WIREs COMPUTATIONAL STATISTICS

WILEY

The State-of-the-Art on Tours for Dynamic Visualization of High-dimensional Data

Journal:	*WIREs Computational Statistics*
Manuscript ID	EOCS-534
Wiley - Manuscript type:	Advanced Review
Date Submitted by the Author:	20-Apr-2021
Complete List of Authors:	Cook, Dianne; Monash University Lee, Stuart; Monash University Da Silva, Natalia; Universidad de la Republica Uruguay Laa, Ursula; University of Natural Resources and Life Sciences Vienna Spyrison, Nicholas; Monash University Wang, Earo; The University of Auckland Zhang, H. Sherry; Monash University
Keywords:	tours, data visualization, high-dimensional data, data science, exploratory data analysis
Choose 1-3 topics to categorize your article:	Analysis of High Dimensional Data (HBAD) < Statistical and Graphical Methods of Data Analysis (HBAA), Statistical Graphics and Visualization (HBAB) < Statistical and Graphical Methods of Data Analysis (HBAA), Exploratory Data Analysis (HAAC) < Statistical Learning and Exploratory Methods of the Data Sciences (HAAA)

Advanced Review: The State-of-the-Art on Tours for Dynamic Visualization of High-dimensional Data

Stuart Lee

Department of Econometrics and Business Statistics, Monash University
stuart.a.lee@monash.edu
ORCiD: 0000-0003-1179-8436

Dianne Cook*

Department of Econometrics and Business Statistics, Monash University
dicook@monash.edu
ORCiD: 000-0002-3813-7155

Natalia da Silva

Instituto de Estadística (IESTA), Universidad de la República
natalia@iesta.edu.uy
ORCiD: 0000-0002-6031-7451

Ursula Laa

Institute of Statistics, University of Natural Resources and Life Sciences
ursula.laa@boku.ac.at
ORCiD: 0000-0002-0249-6439

Nicholas Spyrison

Faculty of Information and Technology, Monash University
nicholas.spyrison@monash.edu

Earo Wang

Department of Statistics, The University of Auckland
earo.wang@auckland.ac.nz
ORCiD: 0000-0001-6448-5260

H. Sherry Zhang

Department of Econometrics and Business Statistics, Monash University
huize.zhang@monash.edu
ORCiD: 0000-0002-7122-1463

Abstract

This article discusses a high-dimensional visualization technique called the tour, which can be used to view data in more than three dimensions. We review the theory and history behind the technique, as well as modern software developments and applications of the tour that are being found across the sciences and machine learning.

Keywords: tours, data visualization, high-dimensional data, data science, exploratory data analysis

1 Introduction

Data commonly arrives with more than two measured variables, which makes it more complicated to plot on a page. With multiple variables, especially if there is some association between variables, this would be called multivariate or high-dimensional data. When the variables are all numeric, or quantitative, visualization often relies on some form of dimension reduction. This can be done by taking linear projections, for example, principal component analysis (Hotelling, 1933) or linear discriminant analysis (Fisher, 1936). It is also common to reduce dimension with nonlinear techniques like multidimensional scaling (MDS) (Kruskal, 1964) or t-Distributed Stochastic Neighbor Embedding (t-SNE) (van der Maaten & Hinton, 2008).

The term "high-dimensional" here means Euclidean space. Figure 1 shows a way to imagine this. It shows a sequence of cube wireframes, ranging from 1D through to 5D, where beyond 2D is a linear projection of the cube. As dimension increase, a new orthogonal axis is added. For cubes, this is achieved by doubling the cube: a 2D is two 1D cubes, a 3D is two 2D cubes, and so forth.

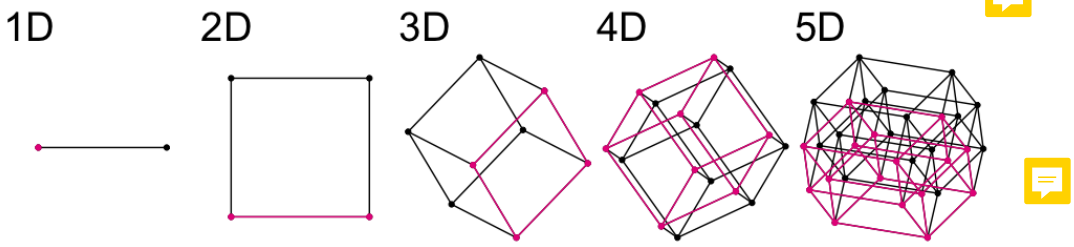

Figure 1: Illustrating what is meant by "high-dimensional" in this paper, and a "linear projection". From a sequence of increasing dimension cubes, from 1D to 5D, as wireframes, it can be seen that the dimension increase by 1, the cube doubles.

The focus of this review will be on visualizing high-dimensional numerical data using linear projections, in particular, as provided by the grand tour (Asimov, 1985; Buja et al., 2005). The

reason being that it is not feasible to adequately review the very large area of visualizing high-dimensions, and there have been numerous developments in tours recently. An overview of the technique, and new modifications is provided, along with how these techniques can be used in a variety of applications.

A tour can be considered to be a dynamic graphic, because it shows a smooth sequence of projections over time, ideally with controls that allow stopping, reversing, changing direction, or going forward again. It can be useful to embed a tour into an interactive graphics system, where plots can be queried and elements highlighted (see for e.g. Swayne et al. (2003) or Tierney (1991)). To create the smooth sequence, a geodesic interpolation is computed between consecutive frames. It allows the viewer to extrapolate from the low-dimensional to shapes corresponding to multivariate distribution, and is particularly useful for detecting clusters, outliers and non-linear dependence.

While tours are invaluable for assessing the geometry of data, they are by no means the only technique available for visualizing structure in high dimensional data. An early technique proposed for assessing pairwise relationships between variables is the scatterplot matrix (SPLOM) (Becker & Cleveland, 1987; Carr & Nicholson, 1984; Chambers et al., 1983; J. W. Tukey & Tukey, 1983; P. A. Tukey & Tukey, 1981). The SPLOM allows the viewer to assess correlation structure but does not scale to large numbers of variables. Similarly, parallel coordinates plots (PCP) can be used to explore correlation and collinearity (Inselberg, 1985; Wegman, 1990). By placing multiple variables side by side in a PCP higher order structure like clustering or lower dimensional embeddings are revealed, however the ordering of variables along the axis changes what can be learned. Another display that relies on variable ordering is the heatmap, which is widely used to visualize cluster structure in bioinformatics. Wilkinson and Friendly (2009) provides a comprehensive history of this display in the social and natural sciences.

All of the aforementioned techniques can be enhanced through the use of interactivity. By combining views with interactive elements like tool tips or highlighting the analyst is able to quickly interrogate interesting features of the data. One particularly important interaction technique in the history of statistical graphics is called brushing (Becker & Cleveland, 1987). When brushing an analyst drags their mouse over the view which results in a region being drawn onto the canvas. When there are multiple views present the act of brushing can be thought of as a database query; points that fall inside the brush can be used to highlight or filter data on adjacent views (Figure 2). This technique is particularly useful when combined with the tour (Section 3.3).

Nonlinear dimension reduction techniques such as t-SNE and Uniform Manifold Alignment and Projection (UMAP) have become very popular in recent years, primarily for the ability to capture cluster structure in a succinct visual summary (McInnes et al., 2020; van der Maaten & Hinton, 2008). However, it is only a summary, and it likely involves substantial warping of the original data space. Using the tour along with these techniques can illuminate the nature of the warped space, and reveal other structure lost in the dimension reduction. Figure 3 illustrates the difference in what can be learned from a tour in comparison with nonlinear dimension reduction using t-SNE. The 10-dimensional data comes from Rauber (2009). The t-SNE view (plot A) shows six clearly separated clusters, all spherical with different sizes. The tour shows that the clusters do not look like this in the full data space. The clusters are at various distances apart and very different sizes, as can be seen from the four projections from a tour. The two green clusters are large and almost spherical, and far from the orange clusters. The orange clusters have one larger one, and three smaller, very close to each other. All of these are elliptical, which

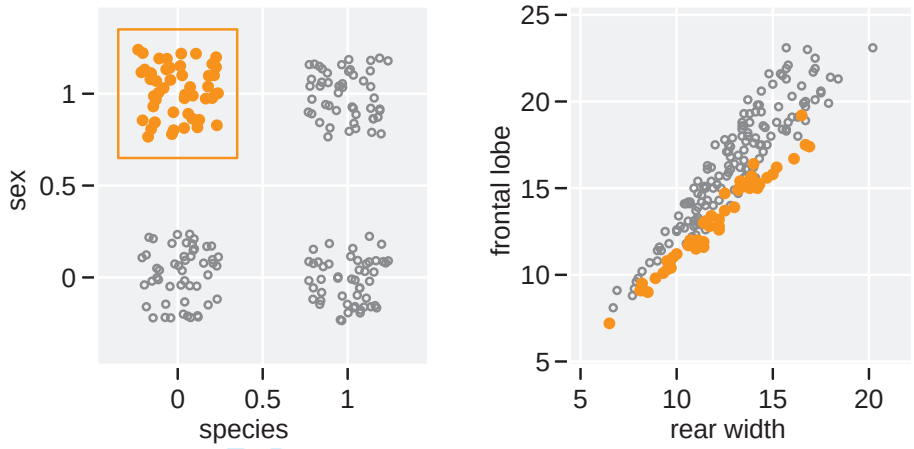

Figure 2: When a user brushes data on the left panel, points that fall inside the region are highlighted in orange. The view on the right responds by highlighting the corresponding points. Adapted from Cook et al., 2007 Figure 2.12.

means that they have very little variability, actually no variability if you watch the full tour, in some of the 10 dimensions. This gives some deeper perspective to what is learned from t-SNE, and illustrates what the t-SNE dimension reduction has done: it has found small gaps between points and expanded these gaps to yield the representation. It should be noted, though, that the methods (t-SNE and tour) complement each other. The t-SNE view gives a clear indication of six clusters, which may have been overlooked on initial viewing with a tour. With this information is the invitation to look closer at the data in the tour, to see, that yes, indeed, there are three, tiny, tiny clusters very close to each other.

The rest of the review is structured as follows: Section 2 defines the notation and components of a tour displays. By its nature a tour is most effective to analyst when combined with interactivity; Section 3 reviews the components of user interfaces for manipulating tour views, including manual tours (Section 3.2) which are useful to test structure of selected features. The implementation of tour paths in statistical software is reviewed in Section 4. Section 5 shows the diverse applications of the tour in the natural sciences, machine learning and applied statistics. Finally, Section 6 discusses future research directions for tours.

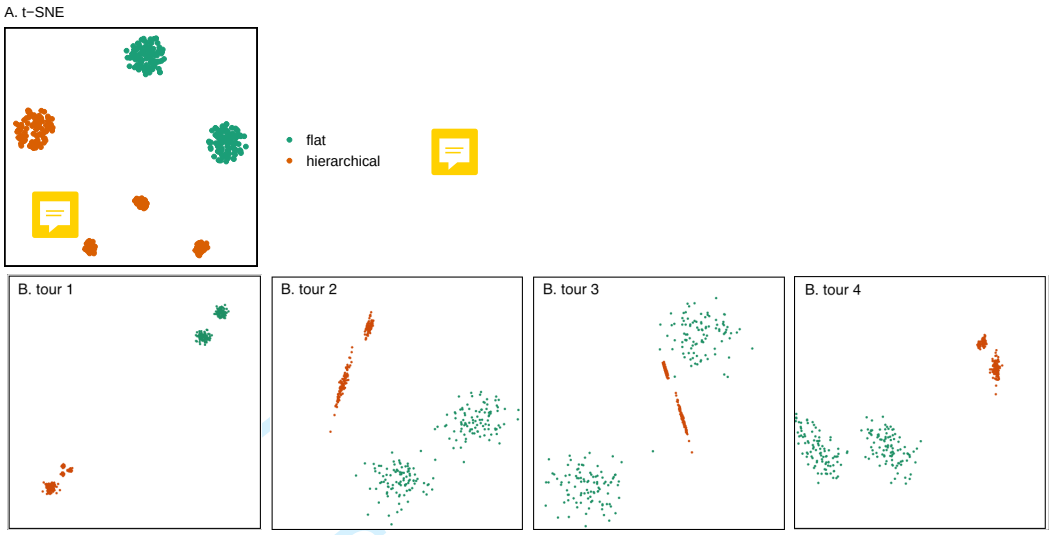

Figure 3: Comparison of structure perception between nonlinear dimension reduction from 10-d, using t-SNE (A) and a tour (B). There are six clusters, as seen in the t-SNE view, but the relative distance between the clusters is extremely varied. This can be seen in a sample of tour projections shown. The two green clusters are (almost) spherical in shape, and very distance from the orange clusters. Three of the orange clusters are very close to each other (just visible in B.1), and all orange clusters are elliptical. The tour provides a more accurate rendering of the clusters in the high-dimensional space, and complements what is learned from the dimension reduction.

2 Tours for high dimensional visualization

2.1 Notation

When using a tour, a sequence of d-dimensional linear projections are obtained from a p-dimensional space, where $d \ll p$. Let $X_{n\times p}$ be the data matrix, consisting of n observations and p variables, whose projections are of interest. A projection basis, $A_{p\times d}$, is a matrix that characterizes the direction from which the data are projected and needs to satisfy an orthonormality condition, which requires that each column in A has unit length and are perpendicular. With a data matrix and a projection basis, a projection of the data can be defined as $Y = X \cdot A$. Figure 4 shows two examples of low-dimensional projections of the palmer penguins data (Horst et al., 2020) in a Huber plot (Huber, 1990) and a histogram.

2.2 Finding targets

A tour path, that is a sequence of projections, can be generated by geodesic interpolation between a set of target planes (Figure 5). Different methods for choosing target planes provide different tour paths. The grand tour is generated using randomly selected targets. A guided tour is generated by choosing particularly structured projections from a projection pursuit optimization. The little tour uses all variable bases as targets, and a local tour rocks back and forth from a particular plane to randomly chosen targets in a small neighborhood.

Figure 6 shows a representation of 1D tour paths of 5D data, drawn on a PCA space. A grand

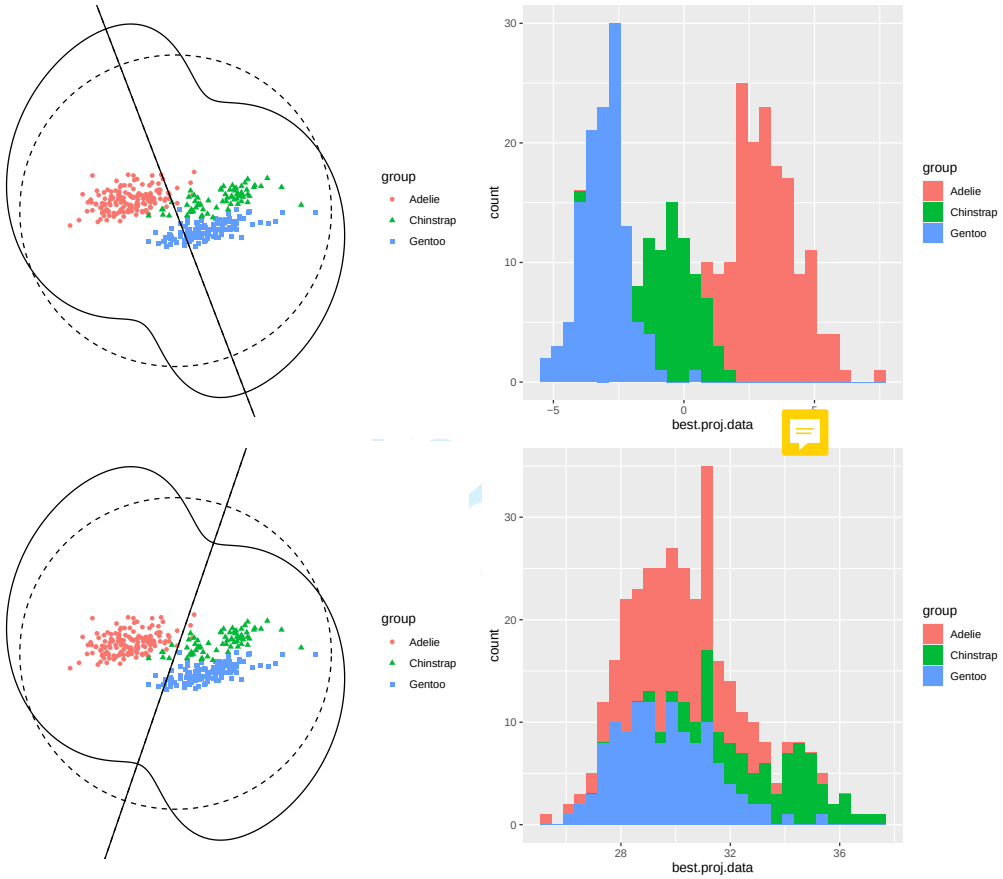

Figure 4: Huber's plot (left plots) of the penguins data (bill length and depth only). The solid line represents how well the projection onto different directions, separates the three species, and the dashed circle is a reference guide. The histogram (right plots) is the 1D projection of the data onto the direction outlined as the solid line in the Huber's plot. Adapted from E.-K. Lee et al., 2005, Figure 1.

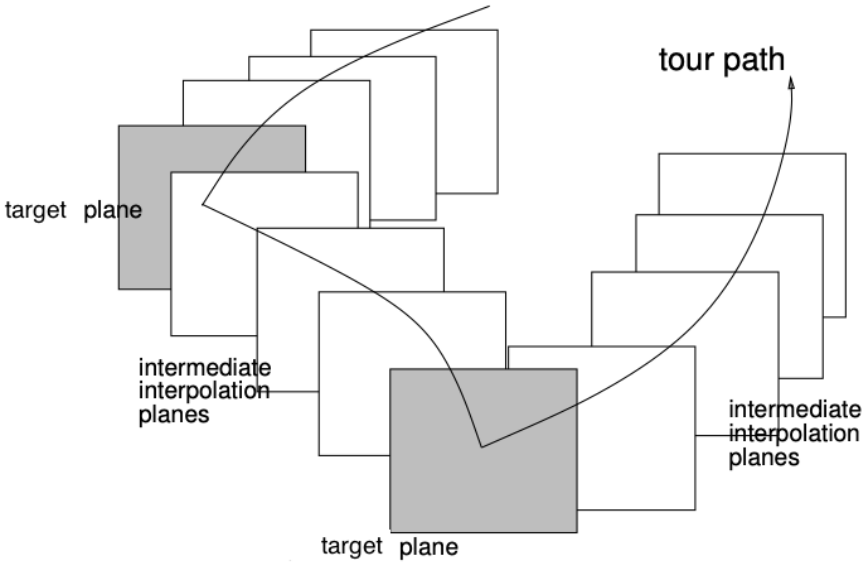

Figure 5: Illustration of a tour path, indicating target planes, and interpolation planes. Adapted from Buja et al., 2005, Figure 1.

tour path is on the left and a guided tour path is on the right. The grand tour path should be more wide-spread because it is attempting to show all possible projections, if left to run long enough. The guided tour path should be short, as the projection pursuit optimization zeroes on the optimally interesting projection.

There are other tour types which don't fit this style of tour. The manual tour (Cook & Buja, 1997) allows the user to change the projection coefficients manually, to rotate a variable into and out of a project, and is discussed later. The recently developed slice tour (Laa, Cook, & Valencia, 2020) can be applied to any of the above tour types. It displays a slice through the orthogonal space, as opposed to a projection, and is explained in more details later.

2.3 Geodesic interpolation

The smooth progression in the tour path is due to geodesic interpolation between the target planes. This takes into consideration two important aspects: (1) it maintains the orthonormality of the projection bases, (2) contains all the rotation to be between planes, not a particular basis in any plane. The first is clearly important because it ensures that we are looking at low-dimensional projections of the data always. The latter is harder to explain, but really important from a visual perspective. It stops any within-plane spin, and could be considered stabilizing the view. More details explanation of this can be found in Buja et al. (2005).

Figure 7 shows tour paths of 2D projections of 6D data. The space of all tour paths is a high-dimensional torus, as represented by the gray points. The tour paths are shown in green and orange, and each dot indicates a projection in the sequence. The plots in this figure could be considered to be a tour looking at itself, because each plot is a selected view from a tour of the torus with the paths overlaid.

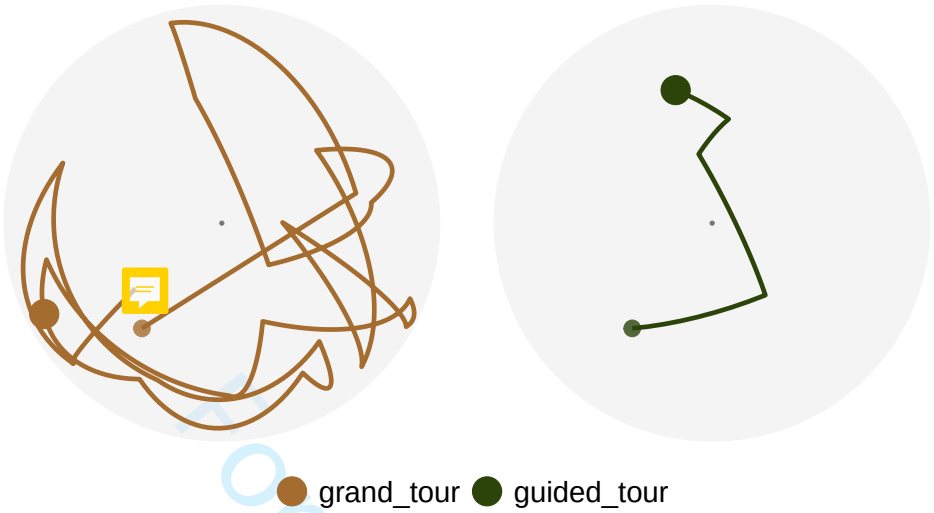

Figure 6: A comparison of guided and grand tour paths in the PCA reduced 2D space. The start point of tour paths is smaller and lighter than the end point. Guided tour optimises the index value iteratively and finishes the search quickly while grand tour wanders in the parameter space to look at possible interesting projections.

2.4 Display

Various displays are available to show projections in 1D, 2D and higher dimensions. A 1D projection displays the data in a histogram analogous to a shadow puppet being projected onto a wall, while 2D projections are displayed as scatterplots. Higher dimensional projections can be shown with multivariate displays like SPLOMs or parallel coordinates. Figure 8 shows a variety of displays for projections and these displays are useful to show the non-normal distributions in the projected data, clustering structure, and multivariate relationship between variables.

2.5 Slices, sections and projections

Interactive systems like GGobi (Swayne et al., 2003) provide the option of adding sectioning information via linked brushing. Sectioning means we select points that fall in a section of the full parameter space, for example for the purpose of highlighting them in a tour display. This often reveals complementary information, and combining sections and projections can for example provide insights into the dimensionality of a data structure (Furnas & Buja, 1994).

Sectioning a high-dimensional space in systematic manner (without interactive selection of data sections) is challenging, since there is a lot of freedom in choosing a section. One approach is to define sections based on projection planes, we refer to such sections as "slices" of the data. This is implemented in the slice tour display (Laa, Cook, & Valencia, 2020), and uses the orthogonal distance of each data point from the current projection plane (typically placed such that it passes through the data mean) to highlight points that are nearby and fade out points further from the plane.

The slice tour can reveal concave or non-linear structures obscured in projections, as well as small structures hidden near the center of a distribution. Figure 9 shows snapshots of the slice tour of points distributed on the surface of geometric shapes. Along with projections, an index

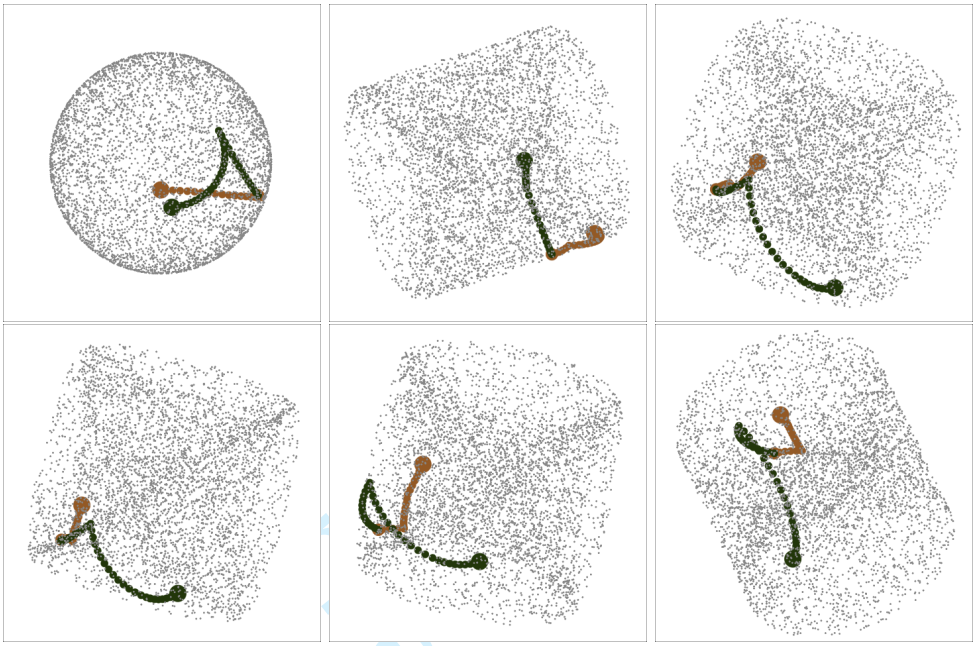

Figure 7: Still views of a tour showing the space of possible two-dimensional projections as a torus shape in grey, and the trace of two short paths obtained via geodesic interpolation traced in color.

function can be used to select target planes to show interesting slices of the data, defined as a section pursuit guided tour (Laa, Cook, Buja, et al., 2020).

2.6 Transformations on projections

Transformations of data are often useful prior to touring such as scaling, sphering, or a logarithmic transformation for skewed distributions. However, we may also want to transform the data after projecting onto lower dimensions, to correct for unwanted effects of the projections. One example is the piling effect that is observed when projecting a high-dimensional distribution (Diaconis & Freedman, 1984): projected points are approximately Gaussian in most views, and increasingly concentrating near the center as dimensionality increases. The sage display (Laa, Cook, & Lee, 2020) proposes a solution to this issue, via a radial transformation of the projected data. This nonlinear transformation is defined in each point of the projection plane and is sensitive to the overall scale of the data as well as the original dimensionality p. In addition, tuning parameters can be used to obtain a more (or less) aggressive redistribution of the points. This is illustrated in Figure 10 showing the application of the sage transformation to the classical pollen dataset (Coleman, 1986). The projection without any rescaling looks similar to what is found when rescaling with default options (left), and we can use either of the two tuning parameters to better resolve the distribution near the center and reveal the hidden structure (middle and right).

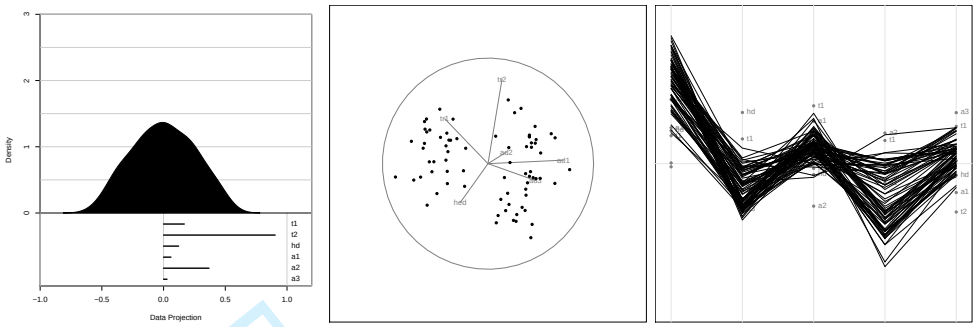

Figure 8: Various displays of the projections: 1D display in histogram (left), 2D display in scatterplot (middle), and 5D display in parallel coordinate plot (right).

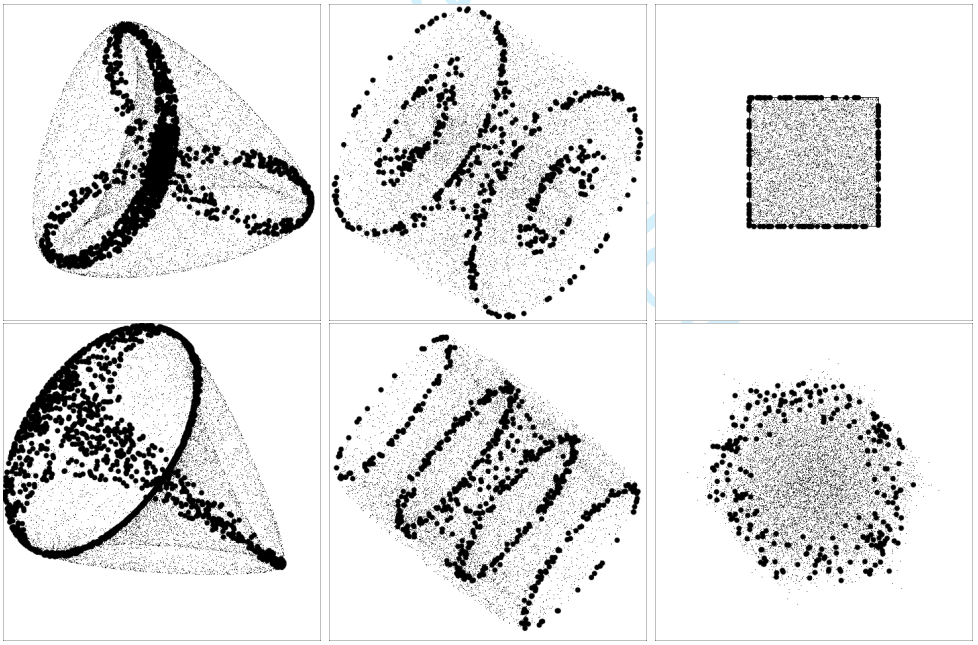

Figure 9: Still views of the slice tour display showing geometric shapes: Roman surface (left), 4D torus (middle) and a 6d cube (right). Adapted from Laa, Cook, Buja, et al., 2020, Figure 4.

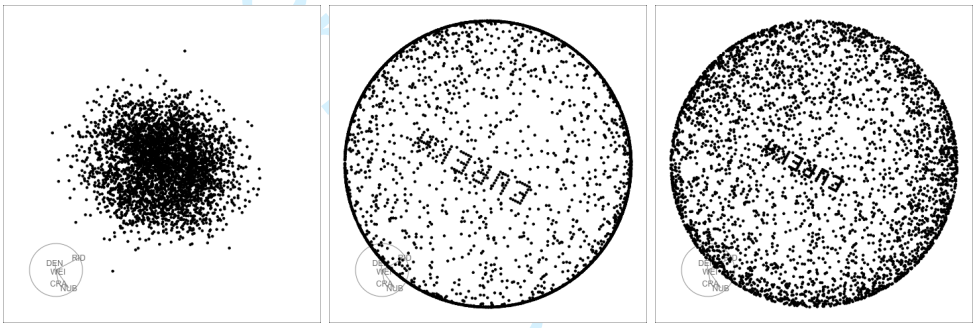

Figure 10: Snapshots of the same projection from the sage tour on the pollen data, illustrating the different tuning options. Left: using the default settings for the transformation results in a similar view as found in projections without the sage transformation. Middle: using the tuning parameter R for better resolution near the center reveals the word "EUREKA". Right: using the γ tuning parameter can also reveal the structure. Adapted from Laa, Cook, and Lee, 2020, Figure 9.

3 Ways of Interacting

3.1 Basic Interactions

Due to the dynamic nature of the tour, user interfaces can enhance the interpretability of the resulting visualization. Often one will want to pause on a particularly interesting projection and return it for use in a downstream analysis or static plot. Basic controls can be implemented so a tour display can be paused, refreshed and replayed over and over, as shown in Figure 11.

3.2 Manual Tours

Manual tours (Cook & Buja, 1997; Spyrison & Cook, 2020) offer a means to interactively control the contribution of a single variable on the projection plane. This is particularly useful for exploring a projection once a feature of interest has been identified. Manual tours can then be employed to test the structure of the feature, with respect to a selected variable. For instance, Figure 12, starts from the orthonormal linear discriminant and explores the sensitivity of class separation as the contribution of a single variable is altered.

3.3 Spin-and-brush

Brushing can be used to aid statistical and geometrical interpretations of the data; brushing can be thought of as conditioning variables on certain regions of the data or used to section lower dimensional views. This aids tasks such as the identification of outlying points or visual cluster analysis when combined with the tour. Here, a "spin-and-brush" approach works well, since different views will reveal different features and a persistent brush helps us to connect this to the previously observed information. As an example we briefly summarize an analysis from Cook et al. (2007) that explores clustering of the physics data previously used to in the context of projection pursuit (Cook et al., 1995; Friedman & Tukey, 1974). This data is well described by its geometric structure: a two-dimensional triangle with two one-dimensional strands linearly extending in different directions from each vertex. These strands can sequentially be identified as clusters using the spin-and-brush approach, as shown in Figure 13. We stop the tour each time a cluster is clearly separate from the main distribution in the current projection and brush the points in a new color. In the end the full structure becomes apparent and can be visualized by replaying the tour with all clusters highlighted in different colors.

Figure 11: The clustering example from Figure 3 using the liminal interface. Here a grand tour is displayed on the right hand side, with buttons allowing users to play, pause and refresh the tour animation. Adapted from S. Lee et al., 2020, Figure 2.

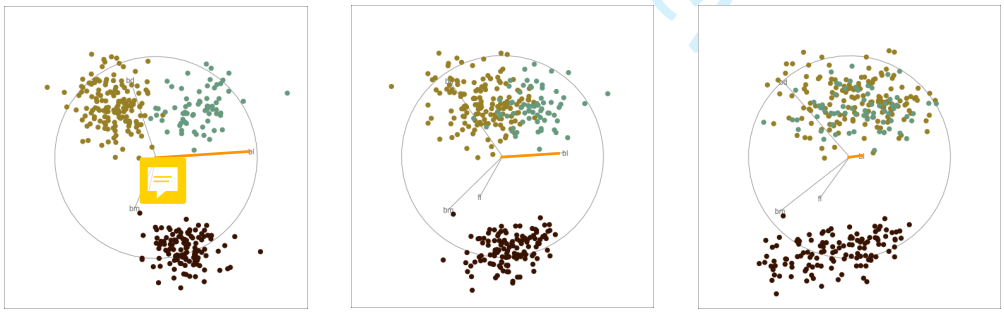

Figure 12: Three projections from a manual tour where variable "bl" is being rotated out of the projection (orange line). When this variable is removed the two light green clusters merge, which informs us that "bl" is an important variable for distinguishing between these two groups.

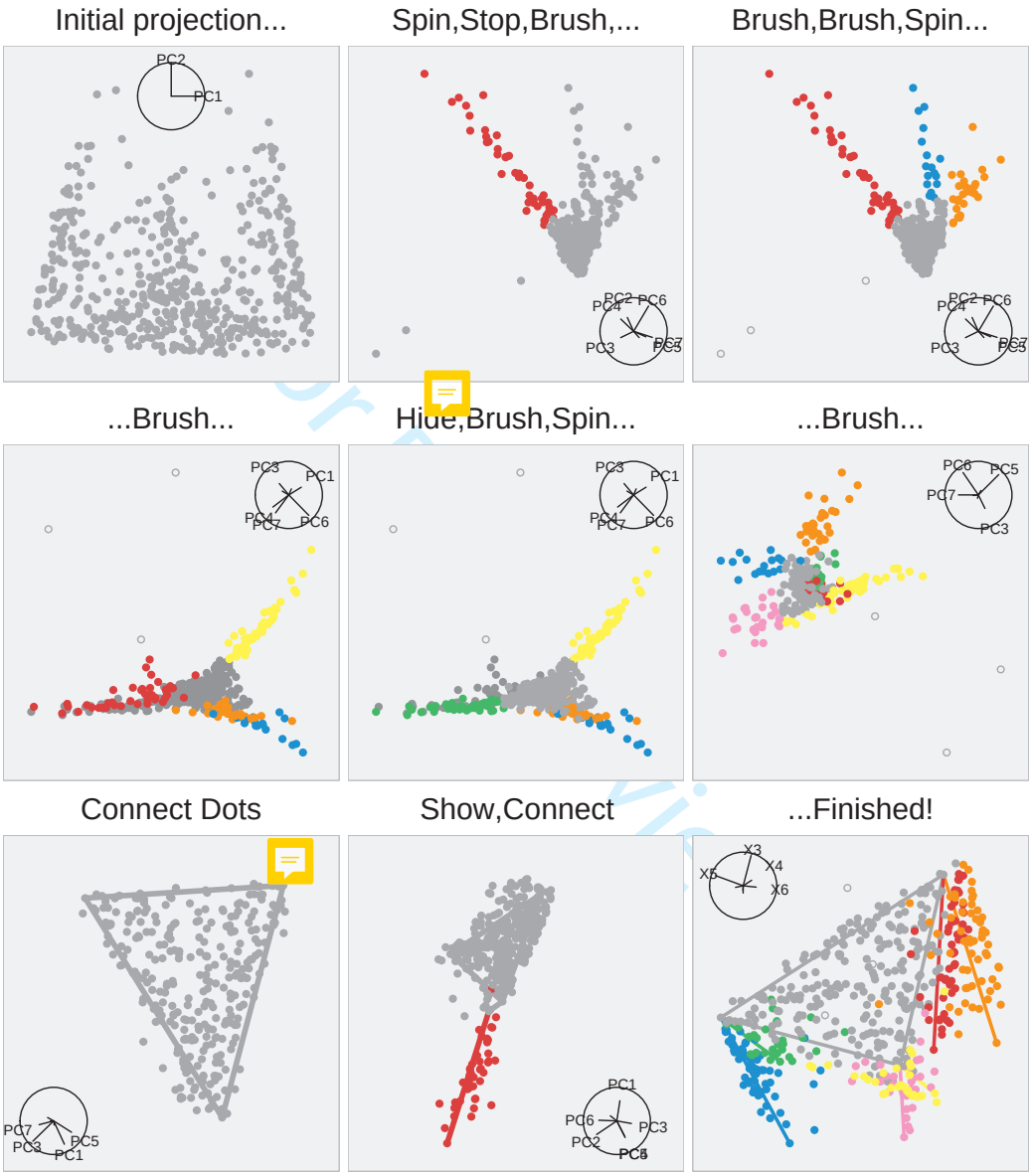

Figure 13: Stepwise identification of clusters in the physics data. Each time a cluster is clearly separated we stop the tour and brush the points, followed by another spin of the data, until we can capture the full distribution. Adapted from Cook et al., 2007, Figure 5.3.

4 Software

There has been a long history of software including tour methods, or something similar. These have laid the foundation for the current tools. The American Statistical Association Statistical Graphics Video Library web site (ASA Statistical Graphics Section, 2021) hosts videos documenting the history. It is fabulous watching the videos titled "Multidimensional Scaling" (Kruskal, 1962), "Real-Time Rotation" (Chang, 1970) and "Prim-9" (Tukey, 1973) show some preliminary methods leading to the development of the grand tour. The video titled "Use of the Grand Tour in Remote Sensing" (McDonald and Willis, 1987) is the first to show a tour, and the video titled "Dataviewer: A Program for Looking at Data in Several Dimensions" (Buja and Tukey, 1987) demonstrates the tour as part of a larger data analysis system. The videos titled "XGobi: Dynamic Graphics for Data Analysis" (Swayne et al, 1991) and "Grand Tour and Projection Pursuit" (Cook et al, 1993) show the tour tools in the XGobi software system. Some work on tours, for example Wegman (1992) and Tierney (1991) is not documented in the video collection.

Recently, some other review papers have summarized different aspects of tour methods. Moustafa and Hadi (2009) explains the relationship between a grand tour and the classical multivariate plot, the Andrews curve. Moustafa (2010) describes a computational shortcut for a tour. Martinez (2013) explains the image grand tour.

The most accessible current software is the R package called tourr (Wickham et al., 2011). This software evolved from the Dataviewer (Buja et al., 1986), XGobi (Swayne et al., 1998), GGobi (Swayne et al., 2003), Orca (Sutherland et al., 2000) and cranvas (Xie et al., 2014) ancestry. The tourr package has a wide range of display types, for different projection dimensions, and a selection of target generating methods including grand, guided, little, local, section, sage and frozen.

The R package spinifex (Spyrison & Cook, 2020) has primary features: it produces manual tours (predefined path or interactive manipulation, identifies orthonormal global feature bases with the use of the Rdimtools (You, 2020) package, renders (manual or other) tours as animations exportable to static .gif (gganimate (Pedersen & Robinson, 2020) package) or interactive .html widgets (plotly (Sievert, 2020) package), lastly it offers interactive shiny application that offers an graphical user interface to quickly sample tour features.

The liminal R package uses the tour to explore the quality of non-linear dimensionality reduction algorithms (S. Lee et al., 2020). The interface consists of two side by side views consisting of a scatter plot displaying a reduced form of the data, and an interactive tour. Controls such as play/stop/restart are implemented allowing a user to pause on interesting projections and return them to their R session for further analysis. Linked brushing is implemented on both views - if users brush on the scatter plot view they can see if and how an algorithm like t-SNE has distorted distances in higher dimensional space, while if they brush on the tour view, the tour is paused and structure like cluster separation can be ascertained.

A Python implementation of the tour written by Nguyen (2020) is also available, and would benefit from extension by other developers.

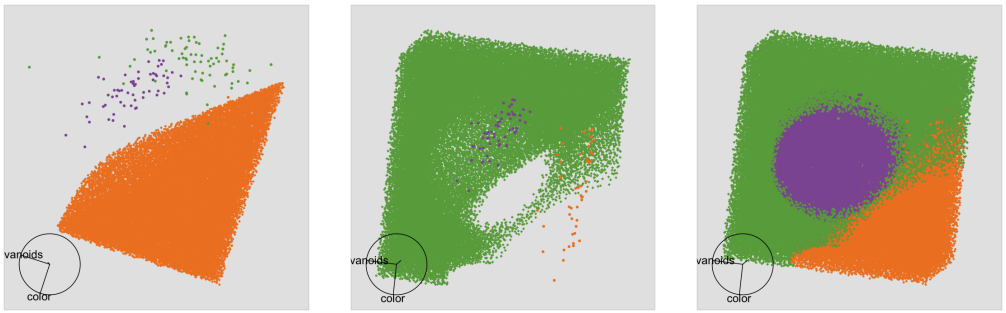

Figure 14: A grand tour run on the prediction boundaries of a radial support vector machine. (Left) The prediction region for points classified as orange produces a slice of a cube. (Middle) The predicition region for points classified as green produces another slice of the cube. (Right) The complete prediction region shows the points classified as purple are embedded within the green slice. Adapted from Wickham et al., 2015, Figure 6.

5 Applications

5.1 Model Visualization

Wickham et al. (2015) espouses the tour as a key component of their approach for visualizing models in data space. They propose using the tour to explore the high dimensional surfaces produced by model fits instead of simple summary statistics. In Figure 14 samples have been taken along the prediction region (i.e. the value the model predicts over a grid of observations) of a support vector machine used to classify three different classes (green, orange purple) by touring over the prediction region. We can see how the classifier splits the decision boundaries.

The grand tour along with direct interaction techniques has been proposed as a technique for understanding the training of deep neural network models (Li et al., 2020). They proposed a novel method for aligning the output of different layers by interpolating the output of different layers of a deep neural network via the the grand tour. They also apply the grand tour to explore the changes in single layers of the network for each training epoch. For example, by running tour over the softmax layer outputs of a classification model, an analyst may understand where in the training confusion between classes occur and gain a view of model performance (Figure 15). Similarly, the difference between training and test sets can be assessed through side by side linked tours of layer outputs. Li and Scheidegger (2020) extends this idea by first reducing the dimensionality of output layers with UMAP (McInnes et al., 2020) and then using the grand tour with manual controls over 15 dimensional embeddings to understand layer topology.

5.2 Physics

Physics models often consider $\mathcal{O}(10)$ free parameters and are compared to a much larger number of experimental observables ($\mathcal{O}(100)$). The comparison typically relies on numeric computation of the model predictions for all observables, potentially obscuring the nature of their dependence on the model parameters. Here, multivariate visualization methods can provide new insights. This was demonstrated in Cook et al. (2018). For example, we explored grouping of experiments based on how they constrain the parameter space and we also identified an interesting multivariate outlier that pointed to potential issues with the data point. For illustration we show static

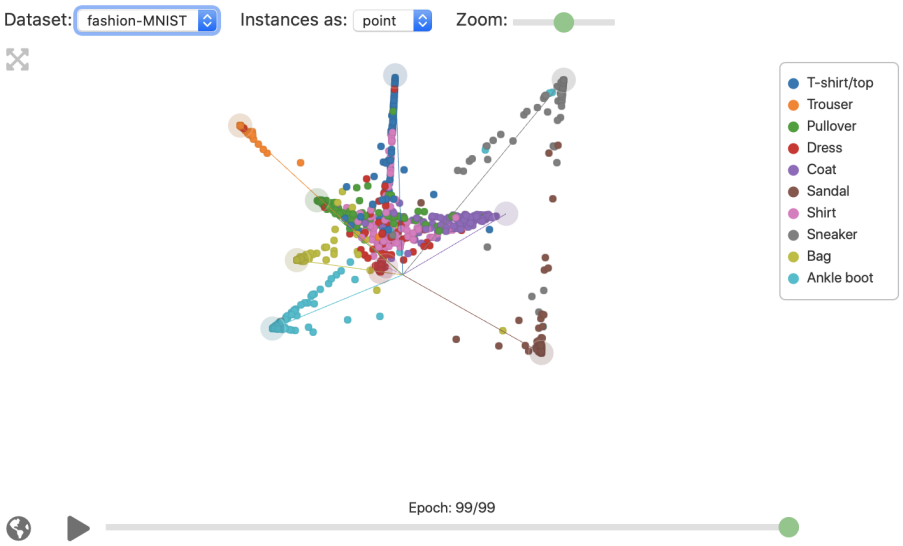

Figure 15: A grand tour run on the output of a deep neural network model for classification. The axes correspond to the probabilities that the model of an image belonging to a certain class. Adapted from Li et al., 2020.

views from the tour showing the orthogonal structure of the three groups (indicated by color), and a display which highlights the "outlyingness" of the data point marked with an asterisk symbol in Figure 16.

5.3 Bioinformatics

In single cell RNA sequencing, scientists are interested in identifying novel cell types and understanding the relationships between cells or their developmental trajectory. To achieve this they perform cluster analysis on a counts matrix or principal components and embed the results via t-SNE and label points according the cluster label. One of the main advantages of t-SNE is the avoidance of over-plotting so clusters can be clearly identified on a scatter plot, however, this can come at the cost of interpretability as global distances are distorted. In Laa, Cook, and Lee (2020), we used radial transformations of the tour projections as an alternative to t-SNE that better preserves global structure while still retaining cluster topology.

5.4 Geometry of data

High-dimensional data is on many analysts' minds in recent years, and a way to build understanding of high-dimensional spaces is to examine common high-dimensional shapes using a tour. Many different shapes can be simulated using the R package geozoo (Schloerke et al., 2016). Figure 18 shows points on the surface of two different types of high-dimensional torii.

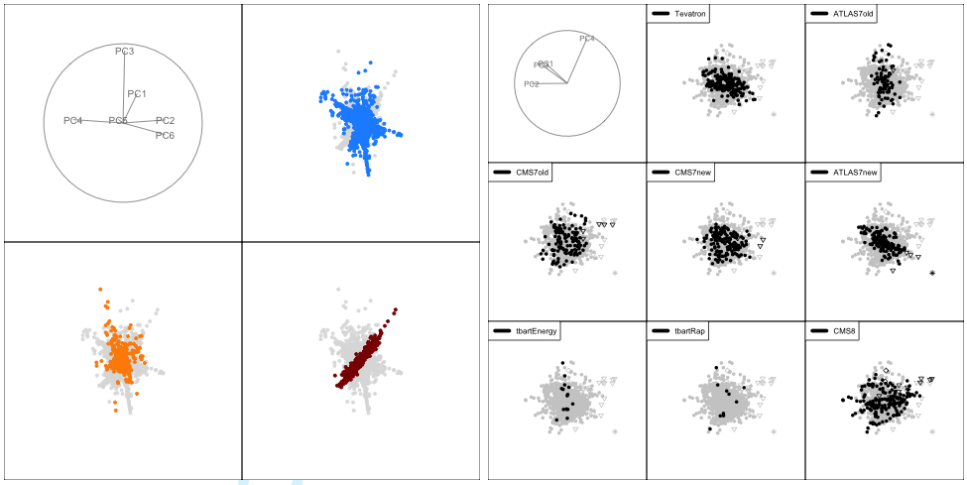

Figure 16: Selected frames from using a grand tour of the physics data. Left: the three different types of measurements (shown in different colors) are aligned along different directions in parameter space. Right: an outlying point is marked with an asterisk symbol in the ATLAS7new facet. Adapted from Cook et al., 2018, Figures 6 and 7.

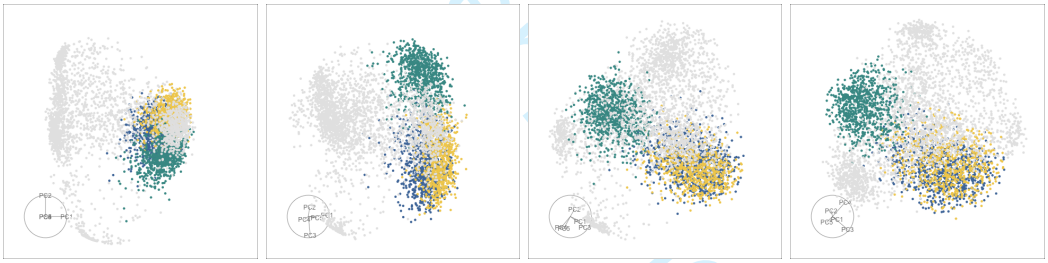

Figure 17: Selected frames from using a grand tour of mouse retina single cell RNA-seq data with the radial transformation. The sage display was used to identify and verify cluster separation between the three highlighted clusters estimated from a clustering algorithm. Adapted from Laa, Cook, and Lee, 2020, Figure 6.

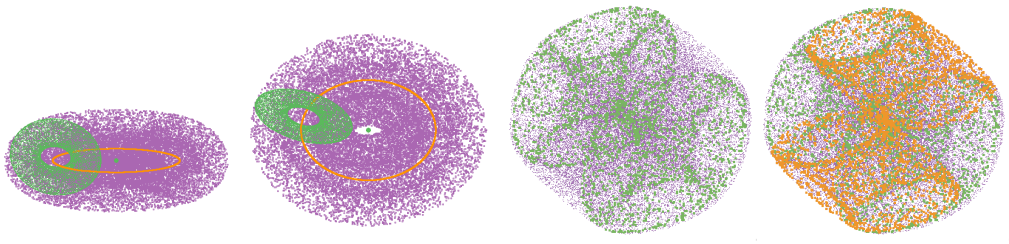

Figure 18: Several views of torii: (two at left) 4D ring torus and (two at right) 6D flat torus. In each case, a subset has been highlighted which illustrates the low-dimensional donut shapes. Adapted from Schloerke et al., 2016, Figures 10 and 13.

6 Discussion

This review highlights the history of tour methods, and modern extensions for the visualization of high dimensional numerical data. The rich ecosystem of tours is being actively developed and applied to a broad range of problems including machine learning. When tours are combined with some interactivity it can assist with analyses like clustering (the spinifex and liminal packages). Modified displays of projected data can be used to expose local structures (the sage algorithm, slicing and section pursuit). New algorithm data collection inside the tour code is being used to better diagnose optimization procedures for projection pursuit (Zhang et al., 2021).

There are many possible future directions for tour research. One big challenge for software engineering is to seamlessly embed the tour with interactivity. Currently, there is no implementation that analysts can use to perform tasks that were easy in software available in the 80s and 90s. More contributions in sectioning, as an alternative to projection, could be broadly useful. An example is conditional model visualization, as implemented in the condviz package (O'Connell et al., 2017), which assists with visual exploration of multivariate fitted models. It could be interesting to define tour paths for different types of nonlinear subspaces, or non-Euclidean spaces.

Acknowledgements

The authors gratefully acknowledge the support of the Australian Research Council. The paper was written in `rmarkdown` (Xie et al., 2018) using `knitr` (Xie, 2017). The source material for this paper is available at https://github.com/dicook/wiley-isghdd.

References

ASA Statistical Graphics Section. (2021). Video Library.

Asimov, D. (1985). The Grand Tour: A Tool for Viewing Multidimensional Data. *SIAM Journal of Scientific and Statistical Computing*, *6*(1), 128–143. https://doi.org/10.1137/0906011

Becker, R. A., & Cleveland, W. S. (1987). Brushing scatterplots. *Technometrics*, *29*(2), 127–142. https://doi.org/10.2307/1269768

Buja, A., Cook, D., Asimov, D., & Hurley, C. (2005). Computational methods for High-Dimensional rotations in data visualization. In C. R. Rao, E. J. Wegman, & J. L. Solka (Eds.), *Data mining and data visualization* (pp. 391–413). Elsevier. https://doi.org/10.1016/S0169-7161(04)24014-7

Buja, A., Hurley, C., & McDonald, J. (1986). A data viewer for multivariate data, In *Computing science and statistics: Proceedings of the 18th symposium on the interface*, Washington, American Statistical Association.

Carr, D. B., & Nicholson, W. L. (1984). *Graphical interaction tools for multiple 2- and 3-dimensional scatterplots* (tech. rep. PNL-SA-12095; CONF-8405194-2). Pacific Northwest Lab., Richland, WA (USA). https://www.osti.gov/biblio/6872143

Chambers, J. M., Cleveland, W. S., Kleiner, B., & Tukey, P. A. (1983). Plotting multivariate data. In *Graphical methods for data analysis* (pp. 129–190). Boston, Duxbury Press. https://doi.org/10.1201/9781351072304-5

Coleman, D. (1986). Geometric features of pollen grains. http://lib.stat.cmu.edu/data-expo/

Cook, D., & Buja, A. (1997). Manual Controls for High-Dimensional Data Projections. *Journal of Computational and Graphical Statistics*, *6*(4), 464–480. https://doi.org/10.2307/1390747

Cook, D., Buja, A., Cabrera, J., & Hurley, C. (1995). Grand tour and projection pursuit. *J. Comput. Graph. Stat.*, *4*(3), 155–172. https://doi.org/10.1080/10618600.1995.10474674

Cook, D., Laa, U., & Valencia, G. (2018). Dynamical projections for the visualization of PDF-Sense data. *The European Physical Journal C*, *78*(9)arXiv:1806.09742, 742. https://doi.org/10.1140/epjc/s10052-018-6205-2

Cook, D., Swayne, D. F., & Buja, A. (2007). *Interactive and dynamic graphics for data analysis: With R and GGobi*. Springer Science & Business Media. https://doi.org/10.1007/978-0-387-71762-3

Diaconis, P., & Freedman, D. (1984). Asymptotics of graphical projection pursuit. *Ann. Stat.*, *12*(3), 793–815. https://doi.org/10.1214/aos/1176346703

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, *7*(2), https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1469-1809.1936.tb02137.x, 179–188. https://doi.org/https://doi.org/10.1111/j.1469-1809.1936.tb02137.x

Friedman, J. H., & Tukey, J. W. (1974). A projection pursuit algorithm for exploratory data analysis. *IEEE Trans. Comput.*, *C-23*(9), 881–890.

Furnas, G. W., & Buja, A. (1994). Prosection views: Dimensional inference through sections and projections. *J. Comput. Graph. Stat.*, *3*(4), 323–353. https://doi.org/10.1080/10618600.1994.10474649

Horst, A. M., Hill, A. P., & Gorman, K. B. (2020). *Palmerpenguins: Palmer archipelago (antarctica) penguin data* [R package version 0.1.0]. R package version 0.1.0. https://allisonhorst.github.io/palmerpenguins/

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol.*, *24*(6), 417. https://doi.org/10.1037/h0071325

Huber, P. (1990). Data analysis and projection pursuit. In *Technical report pjh-90-1*. Dept. of Mathematics, Massachusetts Institute of Technology Cambridge, MA.

Inselberg, A. (1985). The plane with parallel coordinates. *Vis. Comput.*, *1*(2), 69–91. https://doi.org/10.1007/BF01898350

Kruskal, J. B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, *29*(1), 1–27. https://doi.org/10.1007/BF02289565

Laa, U., Cook, D., Buja, A., & Valencia, G. (2020). *Hole or grain? a section pursuit index for finding hidden structure in multiple dimensions*.

Laa, U., Cook, D., & Lee, S. (2020). *Burning sage: Reversing the curse of dimensionality in the visualization of high-dimensional data*.

Laa, U., Cook, D., & Valencia, G. (2020). A slice tour for finding hollowness in high-dimensional data. *J. Comput. Graph. Stat.*, (ja)arXiv https://doi.org/10.1080/10618600.2020.1777140, 1–10. https://doi.org/10.1080/10618600.2020.1777140

Lee, E.-K., Cook, D., Klinke, S., & Lumley, T. (2005). Projection pursuit for exploratory supervised classification. *J. Comput. Graph. Stat.*, *14*(4), 831–846.

Lee, S., Laa, U., & Cook, D. (2020). *Casting multiple shadows: High-Dimensional interactive data visualisation with tours and embeddings*.

Li, M., & Scheidegger, C. (2020). Toward Comparing DNNs with UMAP Tour. Retrieved March 11, 2021, from https://tiga1231.github.io/umap-tour/

Li, M., Zhao, Z., & Scheidegger, C. (2020). Visualizing neural networks with the grand tour. *Distill*, . https://doi.org/10.23915/distill.00025

Martinez, W. L. (2013). Image grand tour. *WIREs Computational Statistics*, *5*(3), http[illegible]/onlinelibrary.wiley.com/doi/pdf/1 198–206. https://doi.org/https://doi.org/10.1002/wics.1253

McInnes, L., Healy, J., & Melville, J. (2020). *Umap: Uniform manifold approximation and projection for dimension reduction*.

Moustafa, R. E. (2010). Pseudograd tour. *WIREs Computational Statistics*, *2*(6), https[illegible]onlinelibrary.wiley.com/doi/pdf/ 711–718. https://doi.org/https://doi.org/10.1002/wics.133

Moustafa, R. E., & Hadi, A. S. (2009). Grand tour and the andrews plot. *WIREs Computational Statistics*, *1*(2), https://onlinelibrary.wiley.com/doi/pdf/10.1002/wics.30, 245–250. https://doi.org/https://doi.org/10.1002/wics.30

Nguyen, W. (2020). An implementation of the tour algorithm in python [Accessed: 2021-4-14]. https://github.com/wgnu6/tourpy

O'Connell, M., Hurley, C. B., & Domijan, K. (2017). Conditional visualization for statistical models: An introduction to the condvis package in r. *Journal of Statistical Software, Articles*, *81*(5), 1–20. https://doi.org/10.18637/jss.v081.i05

Pedersen, T. L., & Robinson, D. (2020). *Gganimate: A grammar of animated graphics* [R package version 1.0.7]. R package version 1.0.7. https://CRAN.R-project.org/package=gganimate

Rauber, A. (2009). Multi-challenge data set [Accessed: 2020-9-17]. http://ifs.tuwien.ac.at/dm/dataSets.html

Schloerke, B., Wickham, H., Cook, D., & Hofmann, H. (2016). Escape from Boxland. *The R Journal*, *8*(2), 243–257. https://doi.org/10.32614/RJ-2016-044

Sievert, C. (2020). *Interactive Web-Based Data Visualization with R, plotly, and shiny* [Chapman Hall/CRC. https://plotly-r.com

Spyrison, N., & Cook, D. (2020). Spinifex: An R Package for Creating a Manual Tour of Low-dimensional Projections of Multivariate Data. *The R Journal*, *12*(1), 243[illegible]https://doi.org/10.32614/RJ-2020-027

Sutherland, P., Rossini, A., Lumley, T., Lewin-Koh, N., Dickerson, J., Cox, Z., & Cook, D. (2000). Orca: A Visualization Toolkit for High-Dimensional Data. *Journal of Computational and Graphical Statistics*, *9*(3), 509–529.

Swayne, D. F., Cook, D., & Buja, A. (1998). XGobi: Interactive Dynamic Graphics in the X Window System. *Journal of Computational and Graphical Statistics*, *7*(1), 113–130.

Swayne, D. F., Temple Lang, D., Buja, A., & Cook, D. (2003). Ggobi: Evolving from xgobi into an extensible framework for interactive data visualization. *Journal of Computational Statistics and Data Analysis*, *43*, 423–444.

Tierney, L. (1991). *LispStat: An Object-Orientated Environment for Statistical Computing and Dynamic Graphics*. New York, NY, Wiley.

Tukey, J. W., & Tukey, P. A. (1983). Some graphics for studying four-dimensional data. In *Computer science and statistics: Proceedings of the 14th symposium on the interface*. Springer-Verlag.

Tukey, P. A., & Tukey, J. W. (1981). Graphical display of data sets in three or more dimensions. In V. Barnett (Ed.), *Interpreting multivariate data* (pp. 189–213). Wiley.

van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *J. Mach. Learn. Res.*, *9*(Nov), 2579–2605. http://www.jmlr.org/papers/v9/vandermaaten08a.html

Wegman, E. J. (1990). Hyperdimensional data analysis using parallel coordinates. *J. Am. Stat. Assoc.*, *85*(411), 664–675. https://doi.org/10.1080/01621459.1990.10474926

Wegman, E. J. (1992). The grand tour in k-dimensions (C. Page & R. LePage, Eds.). In C. Page & R. LePage (Eds.), *Computing science and statistics*, New York, NY, Springer New York.

Wickham, H., Cook, D., & Hofmann, H. (2015). Visualizing statistical models: Removing the blindfold. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, *8*(4), 203–225. https://doi.org/10.1002/sam.11271

Wickham, H., Cook, D., Hofmann, H., & Buja, A. (2011). tourr: An R package for exploring multivariate data with projections. *Journal of Statistical Software, Articles*, *40*(2), 1–18. https://doi.org/10.18637/jss.v040.i02

Wilkinson, L., & Friendly, M. (2009). The history of the cluster heat map. *The American Statistician*, *63*(2), 179–184. https://doi.org/10.1198/tas.2009.0033

Xie, Y. (2017). *Dynamic documents with R and knitr* (2nd). Boca Raton, Florida, Chapman; Hall/CRC. https://doi.org/10.1201/9781315382487

Xie, Y., Allaire, J. J., & Grolemund, G. (2018). *R markdown: The definitive guide*. Chapman; Hall/CRC. https://bookdown.org/yihui/rmarkdown

Xie, Y., Hofmann, H., & Cheng, X. (2014). Reactive Programming for Interactive Graphics. *Statistical Science*, *29*(2), 201–213. https://doi.org/10.1214/14-STS477

You, K. (2020). *Rdimtools: An R package for dimension reduction and intrinsic dimension estimation*.

Zhang, H. S., Cook, D., Laa, U., Langrené, N., & Menéndez, P. (2021). *Visual diagnostics for constrained optimisation with application to guided tours*.

This article discusses a high-dimensional visualization technique called the tour, which can be used to view data in more than three dimensions. We review the theory and history behind the technique, as well as modern software developments and applications of the tour that are being found across the sciences and machine learning.