

Dynamic visualization of high-dimensional functions via low-dimension projections and sectioning across 2D and 3D display devices

A thesis submitted for the degree of

Doctor of Philosophy

by

Nicholas S Spyrison

B.Sc. Statistics, Iowa State University



Department of Econometrics and Business Statistics

Monash University

Australia

January 2019

Contents

Acknowledgements	v
Declaration	vii
Preface	ix
Abstract	xi
1 Introduction	1
2 Literature review	3
2.1 Touring	3
2.2 Virtual reality	7
3 Tour methodology	9
3.1 Spinnifex	9
4 Display dimensionality	11
4.1 My work	11
5 Human-computer interaction of 3d projections	13
5.1 Tour in 3D	13
A Additional stuff	15
Bibliography	17

Acknowledgements

I would like to thank ...

Declaration

I hereby declare that this thesis contains no material which has been accepted for the award of any other degree or diploma in any university or equivalent institution, and that, to the best of my knowledge and belief, this thesis contains no material previously published or written by another person, except where due reference is made in the text of the thesis.

Nicholas S Spyrison

Preface

The material in Chapter 1 has been submitted to the journal *Journal of Impossible Results* for possible publication.

The contribution in Chapter ?? of this thesis was presented in the International Symposium on Nonsense held in Dublin, Ireland, in July 2015.

Abstract

This thesis is about ...

Chapter 1

Introduction

This is where you introduce the main ideas of your thesis, and an overview of the context and background.

In a PhD, Chapter 2 would normally contain a literature review. Typically, Chapters 3–5 would contain your own contributions. Think of each of these as potential papers to be submitted to journals. Finally, Chapter 6 provides some concluding remarks, discussion, ideas for future research, and so on. Appendixes can contain additional material that don't fit into any chapters, but that you want to put on record. For example, additional tables, output, etc.

Chapter 2

Literature review

2.1 Touring

2.1.1 Overview

In univariate datasets histograms, or smoothed density curves are employed to visualize data. In bivariate data scatterplots and contour plots (2-d density) can be employed. In three dimensions the two most common techniques are: 2-d scatter plot with the 3rd variable as an aesthetic (such as, color, size, height, *etc.*) or rendering the data in a 3-d volume using some perceptive cues giving information describing the seeming depth of the image ¹. When there are 4 variables: 3 variables as spatial-dimensions and a 4th as aesthetic, or a scatterplot matrix consisting of 4 histograms, and 6 unique combinations of bivariate scatterplots.

Let p be the number of numeric variables; how do we visualize data for even modest values of p (say 6 or 12)? It's far too common that visualizing in data-space is dropped altogether in favor of modeling parameter-space, model-space, or worse, long tables of statistics without visuals (Wickham, Cook, and Hofmann, 2015). Yet, we all know of the risks and possible mis-leadingness of relying too heavily on parameters alone (Anscombe, 1973; Matejka and Fitzmaurice, 2017). So why do we move away from visualizing in

¹Graphs of data depicting 3 dimension are typically printed on paper, or rendered on a 2-d monitor, they are intrinsically 2-d images. They are sometimes referred to as 2.5-d, or more frequently erroneously referred to as 3-d, more on this later.

data-space? Scalability, in a word, we are not familiar with methods that allow us to concisely depict and digest $p \geq 5$ or so dimensions. This is where dimensionality reduction comes in. Specifically, we will be focusing on touring. In the interest of time I will not belabor the diversity of dimensionality reduction, however, see (Grinstein, Trutschl, and Cvek, 2002; Carreira-Perpinán, 1997) for a quick summary. Suffice it to say that touring has a couple of salient features: linear transformations such that we can interpolate back to the original variable space and scalable, non-discarding use of dimensions. By exploring the breadth of touring techniques we are able to preserve the visualization of data-space, and the intrinsic understanding of structure and distribution of data that is more succinct or beyond the reach of statistic values alone.

Touring is a linear dimensionality reduction technique that orthogonally projects p -space down to d -space. Many of such projections are interpolated while varying the rotation of p -space and viewed in order to the effect of watching an animation of the lower dimensional embedding changing as p -space is manipulated. Shadow puppets offer a useful analogy to aid in conceptualizing touring. Imagine a fixed light source facing a wall. When a hand or puppet is introduced the 3-dimensional object projects a 2-dimensional shadow onto the wall. This is a physical representation of a simple projection, that from $p = 3$ down to $d = 2$. If the object rotates then the shadow correspondingly changes. Observers watching only the shadow are functionally watching a 2-dimensional tour as the 3-dimensional object is manipulated.

2.1.2 History

Touring was first introduced by Asimov in 1985 with his purposed Grand Tour (Asimov, 1985) at Stanford University. In which, Asimov suggested three types of Grand Tours: torus, at-random, and random-walk. The specifics of which will be discussed below in the Typology section.

Note that the the above methods have no input from the user aside from the starting basis. The bulk of touring development has largely been around dynamic display, user interaction, geometric representation, and application.

Below is a non-exhaustive list of software implementing touring in some degree, ordered by descending year:

- Spinifex (**spinifex**) – for Linux, Unix, and Windows.
- Tourr (Wickham et al., 2011) – for Linux, Unix, and Windows. R package.
- CystalVision (Wegman, 2003) – for Windows.
- GGobi (Swayne et al., 2003) – for Linux and Windows.
- DAVIS (Huh and Song, 2002) – Java based, with GUI.
- VRGobi (Nelson, Cook, and Cruz-Neira, 1998) – for use with the C2 in stereoscopic 3d displays.
- ExplorN (Carr, Wegman, and Luo, 1996) – for SGI Unix.
- XGobi (Swayne, Cook, and Buja, 1991) – for Linux, Unix, and Windows (via emulation).
- XLispStat (Tierney, 1990) – for Unix, and Windows.
- Prim-9 (Asimov, 1985; Fisher, Friedman, and Tukey, 1974) – on an internal operating system.

Support and maintenance of such implementations give them a particularly short life span, while conceptual abstraction and technically heavier implementations have hampered user growth. There have been notable efforts to diminish the barriers to entry and make touring more approachable as a data exploration tool [Huh and Song (2002); Swayne et al. (2003); Wegman (2003); Wickham et al. (2011); huang_tourgui: 2012].

2.1.3 Typology

Movement

A fundamental aspect of touring is the path of rotation. Of which there are four primary distinctions (Buja et al., 2005): random choice, precomputed choice, data driven, and manual control.

- *grand tour*, a constrained random choice such as Asimov's grand tour.
 - torus-surface (Asimov, 1985)

- Geodesic
- at-random
- random-walk
- *guided tour*, data driven tour optimiazing some objective function via (stochastic) gradient descent [Hurley and Buja (1990);].
 - holes (Cook, Buja, and Cabrera, 1993) - iterates projections that add more white space to the center of the projection.
 - cmass (Cook, Buja, and Cabrera, 1993) - find the projection with the most density or mass in the center.
 - lda (Lee et al., 2005) - linear discriminant analysis, seeks a projection where 2 or more classes are most separated.
 - pda - principal component analysis finding where the data is most spread (1d only).
 - other user-defined objective function (Wickham et al., 2011).
- *planned tour*, Precomputed choice, In which the path has already been generated or defined.
 - *little tour* (McDonald, 1982), where every permutation of variables is stepped through in order, analogous to a brute-force or exhaustive search.
- *manual tour* - Manual control, a constrained rotation on selected manipulation variable and magnitude (Cook and Buja, 1997). Typically used to explore the local area after identifying an interesting feature from another tour.
- *dependance tour*, combination of n independent 1d tours. A vector describes the axis each variable will be displayed on. ie $c(1, 1, 2, 2)$ is a 4 to 2d tour with the first 2 variables on on the first axis, and the remaining on the second.
 - *correlation tour* (Buja, Hurley, and McDonald, 1987), a special case of the dependance tour, analogous to canonical correlation analysis

... * torus: where a p -dimensional torus, T^p is created from a Cartesian product of p unit circles with $T^p \in \mathbb{R}^p$. Unfortunately uniformity of the parameters do not correlate

to uniform points on the surface of the torus. If step distance between frames is fixed, disproportionate time is spent between subspaces. If step distance is change to account for uniform points on the torus then the continuity of the tour is lost.

* at-random: where each 2-frame is chosen at random without replacement. This affords an assured uniform distribution of subspaces, but is far to discontinuous for observation. It also leaves no parameters to control. * random-walk: combines the continuity of the torus method and the uniformity of the at-random method while leaving room for a control parameter.

Geoms

Scatterplots offer a simple, general case for viewing lower-dimension embeddings of higher-dimensions. Such visualization offer p -dim down to d -dim, typically two in the case of a standard monitor. Yet, no intrinsic value stops touring being used in other graphics or geoms (geometrics). For instance using parallel coordinate plots (PCP)(Ocagne, 1885), Andrews plots (Andrews, 1972), Chernoff faces (Chernoff, 1973), anaglyphs (Rollmann, 1853), star glyphs (Siegel et al., 1972), scatterplot matrix (Chambers et al., 1983), and even pictures all offer perfectly valid graphs in p -dimensions.

This works well when the number of dimensions being toured is small (in the neighborhood of 5-10), yet the number of view, or 2-frames and we can produce from p -space suffers from the so called blessing/curse of dimensionality. In which the plethora of degrees of freedom either offer many (non-unique) solutions to a problem or something that becomes ever increasing unlikely, *ie.* a bivariate square needs .

2.1.4 Linear vs non-linear dimensionality reduction

2.2 Virtual reality

Chapter 3

Tour methodology

3.1 Spinnifex

3.1.1 Tourr

3.1.2 Plottly

Chapter 4

Display dimensionality

4.1 My work

4.1.1 XGobbi vs the C2

Chapter 5

Human-computer interaction of 3d projections

5.1 Tour in 3D

5.1.1 ImAxes / IATK

Appendix A

Additional stuff

You might put some computer output here, or maybe additional tables.

Note that line 5 must appear before your first appendix. But other appendices can just start like any other chapter.

Bibliography

- Andrews, DF (1972). Plots of High-Dimensional Data. *Biometrics* **28**(1), 125–136. (Visited on 12/19/2018).
- Anscombe, FJ (1973). Graphs in Statistical Analysis. *The American Statistician* **27**(1), 17–21. (Visited on 12/19/2018).
- Asimov, D (1985). The grand tour: a tool for viewing multidimensional data. *SIAM journal on scientific and statistical computing* **6**(1), 128–143.
- Buja, A, D Cook, D Asimov, and C Hurley (2005). “Computational Methods for High-Dimensional Rotations in Data Visualization”. en. In: *Handbook of Statistics*. Vol. 24. Elsevier, pp.391–413. [http : / / linkinghub . elsevier . com / retrieve / pii / S0169716104240147](http://linkinghub.elsevier.com/retrieve/pii/S0169716104240147) (visited on 04/15/2018).
- Buja, A, C Hurley, and JA McDonald (1987). A data viewer for multivariate data. In: *Colorado State Univ, Computer Science and Statistics. Proceedings of the 18 th Symposium on the Interface p 171-174(SEE N 89-13901 05-60)*.
- Carr, D, E Wegman, and Q Luo (1996). ExplorN: Design considerations past and present. **129**.
- Carreira-Perpinán, MA (1997). A review of dimension reduction techniques. *Department of Computer Science. University of Sheffield. Tech. Rep. CS-96-09* **9**, 1–69.
- Chambers, J, W Cleveland, B Kleiner, and P Tukey (1983). Graphical Methods for Data Analysis. eng.
- Chernoff, H (1973). The Use of Faces to Represent Points in K-Dimensional Space Graphically. *Journal of the American Statistical Association* **68**(342), 361–368. (Visited on 01/05/2019).

- Cook, D and A Buja (1997). Manual Controls for High-Dimensional Data Projections. *Journal of Computational and Graphical Statistics* 6(4), 464–480. (Visited on 04/15/2018).
- Cook, D, A Buja, and J Cabrera (1993). Projection Pursuit Indexes Based on Orthonormal Function Expansions. *Journal of Computational and Graphical Statistics* 2(3), 225–250. (Visited on 01/07/2019).
- Fisher, K, MA, JH Friedman, and JW Tukey (1974). PRIM-9: An Interactive Multidimensional Data Display and Analysis System.
- Grinstein, G, M Trutschl, and U Cvek (2002). High-Dimensional Visualizations. en, 14.
- Huh, MY and K Song (2002). DAVIS: A Java-based Data Visualization System. en. *Computational Statistics* 17(3), 411–423. (Visited on 01/06/2019).
- Hurley, C and A Buja (1990). Analyzing High-Dimensional Data with Motion Graphics. *SIAM Journal on Scientific and Statistical Computing* 11(6), 1193–1211. (Visited on 11/27/2018).
- Lee, EK, D Cook, S Klinke, and T Lumley (2005). Projection Pursuit for Exploratory Supervised Classification. *Journal of Computational and Graphical Statistics* 14(4), 831–846. (Visited on 01/07/2019).
- Matejka, J and G Fitzmaurice (2017). Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics through Simulated Annealing. en. In: *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems - CHI '17*. Denver, Colorado, USA: ACM Press, pp.1290–1294. <http://dl.acm.org/citation.cfm?doid=3025453.3025912> (visited on 12/19/2018).
- McDonald, JA (1982). INTERACTIVE GRAPHICS FOR DATA ANALYSIS.
- Nelson, L, D Cook, and C Cruz-Neira (1998). XGobi vs the C2: Results of an Experiment Comparing Data Visualization in a 3-D Immersive Virtual Reality Environment with a 2-D Workstation Display. en. *Computational Statistics* 14(1), 39–52.
- Ocagne, Md (1885). *Coordonnées parallèles et axiales. Méthode de transformation géométrique et procédé nouveau de calcul graphique déduits de la considération des coordonnées parallèles, par Maurice d'Ocagne, ...* French. OCLC: 458953092. Paris: Gauthier-Villars.
- Rollmann, W (1853). Zwei neue stereoskopische Methoden. *Annalen der Physik* 166, 186–187. (Visited on 01/05/2019).

- Siegel, JH, EJ Farrell, RM Goldwyn, and HP Friedman (1972). The surgical implications of physiologic patterns in myocardial infarction shock. English. *Surgery* **72**(1), 126–141. (Visited on 01/05/2019).
- Swayne, DF, D Cook, and A Buja (1991). *Xgobi: Interactive Dynamic Graphics In The X Window System With A Link To S*.
- Swayne, DF, DT Lang, A Buja, and D Cook (2003). GGobi: evolving from XGobi into an extensible framework for interactive data visualization. *Computational Statistics & Data Analysis*. Data Visualization **43**(4), 423–444. (Visited on 12/19/2018).
- Tierney, L (1990). *LISP-STAT: An Object Oriented Environment for Statistical Computing and Dynamic Graphics*. eng. Wiley Series in Probability and Statistics. New York, NY, USA: Wiley-Interscience.
- Wegman, EJ (2003). Visual data mining. en. *Statistics in Medicine* **22**(9), 1383–1397. (Visited on 12/19/2018).
- Wickham, H, D Cook, and H Hofmann (2015). Visualizing statistical models: Removing the blindfold: Visualizing Statistical Models. en. *Statistical Analysis and Data Mining: The ASA Data Science Journal* **8**(4), 203–225. (Visited on 03/16/2018).
- Wickham, H, D Cook, H Hofmann, and A Buja (2011). **tourr** : An R Package for Exploring Multivariate Data with Projections. en. *Journal of Statistical Software* **40**(2). (Visited on 11/23/2018).