CLIPPR: Maximally Informative CLIPped PRojections with Bounding Regions

Category: Research

Paper Type: Algorithm

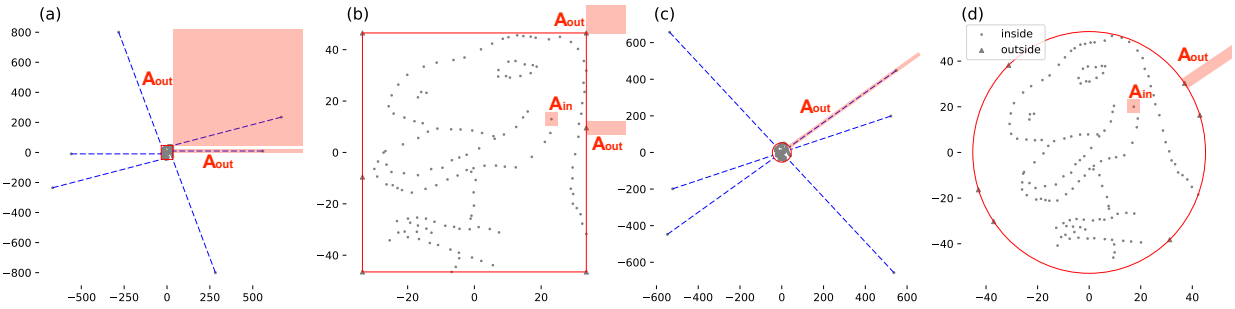

Fig. 1. CLIPPR is an algorithm for generating maximally informative clipped projections with bounding regions. The projection shown in (a) features a dense core of occluded points as well as a set of outliers. By finding an approprate bounding box, a *clipped projection* is obtained in (b). In the clipped projections, points outside the bounding region are shown with triangular markers on the boundary. CLIPPR can also produce elliptical bounding regions. The data consists of six artificial outliers plus Alberto Cairo's figure: `http://www.thefunctionalart.com/2016/08/download-datasaurus-never-trust-summary.html`.

Abstract—Projecting data down to two dimensions to visualize in a scatterplot is one of the basic building blocks of visualization. While there are various established methods used for projection, many projections fail to capture phenomena at different scales, due to occlusion or overplotting. A trade-off emerges between showing small and large-scale structure. In this work, we present an algorithm that parameterizes this tradeoff to calculate multiple projections that vary by the scale of the highlighted structure. By jointly optimizing both the *information theoretic content* of the projection and the clipped bounding region of the resulting view, we can empirically find relevant structure to show to a user. We demonstrate the use of this algorithm on several synthetic and real datasets. We also describe how this method would be useful in a visual analytics system for providing a *grand tour* for both low- and high-dimensional datasets. By exposing a simple resolution parameter to the user, the user is able to guide their own path through their data, enabling them to glean multiple levels of insight in a way that other static projection techniques could not allow.

1 Introduction

Imagine a scenario where an analyst is examining a new dataset of real numbers. The analyst loads the dataset in her favorite command-line statistical analysis environment and plots the dataset as a 2-dimensional scatterplot. However, instead of seeing interesting patterns within the data, the analyst sees a small blob of overplotted points and a few outliers such as the image shown in Figure 1(a).

More formally, we can state that the *information content* of Figure 1(a) is lower than that of Figure 1(b). The need to show all data points, including the outliers, results in overplotting of many of the data points. Overplotting is commonly considered to be a function of display resolution. However, due to the limits of human perception, even if no two data points in Figure 1(a) appear on the same pixel (possibly because of an ultra-high resolution display), the result still appears to be cluttered. The user would not be able to discern the underlying pattern as easily in Figure 1(a) as if the data is presented in a "zoomed in" way as in Figure 1(b).

In this poster submission, we propose a technique to automatically identify *clipped projections* of a high-dimensional data in a 2-dimensional subspace. Our technique is based on the maximization of the information content within the clipped projections, but also allows for control to account for the resolution of the displayed plot or human perception (whichever is worse). Our technique, CLIPPR, exposes a *resolution parameter* that can be set by the user or by the visualization designer to control the scale of visual phenomenon found in the dataset. We demonstrate the use of this technique on several synthetic and real datasets.

2 Related Work

Static Projection Techniques: We compare CLIPPR to other popular projection techniques. Most projection techniques optimize a particular metric that is defined over every possible projection. Principal Component Analysis [3] optimizes variance of each dimension in order. If we ignore clipping (or if the clipped projection includes all points inside the bounding box), our technique can be made equivalent to PCA [4]. Multidimensional Scaling [5] optimizes the ratio of differences between data space and projected space. More similar to CLIPPR are Stochastic Neighbor Embedding (SNE) techniques [2, 6], in which the location of a projected point is expressed as a random variable, and the dissimilarity (KL-divergence) of the joint distributions (over the random variables) between data space and projected space is minimized. Note, however, that CLIPPR results in *linear* projections, while SNE projections are non-linear and thus distort distances and shapes in data.

3 Method

Any projection of data requires a metric that determines how to rank potential projections [1]. For example, PCA maximizes the variance found in the first two dimensions with the assumption that, with no other prior knowledge of the data, maximum variance in the data means showing the most information to the user that is possible to see in two dimensions. However, in many cases, the variance in the dataset may not be particularly interesting. Consider outliers as in Figure 1. They contribute largely to the variance of the dataset, but they do not add to the user's understanding of the relevant structure in the data (i.e., the dinosaur in the center).

Instead, we propose a metric that rewards projections that are unexpected (and thus interesting) under a user's prior belief about the layout of the data. We can optimize the Information Content (IC) of a clipped projection, where IC is defined as a function of the *visible* points inside a bounding box, with respect to a background distribution.

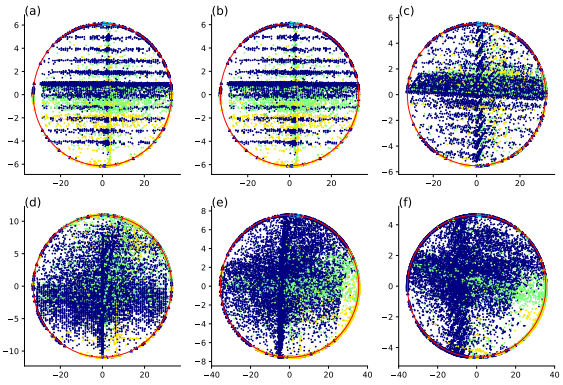

Fig. 2. Grand tour on UCI shuttle dataset. (a) The 1st frame of the grand tour, $f = 0.01$. (b) The 5th frame of the grand tour, $f = 0.05$. (c) The 6th frame, $f = 0.06$. (d) The 7th frame, $f = 0.07$. (e) The 20th frame, $f = 0.2$. (f) The 100th frame, $f = 1.0$. Points are colored by their class label. By varying the f parameter, different relationships between classes emerge.

Let s be the bounding box parameters and f be a resolution parameter controlling how far away two points need to be in order to be *discernable*. We will denote the projections of a data set $\hat{\mathbf{X}}$ onto the column vectors of $\mathbf{W}$ as $\hat{\mathbf{\Pi}}_{\mathbf{W}} \in \mathbb{R}^{n \times k}$, and $\mathbf{A}(\hat{\mathbf{\Pi}}_{\mathbf{W}}, f, \mathbf{s})$ be the union of discernible regions of points in the projection. We can define the probability that a weight matrix $\mathbf{W}$ generated a clipped projection.

$$\begin{aligned}
&\mathbf{Pr}\left(\hat{\mathbf{\Pi}}_{\mathbf{W}} \in \mathbf{A}(\hat{\mathbf{\Pi}}_{\mathbf{W}}, f, \mathbf{s})\right) \\
&= \int_{\mathbf{A}(\hat{\mathbf{\Pi}}, f, \mathbf{s})} p_{\mathbf{\Pi}_{\mathbf{W}}}(\mathbf{\Pi}_{\mathbf{W}})\, d\mathbf{\Pi}_{\mathbf{W}} \\
&= \prod_{i=1,2,\ldots,n} \left[\int_{\mathbf{z}_i \in \mathbf{A}(\hat{\mathbf{z}}_i, f, \mathbf{s})} p_{\mathbf{\Pi}_{\mathbf{W}}}(\mathbf{z}_i) d\mathbf{z}_i \right]. \qquad (1)
\end{aligned}$$

The goal of finding the most informative clipped projection can be formalized as an optimization problem:

$$\begin{aligned}
&\mathbf{IC}(\mathbf{W}, \hat{\mathbf{\Pi}}_{\mathbf{W}}, \mathbf{s}) = -\log \mathbf{Pr}\left(\hat{\mathbf{\Pi}}_{\mathbf{W}} \in \mathbf{A}(\hat{\mathbf{\Pi}}_{\mathbf{W}}, f, \mathbf{s})\right). \\
&\underset{\mathbf{W},\mathbf{s}}{\mathbf{argmax}}\ \mathbf{IC}(\mathbf{W}, \hat{\mathbf{\Pi}}_{\mathbf{W}}, \mathbf{s}), \\
&\mathbf{s.t.}\ \ \mathbf{W}'\mathbf{W} = \mathbf{I},\ \ \mathbf{s} > \mathbf{0}.
\end{aligned}$$

4 Experiments

The experiments are conducted using a Python implementation of CLIPPR, which, along with the datasets used, have been made publicly available for the purpose of reproducibility.[1]

For the sake of brevity, we provide information on two experiments. First, we demonstrate two types of bounding regions supported by CLIPPR, rectangles and ellipses. These are demonstrated on a synthetic dataset, found in Fig. 1. In both types of bounding regions, outliers are shifted to the boundary lines, so that the primary structure, the dinosaur figure, is easily visible in the projection. Outliers are signified by arrow glyphs that communicate their direction from the center of the projection.

We also experimented with varying the resolution parameter to uncover more than one scale of phenomenon in multiple projections. To explore this, we ran CLIPPR on a real dataset at six different settings of the resolution parameter, visible in Fig. 3. The UCI Shuttle dataset[2] consists of 14,500 testing examples from a shuttle flight. Initially, we see that data items of the blue class seem to differ by regular intervals, suggesting that there exists a discrete binning of one variable in the dataset that varies across the blue class. As the f parameter is swept higher, the blue points cluster into a single horizontal and single vertical component. The two scales reveal two different structures in the data.

5 Discussion

Data often contains structures in different scales. While various projection methods are used to visualize the data, many of them fail to capture phenomena at different scales. A trade-off emerges between showing small and large-scale structure. Methods like PCA and MDS that optimize for showing the large-scale structure can over-represent outliers, whereas methods like tSNE emphasize the small-scale structure but can present spurious global structures. In this paper, we present CLIPPR, an algorithm that aims to optimize the information content of the visualization, searching for the most informative structure of a given dataset. Results from several case studies show how CLIPPR allows user to drive their own tour of exploring the structures with different scales in the data. The experiments also suggests that CLIPPR can be used by visualization designers to carefully craft their own projection.

CLIPPR is not yet efficient enough to provide real-time data exploration. Currently, CLIPPR needs to pre-render a series of projections to provide a grand tour. Improving the scalability of CLIPPR is one of the important directions for further investigation. Also, optimizing the information content of clipped projections with bounding box that has higher degrees of freedom (e.g., shifting center) is also worth examining.

In our experiments, the resolution parameter seemed to be related to the perceptual ability of a user to differentiate points. In this work, the value of the resolution parameter is set heuristically by searching through the resulting projections as the parameter changes. In recent years, research at the intersection of perceptual psychology, psychophysics, and visualization has informed the choice of visualization parameter. A similar approach could inform an automated choice for the resolution parameter.

References

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

[2] G. E. Hinton and S. T. Roweis. Stochastic neighbor embedding. In *Advances in neural information processing systems*, pp. 857–864, 2003.

[3] I. T. Jolliffe. Principal component analysis and factor analysis. In *Principal component analysis*, pp. 115–128. Springer, 1986.

[4] B. Kang, J. Lijffijt, R. Santos-Rodríguez, and T. De Bie. SICA: subjectively interesting component analysis. *Data Mining and Knowledge Discovery*, online ahead of print. doi: 10.1007/s10618-018-0558-x

[5] J. B. Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1):1–27, 1964.

[6] L. v. d. Maaten and G. Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[1] Implementation of CLIPPR and the datasets can be found online at:`https://clipped-projections-test.herokuapp.com/static/VAST18_submission_1237_CLIPPR_code.zip`

[2] `https://archive.ics.uci.edu/ml/datasets/Statlog+(Shuttle)`, 'shuttle.tst'