An Awesome Approach to Partitioning/Faceting Views

Anushka Anand and Justin Talbot

Abstract— Faceting/Partitioning a data view is a useful, easily interpretable multivariate display technique (trellis). Selecting a variable to partition a view such that it provides a visual explanation for the structure seen in the large single view is expensive/hard given high-dimensional datasets and limited prior research. We describe a set of goodness criteria for splits resulting from using a particular variable to partition the original view into a set of facets. Then we propose an algorithm for applying these criteria to select interesting partitions and evaluate the approach on a few real world datasets.

Index Terms—High-dimensional data, Faceting, Visual Analytics

1 Introduction

Exploratory Data Analysis (EDA), championed by Tukey [34], 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 [29, 40, 30]. Another popular approach is examining axis-unaligned projections through dimensionality reduction techniques [14, 43, 27]. 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 [39, 22] 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* [33] or *Collections*[6] and *Trellis* displays [5]. 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 [15, 20] and medicine [19, 24]. Furthermore, visual analysis tools such as the ggplot2 library in the R language [37] and Tableau [31] 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.

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 [38, 21]. 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 [10].

”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.

Multivariate Visual Explanations [4] explicitly reveal the hidden multivariate relationships in a simple manner to fill the WorldView gap [2] in visualization tools that fail to provide support for the discovery of useful correlative relationships in multivariate data. The MVE [4] approach tightly integrates partial derivatives computation and visual inspection to reveal multivariate correlations and as the structure of interest. We investigate a general approach to multivariate visual explanations that can be used to discover various structures of interest specified by quantitative data visualization quality metrics.

2.2 Data Visualization Quality Metrics

A large number of quality metrics are used in methods for high-dimensional data analysis [8]. Many metrics for scatterplots determine projections of the data to be displayed, often for particular tasks: cluster separation [28, 32], class consistency and separation [30, 25], interesting visual shapes [40] or statistical properties [17, 29]. Metrics for parallel coordinate plots [3, 13, 16, 43] also focus on the ordering of dimensions. Some metrics [7, 11] focus on the level of abstraction, including aggregration and sampling, in these chart types. Others [1, 3, 26, 43] 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

- *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.

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 [29] (rank all 2D projections). a follow up [23] 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 [9] of linear and nonlinear 2D projections.

2.3 Scagnostics

Graph-theoretic scagnostics [40], Scagnostic distributions [41], ScagExplorer [12]

Experimental evaluation [18] of scagnostics and other metrics, showing that scagnostics best matches human rankings.

AutoVis [42]

3 Method

Statistical methods like ridge and lasso regression help automatically select subsets of variables that produce good explanatory models of a set of multivariate observations. However, these methods make a number of assumptions about the errors in the model given the sample and about the independence of predictor variables. Using lasso regression to progressively add explanatory variables assumes interest in a linear model and may not translate to visually interesting different patterns in consecutive steps.

Given a user-selected data relationship that captures the set of dependent variables of interest, we seek to add explanatory variables that help explain the patterns seen in the visual representations. For simplicity, we describe a scenario where the user selects a bivariate relationship of interest and visualizes it as a scatterplot. Adding explanatory variables partitions the scatterplot into multiple plots (small multiples) such that there as many plots as there are discrete categories of a categorical Partitioner or discrete bins of a quantitative Partitioner variable. Therefore, a Partitioner splits or facets a particular visual representation into similar views of subsets of the data. We propose a number of criteria to evaluate the goodness of the splits resulting from adding a Partitioner variable to explain the structure seen in the original visual representation.

3.1 Split Criteria

We set up the following goodness-of-split criteria to guide our work:

- Visually Interesting Pattern: The splits of 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

- Diversity: Robustness to overfitting? The pattern in the splits would be different from the original... and from each other.

- 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. With a small set of points, 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.

- 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's effect on the data being modeled.

Support is dependent on the degrees of freedom criterion as the number of points per split would decrease as the number of splits increase given a constant number of observations to start with.

3.2 Algorithm

We describe the design of an algorithm that constructs good splits given the criteria outlined above. We assume we start with the user specified visual representation of a set of data and seek to split the display into facets that reveal useful visual structure.

Filtering the large number of views of a high-dimensional dataset motivated Tukey's proposal of *cognostics* [35, 36] - diagnostic metrics to evaluate the usefulness of views - so users would only manually investigate a small set of high-ranked, potentially useful views. As described in Section 2, there are numerous metrics to evaluate various view types (from scatterplots to radial views) based on various tasks (from finding outliers to separating classes or groups).

We can explain the distribution of observations by splitting the dataset into groups based on the Partitioner such that each group is relatively homogenous (has low variance) and the mean of each group is distinct. Then we could apply ANOVA to we compare the means of these groups and look for significant differences as indicators of interesting splits. This type of analysis assumes that the groups being compared are statistically independent and are balanced in size which will not be the case for arbitrary categorical fields in datasets.

Another visual pattern metric could be to use the correlation or slopes from linear regression fits to help distinguish splits where subsets of the data determined by the Partitioner variable have particular linear relationships. This could help in the discovery of confounding covariates, the unexamined fields that have an effect on the data pattern. Simpson's paradox is a classic example of such mix-effects when aggregate numbers are affected by changes in the relative size and value of the subpopulations.

Going towards non-parametric metrics, visual pattern salience is often captured by entropy on the binned visual representation. However, entropy does not consider the adjacency pattern in the grid of points so a sine wave pattern might be just as interesting a small Gaussian pattern if they sit in the same number of bins. Here, we would like to be able to differentiate visual distinct patterns.

The distributional shape of visual patterns in a scatterplot are quantitatively captured through graph-theoretic scagnostics [40]. These metrics have the benefit of being non-parametric and robust to the number of points as they first bin the data. However, when considering robustness, these metrics are not altogether scale-invariant nor do they capture location specifics if we are interested in particular positional patterns. Selecting non-parametric metrics for the visual pattern would allow for a more generally applicable algorithm.

After selecting a cognostic metric that quantitatively captures the visual pattern of a set of data observations, we use the metric to determine a set of scores for the set of splits that result from applying the facet operator with a particular Partitioner variable. The diversity requirement is captured by comparing the set of scores to a similar set of scores for random splits. Let the Partitioner variable d_p create k splits with sizes $\{s_1, s_2, ..., s_k\}$. We generate r random sets of k splits bootstrapped without replacement from the original data and of the same sizes $\{s_1, s_2, ..., s_k\}$. From these bootstrapped random splits we compute the cognostic on each split and get k distributions of cognostics. We then compare the cognostic metrics from the Partitioner split to these random splits.

The support criterion is triggered in how we determine if the difference in cognostic metrics between the Partitioner splits and random splits is meaningful. The bootstrapped metric distributions function as reference ”null distributions” and we apply Chebyshev's inequality to determine how significant the difference of the Partitioner's splits are from random. To account for the observation that the distributions for various cognostics were not Gaussian, we use a non-parametric significance test. Now, for splits of small size s_i, there will be a wide spread of values for most cognostics as it is easy to find a pattern with a few points. However, these patterns are not “real” or significant patterns because of low support and we expect the cognostics for the splits from Partitioner variable to fall within the reference distribution from random splits. This implies that we will be conservative in accepting visual patterns from splits with low support as truly revealing interesting structure. Conversely, cognostics with high support are likely to be meaningful and a significant difference from the reference distribution would mark a split with interesting visual patterns.

We want to penalize a large number of splits as this negatively affects the support and it provides a heavier investment from the user in terms of visual comparisons. The z-score from applying Chebyshev's inequality penalizes the increase in number of splits k (as k goes to ∞, the z-score goes to 0)

$$\sum_{i=1}^{n} \frac{(X_i - \mu_i)^2}{\sigma_i^2}$$

3.3 Handling Continuous Partitioners

Determining discrete splits for a categorical variable is trivial as the observations are naturally partitioned into subsets for each discrete choice the variable offers. For continuous variables, discrete partitions can be created through a binning technique. There are various binning techniques [**?**, **?**] employed in histograms. An alternative binning strategy is one with overlapping bins of roughly equal count called shingles [**?**].

3.4 Multiple Partitioners

Extending the algorithm to pick multiple partitioners – useful for small multiples/trellis

4 Examples

Data characteristics

Visually interesting patterns – scagnostics (works for different ones) Different – Can it pick out "confounding variates" such as those that cause Simpson's Paradox? support – small n (ourworld.csv) vs large n. small -¿ inside distributions so we'll be more conservation in judgements of "real" patterns number splits – State vs Region

5 Discussion

Our approach is generalizable.. The wide range of metrics [8] can be used w/ our method to describe the visual pattern of interest to find.

UI Ideas like Decision tree approach of guided EDA. Trellis displays - choice of layout

6 Conclusion

We explore...

Acknowledgments

The authors wish to thank A, B, C.

References

[1] G. Albuquerque, M. Eisemann, D. J. 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. IEEE, 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] S. Barlowe, T. Zhang, Y. Liu, J. Yang, and D. Jacobs. Multivariate visual explanation for high dimensional datasets. In *Visual Analytics Science and Technology, 2008. VAST '08. IEEE Symposium on*, pages 147–154, Oct 2008.

[5] 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.

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

[7] 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.

[8] 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.

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

[10] 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.

[11] 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.

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

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

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

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

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

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

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

[30] 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.

[31] 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.

[32] 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.

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

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

[35] J. W. Tukey and P. A. Tukey. Some graphics for studying four-dimensional data. In *Computer Science and Statistics: Proceedings of the 14th Symposium on the Interface*, 1982.

[36] J. W. Tukey and P. A. Tukey. Computer graphics and explaoratory data analysis: An introduction. In *Proceedings of the Sixth Annual Conference and Exposition: Computer Graphics*, 1985.

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

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

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

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

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

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

[43] 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.