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 foundamental 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 = 108, 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 2008; Coleman 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 matricies 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-linear 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 interoperabilty 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.

In multivariate spaces, performance measures and computational complexity are regularly compare to like algorithms and models. Human perception and inference from visuals is notably missing. We perform a within-participant, crowd sourced user study exploring the efficacy of 3 methods of linear embedding visualizations.

Section discusses the visualization methods. Section goes into the user study. The subsection digs into the task and its evaluation. The results of the study are in section .Discussion is covered in section . An accompanying tool is discussed in section .

Background, visual methods

Linear projection notation

Consider a numeric data matrix with n observations of p variables,

$$\begin{gathered}\mathbf{X}_{[n,p]} = (\mathbf{x}_1,\ ...,\ \mathbf{x}_p) \\ \mathbf{x}_i = (x_{1i},\ ...,\ x_{ni}) \mid i \in [1,p]\end{gathered}$$

Let $Y_{[n,d]}$ be the d-dimensional projection or embedding of $\mathbf{X}_{[n,p]}$ via matrix multiplication of a particular orthonormal basis matrix $\mathbf{B}_{[p,d]}$.

$$\begin{gathered}\mathbf{Y}_{[n,d]} = \mathbf{X}_{[n,p]}\mathbf{B}_{[p,d]} \mid \mathbf{B}\ \textit{is orthonormal} \\ \mathbf{y}_j = (y_{1j},\ ...,\ y_{nj}) \mid j \in [i,d]\end{gathered}$$

A matrix is said to be orthonormal if and only if they are 1) orthogonal, that is all column pairs are independent, having a cross product of 0, and 2) normal, each columns has a norm distance of 1.

Principal Component Analysis

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.

Grand Tours

Later, Asimov (Asimov 1985), coined data visualization *tour*, an animation of many linear projections across local changes in the basis. One of key features of the tour is the object permanence of the data points. That is to say by watching near by, orthogonally-interpolated frames one can track the relative changes of observations as variable contributions change.

Asimov originally purposed the *grand* tour. To start, several target bases are randomly selected. These target bases are then orthogonally-interpolated between with a fixed target distance between interpolation frames. The data matrix is premultiplied to the array of interpolated bases and rendered into an animation. There is no user interaction in a grand tour and the target.

Manual Tours

The *manual* tour (Cook and Buja 1997; Spyrison and Cook 2020) defines its basis path by manipulating the basis contribution of a single selected variable. A manipulation dimension is appended onto the projection plane, with a full contribution given to the selected variable. The target bases are then selected based on rotating this newly created manipulation space. The target bases are then similarly orthongally-interpolated, data projected, and rendered into an animation. In order for variables to remain independent of each other the contributions of the other variables must also change, *ie.* the orthonormally of the dimension space should be preserved. A key feature of the manual tour is that it affords users a way to control the variable contributions of the next target basis. This means that such manipulations can be select and queued in advance or select on the spot for human-in-the-loop analysis(Karwowski 2006). Due to the huge volume of p-space (an aspect of the curse of dimensionality (Bellman 1957)) and the abstraction constrained interpolation of the basis navigating large changes in the basis can become cumbersome. It is advisable to first identify a basis of particular interest and then use a manual tour as a finer, local exploration tool to observe how the contributions of the selected variable does or does not contribute to the feature of interest.

In order to simplify the task and keep its duration realistic we consider a variant of the manual tour, called a *radial* tour. In a radial tour the selected variable is allowed to change its magnitude of contribution, but not it's angle; it must move along the direction of its original radius.

User study

Hypothesis

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

PCA will be compared as a baseline as it is a popular stationary linear embedding. The grand tour will act as a secondary control that includes the object permanence of the data to near by frames, but with the ability to check individual variable or influence it's path. using these as comparisons we want to identify how much, if any, the radial tour helps an analyst to interpret the contributions of individual variables.

Task and evaluation

The display was a 2D scatterplot with observations supervised with the shape and color of the data points mapped to their cluster. There were either 3 or 4 clusters with even number of observations. Participants were asked to 'check any/all variables that contribute more than average to the cluster separation green circles and orange triangles,' which was further explained in the explanatory video as 'mark and and all variable that carry more than their fair share of the weight, or 1/4th in the case of 4 variables.'

The instructions iterated several times in the video was: 1) Use the input controls to find a frame that contains separation between the clusters of green circles and orange triangles, 2) look at the orientation of the variable contributions in the gray circle, a visual depiction of basis, and 3) select all variables that contribute more than average in the direction of the separation in the scatterplot.

The evaluation measure of this task was designed to have a few of features: 1) the sum of squares of the individual variable marks should be 1. 2) The sum of the correct variable(s) is 1, incorrect variables sum to

-1, a selection of all or none (disallowed in application) should sum to 0. With these in mind we define the following measure for evaluating the task:

Given a specific dataset $\mathbf{X}$ containing k clusters, let $\mathbf{S}_j$ be the set of observations in cluster j and $\overline{S_j}$ be a vector of variable means for cluster j.

$$\mathbf{X}_{[n,\ p]} = (x_1,\ ....,\ x_p)$$
$$\mathbf{S}_j = (s_{j1},\ ....,\ s_{ji}) \mid j \in [1, k],\ \mathbf{S}_j \in \mathbf{X}$$
$$\overline{S_j} = (\overline{s_1},\ ....,\ \overline{s_p})$$

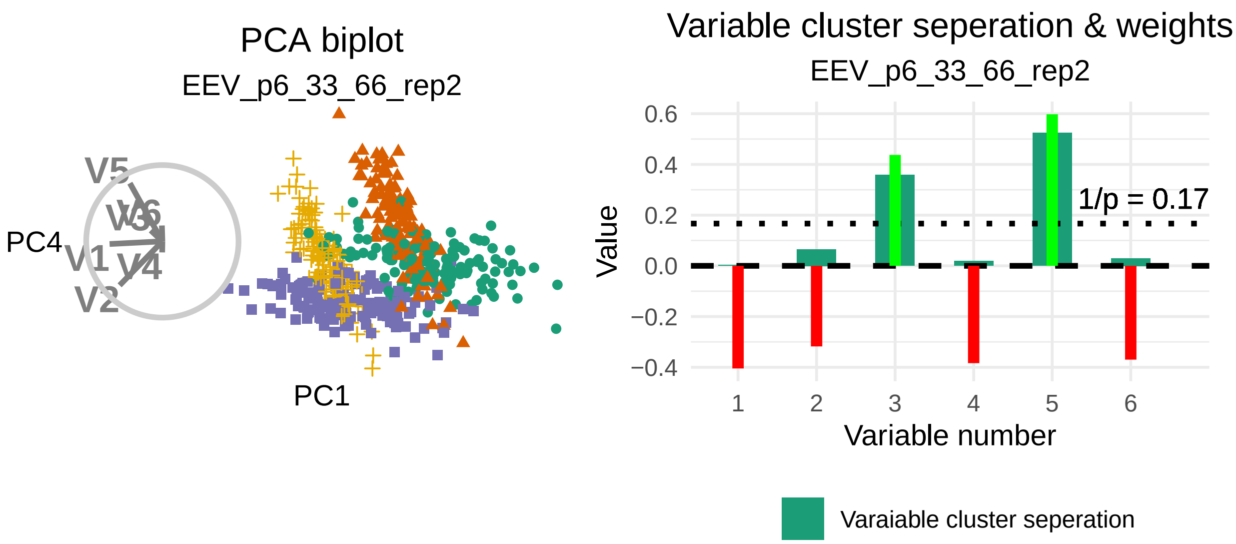
We define weights, W to be a vector of variable marks of each variable if selected, the fraction of the difference in variable means, less $1/p$, the weight each variable would hold is the signal was uniformly distributed among the variables.

$$\begin{aligned} W &= (w_1,\ ...,\ w_p) \\ &= \frac{(\overline{s_{b1}} - \overline{s_{a1}},\ ...,\ \overline{s_{bp}} - \overline{s_{ap}})}{\sum_{i=1}^{p}(|\overline{s_{bi}} - \overline{s_{ai}}|)} - \frac{1}{p} \end{aligned}$$

Participant responses, R are a vector of true/false values indicating if the participant thinks the variable separates the two clusters more than if the separation was spread uniformly between variables. Then M is a vector of variable marks.

$$\begin{aligned} M &= (m_1,\ ...,\ m_p) \\ &= I(r_i) * sign(w_i) * \sqrt{|w_i|} \end{aligned}$$

Where I is the indicator function. Then the total marks for this task is the sum of the marks vector.



TODO: XXX, NEED caption and title for figure

Experimental design

TODO: XXX CONTINUE WRITING HERE Below we discuss the n = 108 within-participant exploratory study across 3 factors,

Factor application

We explored performance across three factors. The first factor is 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 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 and grouping

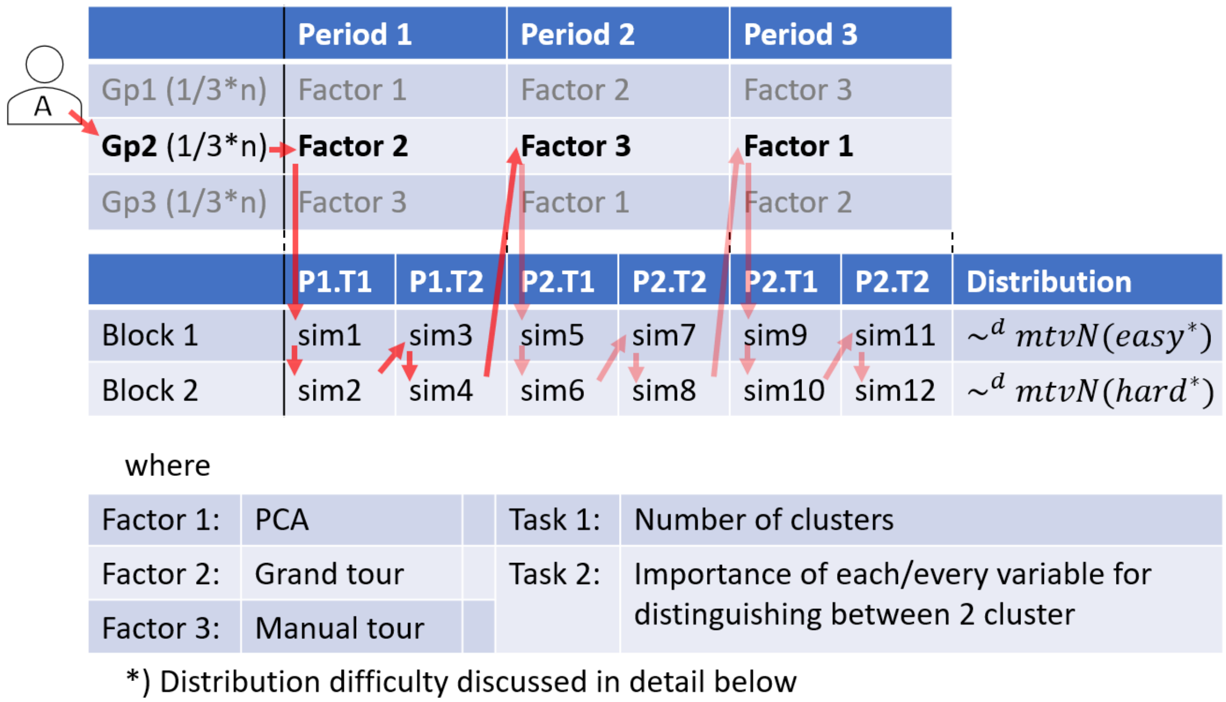


Figure 1: Example case. Person 'A' is assigned to group 2, where they will use factor 2 (grand tour) for the first period. They perform 3 block difficulties of task 1 on simulations of increasing difficulty. Then 3 block difficulties of task 2 on unique simulations sampled from the same distributions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

Synthetic data and fixed parmeters

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 $X_i \overset{d}{\sim} \mathcal{N}(\mu = 0, \sigma^2 = 1)|\mathbf{K}$. 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 difficultly 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).

Post study survey

The plot display of the first task was limited to 1 minute and 3 minutes on the second task. Responses we 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 5-point Likert scale. The questions and possible responses are as follows:

1) What are your preferred pronouns? [decline to answer, he/him, she/her, they/them or other]
2) Which age group do you belong to? [decline to answer, 18 to 24, 25 to 35, 36 to 45, 45 to 60, 60 and up]
3) What is your highest completed education? [decline to answer, Undergraduate degree (BA/BSc/other), Graduate degree (MA/MSc/MPhil/other), Doctorate degree (PhD/other); prolific.co participants were filtered to those stating they had an least an undergraduate degree]

likert scale [1-5], least agreement to most agreement: 4) I understand the how to perform the task. 5) I am experienced with data visualization. 6) I am educated in multivariate statistical analysis.

for each factor:

7-9) I was already familiar with visualization. 10-12) I found this visualization easy to use. 13-15) I felt confident in my answers with this visualization. 15-18) I liked using this visualization.

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

Sampling population

cite Kadek's paper or another pointing to prolific.co?

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.

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.

Results

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

Discussion

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 2020) R package contains a free, open-source `shiny` (Chang et al. 2020) 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.

Acknowledgments

This article was created in R (R Core Team 2020), using `knitr` (Xie 2014) and `rmarkdown` (Xie, Allaire, and Grolemund 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/https://doi.org/10.1137/0906011.

Bellman, Richard Ernest. 1957. *Dynamic Programming*. Princeton University Press. https://books.google.com.au/books?id=fyVtp3EMxasC&redir_esc=y.

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. 2020. *Shiny: Web Application Framework for r.* https://CRAN.R-project.org/package=shiny.

Coleman, David. 1986. "Geometric Features of Pollen Grains." In. Statistical Computing Statistical Graphics. http://stat-computing.org/dataexpo/1986.html.

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.

Goodman, Steven. 2008. "A Dirty Dozen: Twelve p-Value Misconceptions." *Seminars in Hematology*, Interpretation of quantitative research, 45 (3): 135–40. https://doi.org/10.1053/j.seminhematol.2008.04.003.

Karwowski, Waldemar. 2006. *International Encyclopedia of Ergonomics and Human Factors, -3 Volume Set.* CRC Press.

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*. 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. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Spyrison, Nicholas, and Dianne Cook. 2020. "Spinifex: An r Package for Creating a Manual Tour of Low-Dimensional Projections of Multivariate Data." *The R Journal* 12 (1): 243. https://doi.org/10.32614/RJ-2020-027.

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 Grolemund. 2018. *R Markdown: The Definitive Guide*. Boca Raton, Florida: Chapman; Hall/CRC. https://bookdown.org/yihui/rmarkdown.