Scagnostics Distributions

Leland WILKINSON and Graham WILLS

Scagnostics is a Tukey neologism for the term *scat*terplot dia*gnostics*. Scagnostics are characterizations of the 2D distributions of orthogonal pairwise projections of a set of points in multidimensional Euclidean space. These characterizations include such measures as density, skewness, shape, outliers, and texture. We introduce a set of scagnostics measures based on graph theory and we analyze their distributions and performance. Our analysis is based on a restrictive set of criteria that must be met in order to have scagnostics measures that can be used effectively in exploratory data analysis.

Key Words: AUTHOR: Please give 3–5 key words that do not appear in the title.

1. INTRODUCTION

In the mid-1980s, John and Paul Tukey introduced an exploratory graphical method called *scagnostics*. This method rested on a set of measures characterizing a 2D scatterplot. While they referred to their idea in Tukey and Tukey (1985), they never published an article or released a computer program on scagnostics. Paul Tukey offered some details at an Institute for Mathematics and its Applications (IMA) visualization workshop a few years later, but he did not include the talk in the workshop volume he and Andreas Buja edited (Buja and Tukey 1993).

Recently, based on the first author's recollection of the IMA workshop and subsequent conversations with Paul Tukey, Wilkinson et al. (2005) developed nine scagnostics measures defined on planar proximity graphs. These measures were scalable to large datasets and therefore suitable for practical applications. Although Wilkinson et al. (2005) documented their algorithms, they did not discuss the empirical distributions of the scagnostic measures themselves. This article investigates those distributions and provides some justification for the use of scagnostics in exploratory data analysis.

1.1 THE TUKEY IDEA

The Tukeys proposed characterizing a large collection of 2D scatterplots through a small number of measures of the arrangement of points in these plots. These measures

Leland Wilkinson is TKKK, University of Illinois at Chicago, Department of Computer Science, 851 S. Morgan Street, Chicago, IL 60606 (E-mail: *leland.wilkinson@gmail.com*). Graham Wills is TKKK, SPSS Inc.

Journal of Computational and Graphical Statistics, Volume 17, Number 2, Pages 1–19
DOI: 10.1198/106186008X320465

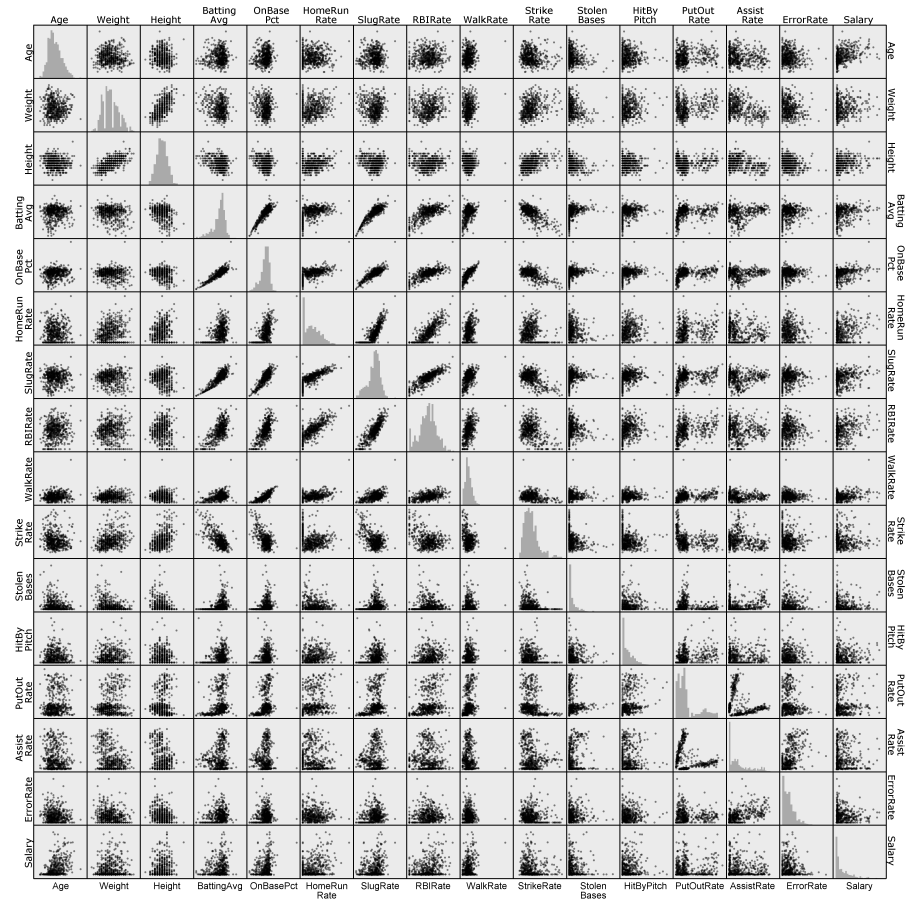

Figure 1. Scatterplot matrix of baseball player statistics. Even this moderate-sized matrix is difficult to read and individual scatterplots cannot easily be compared. Note the diversity of scatters, with few resembling bivariate normal distributions.

included the area of the peeled convex hull (Tukey 1974), the perimeter length of this hull, the area of closed 2D kernel density isolevel contours (Silverman 1986), the perimeter length of these contours, and a nonlinearity measure of association based on principal curves (Hastie and Stuetzle 1989). By using these measures, the Tukeys aimed to detect anomalies in density, shape, association, and other features of 2D scatterplots.

After computing these measures, the Tukeys made a scatterplot matrix (SPLOM) of the scagnostics themselves. This display, invented by Hartigan (1975) and popularized by Chambers et al. (1983), organizes scatterplots in the layout of a covariance matrix. The Tukeys intended to identify unusual scatterplots by linking each 2D plot to its corresponding point in the scagnostics SPLOM. With brushing and linking tools, a user could systematically navigate through a large collection of scatterplots by examining points in the scagnostic SPLOM.

Figures 1 through 3 show how this was expected to work. Figure 1 shows a SPLOM based on a dataset of baseball player statistics collected from various sites on the Web,

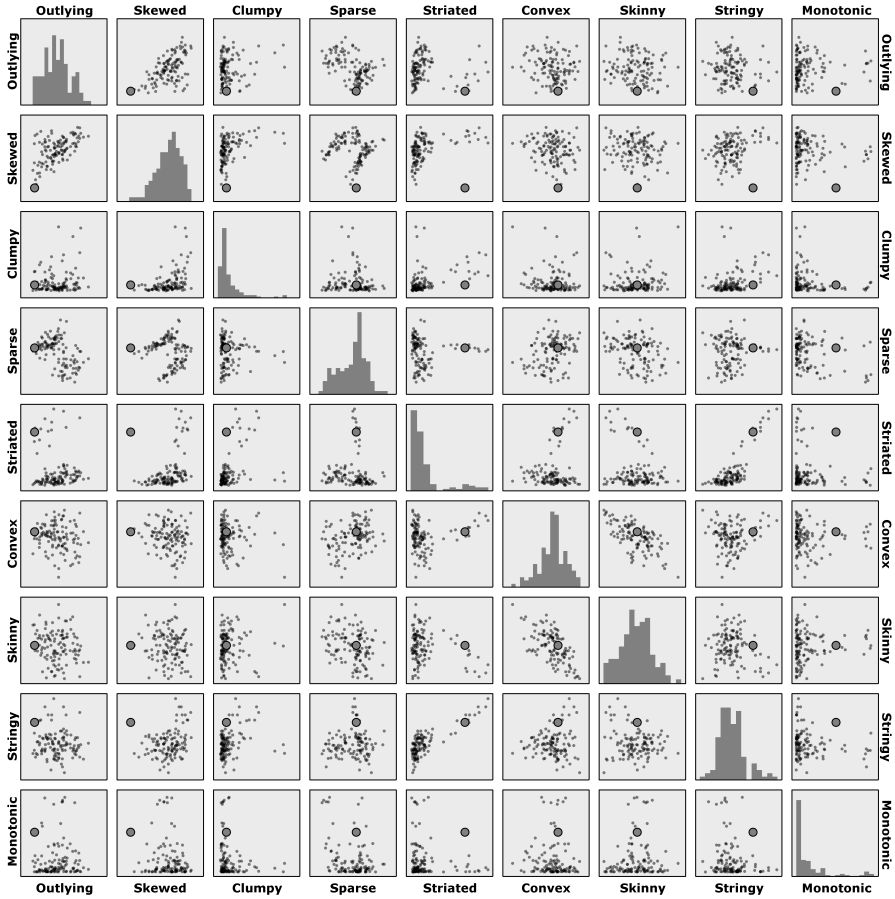

Figure 2. Scatterplot matrix of scagnostics measures derived from the baseball data in Figure 1. Each panel contains 120 points, one for each scatterplot in Figure 1. The point representing one unusual scatterplot is highlighted with a large circular symbol.

including variables such as Batting Average, On-Base Percent, Home Run Rate, etc. We chose a relatively small dataset for Figure 1 to show that, even with lensing or pan-and-zoom tools, navigating a SPLOM with more than 15 to 20 variables is impractical. The labels become too small and even the shapes of some of the scatterplots are difficult to discern. Scagnostics provides an alternative for identifying subsets of plots that share common features so that regularities and irregularities can be detected.

Irregularities in this context do not imply nonnormal distributions. Notice, for example, that the three scatterplots in the upper left corner of the SPLOM are plausibly bivariate normal (on the basis of the elliptical appearance of the scatters and on the basis of what we know about the distribution of these variables—Age, Weight, Height). There are few other bivariate normal plots in the SPLOM, so we should consider these three plots as unusual – outliers, as it were, among all the plots. Scagnostics should be based on measures that help us to identify both regularities *and* anomalies.

Figure 2 shows a SPLOM of the scagnostics computed on this dataset using the defini-

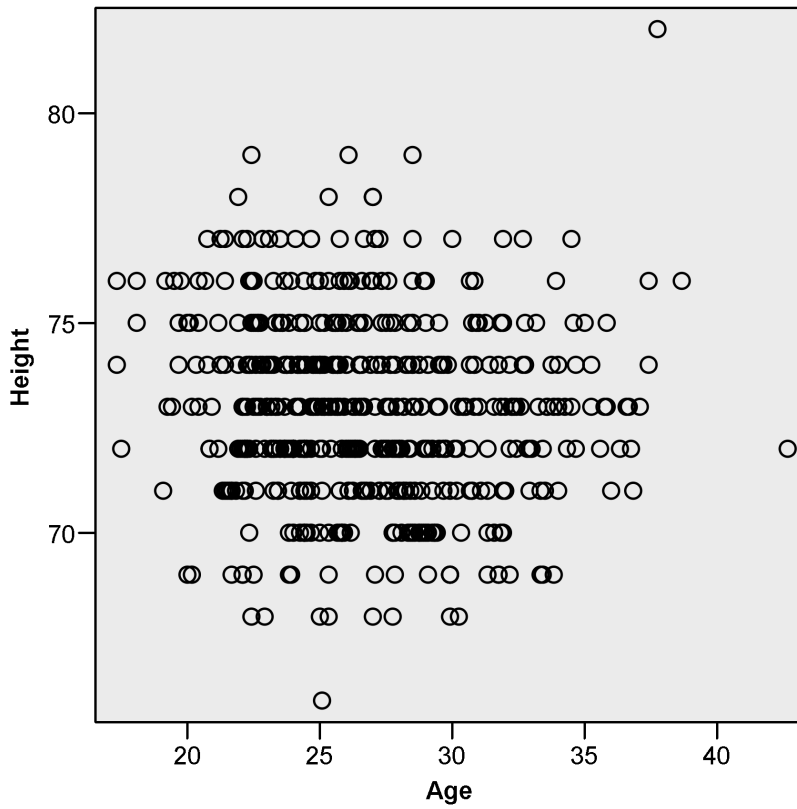

Figure 3. Scatterplot of age versus height of baseball players. This scatterplot is the one represented by the highlighted point in Figure 2.

tions in Wilkinson et al. (2005). There are nine scagnostic measures—Outlying, Skewed, Clumpy, Sparse, Striated, Convex, Skinny, Stringy, Monotonic—that determine the cells of the SPLOM. These measures (presented in Appendix A) were devised to cover a wide variety of distributions among real datasets.

Each scagnostics SPLOM cell contains 120 points, one for each 2D plot below the diagonal of the raw data SPLOM in Figure 1. We have highlighted one scagnostics point (plotted with a large circular symbol) that represents the Age-Height scatterplot shown in Figure 3. This scatterplot has few outliers, has an unusually small proportion of small interpoint distances (i.e., is not skewed), is not clumpy, and is quite striated. This profile is largely a consequence of the rounding of heights to the nearest inch, and a consequence of the bivariate normal distribution that appears to underly the data. The scagnostics reveal that this scatterplot is rather unusual when compared to most of the others. The striation and other unusual features would have been difficult to detect in the original SPLOM.

Wilkinson et al. (2005) discussed several other interesting aspects of the baseball data that are revealed by scagnostics. We will not repeat them here, except to note that their scagnostics were based on the ordinary words statisticians use to describe scatterplots. As

such, their scagnostics can be used to characterize the aspects of scatterplots that analysts frequently observe when devising models, diagnosing residuals, and searching for anomalies.

1.2 CONSIDERATIONS

The Tukeys' idea is powerful and simple, but implementing it involves many details. There are several criteria that should be met by candidate scagnostics:

1. We want to distinguish many types of point distributions: multivariate normal, log-normal, multinomial, sparse, dense, convex, clustered, etc.
2. We want a small number of scagnostics characterizing these distributions.
3. We want our scagnostics on a common scale so we can compare them to each other.
4. We want our scagnostics to have comparable distributions so we can compare them to a standard.
5. We want the intrinsic dimensionality of these scagnostics, when calculated over a large number of heterogeneous scatterplots, to be as large as possible.
6. We want our scagnostics to be efficiently computable so they are scalable to large numbers of points and dimensions.

Wilkinson et al. (2005) outlined an approach to meeting these criteria. They defined nine scagnostic measures (detailed in Appendix A), developed a scalable algorithm for computing them, and implemented a Java application for using them interactively. See the Appendix for available software.

Wilkinson et al. (2005) did not analyze the behavior or evaluate the effectiveness of the nine measures, however. This article reviews these measures and evaluates them on the above criteria. To conduct this evaluation, we used Monte Carlo methods on real and artificial datasets.

2. ASSESSMENT

Our assessments of the effectiveness of these graph-theoretic scagnostics necessarily involve Monte Carlo simulation and real datasets. There are scant small-sample or asymptotic results for the distribution of these statistics at this time. Furthermore, the universe of alternative distributions makes results based on only one or two reference distributions (uniform, normal, etc.) not particularly useful. Instead, we are interested in the performance of these scagnostics against the background of a wide variety of 2D point distributions likely to be found in real data. For our tests, therefore, we simulated disparate distributions and selected real datasets that are highly heterogeneous in their 2D marginal point distributions.

We examined four aspects of scagnostics behavior. First, we assessed consistency of the scagnostics across different sample sizes. Are they biased with regard to sample size?

Second, we investigated whether the distributions of the scagnostic measures are relatively homogeneous across a variety of point distributions. Can we approximate scagnostics distributions with a parametric family? Third, we looked at the sensitivity of the scagnostics. Do they respond sensitively to differences in 2D point distributions? Finally, we looked at the dimensionality of the scagnostics measures. Do these scagnostics correlate so highly with each other that they measure only a few aspects of 2D point clouds, or do they represent roughly nine dimensions of possible variation?

2.1 CONSISTENCY

To assess consistency, we ran a Monte Carlo simulation. We constructed ten 2D point distributions varying in their topology, density, and other critical aspects. These were, respectively,

1. Uniform (2D Poisson process)
2. Spherical (spherical normal)
3. Binormal (bivariate normal with $\rho = 0.6$)
4. Funnel (bivariate log-normal with $\rho = 0.6$)
5. Exponential (exponential function plus random error)
6. Quadratic (negative quadratic function plus random error)
7. Clustered (three separated spherical normals at the vertices of an equilateral triangle)
8. Doughnut (two polar uniforms separated by a moat of white space)
9. Stripe (product of Uniform and integer $[1, 5]$)
10. Sparse (product of integer $[1, 3]$ with itself)

From each of these distributions, we sampled 100 point configurations at nine different sample sizes ranging from 100 to 900. We then rescaled each sample to the unit square and computed each of the nine scagnostics.

Figure 4 shows boxplots of the scagnostics across different sample sizes for all the distributions. Overall, it appears that there are no substantial global trends across sample size. This is relatively reassuring because, as the computations in the Appendix show, we needed to adjust for bias due to hexagonal binning (which is required for scalability on large datasets). The adjustment appears successful.

There are several noteworthy aspects to Figure 4. Several of the boxplots have outliers. We expect this to be the case for the Outlying scagnostic. Furthermore, we expect to find a smaller proportion of outliers in larger samples from the distributions we devised; this is a sampling, not a binning effect. The remaining outliers are due to some of the unusual distributions we included in our study. The Skewed scagnostic has outliers in the lower tails due entirely to the Sparse dataset, which has only a mildly skewed distribution of

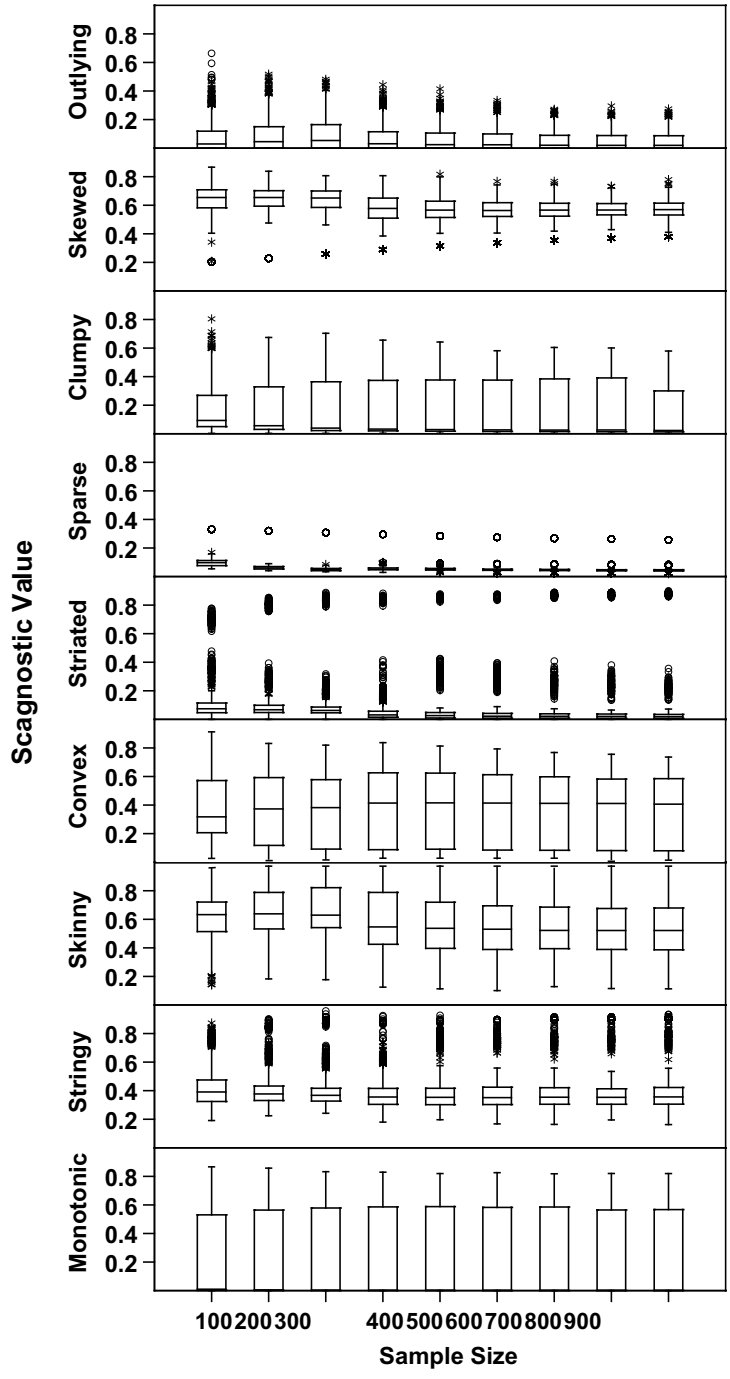

Figure 4. Boxplots of scagnostics measures for different samples from a wide variety of distributions. The horizontal axis represents sample size. The vertical axis is stratified by type of scagnostic. Scagnostic values vary between 0 and 1. The general lack of an overall trend across sample size indicates that the scagnostic measures are consistent.

interpoint distances. It appears we have not been able to adjust for binning bias in this one case. The Sparse scagnostic has outliers in the upper tails. Again, these are confined to the Sparse dataset. The binning bias is less apparent here. The Striated scagnostic has extreme outliers in the upper tails. The highest group is due, not surprisingly, to the Stripe dataset. There is a slight binning bias at the smallest sample sizes, but not elsewhere. Finally, the Stringy scagnostic has outliers due to the Stripe dataset. There does not appear to be much evidence of a binning bias here.

2.2 HOMOGENEITY

To assess homogeneity, we investigated whether the scagnostic measures have similar distributions. We expected several of these measures (especially the squared Spearman correlation) to follow roughly a beta distribution because they were defined on the unit interval and had shapes consistent with the beta in preliminary analyses. Consequently, we fit a beta to the sample scagnostics histograms. We used the same Monte Carlo design outlined in the previous section. This time, however, we set the sample size to 1,000 and selected three representative point distributions for each scagnostic. The goal here was to select a generating point set that would cause a sample to score low, medium, or high on each scagnostic. We then generated 1,000 pseudo-random samples for each configuration and computed the corresponding scagnostic for which that point set was designed.

Figure 5 shows the result of our simulation. The fitted beta distributions are gray and the sample kernel densities are black. The fits are close. Only one (the largest density on Sparse) had a Kolmogorov–Smirnov statistic larger than 0.1.

Several aspects of Figure 5 are noteworthy. First, it is clear that the α and β parameters for the fitted beta distributions are not related through a simple function (e.g., $\alpha + \beta = c$ or $\alpha\beta = c$). Unlike the distribution of the squared Pearson correlation, these scagnostics are not simply transformable to constant variance. The Skewed statistic, for example, has large variance in the same range that the Clumpy statistic has relatively small variance. This is because it is rare to find a 2D point distribution with negative skew in its interpoint distances. Slight perturbations in point locations can drastically affect the overall negative skew in these cases. Similarly, large values of the Outlying statistic have large variances. We generated these values by computing scagnostics on bivariate spherical normal data raised to the fifth power. Needless to say, we should expect high variability in the frequency and location of outliers in this context. By contrast, the low-outlier point set was generated from a uniform distribution. We expect the distribution of sample outliers to be near zero in this case.

Despite these understandable exceptions, the overall picture is relatively homogeneous. The simulation results reduce our concern that we might be mixing heterogeneous distributions when we construct composites of these scagnostic measures or use them in multivariate analyses for clustering, dimension reduction, and so on.

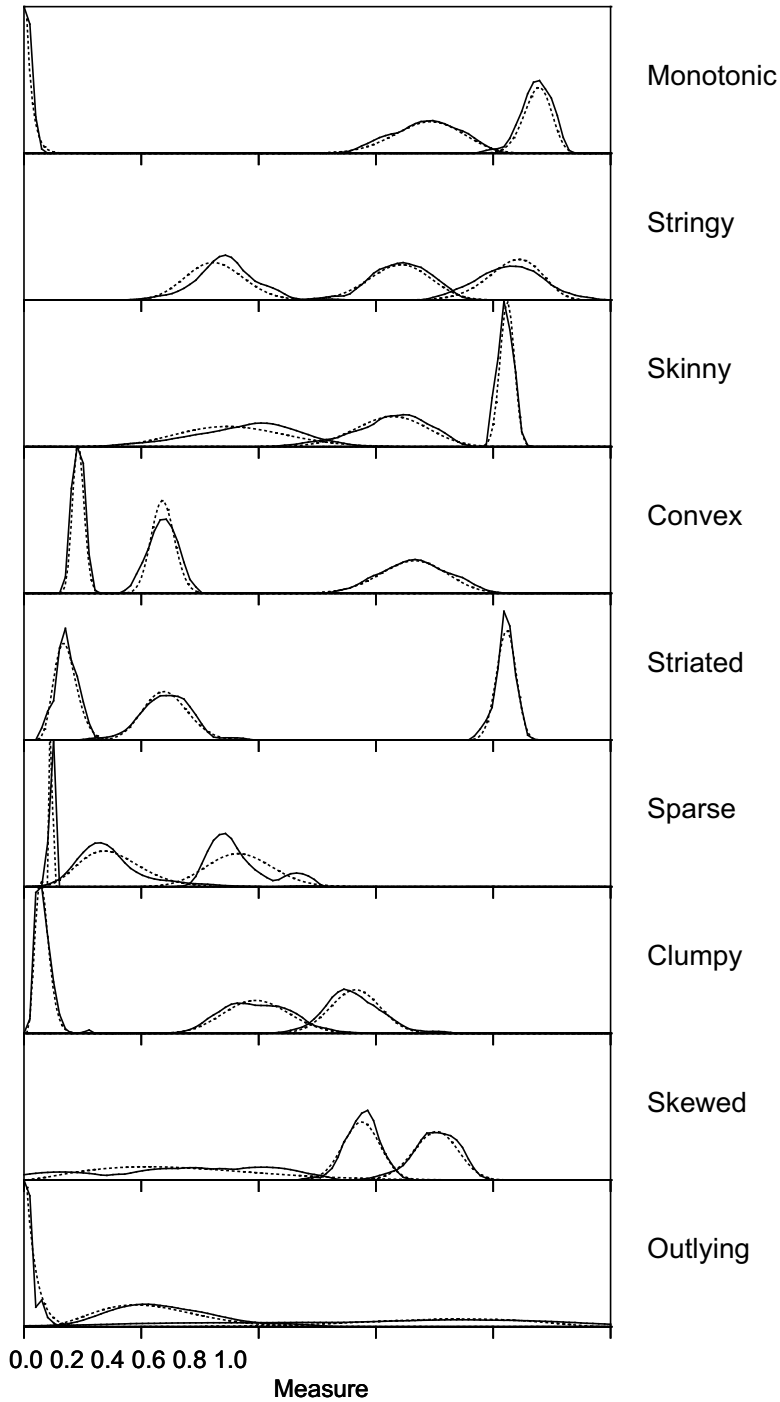

Figure 5. Scagnostics kernel densities (solid lines) and beta distribution fits (dotted lines). Each kernel density is based on 1,000 samples from one of the distributions used in Figure 4. The distributions were selected to yield a relatively low, medium, or high value on each scagnostic. The beta distribution is a relatively good fit to the sample distributions.

2.3 SENSITIVITY

Now we ask if each scagnostic responds effectively to point sets that exemplify the aspects it represents. If we had identifiable noise and signal distributions, we could do a signal detection analysis to determine each scagnostic's receiver operating characteristic. Since we do not, we might instead conduct a perceptual study in which human subjects classified point scatters and judged the success of the scagnostic measures themselves.

We offer a simpler approach at this point. Figure 6 shows a layout of sample 2D scatterplots arranged on the scagnostic scale. These scatterplots were selected from five real datasets:

1. Baseball—baseball player statistics cited in Wilkinson et al. (2005)
2. Boston—Boston housing statistics cited in Breiman et al. (1984)
3. Abalone—measurements of Abalone specimens from Nash et al. (1994)
4. Wind—measurements from a Greenland weather station, cited in Wilkinson (2005)
5. Ourworld—UN statistics on world countries, cited in Wilkinson (2005)

The location of the midpoint of each scatterplot on the horizontal scale represents approximately the value of the scagnostic measure represented in each row. It is important to keep in mind that there is not a unique 2D scatterplot that exemplifies each scagnostic value. We could have chosen a relatively uniform scatter of points to represent a low value of the Monotonicity measure, for example. And we could have chosen a monotonic functional dataset to represent a high Monotonicity value instead of the linear one we chose. There are many ways to be monotonic, stringy, skinny, convex, clumpy, and so on. Nevertheless, Figure 6 indicates that each scagnostic is sensitive to the kind of variation it is intended to represent.

2.4 DIMENSIONALITY

Finally, we would like to assess whether the nine scagnostics measure relatively uncorrelated aspects of point scatters. First, though, we must ask the question, "Uncorrelated over what?" At this point, we believe the best answer to that question is, "Uncorrelated over real data." Consequently, we assembled five heterogeneous datasets and computed scagnostics on pairwise scatters from each. Then we computed principal components on these scagnostics and examined the distribution of eigenvalues corresponding to each component.

Figure 7 shows a scree plot of these eigenvalues. To assist judging the distribution of eigenvalues, we have included a scree for random data consisting of 100 pseudo-random values on a nine-dimensional spherical Gaussian. We have square-rooted the vertical scale to reduce overlap among the profiles. And we have kept the symbols light gray to reduce clutter.

There is considerable variability in the screes. This is to be expected. The steepest is for the Abalone dataset. This is because almost every scatter in this dataset has the

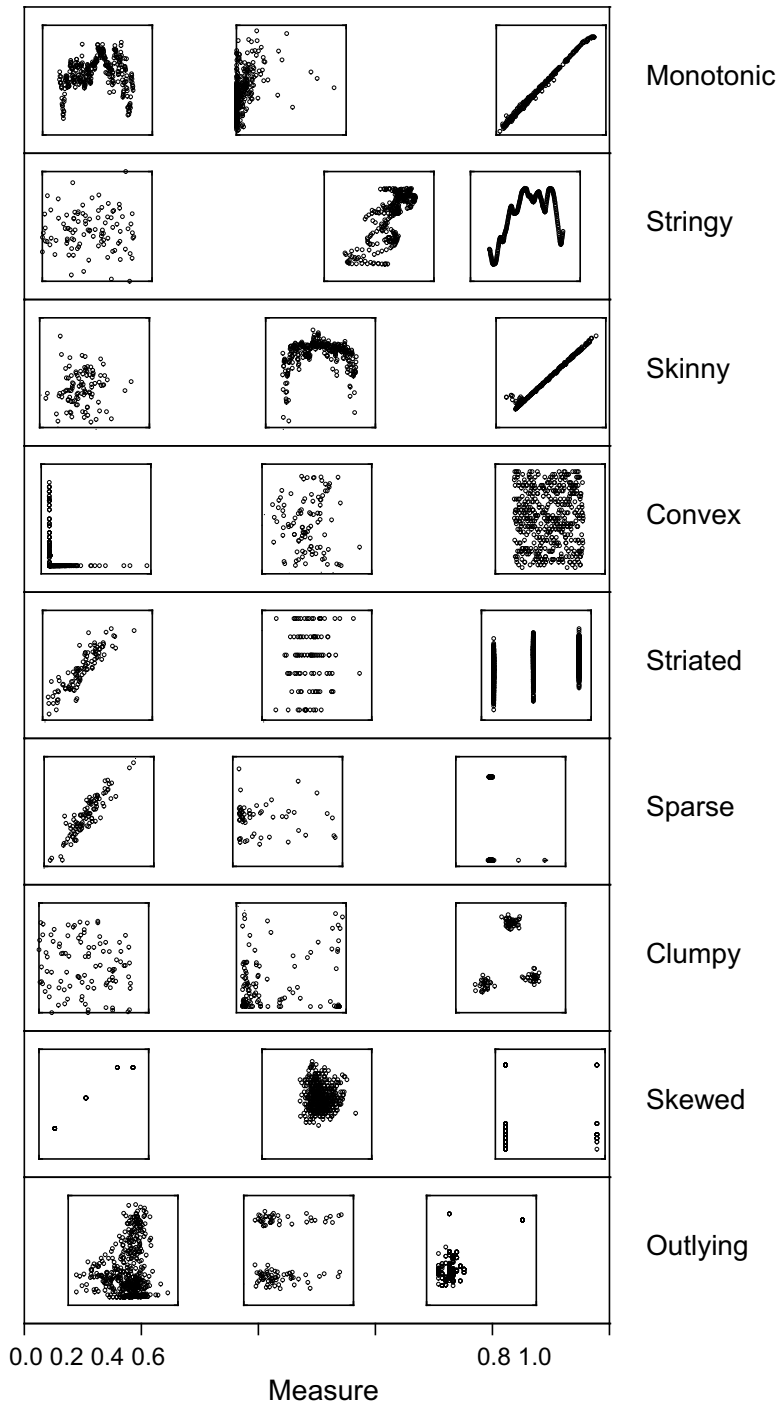

Figure 6. Scatterplots from a variety of real datasets aligned on scagnostics scale for each scagnostic. The scatterplots were selected for having a relatively low, medium, or high value on each scagnostic. This figure shows that high-value scatterplots are reasonable exemplars for the descriptive names (Monotonic, Stringy, etc.) and low-value scatterplots correspondingly lack the feature described by each scatterplot name.

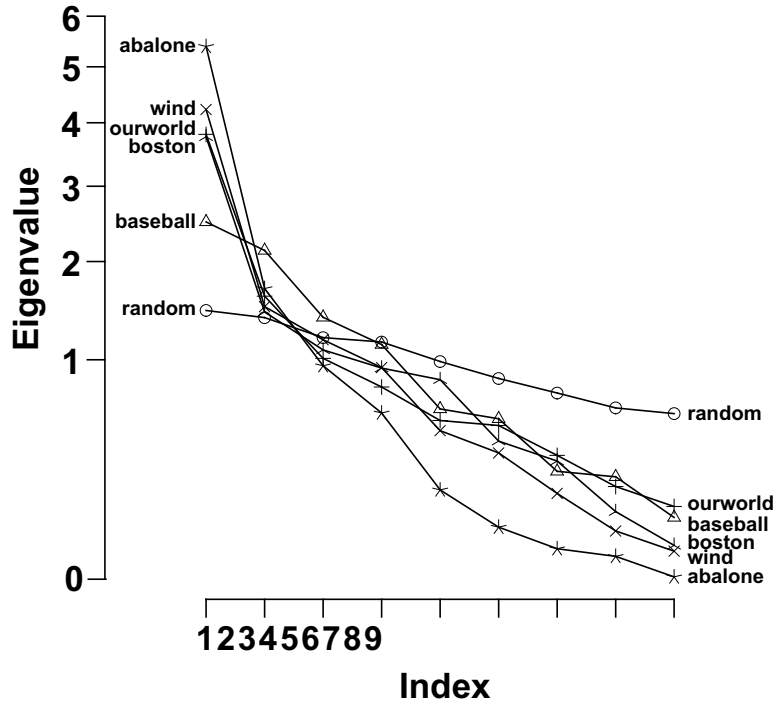

Figure 7. Scree plot of eigenvalues for each of five sample datasets. The vertical axis has been logged in order to separate the curves. The scree for bivariate random normal data is relatively flat. The scree for the baseball dataset comes closest to the random scree because the baseball scatterplots are heterogeneous on the scagnostics measures. By contrast, the abalone dataset scree has a steep slope because almost all the abalone pairwise scatterplots have the same shape. The range of eigenvalues indicates that the scagnostics measures are not derived from one or two underlying components.

same teardrop shape. The shallowest scree is for the Baseball dataset. This dataset has the greatest variety (by inspection) of scatter patterns. Our world and Boston are somewhat less heterogeneous. It should be clear that as the universe of test patterns expands, the dimensionality of the scagnostics increases. When applied to heterogeneous data, these scagnostic indices measure more than a few latent factors.

3. CONCLUSION

Wilkinson et al. (2005) demonstrated scagnostics applications in visualization, multivariate sorting, clustering, and outlier identification. Although the Tukeys originally intended it to be a graphical tool for inspecting scatterplot matrices, it is clear they expected it to be more (Tukey and Tukey 1985). We do as well. For example, we are now considering the application of scagnostics to high-dimensional tomography. We hope to detect low-dimensional structure in high-dimensional point embeddings by using scagnostics.

Before we develop scagnostic applications, however, it is critical to determine if the measures have the properties we introduced in this article. They should be sensitive to a wide variety of distributions. They should be on a common scale. They should have com-

parable distributions. They should be small in number but large in effective dimensionality. And they should be computationally efficient.

Meeting all these criteria is not a trivial task. We considered a large number of candidate measures before winnowing the number to nine. There is room for more, of course, but additions should meet these criteria as well as the original nine do. The approach in this article should serve as a first step toward this goal.

A. COMPUTING SCAGNOSTICS

For more detail on the material in this Appendix, see Wilkinson et al. (2005). A Java interactive scagnostics program is available from the second author (*leland.wilkinson@gmail.com*). Hadley Wickham has developed a scagnostics R function (based on a C++ translation of the Java code). It is available in CRAN (*http://cran.r-project.org/*). Other material and datasets are available on the *JCGS* website.

A.1 GEOMETRIC GRAPHS

Our scagnostic measures are based on the following definitions. 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 straight line segments connecting pairs of points. We restrict our graphs to 2D Euclidean geometric graphs and omit the asterisk in subsequent notation.

Our measures are derived from several features of 2D Euclidean 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 successively adjacent, distinct edges. A path is *closed* if its first and last vertex are the same. A $polygon$, P, is a region bounded by a closed path. A *simple polygon* is a polygon bounded by exactly one closed path that has no intersecting edges. We restrict P to simple polygons. The perimeter of a simple polygon, perimeter(P), is the length of its boundary. The area of a simple polygon, area(P) is the area of its interior.

A.1.1 Minimum Spanning Tree

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 is least of all spanning trees on a given set of points (Kruskal 1956). We restrict ourselves to the geometric MST computed from Euclidean distances between points in a 2D Euclidean geometric graph.

A.1.2 Convex Hull

A *hull* of a set of points embedded in 2D Euclidean space is a collection of the boundaries of one or more simple polygons that have a subset of the points for their vertices and that collectively contain all the points. This definition includes entities that range from the boundary of a single simple polygon to a collection of boundaries of simple 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. By definition, the convex hull bounds a single polygon. A *peeled convex hull* is a convex hull computed after deleting points on the convex hull.

A.1.3 Alpha Hull

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 (Jaromczyk and Toussaint 1992). 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 $D(r)$.

An alpha shape is a collection of one or more simple polygons. In an *alpha shape graph* (Edelsbrunner et al. 1983), an edge exists between any pair of points that can be touched by an open disk $D(\alpha)$ containing no points. Marchette (2004) recommended a value of α to be the average value of the edge lengths in the MST. To reduce noise, we use a larger value, namely, the 90th percentile of the MST 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.

A.2 PREPROCESSING

We bin our data and delete outliers before computing our geometric graphs. This preprocessing improves performance of our algorithms and robustness of our measures.

A.2.1 Binning

We begin by normalizing the data to the unit interval and then use a 40 by 40 hexagonal grid to aggregate the points in each scatterplot. If there are more than 250 nonempty cells, we reduce the bin size by half and rebin. We rebin until there are no more than 250 nonempty cells. The choice of bin size is constrained by efficiency (too many bins slow down calculations of the geometric graphs) and sensitivity (too few bins obscure features in the scatterplots).

We use hexagon binning (Carr et al. 1987) to improve performance. Hexagon binning is slightly slower than rectangular binning, but reduces anisotropy of local neighborhoods because of the near-circular shape of hexagons. This bias reduction is important for keeping scagnostics orientation-independent.

Binning, like other aggregation methods, can affect statistical estimates. A well-known instance of such an effect is the ecological correlation (Freedman 2001). Coarse binning can make dense point sets look sparse, nonconvex distributions look convex, and so on. Consequently, we apply a stabilizing transformation on some of the scagnostics computed from binned data to attenuate the influence of binning. Our weight function is

$$w = 0.7 + \frac{0.3}{1 + t^2}, \tag{A.1}$$

where $t = n/500$. This function is fairly constant for $n > 2000$. We determined its shape and parameters by hex binning and computing scagnostics on a wide variety of datasets. We use this function to adjust for bias in the Skewed, Sparse, and Convex scagnostics formulas below.

A.2.2 Deleting Outliers

We delete outliers to improve robustness of our scagnostics. Classical outlier detection methods (Barnett and Lewis 1994) are of little use for this purpose because they presume parametric densities. To avoid distributional assumptions, Tukey (1974) used the recursively peeled convex hull to delete extreme points. For 1D points, this amounts to Winsorizing, or successive symmetric trimming of extreme observations.

Because we do not assume convex support for our point sets, we cannot expect outliers will be outside the edges of a peeled convex hull. We want to identify points located in relatively sparse interior regions, for example. Consequently, we peel the MST instead of the convex hull. We consider an outlier to be a vertex whose adjacent edges in the MST all have a weight (length) greater than ω.

There are theoretical results on the distribution of the largest edge for an MST on normally distributed data (Penrose 1998), but we work instead with a nonparametric criterion for simplicity. Following Tukey (1977), we choose

$$\omega = q_{75} + 1.5(q_{75} - q_{25}), \tag{A.2}$$

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.

A.3 COMPUTING SCAGNOSTIC MEASURES

We now present the scagnostic measures computed on our three geometric graphs. In the formulas below, we use H for the convex hull, A for the alpha hull, and T for the minimum spanning tree.

We are interested in assessing three aspects of scattered points: *density*, *shape*, and *association*.

A.3.1 Density Measures

The following measures detect different distributions of points.

- *Outlying*

The Outlying scagnostic measures the proportion of the total edge length of the minimum spanning tree accounted for by the total length of edges adjacent to outlying points (as defined above). Note that we do this calculation before deleting outliers for the other measures.

$$c_{\text{outlying}} = \text{length}(T_{\text{outliers}})/\text{length}(T). \tag{A.3}$$

- *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 (Adami and Mazure 1999). However, theoretical results (Steele 1988; Penrose 1998) 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 therefore use two measures of relative density. The first is a relatively robust measure of skewness in the distribution of edge lengths.

$$q_{\text{skew}} = (q_{90} - q_{50})/(q_{90} - q_{10}). \tag{A.4}$$

Because Skewed tends to *decrease* with n after adaptive binning, we invert the weight in (A.1) to compute the Skewed scagnostic.

$$c_{\text{skew}} = 1 - w(1 - q_{\text{skew}}) \tag{A.5}$$

- *Sparse*

The second edge-length statistic, Sparse, 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 is the same value we use for the α statistic.

$$c_{\text{sparse}} = w q_{90}, \tag{A.6}$$

where w is the weight function in (A.1). 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.

- *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 RUNT statistic (Hartigan and Mohanty 1992). 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 (Gower and Ross 1969), we can associate a runt size (r_j) with each edge (e_j) in the MST, as described by Stuetzle (2003). 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).

The RUNT-based measure responds to clusters with small maximum intracluster distance relative to the length of their nearest-neighbor inter-cluster distance. In the formula below, j runs over all edges in T and k runs over all edges in R_j.

$$c_{\text{clumpy}} = \max_j \left[1 - \max_k \left[\text{length}(e_k) \right] / \text{length}(e_j) \right] \tag{A.7}$$

- *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. Another common example is the pattern of parallel lines of points produced by the product of categorical and continuous variables.

We could recognize parallel lines with a Hough transform (Illingworth and Kittler 1988). Other configurations of points that represent vector flows or striated textures might not follow parallel or even straight paths, however. We use 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

$$c_{\text{striate}} = \frac{1}{|V|} \sum_{v \in V^{(2)}} I(\cos \theta_{e(v,a)e(v,b)} < -0.75). \tag{A.8}$$

A.3.2 Shape Measures

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, and so forth. Of course, scattered points are by definition *not* these things, so we need additional machinery (based on geometric graphs) to allow us to make such inferences. In particular, we will measure aspects of the convex hull, the alpha hull, and the minimum spanning tree.

- *Convex*

Our convexity measure is based on 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.

$$c_{\text{convex}} = w[\text{area}(A)/\text{area}(H)], \tag{A.9}$$

where w is the weight function in (A.1).

- *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.

$$c_{\text{skinny}} = 1 - \sqrt{4\pi\,\text{area}(A)}/\text{perimeter}(A). \tag{A.10}$$

- *Stringy*

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

$$c_{\text{stringy}} = \frac{|V^{(2)}|}{|V| - |V^{(1)}|}. \tag{A.11}$$

We cube the Stringy measure to adjust for negative skew in its conditional distribution on n.

A.3.3 Association Measure

We are interested in a symmetric and relatively robust measure of association.

- *Monotonic*

We use the squared Spearman correlation coefficient 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.

$$c_{\text{monotonic}} = r^2{}_{\text{Spearman}} \tag{A.12}$$

This is the only coefficient not based on a subset of the Delaunay graph.

[Received August 2006. Revised July 2007.]

REFERENCES

Adami, C., and Mazure, A. (1999), "The Use of Minimal Spanning Tree to Characterize the 2D Cluster Galaxy Distribution," *Astronomy & Astrophysics Supplement Series*, 134, 393–400.

Barnett, V., and Lewis, T. (1994), *Outliers in Statistical Data*, New York: Wiley.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984), *Classification and Regression Trees*, Belmont, CA: Wadsworth.

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

Carr, D. B., Littlefield, R. J., Nicholson, W. L., and Littlefield, J. S. (1987), "Scatterplot Matrix Techniques for Large *n*," *Journal of the American Statistical Association*, 82, 424–436.

Chambers, J. M., Cleveland, W. S., Kleiner, B., and Tukey, P. A. (1983), *Graphical Methods for Data Analysis*, New York: Chapman and Hall.

Edelsbrunner, H., Kirkpatrick, D. G., and Seidel, R. (1983), "On the Shape of a Set of Points in the Plane," *IEEE Transactions on Information Theory*, 29, 551–559.

Freedman, D. (2001), "Ecological Inference and the Ecological Fallacy," in *International Encyclopedia of the Social and Behavioral Sciences*, eds. N. Smelser and P. Baltes, Oxford: Pergamon Press, pp. 4027–4030.

Gower, J. C., and Ross, G. J. S. (1969), "Minimal Spanning Trees and Single Linkage Cluster Analysis," *Applied Statistics*, 18, 54–64.

Hartigan, J. A. (1975), "Printer Graphics for Clustering," *Journal of Statistical Computation and Simulation*, 4, 187–213.

Hartigan, J. A., and Mohanty, S. (1992), "The Runt Test for Multimodality," *Journal of Classification*, 9, 63–70.

Hastie, T., and Stuetzle, W. (1989), "Principal Curves," *Journal of the American Statistical Association*, 84, 502–516.

Illingworth, J., and Kittler, J. (1988), A Survey of the Hough Transform," *Computer Vision, Graphics, and Image Processing*, 44, 87–116.

Jaromczyk, J., and Toussaint, G. (1992), "Relative Neighborhood Graphs and their Relatives." ¿¿JOURNAL INFO¡¡

Kruskal, J. (1956), "On the Shortest Spanning Subtree of a Graph and the Travelling Salesman Problem," *Proceedings of the American Mathematical Society*, 7, 48–50.

Marchette, D. (2004), *Random Graphs for Statistical Pattern Recognition*, New York: Wiley.

Nash, W. J., Sellers, T. L., Talbot, S. R., Cawthorn, A. J., and Ford, W. B. (1994), "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.

Penrose, M. D. (1998), "Extremes for the Minimal Spanning Tree on Normally Distributed Points," *Advances in Applied Probability*, 30, 628–639.

Silverman, B. (1986), *Density Estimation for Statistics and Data Analysis*, New York: Chapman & Hall.

Steele, J. M. (1988), "Growth Rates of Euclidean Minimal Spanning Trees with Power Weighted Edges," *The Annals of Probability*, 16, 1767–1787.

Stuetzle, W. (2003), "Estimating the Cluster Tree of a Density by Analyzing the Minimal Spanning Tree of a Sample," *Journal of Classification*, 20, 25–47.

Tukey, J. W. (1974), "Mathematics and the Picturing of Data," in *Proceedings of the International Congress of Mathematicians*, Vancouver, Canada, pp. 523–531.

——— (1977), *Exploratory Data Analysis*, Reading, MA: Addison-Wesley Publishing Company.

Tukey, J. W., and Tukey, P. (1985), "Computer Graphics and Exploratory Data Analysis: An Introduction," in *Proceedings of the Sixth Annual Conference and Exposition: Computer Graphics '85*, Fairfax, VA: National Computer Graphics Association.

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

Wilkinson, L., Anand, A., and Grossman, R. (2005), "Graph-Theoretic Scagnostics," in *Proceedings of the IEEE Information Visualization 2005*, pp. 157–164.