

The effect of user interaction for understanding variable contributions to structure in linear projections

Nicholas Spyrison, Dianne Cook, Kimbal Marriott

Abstract

Viewing data in its original variable space is fundamental to the exploratory data analysis. For multivariate data this is an complex task. We perform a between-participant user study to evaluate 3 types of linear embeddings, namely, biplots of principal componts, grand tours, and radial tours. Crowdsourced participants (n=XXX, via prolific.co) were asked to identify which variable(s) explain the difference between 2 clusters of data. We find that...

Introduction

Multivariate data is ubiquitous. Yet exploratory data analysis (EDA) (Tukey 1977) of such spaces becomes difficult, increasingly so as dimension increases. Numeric statistic summarization of data often doesn't explain the full complexity of the data or worse, can lead to missing obvious visual patterns (Anscombe 1973; Matejka and Fitzmaurice 2017; **goodman_dirty_2008?**; **coleman_geometric_1986?**). Data should be visually inspected in it's original variable-space before applying models or summarizations. This allows users to validate assumptions, identify outliers, and facilitates the identification of visual peculiarities.

For these reasons, it is important to use visualizations of data spaces and extend the diversity of its application. However, visualizing data containing more than a handful of variables is not trivial. Scatterplot matrices or small multiples (Chambers et al. 1983) looks at all permutation pairs of variables, but quickly becomes to vast a number of images to consider. On the other extreme, parallel coordinates plot (Ocagne 1885) and its radial variants, plot observations as lines varying across scaled variables as displayed in a line or circle. This scales well with dimensionality, while suffering from couple issues. The larger issue, being the loss of mapping multiple variables to graphic position, which is perhaps the most important visual cue for human perception (Munzner 2014). The lesser being that they suffer from asymmetry, as their interpretation is dependent on variable ordering.

Using a linear combinations of variables will allow us to keep position in 2 display axes while peering into information not contained in any one dimension. The idea of using a combination of variables may appear daunting at first, however we do it almost exclusively in the spatial dimensions. That is to say we are rarely completely aligned with rectangular objects at any one point in time. Consider a book or a filing cabinet any orientation that isn't fully a 2D rectangle, you are seeing as a linear combination of its variables, height, width, and depth. Generalizing this to arbitrary data dimensions we can project or embed a 2D profile of p -dimensional data. Its worth noting that the number of these embedded profiles, and thus the time it takes to explore them, increase exponentially with the dimensionality of the data.

Non-lienear emeddings, the compliment of the linear embedding, have also been well received recently especially with the emergence of t-Distributed stochastic neighbor embedding (Maaten and Hinton 2008). Such techniques distort the fully dimensionality on to a low, typically 2D plane. The issue with doing so is that unit of distance is not consistent with location in the embedded space, which severely hinders the interoperability of these embeddings. Additionally they often have hyperparameters that need tuning. Doing so results in completely different or contradicting embeddings. Suffice it to say we exclude their consideration for such broad application for multivariate EDA.

Additionally there are many methods suitable for data with known classes. Linear discriminant analysis (Fisher 1936) for instance also produces linear combinations of variables, based not in order of variation of the data, but rather on the separation of known classes. In this work we want to be fully agnostic of any such class supervision and preclude them from our comparison as well.

TODO: XXX CONTINUE HERE, link In this paper we explore the

Exploring and understanding the finer structural details is an under-served aspect of multivariate data analysis. This work contained below performs a within-participant exploratory study to shed light on techniques that may be most suited for such a task.

Section formalizes the hypothesis statement. Section explains the experimental design, with sections and explaining the design factors and tasks. The results of the study are found in section . An accompanying tool is discussed in section . Discussion is covered in section .

Background

Considering that we want to explore multivariate data space, while maintaining position mapping of points. Linear combinations of variables becomes an ideal candidate. Principal component analysis (PCA) (Pearson 1901) creates new components that are linear combinations of the original variables. The creation of these variables is ordered by decreasing variation which is orthogonally constrained to all previous components. while the full dimensionality is in tact the benefit comes from the ordered nature of the components. For instance if nearly all of the variation in a data-space can be explained in the first half of its components than the complexity of viewing such a space is exponentially simplified.

Visual methods

Principal component analysis

Grand tour

Later, Asimov (Asimov 1985), coined *tour*, an animation of many projections across *continuous* changes in the basis. Exploring multivariate spaces this way offers a number of desirable features including more depth visual cues and extensible phase space exploration.

The various types of tours are distinguished by the method defining the path the basis animates. The original, and widest know, is the *grand* tour (Asimov 1985). In a grand tour, several target bases are identified by a constrained random walk. These target bases are then interpolated into many interim frames to be viewed as a more continuous animation.

Radial tour

The *manual* tour [Cook and Buja (1997);] defines its basis path by manipulating the basis contribution of a selected variable. Many such manipulations may be predefined and animated. Alternatively, these parameterized steps can allow human-in-the-loop (Karwowski 2006) interactive use.

User study

Hypothesis

Task and evaluation

Simulating data

Data collection

Sampling population

Supporting and extending the applicability of data visualization is an important endeavor. There exist various linear projection techniques to explore multivariate data spaces.

Does the animated removal of single variables via the manual tour improve the ability of the analyst to understand the importance of variables contribution to the separation of clusters?

More recently there have been advances and fanfare in non-linear projections such as self-organizing maps (Kohonen 1990), and t-SNE (Maaten and Hinton 2008). Because of the use of non-affine transformations, they offer arbitrary model spaces, without inter-operability back to variable space. This precludes them as candidates for exploratory data analysis of the multivariate data in question. They can be useful for the rapid identification of possible candidates for outliers or classifications. However they can suffer from overfitting, and crucially cannot be interpreted in terms of the original variables.

Experimental design

Below we discuss the $n = \text{XXX}$ within-participant exploratory study across 3 factors,

Groups


Each participant was randomly split into one of three even groups. The group controls the order of the factors that the participant was evaluated in for a Latin square of the 3 factors. For instance, the order of the first group was PCA, grand, manual. Group level only impacts the order the factors are displayed while task, block, and simulation order remained the same.

Factors

We explored performance across three factors. The first factor is Principal Component Analysis (PCA). The second factor is an animated walk of interpolation frames between target bases, called a *grand* tour. The third factor allows for the manual control of the individual variable's contribution to the projection, performing a *manual* tour.

All factors are shown as a scatterplot. The basis axes projection was also illustrated to the left of the plot. They are shown in a unit circle and show the magnitude and direction each variable contributes to the projection.

The user interface was kept the same whenever possible, but the control inputs did change slightly to accommodate the differences between factors. PCA had 2 side-by-side radio button inputs that select principal components to display on the x- and y-axes. The manual tour had the same axes selection, with the addition of a drop-down bar and slider control. The drop-down selects the variable to manipulate the



	Period 1	Period 2	Period 3
Gp1 (1/3*n)	Factor 1	Factor 2	Factor 3
Gp2 (1/3*n)	Factor 2	Factor 3	Factor 1
Gp3 (1/3*n)	Factor 3	Factor 1	Factor 2

	P1.T1	P1.T2	P2.T1	P2.T2	P2.T1	P2.T2	Distribution
Block 1	sim1	sim3	sim5	sim7	sim9	sim11	$\sim^d mtvN(easy^*)$
Block 2	sim2	sim4	sim6	sim8	sim10	sim12	$\sim^d mtvN(hard^*)$

where

Factor 1:	PCA	Task 1:	Number of clusters
Factor 2:	Grand tour	Task 2:	Importance of each/every variable for distinguishing between 2 cluster
Factor 3:	Manual tour		

*) Distribution difficulty discussed in detail below

Figure 1: Example case. Person 'A' is assigned to group 2, where they will use factor 2 (grand tour) for the first period. They perform 3 block difficulties of task 1 on simulations of increasing difficulty. Then 3 block difficulties of task 2 on unique simulations sampled from the same distributions of increasing difficulty. After this, they proceed to period 2, where they use factor 3 (manual tour) to perform 3 block difficulties of each task. Lastly, in the third period, they use factor 1 (PCA) to perform the tasks.

contribution of, while the slider controlled the magnitude [0-1] of the contribution of that variable on the projection. Performing this manipulation does require the contributions of the other variables to change if they are to keep their orthogonal relationship. The grand tour has no axis or variable inputs and comes precompiled as an animation of a 15 second showing 90 frames at 6 frames per second. The user can control the location or play/pause the animation at will. Each frame is a geodesic interpolation that is close to 0.1 radians away from the previous frame. These frames will typically include 6 or 7 bases identified randomly.

Blocks

Participants were randomly assigned to 1 of 3 even groups. Each group had a different factor order containing all factors. Both tasks were performed in the same order. Each task had 3 repetitions performed on new simulations that were drawn from 3 parameterizations in increasing difficulty. Each participant went through the simulations in the same order, while their factor order will vary. Fixing block difficulty order while varying factors should mitigate potential learning bias.

Fixed parameters

XXX

Tasks

Within each factor, participants performed 2 tasks in a fixed order. The first task asked participants to identify the number of clusters present in the data. In this task, clusters were unsupervised, where all observations appeared as black circles. This task does not give insight to the hypothesis, but rather served as a standard for assessing the general aptitude for this sort of high dimensional analysis as it was simpler. In application, linear discriminant analysis (Fisher 1936) or penalized discriminant analysis (Hastie, Buja, and Tibshirani 1995) are better suited for classifying such unsupervised data.

The second task is focused on the hypothesis of the study, it asked participants to identify any/all variables that were very important and somewhat important for distinguishing a given cluster from the others. For instance, which variables are very- and somewhat- important for distinguishing clusters ‘A’ and ‘B.’ This task was supervised by cluster; observations were assigned shape and (color-blind friendly) color according to their cluster. A legend identifying cluster by letter is used for the second task.

Data simulations

The data used for the study were sampled from 3 multivariate normal distributions. The distributions were parameterized with the number of clusters, the number of noise variables, and the number of variables. Simulations with 4 dimensions contained 3 clusters, while those with 6 dimensions were given 4 clusters. Each cluster containing 140 observations each. Each simulation contained 3 or 4 noise variables, which were distributed as $\mathcal{N}(0, \sigma^2)$. Non-noise variables were distributed $X_i \stackrel{d}{\sim} \mathcal{N}(\mu = 0, \sigma^2 = 1) | \mathbf{K}$. The variance-covariance matrix was constrained with non-diagonal elements selected between -0.1 to 0.6, before being constrained into a positive definitive matrix.

From the 4 sets of parameterizations, 20 simulations were drawn. The 2 most simple simulations were used during the training section of the study. All participants were exposed to the same training data sets, shown in the same order to standardize training. The remaining 18 simulations were drawn such that the remaining 3 parameterizations were sampled 6 times each. These correspond to the 3 block difficulties of a given factor and task with increasing difficulty. Referring to the middle of figure ??, a participant would perform each factor-task for 3 block difficulties with increasing difficulty before proceeding. The next factor-task has 3 new data sets but parameterized for the same order of increasing difficulty. All participants experience

the same order of simulations while the order of the factor (visualization) was changed as controlled by a partition into 3 even groups (top of the same figure).

Measures and survey

The plot display of the first task was limited to 1 minute and 3 minutes on the second task. Responses were available during and after the timer was running. The value and time of each response were captured in a temporary variable that was written to the response table once the user proceeded to the next page. The number of plot manipulations and response entries was also captured for each page including training.

After responses for each task were collected, participants were given a short survey containing questions gauging demographics, experience, and subjective evaluation of each factor on a 9-point Likert scale. The questions and possible responses are as follows:

- Which sex are you? [decline to answer, female, male, intersex/other]
- What is your age group? [decline, 19 or younger, 20 to 29, 30 to 39, 40 or older]
- What is your English proficiency?, [decline to answer, English first language, Multilingual including English from a young age, English not first language]
- What is your highest completed education? [decline, high school, undergraduate, honors/masters/MBA, doctorate]

likert-like scale [1-9], least agreement to most agreement: * I am experienced with data visualization. * I am educated in multivariate statistical analysis * I was already familiar with visualization, for each of the 3 factors * I found this visualization easy to use, for each of the 3 factors * I felt confident in my answers with this visualization, for each of the 3 factors * I liked using this visualization, for each of the 3 factors

The code, response files, their analyses, and study application are made publicly available at on GitHub at github.com/nspyrisn/spinifex_study.

Training

The training was controlled for all participants as much as possible. All participants received the same written interface instructions and watched the same training video introducing the methods and the same task prompts were displayed for their respective tasks. The factor-, interface-, and task- training took place in a continuous block where questions were invited. Questions were disallowed once the formal evaluation section started.

Participant population

A sample of convenience was taken from postgraduate students in the department of econometrics and business statistics and the faculty of information technology at Monash University, based in Melbourne, Australia. Participants were required to have prior knowledge of multivariate data visualizations.

Results

#TODO: XXX Need to run study and add results here.

Accompanying tool: spinifex application

To accompany this study we have produced a more general use tool to perform such exploratory analysis of high dimensional data. The R package, **spinifex**, {Spyrison and Cook (2019)} R package contains a free, open-source **shiny** (Chang et al. 2018) application. The application allows users to explore their data with either interactive or predefined manual tours without the need for any coding. Limited implementations of grand, little, and local tours are also made available. Data can be imported in .csv and .rda format, and projections or animations can be saved as .png, .gif, and .csv formats where applicable. Run the following R code for help getting started.

```
install.packages("spinifex", dependencies = TRUE)
spinifex::run_app("intro")
spinifex::run_app("primary")
```

Discussion

Acknowledgments

This article was created in R (R Core Team 2019), using **knitr** (Xie 2014) and **rmarkdown** (Xie, Allaire, and Golemund 2018), with code generating the examples inline. The source files for this article, application, data, and analysis can be found at github.com/nspyrison/spinifex_study/. The source code for the **spinifex** package and accompanying shiny application can be found at github.com/nspyrison/spinifex/.

Bibliography

- Anscombe, F. J. 1973. "Graphs in Statistical Analysis." *The American Statistician* 27 (1): 17–21. <https://doi.org/10.2307/2682899>.
- Asimov, Daniel. 1985. "The Grand Tour: A Tool for Viewing Multidimensional Data." *SIAM Journal on Scientific and Statistical Computing* 6 (1): 128–43. <https://doi.org/10.1137/0906011>.
- Chambers, J. M., W. S. Cleveland, B. Kleiner, and P. A. Tukey. 1983. "Graphical Methods for Data Analysis."
- Chang, Winston, Joe Cheng, J. J. Allaire, Yihui Xie, and Jonathan McPherson. 2018. *Shiny: Web Application Framework for r*. <https://CRAN.R-project.org/package=shiny>.
- Cook, Dianne, and Andreas Buja. 1997. "Manual Controls for High-Dimensional Data Projections." *Journal of Computational and Graphical Statistics* 6 (4): 464–80. <https://doi.org/10.2307/1390747>.
- Fisher, Ronald A. 1936. "The Use of Multiple Measurements in Taxonomic Problems." *Annals of Eugenics* 7 (2): 179–88. <https://doi.org/10.1111/j.1469-1809.1936.tb02137.x>.
- Hastie, Trevor, Andreas Buja, and Robert Tibshirani. 1995. "Penalized Discriminant Analysis." *The Annals of Statistics*, 73–102.
- Karwowski, Waldemar. 2006. *International Encyclopedia of Ergonomics and Human Factors, -3 Volume Set*. CRC Press.
- Kohonen, Teuvo. 1990. "The Self-Organizing Map." *Proceedings of the IEEE* 78 (9): 1464–80.
- Maaten, Laurens van der, and Geoffrey Hinton. 2008. "Visualizing Data Using t-SNE." *Journal of Machine Learning Research* 9: 2579–2605.

- Matejka, Justin, and George Fitzmaurice. 2017. “Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics Through Simulated Annealing.” In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems - CHI '17*, 1290–94. Denver, Colorado, USA: ACM Press. <https://doi.org/10.1145/3025453.3025912>.
- Munzner, Tamara. 2014. *Visualization Analysis and Design*. AK Peters/CRC Press.
- Ocagne, Maurice d'. 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, ...* Paris: Gauthier-Villars.
- Pearson, Karl. 1901. “LIII. On Lines and Planes of Closest Fit to Systems of Points in Space.” *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2 (11): 559–72.
- R Core Team. 2019. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. <https://www.R-project.org/>.
- Spyrison, Nicholas S., and Dianne Cook. 2019. *Spinifex: Manual Tours, Manual Control of Dynamic Projections of Numeric Multivariate Data* (version 0.1.0.9000). <https://github.com/nspyrison/spinifex/>.
- Tukey, John W. 1977. *Exploratory Data Analysis*. Vol. 32. Pearson.
- Xie, Yihui. 2014. “Knitr: A Comprehensive Tool for Reproducible Research in r.” In *Implementing Reproducible Computational Research*, edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC. <http://www.crcpress.com/product/isbn/9781466561595>.
- Xie, Yihui, J. J. Allaire, and Garrett Grolemond. 2018. *R Markdown: The Definitive Guide*. Boca Raton, Florida: Chapman; Hall/CRC. <https://bookdown.org/yihui/rmarkdown>.