THE MOMENT PRESERVATION METHOD OF CLUSTER ANALYSIS

By

Bernard Harris
Department of Statistics
University of Wisconsin-Madison
THE MOMENT PRESERVATION METHOD OF CLUSTER ANALYSIS

BERNARD HARRIS

UNIVERSITY OF WISCONSIN
MADISON, WISCONSIN, U.S.A.

Abstract. A technique for cluster analysis, known as the moment preservation method is found primarily in the engineering literature. The method appears to be motivated by mixture models. A brief description of this technique is as follows. If q clusters are to be determined, then 2q-1 linearly independent functions of the data are calculated. (In many of the applications that I have encountered, these are the first 2q-1 sample moments.) For univariate data, the values of these functions are used to determine thresholds. The thresholds are chosen so that these moments are preserved. For multivariate data, the thresholds are replaced by their natural analogue, linear manifolds. The theoretical properties of this process can be determined from a substantial body of mathematical analysis, known as the "Reduced Moment Problem". These properties facilitate determining the solutions sets and their properties, including the determination of optimal solutions.

1. Introduction and Summary.

A technique for determining clusters, known as the moment preservation technique for cluster analysis has appeared in the engineering literature; in particular, see A. J. Tabatai and O. R. Mitchell (1984), W. H. Tsai (1985), Soo-Chang Pei and Ching-Min Cheng (1986), E. J. Delp and O. R. Mitchell (1979), A. J. Tabatai (1981), J. C. Lin and W. H. Tsai (1994). The proponents of this method claim that it is highly accurate and easily computable. The authors of the above papers appeared to be unaware that there is a substantial mathematical literature applicable to their procedure and that these results can be readily extended to more general situations.

2. The Moment Preservation Method.

The present form of the moment preservation method will now be described for univariate data. Let \( x_1, x_2, \ldots, x_n \) be the realizations of a random sample from a continuous distribution. The number of clusters to be obtained needs to be specified in advance. Assume that \( q \) (\( q \leq n \)) clusters are desired. Then the first \( 2q-1 \) non-central sample moments are calculated. The unique discrete probability distribution with positive probability at \( q \) points, which possesses these moments, is determined. Thus, to obtain two clusters, the first three sample moments are calculated, which determines the two point distribution, \( (p_1, y_1; 1-p_1, y_2, 0 < p_1 < 1, y_1 < y_2) \). In general, the solution is given by \( (p_i, y_i, i=1,2,\ldots,q) \).
p_i \geq 0, \sum_{i=1}^{q} p_i = 1, y_1, y_2, \ldots, y_q; y_i is the center of the i-th cluster. The data points x_i, i=1,2,\ldots,n, are assigned to the q clusters as follows. Arrange the n data points in increasing order. The observations whose rank does not exceed np_1 are placed in the first cluster. From the remaining observations, those whose rank does not exceed n(p_1+p_2) are placed in the second cluster. This process is continued until all observations have been assigned to a cluster. The process of obtaining the q-point probability distribution, referred to above, is computationally identical to solving the "classical" reduced moment problem.

3. The Reduced Moment Problem and Its Solutions.

Let M_k = \{\mu_1, \mu_2, \ldots, \mu_k\} be a given realizable sequence of the first k non-central moments. Realizable means that there exists a cumulative distribution function F(x) such that

\int_{-\infty}^{\infty} x^j dF(x) = \mu_j, j=1,2,\ldots,k.

F(x) is unknown. The problem of determining the set of possible probability distributions with these moments is called the (classical) reduced moment problem. There are three cases.

The Hausdorff moment problem. The carrier set of F(x) is a subset of [0,1].
The Stieltjes moment problem. The carrier set is a subset of [0,\infty).
The Hamburger moment problem. The carrier set is a subset of (-\infty,\infty).

There is an extensive body of literature on this subject as well as on various modifications and extensions of it. Space limitations prevent even a cursory discussion of this subject, nevertheless, for interested readers, the following references may be of interest: J.A. Shohat and J.D. Tamarkin (1943), S. Karlin and L.S. Shapley (1953), N. Ahiezer and M.G. Krein (1938).

To characterize the solution set, we will use a geometric approach, based on the theory of convex functions and convex sets. The following basic results are relevant. Let \mathcal{X} denote a subset of the real line (such as \{0,1\}, [0,\infty)), and so forth.

1. The set of probability distributions on \mathcal{X} is a convex set.
2. The set of discrete probability distributions on \mathcal{X} is a convex set.
3. The set of probability distributions on \mathcal{X} with finite support is a convex set.

In each of the above situations, the set of extreme points is the set of degenerate (one-point) distributions. In particular, the set of probability distributions on \mathcal{X} is the weak closure of the set of finite convex combinations of the degenerate distributions. That is, all distributions are "limits" of finite convex combinations of the degenerate distributions.
4. Let $\mathcal{M}_k$ be the set of realizable moment sequences of length $k$ for probability distributions on $\mathcal{X}$. Then $\mathcal{M}_k$ is a convex set in $E_k$, (k-dimensional Euclidean space).
5. The set of probability distributions $\mathcal{F}(\mathcal{M}_k)$ on $\mathcal{X}$ with a given realizable moment sequence $\mathcal{M}_k$ is a convex set.

The extreme points of $\mathcal{M}_k$ are the vectors of the form $(a,a^2,\ldots,a^k)$, that is, the reduced moment sequences of the degenerate distributions. Thus, for example, for $k=2$, $\mathcal{X}=[0,1]$, the set of realizable moments is the set bounded by the line from $(0,0)$ to $(1,1)$ and the curve $\{(x,x^2), 0 \leq x \leq 1\}$. A useful characterization of moment sequences in terms of the extreme points is provided by a theorem of Carathéodory, which we will now state and subsequently, we will show how this theorem can be utilized to select solutions of the reduced moment problem with specific properties.

**Carathéodory's Theorem.** Let $C$ be a compact convex set in $E_k$. Let $u \in C$. Then $u$ has a representation as a convex combination of $\leq k+1$ extreme points of $C$. If $u$ is a boundary point of $C$, then $u$ has a representation as a convex combination of $\leq k$ extreme points of $C$.

To see how Carathéodory's Theorem can be exploited, we consider the specific case $k=2$ and $\mathcal{X}=[0,1]$. If a point $u$ is on the boundary, there are two possibilities. A point on the "lower boundary" is the image of a degenerate distribution, which has probability one at some value in $[0,1]$. A point on the upper boundary is a convex combination of degenerate distributions at 0 and 1. Now assume that $u$ is in the interior of $C$. $u$ can be represented as a convex combination of two points on the lower boundary or one point on the lower boundary and one point on the upper boundary or three points on the lower boundary. The coefficients of the convex combination can be readily determined by comparing the distances of $u$ to each of the extreme points in the convex combination.

4. **Some Extensions and Generalizations.**

As far as the mathematical theory is concerned, there is no reason to restrict to the first $k$ monomials; $x, x^2, \ldots, x^k$, which determine the first $k$ non-central moments. In principle any set of $k$ linearly independent functions could be used. However, the computations necessary for determining the solutions are more complicated if the functions other than $x$ are not strictly convex.

Some classes of variational problems can be treated using the same theory. For background material, see J. S. Rustagi (1976), B. Harris (1959), B. Harris (1962).

Let $g(x)$ be a real valued function, which is integrable on $\mathcal{F}(\mathcal{M}_k)$. Then we wish to maximize (minimize) $\int g(x)d\mathcal{F}(x), \mathcal{F} \in \mathcal{F}(\mathcal{M}_k)$. This is of interest in the clustering problem, which motivated this study, in the following way. If $g(x)$ is some measure of merit of the clustering procedure, such as the probability of misclassification, then we would wish to choose the best solution from among the solutions of the reduced moment problem. The procedure requires considering the "reduced moment set" generated by $\{g(x), x, x^2,\ldots, x^k\}$.
It is not possible to describe the possible procedures in the available space. The methodology uses the supporting hyperplane theorem for convex sets and/or the Hahn-Banach extension theorem. However, the following brief remarks can be provided. In general, extremal solutions tend to be generated by distributions with small carrier sets. The computations are easily carried out if g(x) is a well-behaved function. Depending on the particular problem, such include completely monotonic functions, absolutely monotonic functions, or strictly convex functions.

In the preceding material, only the univariate case was discussed. The multivariate case is feasible, but substantially more complicated. There are some problems with identifiability in multidimensional situations. One alternative that has been utilized is to subject the data to principal component analysis. This permits the reduction of the multivariate problem to a sequence of one dimensional problems. Treating the first principal component as a one dimensional problem, carry out the moment preservation method for projections of the data onto the first principal component. Then project the residuals onto the second principal component and so forth. There is also some work in progress in using binary decision trees for such problems, but this is in an early stage.

5. A Numerical Illustration.

We consider two problems using the same data. For simplicity, the data is artificial, so that the solutions are particularly simple. For the first, we determine two clusters using the first three sample moments. This is precisely the engineering application that motivated this study. The second problem will utilize two moments and characterize the set of distributions that have these moments. We will obtain a parametric representation for the elements of this set. If we choose a specific parameter, then we get the unique solution that is obtained as the solution to the first problem.

Denote the sample moments by \( m_1, m_2, m_3 \). Then the solution is given by

\[
a = \frac{m_1 m_3 - m_2^2}{m_2 - m_1^2}, \quad b = \frac{m_1 m_2 - m_3}{m_2 - m_1^2},
\]

\[
y_1 = .5\{-b-(b^2-4a)^{1/2}\}, \quad y_2 = .5\{-b-(b^2-4a)^{1/2}\},
\]

\[
p_1 = \frac{y_2-m_1}{y_2-y_1}, \quad p_2 = 1-p_1.
\]

Specifically, let \( m_1 = .5, m_2 = .375, m_3 = .3125 \). We restrict to distributions on \([0,1]\). Proceeding according to the above recipe, we get \( a = .125, b = -.1 \). Then \( y_1 = .1464 \),
Finally, we obtain $p_1 = p_2 = .5$. It is easily seen that these satisfy the moment conditions. Assume that $n$, the sample size is an even number. This generates the following two clusters as follows: rank the data, split the data at the sample median and put the $n/2$ smallest observations in the first cluster and the remainder in the second cluster.

Now we characterize all two point distributions on $[0,1]$ with mean $.5$ and second moment $.375$. The point $(.5,.375)$ is in the interior of $M_2$. The totality of two point representations can be constructed as follows. Choose any $x \in [0.1]$. Draw the line segment from $(x,x^2)$ through $(.5,.375)$ and continue until it reaches the boundary of $M_2$. The set of two point distributions are characterized by the set of such lines which reach the lower boundary. It is easily seen that every solution has $y_2 \in [.75, 1]$ and symmetrically $y_1 \in [0, .25]$. Also, at the extremal solutions, $(0,.75), p_1 = 1/3$ and $p_2 = 2/3; (.25, 1), p_1 = 2/3$ and $p_2 = 1/3$. For every $x$ in this interval, $p_1$ and $p_2$ are easily calculated. A simple parametric representation is possible which can be used to solve variational problems.


This paper provides a very brief description of the moment preservation method of cluster analysis and some of the ramifications of this technique. Space limitations prevented an extensive discussion. An extensive discussion, which will include mathematical details and proofs will be prepared in the near future.

REFERENCES


Shohat, J. A. and Tamarkin, J. D. (1943) : The Problem of Moments, American Mathematical Society, Providence, Rhode Island

Tabatai, A. J. (1981) : Edge Location and Data Compression for Digital Imagery, PhD. Dissertation, School of Electrical Engineering, Purdue University, Lafayette, Indiana, U.S.


Keywords: Cluster Analysis, Moment Preservation, Reduced Moment Problem.