Transforming Scagnostics to Reveal Hidden Features

Tuan Nhon Dang, Leland Wilkinson

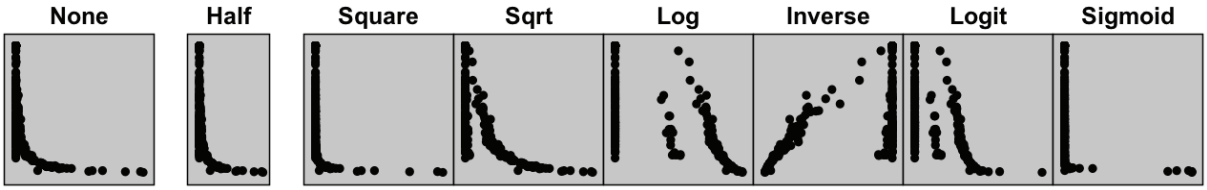

Fig. 1. Scatterplots derived from the Subway data: Different transformations are applied on the horizontal axis revealing different data patterns. In particular, the *log* and *logit* transformations reveal the existence of three distinct clusters. The leftmost cluster contains multiple instances of zero on the horizontal variable; the plot warns us that care must be taken in statistical modeling of this variable. These features are not evident in plots of the raw data.

Abstract— Scagnostics (Scatterplot Diagnostics) were developed by Wilkinson et al., based on an idea of Paul and John Tukey, in order to discern meaningful patterns in large collections of scatterplots. The Tukeys' original idea was intended to overcome the impediments involved in examining large scatterplot matrices (multiplicity of plots and lack of detail). Wilkinson's implementation enabled for the first time scagnostics computations on many points as well as many plots. Unfortunately, scagnostics are sensitive to scale transformations. We illustrate the extent of this sensitivity and show how it is possible to pair statistical transformations with scagnostics to enable discovery of hidden structures in data that are not discernible in untransformed visualizations.

Index Terms—Scagnostics, Scatterplot matrix, Transformation, High-Dimensional Visual Analytics

1 Introduction

While interactive visualization systems have been said to be effective for conducting visual analytics on big data [28, 22], the scale of many datasets precludes exploring visually every variable or relationship among variables. Consequently, systems designed for big data exploration need to filter irrelevant material and discriminate among alternatives in order to present meaningful visualizations in response to user queries. Among several approaches to this problem, Scagnostics [43] were designed to help users navigate through large collections of scatterplots in order to discern meaningful patterns. Figure 2 shows an example.

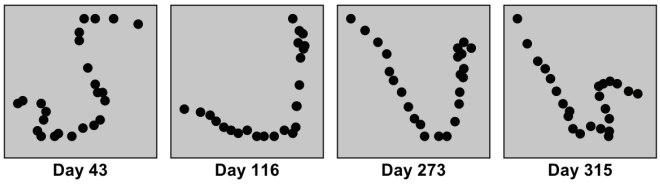

Fig. 2. Scatterplots of barometric pressure vs. air temperature on days 43, 116, 273, and 315 in the Weather data.

The data in Figure 2 consist of hourly meteorological measurements (24 data points in each scatterplot) in 2008 from the Gulf of Maine [23]. There are 50,000 scatterplots derived from this dataset and our query was to find scatterplots with a high degree of "stringy" behavior. Using scagnostics, which parsimoniously encode significant features in the plots, we were able to locate these plots in real time.

Suppose, however, that a stringy feature is embedded in a dense region of a scatterplot so that it is obscured from normal view. In this case, graph-theoretic scagnostics would be unable to detect this feature. If we were to enlarge the local subregion of the plot, however, we might detect this feature with scagnostics after the transformation. That conjecture is the subject of this paper. We propose to transform scagnostics nonlinearly to reveal hidden features. We evaluate the quality of transforms for a particular plot by measuring them against individual scagnostics; the higher the scagnostic value, the more suitable a given transform is for revealing it.

The paper is structured as follows: We describe related work in the following section. Then we introduce our testbed and illustrate it on real datasets. We present test results in Performance. In our Conclusion, we argue that by going beyond classic statistical summaries (means, standard deviations, correlations, etc.), our approach makes it possible to uncover unusual distributions, mixtures of distributions, and other important features hidden in real datasets.

- *Tuan Nhon Dang is with Department of Computer Science, University of Illinois at Chicago, E-mail: nhontuan@gmail.com.*
- *Leland Wilkinson is with Skytree Software Inc. and Department of Computer Science, University of Illinois at Chicago, E-mail: leland@skytree.net.*

Manuscript received 31 Mar. 2014; accepted 1 Aug. 2014. Date of publication 11 Aug. 2014; date of current version 9 Nov. 2014. For information on obtaining reprints of this article, please send e-mail to: tvcg@computer.org.
Digital Object Identifier 10.1109/TVCG.2014.2346572

2 Related Work

We review in this section both feature-based characterizations of scatterplots and the use of transformations to reveal structure or improve statistical inference.

2.1 Scagnostics

In 1989 at a workshop on Computational Statistics, Robustness, and Diagnostics, Paul Tukey presented an idea for characterizing scatterplots [7]. He called it *scagnostics* in order to position it as a special case of what John Tukey had called *cognostics*. The Tukeys intended to characterize a collection of 2D scatterplots through a small number of measures of the pattern of points in these plots. These measures were designed to detect anomalies in density, shape, association, and other features. The Tukeys never published a paper on this topic. Researchers in the workshop were enthusiastic about the idea, although implementing it for larger scatterplots was not practical at the time.

Some years later, Leland Wilkinson, who had been at the workshop, attended a session at the 2003 InfoVis conference and described

the Tukeys' idea. Jinwook Seo and Ben Shneiderman enthusiastically picked up on this comment and presented a paper the following year called *Rank by Feature* [27]. This method relied on classical statistics (means, medians, correlations, etc.) instead of the Tukeys' nonparametric shape descriptors (clumpy, monotonic, convex, etc.), but their implementation supported the effectiveness of charactering scatterplots in order to navigate a large corpus. Subsequently, Wilkinson decided to implement the original Tukey idea through nine Scagnostics defined on planar proximity graphs [43].

Wilkinson's scagnostics measures depend on proximity graphs that are all subsets of the Delaunay triangulation: the minimum spanning tree (MST), the alpha complex [16], and the convex hull [30]. The scagnostics measures are named Outlying, Skewed, Clumpy, Dense, Striated, Convex, Skinny, Stringy, and Monotonic. Figure 3 shows some example scatterplots and their scagnostics. In particular, the scatterplots with a low score on the associated scagnostic are on the left while the scatterplots with a high score on the associated scagnostic are on the right. The implementation of scagnostics are described in detail in [44].

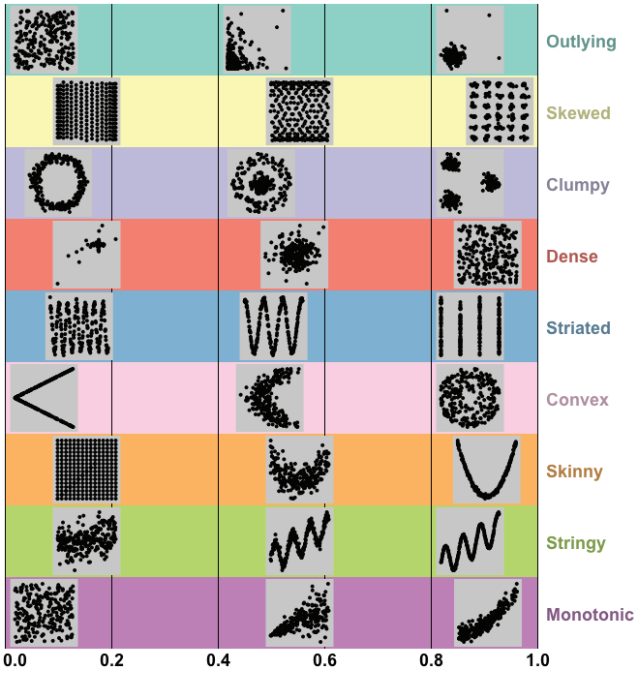

Fig. 3. Some example scatterplots and their scagnostics measures.

Following Wilkinson's work, researchers developed scagnostics-type measures for parallel coordinates [13], pixel displays [26], 3D scatterplots [19], and other graphics [29, 34, 1].

2.2 Transformations

Transformations have been exploited for a number of purposes in statistical and visual analytics.

2.2.1 Statistical Transformations

The use of nonlinear transformations in statistics extends back more than a century. There have been three principal motivations for using these transformations prior to statistical analysis: stabilizing variance (mitigating the dependency between the means and variances in a family of distributions), normalizing or inducing symmetry (ameliorating the biasing effects of skewness on Gaussian statistical methods), and representing discrete data with a continuous model (e.g., *logit* and *probit* analysis).

The classic variance stabilizing transformations have been the arcsine (angular) for proportions and the square root for counts [3]. Examples of normalizing transformations have been Fisher's z for correlations [18] and the Tukey *ladder of powers* set of re-expressions [36]. The famous Box-Cox power transformation [6] is a restriction on the Tukey ladder of powers designed for normally distributed variables (a rare circumstance with real data, despite its popularity in practice). The *logit* and *probit* transformations [5] are used as link functions to enable the fitting of continuous models to discrete data generated by distributions such as the Binomial.

2.2.2 Interactive Graphical Transformations

Statistical software packages were the first to incorporate nonlinear transformations into interactive graphical displays [38, 35, 11, 41]. They chose the Tukey ladder of powers function. Tukey's function is of the form $x^* = x^p$; the parameter p governs the members of the family. These packages use a slider control that maps to this parameter to enable real-time transformations of histograms.

2.2.3 Lensing

Furnas [20] introduced a method for exploring local detail in visualizations (tables, scatterplots, etc.) that employed a lens model to magnify local detail. His method is general because it allows a lens to take a variety of shapes (circle, square, etc.) depending on the distance function chosen for specifying the geometry of the lens. Based on Furnas' idea, researchers have developed specialized lenses for handling very large trees, maps, tables, and documents [10]. The transitions between focus and context of these techniques are achieved through a single dimension (space). Sigma Lens [24] combines other dimensions (translucence and time) to achieve more efficient transitions. Pietriga et al. [25] extends Sigma Lenses framework to provide a unified model that makes it possible to define new focus+context interaction techniques as independent as possible of the representation and graphics library employed.

2.2.4 Aspect Ratio Selection

Changing the aspect ratio of a plot is a linear transformation. In 1988, the problem of selecting the aspect ratio of a line chart was first discussed rigorously by Cleveland et al. [9]. The authors demonstrated how the choice of a line chart's aspect ratio can impact graphical perception of trends in time series data. They then proposed a technique called *banking*. The basic idea underlying this technique is that the slopes in a line chart are most readable when the average orientation of all line segments in a chart is 45°. This suggests that the aspect ratio of plots could be determined by setting the median slope of all line segments to 1. They call this approach *median absolute slope* [9]. Cleveland later suggested a weighted version of this method, *length-weighted average orientation* [8]. In this new version, the length-weighted mean of the absolute orientations of the line segments is set to 45 degrees.

Heer and Agrawala [21] extended Cleveland's work in two ways. First, the authors described 12 different banking algorithms that represent alternate optimization criteria for Cleveland's banking procedure. These criteria are designed to find an aspect ratio that further improves the visual perception of line segment orientations. Second, they developed *multi-scale banking*, a technique that combined spectral analysis and banking to 45° to automate the selection of aspect ratios for different levels of granularity. This technique automatically identifies trends at various frequencies that may be of interest and then generates a banked chart for each of these scales.

The above approaches for automatically selecting the aspect ratio of a line chart have several shortcomings. First, the way a curve is approximated by line segments can dramatically change the selected aspect ratio. Second, they don't preserve semantically symmetric shapes. To address these problems, Talbot et al. [33] offered a method for selecting the aspect ratio for line charts that minimizes the arc length of the plotted curve while keeping the area of the plot constant. This approach is parameterization invariant (redundant line vertices do not influence the result), robust to a wide range of inputs, and responsive to visual symmetries in the data.

Recently, Fink et al. [17] extended Cleveland's aspect ratio argument to general scatterplots. The basic idea behind this method is to select an aspect ratio such that the resulting scatter plot optimizes

a feature of a Delaunay triangulation on the points in the plot. The authors defined six different measures on the Delaunay triangulation. They chose the Delaunay triangulation because they believed it to be a conceptually meaningful structure for representing human perceptual grouping [14]. However, while the scagnostics features were based on the types of configurations statisticians and analysts attend to, and while the Cleveland paradigm was based on the accuracy of slope judgments, the measures used in this study were based on subjective judgments, by relative novices, of the "meaningfulness" of given plots under various aspect-ratio transformations. It is not clear how their methods could be appropriately applied to the wide range of configurations encountered in an exploratory data analysis environment.

3 Scagnostic Transformations

Transforming variables affects densities, relative distances, and orientations of points within a scatterplot. Consequently, it has a significant impact on our ability to perceive patterns in the data. Figure 4 shows examples of the proximity graphs of *square root* and *log* transformations applied on the same set of data points in the New York City subway dataset [40]. The plotted variables are station versus ridership in 1918. Different proximity graphs produce different sets of scagnostics.

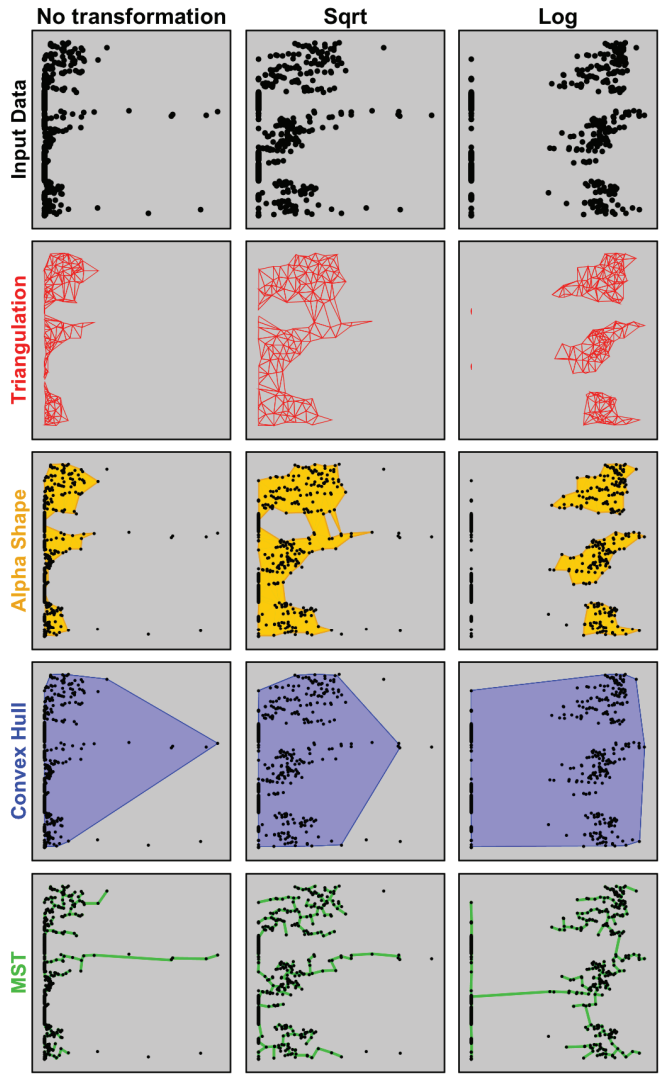

Fig. 4. Visualization of the Subway data: Transforming a point set changes its Delaunay triangulation and the three geometric graphs. The top left frame is the unaltered scatterplot of station and subway ridership in New York City in 1918. The next two scatterplots are the results of applying *square root* and *log* transformations on X-axis.

3.1 Choice of Transformation

Choosing appropriate transformations must be guided by some justifiable principles. First of all, the classical statistical transformations arose out of experiences applying models based on theoretical distributions to real data; we should give them serious consideration. Second, the transformations we choose ought to cover the full range of negative to positive skewness as well as mixtures of distributions that are relatively symmetric. Third, we should take care to make our portfolio of transformations approximately mirror-symmetric; in a plot of x^* against x, where x^* is the transformed value of x, they should reflect around a diagonal representing the identity transformation.

The transformations we chose are shown in Figure 5.

- *none*: $x^* = x$ (leaves points unchanged)
- *half*: $x^* = x/2$ (squeezes all points together)
- *square*: $x^* = x^2$ (pulls points toward left of frame)
- *square root*: $x^* = \sqrt{x}$ (mildly pulls points toward right of frame)
- *log*: $x^* = log(x)$ (strongly pulls points toward right of frame)
- *inverse*: $x^* = 1/x$ (reverses scale and squeezes points into left of frame)
- *logit*: $x^* = (log(x/(1-x)) + 10)/20$ (squeezes points toward middle of frame)
- *sigmoid*: $x^* = 1/(1 + exp(-20x + 10))$ (expands points away from middle of frame)

The *half* transformation implements an aspect ratio of 1:2 when applied to Y and 2:1 when applied to X. The parameters for the *logit* and *sigmoid* functions govern their curvature; they were designed to make the curves mirror-symmetric and the slopes sufficient to cover areas not spanned by the other functions.

We apply these 8 transformations independently on both X and Y axes. Therefore, we have 64 different combinations to consider when we compute scagnostics. Before transformation, the values are normalized to the unit interval (as in the original scagnostics paper [43]). After transformation, the values are renormalized to the unit interval.

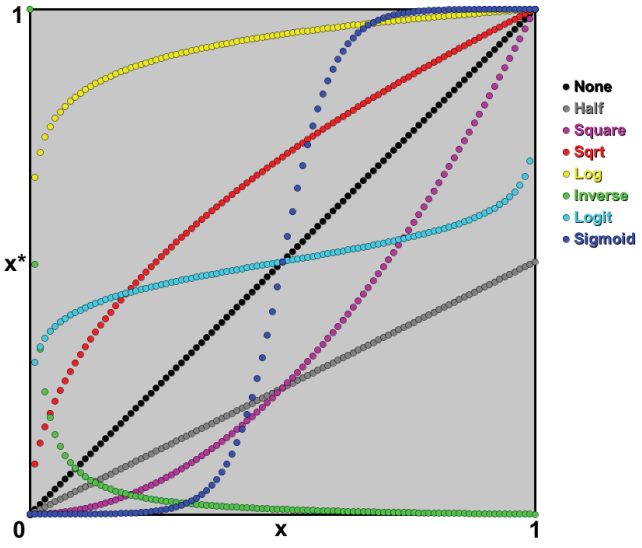

Fig. 5. Transformation functions.

3.2 Datasets

We will illustrate the effects of these transformations mainly through real data examples. We use datasets retrieved from the UCI Repository [4] and other sources to demonstrate the performance of our approach. Table 1 summarizes prominent aspects of these datasets ordered by the number of attributes.

Table 1. Characteristics of datasets used for testings and demonstrations in the following sections.

Datasets	#Instances	#Variables	#Scatterplots
Sleep	62	10	45
Page Blocks	5,473	10	45
Segment	2,309	20	190
Web Statistics	300	22	231
Water Treatment	527	38	703
Subway	423	104	5,356
Communities	1,994	128	8,128
Gas Sensor	3,600	128	8,128
Musk	476	167	13,861
Isolet	1,559	167	13,861
Madelon	1,042	500	124,750

3.3 Examples

In this section, we demonstrate the application on real datasets using a simple testbed. A user first selects a scagnostic measure such as Clumpy. The program then applies the combinations of transformations on the X and Y axes and computes the selected measure on each combination.

3.3.1 Striated

Figure 6(a) shows an example of computing the Striated measure on the Madelon data [4]. We have selected variable 41 vs. variable 48 to illustrate how transformations can reveal striations that are not evident in the raw plot. The gray color scale is adopted to highlight the Striated measure (dark plots are high Striated, white plots are low Striated). Because there are 1,042 data points highly concentrated in the middle of the scatterplot, we do not notice the striation; in fact, our impression is that the raw plot in the upper left is spherical Gaussian. Only after applying the *sigmoid* transformation do we discern the regular stripes in the middle. This plot is not based on Gaussians.

Figure 6(b) shows another example of computing the Striated measure on the Water Treatment data [4]. We have selected conductivity vs. suspended solids to primary settler. In this example, the striation is revealed when applying *inverse* on the X-axis.

3.3.2 Clumpy

Figure 7 shows an example of the Clumpy measure on the Web Statistics data[1]. We have selected a pair of variables (mean of connecting time vs. standard deviation of connecting time) to illustrate the potential for discerning clusters after transformation. The gray color scale is adopted to highlight the Clumpy measure (dark backgrounds are high Clumpy, white backgrounds are low Clumpy). When we examine the *inverse*/*inverse* (or *log*/*inverse*) plot, the two clusters are most visible.

The *sigmoid* tends to draw points apart, but it will not produce clusters when they do not exist in the data. Unless there is a margin at the center separating points, a gap will not appear under these transformations. This observation applies to nonlinear transformations used in SVMs as well. Figure 8 shows an example where *sigmoid* transformation reveals the three clusters which are not apparent in the regular scatterplot due to the high concentration of data points in the center of Y-axis.

3.3.3 Dense

Figure 9 shows an example of the Dense measure applied to the Page Blocks data. We have selected a pair of variables to highlight this feature (eccentricity of the block (length / height) vs. total number of black pixels in the original bitmap of the block). As usual, the gray color scale is adopted to highlight the measure (black plots are high Dense, white plots are low Dense). The data points are colored by their classes. We note that the *inverse*/*inverse* transformation reveals class separation not evident in the regular scatterplot.

[1] http://davis.wpi.edu/xmdv/datasets/webstats.html

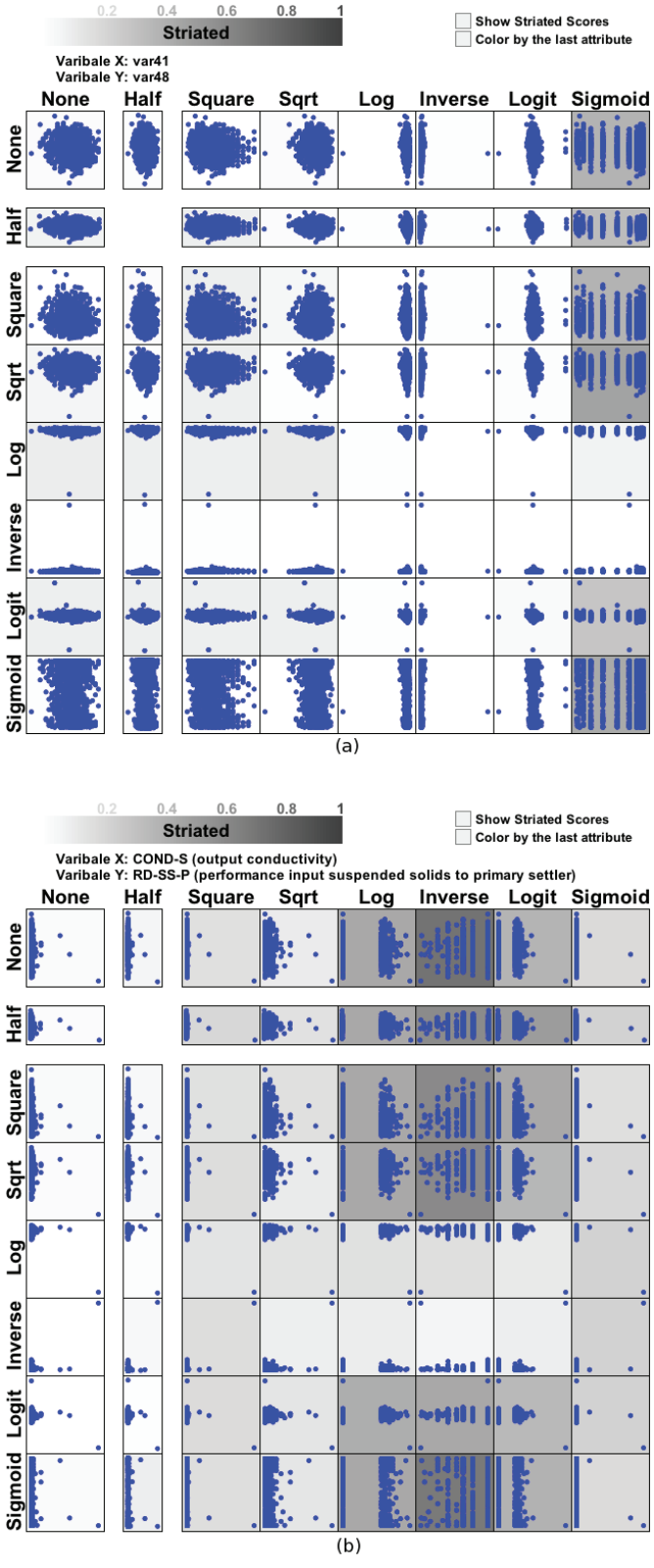

Fig. 6. Inspecting Striated measure on: a) The Madelon data: The selected variables are variable 41 vs. variable 48 b) The Water Treatment data: The selected variables are output conductivity vs. performance input suspended solids to primary settler.

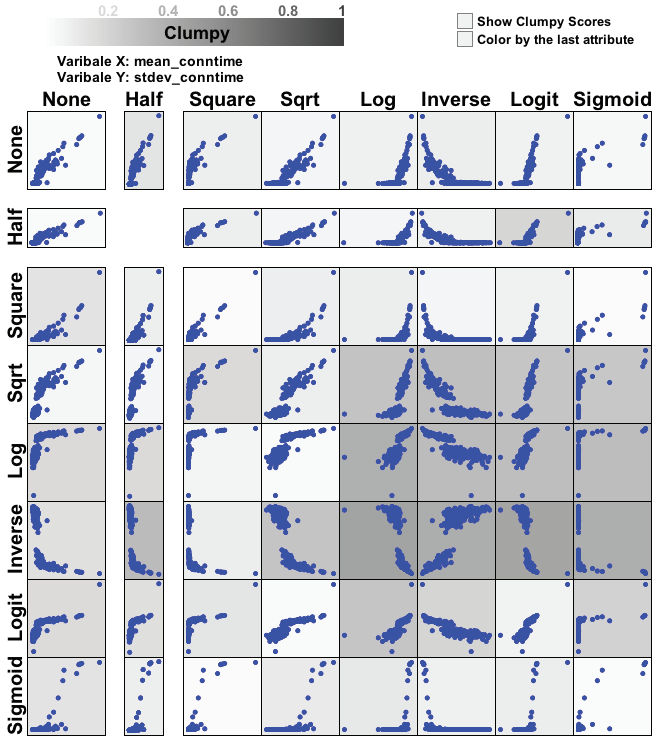

Fig. 7. Inspecting the Clumpy measure on the Web Statistics data: The selected variables are mean of connecting time vs. standard deviation of connecting time. The two clusters are most visible in the *inverse*/*inverse* (or *log*/*inverse*) plot.

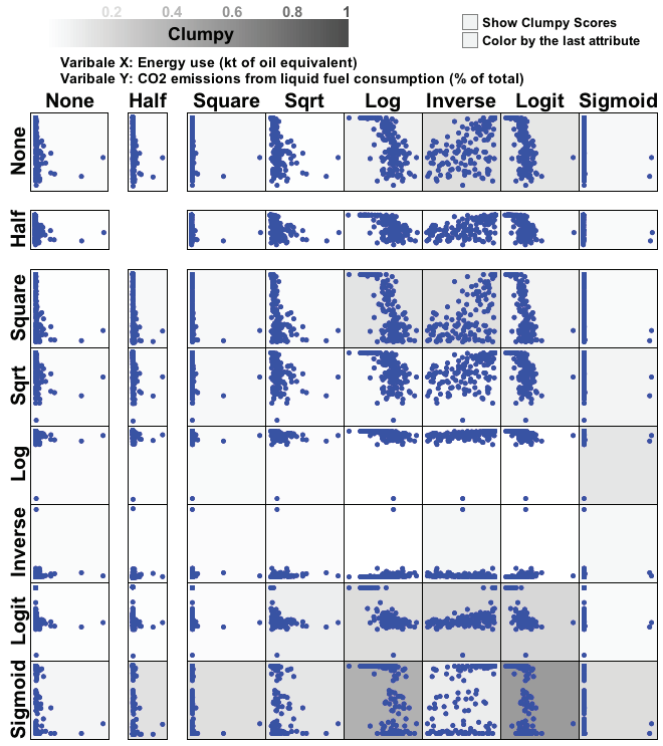

Fig. 8. Inspecting the Clumpy measure on the World Bank data: The plotted variables are energy use vs. CO2 emissions from liquid fuel consumption. Each data point is a country.

This is an especially interesting example because it illustrates the importance of transformation prior to classification when there are nonlinear boundaries separating classes. Support Vector Machines, for example, exploit nonlinear transformations in reproducing kernel Hilbert space in order to handle nonlinear margins between classes [37]. We ran a canonical discriminant analysis on these variables in SYSTAT [42]. The error rate on the untransformed variables was 61 percent (worse than chance)! The error rate on the *inverse* transformed variables was 28 percent.

A denser scatterplot does not suggest class separation, but there is a better chance to separate classes when data points are spread out. We can also define a new feature, says class separation feature, which measures the overlapped area among alpha shapes of different classes [39]. That is, the use of our proposed approach is not limited to the nine scagnostics. We might define other measures specialized for an application domain and use the proposed approach to find the transformation maximizing/minimizing these features. Moreover, we can combine multiple features to highlight a pattern. For example, maximizing the Dense feature before minimizing overlapped area among alpha shapes of different classes results in better chances of finding class separation in a scatterplot.

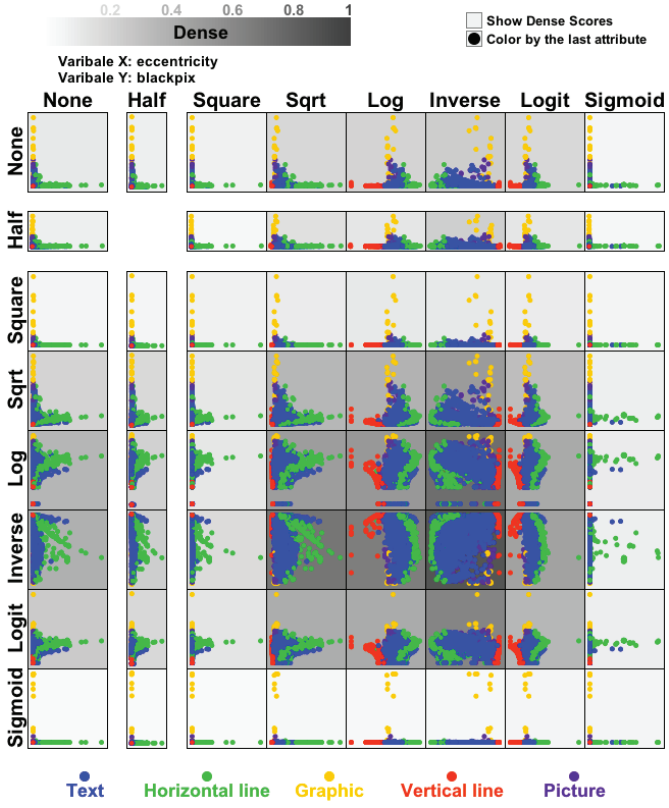

Fig. 9. Inspecting the Dense measure in the Page Blocks data: The selected variables are eccentricity of the block (length / height) vs. total number of black pixels in the original bitmap of the block. The data points are colored by their classes as depicted at the bottom.

3.3.4 Outlying

Figure 10 shows an example of the Outlying measure applied to the Sleep data [2]. We have selected the body weight and brain weight variables. This example gives us the opportunity to invert our usual process. In this case we want to find a transformation that *minimizes* outliers. Thus, we want to look at the frames with the lightest background. The *log*/*log* transformation is consistent with the appropriate statistical analysis of these two variables in the original paper and is indeed among the lightest of the frames. Note that this *log*/*log* plot

linearizes the relationship as well. The so-called outliers in the raw plot (upper left) are not outliers after proper transformation.

Since most transformations, such as *square root*, *log*, and *inverse* squeeze points together, most outliers are artifacts when these transformations are applied. We don't suggest to apply transformations to find outliers.

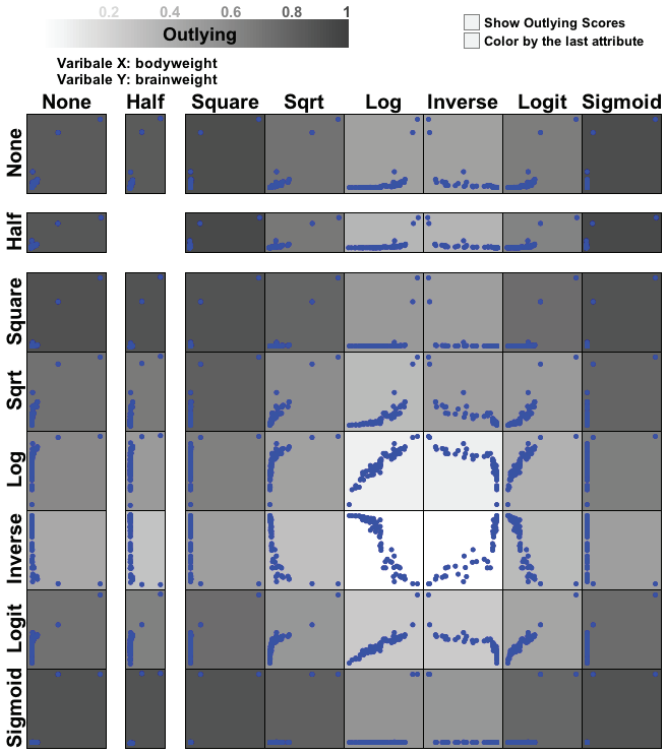

Fig. 10. Inspecting the Outlying measure on the Sleep data: The selected variables are body weight vs. brain weight.

3.3.5 Monotonic

Figure 11 shows an example of the Monotonic measure applied to the World Bank data[2]. We have selected rural population vs. urban population. The Monotonic is our only coefficient not based on a subset of the Delaunay graph. However similar to other visual features, a few outliers are sufficient to distort this property as depicted in the regular scatterplot (top left). The outliers in this case are China and India. In this example, the strong correlation of the two variables is revealed when we squeeze together initially high values (this helps to reduce the impact of outliers).

4 Performance

In this section we focus on evaluating the performance of scagnostics transformations. We investigated the performance of our approach on large data in terms of n (number of observations) and p (number of scatterplots). All tests were performed on a 2.3 GHz Intel Core i5, Mac OS X Version 10.7.5, 4 GB RAM running Java 1.6 and Processing 1.5.1.

4.1 Overall Running Times

The graphs in Figure 12 show computation time broken down into the time to read and transform data, the time to bin the data points, and the time to compute scagnostics. Transformations are applied on one dimension at a time. Here are some observations from empirical analysis:

[2] http://data.worldbank.org/indicator

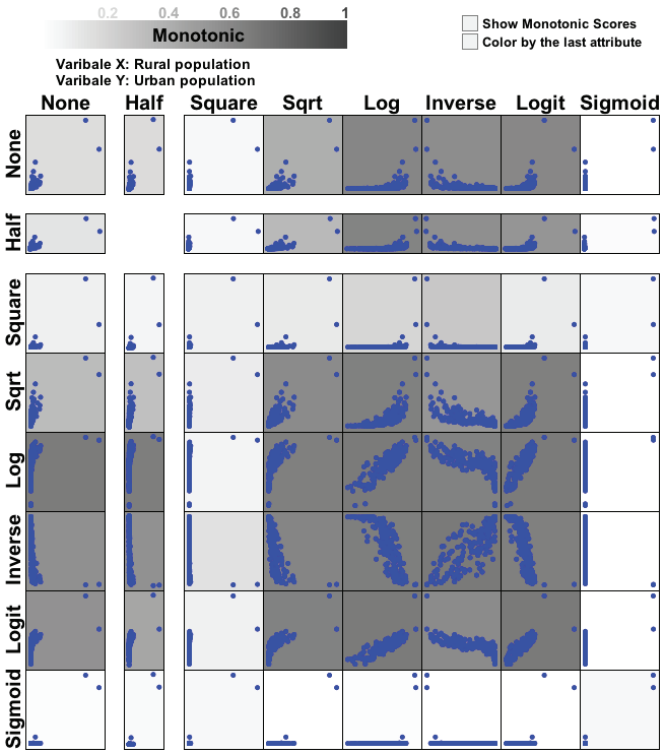

Fig. 11. Inspecting the Monotonic measure on the World Bank data: The selected variables are rural population vs. urban population in 2010. Each data point is a country.

- The bottleneck of our approach is at the stage of computing scagnostics. However, this stage is completely parallelizable.
- The time for binning is linearly dependent on n. In contrast, the time for computing scagnostics is almost independent on n since the three proximity graphs are computed on binned data, not the original data points.
- The scagnostics computation time for each dataset is linearly dependent on p. In Figure 12, the two datasets at the bottom (Communities and Gas Sensor) contain the same number of variables (128 variables or 8,128 scatterplots), and contain respectively 1,994 and 3,600 observations. The two datasets on the top (Musk and Isolet) contain the same number of variables (167 variables or 13,861 scatterplots), and contain respectively 476 and 1,559 observations. On average, the scagnostics computation time for the Musk and Isolet datasets is nearly twice of the computation time for the Communities and the Gas Sensor datasets.

4.2 Scagnostics computation times of different transformation combinations

We now inspect the Scagnostics computation times of different transformation combinations. The four datasets used in Section 4.1 are reused for this purpose. However, we use only the first 100 variables and the first 400 instances in each dataset so that we can compare the Scagnostics computation times across the four datasets.

Figure 13 shows the test results. In particular, the matrices on the right show the Scagnostics computation times of 64 combinations. The layouts on the left summarize 4950 scatterplots (100 variables) in each dataset [12]. In other words, the layouts display 10 to 15 exemplar scatterplots (for each dataset) which represent 4950 scatterplots. The size of an exemplar scatterplot denotes the size of its cluster.

Here are some observations from empirical analysis:

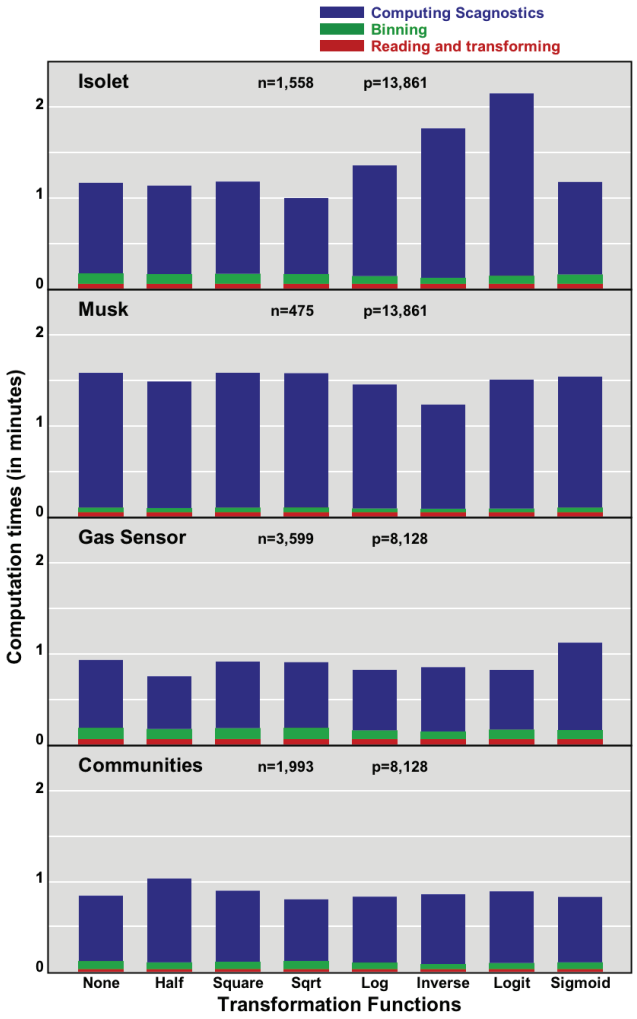

Fig. 12. Computation times (in minutes) for datasets retrieved from the UCI Repository: n is the number of observations and p is the number of scatterplots.

- On average of the four datasets, the *inverse*/*inverse* transformation is the lowest computation time because this transformation combination tries to pull the data points into one corner of a scatterplot. This effect is prominent on the Isolet dataset.

- With the same number observations (n=400), denser scatterplots require more time to compute scagnostics. In the left layouts of Figure 13, the Isolet dataset contains much denser scatterplots compared to the Gas Sensor dataset.

4.3 Demonstrative case study

This section presents a demonstration of our approach being meaningfully deployed in practice. We use the World Bank data. In particular, we inspect 100 variables in the Economy and Growth section, such as exports of goods and services, gross capital formation, and GDP per capita. We first obtain scagnostics of all scatterplots and their transformations. Then we compute the average for each transformation as depicted in Figure 14. The plotted measure is Monotonic. The most significant gain on Monotonic is *log*/*log* transformation. This suggests that the data can benefit from the *log*/*log* transformation when we look for trends.

In Figure 15, we plot the first twelve variables in the data. The lower triangle shows the regular scatterplot matrix. The upper triangle shows their *log*/*log* transformation. The gray color scale is used

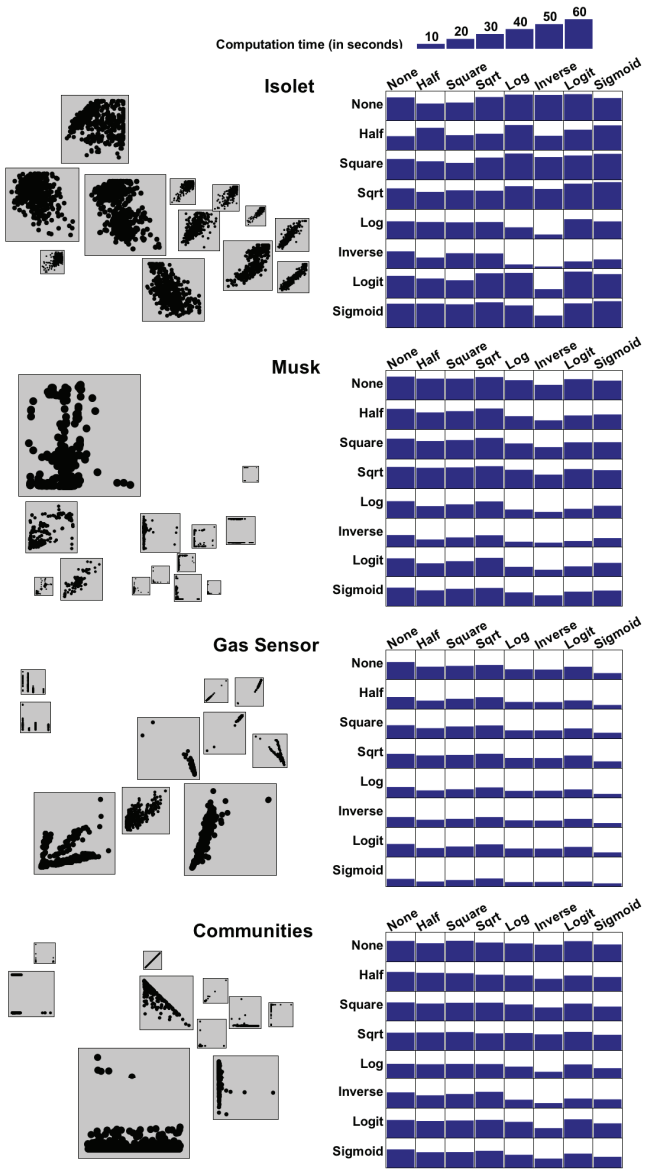

Fig. 13. Computation times of different transformation combinations for four datasets retrieved from the UCI Repository.

	None	Half	Square	Sqrt	Log	Inverse	Logit	Sigmoid
None	0.1	0.09	0.06	0.15	0.26	0.22	0.2	0.03
Half	0.09		0.06	0.13	0.22	0.2	0.16	0.04
Square	0.07	0.07	0.07	0.05	0.07	0.07	0.05	0.06
Sqrt	0.14	0.12	0.05	0.23	0.38	0.25	0.34	0.01
Log	0.23	0.2	0.07	0.38	0.42	0.37	0.41	0.01
Inverse	0.22	0.21	0.08	0.31	0.39	0.38	0.37	0.02
Logit	0.18	0.15	0.06	0.33	0.41	0.33	0.4	0.01
Sigmoid	0.03	0.04	0.06	0.01	0.01	0.02	0.01	0.05

Fig. 14. Average Monotonic for 100 variables in the Economy and Growth section of the World Bank data.

to highlight the Monotonic measure. The strong correlation of different economical sectors is distorted in the regular scatterplot matrix because of outliers (strong countries such as the US and China).

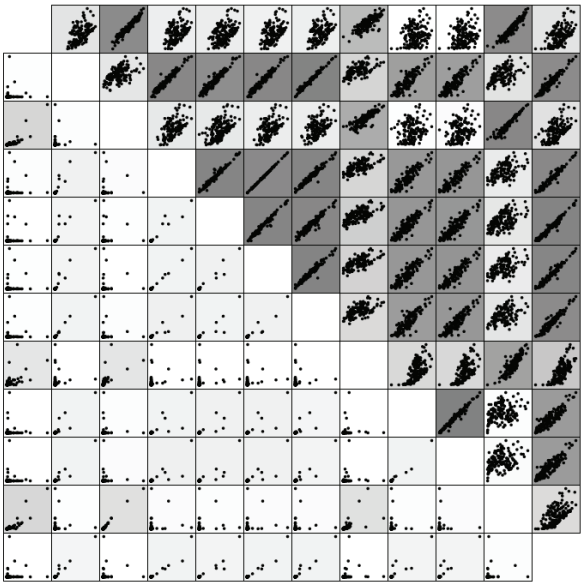

Fig. 15. Inspecting the Monotonic measure of 12 variables in the Economy and Growth section of the World Bank data. The lower triangle shows the regular scatterplot matrix. The upper triangle shows their *log*/*log* transformation.

5 Conclusion

We have presented a method for selecting the best transformation to reveal scagnostics hidden in a scatterplot. The basic idea behind our method is to use scagnostics as the measure of goodness for selecting a transformation. We illustrated our approach on real datasets in Section 3 and evaluated it in Section 4.

While we developed a testbed for illustrating scagnostics transforms in this paper, we do not regard it as an end-user application. Instead, we believe that the algorithms outlined in this paper could be used to embed transformed scagnostics analytics in visual analytics platforms such as Jigsaw [31], Tableau [32], or Xmdv [15].

Figure 16 shows one possible interface that provides filtering, brushing and linking for exploratory scagnostics transformations. This figure is based on the Segment dataset from the UCI Repository [4]. In particular, Figure 16(a) shows the scatterplot matrix of 20 variables in the Segment dataset. Each scatterplot is colored by its Dense score. In Figure 16(b), we filter and display only the scatterplots which gain at least 0.3 on the Dense feature when applying transformations. The brushed cell is bordered in red. The transformation matrix plot in Figure 16(c) instantly shows the selected scatterplot under all pairs of transformations. We say instantly, because the scagnostic plots are buffered as thumbnail bitmaps at this stage of processing. Preprocessing involves computing all the scagnostics under all the transformations, which takes more time (see Figure 12). This preprocessing step must be done only once for a given set of data, however. In Figure 16(c), we also request to show the Dense scores on top of each transformed scatterplot. Data points are colored by their class attribute.

One might ask what the difference is between our approach and the use of transformations in classical statistics or machine learning. First of all, any transformation used in an analytic must be evaluated in terms of a figure of merit. For classical statistical analyses, the figure of merit is the closeness of the distribution of residuals to a specified distribution (normal, exponential, etc.). In the case of the popular Box-Cox method for analysis of variance [6], this figure is arrived at through maximum likelihood. In other cases, as in the Tukey ladder of powers, it is inspected through plots. Alternative evaluation

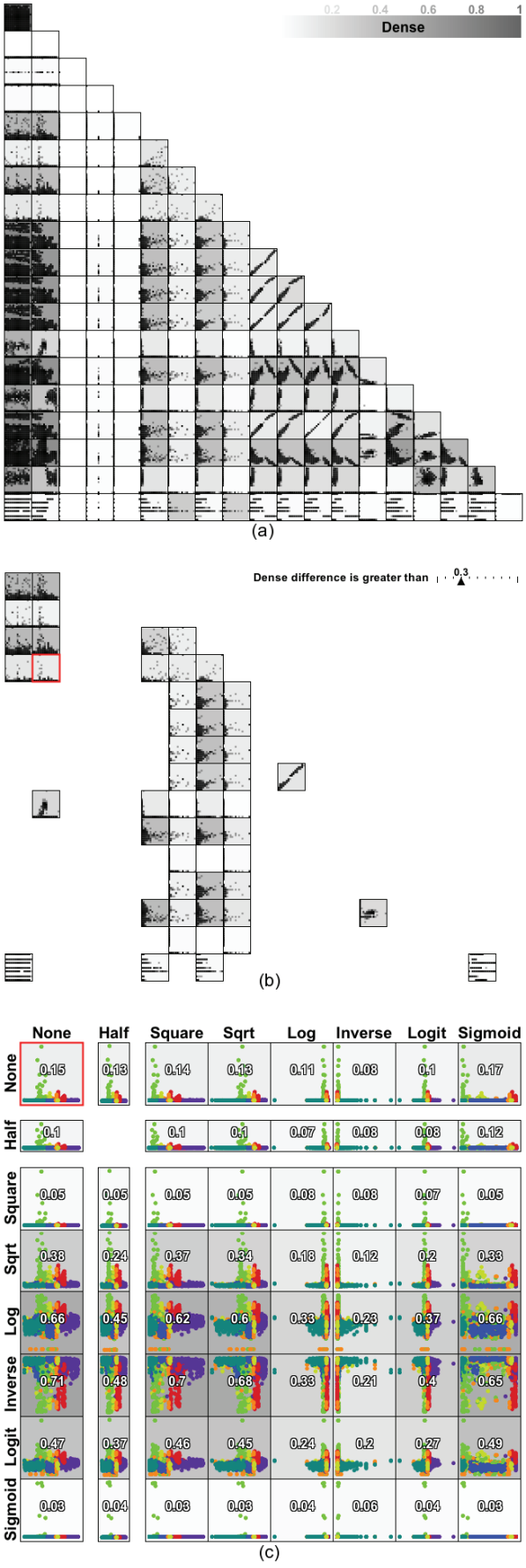

Fig. 16. Possible application allowing filtering, brushing and linking. This figure is based on the Segment dataset from the UCI Repository.

criteria for classical or exploratory methods include linearity of the model, homoscedasticity, or independence of residuals.

Scagnostics transformations involve a different figure of merit for each scagnostic, and none of these, to the best of our knowledge, is found in the analytic transformation literature. We have seen in some cases (e.g., the Dense measure) that a scagnostic transformation might be of use in standard parametric or nonparametric analyses. The primary goal of these transformations, however, is to serve a visual analytics strategy that usually precedes formal modeling. Visual analytics can be useful not only for guiding us to supportable conclusions, but also for guiding us to proper models.

REFERENCES

[1] G. Albuquerque, M. Eisemann, and M. Magnor. Perception-based visual quality measures. In *IEEE VAST*, pages 13–20, 2011.

[2] T. Allison and D. Cicchetti. Sleep in mammals: Ecological and constitutional correlates. *Science*, 194:732–734, 1976.

[3] F. Anscombe. The transformation of poisson, binomial and negative-binomial data. *Biometrika*, 35:246–254, 1948.

[4] A. Asuncion and D. Newman. UCI machine learning repository, 2007.

[5] J. Berkson. Application of the logistic function to bio-assay. *Journal of the American Statistical Association*, 39:357–365, 1944.

[6] G. Box and D. Cox. An analysis of transformations. *Journal of the Royal Statistical Society, B*, 26:211–252, 1964.

[7] A. Buja and P. Tukey. *Computing and Graphics in Statistics*. IMA volumes in mathematics and its applications. Springer-Verlag, 1991.

[8] W. S. Cleveland. A Model for Studying Display Methods of Statistical Graphics. *Journal of Computational and Statistical Graphics*, 2:323–364, 1993.

[9] W. S. Cleveland, M. E. McGill, and R. McGill. The shape parameter of a two-variable graph. *Journal of the American Statistical Association*, 83:289–300, 1988.

[10] A. Cockburn, A. Karlson, and B. B. Bederson. A review of overview+detail, zooming, and focus+context interfaces. *ACM Comput. Surv.*, 41(1):2:1–2:31, Jan. 2009.

[11] A. B. D. F. Swayne, D. Cook. Xgobi: Interactive dynamic graphics in the x window system with a link to s]. In *Proceedings of the 1991 American Statistical Association Meetings*. American Statistical Association, 1991.

[12] T. N. Dang and L. Wilkinson. Scagexplorer: Exploring scatterplots by their scagnostics. In *Proceedings of the 7th Pacific Visualization Symposium (PacificVis)*, 2014.

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

[14] M. Dry, D. Navarro, K. Preiss, and M. Lee. The Perceptual Organization of Point Constellations. In *Annual Meeting of the Cognitive Science Society*, 2009.

[15] J. Y. E. A. Rundensteiner, M. O. Ward and P. R. Doshi. Xmdv-Tool: visual interactive data exploration and trend discovery of high dimensional data sets. In *Proceedings of the 2002 ACM SIGMOD international conference on Management of data*, pages 631–631. ACM, 2002.

[16] 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.

[17] M. Fink, J.-H. Haunert, J. Spoerhase, and A. Wolff. Selecting the aspect ratio of a scatter plot based on its delaunay triangulation. *Visualization and Computer Graphics, IEEE Transactions on*, 19(12):2326–2335, 2013.

[18] R. Fisher. The general sampling distribution of the multiple correlation coefficient. *Proceedings of the Royal Society of London, Series A*, 121:654–673, 1928.

[19] L. Fu. Implementation of three-dimensional scagnostics. Master's thesis, University of Waterloo, Department of Mathematics, 2009.

[20] G. W. Furnas. Generalized fisheye views. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '86, pages 16–23, New York, NY, USA, 1986. ACM.

[21] J. Heer and M. Agrawala. Multi-scale banking to 45 degrees. *Visualization and Computer Graphics, IEEE Transactions on*, 12(5):701–708, 2006.

[22] D. A. Keim, H. Qu, and K.-L. Ma. Big-data visualization. *IEEE Computer Graphics and Applications*, 33:20–21, 2013.

[23] U. of Maine School of Marine Sciences. Weather dataset. `http://gyre.umeoce.maine.edu/buoyhome.php`.

[24] E. Pietriga and C. Appert. Sigma lenses: Focus-context transitions combining space, time and translucence. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, pages 1343–1352, New York, NY, USA, 2008. ACM.

[25] E. Pietriga, O. Bau, and C. Appert. Representation-independent in-place magnification with sigma lenses. *IEEE Transactions on Visualization and Computer Graphics*, 16(3):455–467, May 2010.

[26] 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.

[27] J. Seo and B. Shneiderman. A rank-by-feature framework for interactive exploration of multidimensional data. *Information Visualization*, 4(2):96–113, July 2005.

[28] B. Shneiderman. The big picture for big data: Visualization. *Science*, 343:730, February 2014.

[29] M. Sips, B. Neubert, and J. Lewis. Selecting good views of high-dimensional data using class consistency. *Computer Graphics Forum*, 28:831–838, 2009.

[30] S. S. Skiena. *The Algorithm Design Manual*. Springer-Verlag New York, Inc., New York, NY, USA, 1998.

[31] J. Stasko, C. Görg, and Z. Liu. Jigsaw: Supporting investigative analysis through interactive visualization. *Information Visualization*, 7(2):118–132, Apr. 2008.

[32] Tableau Software. Tableau. http://www.tableausoftware.com.

[33] J. Talbot, J. Gerth, and P. Hanrahan. Arc length-based aspect ratio selection. *Visualization and Computer Graphics, IEEE Transactions on*, 17(12):2276–2282, 2011.

[34] A. Tatu, G. Albuquerque, M. Eisemann, P. Bak, H. Theisel, M. Magnor, and D. Keim. Automated analytical methods to support visual exploration of high-dimensional data. *IEEE Transactions on Visualization and Computer Graphics*, 17:584–597, 2011.

[35] L. Tierney. *LISP-STAT: An Object-Oriented Environment for Statistical Computing and Dynamic Graphics*. John Wiley & Sons, New York, 1990.

[36] J. W. Tukey. On the comparative anatomy of transformations. *Annals of Mathematical Statistics*, 28:602–632, 1957.

[37] V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag New York, Inc., New York, NY, USA, 2nd edition, 1999.

[38] P. Velleman. *Data Desk [computer software]*. Ithaca, NY, 1998.

[39] T. von Landesberger, S. Bremm, P. Rezaei, and T. Schreck. Visual analytics of time dependent 2d point clouds. In *Proceedings of the 2009 Computer Graphics International Conference*, CGI '09, pages 97–101, New York, NY, USA, 2009. ACM.

[40] M. O. Ward. New york city dataset. `http://davis.wpi.edu/~xmdv/datasets/subway.html`.

[41] L. Wilkinson. *SYSTAT for Macintosh, Version 5.0*. SYSTAT Inc., Evanston, IL, 1986.

[42] L. Wilkinson. *SYSTAT, Version 10*. SPSS Inc., Chicago, IL, 1998.

[43] 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.

[44] 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.