An Awesome Title

Anushka Anand and Justin Talbot

Abstract— Some place holder text...

Index Terms—High-dimensional data, Decision trees, Visual Analytics

1 Introduction

Exploratory Data Analysis (EDA), championed by Tukey [33], relies heavily on visual representations in the search for informative and potentially surprising structure in data. Such analysis usually starts with an overview of the data dimensions of interest and follows a path of progressive refinement getting more focused or detailed based on the question being asked. A common strategy for visual exploration and analysis of multidimensional datasets is to examine two-dimensional projections often through selecting "interesting" axis-aligned projections [28, 37, 29]. Another popular approach is examining axis-unaligned projections through dimensionality reduction techniques [13, 40, 26]. However, there is little work in exploring how to slice two-dimensional projections into multiple plots with "interesting" visual structure. Here we propose a methodology, applying the principle of statistical significance testing, to determine how to select such slices.

Faceting is a slicing operator [36, 21] to split up a dataset into subsets that are examined together. Often these subsets are determined by discrete values of a categorical dimension or bins of a quantitative dimension. The result of faceting is expressed visually as *Small Multiples* [32] or *Collections*[5] and *Trellis* displays [4]. Small multiples facilitate finding structure and patterns in complex data by forming tables of simple views directly depicting comparisons across dimensions in the data. They increase the number of dimensions that can be easily visually processed and are applied in visual data analysis tools across different application domains such as geography [14, 19] and medicine [18, 23]. Furthermore, visual analysis tools such as the ggplot2 library in the R language [34] and Tableau [30] allow users to rapidly generate small multiple visualizations to explore data. However, with limited prior knowledge about the interaction effects of dimensions in a dataset, it becomes an exercise of trial and error to find the *Partitioner* dimension that reveals "interesting" structure in the subsets resulting from faceting.

We propose a method of selecting a Partitioner dimension to facet a given bivariate relationship by using significance tests of a particular score metric of the resulting splits. Our contributions are:

- A set of goodness criteria for the collection of splits resulting from faceting with a particular Partitioner dimension.
- A method for quantitatively evaluating the quality of the splits. We compare the score metric against reference distributions of the metric from bootstrapped random sample splits which act as "null splits".

The next section summarizes related work on visual explanations for multivariate data analysis. This is followed by a description of the method we propose and discussion of metrics we use. Then we describe examples using our method to explain high-dimensional structure in a number of datasets. Finally, we draw conclusions from this research and outline future work.

- *Anushka Anand is with Tableau Research. E-mail: aanand@tableau.com.*
- *Justin Talbot is with Tableau Research. E-mail: jtalbot@tableau.com.*

Manuscript received 31 Mar. 2015; accepted 1 Aug. 2015; date of publication xx xxx 2015; date of current version xx xxx 2015.
For information on obtaining reprints of this article, please send e-mail to: tvcg@computer.org.

2 Related Work

In this section, we first summarize previous work in three related areas: visual inference methods, quality metrics for data visualizations and work related to Scagnostics.

2.1 Visual Inference Methods

Bridging the gap between exploratory and confirmatory statistics is work that investigates statistical significance testing in the hypothesis testing of visual findings [35, 20]. Human subjects are asked whether the observed dataset looks anything like random bootstrapped samples in lineup or Rorschach protocols to enable simulation based statistical inference of visual patterns [9].

"New Procedures for Visualizing Data and Diagnosing Regression Models" by R Menjoge. This is a thesis. Chapter 2 has a very interesting approach using bootstrapping to get a 95% visual confidence interval for a single plot. It's like the effect size version of Buja's null hypothesis approach. I can't find a published paper other than the thesis.

2.2 Data Visualization Quality Metrics

A large number of quality metrics are used in methods for high-dimensional data analysis [7]. Many metrics for scatterplots determine projections of the data to be displayed, often for particular tasks: cluster separation [27, 31], class consistency and separation [29, 24], interesting visual shapes [37] or statistical properties [16, 28]. Metrics for parallel coordinate plots [3, 12, 15, 40] also focus on the ordering of dimensions. Some metrics [6, 10] focus on the level of abstraction, including aggregration and sampling, in these chart types. Others [1, 3, 25, 40] offer metrics for radial displays, pixel maps, table lens and other visualization types.

Tableau's Show Me [**?**] suggests appropriate chart types based on a set of two or more user-selected data fields. When the user selects a chart type, the system automatically creates a view for that chart type. However, for each chart type there are multiple possible effective views based on different visual mappings and dimension orderings of the selected fields. Our work extends automatic data visualization generation to support small multiples alternatives by determining appropriate visual encodings, level of abstraction and ordering of data fields. We contribute scoring criteria to rank data views considering data properties and visual layout.

Rank-by-feature [28] (rank all 2D projections). a follow up [22] that claims they do rank-by-feature on subsets and compares subsets. I thought this would be closely related to our work, but I can't figure out what the paper is doing.

Evolutionary selection [8] of linear and nonlinear 2D projections.

2.3 Scagnostics

Graph-theoretic scagnostics [37], Scagnostic distributions [38], ScagExplorer [11]

Experimental evaluation [17] of scagnostics and other metrics, showing that scagnostics best matches human rankings.

AutoVis [39]

3 Methodology

Multivariate explanations as a worldview task [2] often not supported by visualization tools. However, their definition is limited in scope to correlation models involving more than two measures...

The goal is to find a Partitioner variable that produces splits of the data that are more informative than random splits. We set up the following goodness-of-split criteria to guide our work:

- Support: An indicator of the strength of the relationship or visual pattern is the proportion of data points that occur in the split and contribute to the pattern.
- Diversity: Robustness to overfitting? The pattern in the splits would be different from the original and from each other.
- Degrees of Freedom: The number of splits captures the dimension of the domain as it is the number of components that fully determine the Partitioner variable's effect on the data being modeled.

3.1 Metrics

The appropriate metrics are often determined by the type of plot being used for visual analysis and capture the idea that the visual pattern can be simply described.

- Non-parametric
- Robust to the number of points
- Scale-invariant

We could explore ANOVA type analysis where we compare the means of more than two groups but this assumes that the groups being compared are statistically independent and are balanced in size. We can explain the distribution of observations by splitting the dataset into groups based on the Partitioner such that group is relatively homogenous (has low variance) and the mean of each group is distinct

Mix effects when aggregate numbers are affected by changes in the relative size and value of the subpopulations. Find the confounding covariate, the unexamined field that has an effect on the data pattern.

Considering correlations or slopes from linear regression fits would allow us to consider Simpson's paradox where an aggregate measure contradicts all the subpopulation measures.

Moving towards non-parametric metrics would allow for this Partitioner selection mechanism to be more generally applicable. Entropy does not consider the adjacency pattern in the grid of points. Therefore, a tighter Gaussian pattern is more interesting because it sits in fewer bins.

Graph-theoretic Scagnostics [37] are a non-parametric alternative that considers distributional shape.

3.2 Algorithm

We bootstrap

wide range of metrics cite – can be used w/ our metric

stats... visual null hypothesis – heike effect size paper...

different metrics piece– – about plot type

random split will have a pattern...misled small amounts of data. 1 example plot. small amount. random split looks like a a pattern.. well inside distribution successful detected not real pattern.

more points...less conservative.

by construction it'll work...proof by demonstration obviously generalizes

shingling case – test this out... continuous variables..

4 Examples

Data characteristics Can it pick out "confounding variates" such as those that cause Simpson's Paradox?

5 Extensions

Our approach is generalizable..

Decision tree approach of guided EDA. Small multiples

6 Conclusion

We explore...

Acknowledgments

The authors wish to thank A, B, C.

References

[1] G. Albuquerque, M. Eisemann, D. Lehmann, H. Theisel, and M. Magnor. Improving the visual analysis of high-dimensional datasets using quality measures. In *Visual Analytics Science and Technology (VAST), 2010 IEEE Symposium on*, pages 19–26, Oct 2010.

[2] R. Amar and J. Stasko. A knowledge task-based framework for design and evaluation of information visualizations. In *Proceedings of the IEEE Symposium on Information Visualization*, pages 143–150, 2004.

[3] M. Ankerst, S. Berchtold, and D. A. Keim. Similarity clustering of dimensions for an enhanced visualization of multidimensional data. In *Information Visualization, 1998. Proceedings. IEEE Symposium on*, pages 52–60. IEEE, 1998.

[4] R. A. Becker, W. S. Cleveland, and M.-J. Shyu. The visual design and control of trellis display. *Journal of computational and Graphical Statistics*, 5(2):123–155, 1996.

[5] J. Bertin. *Semiology of Graphics*. University of Wisconsin Press, 1983.

[6] E. Bertini and G. Santucci. Give chance a chance: modeling density to enhance scatter plot quality through random data sampling. *Information Visualization*, 5(2):95–110, 2006.

[7] E. Bertini, A. Tatu, and D. Keim. Quality metrics in high-dimensional data visualization: an overview and systematization. *Visualization and Computer Graphics, IEEE Transactions on*, 17(12):2203–2212, 2011.

[8] N. Boukhelifa, W. Cancino, A. Bezerianos, and E. Lutton. Evolutionary visual exploration: Evaluation with expert users. *Computer Graphics Forum*, 32(3pt1):31–40, 2013.

[9] A. Buja, D. Cook, H. Hofmann, M. Lawrence, E.-K. Lee, D. F. Swayne, and H. Wickham. Statistical inference for exploratory data analysis and model diagnostics. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 367(1906):4361–4383, 2009.

[10] Q. Cui, M. O. Ward, E. A. Rundensteiner, and J. Yang. Measuring data abstraction quality in multiresolution visualizations. *Visualization and Computer Graphics, IEEE Transactions on*, 12(5):709–716, 2006.

[11] T. N. Dang and L. Wilkinson. Scagexplorer: Exploring scatterplots by their scagnostics. In *Pacific Visualization Symposium (PacificVis), 2014 IEEE*, pages 73–80, March 2014.

[12] A. Dasgupta and R. Kosara. Pargnostics: Screen-space metrics for parallel coordinates. *Visualization and Computer Graphics, IEEE Transactions on*, 16(6):1017–1026, 2010.

[13] J. H. Friedman and J. W. Tukey. A Projection Pursuit algorithm for exploratory data analysis. *IEEE Trans. on computers*, 23(9):881–890, 1974.

[14] D. Guo, J. Chen, A. M. MacEachren, and K. Liao. A visualization system for space-time and multivariate patterns (vis-stamp). *Visualization and Computer Graphics, IEEE Transactions on*, 12(6):1461–1474, 2006.

[15] S. Johansson and J. Johansson. Interactive dimensionality reduction through user-defined combinations of quality metrics. *Visualization and Computer Graphics, IEEE Transactions on*, 15(6):993–1000, 2009.

[16] S. Kandel, R. Parikh, A. Paepcke, J. M. Hellerstein, and J. Heer. Profiler: Integrated statistical analysis and visualization for data quality assessment. In *Proceedings of the International Working Conference on Advanced Visual Interfaces*, pages 547–554. ACM, 2012.

[17] D. Lehmann, S. Hundt, and H. Theisel. A study on quality metrics vshuman perceptions: Can visual measures help us to filter visualizations of interest? *Information Technology*, 57(1):11–21, 2015.

[18] A. Lunzer, R. Belleman, P. Melis, and G. Stamatakos. Preparing, exploring and comparing cancer simulation results within a large parameter space. In *Information Visualisation (IV), 2010 14th International Conference*, pages 258–264. IEEE, 2010.

[19] A. MacEachren, D. Xiping, F. Hardisty, D. Guo, and G. Lengerich. Exploring high-d spaces with multiform matrices and small multiples. In *Information Visualization, 2003. INFOVIS 2003. IEEE Symposium on*, pages 31–38. IEEE, 2003.

[20] M. Majumder, H. Hofmann, and D. Cook. Validation of visual statistical inference, applied to linear models. *Journal of the American Statistical Association*, 108(503):942–956, 2013.

[21] T. Munzner. *Visualization Analysis and Design*. AK Peters Visualization Series. CRC Press, 2014.

[22] H. Piringer, W. Berger, and H. Hauser. Quantifying and comparing features in high-dimensional datasets. In *Information Visualisation, 2008. IV '08. 12th International Conference*, pages 240–245, July 2008.

[23] S. Sarni, A. Maciel, R. Boulic, and D. Thalmann. A spreadsheet framework for visual exploration of biomedical datasets. In *Computer-Based Medical Systems, 2005. Proceedings. 18th IEEE Symposium on*, pages 159–164. IEEE, 2005.

[24] M. Schäfer, L. Zhang, T. Schreck, A. Tatu, J. A. Lee, M. Verleysen, and D. A. Keim. Improving projection-based data analysis by feature space transformations. In *IS&T/SPIE Electronic Imaging*, pages 86540H–86540H. International Society for Optics and Photonics, 2013.

[25] J. Schneidewind, M. Sips, and D. A. Keim. Pixnostics: Towards measuring the value of visualization. In *Visual Analytics Science And Technology, 2006 IEEE Symposium On*, pages 199–206. IEEE, 2006.

[26] M. Sedlmair, T. Munzner, and M. Tory. Empirical guidance on scatterplot and dimension reduction technique choices. *IEEE Trans. Vis. Comput. Graph.*, 19(12):2634–2643, 2013.

[27] M. Sedlmair, A. Tatu, T. Munzner, and M. Tory. A taxonomy of visual cluster separation factors. *Comp. Graph. Forum*, 31(3pt4):1335–1344, June 2012.

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

[29] M. Sips, B. Neubert, J. P. Lewis, and P. Hanrahan. Selecting good views of high-dimensional data using class consistency. *Computer Graphics Forum*, 28(3):831–838, 2009.

[30] C. Stolte, D. Tang, and P. Hanrahan. Polaris: A system for query, analysis, and visualization of multidimensional relational databases. *IEEE Trans. Vis. Comput. Graph.*, 8(1):52–65, 2002.

[31] A. Tatu, G. Albuquerque, M. Eisemann, J. Schneidewind, H. Theisel, M. Magnor, and D. Keim. Combining automated analysis and visualization techniques for effective exploration of high-dimensional data. In *Visual Analytics Science and Technology, 2009. VAST 2009. IEEE Symposium on*, pages 59–66. IEEE, 2009.

[32] E. R. Tufte. *The Visual Display of Quantitative Information*. Graphics Press, Cheshire, CT, USA, 1986.

[33] J. W. Tukey. *Exploratory Data Analysis*. Addison-Wesley, 1977.

[34] H. Wickham. ggplot: An implementation of the grammar of graphics. *R package version 0.4. 0*, 2006.

[35] H. Wickham, D. Cook, H. Hofmann, and A. Buja. Graphical inference for infovis. *IEEE Trans. Vis. Comput. Graph.*, pages 973–979, 2010.

[36] L. Wilkinson. *The Grammar of Graphics*. Springer-Verlag New York, Inc., 2005.

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

[38] L. Wilkinson and G. Wills. Scagnostics distributions. *Journal of Computational and Graphical Statistics*, 17(2):473–491, 2008.

[39] G. Wills and L. Wilkinson. Autovis: Automatic visualization. *Information Visualization*, 9(1):47–69, 2010.

[40] J. Yang, M. O. Ward, E. A. Rundensteiner, and S. Huang. Visual hierarchical dimension reduction for exploration of high dimensional datasets. In *Proceedings of the Symposium on Data Visualisation 2003*, pages 19–28, 2003.