

NEW TOUR METHODS FOR VISUALIZING HIGH-DIMENSIONAL DATA

Ursula Laa¹ and Dianne Cook²

¹ University of Natural Resources and Life Sciences, Vienna, Institute of Statistics, (e-mail: ursula.laa@boku.ac.at)

² Department of Econometrics and Business Statistics, Monash University, (e-mail: dicook@monash.edu)

ABSTRACT: Tour methods visualize high-dimensional spaces as animated sequences of low-dimensional projections. Viewing the projected data allows us to uncover and understand shapes and patterns in such high-dimensional spaces. Typically we create this animation by first selecting a target plane, and then we interpolate to gradually move to the selected target. While several methods for target selection are available in the R package *tourr*, it currently only implements geodesic interpolation. Here we present recent developments in tour methods. We first describe a manual user-guided control for target selection, that also includes the interactive selection of sections in the context of a slice tour. We then present a new interpolation method for frame-to-frame transitions instead of plane-to-plane, important for projection pursuit applications.

KEYWORDS: data visualization, grand tour, dynamic graphics, projection pursuit

1 Introduction

The grand tour (Asimov, 1985) visualizes multivariate data distributions as animated sequences of interpolated low-dimensional projected views, and by following such an animation we can build intuition about a data distribution in a high-dimensional space. We may discover patterns of interest, for example, clustering, or we can detect outlying points. For a summary of the current state-of-the-art on tour methods and applications see Lee *et al.*, 2022.

Constructing the animation typically iterates between two steps: first, we select a target plane onto which the data should be projected (*target selection*), and then we compute the interpolated path between the current viewing plane and the selected target (*interpolation algorithm*).

Examples of target selection include a random selection (the *grand tour*, which provides an overview of the distribution across the full space), and a selection that optimizes a projection pursuit index (the *guided tour* (Cook *et al.*, 1995), which moves towards more “interesting” views of the data). The latter is important when patterns of interest may be hidden and only visible in a small

part of a much larger data space. These as well as additional approaches are implemented in the *tourr* R package (Wickham *et al.*, 2011).

The interpolation algorithm should then find a path from one target plane to the next, such that we can view the projected data as a smooth animation while gradually changing the viewing angle. Importantly each intermediate step also needs to be defined by an orthonormal projection matrix such that the data is not distorted in the visualization. The preferred approach is typically a geodesic interpolation (Buja *et al.*, 2005) which finds the shortest path between two planes, independent of the orientation of the target frame. This has the advantage that any within-plane rotation is avoided during the interpolation, but can limit applications of the guided tour when the considered projection pursuit index is not invariant to rotation within the plane (Laa & Cook, 2020).

2 Manual tour in Mathematica

With a manual tour, the user can alter the contribution of a selected variable to the target projection. This is, in particular, useful to interpret patterns found in one projection, for example through projection pursuit, to understand the sensitivity of the pattern to the input variables.

The previous approach to the manual tour (Cook & Buja, 1997; Spyrisson & Cook, 2020) was overly complicated since it requires the construction of a manipulation space as an intermediate step. Here we present a simpler approach described in Laa *et al.*, 2023. The new method uses the interactive graphics interface available in Mathematica, for illustration, we show a screenshot in Fig. 1. We can manually change the projection by dragging one of the variables in an axes display of the projection matrix. The main part of the display shows the projected data (or slices of it, defined according to Laa *et al.*, 2020), tracking changes to the projection matrix while ensuring its orthonormality.

In the presentation, we will show an example of how the new approach can be used for a detailed inspection of a fitted classification model, by comparing the decision boundaries generated by two different models.

3 Alternative interpolation methods

We need an alternative to geodesic interpolation for a guided tour that is optimizing a projection pursuit index that is not rotation invariant. This situation is illustrated in Fig. 2: the data is simulated to have a functional dependence between two of the variables (V5 and V6). We define an index that captures functional dependence in the projection: we compute the residuals

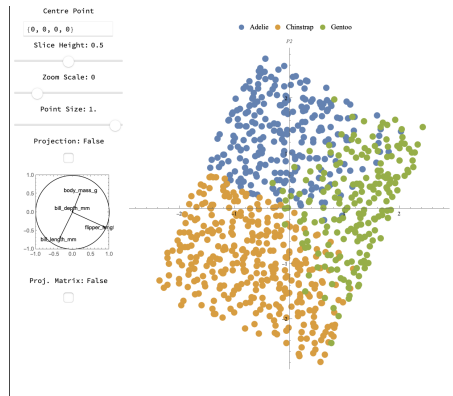


Figure 1. Screenshot of the interactive Mathematica manual tour interface.

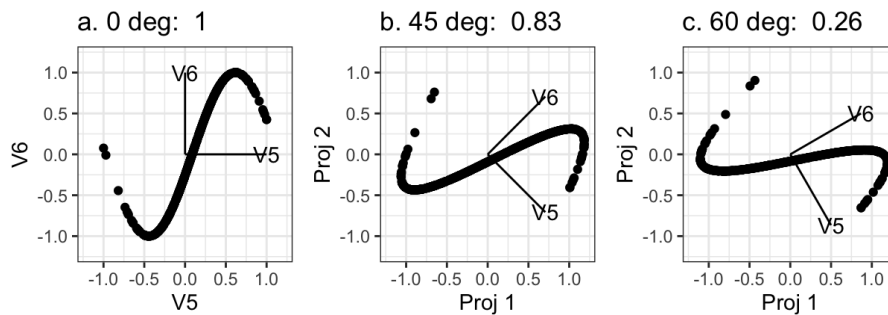


Figure 2. Simple spline index computed on within-plane rotations of the same projection, resulting in very different index values.

of a splines model using values along the projected x- and y-axis as independent and dependent variables, and normalize by the variance along the y-direction (Grimm, 2016). The index value changes dramatically when rotating within the plane: on the left we see the index taking its maximum value of 1, while on the right the index value has dropped to 0.26.

To offer an alternative interpolation method in those settings we have implemented a frame-to-frame interpolation based on Givens rotations, as suggested in Buja *et al.*, 2005. The algorithm is available through the R package *woylieR*. The presentation will give a brief overview of the algorithm and its implementation. We then show how it can be used to improve the results of a guided tour for the example of exchange rate data.

References

- ASIMOV, DANIEL. 1985. The Grand Tour: A Tool for Viewing Multidimensional Data. *SIAM Journal of Scientific and Statistical Computing*, **6**(1), 128–143.
- BUJA, ANDREAS, COOK, DIANNE, ASIMOV, DANIEL, & HURLEY, CATHERINE. 2005. Computational Methods for High-Dimensional Rotations in Data Visualization. *Chap. 14, pages 391–413 of: RAO, C R, WEGMAN, E J, & SOLKA, J L (eds), Data Mining and Data Visualization*. Handbook of Statistics, vol. 24. Elsevier.
- COOK, DIANNE, & BUJA, ANDREAS. 1997. Manual Controls for High-Dimensional Data Projections. *Journal of Computational and Graphical Statistics*, **6**(4), 464–480.
- COOK, DIANNE, BUJA, ANDREAS, CABRERA, JAVIER, & HURLEY, CATHERINE. 1995. Grand Tour and Projection Pursuit. *Journal of Computational and Graphical Statistics*, **4**(3), 155–172.
- GRIMM, KATRIN. 2016. *Kennzahlenbasierte Grafikauswahl*. doctoral thesis, Universität Augsburg.
- LAA, URSULA, & COOK, DIANNE. 2020. Using tours to visually investigate properties of new projection pursuit indexes with application to problems in physics. *Comput Stat* **35**, 1171–1205.
- LAA, URSULA, COOK, DIANNE, & VALENCIA, GERMAN. 2020. A Slice Tour for Finding Hollowness in High-Dimensional Data. *Journal of Computational and Graphical Statistics*, **29**(3), 681–687.
- LAA, URSULA, AUMANN, ALEX, COOK, DIANNE, & VALENCIA, GERMAN. 2023. New and simplified manual controls for projection and slice tours, with application to exploring classification boundaries in high dimensions. *Journal of Computational and Graphical Statistics*, to appear.
- LEE, STUART, COOK, DIANNE, DA SILVA, NATALIA, LAA, URSULA, SPYRISON, NICHOLAS, WANG, EARO, & ZHANG, H. SHERRY. 2022. The state-of-the-art on tours for dynamic visualization of high-dimensional data. *WIREs Computational Statistics*, **14**(4), e1573.
- SPYRISON, NICHOLAS, & COOK, DIANNE. 2020. spinifex: an R Package for Creating a Manual Tour of Low-dimensional Projections of Multivariate Data. *The R Journal*, **12**(1), 243.
- WICKHAM, HADLEY, COOK, DIANNE, HOFMANN, HEIKE, & BUJA, ANDREAS. 2011. tourr: An R Package for Exploring Multivariate Data with Projections. *Journal of Statistical Software*, **40**(2), 1–18.