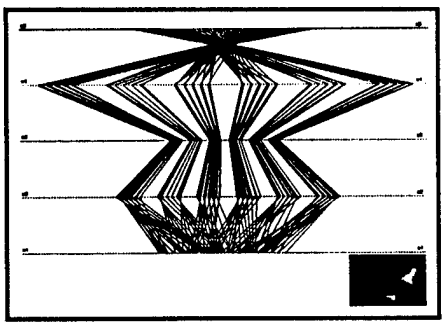
High Dimensional Clustering Using Parallel Coordinates and the Grand Tour

Edward J. Wegman
and
Qiang Luo

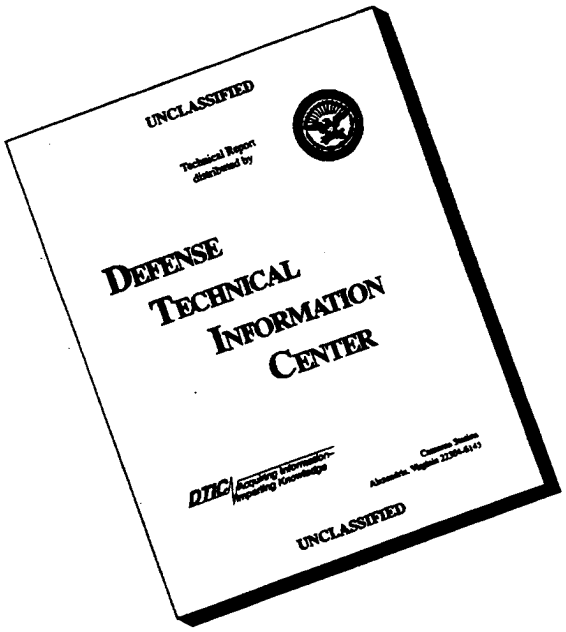
Technical Report No. 124
April, 1996

Center for Computational Statistics



George Mason University
Fairfax, VA 22030

DISCLAIMER NOTICE



THIS DOCUMENT IS BEST QUALITY AVAILABLE. THE COPY FURNISHED TO DTIC CONTAINED A SIGNIFICANT NUMBER OF PAGES WHICH DO NOT REPRODUCE LEGIBLY.

CENTER FOR COMPUTATIONAL STATISTICS
TECHNICAL REPORT SERIES

TTR 110. Edward J. Wegman, Huge Data Sets and the Frontiers of Computational Feasibility, November, 1994. published ***Journal of Computational and Graphical Statistics***, 4(4), 281-195, 1995.

TR 111. Winston C. Chow, Fractional Process Modeling, November, 1994.

TR 112. Mark C. Sullivan, ***Computationally Efficient Statistical Signal Processing Using Nonlinear Operators*** (Ph.D. Dissertation), December, 1994.

TR 113. Irwin Greenberg, Some Simple Approximation Methods in Level Crossing Problems, December, 1994.

TR 114. Jeffrey L. Solka, Wendy L. Poston and Edward J. Wegman, A New Visualization Technique to Study the Time Evolution of Finite and Adaptive Mixture Estimators, December, 1994. published ***Journal of Computational and Graphical Statistics***, 4(3), 180-198, 1995.

TR 115. D. B. Carr and A. R. Olsen, Representing Cumulative Distributions with Parallel Coordinate Plots, August, 1995

TR 116. Jeffrey L. Solka, ***Matching Model Information Content to Data Information*** (Ph.D. Dissertation), August, 1995.

TR 117. Wendy L. Poston, ***Optimal Subset Selection Methods***, (Ph.D. Dissertation), August, 1995.

TR 118. Clifton D. Sutton, Sphere Packing, August, 1995.

TR 119. Wendy L. Poston, Edward J. Wegman, and Jeffrey L. Solka, A Parallel Algorithm for Subset Selection, August, 1995.

TR 120. Barnabas Takacs, Harry Wechsler, and Edward J. Wegman, A Model of Active Perception and its Implementation on the Intel Paragon XP/S, August, 1995.

TR 121. Shan-chuan Li, Walter Dyar, and Mary-Ellen Verona, GRASS Database Explored and Applied to Biodiversity Query with Splus, August, 1995, to appear ***Computing Science and Statistics***, 27, 1995.

TR 122. Kathleen Golitko Perez-Lopez, ***Management of Scientific Image Databases Using Wavelets*** (Ph.D. Dissertation), August, 1995.

TR 123. Edward J. Wegman, Jeffrey L. Solka and Wendy L. Poston, Immersive Methods for Mine Warfare, April, 1996.

TR 124. Edward J. Wegman and Qiang Luo, High Dimensional Clustering using Parallel Coordinates and the Grand Tour, April, 1996.

TR 125. Kletus A. Lawler, ***Linear and Nonlinear Regression Estimates for a Cobb-Douglas Model***, (M.S. Thesis), April, 1996.

TR 126. Ehsan S. Soofi, Information Theoretic Regression Methods, April, 1996.

High Dimensional Clustering Using Parallel Coordinates and the Grand Tour

Edward J. Wegman, Qiang Luo

Center for Computational Statistics
George Mason University
Fairfax, VA 22030 USA

Abstract

In this paper, we present some graphical techniques for cluster analysis of high-dimensional data. Parallel coordinate plots and parallel coordinate density plots are graphical techniques which map multivariate data into a two-dimensional display. The method has some elegant duality properties with ordinary Cartesian plots so that higher-dimensional mathematical structures can be analyzed. Our high interaction software allows for rapid editing of data to remove outliers and isolate clusters by brushing. Our brushing techniques allow not only for hue adjustment, but also for saturation adjustment. Saturation adjustment allows for the handling of comparatively massive data sets by using the α-channel of the Silicon Graphics workstation to compensate for heavy overplotting.

The grand tour is a generalized rotation of coordinate axes in a high-dimensional space. Coupled with the full-dimensional plots allowed by the parallel coordinate display, these techniques allow the data analyst to explore data which is both high-dimensional and massive in size. In this paper we give a description of both techniques and illustrate their use to do inverse regression and clustering. We have used these techniques to analyze data on the order of 250,000 observations in 8 dimensions. Because the analysis requires the use of color graphics, in the present paper we illustrate the methods with a more modest data set of 3848 observations. Other illustrations are available on our web page.

1. Introduction

Visualization of high-dimensional, multivariate data has enjoyed a considerable development with the introduction of the grand tour by Asimov (1985) and Buja and Asimov (1985) and the parallel coordinate display by Inselberg (1985) and Wegman (1990). The former technique is an animation methodology for viewing two-dimensional projections of general d-dimensional data where the animation is determined by a space-filling curve through all possible orientations of a two-dimensional coordinate system in d-space. Viewed as a function of time, the grand-tour animation reveals interesting projections of the data, projections that reveal underlying structure. This in turn allows for the construction of models of data's underlying structure.

The parallel coordinate display is in many senses a generalization of a two-dimensional Cartesian plot. The idea is to sacrifice orthogonal axes by drawing the axes parallel to each other in order to obtain a planar diagram in

which each d-dimensional point has a unique representation. Because of elegant duality properties, parallel coordinate displays allow interpretations of statistical data in a manner quite analogous to two-dimensional Cartesian scatter plots. Wegman (1991) formulated a general d-dimensional form of the grand tour and suggested using the parallel coordinate plot as a visualization tool for the general d-dimensional animation. We have found this combination of multivariate visualization tools to be extraordinarily effective in the exploration of multivariate data. In Section 2, we briefly describe parallel coordinate displays including interpretation of parallel coordinate displays for detecting clusters. In Section 3, we describe the generalized d-dimensional grand tour and a partial sub-dimensional grand tour. In Section 4, we discuss brushing with hue and saturation including a discussion of perceptual considerations for visual presentation. Finally we close in Section 5 with a sequence of illustrations of the use of these techniques to remove noise and isolate clusters in a five-dimensional data set.

2. Parallel Coordinate and Parallel Coordinate Density Plots

The parallel coordinate plot is a geometric device for displaying points in high- dimensional spaces, in particular, for dimensions above three. As such, it is a graphical alternative to the conventional scatterplot. The parallel coordinate density plot is closely related and addresses the situation in which there would be heavy overplotting. In this circumstance, the parallel coordinate plot is replaced with its density and so is much more appropriate for very large, high-dimensional data sets. In place of the conventional scatter plot which tries to preserve orthogonality of the d-dimensional coordinate axes, draw the axes as parallel. A vector $(x_1, x_2, \ldots, x_d)$ is plotted by plotting x_1 on axis 1, x_2 on axis 2 and so on through x_d on axis d. The points plotted in this manner are joined by a broken line. The principal advantage of this plotting device is that each vector $(x_1, x_2, \ldots, x_d)$ is represented in a planar diagram in which each vector component has essentially the same representation.

The parallel coordinate representation enjoys some elegant duality properties with the usual Cartesian orthogonal coordinate representation. Consider a line $\mathcal{L}$ in the Cartesian coordinate plane given by $\mathcal{L}$: $y = mx+b$ and consider two points lying on that line, say $(a, ma+b)$ and $(c, mc+b)$. Superimpose a Cartesian coordinate axes t, u on the xy parallel axes so that the y parallel axis has the equation $u = 1$. The point $(a, ma + b)$ in the xy Cartesian system maps into the line joining $(a, 0)$ to $(ma + b, 1)$ in the tu coordinate axes. Similarly, $(c, mc+b)$ maps into the line joining $(c, 0)$ to $(mc+b, 1)$. A straightforward computation shows that these two lines intersect at a point (in the tu plane) given by $\bar{\mathcal{L}} : (b(1-m)^{-1}, (1-m)^{-1})$. This point in the parallel coordinate plot depends only on m and b, the parameters of the original line in the Cartesian plot. Thus $\bar{\mathcal{L}}$ is the dual of $\mathcal{L}$ and one has the interesting duality result that points in Cartesian coordinates map into

lines in parallel coordinates while lines in Cartesian coordinates map into points in parallel coordinates. This duality is discussed in further detail in Wegman (1990).

The point-line, line-point duality seen in the transformation from Cartesian to parallel coordinates extends to conic sections. The most significant of these dualities from a statistical point of view is that an ellipse in Cartesian coordinates maps into a hyperbola in parallel coordinates. A distribution which has ellipsoidal level sets would have hyperbolic level sets in the parallel coordinate presentation. It should be noted that the quadratic form does not describe a locus of points, but a locus of lines, a line conic. The notion of a line conic is, perhaps, a strange notion. By this is meant a locus of lines whose coordinates satisfy the equation for a conic. These may be more easily related to the usual notion of a conic when it is realized that the envelope of this line conic is a point conic. As mentioned there is a duality between points and lines and between conics and conics. It is worthwhile to point out two other nice dualities. Rotations in Cartesian coordinates become translations in parallel coordinates and vice versa. Perhaps more interesting from a statistical point of view is that points of inflection in Cartesian space become cusps in parallel coordinate space and vice versa. Thus the relatively hard-to-detect inflection point property of a function becomes the notably more easy to detect cusp in the parallel coordinate representation. Inselberg (1985) discusses these properties in detail.

Since ellipses map into hyperbolas, one has an easy template for diagnosing uncorrelated data pairs. With a completely uncorrelated data set, one would expect the 2-dimensional scatter diagram to fill substantially a circumscribing circle. The parallel coordinate plot would approximate a figure with a hyperbolic envelope. As the correlation approaches negative one, the hyperbolic envelope would deepen so that in the limit one would have a pencil of lines, what is called by Wegman (1990) the cross-over effect.

Most importantly for the present paper, it should be noted that clustering is easily diagnosed using the parallel coordinate representation. The individual parallel coordinate axes represent one-dimensional projections of the data. Thus, separation between or among sets of data on any one axis or between any pair of axes represents a view of the data which isolates clusters. An elementary view of this idea is seen in Figure 1, where we illustrate the appearance of three distinct clusters in a four dimensional space. Because of the connectedness of the multidimensional parallel coordinate diagram, it is usually easy to see whether or not this clustering propagates through other dimensions.

Some of the data analysis features of the parallel coordinate representation include the ability to diagnose one-dimensional features such as **marginal densities**, two-dimensional features such as **correlations** and **nonlinear structures**, and multi-dimensional features such as **clustering, hyperplanes**, and the **modes**. These interpretations are discussed in more detail in Wegman (1990) while parallel coordinate density plots are discussed in

Miller and Wegman (1991).

3. The Grand Tour Algorithm in d-space

Let $\boldsymbol{e}_j = (0, 0, \ldots, 0, 1, 0, \ldots, 0)$ be the canonical basis vector of length d. The 1 is in the j^{th} position. The $\boldsymbol{e}_j$ are the unit vectors for each coordinate axis in the initial position. We want to do a general rigid rotation of these axes to a new position with basis vectors $\boldsymbol{a}_j(t) = (a_1^j(t), a_2^j(t), \ldots, a_d^j(t))$, where, of course, t is a time index. The strategy then is to take the inner product of each data point, say $\boldsymbol{x}_i$, $i = 1, \ldots, n$ with the basis vectors, $\boldsymbol{a}_j(t)$. The j subscript on $\boldsymbol{a}_j(t)$ means that $\boldsymbol{a}_j(t)$ is the image under the generalized rotation of the canonical basis vector $\boldsymbol{e}_j$. Thus the data vector $\boldsymbol{x}_i$ is $(x_1^i, x_2^i, \ldots, x_d^i)$, so that the representation of $\boldsymbol{x}_i$ in the $\boldsymbol{a}_j$ coordinate system is

$$\boldsymbol{y}_i(t) = (y_1^i(t), y_2^i(t), \ldots, y_d^i(t)), \ i = 1, \ldots, n$$

where

$$y_j^i(t) = \sum_{k=1}^{d} x_k^i a_k^j(t),$$

$j = 1, \ldots, d$ and $i = 1, \ldots, n$. The vector $\boldsymbol{y}_i(t)$ is then the linear combination of basis vectors representing the i^{th} data point in the rotated coordinate system at time t.

The goal thus is to find a generalized rotation, $\boldsymbol{Q}$, such that $\boldsymbol{Q}(\boldsymbol{e}_j) = \boldsymbol{a}_j$. We can think of $\boldsymbol{Q}$ as either a function or as a matrix $\boldsymbol{Q}$ where $\boldsymbol{e}_j \times \boldsymbol{Q} = \boldsymbol{a}_j$. We implement this by choosing $\boldsymbol{Q}$ as an element of the special orthogonal group denoted by SO(d) of orthogonal $d \times d$ matrices having determinant of $+1$. In order to find a continuous, space-filling path through the Grassmannian manifold of d-flats, we must find a continuous, space-filling path through the SO(d).

In general d-dimensional space, there are $d-2$ axes orthogonal to each two-flat. Thus rather than rotating around an axis as we are used to in ordinary three-dimensional space, we must rotate in a plane in d-dimensional space. The generalized rotation matrix, $\boldsymbol{Q}$, is built up from a series of rotations in individual two-flats. In d-space, there are d canonical basis vectors and, thus, $\binom{d}{2} = \frac{1}{2}(d^2 - d)$ distinct two-flats formed by the canonical basis vectors. We let $R_{ij}(\theta)$ be the element of SO(d) which rotates in the $e_i e_j$ plane through an angle of θ. We define $\boldsymbol{Q}$ by

$$\boldsymbol{Q}(\boldsymbol{\theta}_{1,2}, \boldsymbol{\theta}_{1,3}, \ldots, \boldsymbol{\theta}_{d-1,d}) = \boldsymbol{R}_{12}(\boldsymbol{\theta}_{1,2}) \times \cdots \times \boldsymbol{R}_{d-1,d}(\boldsymbol{\theta}_{d-1,d}).$$

There are $p = \frac{1}{2}(d^2 - d)$ factors. The restrictions on θ_{ij} are $0 \leq \theta_{ij} \leq 2\pi$, $1 \leq i < j \leq d$. The vector $(\theta_{1,2}, \theta_{1,3}, \ldots, \theta_{d-1,d})$ can thus be thought of as a point on a p-dimensional torus. This is the origin of the description of

this method as the torus method. The individual factors $R_{ij}(\theta)$ are $d \times d$ matrices given by

$$\begin{bmatrix} 1 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\ \vdots & & \vdots & & \vdots & & \vdots \\ 0 & \cdots & cos(\theta) & \cdots & -sin(\theta) & \cdots & 0 \\ \vdots & & \vdots & & \vdots & & \vdots \\ 0 & \cdots & sin(\theta) & \cdots & cos(\theta) & \cdots & 0 \\ \vdots & & \vdots & & \vdots & & \vdots \\ 0 & \cdots & 0 & \cdots & 0 & \cdots & 1 \end{bmatrix}$$

where the cosine and sine entries are in the i^{th} and j^{th} columns and rows.

The final step in the algorithm is to describe a space filling path on the p-dimensional torus, T^p. This can be done by a mapping α: $\boldsymbol{R} \rightarrow T^p$ given by

$$\alpha(t) = (\lambda_1 t, \lambda_2 t, \ldots, \lambda_p t)$$

where $\lambda_1, \ldots, \lambda_p$ is a sequence of mutually irrational real numbers and the $\lambda_i t$ are interpreted modulo 2π. The composition of α with $\boldsymbol{Q}$ will describe a space filling path in $\mathrm{SO}(d)$. Thus our final algorithm is given by

$$\boldsymbol{a}_j(t) = \boldsymbol{e}_j \times \boldsymbol{Q}(\lambda_1 t, \ldots, \lambda_p t)$$

The canonical unit vector for each coordinate axis at time t described by the grand tour algorithm is an orthogonal linear combination, $\boldsymbol{a}_j(t) = \boldsymbol{e}_j \times \boldsymbol{Q}(t)$, of the original unit vectors. This has several important implications for the utility of this methodology. First, it should be immediately clear that one can do a grand tour on any subset of the original coordinate axes simply by fixing the appropriate two-planes in the rotation matrix given above by $\boldsymbol{Q}$. That is, if we wish the j^{th} variable to not be included in the grand tour rotation, we simply put a 1 in the $\boldsymbol{Q}_{jj}(t)$ entry with 0 in the remaining positions in the j^{th} row and the j^{th} column. Thus it is straightforward to do a partial grand tour. The interest in doing a partial grand tour will be discussed in the next section.

The second important implication relates to the connection with the parallel coordinate display. An immediate concern with the parallel coordinate display is the preferential ordering of the axes. In our discussion above we indicated that the axis for variable one is adjacent to the axis for variable two, but not for variable three. In general the axis for variable j is adjacent to the axes for variables $j-1$ and $j+1$ but for no other axes. It is easy to see pair wise relationships for adjacent variables, but less easy for non-adjacent variables. Wegman (1990) has a substantial discussion on methods for considering all possible permutations. This concern is immaterial when one does the grand tour since eventually $\boldsymbol{a}_j(t) = \boldsymbol{e}_j \times \boldsymbol{Q}(t) = \boldsymbol{e}_i$, for every i. Thus eventually every possible permutation of the axes will appear in the

grand tour.

4. Some Additional Visualization Devices

4.1 Brushing with Hue and Saturation

A powerful method in high interaction graphics is the brushing technique. The idea is to isolate clusters or other interesting subsets of a data set by, in effect, painting that subset with a color. This is usually done in two settings: 1) with co-plots and 2) with animations. The brushed color becomes an attribute of the data point and is maintained in all representations. The idea of co-plots is that a particular data set may be presented in more that one way, for example in a scatter plot matrix or say with a scatter plot, a histogram and a dot plot. Points colored the same way in all presentations allow the data analyst the ability to track coherent clusters or subsets of the data through different representations. Of course, with an animation, the coloring allows the data analyst to follows clusters or subsets of the data through the time evolution of the animation.

In general, colors may assume of range of saturations depending on the relative proportion of gray to chroma. We implement the following device. We de-saturate the hue with black so that the brushing color is nearly black when desaturated. This by itself would not be fully useful. However, when points are overplotted, we add the hue components. Say, for example, we use an approximate one and one half percent hue component of blue. This would mean (on an eight bit scale) approximately 2 bits of blue and 0 bits each of red and green. Thus if approximately 67 observations were overplotted at a given pixel, that pixel would be fully saturated with blue. Fewer observations mean a less saturated color. The level of saturation of the brushing color is controllable by the user. Larger data sets suggest lower saturation levels. The level of saturation thus reflects the degree of overplotting. This device is in essence a way of creating a parallel coordinate (or any other kind of) density plot. (See Miller and Wegman, 1991). The advantage of this technique for creating a density plot is that it does not depend on smoothing algorithms so that individual data points are still resolvable.

The addition of saturations is implementable in hardware on Silicon Graphics workstations by means of the α-channel. The α-channel is a hardware device for blending pixel intensities and has its primary use for transparency algorithms. However, by blending pixels intensities of the same color, we can in effect add the pixel intensities and achieve brushing with hue and saturation with no speed penalty whatsoever. This technique is incredibly powerful in resolving structure in large data sets with heavy overplotting as we hope to illustrate in the next section.

4.2 Some Perceptual Issues.

Brushing with hue and saturation leads to an interesting question concerning perception of the resulting plots. When viewed against a black background, the low saturation observations, i.e. those that are not heavily overplotted, blend with that background. This is quite useful when trying to understand the internal structure of the high density regions of the plot. Our usual technique is to brush with a white (actually a very dark gray) color. Then the internal structure appears as white in the highest density regions as illustrated in Figure 4. The resulting plot looks rather like an x-ray of the internal structure of the data set.

When viewed against a white background, the low saturation level observations are nearly black and so are quite visible. This is extremely useful when looking for outliers, which would tend to be invisible against the black background. The white background is also extremely useful for data editing. Our implementation on the Silicon Graphics workstations supports a scissors feature so that we can prune away low density regions. This feature allows for rapid visual data editing which may be useful for eliminating outliers, transcription errors, and data with missing values that could impair the ability to reach sensible conclusions. Obviously, when using a white background, it is important to brush with some hue, since when brushed with a gray the result would be that highest density regions would be white again blending with the background. Because the apparent brightness of normal hues (red, blue, green, etc.) is lower than the apparent brightness of white, the internal structure of the data set is less apparent with a white background than with a black background. Thus it is clear that both backgrounds have their utility depending on the task at hand.

4.3 Visual Regression and Clustering Decision Rules.

The combination of hue and saturation brushing and the partial grand tour creates a device for visual regression and clustering decision rules. Consider a response variable of interest, let us say, for example, profit in a financial setting. Let us suppose we wish to answer the following question, "What combination of customer demographics variables is likely to cause the corporation to lose money?" We brush the profit variable as follows: for negative profits, we brush the observations red, for positive profits we brush them green. Where the variables overlap, the combination of red and green sum to yellow. Where there are observations primarily leading to losses, the result will be generally red and where profits, primarily green. Since we are interested in the covariates leading to profit, we fix the profit variable so that it does not enter into the partial grand tour rotation. We may rotate on any combination of explanatory covariates we wish. For example, we may have data on customer's average account balance, sex, race, age and annual income. While all of these may affect the profitability of the corporation,

the prohibitions against discrimination on the basis of race and sex would lead us to generate decision rules which do not consider these factors. Similarly, some data may be extremely difficult or expensive to collect. Thus while it may be an extremely helpful covariate, it may be missing so often that its value is substantially diminished in forming decision rules.

The partial grand tour is done on the explanatory covariates of interest while keeping the response variable and any other explanatory covariates that we wish to exclude fixed. Because the grand tour automatically forms orthogonal linear combinations of desired explanatory covariates, the color coding allows us in effect to see the response variable in terms of the orthogonal linear combinations of the explanatory variables. Thus when we see a linear combination of explanatory variables that is intensely red in our example, we know that this is a combination of variables which leads to a negative profit. We can thus isolate the range of the linear combination of covariates that is colored red and this will be a component of the decision tree in terms of demographic variables that causes the organization to lose money. We can then edit that particular cluster of observations from our data set and resume the partial grand tour. Repeating this process recursively allows us to determine a sequence of decision rules that isolate customers likely to cause financial loss to the organization. Because this methodology is so intensively dependent on color, it is not possible to easily illustrate these techniques in this paper. However, we have included an example based on this idea as well as other examples in our web server at URL

http : //www.galaxy.gmu.edu/images/gallery/research_arcade.html.

This method thus leads to a high interaction techniques for rapidly identifying a decision rule based on visual display. The rules are sophisticated in the sense that they need not be simple binary decision rules and they need not be based on simply the original covariates.

5. An Example

In this example we would like to consider a synthetic dataset about the geometric features of pollen grains. There are 3848 observations on 5 variables. This data is the 1986 ASA Data Exposition dataset, made up by David Coleman of RCA Labs. The data set is available from STALIB at URL=http://www.stat.cmu.edu/datasets/. Figure 2 is the scatterplot matrix for this five-dimensional data. Note that in all presentations the data appear to have elliptical contours. This is true even when all five variables are rotated through the grand tour. This is suggestive of the fact that the point cloud is sampled from a five-dimensional multivariate distribution with ellipsoidal level sets, perhaps a multivariate normal. Figure 3 is the corresponding parallel coordinate display with pure black on a white background. In this display each of the five variables have been rescaled so as to fill the parallel coordinate axes. Note that in the parallel coordinate display,

the variables exhibit hyperbolic envelopes, the dual of elliptical contours in Cartesian plots. This confirms our observation of the ellipsoidal level sets.

Figure 4 represents a highly desaturated version of the same parallel coordinate plot, this time white on a black background. With this desaturated view, it is clear that there is an interesting internal structure buried in the noise. Figure 5 represents a partially pruned view with much of the noise removed. Figure 6 is the result of a second pruning edit in which the internal structure is fully revealed. In both Figures 5 and 6 the data have been rescaled to fill the axes. Figure 7 represents the results of grand tour in which it is clear from the gaps seen in axes two, three and four that this data forms six clusters separable in at least three dimensions of the five. Our software permits this edit to be accomplished in less than three minutes. Figure 8 displays the edited data in a scatter plot display. The 99 remaining points from the original 3848 points are perfectly isolated from the noise and spell the word EUREKA. The six letters are of course the six clusters isolated in Figure 7. The 99 points are only about 2.7% of the data set and yet we were able to isolate these points in six clusters using the techniques described in this paper.

6. Acknowledgment

This work was supported by the U.S. Army Research Office under Grant DAAH04-94-G-0267. The software to accomplish the tasks described in this article is known as ExplorN and is a research product of the Center for Computational Statistics at George Mason University. It has been principally developed by the authors of this article with additional contributions from Mr. Ji Shen and Professor Daniel Carr. We thank Shan-chuan Li for converting our manuscript to TEX form.

7. References

Asimov, D. (1985), "The grand tour: a tool for viewing multidimensional data", **SIAM J. Sci. Statist. Comput.**, 6, 128-143.

Buja, A. and Asimov, D. (1985), "Grand tour methods: an outline", **Computer Science and Statistics: Proceedings of the Seventeenth Symposium on the Interface,** 63-67, (D. Allen, ed.), New York: North Holland Publishing Company.

Inselberg, A. (1985), "The plane with parallel coordinates", **The Visual Computer**, 1, 69-91.

Miller, J. J. and Wegman, E. J. (1991), "Construction of line densities for parallel coordinate plots", **Computing and Graphics in Statistics,** (A. Buja and P. Tukey, eds.), 107-123, Springer-Verlag: New York.

Wegman, E. J. (1990), "Hyperdimensional data analysis using parallel co-

ordinates", **J. American Statist. Assoc.**, 85, 664-675.

Wegman, E. J. (1991), "The grand tour in k-dimensions", **Computing Science and Statistics: Proceedings of the 22nd Symposium on the Interface,** 127-136.

Legends for Figures

Figure 1. a. Scatterplot matrix of three clusters in four dimensions. b. Parallel coordinate plot corresponding to the scatterplot matrix in 1.a. Note that a separation along any axis or in between axes is indicative of a cluster. Note also that distinctive slopes of the line segments between pairs of axes also separate clusters.

Figure 2. The scatterplot matrix of 3848 observations on 5 variables from a synthetic dataset about the geometric features of pollen grains. The level sets appear to be elliptical in all five dimensions suggesting a five-dimensional ellipsoidal shape. One might be tempted to guess multivariate Gaussianity.

Figure 3. The fully saturated parallel coordinate plot of the same 3848 observations in five space. The hyperbolic envelope tends to confirm the conclusions about a five dimensional ellipsoidal level set. However, little can be seen from either Figure 2 or Figure 3 about the internal structure of this data.

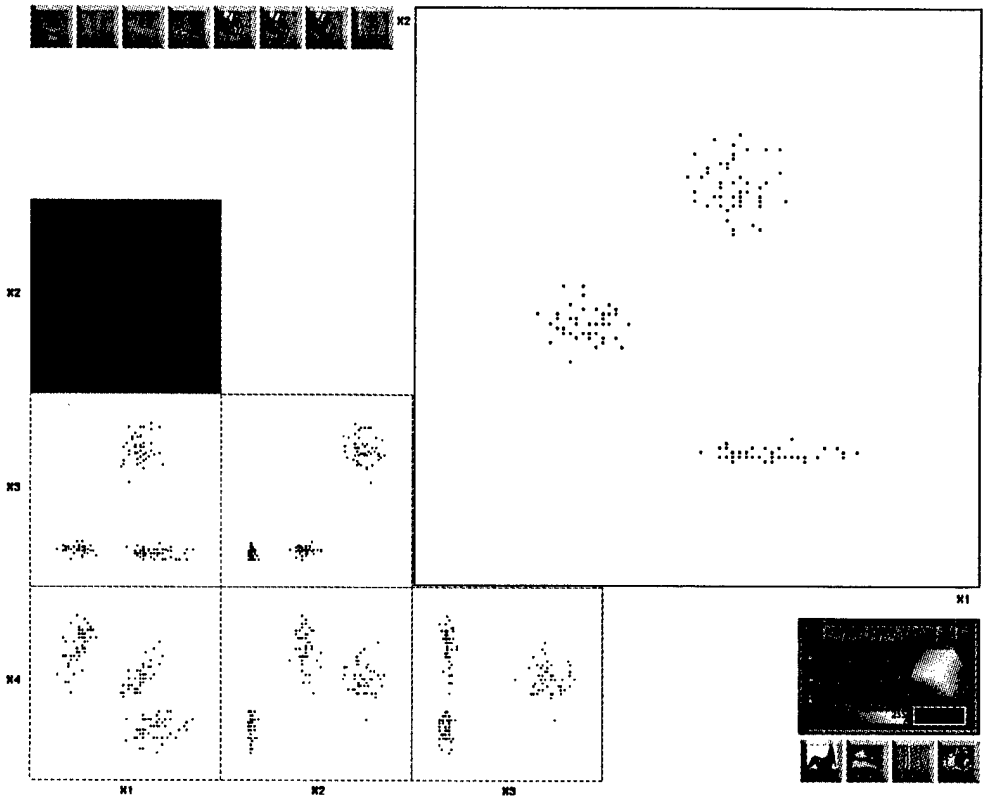
Figure 4. The desaturated parallel coordinate plot of the 3848 observations this time plotted on a black background. Notice the internal structure and the x-ray like appearance of this density plot.

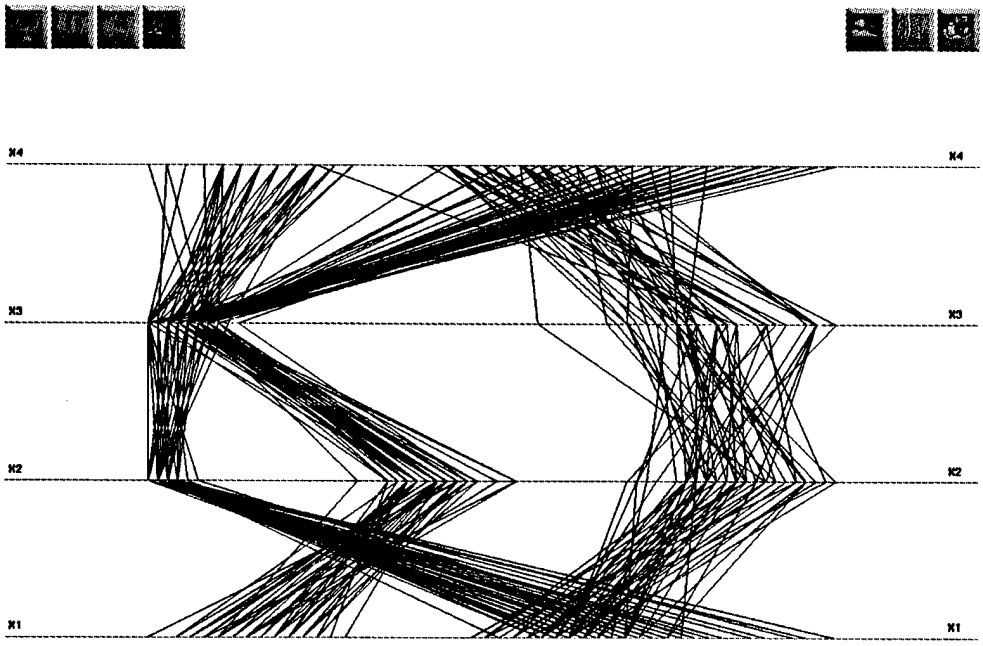
Figure 5. An intermediate parallel coordinate plot pruned to remove observations away from the internal structure. The plot is rescaled to fill the same scale as in Figure 4.

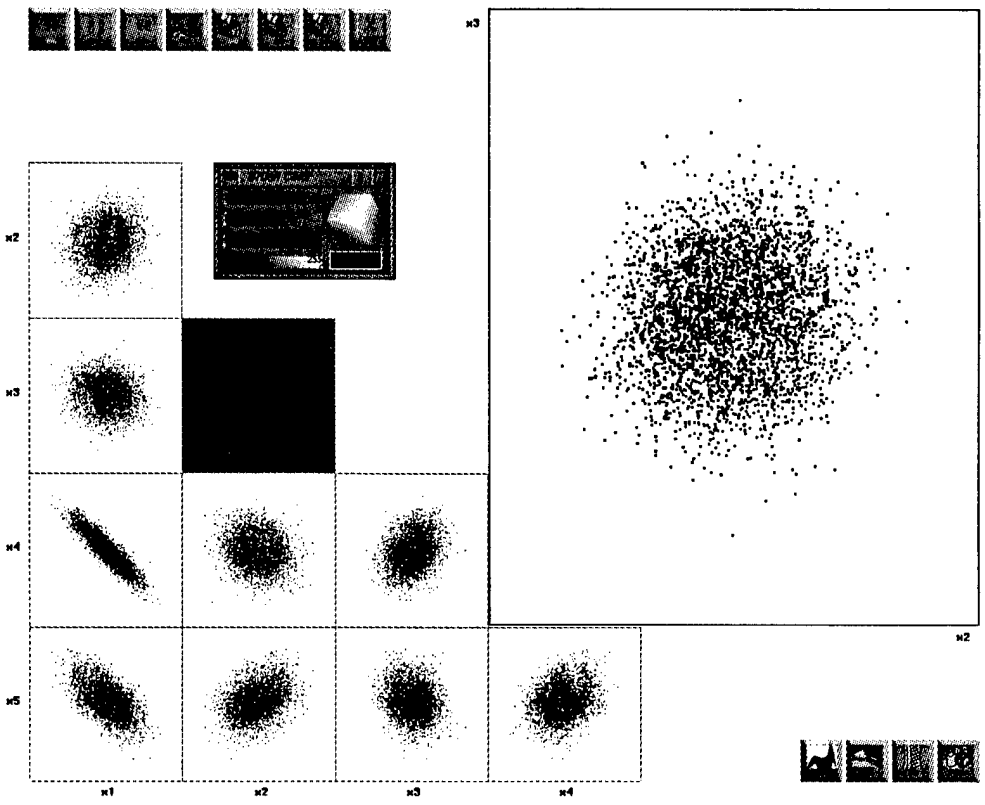
Figure 6. The final pruned parallel coordinate plot with all observations removed except those corresponding to the internal structure. The plot is again rescaled. The five gaps on axes two and three are suggestive of six clusters.

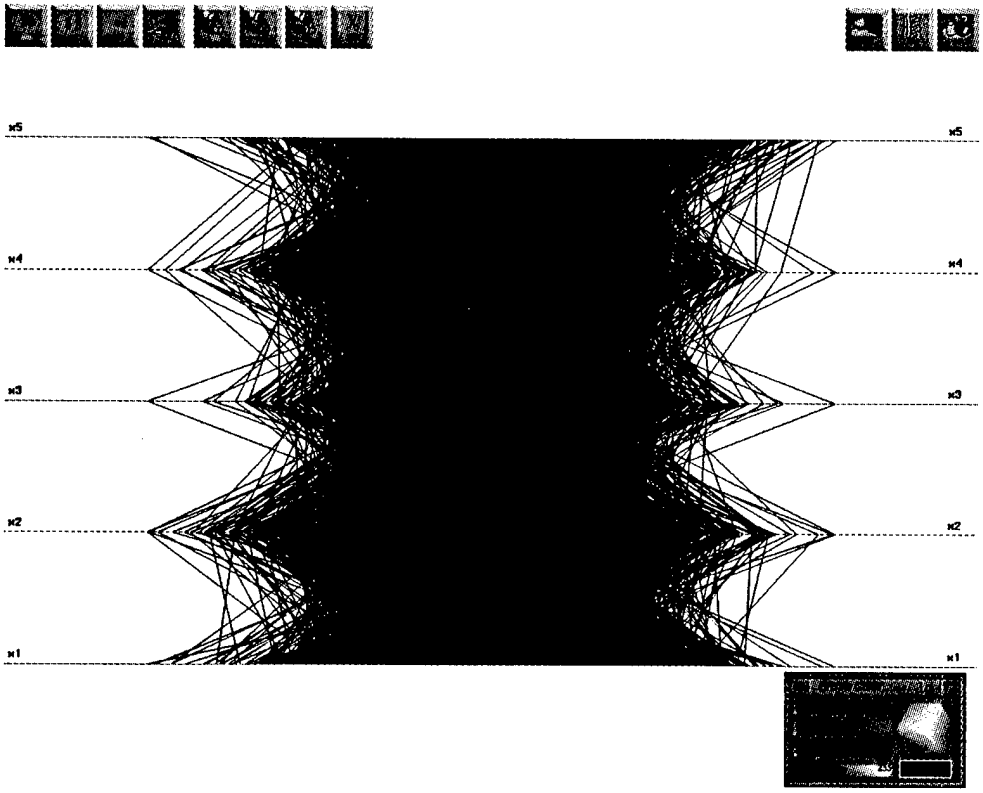
Figure 7. The result of a grand tour rotation of the data in Figure 6. The rotation confirms that these are six clusters completely separable in at least three of the five dimensions.

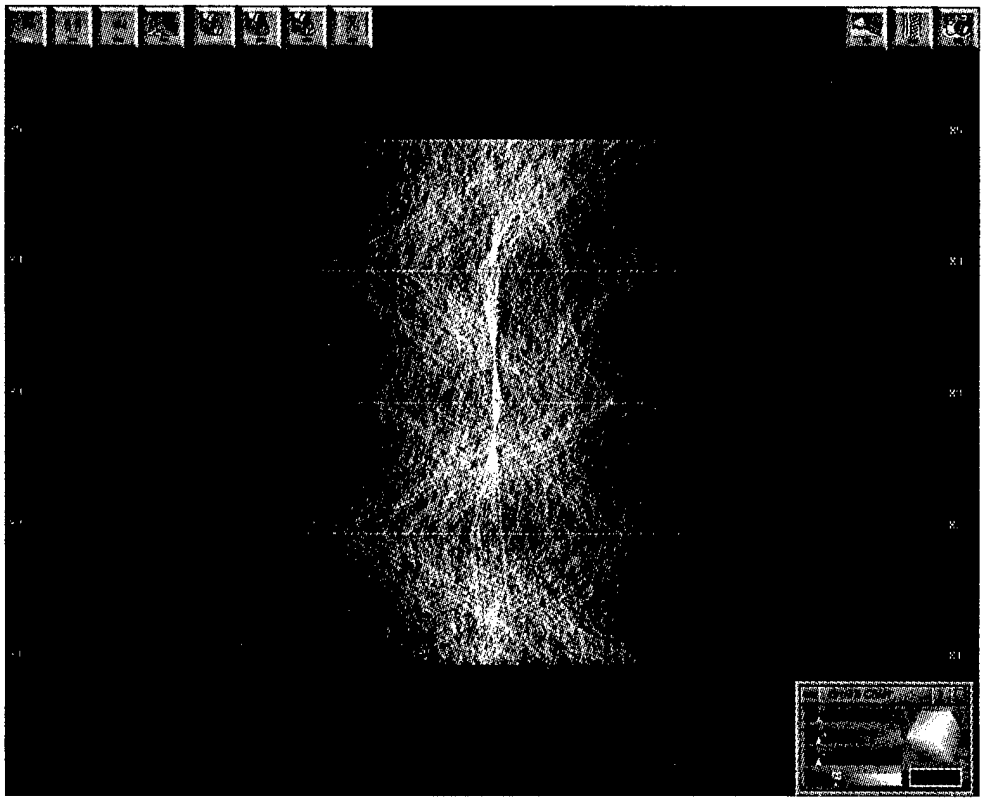
Figure 8. The result of plotting the data isolated in the parallel coordinate display back into the scatterplot matrix. It is now apparent that the six clusters for the letters E U R E K A. The six letters are made up of 99 points of the 3848 in the original data set, less than 2.7% of the total observations.

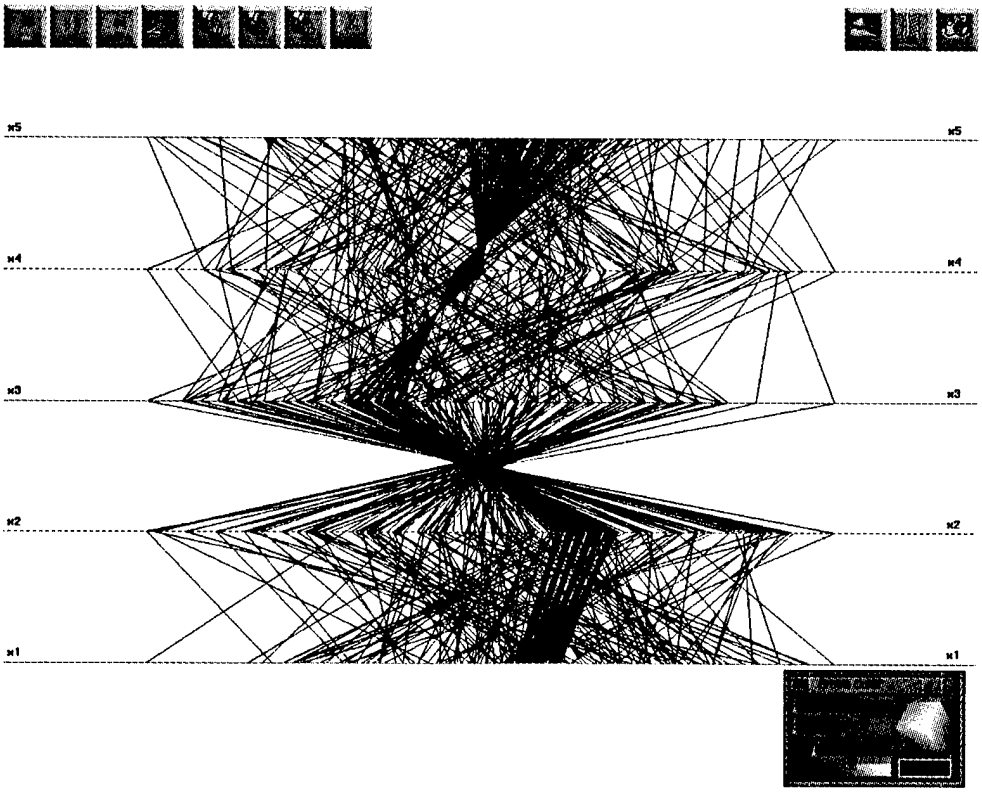

X2
X3
X4
X1
X2
X3
X1

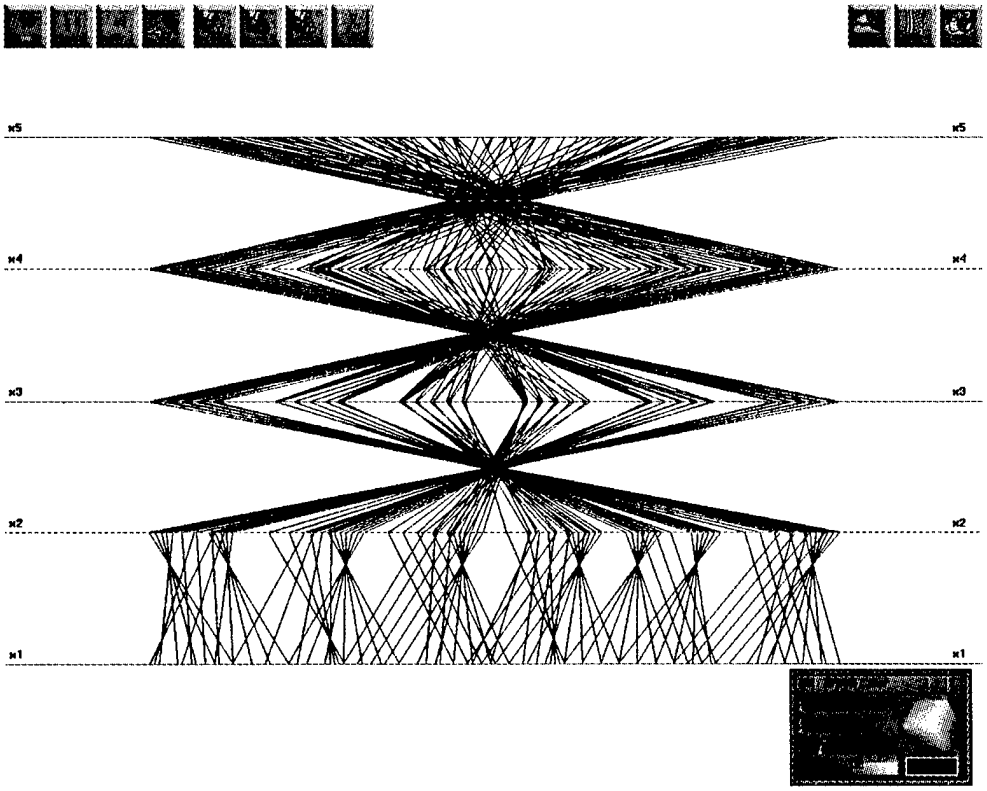

X4
X3
X2
X1
X4
X3
X2
X1

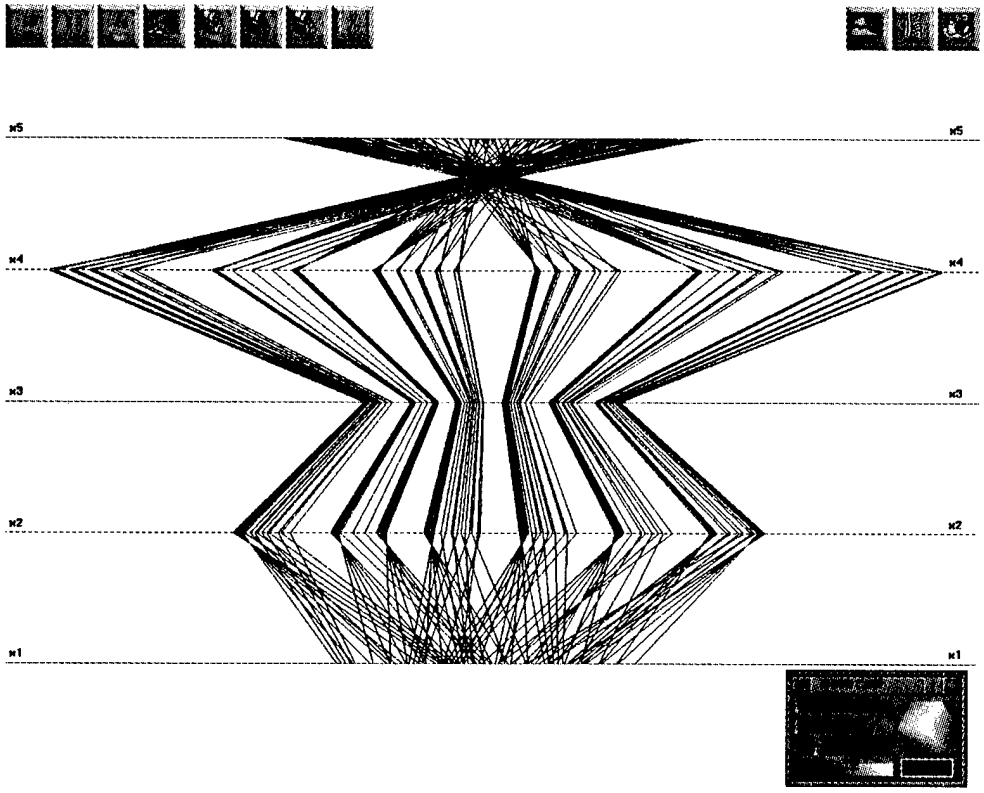

x2
x3
x4
x5
x1
x2
x3
x4
x3
x2

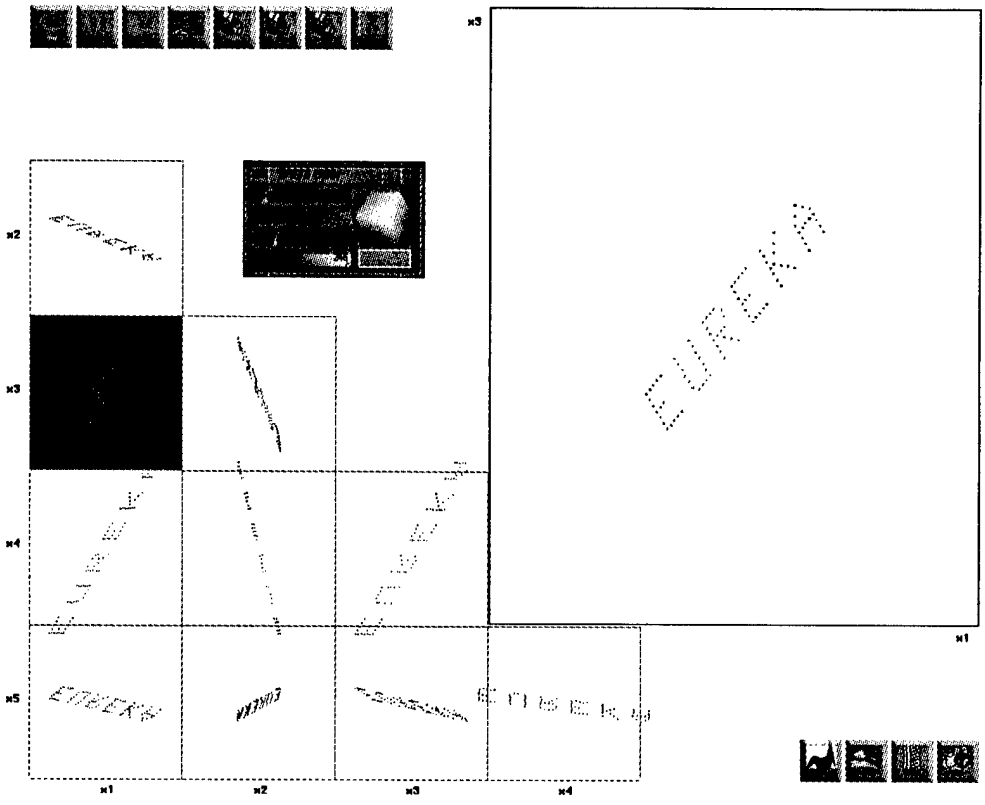

x5
x4
x3
x2
x1
x5
x4
x3
x2
x1




x5
x4
x3
x2
x1
x5
x4
x3
x2
x1


к5
к4
к3
к2
к1
к5
к4
к3
к2
к1


x5
x4
x3
x2
x1
x5
x4
x3
x2
x1


EUREKA

SF 298 MASTER COPY KEEP THIS COPY FOR REPRODUCTION PURPOSES

REPORT DOCUMENTATION PAGE

Form Approved OMB NO. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	April 1996	Technical

4. TITLE AND SUBTITLE
High Dimensional Clustering Using Parallel Coordinates and the Grand Tour

5. FUNDING NUMBERS
DAAH04-94-G-0267

6. AUTHOR(S)
Edward J. Wegman and Qiang Luo

7. PERFORMING ORGANIZATION NAMES(S) AND ADDRESS(ES)
Center for Computational Statistics
George Mason University
Fairfax, VA 1996

8. PERFORMING ORGANIZATION REPORT NUMBER
#124

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSORING / MONITORING AGENCY REPORT NUMBER
ARO 32850.11-MA

11. SUPPLEMENTARY NOTES
The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision, unless so designated by other documentation.

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution unlimited.

12 b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)* In this paper, we present some graphical techniques for cluster analysis of high-dimensional data. Parallel coordinate plots and parallel coordinate density plots are graphical techniques which map multivariate data into a two-dimensional display. The method has some elegant duality properties with ordinary Cartesian plots so that higher-dimensional mathematical structures can be analyzed. Our high interaction software allows for rapid editing of data to remove outliers and isolate clusters by brushing. Our brushing techniques allow not only for hue adjustment, but also for saturation adjustment. Saturation adjustment allows for the handling of comparatively massive data sets by using the alpha-channel of the Silicon Graphics workstation to compensate for heavy overplotting.

The grand tour is a generalized rotation of coordinate axes in a high-dimensional space. Coupled with the full-dimensional plots allowed by the parallel coordinate display, these techniques allow the data analyst to explore to explore data which is both high-dimensional and massive in size. In this paper we give a description of both techniques and illustrate their use to do inverse regression and clustering. We have used these techniques to analyze data on the order of 250,000 observations in 8 dimensions. Because the analysis requires the use of color graphics, in the present paper we illustrate the methods with a more modest data set of 3848 observations. Other illustrations are available on our web page.

14. SUBJECT TERMS	15. NUMBER IF PAGES
data analysis,saturation brushing, cluster analysis, graphical decision rules	21
	16. PRICE CODE

17. SECURITY CLASSIFICATION OR REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UL

GENERAL INSTRUCTIONS FOR COMPLETING SF 298

The Report Documentation Page (RDP) is used in announcing and cataloging reports. It is important that this information be consistent with the rest of the report, particularly the cover and title page. instructions for filling in each block of the form follow. It is important to ***stay within the lines*** to meet ***optical scanning requirements.***

Block 1. Agency Use Only *(Leave blank)*

Block 2. Report Date. Full publication date including day, month, and year, if available (e.g. 1 Jan 88). Must cite at least year.

Block 3. Type of Report and Dates Covered. State whether report is interim, final, etc. If applicable, enter inclusive report dates (e.g. 10 Jun 87 - 30 Jun 88).

Block 4. Title and Subtitle. A title is taken from the part of the report that provides the most meaningful and complete information. When a report is prepared in more than one volume, repeat the primary title, add volume number, and include subtitle for the specific volume. On classified documents enter the title classification in parentheses.

Block 5. Funding Numbers. To include contract and grant numbers; may include program element number(s), project number(s), task number(s), and work unit number(s). Use the following labels:

C	- Contract	**PR**	- Project
G	- Grant	**TA**	- Task
PE	- Program Element	**WU**	- Work Unit Accession No.

Block 6. Author(s). Name(s) of person(s) responsible for writing the report, performing the research, or credited with the content of the report. If editor or compiler, this should follow the name(s).

Block 7. Performing Organization Name(s) and Address(es). Self-explanatory.

Block 8. Performing Organization Report Number. Enter the unique alphanumeric report number(s) assigned by the organization performing the report.

Block 9. Sponsoring/Monitoring Agency Name(s) and Address(es). Self-explanatory.

Block 10. Sponsoring/Monitoring Agency Report Number. *(If known)*

Block 11. Supplementary Notes. Enter information not included elsewhere such as; prepared in cooperation with...; Trans. of...; To be published in.... When a report is revised, include a statement whether the new report supersedes or supplements the older report.

Block 12a. Distribution/Availability Statement. Denotes public availability or limitations. Cite any availability to the public. Enter additional limitations or special markings in all capitals (e.g. NORFORN, REL, ITAR).

DOD	- See DoDD 4230.25, "Distribution Statements on Technical Documents."
DOE	- See authorities.
NASA	- See Handbook NHB 2200.2.
NTIS	- Leave blank.

Block 12b. Distribution Code.

DOD	- Leave blank
DOE	- Enter DOE distribution categories from the Standard Distribution for Unclassified Scientific and Technical Reports
NASA	- Leave blank.
NTIS	- Leave blank.

Block 13. Abstract. Include a brief *(Maximum 200 words)* factual summary of the most significant information contained in the report.

Block 14. Subject Terms. Keywords or phrases identifying major subjects in the report.

Block 15. Number of Pages. Enter the total number of pages.

Block 16. Price Code. Enter appropriate price code *(NTIS only).*

Block 17. - 19. Security Classifications. Self-explanatory. Enter U.S. Security Classification in accordance with U.S. Security Regulations (i.e., UNCLASSIFIED). If form contains classified information, stamp classification on the top and bottom of the page.

Block 20. Limitation of Abstract. This block must be completed to assign a limitation to the abstract. Enter either UL (unlimited) or SAR (same as report). An entry in this block is necessary if the abstract is to be limited. If blank, the abstract is assumed to be unlimited.