = \{B, B^c\}$ with $B = A \setminus \{x^*\}$ is superior to the partition $\mathcal{G} = \{A, A^c\}$.

**Proof.** We have:

$$\Psi(\mathcal{G}) = E[D_\phi(X, E[X|\mathcal{G}])]
= P[A]E[D_\phi(X, E[X|A])|A] + P[A^c]E[D_\phi(X, E[X|A^c])|A^c]
= P[A]E[D_\phi(X, m_A)|A] + P[A^c]E[D_\phi(X, m_{A^c})|A^c]
= \sum_A D_\phi(x, m_A) p(x) + \sum_{A^c} D_\phi(x, m_{A^c}) p(x)
> \sum_B D_\phi(x, m_A) p(x) + D_\phi(x^*, m_A) p(x^*) + \sum_{A^c} D_\phi(x, m_{A^c}) p(x)
= \sum_B D_\phi(x, m_A) p(x) + \sum_{B^c} D_\phi(x, m_{A^c}) p(x)
= P[B]E[D_\phi(X, E[X|A])|B] + P[B^c]E[D_\phi(X, E[X|A^c])|B^c]
\geq P[B]E[D_\phi(X, E[X|B])|B] + P[B^c]E[D_\phi(X, E[X|B^c])|B^c]
= E[D_\phi(X, E[X|\mathcal{F}])] = \Psi(\mathcal{F}),$$

with the last inequality due to (6). □

From Proposition 1 therefore, we can deduce that the group $A$ has to consist of points whose divergences from their group mean are low.

According to Remark 1, it is sensible to start the creation of the group $A$, from the most extreme point $x_e$, and then based on Proposition 1 to introduce additional points with low divergences from the current group mean of $A$ until we meet with the constraint $|A| = k$. This is exactly what the Importance Partitioning algorithm (IP), Kokolakis and Fouskakis (2005), is doing, an algorithm that will be used in our experimental section.

Proposition 1 provides us with a criterion for the optimal partition into two groups $A$ and $A^c$, without cardinality constraints. Inequality (16) implies that the two groups of the optimal partition are separated by a hyperplane. Therefore, the convex hulls of the two groups have no intersection. Such a partition of $\Omega$ is called convex.

When dealing with cardinality constraints, as it will be seen in the following section, convexity of the optimal partition is not always the case. The following proposition, applied after the creation of a $(k, n-k)$ partition $\mathcal{G}$, provides us with a very useful tuning that does not violate the cardinality requirement $|A| = k$.

**Proposition 2.** Let $\mathcal{G} = \{A, A^c\}$ be a $(k, n-k)$ partition of $\Omega$. Let also $x^* \in A$ and $y^* \in A^c$ such that:

$$D_\phi(x^*, m_A) p(x^*) + D_\phi(y^*, m_{A^c}) p(y^*) > D_\phi(x^*, m_{A^c}) p(x^*) + D_\phi(y^*, m_A) p(y^*).$$

(17)

Then, the partition $\mathcal{F} = \{B, B^c\}$ with $B = (A \setminus \{x^*\}) \cup \{y^*\}$ is superior to the partition $\mathcal{G} = \{A, A^c\}$. 

Proof. We have:

\[ \Psi(\emptyset) = E \left[ D_\emptyset(X, E[X|\emptyset]) \right] \]

\[ = P[A]E \left[ D_\emptyset(X, E[X|A])|A \right] + P[A^c]E \left[ D_\emptyset(X, E[X|A^c])|A^c \right] \]

\[ = P[A]E \left[ D_\emptyset(X, m_A) | A \right] + P[A^c]E \left[ D_\emptyset(X, m_{A^c}) | A^c \right] \]

\[ = \sum_A D_\emptyset(x, m_A) p(x) + \sum_{A^c} D_\emptyset(x, m_{A^c}) p(x) \]

\[ = \sum_{A \setminus \{x^*\}} D_\emptyset(x, m_A) p(x) + D_\emptyset(y^*, m_A) p(x^*) + D_\emptyset(y^*, m_{A^c}) p(x^*) + \sum_{A \setminus \{y^*\}} D_\emptyset(x, m_{A^c}) p(x) \]

\[ = P[B]E \left[ D_\emptyset(X, E[X|B])|B \right] + P[B^c]E \left[ D_\emptyset(X, E[X|B^c])|B^c \right] \]

\[ \geq P[B]E \left[ D_\emptyset(X, E[X|B])|B \right] + P[B^c]E \left[ D_\emptyset(X, E[X|B^c])|B^c \right] \]

\[ = E \left[ D_\emptyset(X, E[X|\emptyset]) \right] = \Psi(\emptyset). \quad \square \]

From Proposition 2 we conclude, that the \((k, n-k)\) partition \(\emptyset = \{A, A^c\}\) in order to be optimal it is necessary the following condition to hold:

\[ D_\emptyset(x, m_A) p(x) + D_\emptyset(y, m_{A^c}) p(y) \leq D_\emptyset(x, m_{A^c}) p(x) + D_\emptyset(y, m_A) p(y), \quad (18) \]

for every \(x \in A\) and \(y \in A^c\).

5. The optimal \((m \times k)\) partition

In this section we examine the consequences of the condition (18) and we generalize our findings for the general \((m \times k)\)-partitioning problem. Condition (18), after some algebra, can be written as follows:

\[ \{ (x, c) + \alpha \} p(x) \geq \{ (y, c) + \alpha \} p(y), \quad (19) \]

where

\[ c = \nabla \{ \varphi(m_A) - \varphi(m_{A^c}) \} \quad (20) \]

and

\[ \alpha = \varphi(m_A) - \varphi(m_{A^c}) + \langle m_{A^c}, \nabla \varphi(m_{A^c}) \rangle - \langle m_A, \nabla \varphi(m_A) \rangle \]

\[ = \frac{1}{2} \{ D_\emptyset(m_A, m_{A^c}) - D_\emptyset(m_{A^c}, m_A) - \langle m_A + m_{A^c}, c \rangle \}. \quad (21) \]

Thus, we have proved the following.
Table 1
Six points in $\mathbb{R}^2$ with the corresponding probabilities

<table>
<thead>
<tr>
<th>$x$</th>
<th>$p(x)$</th>
<th>$y$</th>
<th>$p(y)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(-1.0, 2.5)</td>
<td>0.20</td>
<td>(-22.0, -0.5)</td>
<td>0.20</td>
</tr>
<tr>
<td>(-2.0, -2.0)</td>
<td>0.20</td>
<td>(-6.0, 1.0)</td>
<td>0.05</td>
</tr>
<tr>
<td>(10.5, 1.4)</td>
<td>0.15</td>
<td>(40.9, -3.7)</td>
<td>0.20</td>
</tr>
</tbody>
</table>

Fig. 1. The optimal $(3, 3)$ partition with data provided by Table 1.

**Theorem 1.** A necessary condition for a $(k, n-k)$ partition $\mathcal{G} = \{A, A^c\}$ to be optimal is the following:

$$\langle \lambda x - (1-\lambda)y, e \rangle \geq \lambda(1-2\lambda),$$

(22)

for every pair of points $x \in A$ and $y \in A^c$, with $\lambda = p(x)/(p(x) + p(y))$.

From the above theorem we can conclude the following:

- The two groups $A$ and $A^c$ of the optimal $(k, n-k)$ partition of $\Omega$, might not be separated by a hyperplane, depending on the involved probabilities of the points. As a simple example, consider six points in $\mathbb{R}^2$ with the corresponding probabilities as given in Table 1. Using the Euclidean distance and a fixed group-size $k = 3$, we find out that the first group $A$, of the optimal partition, consists of the first three data points in Table 1. We realize then, from Fig. 1, that this optimal $(3, 3)$ partition is not a convex one.
- The normal vector, $n = e/\|e\|$, of the involved hyperplane depends only on the group-means. This is due to the Pythagorean property of Bregman divergences.
- With uniform probability measure condition (22) takes the form

$$\langle x - y, e \rangle \geq 0.$$  

(23)

Thus, the two groups are separated by a hyperplane.

We are considering now two special cases for the function $\varphi$. 

Euclidean distance. When \( \varphi(x) = \|x\|^2 \), then condition (22) takes the form
\[
\langle \lambda x - (1 - \lambda) y, m_A - m_{A'} \rangle \geq \frac{1 - 2\lambda}{2} \left\{ \|m_A\|^2 - \|m_{A'}\|^2 \right\},
\]
and under uniformity of the probability measure condition (24) takes the form:
\[
\langle x - y, m_A - m_{A'} \rangle \geq 0.
\]
We realize therefore, that in the case of Euclidean distance and with uniform probability measure, the two groups are separated by a hyperplane perpendicular to the line joining their means.

Mahalanobis distance. When \( \varphi(x) = x^T \Sigma^{-1} x \), with \( \Sigma \) a \((d \times d)\) symmetric and positive definite matrix, then condition (22) takes the form
\[
\langle \lambda x - (1 - \lambda) y, \Sigma^{-1} (m_A - m_{A'}) \rangle \geq \frac{1 - 2\lambda}{2} \left\{ m_A^T \Sigma^{-1} m_A - m_{A'}^T \Sigma^{-1} m_{A'} \right\},
\]
and under uniformity of the probability measure condition (26) takes the form:
\[
\langle x - y, \Sigma^{-1} (m_A - m_{A'}) \rangle \geq 0.
\]
The above methodology can be applied in the \((m \times k)\)-partitioning problem; therefore, we have the following theorem which is an extension of Proposition 2.

**Theorem 2.** A necessary condition for an \((m \times k)\) partition \( \mathcal{G} = \{G_1, \ldots, G_m\} \) to be optimal is the following:
\[
\sum_{j=1}^{\ell} D_\varphi (x_{ij}, m_{ij}) p(x_{ij}) \leq \sum_{j=1}^{\ell} D_\varphi (x_{ij}, m_{ij+1}) p(x_{ij}),
\]
for any set of groups \( \{G_{i1}, \ldots, G_{i\ell}\}\) and every set of points \(x_{ij} \in G_{ij}, j = 1, \ldots, \ell, \ell = 2, \ldots, m,\) with \(m_{ij}\) the mean of the group \(G_{ij}\) and \(m_{i\ell+1} = m_i\).

The quantity \(\ell\) can be considered as the depth of the search that will be performed in order to come closer to the optimal \((m \times k)\) partition. The larger it is, the closer to the optimal partition we get, but additional computational time is required. In the experimental section that follows we take \(\ell = 2\) for computational reasons, and therefore it is crucial to apply our methodology starting with a good initial partition.

6. Experimental results

A real data-set of 1000 households in Greece was used, based on the “1998’s Household Budget Continuous Survey (HBCS)” collected by the National Statistical Service of Greece (NSSG). For each household 12 quantitative variables were collected, representing expenses for: Food and Non-Alcoholic Beverages (E1), Alcoholic Beverages and Tobacco (E2), Clothing and Footwear (E3), Housing (E4), Furniture and Household Equipment (E5), Health (E6), Transport (E7), Communications (E8), Recreation and Culture (E9), Education (E10), Hotels, Cafes and Restaurants (E11) and Miscellaneous (E12).

Two special cases for the function \(\varphi\) are considered here; the Euclidean and the Mahalanobis distance with \(\Sigma\) the covariance matrix of the data. The uniform probability measure has been applied in both cases. Finally, we used the information loss criterion (13) as a measure of the quality of the partition in every case. Since there were missing data in some of the variables, we performed two different comparisons, one with 880 households and six variables, namely (E1), (E4), (E5), (E8), (E9) and (E12), and another one with 640 households and 8 variables, namely (E1), (E3), (E4), (E5), (E8), (E9), (E11) and (E12), for fixed group sizes with \(k = 3, k = 4, k = 5\) and \(k = 6\) elements.
Table 2
Comparison of information loss with different group sizes and distances, using IP and ILS to create the initial partitions and afterwards applying the Inner Product algorithm. In the first part of the table we have 880 data points of dimension 6 and in the second part we have 640 data points of dimension 8.

<table>
<thead>
<tr>
<th>Group size (k)</th>
<th>Euclidean distance</th>
<th>Mahalanobis distance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>IP</td>
<td>IP+INPR</td>
</tr>
<tr>
<td>n = 880, d = 6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>k = 3</td>
<td>0.0899</td>
<td>0.0893</td>
</tr>
<tr>
<td>k = 4</td>
<td>0.1212</td>
<td>0.1205</td>
</tr>
<tr>
<td>k = 5</td>
<td>0.1456</td>
<td>0.1449</td>
</tr>
<tr>
<td>k = 6</td>
<td>0.1681</td>
<td>0.1644</td>
</tr>
<tr>
<td>n = 640, d = 8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>k = 3</td>
<td>0.1314</td>
<td>0.1300</td>
</tr>
<tr>
<td>k = 4</td>
<td>0.1698</td>
<td>0.1688</td>
</tr>
<tr>
<td>k = 5</td>
<td>0.2014</td>
<td>0.2006</td>
</tr>
<tr>
<td>k = 6</td>
<td>0.2300</td>
<td>0.2274</td>
</tr>
</tbody>
</table>

At the beginning we created initial partitions according to two different heuristic algorithms: (a) the algorithm IP, Kokolakis and Fouskakis (2005) and (b) the Initial Linear Separability algorithm (ILS). Both these algorithms make use of Remark 1, starting from the most extreme point $x_e$, and of Proposition 1, creating a group $A$ with a small within-group divergence. In the first algorithm this is achieved on the basis of the divergence of the points from the extreme, while in the second according to the divergence of the projections of the points on the line $\{ax_e : a \in \mathbb{R}\}$ from the extreme. The two algorithms are described as follows:

**Algorithm 1** (Importance Partitioning (IP)).

- Choose the most extreme point $x_e$.
- Construct the first group by finding initially the closest point to $x_e$, and then introduce each time a new one that is closer to the mean of the already selected points, until a group of $k$ points is formed.
- Exclude this group from the data-set. If they remain more than $k$ data points, find the new extreme point and repeat Step 2.
- If they remain $k$ data points, these form the last group.
- If they remain less than $k$ data points, introduce each one to its closest group.
- End.

**Algorithm 2** (Initial Linear Separability (ILS)).

- Center the data.
- Find the most extreme point $x_e$.
- Find the $(k - 1)$ points whose projections on the line $\{ax_e : a \in \mathbb{R}\}$ are closer to $x_e$. These $(k - 1)$ points together with $x_e$ form a group of $k$ points.
- Exclude this group from the data-set. If they remain more than $k$ data points, repeat Steps 1–3.
- If they remain $k$ data points, these form the last group.
- If they remain less than $k$ data points, introduce each one to its closest group.
- End.

After the creation of the initial partitions, we make use of Theorem 2 with $\ell = 2$, by running an algorithm named Inner Product (INPR). Specifically, we require for any two groups not to exist a pair of points belonging to them, respectively, that violate condition (28).

The partition created from the IP algorithm is expected to be a good starting partition according to Kokolakis and Fouskakis (2005), thus, it would be interesting to see the extent of the improvement that can be accomplished after
applying the INPR algorithm. On the other hand, the ILS algorithm is much faster than IP but is not expected to provide a very good initial partition and, for sensitivity reasons, it would be very useful to examine whether or not the INPR algorithm is affected by this handicap.

Table 2 presents the resulting information loss in each case. We note that, in every case, the INPR improves the initial solution. The improvement is more dramatic when starting from not a very good initial partition (ILS). The best results were achieved, when starting from the partition created by the IP algorithm. Taking into account that the complexity of the IP is quite low, as it has been indicated in Kokolakis and Fouskakis (2005), the combination of IP with INPR provides a very fast and efficient way to construct a near-optimum partition.

7. Conclusions

This paper examines a technique for ensuring the protection of confidential statistical information, while preserving the information content of the data. The basic idea of the method is to aggregate the data in clusters of fixed small size, by minimizing Bregman divergences. We showed that, in the case of non-uniform probability measures the optimal solution is not necessarily convex, i.e. the groups are not separated by hyperplanes, while with uniform it is. We proposed an algorithm, based on our findings, that provides us with a very fast and efficient tool to construct a near-optimum partition for the \((m \times k)\)-partitioning problem. As a future work, it is under consideration to apply, after the creation of the partition based on our methodology, known stochastic optimization techniques, such as tabu search and genetic algorithms, for further improvements.

Acknowledgments

After having submitted the first version of this paper, a referee mentioned a related work by Banerjee et al. (2005b); they have proposed and analyzed parametric hard and soft clustering algorithms based on Bregman divergences. Despite some similarities in the general set up, their methodology and proposed algorithm refer to the unconstrained clustering problem, while here we are dealing with the fixed cardinality partition problem.

We are grateful to the National Statistical Service of Greece for making the data available to us and to the referee and the associate editor for comments and references that improved the paper.

References

