

# A Plot for Visualizing Multivariate Data

Rida E. A. Moustafa<sup>1</sup>

Email: rmoustaf@galaxy.gmu.edu, rmustafa@aalcpas.com

April 8, 2003

## Abstract

Due to rapid changes in the data set sizes, both in dimensions and observations, there is an urgent need for highly sophisticated techniques that allow the analyst to visually discover hidden patterns that might exist in a given data set. In this paper, we introduce a new technique for visualizing large multivariate data by means of dimension reduction. This new technique will allow the user to visually detect hidden structures from any given large, multivariate and complex data set in an efficient and interactive way, compared to widely used dimensional reduction techniques.

**Keyword:** Visual data mining, Pattern Recognition, Nonlinear transformation.

## 1 Introduction

Visualizing multivariate data is an indirect approach in data mining and knowledge discovery processes. It is considered one of the most important steps in exploring hidden structures and attaining a better understanding of the relationships in the data sets. Large dimensions are hard to visualize. Moreover, most of the conventional multivariate visualization techniques generally do not scale very well for large number of observations; hence, it is difficult to interactively visualize large and complex data sets. Therefore, more attention must be expended in developing new techniques and softwares for visual data mining that can deal with complex, large-sized multivariate data.

It is worth mentioning some of the dimensional reduction techniques that widely used to visualize multivariate data. These techniques are, multidimensional scaling (MDS), principle component analysis (PCA), fisher discriminant analysis (FDA), projection pursuit and self-organized maps (SOM).

In this paper, we introduce an efficient algorithm for visualizing large-sized multivariate data sets. It has strong mathematical and statistical foundations that assist the analyst, not only in visually distinguishing classes in the data set, but, also, in understanding the underlying geometric structures of the data set.

Our algorithm is constructed of two consecutive projections: The first is the linear projection ( $m$ ) produced by computing the distance of each observation from the

---

<sup>1</sup>Dr. Rida Moustafa is the Director of Data Mining and Knowledge Discovery Group at AAL.

origin. The second projection is ( $v$ ) produced by constructing a nonlinear function of the data and the first projection and is computed by the following steps:

- find the deviations of the observation from the first projection
- compute the distance of these deviated observations from the origin
- find the relationship between the two projections in a scatter-plot view graph

We demonstrate the efficiency of this  $mv$ -method in visualizing multivariate data and its application in several areas of statistical data mining and pattern recognition.

The rest of the article consists of the following main sections: Section two discusses the theory of the introduced technique with illustrations, using simulated data. In section three, we demonstrate the power of our method on three real data sets and compare our results with well known existing methods for visualizing multivariate data by means of dimension reductions.

## 2 The theory of $mv$ -algorithm

The  $mv$ -plot is a 2-dimensional plot of multivariate large-sized data set. It is designed to assist the analyst visually explore the underlying structure(s) of a given data set. It can be considered a technique for visual clustering and classification. Furthermore, it allows us to recognize geometric objects such as hyper-planes, hyper-spheres, etc. The designed algorithm for the  $mv$ -plot consists of linear and nonlinear invariant transformation processes that map points in  $R^d$  into points in  $R^1$ .

The mapping processes can be described as follows: Given an observation  $x \in R^d$  as  $x = (x_1, x_2, \dots, x_d)$ , then  $m = f(x) = \frac{1}{d} \sum_{j=1}^d |x_j|$  and  $v = \sqrt{g(x, f(x))} = \sqrt{\frac{1}{d} \sum_{j=1}^d |x_j - f(x)|^2}$ . In general, for a multivariate data set of  $n$ -observations in  $d$ -dimensional space, say  $X = \{x_{ij} | i = 1, 2, \dots, n; j = 1, 2, \dots, d, \}$  the mapping process will produce the vectors  $m$  and  $v$ , both of length  $n$ . This is represented by the following equations:

$$m_i = \frac{1}{d} \sum_{j=1}^d |x_{ij}|$$

$$v_i = \left( \frac{1}{d} \sum_{j=1}^d (x_{ij} - m_i)^2 \right)^{\frac{1}{2}}$$

Understanding the mathematical and statistical relationship between  $m$  and  $v$  is essential in order to discover any pattern and to explain the results. Therefore, we investigate the relationship from several perspectives of mathematics and statistics on simulated data in two, three and hyper-dimensional space. Furthermore, we consider several real data sets that have been frequently studied for the same purpose of classification, clustering and visual data mining. Finally, we compare the  $mv$ -plot

results with other effective techniques that have been used for the same purposes, such as: principle component, multidimensional scaling and Fisher discriminant analysis.

## 2.1 The $mv$ plot in two dimensions

In this part, we will discuss the situation where the data  $X \in \mathbb{R}^{n^2}$ . This allows us to show and understand the relationships in 2-dimensional original space and its projection into 2-dimensional  $mv$ -space. We consider scaling data to be positive real numbers and this will be the same for all the cases we consider through this study. In this case, computation of  $m$  and  $v$  values is as follows:

$$\begin{aligned} m_i &= \frac{1}{2}(x_{i1} + x_{i2}), \\ v_i^2 &= \frac{1}{2} (|x_{i1} - m_i|^2 + |x_{i2} - m_i|^2) = \frac{1}{4} (|x_{i1} - x_{i2}|^2 + |x_{i2} - x_{i1}|^2) \\ &\Rightarrow v_i = \frac{1}{2}|x_{i2} - x_{i1}|. \end{aligned}$$

This is exactly like computing the principle component in two dimensions. Assume our data is drawn from a linear model of the form  $x_{i2} = w_1x_{i1} + w_0$ , then:

$$\begin{aligned} m_i &= \frac{1}{2} ((w_1 + 1)x_{i1} + w_0), \\ v_i &= \frac{1}{2} |(w_1 - 1)x_{i1} + w_0|. \end{aligned}$$

Consider the weights  $w$ 's are large enough such that  $w_1 - 1 \approx w_1 + 1 = a_1, w_0 = a_2$  both of which are positive reals. Then, the relationship between  $m$  and  $v$  is linear. Therefore, the linear pattern in the original space is mapped into a linear pattern in the  $mv$ -space by scaling the dataset.

As an example, we consider three cases for single line, intersected lines and parallel lines. These lines are shown at the Figure 1, respectively. Its  $mv$ -plots are shown at the bottom of the same figure. One can notice that the structure is preserved under the projection.

## 2.2 The $mv$ plot in three dimensions

In this subsection, we will study the situation where the data  $X \in \mathbb{R}^{n^3}$ . This of course, helps in the generalization process and shows the relationship between the projection of the three dimensional structures in data set and their correspondence in two dimensional ones. Computing  $m$  and  $v$  values for the data in  $X$ , we get:

$$m_i = \frac{1}{3} (x_{i1} + x_{i2} + x_{i3})$$

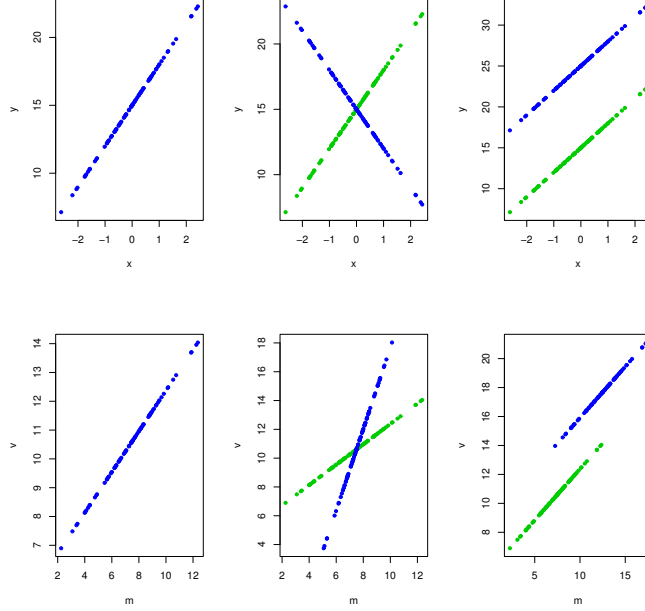


Figure 1:  $mv$ -plot of Line(s).

$$v_i^2 = \frac{1}{3} \left( |x_{i1} - m_i|^2 + |x_{i2} - m_i|^2 + |x_{i3} - m_i|^2 \right) \Rightarrow$$

$$v_i^2 = \frac{1}{3} \left( |2x_{i1} - (x_{i2} + x_{i3})|^2 + |2x_{i2} - (x_{i3} + x_{i1})|^2 + |2x_{i3} - (x_{i2} + x_{i1})|^2 \right).$$

Let us have linear model of the form  $x_{i3} = w_1x_{i1} + w_2x_{i2} + w_0$  then:

$$m_i = \frac{1}{3} \left( (w_1 + 1)x_{i1} + (w_2 + 1)x_{i2} + w_0 \right);$$

$$v_i^2 = \frac{1}{3} \left( |(w_1 - 2)x_{i1} + (w_2 + 1)x_{i2} + w_0|^2 + |(w_2 - 2)x_{i2} + (w_1 + 1)x_{i1} + w_0|^2 + |(2w_1 - 1)x_{i1} + (2w_2 - 1)x_{i2} + 2w_0|^2 \right).$$

Suppose the weights  $w$ 's are large enough such that  $w_1 + 1 \approx w_1 - 2 \approx w_1 - 0.5 = a_1$ ,  $w_2 + 1 \approx w_2 - 2 \approx w_2 - 0.5 = a_2$ ,  $w_0 = a_3 \Rightarrow$ :

$$m_i \approx a_1x_{i1} + a_2x_{i2} + a_3;$$

$$v_i \approx a_1x_{i1} + a_2x_{i2} + a_3;$$

From these two cases, we suspect relationship between  $m_i$  and  $v_i$  to be nearly linear. We will show this case by an example, we consider random data that fit

plane(s) in 3- $d$ . Figure 2, at top-left and bottom-left respectively, we show a plane and 2-planes in the original space. The corresponding line and 2-lines in  $mv$ -space are shown, respectively, at top-right and bottom-right of the figure.

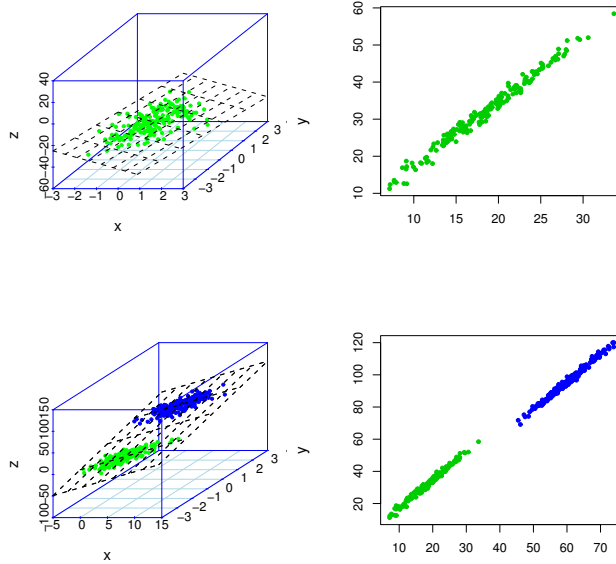


Figure 2:  $mv$ -plot of 3- $d$  plane(s).

### 2.3 The $mv$ plot in $d$ -dimensions

In this part we will consider the data  $X \in \mathbb{R}^{\text{nd}}$  drawn from a linear model given by the following relationship  $x_{id} = w_0 + \sum_{j=1}^{d-1} w_j x_{ij}$ , after some restriction on the weights, we can express  $m$  and  $v$  as:

$$m_i = \frac{1}{d} \left[ \sum_{i=1}^{d-1} (w_j + 1)x_{ij} + w_0 \right],$$

$$v_i = \frac{d-1}{d^2} \left[ \sum_{i=1}^{d-1} ((d-1)w_j - 1)x_{ij} + (d-1)w_0 \right].$$

Of course we can adjust the weights to be high enough as in the 3- $d$  situations. we will have a linear relationship between  $m$  and  $v$  as :  $m_i = \sum_{j=1}^{d-1} a_j x_{ij} + a_d$ , and  $v_i = \sum_{j=1}^{d-1} a_j x_{ij} + a_d$ . Therefore, the linear relationship can be detected, even for very large dimensions, by scaling the weights and, on real data by scaling the variables.

## 2.4 Detecting nonlinear data with $mv$ -plot

In this subsection, we would like to show that it is possible to detect nonlinear patterns using the  $mv$ -technique. Let us consider nonlinear data sampled from the sin and the cos functions on  $[0, 2\pi]$ . Its plot in the original space is shown at the top (left and right) of Figure 3 and its plot in the  $mv$  space is shown at the bottom (left and right). Notice that, at the  $mv$ -space, for  $(x, \cos(x))$  we have  $(m = x + \cos(x), v = |x - \cos(x)|)$  and for  $(x, \sin(x))$ , we have  $(m = x + \sin(x), v = |x - \sin(x)|)$ . This explains why the produced plots are rotated sin and cos in the  $mv$  space.

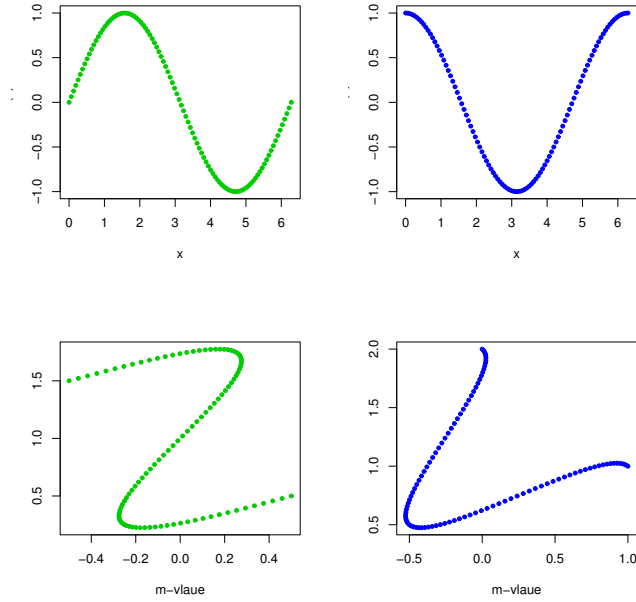


Figure 3:  $mv$ -plot of sin and cos functions.

In fact, we can employ  $mv$ -technique to detect sphere(s) in hyper-dimensions. Let the given data be distributed on a sphere with radius  $r$  and the data centered at zero. Computing the  $v$ -values based on the  $L_2$  norm, we get:

$$v_i^2 = \frac{1}{d} \sum_{j=1}^d (x_{ij} - m_i)^2 = \frac{1}{d} \left( \sum_{j=1}^d x_{ij}^2 - dm_i^2 \right)$$

$$\Rightarrow v_i^2 + m_i^2 = \frac{r^2}{d}.$$

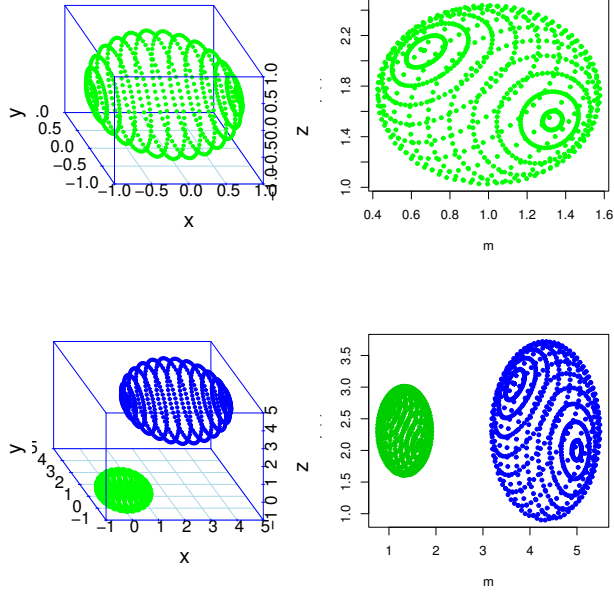


Figure 4:  $mv$ -plot of 3- $d$  sphere(s).

In general, for non-zero centered spheres, we have:

$$v_i^2 = \frac{1}{d} \sum_{j=1}^d (x_{ij} - x_c + x_c - m_i)^2 = \frac{1}{d} \left( \sum_{j=1}^d (x_{ij} - x_c)^2 - d(x_c - m_i)^2 \right)$$

$$\Rightarrow v_i^2 + m_i^2 = \frac{r^2}{d}.$$

Therefore, sphere(s) in hyper-dimension original space will be circle(s) in the  $mv$ -space. Let us consider data points that are randomly distributed on a sphere, shown in 3-dimension at the top-left of Figure 4. Its plot in  $mv$ -space is a circle shown at top-right of the figure. At bottom-right, we have two separated spheres with different radii. The corresponding circles in  $mv$ -space are shown at the bottom-left of the figure.

### 3 Results and Discussions

#### 3.1 The iris data set

In this subsection, we will present our results on the well-known Fisher’s Iris data. This data and it can be obtained from URL <sup>2</sup>. It consists of 150 observations containing four measurements based on the petals and sepal of three species of Iris. These three species are: Iris Setose, Iris Virginia, and Iris Versicolor. This data is an appropriate choice, as it is frequently used to illustrate classification, clustering or visualization techniques (Venables,Ripley 1998),(Martinez(2002).

The visualization shown in Figure 5 demonstrates that *mv*-plot achieve as satisfactory a result as the three well known plots, namely: the Principle component plot, Multidimensional scaling plot, and Fisher’s discriminant plot. All the plots, including ours, show that the data consists of three groups and one of those three (Iris Setose) is highly separated from the other two. Moreover, according to *mv*-plot theory, we can confirm that the three groups lie on three hyper-planes in  $R^4$ .

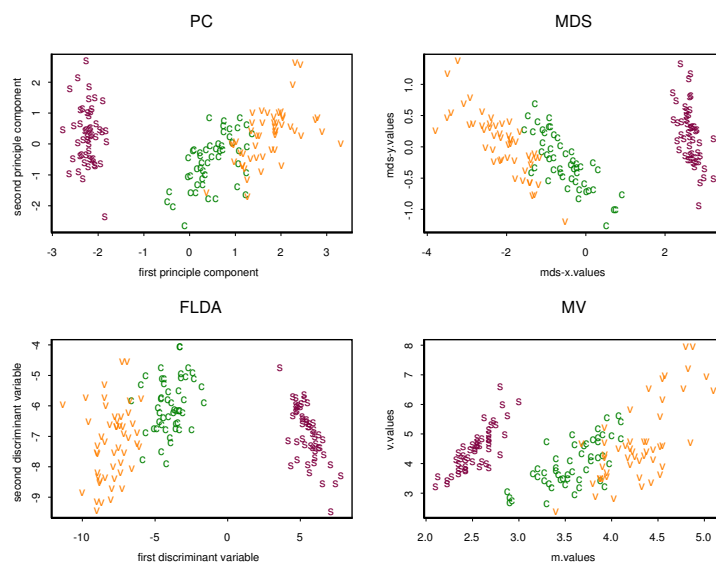


Figure 5: *mv*-plot of the Iris data: comparison.

#### 3.2 The process control data set

The second data set is the Synthetic Control Chart Time Series. It contains 600 examples of control charts synthetically generated by the process in Alcock and

<sup>2</sup><http://www.cmu.edu>

Manolopoulos (1999). The data set lies in 60 dimensional space. There are six different classes of control charts: 1-normal, 2-cyclic, 3-increasing trend, 4-decreasing trend, 5-upward shift, 6-downward shift. This time series data is listed on Knowledge discovery cup-1999 as a challenging problem for the data mining community. It can be obtained from the URL <sup>3</sup>. The data set is generally used for testing classification, clustering and visualization techniques. Several clustering techniques have been tested on this particular data set and did not show the existing clusters(Pham, Chan 1998). Our goal here is to visually explore the existing 6-classes, using *mv*-algorithm, and to compare our results with other techniques. In Figure 6, we can see that the PCP and FLDA methods are able to distinguish the first and second classes properly, but there is no clear discrimination between the classes (three, five) and (four, six). The MDS method did not show groups at all. Comparing this with the *mv*-plot, there is no doubt that one can see that there are six classes in this particular data set. Moreover, the computational effort is negligible compared with other methods studied here.

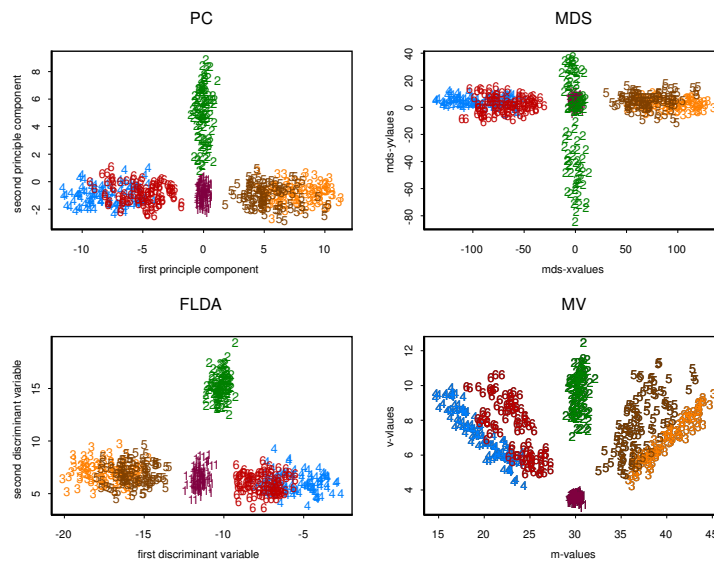


Figure 6: *mv*-plot of the process control time series data: comparison.

### 3.3 The Pollen data set

This data set consists of 3,848 observations in 5 dimensional space. It is a challenging data set for statisticians offered by the American Statistical Association. It is

<sup>3</sup><http://kdd.ics.uci.edu>

worth mentioning here that most of the existing techniques and, also, the ones we report on need more computer power (storage and speed) to complete the process; therefore, we will report only on our plot. In Figure 7, we see a filled circular shape and a line coming out of it. According to the *mv*-plot theory, one can tell that the circular shape is hyper-sphere in 5-dimensions, and, the linear shape is a hyper-plane in 5-dimensions, too. We extract the linear pattern from the data set and visualize it in Figure 8, which shows the word (EUREKA). In fact, our discoveries with the *mv*-plot agree with those found in (Wegman 2002) and with others that have successfully analyzed this data set.

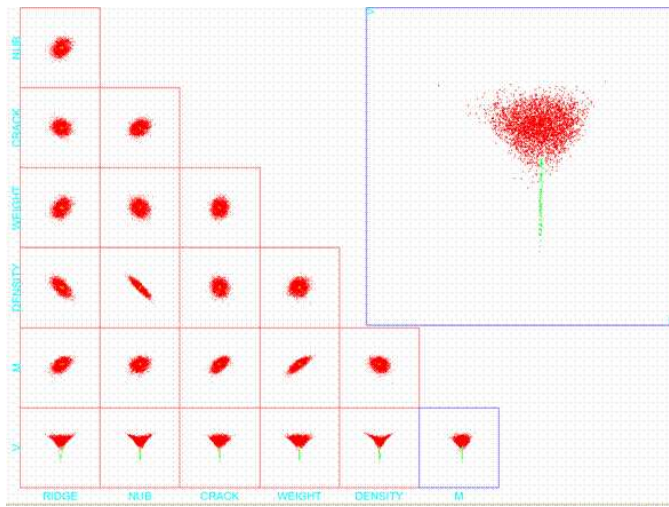


Figure 7: *mv*-plot of Pollen data: detecting both of linear and nonlinear structure.

## 4 Summary and Future work

We introduced an efficient algorithm for visualizing hyper-dimensional data sets. We demonstrated the efficiency of the algorithm on different complex simulated and real data set in very high-dimensional space and were able to detect the existing patterns visually. We compared the algorithm on the same data sets with the principal components analysis, fisher discriminant analysis, and multidimensional scaling. In most cases, our algorithm achieved better results than the afore-mentioned algorithms. In some other cases, these algorithms didn't work at all because of the computer speed and storage limitations of most computers.

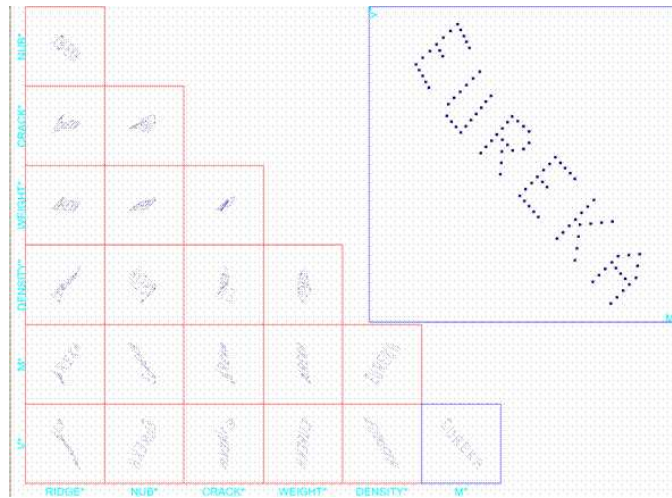


Figure 8: *mv*-plot the hidden linear structure in Pollen data.

## 5 Acknowledgements

The authors would like to thank the computational statistics group at George Mason University for their help and support and especially Dr. John Miller and Dr. Edward Wegman and AAL's Advanced Data Mining Group.

This work was funded in part by the Air Force Office of Scientific Research under the contract F49620-01-1-0274.

## 6 References

- Pham D. , Chan A., **1998** "Control Chart Pattern Recognition using a New Type of Self Organizing Neural Network" Proc. Instn, Mech, Engrs. Vol 212, No 1, pp 115-127.
- Mitchie, Spiegelhalter, and Taylor, **1994** *Machine Learning, Neural and Statistical Classification*, (STATLOG project).
- Martinez, W., Martinez, A., **2002** *Computational Statistics Handbook with Matlab*, CRC.
- Venables, W, Ripley, B., *Modern Applied Statistics with S-PLUS*, Second Edition, Springer, **1998**
- Everitt, B., Dunn, G. ,**2001** *Applied Multivariate Data Analysis*, Second Edition.
- Everitt, B. and Dunn, G. **1983** *Advanced Methods of Data Exploration and Modeling*, Heinman, London.
- Everitt, B. and Dunn, G. **1991** *Applied Multivariate Data Analysis*, John Wiley & Sons, New York.

- Fienberg, S. **1979** "Graphical methods in statistics," *The American Statistician*, **33**, 165-178.
- Fishback, W. T. **1962** *Projective and Euclidean Geometry*, John Wiley & Sons, New York.
- Kaufman, L. and Rousseeuw, P. **1990** *Finding Groups in Data: An Introduction to Cluster Analysis*, John Wiley & Sons, New York.
- Mortenson, M. E. **1995** *Geometric Transformation*, Industrial Press Inc., New York.
- Moustafa, R. E. **2001** *Fast Conceptual Clustering Algorithm For Data Mining and Visualization*, Ph.D. Dissertation, George Mason University, Fairfax, Va.
- Wegman, E., Carr, D., and Luo, Q. **1993** "Visualizing Multivariate Data," *Multivariate Analysis: Future Directions*, (C. R. Rao, ed.), Amsterdam: North Holland, 423-466.
- Wegman, E. and Luo, Q. **1997** "High dimensional clustering using parallel coordinates and the grand tour," *Computing Science and Statistics*, 28 352-360.
- Wegman, E. and Solka, J. **2002** "On some mathematics for visualizing high dimensional data," *Sankhya*, 64(2),429-452. Wegman, E. **2003** "Visual data mining"(in press) *Statistics in Medicine*.