ScagExplorer: Exploring Scatterplots by Their Features

Tuan Nhon Dang
University of Illinois at Chicago

Leland Wilkinson
University of Illinois at Chicago

Abstract

Scatterplot matrix (SPLOM) is a useful tool for displaying bivariate relationships; however it can easily run out of pixels when presenting high-dimensional data. In this poster, we introduce a method for guiding interactive exploration of high-dimensional data. The method is based on characterizations of the 2D distributions of orthogonal pairwise projections on a set of points in multidimensional Euclidean space. These characterizations include measures such as, density, shape, outliers, and texture. Our application, ScagExplorer, is designed to handle the types of multivariate data series that are often found in security, financial, social, and other sectors.

Index Terms: I.5.2 [Pattern recognition]: Design Methodology—Pattern analysis

1 Introduction

Previous researchers have applied data-driven approaches to moderate-sized collections of scatterplots. The scale of these efforts has been constrained by display limits and computational complexity. We have developed an alternative approach in order to deal with huge collections of scatterplots. Our goal is to be able to organize these plots into meaningful subsets in reasonable time and to present these plots to users in a rich exploratory environment.

Many applications generate huge collections of scatterplots. Suppose, for example, we have data consisting of many time series over many variables with many time points. Suppose, further, that we want to identify unusual events at individual time points across all the series. If there is only one time series for each variable, the usual analytic approach to this problem is to perform spectral modeling of cross-covariance functions among the series. A visual analytic analog of this approach is to plot pairs of series and highlight noteworthy features in the pairs that appear to be substantially related. This traditional approach will not work if there is more than one time series for each variable, however. The total number of possible scatterplots expands multiplicatively.

Our contributions in this poster are: We have proposed the scagnostics-based quantitative measurement to help capture similar scatterplots in the collection. Then, we have developed a dynamic algorithm that leverages force-directed graph methods to cluster large collections of scatterplots. This cluster layout provides an overview of the 2D relations of variables in a dataset. Users can then select a cluster to investigate for further details.

2 Related Work

2.1 Scagnostics

Scagnostics were introduced by Tukey and Wilkinson [8]. The graph-theoretic implementation of scagnostics are described in details in [9]. In brief, the scagnostics measures depend on proximity graphs that are all subsets of the Delaunay triangulation: the convex hull, the minimum spanning tree (MST), and the alpha complex [4]. Figure 1 shows an example of the three geometric graphs generated on the same set of data points.

In this poster, we have selected to work with 6 out of 9 scagnostics which are more significant in characterizing bivariate relationships: Outlying, Clumpy, Striated, Sparse, String, and Monotonic.

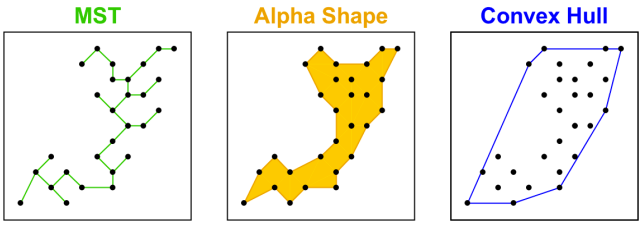

Figure 1: Graphs for computing scagnostics measures.

2.2 Feature-based Approaches

Seo and Shneiderman [5] computed statistical summaries (means, standard deviations, correlations, etc.) on univariate and bivariate distributions and then ranked them in order to identify similar distributions. Other researchers have developed scagnostics-type measures for parallel coordinates, pixel displays, and other graphics [3, 7]. Yang et al.[10] proposed a *Value and Relation* technique grasping the associations among dimensions. However, this technique fails to capture more complicated relations such as Stringy.

TimeSeer [1] uses scagnostics for organizing multivariate time series and for guiding interactive exploration through high-dimensional data. TimeSeer can graph scagnostics time series of up to 10 preselected pairs of variables in single display. TimeSeer [2] resolves this limitation by allowing a user to examine all time series in a corpus simultaneously.

3 ScagExplorer

ScagExplorer is a platform for interactively visualizing scagnostics on multivariate distributions. ScagExplorer is designed to deal with a large scatterplot space containing thousands of scatterplots. While TimeSeer is devoted to multivariate time series, ScagExplorer has nothing to do with time-oriented data. It is, instead, concerned with managing and exploring large collections of scatterplots.

3.1 Dissimilarity of Two Scatterplots

The dissimilarity of two scatterplots (S and P) is computed by the following equation:

$$Dissimilarity(S,P) = \sqrt{\sum_{i=1}^{6}(S_i - P_i)^2} \qquad (1)$$

where S and P are two arrays of six scagnostics of the two scatterplots. The most obvious benefit of our parameterization is to reduce the complexity of searching for similar scatterplots from $O(n)$ to $O(1)$. The tradeoff here is, of course, that we might lose details of scatterplots.

3.2 Exploring a large corpus of plots

There are several solutions to deal with the scalability problem of SPLOMs: **Dimension reduction**: preselect a subset of dimensions. The advantage of this approach is that we downsize the SPLOM. However, information lost is a disadvantage of this approach. **Lensing** [1]: this approach is time consuming to obtain an overview of the whole data,. **SPLOM reordering**: Wilkinson et al.[9] sorted

the variables in the raw data SPLOM using the size of the loadings on the first principal component of the scagnostic measures. However, features sorting does not guarantee that the two adjacent plots are similar. Lehmann et al.[6] used a known quality measure to reorder dimensions. This reordering method concentrates on the best plots in different regions, and the number of relevant regions is arbitrary. Due to these limitations, ScagExplorer allows the scatterplots beak out of the SPLOM to move to their desired location (using the force-directed layout).

We use force-directed layout to place the scatterplots on a 2D view because this layout is intuitiveness, flexibility, and interactivity. The main disadvantage is high running time. Since for every plot, we have to compute attraction or repellant against all other plots, the running time at each iteration is $O(n^2)$.

We first put all scatterplots randomly in the output panel and we then allow them to interact to find clusters based on their scagnostics measures. Figure 2(a) shows an example of our dynamic clustering algorithm on the US economy data from the Bureau of Labor Statistics. The data comprise monthly statistics on various aspects of the US economy over 12 years from 2000 to 2011. There are 44 variables represented in the dataset: Consumer Price Index (CPI), Producer Price Index (PPI), Import/Export Price Indexes, and Employment Rate of major economy sectors. For these data, we have 946 scatterplots with 144 data points (144 months of 12 years) to examine. In particular, we color the scatterplots by their Outlying (high Outlying plots are in red, low Outlying plots are in blue). We then can investigate the details of each cluster. In Figure 2(a), we inspect high Outlying cluster on the top left. ScagExplorer provides a dragging box (Box A) for cluster selection. The details of selected plots in Box A are displayed in Box B on the right. As depicted, most of outliers involve December, January, and February.

The Scagnostic View in Figure 2(b) to reveals interesting distributions of scagnostics. We have superimposed a typical scatterplot within each cluster: cluster containing plot 1 is high Outlying, cluster containing plot 2 is high Stringy and low Monotonic, cluster containing plot 3 is high Stringy and high Monotonic, cluster containing plot 4 is high Striated, and cluster containing plot 5 is high Striated and high Monotonic. Exemplar scatterplots from each cluster clearly show the differences between clusters. Notice that there are also polarities in a big cluster. The color legend for scagnostics is depicted on the right.

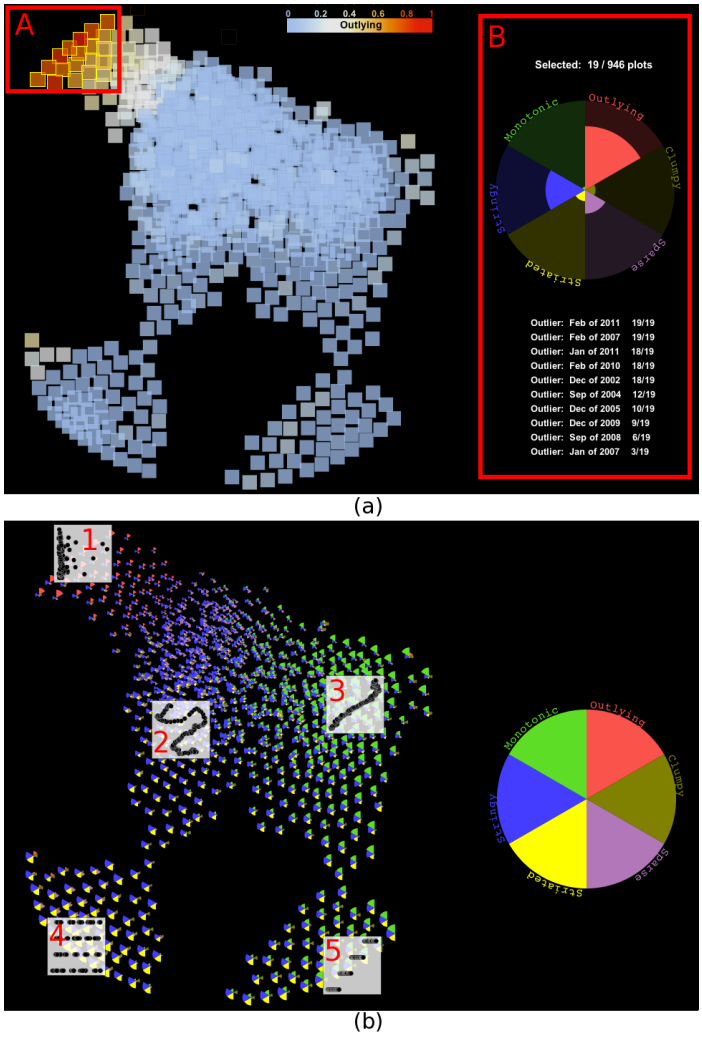

Figure 2: US economy data: a) Cluster View b) Scagnostic View.

4 Conclusions

ScagExplorer is a visual tool for analyzing high-dimensional datasets. It highlights the usefulness of examining raw data distributions rather than reductive statistical summaries or aggregations (as in OLAP cubes). ScagExplorer makes it possible to examine pairwise scatterplots in a coherent framework. While parallel coordinates and other multivariate displays (glyphs, projections, profiles, etc.) have their uses, it is helpful to have a tool that allows a user to explore data in the familiar form of the scatterplot.

Finally, we plan to investigate the use of ScagExplorer on large security databases to assess the gains we claim for its performance. In addition, we expect to investigate how ScagExplorer can be extended to temporal and spatial data.

References

[1] T. N. Dang, A. Anand, and L. Wilkinson. Timeseer: Scagnostics for high-dimensional time series. *IEEE Transactions on Visualization and Computer Graphics*, 99(PrePrints), 2012.

[2] T. N. Dang and L. Wilkinson. Timeseer: detecting interesting distributions in multiple time series data. In *Proceedings of the 5th International Symposium on Visual Information Communication and Interaction*, VINCI '12, pages 43–51, New York, NY, USA, 2012. ACM.

[3] A. Dasgupta and R. Kosara. Pargnostics: Screen-space metrics for parallel coordinates. *IEEE Transactions on Visualization and Computer Graphics*, 16:1017–2626, 2010.

[4] 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.

[5] S. J. and B. Shneiderman. A rank-by-feature framework for interactive exploration of multidimensional data. *Information Visualization*, 4:96–113, March 2005.

[6] D. J. Lehmann, G. Albuquerque, M. Eisemann, M. Magnor, and H. Theisel. Selecting coherent and relevant plots in large scatterplot matrices. *Comp. Graph. Forum*, 31(6):1895–1908, Sept. 2012.

[7] J. Schneidewind, M. Sips, and D. Keim. Pixnostics: Towards measuring the value of visualization. In *Proceedings of IEEE Symposium on Visual Analytics Science and Technology (VAST)*, pages 199–206, Baltimore, MD, 2006.

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

[9] L. Wilkinson, A. Anand, and R. Grossman. High-dimensional visual analytics: Interactive exploration guided by pairwise views of point distributions. *IEEE Transactions on Visualization and Computer Graphics*, 12(6):1363–1372, 2006.

[10] J. Yang, A. Patro, S. Huang, N. Mehta, M. O. Ward, and E. A. Rundensteiner. Value and relation display for interactive exploration of high dimensional datasets. In *Proceedings of the IEEE Symposium on Information Visualization*, INFOVIS '04, pages 73–80, Washington, DC, USA, 2004. IEEE Computer Society.