**DIMACS Technical Report 2008-01**
**February 2008**

# Variance-based Criteria for Clustering and Their Application to the Analysis of Management Styles of Mutual Funds Based on Time Series of Daily Returns

by

Nikita I. Lytkin
Dept. of Computer Science
Rutgers University
New Brunswick, New Jersey 08903

Casimir A. Kulikowski
Dept. of Computer Science
Rutgers University
New Brunswick, New Jersey 08903

Ilya B. Muchnik
DIMACS
Rutgers University
New Brunswick, New Jersey 08903

# ABSTRACT

The problem of clustering is formulated as the problem of minimization of a certain objective function over the set of all possible clusterings. The objective function measures mathematically the quality of a clustering. According to a previously published theoretical result, if the objective function being minimized is strictly convex, then the corresponding clustering surface is strictly convex. As a direct implication of this result followed the construction of a basic gradient algorithm of search for locally optimal solutions (i.e., clusterings). This gradient procedure constitutes the core of the clustering algorithms proposed in this work for minimization of two novel objective functions.

An important direction in statistical sampling theory deals with construction of optimal stratified samples from a population. One of the problems addressed by stratified sampling is the construction of a sample for estimation of the mean value of a particular scalar parameter, such that the variance of the estimate is minimized. For this purpose, a criterion for optimal partitioning of the population into a certain number of groups (strata) was derived. This criterion is known as Neyman's criterion for optimal stratified sampling.

Neyman's criterion was originally developed for one-dimensional data. This criterion is generalized to $n$-dimensional space, and is used as an objective function for clustering. A proof of strict concavity of the generalized objective function is given. Strict concavity provided the basis for a K-means-like gradient-based clustering algorithm proposed in this work.

The other objective function investigated in this work, also originated in statistical sampling theory, where it was derived for optimal selection of representative types for estimation of the mean value of a certain scalar parameter. Selection of representative types differs from stratified sampling in that, under the former sampling scheme, a single representative is sampled from each group. A proof of non-convexity of this objective function is given. An efficient stochastic procedure for minimization of this objective function is proposed. The procedure uses, as an elementary operation, a modification of the basic gradient algorithm mentioned above, and has the same computational complexity as the well-known K-means algorithm.

The proposed clustering methods and the K-means method are similar, both intuitively and mathematically. For this reason, a systematic comparative analysis of these methods was performed. Experimental results obtained on synthetic data illustrate the differences in the forms of the discriminant surfaces constructed by each of the three clustering algorithms. In contrast to K-means, which produces linear discriminant surfaces, the other two criteria produce quadratic discriminant surfaces. These results also help in the interpretation of differences between clusterings produced by the three algorithms.

An application of all three methods to clustering real-world time series data is demonstrated. In this work, time series were treated as static points in $n$-dimensional space. The dataset consisted of time series of daily returns of 6671 mutual funds from May 2005 until May 2006. The results obtained closely corresponded to the outcomes of an informal financial analysis of hidden information on the funds' management styles (dynamics of the funds' portfolios). By applying three different clustering methods, three different results were obtained. The consistent part of these clusterings was interpreted as the most robust and objectively consistent component in the existing classification of mutual funds by their management styles.

Preliminary experimental results obtained in this work suggest that in practice it is useful to apply all three clustering methods together in order to aid in the discovery of consistent cores within the data.

Further plans for applications of clustering methods on time series data in the domain of cyber security are outlined.

# 1 Introduction

We consider clustering as the problem of minimization of a certain criterion (functional), which measures mathematically the quality of a clustering. Clustering methods considered in this work partition an $n$-dimensional space into $K$ non-overlapping regions. Throughout the rest of this paper, we will refer to these regions as classes and to partitions of the corresponding $n$-dimensional space as classifications.

A theoretical result published in [3] states that if a classification criterion being minimized is strictly convex, then the corresponding classification surface is strictly convex, as well. As a direct implication of this result followed the construction of a basic gradient algorithm of search for locally optimal solutions (i.e., classifications). This gradient procedure constitutes the core of the classification algorithms proposed in this work for minimizing two novel criteria.

An important direction in statistical sampling theory deals with construction of optimal stratified samples from a population. One of the problems addressed by stratified sampling is the construction of a sample for estimation of the mean value of a particular scalar parameter, such that the variance of the estimate is minimized [4]. For this purpose, a criterion for optimal partitioning of the population into a certain number of groups (strata) was derived [15]. This criterion is known as Neyman's criterion for optimal stratified sampling.

Neyman's criterion was originally developed for one-dimensional data. In this work, we generalize this criterion to $n$-dimensional space and present a proof of strict concavity of the generalized functional. Strict concavity provided the basis for a K-means-like gradient-based classification algorithm proposed in this work.

The other functional investigated in this work, also originated in statistical sampling theory, where it was derived [11] for optimal selection of representative types (from one-dimensional data) for estimation of the mean value of a certain scalar parameter. Selection of representative types differs from stratified sampling in that, under the former sampling scheme, a single representative is sampled from each group.

Functional [11] was first generalized to $n$-dimensional space in [18], where a classification algorithm for this functional was also given. However, the computational complexity of the algorithm [18] is exponential in $n$ rendering the algorithm impractical for $n > 2$.

We prove that functional [11] is non-convex and give an efficient stochastic procedure for minimization of this functional for $n$-dimensional data. The procedure uses, as an elementary operation, a modification of the basic gradient algorithm mentioned above, and has the same computational complexity as the well-known K-means algorithm [14].

The two classification methods proposed in this work and the K-means method are similar, both intuitively and mathematically. For this reason, we performed a systematic comparative analysis of these methods. Experimental results obtained on synthetic data illustrate the differences in the forms of the discriminant surfaces constructed by each of the three classification algorithms. In contrast to K-means, which produces linear discriminant surfaces, the other two criteria investigated produce quadratic discriminant surfaces. These results also help in the interpretation of differences between classifications produced by the three algorithms.

Finally, we present a substantial real-world application of the algorithms for classification of time series data [2, 6, 8, 10, 12, 13, 16, 17, 19]. In this work, time series were treated as static points in $n$-dimensional space. The dataset consisted of time series of daily returns of 6671 mutual funds from May 2005 until May 2006. Experimental results demonstrated that the obtained classifications closely corresponded to the outcomes of an informal financial analysis of hidden information on the funds' management styles [5, 7] (dynamics of the funds' portfolios). By applying three different classification methods, three different classifications were obtained. The consistent part of these classifications was interpreted as the most robust and objectively consistent component in the existing classification of mutual funds by their management styles.

Preliminary experimental results obtained in this work suggest that in practice it is useful to apply all three classification methods together in order to aid in the discovery of consistent cores within the data.

In Section 2 we provide the theoretical background underlying the development of the classification algorithms described in the succeeding sections. In Section 3 we prove convexity properties of the criteria discussed, and give the corresponding classification algorithms. Experimental results are described in Section 4. Concluding remarks and further plans for applications of classification methods on time series data in the domain of cyber security are outlined in Section 5.

## 2 Mathematical Foundations: Criteria of Optimality for Constructing Classifications

In this work, we consider methods that produce non-overlapping classifications of objects into a specified number $K$ of classes. A classification is termed non-overlapping if every object is assigned to exactly one of $K$ classes. We consider objects as points in an $n$-dimensional Euclidean space. A non-overlapping classification in an $n$-dimensional Euclidean space is a partition of this space into $K$ non-overlapping regions such that points within the same region are assigned to the same class, while points of different regions are assigned to different classes. Throughout this work, the term classification will be used to signify a non-overlapping classification.

Formally, let $\mathcal{X}$ denote an $n$-dimensional Euclidean space where each distinct data instance is uniquely characterized by a vector $x \in \mathcal{X}$. A classification $H$ is a partition of space $\mathcal{X}$ into $K$ non-overlapping regions, and is determined by a set of characteristic functions $H = (h_1(x), \dots, h_K(x))$, where

$$h_\alpha(x) = \begin{cases} 1, & \text{if } x \text{ belongs to class } \alpha \\ 0, & \text{otherwise.} \end{cases}$$

We denote by $\mathcal{H}$ the set of all possible classifications into $K$ non-empty classes. The key role in this work is played by a specific type of classifications where class boundaries in space

$\mathcal{X}$ are specified by smooth[1] functions. We illustrate this below for classifications into two classes ($K = 2$).

A classification $H = (h_1(x), h_2(x))$ is specified by a *discriminant function* $F(x)$ as follows[2]:

$$h_1(x) = \begin{cases} 1, & \text{if } F(x) \geq 0 \\ 0, & \text{otherwise,} \end{cases} \quad \text{and} \quad h_2(x) = \begin{cases} 1, & \text{if } F(x) < 0 \\ 0, & \text{otherwise.} \end{cases} \tag{2.1}$$

The boundary between classes is the *discriminant surface* $F(x) = 0$. Set $\mathcal{H}$ consists of all possible classifications of space $\mathcal{X}$ into two non-empty classes. Further, we only consider classifications into two classes such that certain criteria (functionals) achieve extremal values on these classifications.

Functionals considered are differentiable functions of the non-normalized class moments, of order up to $r$, of the probability distribution function $P(x)$. The non-normalized class moments of the $l$-th order ($l = 0, \ldots, r$) are defined as

$$M_1^{(l)} = \int\limits_{\mathcal{X}} x^l h_1(x) dP(x) \quad \text{and} \quad M_2^{(l)} = \int\limits_{\mathcal{X}} x^l h_2(x) dP(x),$$

where $x^l$ denotes the scalar $\|x\|^l$ when $l$ is even and the vector $x\|x\|^{l-1}$ when $l$ is odd. It is assumed that the probability density function $\Pr(x)$ of occurrence of points $x \in \mathcal{X}$ exists, is continuous and is concentrated in a compact set $R$ of space $\mathcal{X}$, i.e., $\Pr(x) = 0,\ \forall x \notin R$.

Now, we state a theorem published in [1] that characterizes the form of smooth discriminant functions considered in this work.

**Theorem 1.** *Let the quality of a classification $H \in \mathcal{H}$ be measured by a functional of the form*

$$I\left(M_1^{(0)}, M_1^{(1)}, \ldots, M_1^{(r)}, M_2^{(0)}, M_2^{(1)}, \ldots, M_2^{(r)}\right), \tag{2.2}$$

*where $I$ is a differentiable function of the non-normalized class moments of order up to and including $r$, and the probability density* $\Pr(x)$ *is a continuous function that is zero outside a compact set $R$ of space $\mathcal{X}$. Then:*

1. *if functional (2.2) achieves an extremum on some discriminant function, the same extremum is achieved on a polynomial discriminant function of degree $r$ defined as:*

$$F(x) = f_2(x) - f_1(x) = \sum_{l=0}^{r} \left(c_2^{(l)}, x^l\right) - \sum_{l=0}^{r} \left(c_1^{(l)}, x^l\right) = \sum_{l=0}^{r} \left(c_2^{(l)} - c_1^{(l)}, x^l\right), \tag{2.3}$$

*where*

$$c_1^{(l)} = \frac{\partial I}{\partial M_1^{(l)}} \quad \text{and} \quad c_2^{(l)} = \frac{\partial I}{\partial M_2^{(l)}} \tag{2.4}$$

[1] As we will see later, class boundaries discussed in this work are specified by polynomial functions.

[2] To avoid ambiguity, points of the discriminant surface $F(x) = 0$ are always assigned to class "1".

*2. the discriminant function defined by (2.3) and (2.4) endows functional (2.2) with a stationary value.*

In Theorem 1, $c_\alpha^{(l)}$ denote scalars when $l$ is even and vectors with coordinates $\frac{\partial I}{\partial M_{\alpha,i}^{(l)}}$ when $l$ is odd, where $\alpha \in \{1,2\}$ is the class index and $M_{\alpha,i}^{(l)}$ is the $i$-th component of vector $M_\alpha^{(l)}$; $\left(c_\alpha^{(l)}, x^l\right)$ denotes multiplication of scalars $c_\alpha^{(l)}$ and $\|x\|^l$ when $l$ is even and the scalar product of vectors $c_\alpha^{(l)}$ and $x\|x\|^{l-1}$ when $l$ is odd.

We note that Theorem 1 is concerned with partitions of the compact set $R$ of space $\mathcal{X}$ rather than the entire space $\mathcal{X}$.

We also note that a functional of the form (2.2) can be constructed such that classifications minimizing it are of interest. In this case, *polynomial membership functions*

$$f_1(x) = \sum_{l=0}^{r} \left(c_1^{(l)}, x^l\right) \quad \text{and} \quad f_2(x) = \sum_{l=0}^{r} \left(c_2^{(l)}, x^l\right), \tag{2.5}$$

are regarded as measures of distance between a point and a class. On the other hand, a functional of the form (2.2) can be constructed such that classifications maximizing it are sought. Under this condition, membership functions (2.5) are regarded as measures of affinity between a point and a class. The corresponding discriminant function (2.3), in this case, has to be taken with a negative sign in definition (2.1) of characteristic functions.

A generalization of Theorem 1 to classifications into an arbitrary number of classes ($K \geq 2$), follows. We consider the problem of finding a classification minimizing a functional of the form

$$I\left(M_1^{(0)}, M_1^{(1)}, \ldots, M_1^{(r)}, \ldots, M_K^{(0)}, M_K^{(1)}, \ldots, M_K^{(r)}\right), \tag{2.6}$$

where

$$M_\alpha^{(l)} = \int_{\mathcal{X}} x^l h_\alpha(x) dP(x), \tag{2.7}$$

denotes the $l$-th ($l = 0, \ldots, r$) non-normalized moment of class $\alpha$ ($\alpha = 1, \ldots, K$).

Let $c = \left(c_1^{(0)}, c_1^{(1)}, \ldots, c_1^{(r)}, \ldots, c_K^{(0)}, c_K^{(1)}, \ldots, c_K^{(r)}\right)$ denote a vector of coefficients, where $c_\alpha^{(l)}$ denote scalars when $l$ is even and $n$-dimensional vectors when $l$ is odd. Vector $c$ specifies polynomial membership functions $f_1(x), f_2(x), \ldots, f_K(x)$, where

$$f_\alpha(x) = \sum_{l=0}^{r} \left(c_\alpha^{(l)}, x^l\right). \tag{2.8}$$

For a given vector $c$, the *polynomial classification* $H^c = (h_1^c(x), \ldots, h_K^c(x))$ is specified via membership functions (2.8) as follows:

$$h_\alpha^c(x) = \begin{cases} 1, & \text{if } f_\alpha(x) = \min\limits_{i=1,\ldots,K} f_i(x),\ \alpha = \min\limits_{i=1,\ldots,K}\{i : f_i(x) = f_\alpha(x)\} \\ 0, & \text{otherwise.} \end{cases} \tag{2.9}$$

For convenience, let $\mu(H)$ denote the vector of the non-normalized class moments under a classification $H \in \mathcal{H}$:

$$\mu(H) = \left( M_1^{(0)}, M_1^{(1)}, \ldots, M_1^{(r)}, \ldots, M_K^{(0)}, M_K^{(1)}, \ldots, M_K^{(r)} \right).$$

Functional (2.6) can be rewritten as

$$I = I\left(\mu(H)\right). \tag{2.10}$$

Two classifications $H$ and $H^*$ are *equivalent* if $\mu(H) = \mu(H^*)$.

**Theorem 2.** *Let $I(\mu(H))$ be a strictly concave functional that attains a local minimum on a classification $H^*$. Then a polynomial classification $H^c$, equivalent to $H^*$, exists for which the vector $c = \left( c_1^{(0)}, c_1^{(1)}, \ldots, c_1^{(r)}, \ldots, c_K^{(0)}, c_K^{(1)}, \ldots, c_K^{(r)} \right)$ of coefficients is determined as a supergradient[3] of the functional $I(\mu(H))$ at the point $\mu(H^*)$.*

The proof of Theorem 2 is provided in Appendix B, and rests on the following lemma, which is proved in Appendix A.

**Lemma 1.** For an arbitrary vector $c$ and an arbitrary classification $H \in \mathcal{H}$, the following inequality holds:

$$(c, \mu(H^c) - \mu(H)) \leq 0.$$

In Appendix B we also show that set $Z = \{\mu(H) : H \in \mathcal{H}\}$ of vectors of the non-normalized class moments of all possible classifications $H \in \mathcal{H}$ is bounded, closed and convex. Therefore, all local minima of a strictly concave functional (2.10) are attained on the boundary points of set $Z$. Lemma 1 states that polynomial classifications correspond to the boundary points of set $Z$. Theorem 2 specifies the form of the polynomial classifications minimizing a strictly concave functional (2.10). Similar generalization of Theorem 1 was first published in [3] where maximization of a convex functional $I\left( M_1^{(0)}, M_1^{(1)}, \ldots, M_K^{(0)}, M_K^{(1)} \right)$ was considered.

# 3 Classification Methods: Algorithms of Search for Extrema of The Criteria

We consider functionals of the form (2.10) that are functions of the first three non-normalized class moments $M_\alpha^{(0)}$, $M_\alpha^{(1)}$ and $M_\alpha^{(2)}$. These non-normalized moments allow for a clearly interpretable characterization of the classes in terms of their weights (class probabilities), locations of the class centers (class means), and tightness of the classes (class variances).

[3] A supergradient of a concave functional $I$ at a point $z^*$ is a vector $q$ satisfying the condition $I(z) - I(z^*) \leq (q, z - z^*)$ for any point $z$ in the domain of functional $I$.

From now on, we denote by $p_\alpha$ the zeroth non-normalized moment $M_\alpha^{(0)}$, i.e., the probability of class $\alpha$.

In the ensuing discussion, we denote by $\mathcal{M}_\alpha^{(l)}$ the $l$-th normalized class moments defined as

$$\mathcal{M}_\alpha^{(l)} = \frac{M_\alpha^{(l)}}{p_\alpha}.$$

Without loss of generality, we assume that for any classification $H \in \mathcal{H}$, class probabilities and variances are positive, i.e.,

$$p_\alpha > 0$$

and

$$\sigma_\alpha^2 = \left(\mathcal{M}_\alpha^{(2)} - \left(\mathcal{M}_\alpha^{(1)}\right)^2\right) > 0,$$

for $\alpha = 1, \ldots, K$.

In this section we analyze three criteria

$$I_1 = \sum_{\alpha=1}^{K} p_\alpha \sigma_\alpha^2 = \sum_{\alpha=1}^{K} p_\alpha \left(\mathcal{M}_\alpha^{(2)} - \left(\mathcal{M}_\alpha^{(1)}\right)^2\right), \tag{3.1}$$

$$I_2 = \sum_{\alpha=1}^{K} p_\alpha \sigma_\alpha = \sum_{\alpha=1}^{K} p_\alpha \sqrt{\left(\mathcal{M}_\alpha^{(2)} - \left(\mathcal{M}_\alpha^{(1)}\right)^2\right)}, \tag{3.2}$$

and

$$I_3 = \sum_{\alpha=1}^{K} p_\alpha^2 \sigma_\alpha^2 = \sum_{\alpha=1}^{K} p_\alpha^2 \left(\mathcal{M}_\alpha^{(2)} - \left(\mathcal{M}_\alpha^{(1)}\right)^2\right). \tag{3.3}$$

Intuitively and mathematically, criteria (3.1), (3.2) and (3.3) are similar. However, as will be shown later, these criteria produce different discriminant surfaces and present different views of the data.

Criterion (3.1) is that of the well known K-means method for which a minimization algorithm was first given by [14]. Criteria (3.2) and (3.3) originated in statistical sampling theory where they were proposed and studied [4, 11] for one-dimensional data.

Criterion (3.2) is known as Neyman's criterion for stratified sampling. The novelty of our work lies in the extension of this criterion to multi-dimensional data and in the construction of efficient algorithms for minimization of criteria (3.2) and (3.3).

The input data for the classification algorithms is assumed to be given in the form of a sample $X = \{x_1, x_2, \ldots, x_m\}$ of points. We denote by $\tilde{\mathcal{H}}$ the set of all possible classifications into $K$ non-empty classes constructed on the basis of the sample $X$. Given a classification $H = (h_1(x), \ldots, h_K(x))$, $H \in \tilde{\mathcal{H}}$, the vector $\tilde{\mu}(H) = \left(\tilde{p}_1, \tilde{M}_1^{(1)}, \tilde{M}_1^{(2)}, \ldots, \tilde{p}_K, \tilde{M}_K^{(1)}, \tilde{M}_K^{(2)}\right)$ of the non-normalized sample class moments is estimated over the sample $X$ as follows:

$$\begin{aligned}
\tilde{p}_\alpha &= \tfrac{1}{m}\sum_{i=1}^{m} h_\alpha(x_i) = \tfrac{m_\alpha}{m}, \\
\tilde{M}_\alpha^{(l)} &= \tfrac{1}{m}\sum_{i=1}^{m} x_i^l h_\alpha(x_i),\ l \in \{1, 2\},
\end{aligned}$$

where $m_\alpha$ is the number of points in class $\alpha$.

In general, for a given functional $I(\mu(H))$, we are interested in finding a classification $H^*$ such that

$$H^* = \arg\min_{H\in\tilde{\mathcal{H}}} I\left(\tilde{\mu}(H)\right).$$

However, the number [9]

$$S(m,K) = \frac{1}{K!}\sum_{\alpha=1}^{K}(-1)^{K-\alpha}\left(\begin{array}{c}K\\ \alpha\end{array}\right)\alpha^m$$

of distinct partitions of the sample $X$ into $K$ classes grows rapidly with $K$ and $m$. For example [9], there are $S(10,4) = 34,105$ partitions of ten objects into four classes, while there are $S(19,4) \approx 10^{10}$ partitions of nineteen objects into four classes. Thus, exhaustive enumeration of all possible classifications $H \in \tilde{\mathcal{H}}$ into $K$ classes constructed on the basis of $m$ sample points is impractical in most cases. We, therefore, resort to search for classifications that provide functional $I$ with local minima.

Functionals (3.1), (3.2) and (3.3) are differentiable[4] functions of the non-normalized class moments. We denote by $\nabla I(\mu(H))$ the gradient

$$\nabla I = \left(\frac{\partial I}{\partial M_1^{(0)}}, \frac{\partial I}{\partial M_1^{(1)}}, \ldots, \frac{\partial I}{\partial M_1^{(r)}}, \ldots, \frac{\partial I}{\partial M_K^{(0)}}, \frac{\partial I}{\partial M_K^{(1)}}, \ldots, \frac{\partial I}{\partial M_K^{(r)}}\right)$$

of a functional $I$ evaluated at a point $\mu(H)$. From Theorem 2 follows that, in cases where functional $I$ is strictly concave, the Basic Gradient Descent procedure given below is guaranteed to converge to a classification that provides functional $I$ with a local minimum. This procedure constitutes the basis for the classification algorithms developed in this section.

**Algorithm 1** (Basic Gradient Descent)**.**

```
1. Input: Initial classification $H$ (arbitrary)
2. Reclassify:
   (a) Compute vector $\tilde{\mu}(H)$ of the non-normalized sample class moments
   (b) Compute vector $c = \nabla I(\tilde{\mu}(H))$
   (c) Construct classification $H^c$ using characteristic functions $h^c_\alpha(x)$ defined by (2.9)
   (d) for $\alpha = 1, \ldots, K$
       i. if $\sum_{x\in X} h^c_\alpha(x) < b$, put into class $\alpha$, $\left(b - \sum_{x\in X} h^c_\alpha(x)\right)$ points closest (as measured by the corresponding membership function (2.8)) to this class
```

[4] If a concave functional $I$ is differentiable at a point $z^*$, then there exists a unique supergradient of $I$ at the point $z^*$, namely the gradient of functional $I$ at the point $z^*$.

$$b = \begin{cases} 1, & \text{for functionals (3.1) and (3.3)} \\ 2, & \text{for functional (3.2)} \end{cases}$$

```
3. Stop condition:
    (a) if $\tilde{\mu}(H) = \tilde{\mu}(H^c)$, go to step (4)
    (b) For functional (3.3): else if $I(\tilde{\mu}(H^c)) > I(\tilde{\mu}(H))$, go to step (4)
    (c) else, set $H \leftarrow H^c$ and return to step (2)
4. Output: Classification $H$
```

In order to obtain a deeper minimum of functional $I$, we introduce a parameter $N$ that controls the number of partitions (of the sample $X$) examined by a classification algorithm. A classification giving the smallest value of the functional is output by the algorithm. The general form of the classification algorithms proposed in this work for functionals (3.2) and (3.3) is the same as the K-means algorithm.

The difference between the three algorithms lies in the membership functions (2.8) according to which classifications are constructed in step 2c of Algorithm 1. The precise form of membership functions for the K-means criterion (3.1) is given in Section 3.1. Membership functions for criteria (3.2) and (3.3) are derived in Sections (3.2) and (3.3), respectively. Below, we give the general classification algorithm for functionals (3.1), (3.2) and (3.3).

**Algorithm 2** (Classification Algorithm for Functionals (3.1), (3.2) and (3.3))**.**

```
1. Input:
    (a) Sample $X = \{x_1, x_2, \ldots, x_m\}$ of points
    (b) Number $K$ of classes
    (c) Number $N$ of iterations
2. for $i = 1, \ldots, N$
    (a) Generate a random assignment $H_i$ of points to classes (for functionals (3.1) and
        (3.3) each class must be non-empty; for functional (3.2) each class must contain
        at least two points)
    (b) Execute gradient descent Algorithm 1 initialized with $H_i$. Denote the resulting
        classification by $H_i^*$
3. Output: $H^* = \arg \min_{\{H_i^* : i=1,\ldots,N\}} I(\tilde{\mu}(H_i^*))$
```

The total computational complexity of Algorithm 2 is $O(NtKmn)$ scalar additions and multiplications, where $t$ is the maximum number of iterations performed by the gradient descent Algorithm 1 during the $N$ iterations in Algorithm 2.

### 3.1 First Criterion: Criterion of The K-means Method

Criterion (3.1) is that of the classical K-means method. In fact, functional (3.1) is independent of the second non-normalized class moments $M_\alpha^{(2)}$, which can be shown by rewriting this functional as

$$
\begin{aligned}
I_1 = \sum_{\alpha=1}^{K} p_\alpha \sigma_\alpha^2 &= \sum_{\alpha=1}^{K} p_\alpha \left( \mathcal{M}_\alpha^{(2)} - \left(\mathcal{M}_\alpha^{(1)}\right)^2 \right) \\
&= \sum_{\alpha=1}^{K} M_\alpha^{(2)} - \sum_{\alpha=1}^{K} p_\alpha \left(\mathcal{M}_\alpha^{(1)}\right)^2 \\
&= \sum_{\alpha=1}^{K} \int\limits_{\mathcal{X}} x^2 h_\alpha^\pi(x) dP(x) - \sum_{\alpha=1}^{K} p_\alpha \left(\mathcal{M}_\alpha^{(1)}\right)^2 \\
&= \int\limits_{\mathcal{X}} x^2 dP(x) - \sum_{\alpha=1}^{K} p_\alpha \left(\mathcal{M}_\alpha^{(1)}\right)^2 , \\
&= C - \sum_{\alpha=1}^{K} p_\alpha \left(\mathcal{M}_\alpha^{(1)}\right)^2 ,
\end{aligned}
$$

where $C$ is a constant independent of a classification $H \in \mathcal{H}$. It follows that minimization of functional (3.1) corresponds to maximization of functional $I_1' = \sum\limits_{\alpha=1}^{K} p_\alpha \left(\mathcal{M}_\alpha^{(1)}\right)^2$.

We continue the discussion regarding minimization of functional (3.1) in order to give an example of construction of a classification algorithm based on the theoretical framework presented in Section 2, and to draw comparisons with the other two criteria discussed in the succeeding subsections.

**Claim 1.** Functional $I_1$ is *strictly concave.*

*Proof.* We prove the claim by showing that the $\alpha$-th functional $I_{1\alpha} = p_\alpha \sigma_\alpha^2$ in summation (3.1) is strictly concave, from which it follows that functional $I_1$ is concave.

Computing the gradient $\nabla I_{1\alpha} = \left( c_\alpha^{(0)}, c_\alpha^{(1)}, c_\alpha^{(2)} \right)$ of functional $I_{1\alpha}$ yields

$$
\begin{aligned}
c_\alpha^{(0)} &= \tfrac{\partial I_1}{\partial p_\alpha} &&= \tfrac{\left(M_\alpha^{(1)}\right)^2}{p_\alpha^2} &&= \left(\mathcal{M}_\alpha^{(1)}\right)^2 , \\
c_\alpha^{(1)} &= \tfrac{\partial I_1}{\partial M_\alpha^{(1)}} &&= -2\tfrac{M_\alpha^{(1)}}{p_\alpha} &&= -2\mathcal{M}_\alpha^{(1)}, \\
c_\alpha^{(2)} &= \tfrac{\partial I_1}{\partial M_\alpha^{(2)}} &&= 1.
\end{aligned}
\tag{3.4}
$$

Let the non-normalized class moments of class $\alpha$ under a classification $H \in \mathcal{H}$ be denoted by $\mu_\alpha(H) = \left(p_\alpha, M_\alpha^{(1)}, M_\alpha^{(2)}\right)$. For any two classifications $H \in \mathcal{H}$ and $\hat{H} \in \mathcal{H}$ we have

$$\left(\nabla I_{1\alpha}(\mu_\alpha(H)), \mu_\alpha(\hat{H}) - \mu_\alpha(H)\right) = \left(\frac{M_\alpha^{(1)}}{p_\alpha}\right)^2 (\hat{p}_\alpha - p_\alpha) - \frac{2}{p_\alpha}\left(M_\alpha^{(1)}, \hat{M}_\alpha^{(1)} - M_\alpha^{(1)}\right) + \hat{M}_\alpha^{(2)} - M_\alpha^{(2)},$$

and

$$I_{1\alpha}(\mu_\alpha(\hat{H})) - I_{1\alpha}(\mu_\alpha(H)) = \hat{M}_\alpha^{(2)} - \frac{\left(\hat{M}_\alpha^{(1)}\right)^2}{\hat{p}_\alpha} - M_\alpha^{(2)} + \frac{\left(M_\alpha^{(1)}\right)^2}{p_\alpha}.$$

By subtracting the first equation from the second and simplifying, we have

$$\begin{aligned} I_{1\alpha}(\mu_\alpha(\hat{H})) - I_{1\alpha}(\mu_\alpha(H)) - \left(\nabla I_{1\alpha}(\mu_\alpha(H)), \mu_\alpha(\hat{H}) - \mu_\alpha(H)\right) = \\ = \ & -\frac{\left(\hat{M}_\alpha^{(1)}\right)^2}{\hat{p}_\alpha} - \left(\frac{M_\alpha^{(1)}}{p_\alpha}\right)^2 \hat{p_\alpha} + \frac{2}{p_\alpha}\left(M_\alpha^{(1)}, \hat{M}_\alpha^{(1)}\right) \\ = \ & -\hat{p_\alpha}\left(\frac{M_\alpha^{(1)}}{p_\alpha} - \frac{\hat{M}_\alpha^{(1)}}{\hat{p}_\alpha}\right)^2 \\ = \ & -\hat{p_\alpha}\left(\mathcal{M}_\alpha^{(1)} - \hat{\mathcal{M}}_\alpha^{(1)}\right)^2 < 0, \ \mathcal{M}_\alpha^{(1)} \neq \hat{\mathcal{M}}_\alpha^{(1)} \end{aligned}$$

It follows that, by definition of a strictly concave function, functional $I_{1\alpha}$ is strictly concave, i.e.,

$$I_{1\alpha}(\mu_\alpha(\hat{H})) < I_{1\alpha}(\mu_\alpha(H)) + \left(\nabla I_{1\alpha}(\mu_\alpha(H)), \mu_\alpha(\hat{H}) - \mu_\alpha(H)\right), \ \mathcal{M}_\alpha^{(1)} \neq \hat{\mathcal{M}}_\alpha^{(1)}$$

Therefore, functional $I_1 = \sum_{\alpha=1}^{K} I_{1\alpha}$ is strictly concave. □

Using the gradient (3.4) for specifying membership functions (2.8) yields

$$\begin{aligned} f_\alpha(x) &= c_\alpha^{(0)} + \left(c_\alpha^{(1)}, x\right) + c_\alpha^{(2)} x^2 \\ &= \left(\mathcal{M}_\alpha^{(1)}\right)^2 - 2\left(\mathcal{M}_\alpha^{(1)}, x\right) + x^2 \\ &= \left(x - \mathcal{M}_\alpha^{(1)}\right)^2, \end{aligned} \tag{3.5}$$

where $c_\alpha^{(2)} x^2$ is a shorthand notation for scalar multiplication $\left(c_\alpha^{(2)}, x^2\right)$ introduced earlier. Note that these membership functions are independent of the second non-normalized class moments $M_\alpha^{(2)}$.

From the definitions of characteristic functions (2.9) and membership functions (3.5) follows that under a polynomial classification $H^c$, a point is assigned to a class whose mean vector is closest to that point, according to the squared Euclidean distance.

The discriminant surface formed by membership functions (3.5) between two classes $\alpha$ and $\beta$ is the linear equation

$$\begin{aligned} F(x) = f_\alpha(x) - f_\beta(x) &= \left(x - \mathcal{M}_\alpha^{(1)}\right)^2 - \left(x - \mathcal{M}_\beta^{(1)}\right)^2 \\ &= 2\left(\mathcal{M}_\beta^{(1)} - \mathcal{M}_\alpha^{(1)}, x\right) + \left(\mathcal{M}_\alpha^{(1)}\right)^2 - \left(\mathcal{M}_\beta^{(1)}\right)^2 = 0. \end{aligned} \tag{3.6}$$

In other words, the discriminant surface constructed by K-means is the hyperplane that contains the mid point $x = \frac{1}{2}\left(\mathcal{M}_\beta^{(1)} + \mathcal{M}_\alpha^{(1)}\right)$ of the line segment connecting the class means $\mathcal{M}_\alpha^{(1)}$ and $\mathcal{M}_\beta^{(1)}$, and whose norm $2\left(\mathcal{M}_\beta^{(1)} - \mathcal{M}_\alpha^{(1)}\right)$ is collinear with that line segment.

## 3.2 Second Criterion: Generalization of Neyman's Criterion for Stratified Sampling

In stratified sampling, the population is first partitioned into $K$ groups. A number of individuals is then sampled from each group, yielding the sample sought. A study of criteria for constructing stratified samples for estimating the mean value $\bar{g}$ of an unknown scalar parameter $g$ was published in [4]. Each data instance in [4] was described by a value of a known scalar parameter $x$. The statistical relationship between parameters $g$ and $x$ was assumed known. In cases of linear functional dependence between $g$ and $x$ (i.e., the correlation coefficient $\rho(g, x) = 1$), a partitioning obtained by minimizing Neyman's criterion (3.2) was shown to be optimal for performing stratified sampling. However, a constructive method for obtaining such partitioning was not described in [4].

As was stated earlier, data instances considered in this work are represented by $n$-dimensional vectors. The generalization of functional (3.2) to $n$-dimensional data is straightforward and the form of functional remains the same.

**Claim 2.** Functional $I_2$ is *strictly concave.*

*Proof.* We prove the claim by showing that the $\alpha$-th functional $I_{2\alpha} = p_\alpha \sigma_\alpha$ in summation (3.2) is strictly concave, from which it follows that functional $I_2$ is concave. First, we compute the gradient $\nabla I_{2\alpha} = \left(c_\alpha^{(0)}, c_\alpha^{(1)}, c_\alpha^{(2)}\right)$ of functional $I_{2\alpha}$:

$$\begin{aligned} c_\alpha^{(0)} &= \frac{\partial I_2}{\partial p_\alpha} &&= \frac{M_\alpha^{(2)}}{2 p_\alpha \sigma_\alpha} &&= \frac{\mathcal{M}_\alpha^{(2)}}{2\sigma_\alpha}, \\ c_\alpha^{(1)} &= \frac{\partial I_2}{\partial M_\alpha^{(1)}} &&= -\frac{M_\alpha^{(1)}}{p_\alpha \sigma_\alpha} &&= -\frac{\mathcal{M}_\alpha^{(1)}}{\sigma_\alpha}, \\ c_\alpha^{(2)} &= \frac{\partial I_2}{\partial M_\alpha^{(2)}} &&= \frac{1}{2\sigma_\alpha}. \end{aligned} \tag{3.7}$$

Let the non-normalized class moments of class $\alpha$ under a classification $H \in \mathcal{H}$ be denoted by $\mu_\alpha(H) = \left(p_\alpha, M_\alpha^{(1)}, M_\alpha^{(2)}\right)$. For any two classifications $H \in \mathcal{H}$ and $\hat{H} \in \mathcal{H}$ we have

$$\begin{aligned}
\left(\nabla I_{2\alpha}(\mu_\alpha(H)), \mu_\alpha(\hat{H}) - \mu_\alpha(H)\right) &= \left(\frac{\mathcal{M}_\alpha^{(2)}}{2\sigma_\alpha}(\hat{p}_\alpha - p_\alpha) - \frac{1}{\sigma_\alpha}\left(\mathcal{M}_\alpha^{(1)}, \hat{M}_\alpha^{(1)} - M_\alpha^{(1)}\right) + \frac{1}{2\sigma_\alpha}\left(\hat{M}_\alpha^{(2)} - M_\alpha^{(2)}\right)\right) \\
&= \frac{1}{2\sigma_\alpha}\left(\mathcal{M}_\alpha^{(2)}\hat{p}_\alpha - M_\alpha^{(2)} - 2\left(\mathcal{M}_\alpha^{(1)}, \hat{M}_\alpha^{(1)}\right) + 2p_\alpha\left(\mathcal{M}_\alpha^{(1)}\right)^2 + \hat{M}_\alpha^{(2)} - M_\alpha^{(2)}\right) \\
&= \frac{\hat{p}_\alpha}{2\sigma_\alpha}\left(\mathcal{M}_\alpha^{(2)} - 2\left(\mathcal{M}_\alpha^{(1)}, \hat{\mathcal{M}}_\alpha^{(1)}\right) + \hat{\mathcal{M}}_\alpha^{(2)}\right) - p_\alpha\sigma_\alpha \\
&= \frac{\hat{p}_\alpha}{2\sigma_\alpha}\left(\mathcal{M}_\alpha^{(2)} - \left(\mathcal{M}_\alpha^{(1)}\right)^2 + \left(\mathcal{M}_\alpha^{(1)} - \hat{\mathcal{M}}_\alpha^{(1)}\right)^2 - \left(\hat{\mathcal{M}}_\alpha^{(1)}\right)^2 + \hat{\mathcal{M}}_\alpha^{(2)}\right) - p_\alpha\sigma_\alpha \\
&= \frac{\hat{p}_\alpha}{2\sigma_\alpha}\left(\sigma_\alpha^2 + \left(\mathcal{M}_\alpha^{(1)} - \hat{\mathcal{M}}_\alpha^{(1)}\right)^2 + \hat{\sigma}_\alpha^2\right) - p_\alpha\sigma_\alpha,
\end{aligned}$$

and

$$I_{2\alpha}(\mu_\alpha(\hat{H})) - I_{2\alpha}(\mu_\alpha(H)) = \hat{p}_\alpha\hat{\sigma}_\alpha - p_\alpha\sigma_\alpha.$$

By subtracting the first equation from the second and simplifying, we obtain the following inequality

$$\begin{aligned}
&I_{2\alpha}(\mu_\alpha(\hat{H})) - I_{2\alpha}(\mu_\alpha(H)) - \left(\nabla I_{2\alpha}(\mu_\alpha(H)), \mu_\alpha(\hat{H}) - \mu_\alpha(H)\right) = \\
&= \frac{\hat{p}_\alpha}{2\sigma_\alpha}\left(2\hat{\sigma}_\alpha\sigma_\alpha - \sigma_\alpha^2 - \left(\mathcal{M}_\alpha^{(1)} - \hat{\mathcal{M}}_\alpha^{(1)}\right)^2 - \hat{\sigma}_\alpha^2\right) \\
&= -\frac{\hat{p}_\alpha}{2\sigma_\alpha}\left((\sigma_\alpha - \hat{\sigma}_\alpha)^2 + \left(\mathcal{M}_\alpha^{(1)} - \hat{\mathcal{M}}_\alpha^{(1)}\right)^2\right) < 0, \;\; \mu_\alpha(\hat{H}) \neq \mu_\alpha(H).
\end{aligned}$$

From the definition of a strictly concave function follows that functional $I_{2\alpha}$ is strictly concave. Therefore, functional $I_2 = \sum_{\alpha=1}^{K} I_{2\alpha}$ is strictly concave. □

Using the gradient (3.7) for specifying membership functions (2.8) yields

$$\begin{aligned}
f_\alpha(x) &= c_\alpha^{(0)} + \left(c_\alpha^{(1)}, x\right) + c_\alpha^{(2)}x^2 \\
&= \tfrac{\mathcal{M}_\alpha^{(2)}}{2\sigma_\alpha} - \tfrac{1}{\sigma_\alpha}\left(\mathcal{M}_\alpha^{(1)}, x\right) + \tfrac{x^2}{2\sigma_\alpha} \\
&= \tfrac{1}{2\sigma_\alpha}\left(\mathcal{M}_\alpha^{(2)} - \left(\mathcal{M}_\alpha^{(1)}\right)^2\right) + \tfrac{1}{2\sigma_\alpha}\left(\left(\mathcal{M}_\alpha^{(1)}\right)^2 - 2\left(\mathcal{M}_\alpha^{(1)}, x\right) + x^2\right)^2 \\
&= \tfrac{\sigma_\alpha}{2} + \tfrac{1}{2\sigma_\alpha}\left(x - \mathcal{M}_\alpha^{(1)}\right)^2.
\end{aligned} \tag{3.8}$$

Membership functions (3.5) differ from (3.8) in that the latter depend not only on the squared Euclidean distance between a point and the class mean $\mathcal{M}_\alpha^{(1)}$, but also on the class standard deviation $\sigma_\alpha$.

The discriminant surface specified by membership functions (3.8) between two classes $\alpha$ and $\beta$ is the quadratic equation

$$\begin{aligned} F(x) = f_\alpha(x) - f_\beta(x) &= \tfrac{\sigma_\alpha}{2} + \tfrac{1}{2\sigma_\alpha}\left(x - \mathcal{M}_\alpha^{(1)}\right)^2 - \tfrac{\sigma_\beta}{2} - \tfrac{1}{2\sigma_\beta}\left(x - \mathcal{M}_\beta^{(1)}\right)^2 \\ &= \left(\sigma_\beta - \sigma_\alpha\right)x^2 + 2\left(\sigma_\alpha \mathcal{M}_\beta^{(1)} - \sigma_\beta \mathcal{M}_\alpha^{(1)}, x\right) + \\ &\quad + \sigma_\beta\left(\mathcal{M}_\alpha^{(1)}\right)^2 - \sigma_\alpha\left(\mathcal{M}_\beta^{(1)}\right)^2 + \sigma_\alpha\sigma_\beta\left(\sigma_\alpha - \sigma_\beta\right) = 0. \end{aligned} \tag{3.9}$$

In the event when class variances are equal ($\sigma_\alpha^2 = \sigma_\beta^2$), discriminant surface (3.9) coincides with the K-means discriminant surface (3.6).

## 3.3 Third Criterion: Generalization of The Method of Selecting Representative Types

The model of stratified sampling used in [4] and described in Section 3.2 was also used in [11] for studying the problem of selecting representative types from a population. This problem differs from stratified sampling in that only a single instance is sampled from each of the $K$ groups in the population. According to [11], partitioning the population so as to minimize criterion (3.3) yields an optimal grouping for selecting representative types in cases of linear functional dependence between scalar parameters $g$ and $x$.

We prove that functional (3.3) is non-convex. It follows that a situation is possible where the value of functional (3.3) will increase as a result of constructing a new classification on some iteration (step 2d) of Algorithm 1. In order to avoid such situations, the value of functional (3.3) has to be explicitly checked (step 3b of Algorithm 1) during gradient descent. We note that this check was not necessary for functionals (3.1) and (3.2) due to their strict concavity.

**Claim 3.** Functional $I_3$ is *non-convex.*

*Proof.* We prove the claim by showing that each functional $I_{3\alpha} = p_\alpha^2 \sigma_\alpha^2$ in summation (3.3) is non-convex, from which it follows that functional $I_3$ is non-convex. First, we compute the gradient $\nabla I_{3\alpha} = \left(c_\alpha^{(0)}, c_\alpha^{(1)}, c_\alpha^{(2)}\right)$ of functional $I_{3\alpha}$:

$$\begin{aligned} c_\alpha^{(0)} &= \tfrac{\partial I_3}{\partial p_\alpha} &&= M_\alpha^{(2)}, \\ c_\alpha^{(1)} &= \tfrac{\partial I_3}{\partial M_\alpha^{(1)}} &&= -2M_\alpha^{(1)}, \\ c_\alpha^{(2)} &= \tfrac{\partial I_3}{\partial M_\alpha^{(2)}} &&= p_\alpha. \end{aligned} \tag{3.10}$$

The Hessian matrix

$$\nabla^2 I_{3\alpha} = \begin{pmatrix} 0 & 0 & 1 \\ 0 & -2 & 0 \\ 1 & 0 & 0 \end{pmatrix}$$

of functional $I_{3\alpha}$ possesses a positive principal minor $M_{1,2,3} = 2$ of odd order and a negative principal minor $M_2 = -2$ of odd order. It follows that the Hessian matrix $\nabla^2 I_{3\alpha}$ is indefinite, which implies that functional $I_{3\alpha}$ is non-convex. Therefore, functional $I_3 = \sum\limits_{\alpha=1}^{K} I_{3\alpha}$ is non-convex. □

Using the gradient (3.10) for specifying membership functions (2.8) yields

$$\begin{aligned} f_\alpha(x) &= c_\alpha^{(0)} + \left(c_\alpha^{(1)}, x\right) + c_\alpha^{(2)} x^2 \\ &= M_\alpha^{(2)} - 2\left(M_\alpha^{(1)}, x\right) + p_\alpha x^2 \\ &= p_\alpha \left(\mathcal{M}_\alpha^{(2)} - \left(\mathcal{M}_\alpha^{(1)}\right)^2\right) + p_\alpha \left(\left(\mathcal{M}_\alpha^{(1)}\right)^2 - 2\left(\mathcal{M}_\alpha^{(1)}, x\right) + x^2\right) \\ &= p_\alpha \sigma_\alpha^2 + p_\alpha \left(x - \mathcal{M}_\alpha^{(1)}\right)^2 . \end{aligned} \tag{3.11}$$

The discriminant surface specified by membership functions (3.11) between two classes $\alpha$ and $\beta$ is the quadratic equation

$$\begin{aligned} F(x) = f_\alpha(x) - f_\beta(x) &= p_\alpha \sigma_\alpha^2 + p_\alpha \left(x - \mathcal{M}_\alpha^{(1)}\right)^2 - p_\beta \sigma_\beta^2 - p_\beta \left(x - \mathcal{M}_\beta^{(1)}\right)^2 \\ &= (p_\alpha - p_\beta) x^2 + 2\left(p_\beta \mathcal{M}_\beta^{(1)} - p_\alpha \mathcal{M}_\alpha^{(1)}, x\right) + \\ &\quad + p_\alpha \left(\mathcal{M}_\alpha^{(1)}\right)^2 - p_\beta \left(\mathcal{M}_\beta^{(1)}\right)^2 + p_\alpha \sigma_\alpha^2 - p_\beta \sigma_\beta^2 = 0, \end{aligned} \tag{3.12}$$

which coincides with the K-means discriminant surface (3.6) when class probabilities and variances are equal, i.e., $p_\alpha = p_\beta$ and $\sigma_\alpha^2 = \sigma_\beta^2$.

# 4 Experimental Results

## 4.1 Comparative Study of The Algorithms on Synthetic Data

In this section we present classification results produced by Algorithms (3.5), (3.8) and (3.11) on a synthetic dataset. The dataset shown in Figure 1 consisted of three classes, each generated by a Gaussian distribution. The data generator parameters were:

- Class probabilities: $p_1 = 0.4,\ p_2 = 0.2,\ p_3 = 0.4$,
- Class means: $\mu_1 = (0, 0),\ \mu_2 = (12, 6),\ \mu_3 = (12, -6)$,
- Class covariance matrices: $\Sigma_1 = 4.5I,\ \Sigma_{\{2,3\}} = 2I$.

The same set of $N = 50$ randomly generated initial assignments of points to classes was used by each algorithm. Classifications yielding the smallest values of criteria (3.1), (3.2) and (3.3) are shown in Figures 2a, 2b and 2c, respectively. This experimental result demonstrates

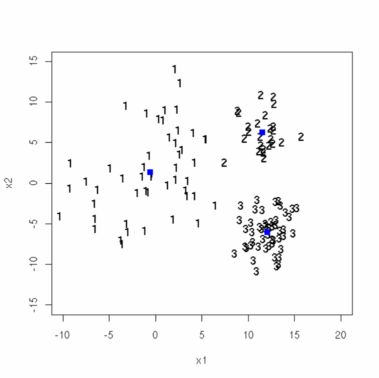
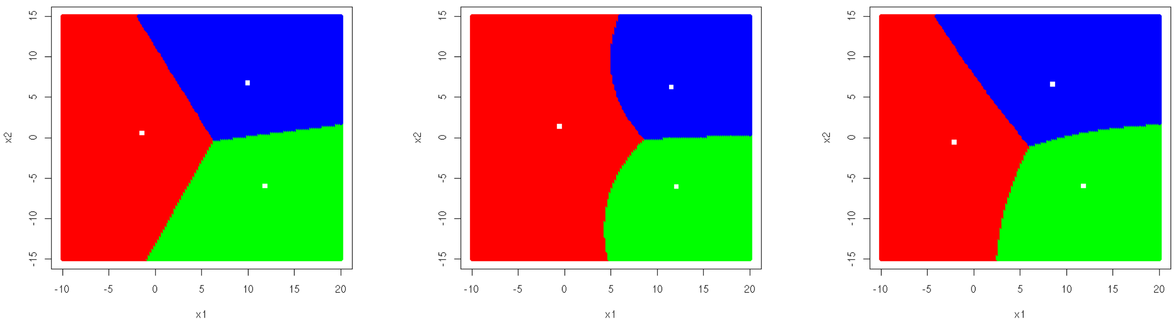


Figure 1: Randomly generated dataset. Integer labels indicate the Gaussian distribution that generated the corresponding point. Sample class means are indicated by blue squares.



(a) Criterion $I_1$ (K-means)

(b) Criterion $I_2$

(c) Criterion $I_3$

Figure 2: Classification results obtained on the dataset shown in Figure 1. Class means are indicated by white squares.

the differences between discriminant surfaces specified by membership functions (3.5), (3.8) and (3.11).

As can be seen from Figure 2a, membership functions (3.5) produced linear discriminant surfaces regardless of class probabilities and variances. Figure 2b, on the other hand, shows that due to larger variance of the red class, membership functions (3.8) produced quadratic discriminant surfaces between the red class and each of the other two classes. Discriminant surface between the blue and the green classes remained linear due to variances of these classes being equal.

The effect class probabilities have on the discriminant surfaces produced by membership functions (3.11) is demonstrated in Figure 2c. Under the classification shown in Figure 2c, the red and the blue classes had nearly equal probabilities $p_r = 0.28$ and $p_b = 0.3$, respectively. Therefore, the discriminant surface between the red and the blue classes approached a linear surface. The probability ($p_g = 0.42$) of the green class was larger than probabilities of each of the other two classes, which resulted in the more pronounced quadratic discriminant surfaces.

Next, we evaluated how well did each of the algorithms reproduce the generator class probabilities, variances and means. We denote by $\mu$ the expected generator mean

$$\mu = \sum_{i=1}^{3} p_i \mu_i,$$

and by $\bar{d}^2$ the average squared Euclidean distance between the generator class means and the expected mean

$$\bar{d}^2 = \sum_{i=1}^{3} p_i (\mu_i - \mu)^2.$$

Table 1 shows that the generator class probabilities, variances and means were best reproduced by Algorithm (3.8), followed by Algorithms (3.5) and (3.11), in that order.

| | $\frac{\lvert p_r - p_1 \rvert}{p_1}$ | $\frac{\lvert p_b - p_2 \rvert}{p_2}$ | $\frac{\lvert p_g - p_3 \rvert}{p_3}$ | max |
|---|---|---|---|---|
| $I_1$ | 0.16 | 0.24 | 0.04 | 0.24 |
| $I_2$ | 0 | 0 | 0 | 0 |
| $I_3$ | 0.30 | 0.52 | 0.04 | 0.52 |

| | $\frac{\lvert \sigma_r^2 - \sigma_1^2 \rvert}{\sigma_1^2}$ | $\frac{\lvert \sigma_b^2 - \sigma_2^2 \rvert}{\sigma_2^2}$ | $\frac{\lvert \sigma_g^2 - \sigma_3^2 \rvert}{\sigma_3^2}$ | max |
|---|---|---|---|---|
| $I_1$ | 0.13 | 1.45 | 0.12 | 1.45 |
| $I_2$ | 0.07 | 0.04 | 0.08 | 0.08 |
| $I_3$ | 0.24 | 2.38 | 0.12 | 2.38 |

| | $\frac{\left(\mathcal{M}_r^{(1)} - \mu_1\right)^2}{\bar{d}^2}$ | $\frac{\left(\mathcal{M}_b^{(1)} - \mu_2\right)^2}{\bar{d}^2}$ | $\frac{\left(\mathcal{M}_g^{(1)} - \mu_3\right)^2}{\bar{d}^2}$ | max |
|---|---|---|---|---|
| $I_1$ | 0.05 | 0.08 | 0.0006 | 0.08 |
| $I_2$ | 0.03 | 0.006 | 0.0003 | 0.03 |
| $I_3$ | 0.09 | 0.24 | 0.0006 | 0.24 |

Table 1: Comparison of Algorithms (3.5), (3.8) and (3.11) on how well they reproduce the generator class probabilities, variances and means. $p_r$, $p_b$, $p_g$, $\sigma_r^2$, $\sigma_b^2$, $\sigma_g^2$, $\mathcal{M}_r^{(1)}$, $\mathcal{M}_b^{(1)}$ and $\mathcal{M}_g^{(1)}$ are the probabilities, means and variances of the red, blue and green classes, respectively.

## 4.2 Study of The Algorithms on Real Data: Classification of Mutual Funds by Style of Management Based on Time Series of Daily Returns

A mutual fund's management style is reflected in continuous adjustments of the fund's portfolio performed by the manager in order to balance risk and profit while, at the same time, keeping with the stated investment objectives of the fund. The information regarding the fund's management style is of great value to investors and other fund managers, but is generally not available to the public. What makes the style analysis more challenging is that, according to the current U.S. regulations, mutual funds are required to report the

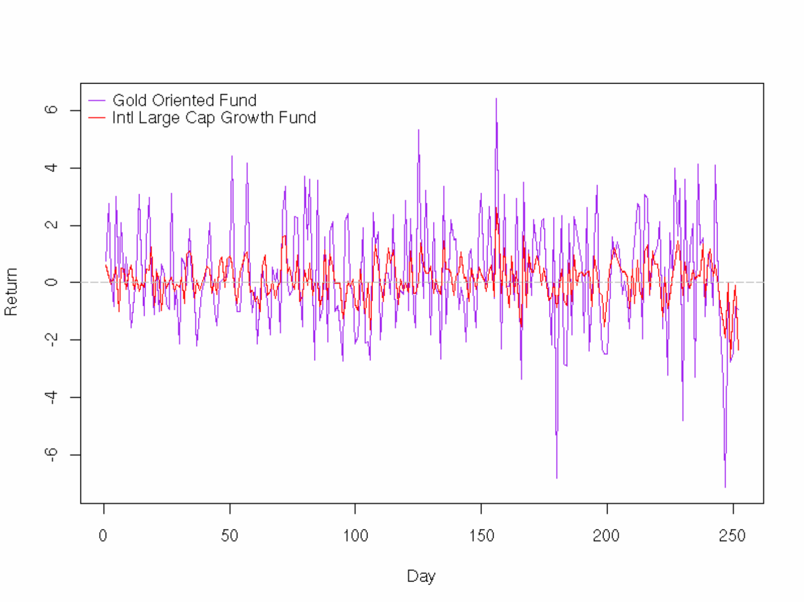


Figure 3: Time series of daily returns of two mutual funds (May 2005 – May 2006).

composition of their portfolios only four times per year (quarterly). However, the return on a fund's portfolio is declared daily.

The return of an individual mutual fund is generally thought of as a stochastic process. However, we hypothesize that funds with similar dynamics of portfolio adjustments performed by the managers would tend to have similar patterns of behavior of the returns.

In this section, we present preliminary results of experiments that test whether mutual funds can be grouped based on time series of their returns such that funds within a group reflect similar portfolio structures.

The dataset used in the experiments contained time series of daily returns of 6671 mutual funds spanning the period from May 2005 until May 2006. Return $r_i^t$ of fund $i$ at day $t$ is determined as

$$r_i^t = \frac{p_i^t - p_i^{t-1}}{p_i^{t-1}},$$

where $p_i^t$ is the closing share price of fund $i$ at day $t$. Sample graphs of returns of two funds are shown in Figure 3, which demonstrates varying dynamics of returns.

Initial experiments performed directly on time series of daily returns did not give robust results. Classes were not well-separated from each other (Euclidean distances between class means were comparable to class standard deviations). Through further investigation, we have learned that returns of a large portion (42%) of mutual funds in our dataset were highly correlated ($|\rho| \geq 0.8$) with the market's return represented by Standard & Poor's 500 (SP500) index. The corresponding histogram of correlation coefficients is shown in Figure 4.

The market trend had to be removed from the time series of daily returns of mutual

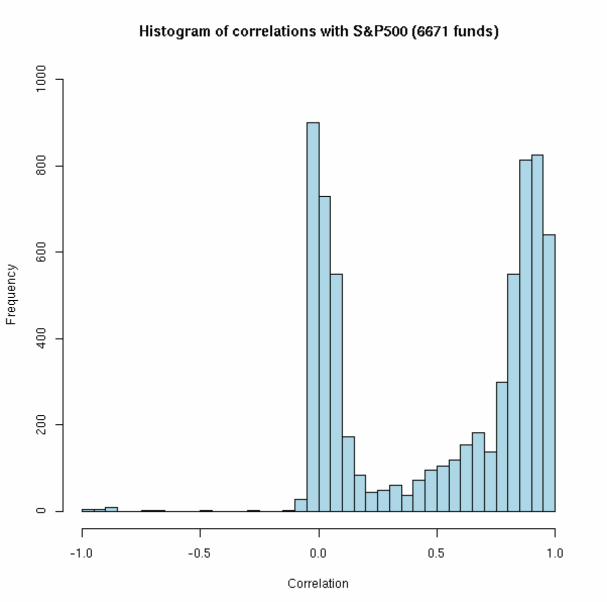


Figure 4: Histogram of coefficients of correlation between returns of the mutual funds and the SP500 index.

funds in order to reveal the specifics of management styles of the funds. We addressed this problem by constructing for each fund $i$ a least squares linear regression model

$$r_i = a_i s + b_i + \varepsilon_i, \tag{4.1}$$

where $s$ is the vector of returns of the market. The residual vector $\varepsilon_i$ reflects the behavior of the fund's return that is not explained by the market trend, and contains a more refined representation of the fund's management style.

We constructed classifications into $K = 20$ classes by Algorithms (3.5), (3.8) and (3.11). The same set of $N = 100$ randomly generated initial classifications was used by each algorithm. Classifications

$$H^1 = \left(h_1^1(x), \ldots, h_{20}^1(x)\right),$$
$$H^2 = \left(h_1^2(x), \ldots, h_{20}^2(x)\right),$$

and

$$H^3 = \left(h_1^3(x), \ldots, h_{20}^3(x)\right)$$

yielding the smallest values of criteria (3.1), (3.2) and (3.3), respectively, were obtained. All three classifications were different.

An intuitive interpretation of functionals (3.1), (3.2) and (3.3) suggests that classifications produced according to these criteria may disagree on class outlier points. Therefore, we combined classifications $H^1$, $H^2$ and $H^3$ in order to extract consistent cores (groups) of mutual funds.

Given a triplet $(\alpha, \beta, \gamma)$ of classes, where each class comes from a different classification, the core $r(\alpha, \beta, \gamma)$ of funds formed by this triplet is the maximal set of funds common to all three classes. In other words, the core $r(\alpha, \beta, \gamma)$ is the intersection of classes $\alpha$ ,$\beta$ and $\gamma$:

$$r(\alpha, \beta, \gamma) = \{x \in X : h^1_\alpha(x) h^2_\beta(x) h^3_\gamma(x) = 1\}.$$

Given three classifications $H^1$, $H^2$ and $H^3$, we are interested in extracting a set of cores of mutual funds, such that the total number of funds contained in these cores is maximized. For two classifications, the extraction of optimal cores corresponds to the well-known optimal assignment problem for which a polynomial-time algorithm exists. However, the problem of extracting optimal cores based on three classifications is an open problem that constitutes one of our future research directions. In this work, we applied a greedy algorithm for extracting cores of mutual funds.

The total of fifteen cores were extracted that together contained 27% of the total number of mutual funds in the dataset. Sizes of the cores are given in Table 2.

| Core | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Size | 494 | 408 | 369 | 256 | 102 | 64 | 52 | 31 | 25 | 16 | 8 | 4 | 3 | 1 | 1 | 1834 |

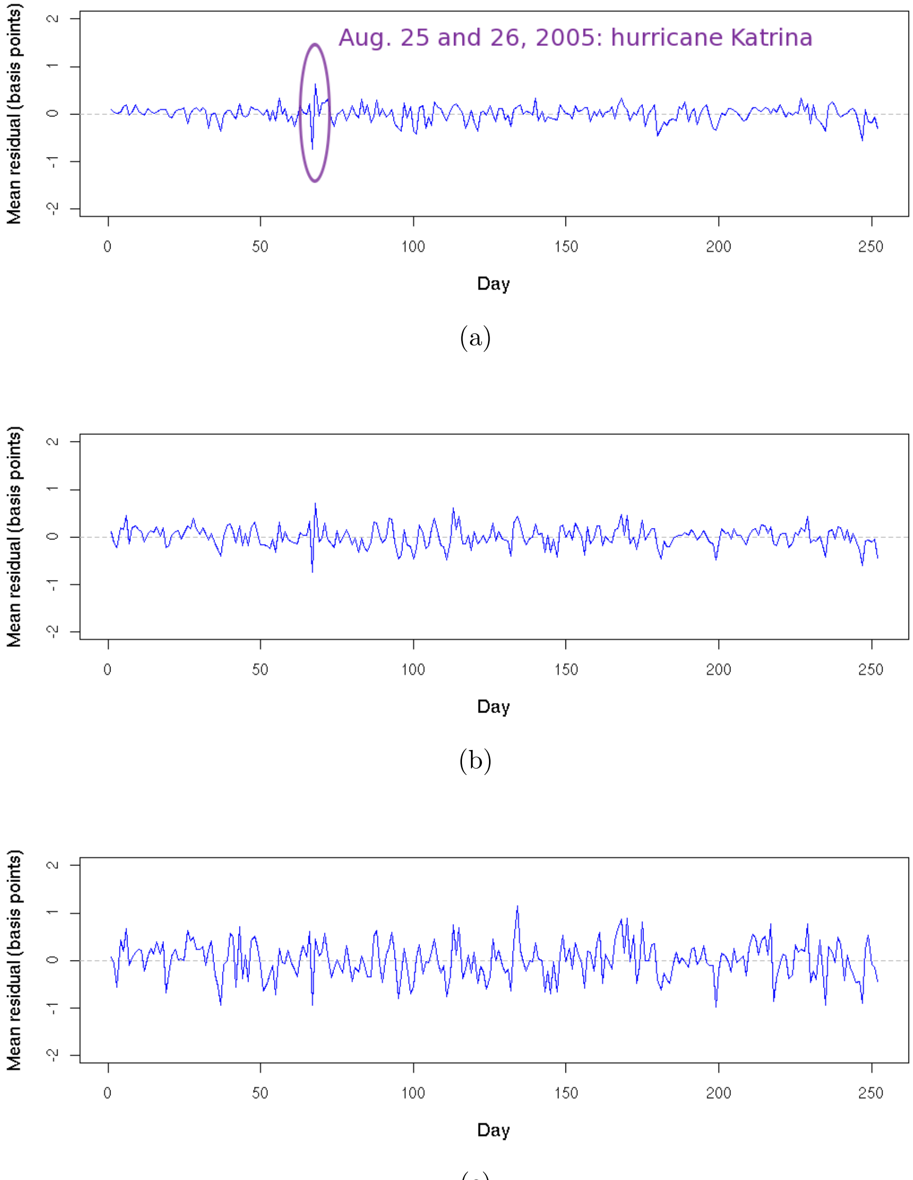
Table 2: Sizes of the cores of mutual funds

The cores were interpreted as the most robust and objectively consistent component in the existing portfolio- and prospectus-based classifications of mutual funds. Portfolio-based classification of mutual funds is obtained on the basis of expert financial analysis of the funds' portfolios. Portfolio-based classification was provided by Lipper, a financial analytics company, for diversified funds, i.e., those funds that invest across economic sectors (e.g. Large Cap Growth).

Prospectus-based classification of mutual funds is provided by the fund managers. We used prospectus-based classification for non-diversified (e.g. Real Estate) funds, since portfolio-based classification for such funds was not available.

Figures 5, 6, 7, 8 and 9 show the mean vectors of residuals $\varepsilon_i$ obtained by model (4.1) for mutual funds comprising the corresponding cores. On these figures, each core is annotated by either portfolio- or prospectus-based classification of the funds contained in the core. Homogeneity of the cores with respect to expert classification of mutual funds within the cores demonstrates the plausibility of the proposed approach to grouping funds based on time series of daily returns such that funds within a group have similar portfolio structures.

It is interesting to note the evident downward spike at time point 68 followed by an upward surge at time point 69 that can be seen clearly on Figures 5a and 5b. These time points correspond to August 25 and 26, 2005 — the midst of hurricane Katrina, which made its initial landfall in the U.S. on August 23, 2005.

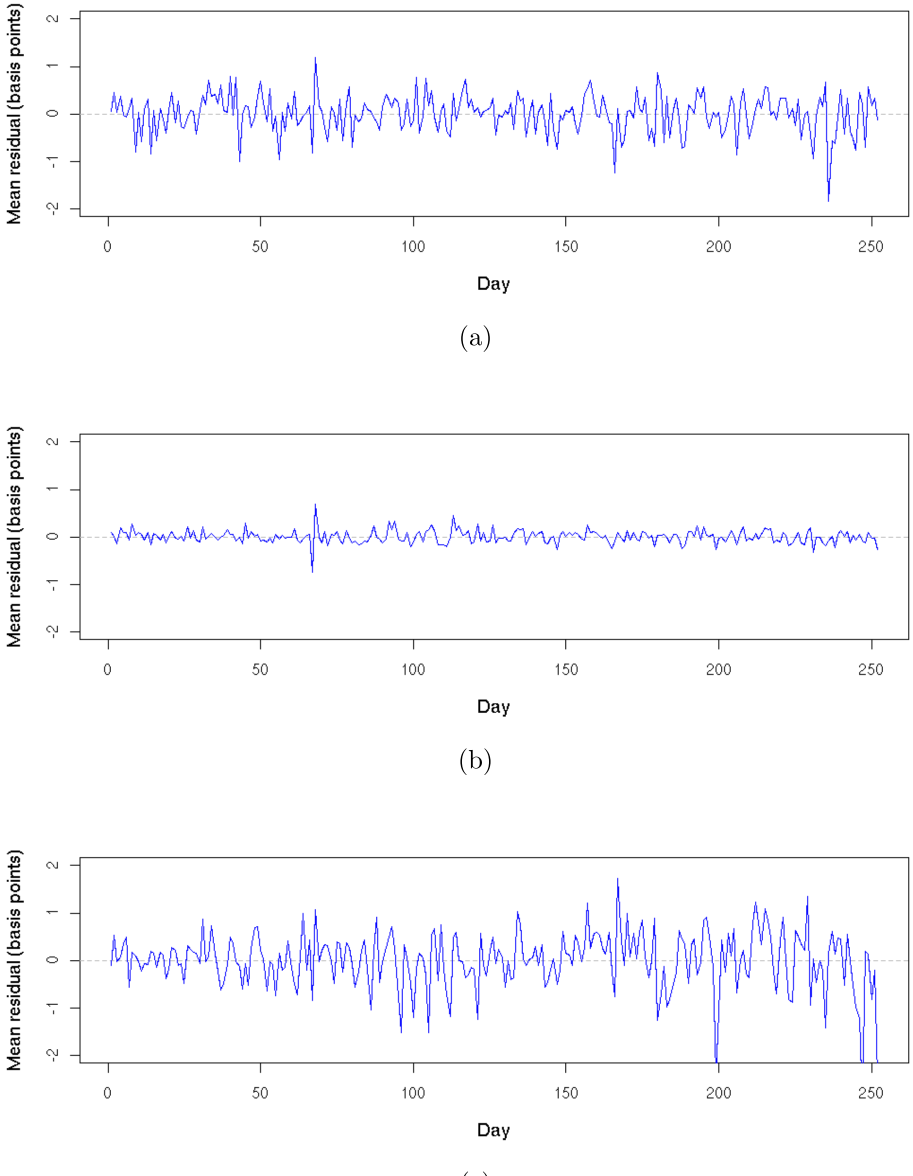


(a)

(b)

(c)

Figure 5: The extracted consistent groups of mutual funds (mean vectors are shown): (a) Large Cap (Core, Value), Multi Cap (Core, Value); (b) Mid Cap (Core, Value); (c) Small Cap (Core, Value).

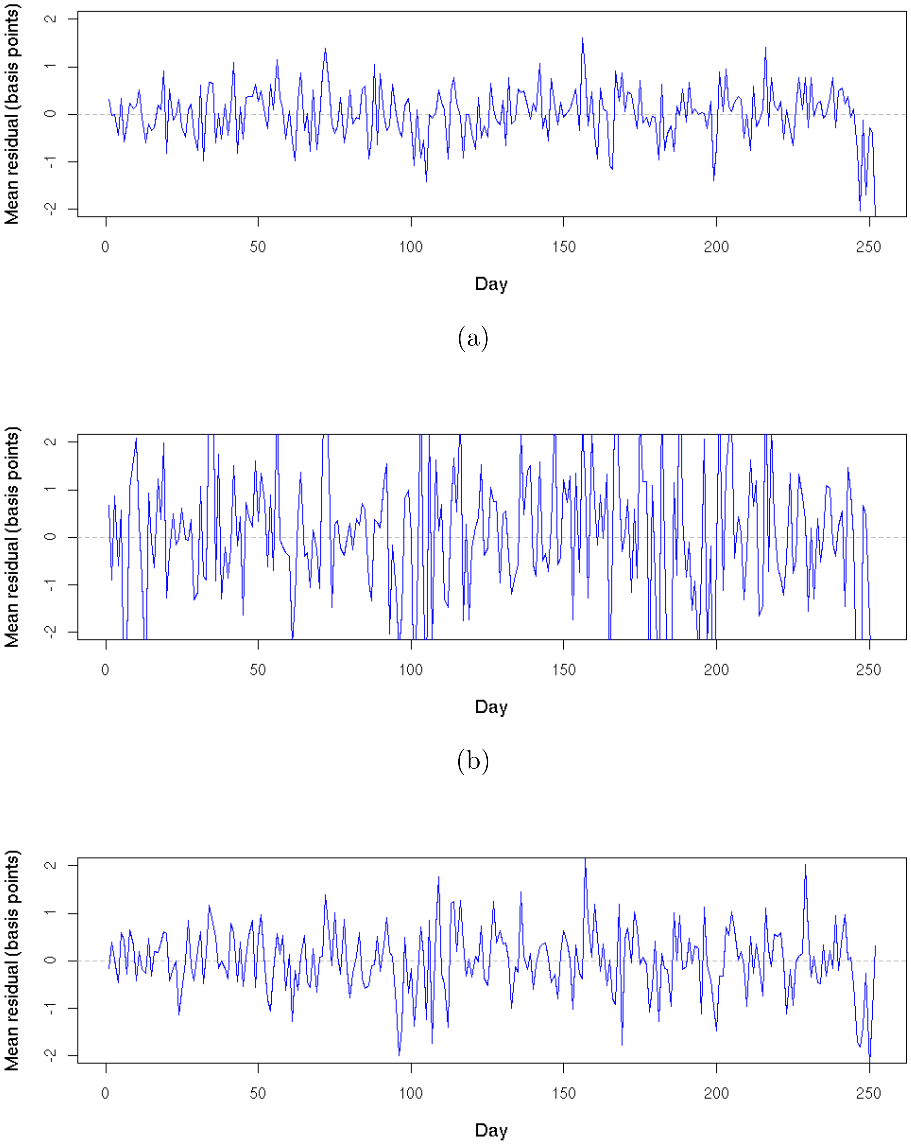


(a)

(b)

(c)

Figure 6: The extracted consistent groups of mutual funds (mean vectors are shown): (a) Multi Cap Growth; (b) Large Cap Growth, Multi Cap Core; (c) Mid Cap Core, Small Cap Core.

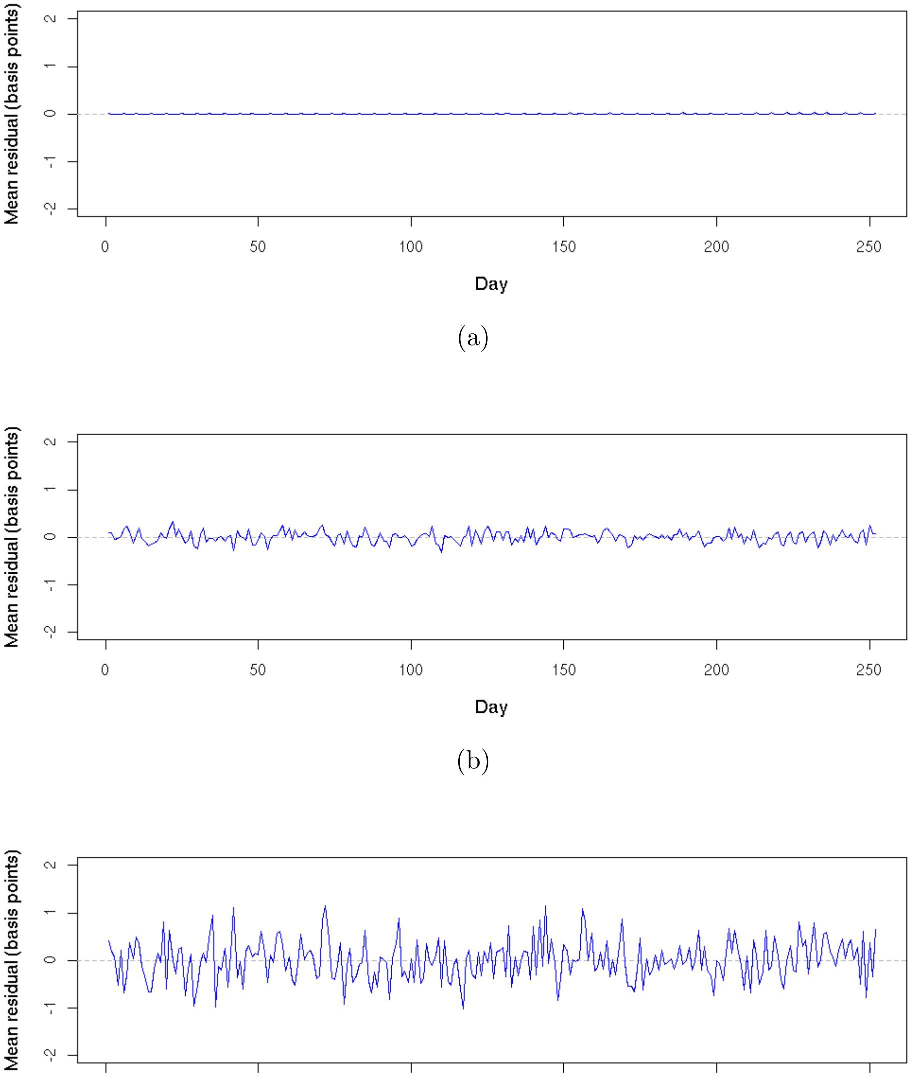


(a)

(b)

(c)

Figure 7: The extracted consistent groups of mutual funds (mean vectors are shown): (a) International Large, Small/Mid, Multi Cap (Core, Growth, Value), European Region; (b) Emerging Markets; (c) Multi Cap Core (Domestic, International, Global).

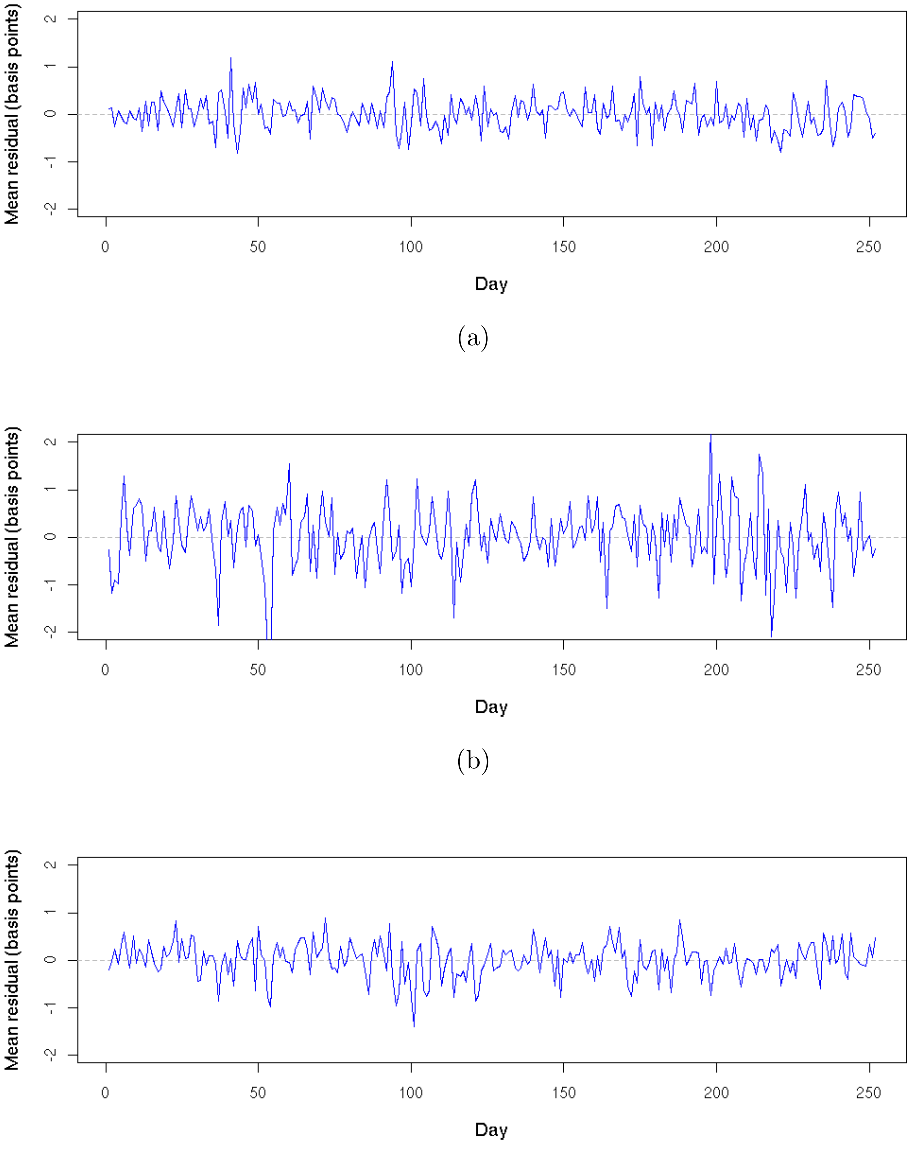


(a)

(b)

(c)

Figure 8: The extracted consistent groups of mutual funds (mean vectors are shown): (a) Money Markets; (b) Municipal Debt; (c) International Income, Global Income.



(a)

(b)

(c)

Figure 9: The extracted consistent groups of mutual funds (mean vectors are shown): (a) Healthcare/Biotechnology; (b) Real Estate; (c) Utility.

# 5 Conclusion and Future Research Directions

In this work, we have generalized functional (3.2) to multi-dimensional data and proved convexity properties of functionals (3.2) and (3.3). First algorithms that work on $n$-dimensional data for functionals (3.2) and (3.3) were developed.

Experimental results obtained on synthetic data, demonstrated the differences in the forms of discriminant surfaces constructed by membership functions (3.5), (3.8) and (3.11) under functionals (3.1), (3.2) and (3.3), respectively.

Preliminary experimental results obtained on real-world time series of daily returns of 6671 mutual funds demonstrated the plausibility of the presented approach to classification of mutual funds by style of management. In the experiments conducted, Algorithms (3.5), (3.8) and (3.11) produced consistent classifications of mutual funds. This suggests the usefulness of applying all three algorithms simultaneously for extracting consistent cores of mutual funds with similar portfolio structures.

Our future research work is comprised of three main directions all of which are related to analysis of time series data. We intend to further study the methods of simultaneous analysis of classifications produced according to functionals (3.1), (3.2) and (3.3).

In order to facilitate expert validation, we plan to develop visualization methods for expert analysis of classification results on time series data, particularly in relation to mutual funds.

We also intend to study applications of the classification methods presented here, on time series data in the area of cyber security for intrusion detection in wireless networks. This research will be conducted in collaboration with Dr. William Pottenger.

**APPENDIX**

# A Proof of Lemma 1

The proof of Lemma 1 follows from the method (2.9) of constructing polynomial classifications. Suppose, $H = (h_1(x), \dots, h_K(x))\,, H \in \mathcal{H}$, is an arbitrary classification whose vector of the non-normalized class moments is $\mu(H) = \left(M_1^{(0)}, M_1^{(1)}, \dots, M_1^{(r)}, \dots, M_K^{(0)}, M_K^{(1)}, \dots, M_K^{(r)}\right)$. Then, for an arbitrary vector $c = \left(c_1^{(0)}, c_1^{(1)}, \dots, c_1^{(r)}, \dots, c_K^{(0)}, c_K^{(1)}, \dots, c_K^{(r)}\right)$ of coefficients and the corresponding polynomial classification $H^c = (h_1^c(x), \dots, h_K^c(x))$, the scalar product $(c, \mu(H^c) - \mu(H))$ is

$$(c, \mu(H^c) - \mu(H)) = \sum_{\alpha=1}^{K}\sum_{l=0}^{r} \left(c_\alpha^{(l)}, M_\alpha^{(l)c} - M_\alpha^{(l)}\right),$$

where $\mu(H^c) = \left(M_1^{(0)c}, M_1^{(1)c}, \dots, M_1^{(r)c}, \dots, M_K^{(0)c}, M_K^{(1)c}, \dots, M_K^{(r)c}\right)$ is the vector of the non-normalized class moments under polynomial classification $H^c$.

By definition (2.7) of the non-normalized class moments, we obtain

$$\begin{aligned}
(c, \mu(H^c) - \mu(H)) &= \sum_{\alpha=1}^{K}\sum_{l=0}^{r} \left(c_\alpha^{(l)}, \int\limits_{\mathcal{X}} x^l h_\alpha^c(x) dP(x) - \int\limits_{\mathcal{X}} x^l h_\alpha(x) dP(x)\right) \\
&= \sum_{\alpha=1}^{K} \left[\int\limits_{\mathcal{X}} \sum_{l=0}^{r} \left(c_\alpha^{(l)}, x^l\right) h_\alpha^c(x) dP(x) - \int\limits_{\mathcal{X}} \sum_{l=0}^{r} \left(c_\alpha^{(l)}, x^l\right) h_\alpha(x) dP(x)\right] \\
&= \sum_{\alpha=1}^{K} \left[\int\limits_{\mathcal{X}} f_\alpha(x) h_\alpha^c(x) dP(x) - \int\limits_{\mathcal{X}} f_\alpha(x) h_\alpha(x) dP(x)\right] \\
&= \int\limits_{\mathcal{X}} \left[\sum_{\alpha=1}^{K} f_\alpha(x) h_\alpha^c(x) - \sum_{\alpha=1}^{K} f_\alpha(x) h_\alpha(x)\right] dP(x).
\end{aligned}$$

It follows from definition (2.9) of characteristic functions $h_\alpha^c(x)$ that

$$(c, \mu(H^c) - \mu(H)) \le 0, \ \forall H \in \mathcal{H}.$$

□

# B Proof of Theorem 2

Before proceeding to the proof of Theorem 2, we show that set $Z = \{\mu(H) : H \in \mathcal{H}\}$ of vectors of the non-normalized class moments is bounded, closed and convex. A point $\mu(H) = \left(M_1^{(0)}, M_1^{(1)}, \dots, M_1^{(r)}, \dots, M_K^{(0)}, M_K^{(1)}, \dots, M_K^{(r)}\right)$ belongs to set $Z$ if and only if the following equations are satisfied:

$$\sum_{\alpha=1}^{K} M_\alpha^{(l)} = \int\limits_{\mathcal{X}} x^l dP(x), \; l = 0, \dots, r. \tag{B.1}$$

Set $Z$ is bounded since the probability density function $P(x)$ is zero outside of the bounded region $R$.

Equations (B.1) imply that set $Z$ is closed, because its complement $\bar{Z}$ is defined by strict inequalities and is therefore open.

Convexity of set $Z$ follows from the fact that for any two points $\mu(H), \mu(\hat{H}) \in Z$, and any $\epsilon \in [0, 1]$, the point $\mu^\epsilon = (1-\epsilon)\mu(H) + \epsilon\mu(\hat{H})$ also lies in set $Z$, i.e.,

$$\sum_{\alpha=1}^{K} \left[(1-\epsilon)M_\alpha^{(l)} - \epsilon\hat{M}_\alpha^{(l)}\right] = (1-\epsilon)\int\limits_{\mathcal{X}} x^l dP(x) + \epsilon\int\limits_{\mathcal{X}} x^l dP(x) = \int\limits_{\mathcal{X}} x^l dP(x),$$

where $M_\alpha^{(l)}$ and $\hat{M}_\alpha^{(l)}$ are the non-normalized class moments under classifications $H$ and $\hat{H}$, respectively.

It follows that all local minima of a strictly concave functional $I(\mu(H))$ are attained on the boundary points of set $Z$. Lemma 1 establishes the fact that the boundary points of set $Z$ correspond to polynomial classifications. Now, we prove Theorem 2, which specifies the form of polynomial classifications minimizing the strictly concave functional $I$.

First, we prove that local extremality of a strictly concave functional $I(\mu(H))$ on a classification $H^*$ implies that $\mu(H^*) = \mu(H^c)$, where $H^c$ is the polynomial classification specified using vector $c$ determined as a supergradient of functional $I$ at the point $\mu(H^*)$.

Suppose that classifications $H^*$ and $H^c$ are not equivalent, i.e., $\mu(H^*) \neq \mu(H^c)$. Then, we can construct a point $\mu^\epsilon = (1-\epsilon)\mu(H^*) + \epsilon\mu(H^c), \; \epsilon \in (0, 1), \; \mu^\epsilon \in Z$.

By concavity of functional $I$, the following inequality holds:

$$I(\mu^\epsilon) \leq I(\mu(H^*)) + (c, \mu^\epsilon - \mu(H^*)) \,.$$

Due to strict concavity of functional $I$, the equality is attained if and only if $\mu^\epsilon = \mu(H^*)$, which contradicts the assumption. Therefore, it follows that

$$I(\mu^\epsilon) - I(\mu(H^*)) < \epsilon\,(c, \mu(H^c) - \mu(H^*)) \,,$$

and from Lemma 1 follows that

$$\epsilon\,(c, \mu(H^c) - \mu(H^*)) \leq 0.$$

Thus,

$$I(\mu^{\epsilon}) - I(\mu(H^*)) < 0,$$

which, given that $\epsilon$ was chosen arbitrarily, contradicts local extremality of $H^*$.

We complete the proof by showing that the existence of a polynomial classification $H^c$ equivalent to a classification $H^*$ that provides the functional $I$ with a local minimum implies that the vector $c$ is a supergradient of functional $I$ at the point $\mu(H^*)$.

Suppose that polynomial classifications $H^c$ and $H^k$ are not equivalent, i.e., $\mu(H^c) \neq \mu(H^k)$, for any vector $k$ determined as a supergradient of functional $I$ at the point $\mu(H^*)$. Then, we can form a point $\mu^{\epsilon} = (1-\epsilon)\mu(H^c) + \epsilon\mu(H^k),\ \epsilon \in (0,1),\ \mu^{\epsilon} \in Z$.

From concavity of functional $I$ and equivalence of classifications $H^*$ and $H^c$ follows that

$$I(\mu^{\epsilon}) \leq I(\mu(H^c)) + (k, \mu^{\epsilon} - \mu(H^c)),$$

where, due to strict concavity of functional $I$, the equality is attained if and only if $\mu^{\epsilon} = \mu(H^c)$, which contradicts the assumption.

Therefore, it follows that

$$I(\mu^{\epsilon}) - I(\mu(H^c)) < \epsilon \left(k, \mu(H^k) - \mu(H^c)\right),$$

and from Lemma 1 follows that

$$\epsilon \left(k, \mu(H^k) - \mu(H^c)\right) \leq 0.$$

Thus,

$$I(\mu^{\epsilon}) - I(\mu(H^c)) < 0,$$

which, given that $\epsilon$ was chosen arbitrarily, contradicts local extremality of classification $H^*$.

□

# References

[1] M. A. Aizerman, E. M. Braverman, and L. I. Rozonoer. *The Method of Potential Functions in Machine Training Theory.* Nauka, Moscow, 1970. In Russian.

[2] A. J. Bagnall and G. J. Janacek. Clustering time series from arma models with clipped data. In *KDD '04: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 49–58, New York, NY, USA, 2004. ACM.

[3] E. V. Bauman and A. A. Dorofeyuk. Recursive automatic classification algorithms. *Automation and Remote Control*, (3):345–355, 1982.

[4] E. M. Braverman, B. M. Litvakov, I. B. Muchnik, and S. G. Novikov. Stratified sampling in the organization of empirical data collection. *Automation and Remote Control*, 36(10):1629–1641, 1975.

[5] S. Brown and W. Goetzmann. Mutual fund styles, 1997.

[6] S. Déjean, P. G. Martin, A. Baccini, and P. Besse. Clustering time-series gene expression data using smoothing spl ine derivatives. *EURASIP Journal on Bioinformatics and Systems Biology*, 2007.

[7] A. B. Dor, R. Jagannathan, and I. Meier. Understanding mutual fund and hedge fund styles using return-based style analysis. *Journal of Investment Management*, 1(1):94–134, 2003.

[8] J. Ernst, G. J. Nau, and Z. Bar-Joseph. Clustering short time series gene expression data. *Bioinformatics*, 21(Suppl. 1):i159–168, 2005.

[9] A. K. Jain and R. C. Dubes. *Algorithms for clustering data.* Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1988.

[10] K. Kalpakis, D. Gada, and V. Puttagunta. Distance measures for effective clustering of ARIMA time-series. In *ICDM*, pages 273–280, 2001.

[11] N. E. Kiseleva, I. B. Muchnik, and S. G. Novikov. Stratified samples in the problem of representative types. *Automation and Remote Control*, 47(5):684–693, 1986.

[12] T. W. Liao. Clustering of time series data–a survey. *Pattern Recognition*, 38(11):1857–1874, November 2005.

[13] J. Lin, M. Vlachos, E. J. Keogh, D. Gunopulos, J.-W. Liu, S.-J. Yu, and J.-J. Le. A mpaa-based iterative clustering algorithm augmented by nearest neighbors search for time-series data streams. In T. B. Ho, D. W.-L. Cheung, and H. Liu, editors, *PAKDD*, volume 3518 of *Lecture Notes in Computer Science*, pages 333–342. Springer, 2005.

[14] J. MacQueen. Some methods for classification and analysis of multivariate observations. In L. M. L. Cam and J. Neyman, editors, *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297, Berkley, CA, 1967. University of California Press.

[15] J. Neyman. On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97:558–625, 1934.

[16] F. Pattarin, S. Paterlini, and T. Minerva. Clustering financial time series: an application to mutual funds style analysis. *Computational Statistics & Data Analysis*, 47(2):353–372, 2004.

[17] P. P. Rodrigues, J. Gama, and J. P. Pedroso. Odac: Hierarchical clustering of time series data streams. In J. Ghosh, D. Lambert, D. B. Skillicorn, and J. Srivastava, editors, *Proceedings of the 6th SIAM International Conference on Data Mining (SDM 2006)*, pages 499–503, Bethesda, Maryland, USA, April 2006. SIAM. DBLP, SCOPUS.

[18] L. J. Schulman. Clustering for edge-cost minimization (extended abstract). In *STOC '00: Proceedings of the thirty-second annual ACM symposium on Theory of computing*, pages 547–555, New York, NY, USA, 2000. ACM.

[19] A. Singhal and D. E. Seborg. Clustering multivariate time-series data. *Journal of Chemometrics*, 19:427–438, 2005.