High-Dimensional Visual Analytics: Interactive Exploration Guided by Pairwise Views of Point Distributions[*]

Leland Wilkinson[†]
SPSS Inc.
Northwestern University

Anushka Anand[‡]
University of Illinois at Chicago

Robert Grossman[§]
University of Illinois at Chicago

Abstract

We introduce a method for organizing high-dimensional multivariate displays and for guiding interactive exploration through high-dimensional data.

CR Categories: H.5.2 [User Interfaces]: Graphical User Interfaces—Visualization; I.3.6 [Computing Methodologies]: Computer Graphics—Methodology and Techniques;

Keywords: visualization, statistical graphics

1 Introduction

Visualization is a centuries-old field. Visual analytics is relatively new. What distinguishes visual analytics from ordinary visualization is the active role played by the computer in the presentation of information to the viewer. We have, for the first time, a marriage of *analytic* statistical algorithms and *visual* presentation in an interactive environment. Before visual analytics, exploring high-dimensional data with widgets like rotation controls, slice-and-dice tools, filter sliders, lensing tools, and real-time brushes was a haphazard enterprise. Exploring raw high-dimensional data with such tools (an idea introduced by John Tukey) necessarily falls prey to the curse of dimensionality.

By contrast, visual analytics offers the prospect of *guided* exploration. Given interactive tools and underlying analytic components, a user can explore views of high-dimensional data that are highlighted by statistical algorithms. The result is the blending of the strengths of each approach: the analytic spotlight of statistical models, and the inferential floodlight of visual exploration.

We need visual analytics for three principal purposes.

- *Checking raw data for anomalies*. Anomalies in raw data include outliers caused by coding errors, sensor malfunctions, extreme environmental conditions, and other factors. Anomalies also include missing values, which may occur randomly or deterministically. And anomalies may include biases due to response sets, ceiling and floor effects, and history and maturation effects. These biases can affect the shape of distributions assumed in data analysis.

- *Exploring data to discover plausible models*. We call this exploratory data analysis (EDA), a term invented by John Tukey [46]. EDA is not a fishing expedition. We explore data with expectations. We revise our expectations based on what we see in the data. And we iterate this process.

- *Checking model assumptions*. Checking model assumptions requires plotting residuals and other diagnostic measures. These plots are often specialized, in order to asses distributional assumptions.

All three of these tasks are well-documented in the statistical literature. Outliers, missing data, and other anomalies are covered in [7][6][37][29]. EDA is discussed in [46][48]. Model diagnostics are presented in [2][5][9][14].

These tasks are more difficult for high-dimensional data. To explore a point distribution or diagnose the fit of a multivariate model, we need to find ways to characterize the distribution. Our characterization cannot be so constrained (as in, for example, a single low-dimensional projection like principal components or MDS) that we cannot find anomalies or detect local structure. And our characterization cannot be so general (as in parallel coordinates or scatterplot matrices) as to overwhelm us with complexity.

We have chosen instead to characterize a point distribution by constructing pairwise orthogonal views of points and then characterizing the points in these views. Pairwise views polynomially expand with dimensions, but if we choose our characterizations well, we can examine them efficiently enough to get a rough picture of the overall distribution.

The analysis begins with an $n \times p$ data matrix X. Pairs of columns of this matrix are mapped to a $p(p-1)/2 \times q$ matrix F that contains our q 2D point cloud characterizations. We call F a *feature* matrix. We then analyze this feature matrix for structure that can be mapped back to the domain of the original data matrix X.

Characterization is a vague term. We narrow it by adding constraints:

- We want to distinguish many types of point distributions: multivariate normal, lognormal, multinomial, sparse, dense, convex, clustered, and so on.

- We want a small number of feature measures that characterize these distributions.

- We want our feature measures on a common scale so we can compare them.

- We want our feature measures themselves to have comparable distributions.

- We want our feature measures to be computable efficiently so they are scalable to large numbers of points and dimensions.

These constraints force us to construct a minimal set of features that not only work well in the pairwise domain, but also lead to discovery of structure in the high-dimensional domain.

2 Related work

Our visual analytics depend on several areas of prior research. The first concerns *projection*, which reduces dimensionality in vector

[*] Graham Wills, Dongping Fang, and John Bauer provided valuable suggestions.

[†] e-mail: leland@spss.com

[‡] e-mail:aanand2@uic.edu

[§] e-mail:grossman@cs.uic.edu

or projective spaces in order to reveal structure. The second involves *geometric graphs*, which use graph-theoretic concepts to reveal structure in point distributions. The third concerns *organizing by features*, which structures multivariate displays according to selected features.

2.1 Projections

Visualization researchers have made extensive use of projections in order to represent high-dimensional structures in low-dimensional space. Simple methods involve linear maps, such as orthogonal pairwise projections or principal components. Other methods, such as multidimensional scaling, involve nonlinear maps. Furnas and Buja [19] discuss the characteristics of projective maps frequently used in visualization.

Finding informative projections for high-dimensional point sets is nontrivial. Friedman and Tukey [16] developed a class of loss functions and an iterative method to pursue projections that reveal selected structures (clustering, simplices, etc.) in point clouds. Asimov [4] devised a smooth path through high-dimensional space that defined a set of low-dimensional projections orthogonal to the path, so that viewers could observe an animated "grand tour" through the space. Roweis and Saul [38] modified the distance metric used in MDS in order to identify relatively compact sets of points on a manifold in low-dimensional space. In general, one has to have some prior knowledge of the characteristics of the high-dimensional point set in order to identify a relatively meaningful projection.

Furnas [18] developed an intriguing form of indirection in a projective mapping that resembles our use of characterizing measures. Furnas mapped a distance matrix to a single point in a low-dimensional space. This transformation enabled him to represent a family of distance matrices as a cloud of points. In his words, "Pictures of the cloud form a family portrait, and its characteristic shape and interrelationship with the portraits of other families can be explored."

2.2 Geometric Graphs

A geometric graph is an embedding of a vertex-edge graph in a metric space. These graphs have received a lot of attention recently because of their ability to characterize point sets embedded in high-dimensional spaces. The manifold learning community [38][8][10] has used k-nearest-neighbor graphs to find informative low-dimensional projections. We use geometric graph-theoretic measures for similar reasons. They are efficient to compute and they carry a lot of information about the configuration of points on which they are based. Marchette [31] surveys the use of these graphs for feature detection.

2.3 Organizing by Features

Many have noted that clustering and/or sorting multivariate displays by selected features can increase the interpretability of these displays [17][30][34][40]. Wilkinson [49] discusses much of the research on organizing multivariate displays.

One organizing principle has especially influenced the work in this paper. Around 20 years ago, John and Paul Tukey developed an exploratory visualization method called *scagnostics*. While they briefly mentioned their invention in [47], the specifics of the method were never published. Paul Tukey did offer more detail at an Institute for Mathematics and its Applications visualization workshop a few years later, but he did not include the talk in the workshop volume he and Andreas Buja edited [11].

We summarize the Tukeys' approach here, based on the first author's recollection of the IMA workshop and subsequent conversations with Paul Tukey. The Tukeys proposed to characterize a large number of 2D scatterplots through a small number of measures of the distribution of points in these plots. These measures included the area of the peeled convex hull [45], the perimeter length of this hull, the area of closed 2D kernel density isolevel contours [41][39], the perimiter length of these contours, the convexity of these contours, a modality measure of the 2D kernel densities, a nonlinearity measure based on principal curves [25] fitted to the 2D scatterplots, and several others. By using these measures, the Tukeys aimed to detect anomalies in density, shape, association, and other features in the 2D scatterplots.

After calculating these measures, the Tukeys constructed a scatterplot matrix [22][13] of the measures. With brushing and linking tools, the Tukeys proposed to identify unusual scatterplots. Wilkinson, Anand, and Grossman [50] discuss the method in more detail.

There are two aspects of the Tukeys' approach that can be improved. First, some of the Tukeys' measures, particularly those based on kernels, presume an underlying continuous empirical or theoretical probability function. This is appropriate for scatters sampled from continuous distributions, but it can be a problem for other types of data. Second, the computational complexity of some of the Tukey measures is $O(n^3)$. Since n was expected to be small for most statistical applications of this method, such complexity was not expected to be a problem.

We can ameliorate both these problems by using graph-theoretic measures. Indeed, the Tukeys used a few themselves. First, the graph-theoretic measures we will use do not presume a connected plane of support. They can be metric over discrete spaces. Second, the measures we will use are $O(n\log(n))$ in the number of points because they are based on subsets of the Delaunay triangulation. Third, we employ adaptive hexagon binning [12] before computing our graphs to further reduce the dependence on n.

There is a price for switching to graph-theoretic measures, however. They are highly influenced by outliers and singletons. Whenever practical, the Tukeys used robust statistical estimators to downweight the influence of outliers. We follow their example by working with nonparametric and robust measures. Further, we remove outlying points before computing our graphs and the measures based on them.

We next introduce and define the graphs we use as bases for our feature measures and then we discuss the measures themselves.

3 Computing Features

Our feature measures depend on geometric graphs.

3.1 Geometric Graphs

A *graph* $G = (V,E)$ is a set V (called *vertices*) together with a relation on V induced by a set E (called *edges*). An edge $e(v,w)$, with $e \in E$ and $v,w \in V$, is a pair of vertices. A *geometric graph* $G^\star = [f(V), g(E), S]$ is an embedding of a graph in a metric space S that maps vertices to points and edges to line segments connecting pairs of points. We will omit the asterisk in the rest of this paper and assume all our graphs are geometric. We will also restrict our candidates to geometric graphs that are:

- *undirected* (edges consist of unordered pairs)
- *simple* (no edge pairs a vertex with itself)
- *planar* (there is an embedding in $\mathbf{R}^2$ with no crossed edges)
- *straight* (embedded edges are straight line segments)
- *finite* (V and E are finite sets)

Figure 1 shows instances of the geometric graphs on which we compute our measures. The points are taken from a dataset in [32]. In this section, we define the geometric graphs that are the bases for our measures.

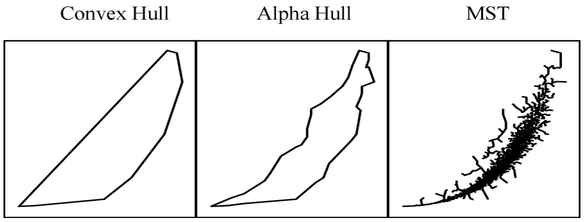

Figure 1: Graphs used as bases for computing scagnostics measures

3.1.1 Convex Hull

A *hull* of a set of points X in $\mathbf{R}^2$ is a collection of the boundaries of one or more polygons that have a subset of the points in X for their vertices and that collectively contain all the points in X. This definition includes entities that range from the boundary of a single polygon to a collection of boundaries of polygons each consisting of a single point. A hull is *convex* if it contains all the straight line segments connecting any pair of points in its interior.

There are several algorithms for computing the convex hull [42]. Since the convex hull consists of the outer edges of the Delaunay triangulation, we can use an algorithm for the Voronoi/Delaunay problem and then pick the outer edges. Its computation thus can be $O(n\log(n))$. We will use the convex hull, together with other graphs, to construct measures of convexity.

3.1.2 Nonconvex Hull (Alpha Shape)

A *nonconvex hull* is a hull that is not the convex hull. This class includes simple shapes like a *star convex* or *monotone convex* hull [3], but it also includes some space-filling, snaky objects and some that have disjoint parts. In short, we are interested in a general class of nonconvex shapes.

There have been several geometric graphs proposed for representing the nonconvex "shape" of a set of points on the plane. Most of these are proximity graphs [27]. A *proximity graph* (or *neighborhood graph*) is a geometric graph whose edges are determined by an indicator function based on distances between a given set of points in a metric space. To define this indicator function, we use an open disk D. We say D *touches* a point if that point is on the boundary of D. We say D *contains* a point if that point is in D. We call an open disk of fixed radius $D(r)$.

In an *alpha shape graph* [15], an edge exists between any pair of points that can be touched by an open disk $D(\alpha)$ containing no points. The alpha shape is relatively efficient to compute because it is a subset of the Delaunay triangulation with a simple inclusion criterion. Marchette [31] recommends a value of α to be the average value of the edge lengths in the minimum spanning tree. To reduce noise, we have chosen a larger value, namely, the 90th percentile of these edge lengths. We clamp this value at one-tenth the width of a frame if the percentile exceeds a tenth. This prevents us from including sparse or striated point sets in a single alpha graph.

3.1.3 Minimum Spanning Tree

A *path* is a list of successively adjacent, distinct edges. A *tree* is a graph in which any two nodes are connected by exactly one path. A *spanning tree* is an undirected graph whose edges are structured as a tree. A *minimum spanning tree* (MST) is a spanning tree whose total length (sum of edge weights) is least of all spanning trees on a given set of points [28]. The edge weights of a geometric MST are computed from distances between its vertices.

The MST is a subgraph of the Delaunay triangulation. There are several efficient algorithms for computing an MST for a set of points in the plane [33], [36].

We will now discuss the feature measures computed on these three graphs. In the formulas below, we use H for the convex hull, A for the alpha hull, and T for the minimum spanning tree. In our feature calculations, we ignore outliers (except for the outlier measure).

3.2 Feature Measures

We are interested in assessing four aspects of scattered points: *outliers*, *density*, *shape*, and *association*. Our measures are derived from several features of geometric graphs:

- The length of an edge, $length(e)$, is the Euclidean distance between its vertices.
- The length of a graph, $length(G)$, is the sum of the lengths of its edges.
- A *path* is a list of vertices such that all pairs of adjacent vertices in the list are edges.
- A path is *closed* if its first and last vertex are the same.
- A closed path is the boundary of a *polygon*.
- The perimeter of a polygon, $perimeter(P)$, is the length of its boundary.
- The area of a polygon, $area(P)$ is the area of its interior.

All our measures are defined to be in the closed unit interval. To compute them, we assume our variables are scaled to the closed unit interval as well.

3.2.1 Outliers

Tukey [45] introduced the use of the peeled convex hull as a measure of the *depth* of a level set imposed on scattered points. For points on the 1D line, this amounts to successive symmetric trimming of extreme observations. Tukey's idea can be used as an outlier identification procedure. We compute the convex hull, delete points on the hull, compute the convex hull on the remaining points, and continue until (one hopes) the contours of successive hulls do not substantially differ.

We have taken a different approach. Because we do not assume that our point sets are convex (that is, comparably dense in all subregions of the convex hull), we cannot expect outliers will be on the edges of a convex hull. They may be located in interior, relatively empty regions. Consequently, we have chosen to peel the MST instead of the hull. We consider an outlier to be a vertex whose adjacent edges in the MST all have a weight greater than ω.

There are theoretical results on the distribution of the largest edge for an MST on normally distributed data [35], but we decided to work with a nonparametric criterion for simplicity. Following Tukey [46], we choose

$$\omega = q_{75} + 1.5(q_{75} - q_{25})$$

where q_{75} is the 75th percentile of the MST edge lengths and the expression in the parentheses is the *interquartile range* of the edge lengths.

- Outlying

This is a measure of the proportion of the total edge length due to extremely long edges connected to points of single degree.

$$c_{outlying} = length(T_{outliers})/length(T)$$

3.2.2 Density

The following indices detect different distributions of points.

- Skewed

The distribution of edge lengths of a minimum spanning tree gives us information about the relative density of points in a scattered configuration. Some have used the sample mean, variance, and skewness statistics to summarize this edge length distribution [1]. However, theoretical results [43][35] show that the MST edge-length distribution for many types of point scatters can be approximated by an extreme value distribution with fewer parameters. We have chosen two measures of relative density. The first is a relatively robust measure of skewness in the distribution of edge lengths.

$$c_{skew} = (q_{90} - q_{50})/(q_{90} - q_{10})$$

We will discuss the second measure in the Sparse section below.

- Clumpy

An extremely skewed distribution of MST edge lengths does not necessarily indicate clustering of points. For this, we turn to another measure based on the MST: the Hartigan and Mohanty RUNT statistic [23]. This statistic is most easily understood in terms of the single-linkage hierarchical clustering tree called a *dendrogram*. The runt size of a dendrogram node is the smaller of the number of leaves of each of the two subtrees joined at that node. Since there is an isomorphism between a single-linkage dendrogram and the MST [21], we can associate a runt size (r_j) with each edge (e_j) in the MST, as described by Stuetzle [44]. The runt graph (R_j) corresponding to each edge is the smaller of the two subsets of edges that are still connected to each of the two vertices in e_j after deleting edges in the MST with lengths less than $length(e_j)$.

Our runt-based measure emphasizes clusters with small intra-cluster distances relative to the length of their connecting edge and ignores runt clusters with relatively small runt size.

$$c_{clumpy} = \max_j \left[1 - \max_k \left[length(e_k)\right] / length(e_j) \right]$$

- Sparse

Our sparseness statistic measures whether points in a 2D scatterplot are confined to a lattice or a small number of locations on the plane. This can happen, for example, when tuples are produced by the product of categorical variables. It can also happen when the number of points is extremely small. We choose the 90th percentile of the distribution of edge lengths in the MST. This was the same value we chose for the α statistic.

$$c_{sparse} = q_{90}$$

In the extremely rare event that this statistic exceeds unity (*e.g.*, when all points fall on either of the two diagonally opposing vertices of a square), we clamp the value to 1.

- Striated

We define coherence in a set of points as the presence of relatively smooth paths in the minimum spanning tree. Smooth algebraic functions, time series, and curves (*e.g.*, spirals) fit this definition. So do points arranged in flows or vector fields. In the examples in this paper, we will see a common striated pattern: parallel strips of points produced by the product of categorical and continuous variables.

We could recognize parallel lines with a Hough transform [26]. Other configurations of points that represent vector flows or striated textures might not follow linear paths, however. We have devised a more general measure. It is based on the number of adjacent edges whose cosine is less than -0.75. Let $V^{(2)} \subseteq V$ be the set of all vertices of degree 2 in V and let $I()$ be an indicator function. Then

$$c_{striate} = \frac{1}{|V|} \sum_{v \in V^{(2)}} I(\cos\theta_{e(v,a)e(v,b)} < -.75)$$

3.2.3 Shape

The shape of a set of scattered points is our next consideration. We are interested in both topological and geometric aspects of shape. We want to know, for example, whether a set of scattered points on the plane appears to be connected, convex, inflated, and so forth. Of course, scattered points are by definition *not* these things, so we are going to need additional machinery (based on our graphs that we fit to these points) to allow us to make such inferences. The measures that we propose will be based on these graphs.

- Convex

This is the ratio of the area of the alpha hull and the area of the convex hull. This ratio will be 1 if the nonconvex hull and the convex hull have identical areas.

$$c_{convex} = area(A)/area(H)$$

- Skinny

The ratio of perimeter to area of a polygon measures, roughly, how skinny it is. We use a corrected and normalized ratio so that a circle yields a value of 0, a square yields 0.12 and a skinny polygon yields a value near one.

$$c_{skinny} = 1 - \sqrt{4\pi area(A)}/perimeter(A)$$

- Stringy

A stringy shape is a skinny shape with no branches. We count vertices of degree 2 and compare them to the overall number of vertices minus the number of single-degree vertices.

$$c_{stringy} = \frac{|V^{(2)}|}{|V| - |V^{(1)}|}$$

3.2.4 Association

The following index helps reveal whether a given scatter is monotonic.

- Monotonic

We have chosen the squared Spearman correlation coefficient, which is a Pearson correlation on the ranks of x and y (corrected for ties), to assess monotonicity in a scatterplot. We square the coefficient to accentuate the large values and to remove the distinction between negative and positive coefficients. We assume investigators are most interested in strong relationships, whether negative or positive.

$$c_{monotonic} = r^2{}_{spearman}$$

This is our only coefficient not based on a subset of the Delaunay graph. Because it requires a sort, its computation is $O(n\log(n))$.

3.3 Binning

We use hexagon binning [12] to improve performance. We begin with a 40 by 40 hexagon grid for each scatterplot. If there are more than 250 nonempty cells, we reduce the bin size by half and rebin. We continue this process until there are no more than 250 nonempty cells.

We examined by Monte Carlo simulation the effect of this adaptive binning on our measures. Three of them – Skewed, Sparse, and Convex– showed a slight binning effect (within .1 in magnitude). We therefore applied a correction factor to these measures:

$$w = .7 + \frac{.3}{1+t^2}$$

where $t = n/500$. Because Skewed tends to *decrease* with n after adaptive binning, we use the transformation

$$1 - w(1 - c_{skew})$$

3.4 Performance

Because we use efficient binning and triangulation, computation time is $O(np^2)$. On a Macintosh G4 PowerBook running Java 1.4.2, computing the measures on 100,000 random cases distributed uniformly on 10 variables required approximately 10 seconds. Computing the measures on 100,000 cases and 25 variables required approximately 75 seconds. Computing on a microarray dataset with 400 cases and 62 highly correlated variables required approximately 425 seconds. Because the effect of sample size is practically negligible, our code can compute roughly four scatterplots per second on this machine.

4 Examples

We begin with a dataset comprising hourly meteorological measurements over a year at the Greenland Humboldt automatic weather station operated by NASA and NSF. These measurements are part of the Greenland Climate Network (GC-Net) sponsored by these federal agencies. The variables include time of year (in hours), low and high temperature, wind speed, humidity, ice temperature, snowfall, wind direction, atmospheric pressure, radiation, and incoming and reflected solar shortwave radiation.

4.1 Data SPLOM

Figure 2 shows a scatterplot matrix of the weather data. We have ordered the variables in the matrix according to the original ordering in the dataset.There are 120 scatterplots in this display. The SPLOM is small enough so we can view all of the scatterplots in a single glance, but organizing them into groups is relatively difficult. We notice that some of the variables appear to be categorical, because they display as stripes when plotted against other variables. We also notice that these variables do not appear to be normally distributed.

4.2 Features Plot

Figure 3 displays the scatterplots in Figure 2 ranked according to each of our features. Above each little plot are two red marks indicating the row and column indices of the location of the plot in the SPLOM pictured in Figure 3. In the interactive program, these plots are linked by mouse-click to the data SPLOM window so a user can navigate between the two.

We now see clearly groups of scatterplots that are similar to each other. This display ranks the top 10 scatterplots in each category, but filters out plots with feature measures less than 0.5. We see two plots which appear to contain a high number of outliers (the leftmost column). These involve snowfall and ice temperature. We are now suspicious that these measurements may be truncated. The Clumpy column picks up these and other variables showing similar behavior.

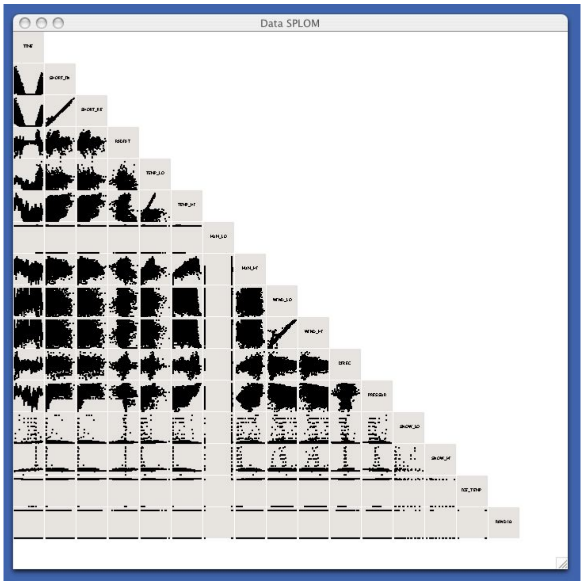

Figure 2: Scatterplot matrix of weather measurements at a Greenland station

We also see in the rightmost column of Figure 3 that two pairs of variables are highly correlated: low and high wind speed and incoming and reflected solar short-wave radiation. The pairs are not intercorrelated, however. The remaining plots in the Monotonicity column appear to be correlated because of measurement artifacts. We are reminded that these data may need considerable cleaning before formal statistical analysis.

4.3 Features SPLOM

The Tukeys originally proposed constructing a scagnostics SPLOM and working with that display to analyze patterns in the raw data SPLOM. We do this in Figure 4. The interactive program allows us to pick a single point in the Features SPLOM and to see the scatterplot that point represents.

We have picked the incoming and reflected short-wave radiation plot. Notice that the point representing this plot is highlighted in red in the Features SPLOM. We saw that this plot was characterized as highly correlated in the Features plot in Figure 3. By inspecting the red points in the Features SPLOM, we can also see that the point configuration is relatively skinny. We also note that, while it is highly monotonically correlated, it is not especially convex. The non-convexity is produced by unusual behavior at the lower left end of the plot. This is clearly not a bivariate normal distribution (which would appear as relatively convex in the Features SPLOM).

4.4 Data SPLOM Permuted by Features Component

Now we use the features to organize our displays. We have sorted the variables in the raw data SPLOM using the size of the loadings on the first principal component of the scagnostic measures. We compute our components on the $p(p-1)/2 \times 9$ features matrix F. Then we sum the component loadings over the indices of the corresponding variables in the original $n \times p$ data matrix X.

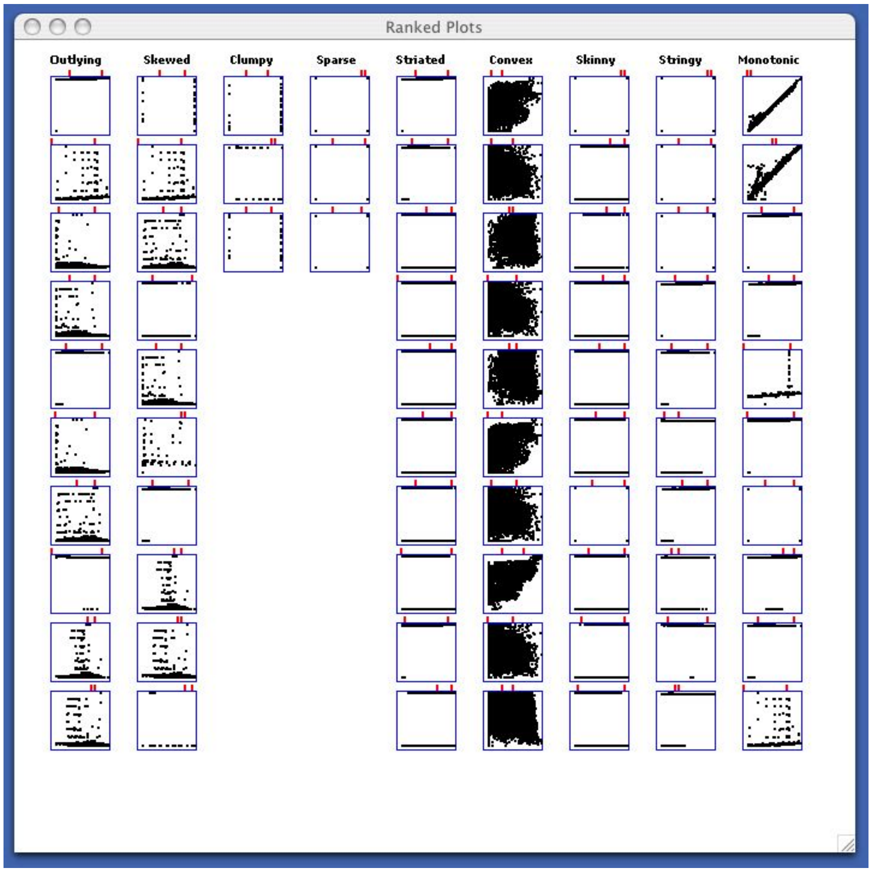

Figure 3: Scatterplots of weather data sorted by features

Figure 5 shows that the sorting segregates the apparently discrete and continuous variables and clusters similar marginal 2D distributions. There appear to be three clusters of plots: relatively dense at the top, relatively striped at the bottom, and mixed in the middle.

4.5 Clustering Features

Features sorting suggests clusters but it is not a clustering procedure. For that we need to cluster the features matrix F directly. A k-means clustering [24] of F yields four clusters. In Figure 6, we color the scatterplots according to the cluster identifiers. There is substantial agreement between the clustering and the layout of the features-sorted SPLOM. Further, we have chosen red for the color of the smallest cluster. This reveals several anomalous scatterplots that do not fit well into the other clusters.

4.6 Features Outliers

Our clustering has led us to consider the possibility of identifying outliers among the scatterplots. This is a second-order concept – namely, are there a few *scatterplots* (not points) that are significantly different from all the others? Answering this question presumes that there is a distribution of scatterplots based on our data and that there are scatterplot outliers from this distribution.

To answer the question, we use the same algorithm we devised for identifying outliers in scatterplots – the MST outlier statistic. Instead of applying it to a single scatterplot, however, we apply it to the multivariate distribution of scatterplot features. That is, we fit a minimum spanning tree to the features in 9 dimensions and we identify outliers in that 9-dimensional empirical distribution.

In Figure 7, we color the outlying scatterplots according to the MST scatterplot outlier statistic. We get a surprising result. The red scatterplots identified in the fourth cluster are not flagged as outliers by our statistic. Instead, several plots involving our suspicious variable ice temperature are flagged.

Inspecting the raw data, we find that ice temperature is not a binary variable. It contains a missing value code (-6999) that swamps all the other values. This artifact affects several of the other variables that appear to be binary in the plots. The scientific metadata accompanying these data would have alerted us to this problem, but we note that the features measures do so as well.

It is important to realize that our outlier procedure does not presume a particular parametric reference distribution. Standard outlier methods often presume multivariate normality, for example. Statisticians look for outliers from multivariate normals when testing assumptions for standard normal models such as linear discriminant analysis and multivariate analysis of variance.

Figure 8 shows, however, that our procedure can find that multivariate normals are outliers when embedded in a batch of real data. The data are statistics for baseball players in 2004. Five plots are identified as outliers by our MST statistic. One is a plot of players' height and weight, which appears to be bivariate normal. It is flagged as an outlier, however, because it is the only untruncated bivariate normal in the whole SPLOM. There are two other plots identified as outliers. One plots batting average against home run rate in an odd-shaped V formation.

Figure 9 shows clearly how a single plot can be identified as an outlier in even a small SPLOM. The data are EPA emissions tests for cars sold in the US. The variables are hydrocarbon emissions, carbon monoxide emissions, horsepower, gallons burned per mile,

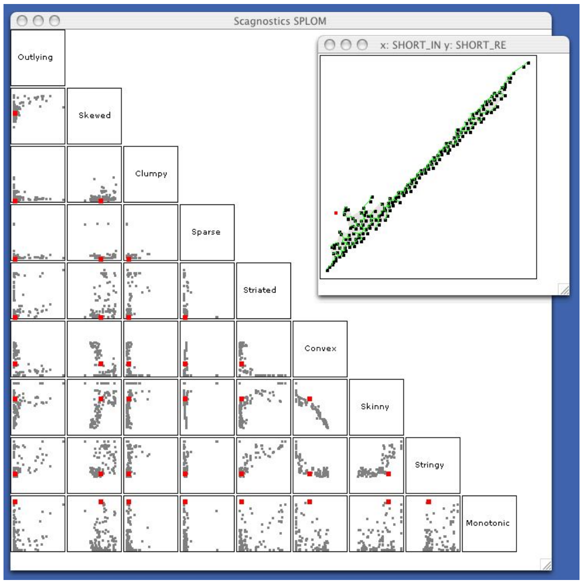

Figure 4: Features SPLOM of weather data

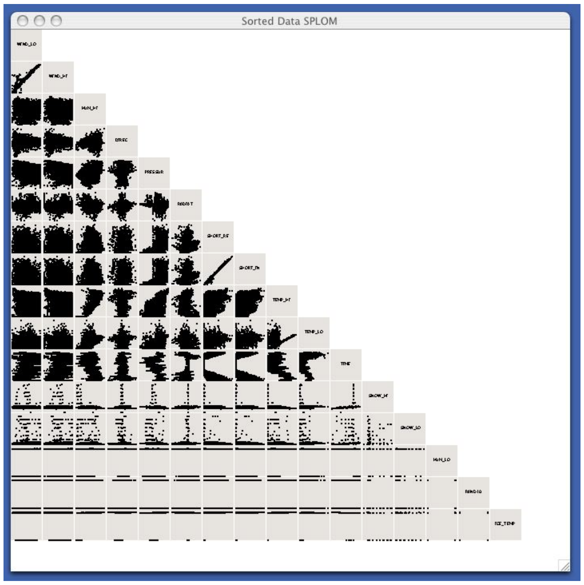

Figure 5: SPLOM of weather data sorted by features component

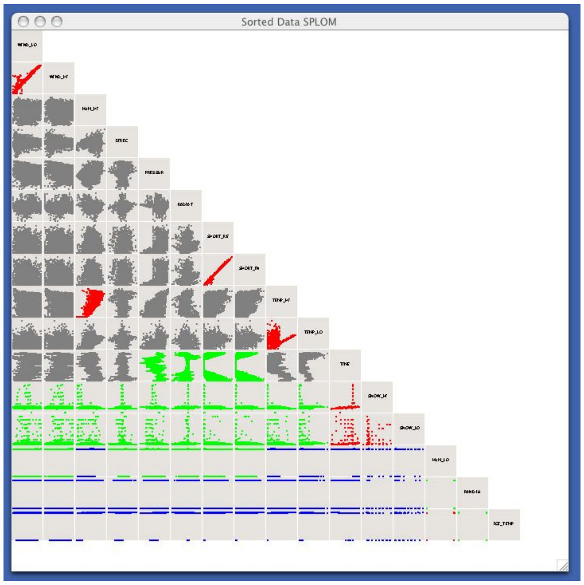

Figure 6: SPLOM of weather data sorted by features component and colored by cluster membership

and carbon dioxide emissions. The flagged red scatterplot clearly indicates the extremely high correlation between carbon dioxide emissions and fuel consumption – a central statistic in the global warming debate.

4.7 Other Multivariate Displays Sorted by Features

Finally, the following examples show that we must not assume that the features matrix F and our analytics based on it are designed only for scatterplot matrices. In Figure 10, we show a parallel coordinates plot of the weather data colored by time of year and sorted by the first principal component based on intercorrelations of the variables. Although this is probably the most prevalent sorting method used on multivariate displays, the sorting doesn't work very well because most of the correlations are near zero and the correlation structure is not particularly informative.

In Figure 11, we sort the parallel coordinates using our features component. The variables follow the same ordering we used in Figure 5. The gaps due to the sparse scatterplots are now pushed toward the top of the display and the parallel coordinate profiles are more coherent.

5 Conclusion

We conclude with a few observations based on our examples. First, we chose relatively small datasets so we could illustrate them in printed form. One should not assume that the data SPLOM shown in Figure 2 is always viewable, however. For many datasets (such as bioinformatics data), there are too many variables to display as a

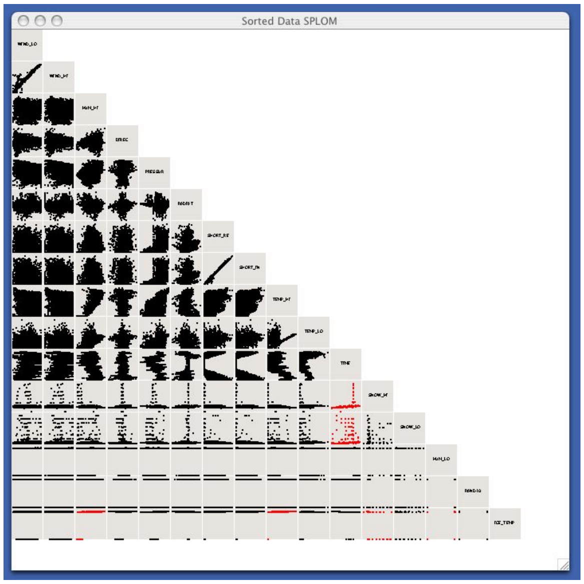

Figure 7: SPLOM of weather data sorted by features component and colored by MST outlier statistic

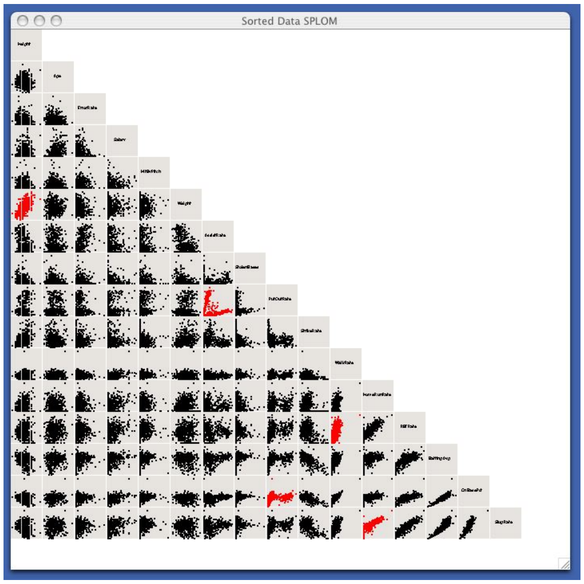

Figure 8: SPLOM of baseball data sorted by features component and colored by MST outlier statistic

SPLOM or in parallel coordinates form. Furthermore, lensing and filtering tools cannot be used effectively with these displays unless the variables are properly sorted. Therefore, we often must rely on features plots to drill into our data.

Second, we hope to have shown that sorting variables or scatterplots on simple statistics such as means, variances, or correlations will not work well on many real datasets. That is because many real datasets are not multinormally (or even "blobbily") distributed. They frequently contain mixtures of categorical and continuous variables, outliers, missing data, and "just plain weird" bivariate distributions. The weather and baseball datasets are good examples of this behavior.

Third (and here we diverge from the original scagnostics formulation), feature-based analytics are not about scatterplot matrices. They are a function of the data that can be used in a variety of analytics to reveal structure in high-dimensional datasets. We have illustrated only a few analytics (sorting, clustering, outlier identification) on the features matrix F. We suspect that this general approach will lead to a variety of other new and useful visual analytics.

References

[1] C. Adami and A. Mazure. The use of minimal spanning tree to characterize the 2d cluster galaxy distribution. *Astronomy & Astrophysics Supplement Series*, 134:393–400, 1999.

[2] F. Anscolmbe and J.W. Tukey. The examination and analysis of residuals. *Technometrics*, pages 141–160, 1963.

[3] E. M. Arkin, Y-J Chiang, M. Held, J. S. B. Mitchell, V. Sacristan, S. Skiena, and T-H Yang. On minimum-area hulls. *Algorithmica*, 21(1):119–136, 1998.

[4] D. Asimov. The grand tour: A tool for viewing multidimensional data,. *Siam Journal on Scientific and Statistical Computing*, 6:128–143, 1985.

[5] A.C. Atkinson. *Plots, Transformations and Regression: An Introduction to Graphical Meth- ods of Diagnostic Regression Analysis*. Oxford University Press, 1985.

[6] A.C. Atkinson. Fast very robust methods for the detection of multiple outliers. *Journal of the American Statistical Association*, 89:1339–1994, 1994.

[7] V. Barnett and T. Lewis. *Outliers in Statistical Data*. John Wiley & Sons, 1994.

[8] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15:1373–1396, 2003.

[9] D. A. Belsley, E. Kuh, and R. E. Welsch. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. John Wiley & Sons, 1980.

[10] M. Brand. Nonlinear dimensionality reduction by kernel eigenmaps. In *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence*, pages 547–552, 2003.

[11] A. Buja and P. Tukey (Eds.). *Computing and Graphics in Statistics*. Springer-Verlag, New York, 1993.

[12] D. B. Carr, R. J. Littlefield, W. L. Nicholson, and J. S. Littlefield. Scatterplot matrix techniques for large n. *Journal of the American Statistical Association*, 82:424–436, 1987.

[13] W. S. Cleveland. *The Elements of Graphing Data*. Hobart Press, Summit, NJ, 1985.

[14] R.D. Cook and S. Weisberg. *An Introduction to Regression Graphics*. John Wiley & Sons, New York, 1994.

[15] H. Edelsbrunner, D. G. Kirkpatrick, and R. Seidel. On the shape of a set of points in the plane. *IEEE Transactions on Information Theory*, 29:551–559, 1983.

[16] J. H. Friedman and J. W. Tukey. A projection pursuit algorithm for exploratory data analysis. *IEEE Transactions on Computers*, C-23:881–890, 1974.

[17] M. Friendly and E. Kwan. Effect ordering for data displays. *Computational Statistics and Data Analysis*, 43(4):509–539, 2003.

[18] G.W. Furnas. Metric family portraits. *Journal of Classification*, 6:7–52, 1989.

[19] G.W. Furnas and A. Buja. Prosection views: dimensional inference through sections and projections. *Journal of Computational and Graphical Statistics*, 3(4):323–385, 1994.

[20] J. Gao and J. M. Steele. Sums of squares of edge lengths and spacefilling curve heuristics for the traveling salesman problem. *Siam Journal on Discrete Mathematics*, 7:314–324, 1994.

[21] J. C. Gower and G. J. S. Ross. Minimal spanning trees and single linkage cluster analysis. *Applied Statistics*, 18:54–64, 1969.

[22] J. A. Hartigan. Printer graphics for clustering. *Journal of Statistical Computation and Simulation*, 4:187–213, 1975.

[23] J. A. Hartigan and S. Mohanty. The runt terrst for multimodality. *Journal of Classification*, 9:63–70, 1992.

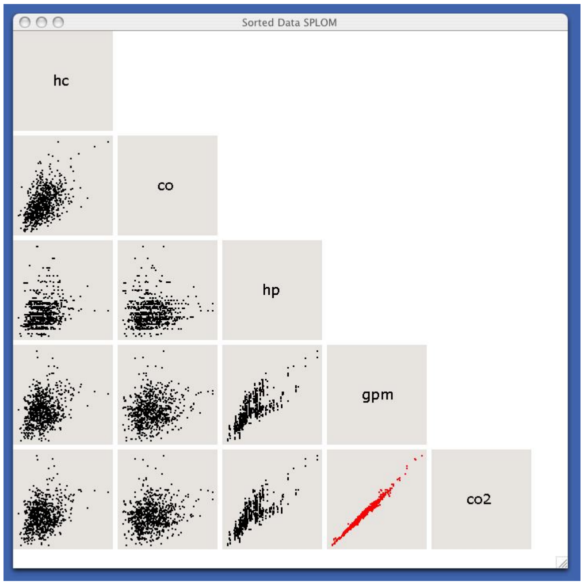

Figure 9: SPLOM of EPA data sorted by features component and colored by MST outlier statistic

[24] J.A. Hartigan. *Clustering Algorithms*. John Wiley & Sons, New York, 1975.

[25] T. Hastie and W. Stuetzle. Principal curves. *Journal of the American Statistical Association*, 84:502–516, 1989.

[26] J. Illingworth and J. Kittler. A survey of the Hough transform. *Computer Vision, Graphics, and Image Processing*, 44(1):87–116, 1988.

[27] J. Jaromczyk and G. Toussaint. Relative neighborhood graphs and their relatives, 1992.

[28] J.B. Kruskal Jr. On the shortest spanning subtree of a graph and the travelling salesman problem. *Proceedings of the American Mathematical Society*, 7:48–50, 1956.

[29] R.J.A. Little and D.B. Rubin. *Statistical Analysis with Missing Data*. John Wiley & Sons, New York, 1987.

[30] A. MacEachren, X. Dai, F. Hardisty, D. Guo, and G. Lengerich. Exploring high-d spaces with multiform matricies and small multiples. In *Proceedings of the IEEE Information Visualization 2003*, pages 31–38, 2003.

[31] D. Marchette. *Random Graphs for Statistical Pattern Recognition*. John Wiley & Sons, New York, 2004.

[32] W. J. Nash, T. L. Sellers, S. R. Talbot, A. J. Cawthorn, and W. B. Ford. The population biology of Abalone (Haliotis species) in Tasmania. I. Blacklip Abalone (H. rubra) from the north coast and islands of Bass Strait. Technical report, Sea Fisheries Division, 1994.

[33] J. O'Rourke. *Computational Geometry in C (2nd ed.)*. Cambridge University Press, Cambridge, 1998.

[34] W. Peng, M. O. Ward, and E. A. Rundensteiner. Clutter reduction in multi-dimensional data visualization using dimension reordering. In *Proceedings of the IEEE Information Visualization 2004*, pages 89–96, 2004.

[35] M. D. Penrose. Extremes for the minimal spanning tree on normally distributed points. *Advances in Applied Probability*, 30:628–639, 1998.

[36] F. P. Preparata and M. I. Shamos. *Computational Geometry: An Introduction*. Springer-Verlag, New York, 1985.

[37] D.M. Rocke and D.L. Woodruff. Identification of outliers in multivariate data. *Journal of the American Statistical Association*, 91:1047–1061, 1996.

[38] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290:2323–2326, 2000.

[39] D. W. Scott. *Multivariate Density Estimation: Theory, Practice, and Visualization*. John Wiley & Sons, New York, 1992.

[40] J. Seo and B. Shneiderman. A rank-by-feature framework for unsupervised multidimensional data exploration using low dimensional projections. In *Proceedings of the IEEE Information Visualization 2004*, pages 65–72, 2004.

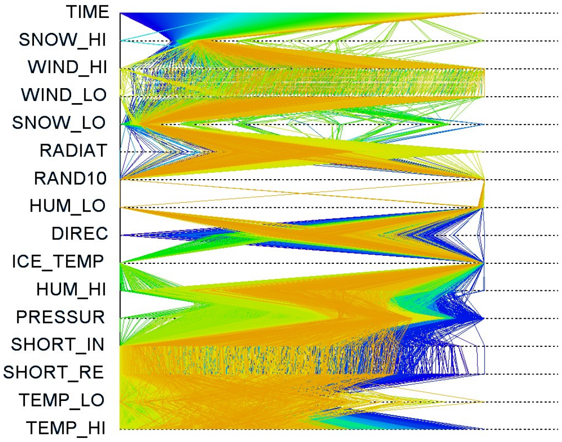

Figure 10: Parallel coordinate plot of weather data sorted by first principal component on correlations

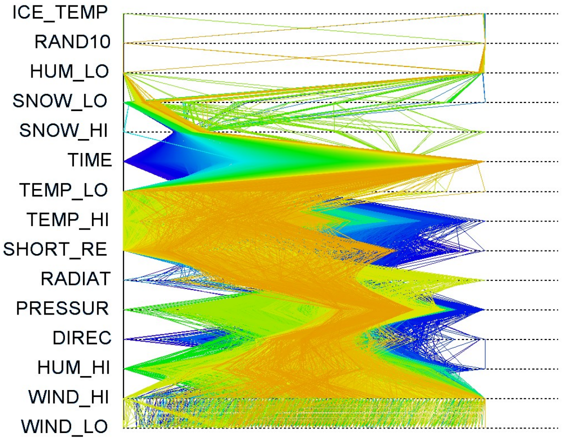

Figure 11: Parallel coordinate plot of weather data sorted by features component

[41] B. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman & Hall, New York, 1986.

[42] S.S. Skiena. *The Algorithm Design Manual*. Springer-Verlag, New York, 1998.

[43] J. M. Steele. Growth rates of Euclidean minimal spanning trees with power weighted edges. *The Annals of Probability*, 16:1767–1787, 1988.

[44] W. Stuetzle. Estimating the cluster tree of a density by analyzing the minimal spanning tree of a sample. *Journal of Classification*, 20:25–47, 2003.

[45] J. W. Tukey. Mathematics and the picturing of data. In *Proceedings of the International Congress of Mathematicians*, pages 523–531, Vancouver, Canada, 1974.

[46] J. W. Tukey. *Exploratory Data Analysis*. Addison-Wesley Publishing Company, Reading, MA, 1977.

[47] J. W. Tukey and P.A. Tukey. Computer graphics and exploratory data analysis: An introduction. In *Proceedings of the Sixth Annual Conference and Exposition: Computer Graphics85*, Fairfax, VA, United States, 1985. National Computer Graphics Association.

[48] P.F. Velleman and D.C. Hoaglin. *Applications, Basics and Computing of Exploratory Data Analysis.* Duxbury Press, 1981.

[49] L. Wilkinson. *The Grammar of Graphics (2nd ed.).* Springer-Verlag, New York, 2005.

[50] L. Wilkinson, A. Anand, and R. Grossman. Graph-theoretic scagnostics. In *Proceedings of the IEEE Information Visualization 2005*, pages 157–164, 2005.