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 = 127, 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 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-linear embeddings, 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,

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

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 $\mathbf{B}_{[p,d]}$.

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

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 orthogonally-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 selected 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.

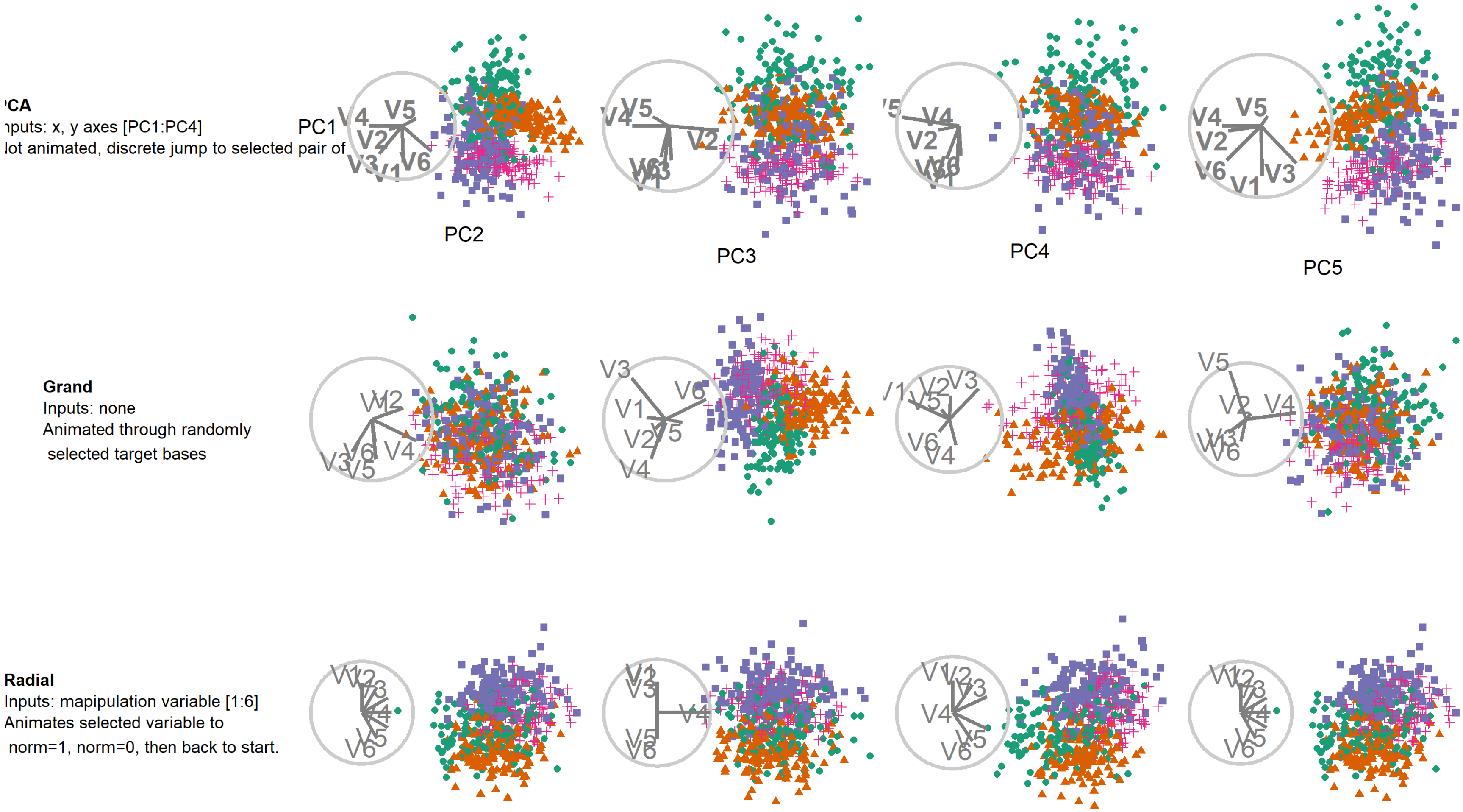

Figure 1: Example permuptaion selection. TODO XXX NEEDS CAPTION AND ROTATE TO LANDSCAPE.

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 quarter 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. Regardless of factor and block values participants were limited to 60 seconds for each evaluation of this task.

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 should sum to 0. With these in mind we define the following measure for evaluating the task:

Let a dataset $\mathbf{X}$ be a simulation containing clusters of observations of different distributions. Let $\mathbf{X}_k$ be the subset of observations in cluster k containing the p variables.

$$\begin{aligned}
&\mathbf{X}_{[n,\ p]} = (x_1,\ ....,\ x_p) \\
&\mathbf{X}_{[n_k,\ p]k} = (x_1,\ ....,\ x_p) \mid n_k \in [1, n],\ \text{an observation subset of } \mathbf{X}
\end{aligned}$$

where

$$x_{i,j,k} \text{ is scalar; observation } i \in (1, ...n),\ \text{variable } j \in (1, ...p),\ \text{cluster } k \in (1, ...K)$$

We define weights, W to be a vector explaining the variable-wise differnce between 2 clusters. Namely the difference of each variable between clusters, as a proportion of the total difference, less $1/p$ the amount of different each variable would hold if it were uniformly distributed.

$$\begin{aligned}
W &= \frac{(\overline{X_{j=1,k=1}} - \overline{X_{j=1,k=2}},\ ...\ (\overline{X_{j=p,k=1}} - \overline{X_{j=p,k=2}})}{\sum_{j=1}^{p}(|\overline{X_{j,k=1}} - \overline{X_{j,k=2}}|)} - \frac{1}{p} \\
&= (w_1,\ ...,\ w_p)
\end{aligned}$$

Participant responses, R are a vector of logical values, whether or not participant thinks the variable separates the two clusters more than if the difference uniformly distributed. Then M is a vector of variable marks.

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

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

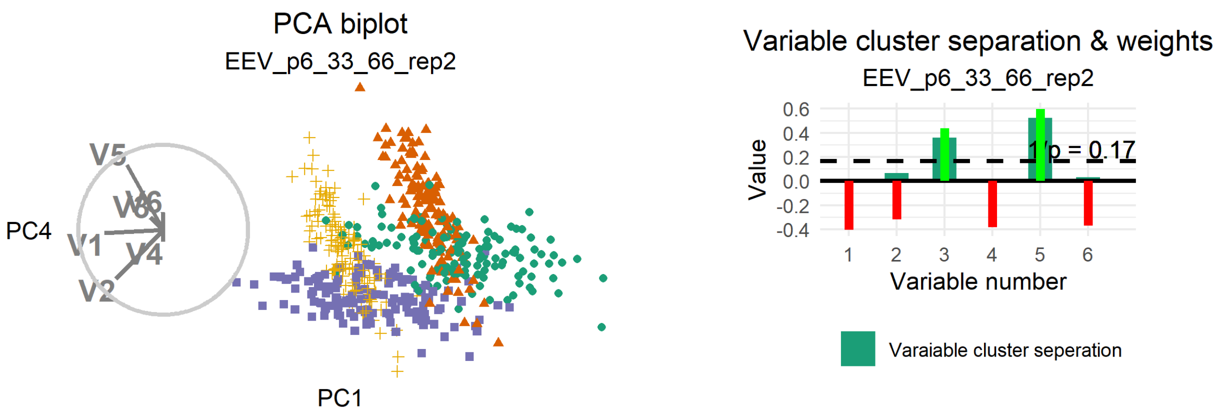

Figure 2: (L), PCA Biplot of the components showing the most cluster separation with (R) Score and weight evaluation. The bar is the absolute standardized separation of cluster means explained by the variable. The dashed line is 1 / dimensionality, the ammount of separation each variable would have if evenly detributed. The green/red lines are the marks of each variable if selected. These are the signed square of the differce between each variable value and the dashed line.

Each of the 3 periods introduced a new factor, where participants were first able to explore an untimed task with data under the simplest parameterization. The training allows the participant to become familiar with the inputs and visual specific to the factor. Upon clicking the proceed button text containing the correct answer displays with visual still intact to explore further. After the training, participant performed 2 evaluation trials. After 60 second the display was removed, though few participants spent 60 second on any particular task. These evaluation trials where performed under different parameterizations as explained in section .

Factor application

Section gives the sources and a description of the visual factors PCA, grand tours, and radial manual tours. Below we cover the aesthetic standardization, as well the unique input and display within each factor.

The visualization methods were selected to standardized wherever possible. All aesthetic values (colors, shapes, sizes, absence of legend, and absence axis titles) were held consistent. Variable contributions where always shown left of the scatterplot embeddings with their aesthetic values consistent as well.

PCA inputs allowed for users to select between the top 4 principal components for both the x and y axis regardless of the data dimensionality (either 4 or 6).

There were no user input for grand tour, users were instead shown a 15 second animation of the same randomly selected path. Users were able to view the same clip up to 4 times within the time limit.

Radial tours were also displayed at 5 frames per second with in interpolation step size of 0.1 radians. Users where able to swap between the 4 or 6 variables, upon which the display would change the the start of radially increasing the contribution of the selected variable till it was full, zeroed and then back to the initial. The complete animation of any 1 variable takes about 20 seconds, and is almost fully in the projection frame at around 6 second. The starting basis of each is initialized to a half-clock design, where the 4 or 6 variables were evenly distributed in half of the circle which is then orthonormalized. This is done to give no variable preference while minimizing variable interactions, as variables opposite of the manipulation variable must lose contribution as the other is rotated to full contribution (and vice versa).

Blocks and parameterization

The volume the parameter-space increase more than exponentially with the dimensionality of the data. Care must be taken to select realistic parameter values. We vary the values for 3 aspects of the simulated data including 1) The dimensionality of the data. 2) the shapes of the clusters, by changing the variance-covariance of the clusters. 3) The location of the difference between clusters, by mixing a signal and a noise variable at different ratio.

We test 2 levels of dimensionality, 4 dimensions containing 3 clusters and 6 dimensions with 4 clusters. Each cluster samples 140 observations. Each dimension is originally distributed as $\mathcal{N}(2 * I(signal), 1) \mid covariances\ \Sigma$ (before signal mixing and standardizing by standard deviation). Signal variables have correlation 0.9 when they have equal orientation and -0.9 when their orientations vary. Noise variables were restricted to 0 correlation. The training always uses 4 dimensions, while the 2 evaluations always contain 4 and 6 dimensions in order of increasing difficulty.

For choosing the shape of the clusters we follow the convention given in by the mclust (Scrucca et al. 2016) who name and categorize 14 variants of distributions of data containing for 3-clustered. The name of the shaped is mapped to the initial for a model's volume, shape, and orientation. We use the EEE, EEV, and EVV, which is further modified by moving 4 fifths of the data out in an "V" or banana-like shape. Figure 3 shows the principal component bi-plot of the 3 three model variants applied here. The training always uses 4 dimensions, while the 2 evaluations always contain 4 and 6 dimensions in order of increasing difficulty. The training data sets use the EEE model. The evaluation periods use EEE, EEV, and EVV-banana respectively in increasing order of difficulty.

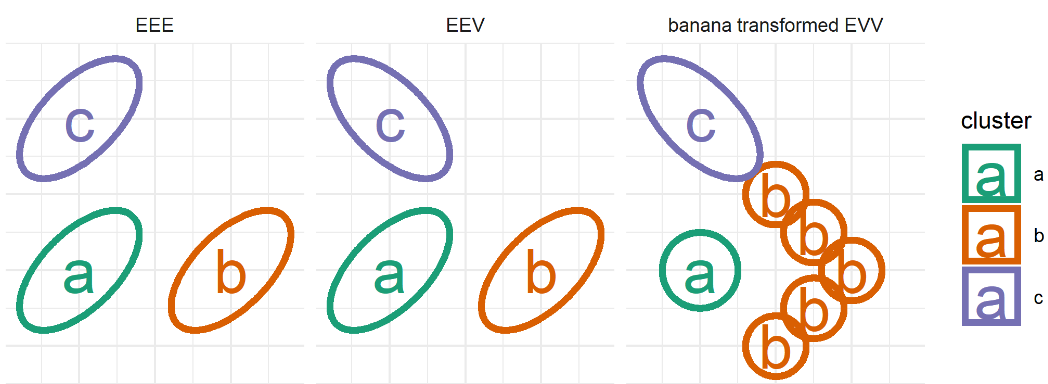

Figure 3: Illustrate the shapes of the vc model families. TODO XXX NEEDS CAPTION

The separation of any two target clusters is currently contained fully within 1 variable at this point. We mix this variable with a noise variable such that the difference in the clusters is mixed at the following respective percentages 100/0% (not mixed), 66/33%, 50/50% (signal evenly mixed). The training always uses 4 dimensions, while the 2 evaluations always contain 4 and 6 dimensions in order of increasing difficulty. The training data does not mix separation. Location mixing within an evaluation period is held constant and rotated through the 6 permutations of their order. Randomizing the order of the location mixing is controlled by iterating once after each of the 6 factor order permutations are evaluated. This is illustrated in figure 4.

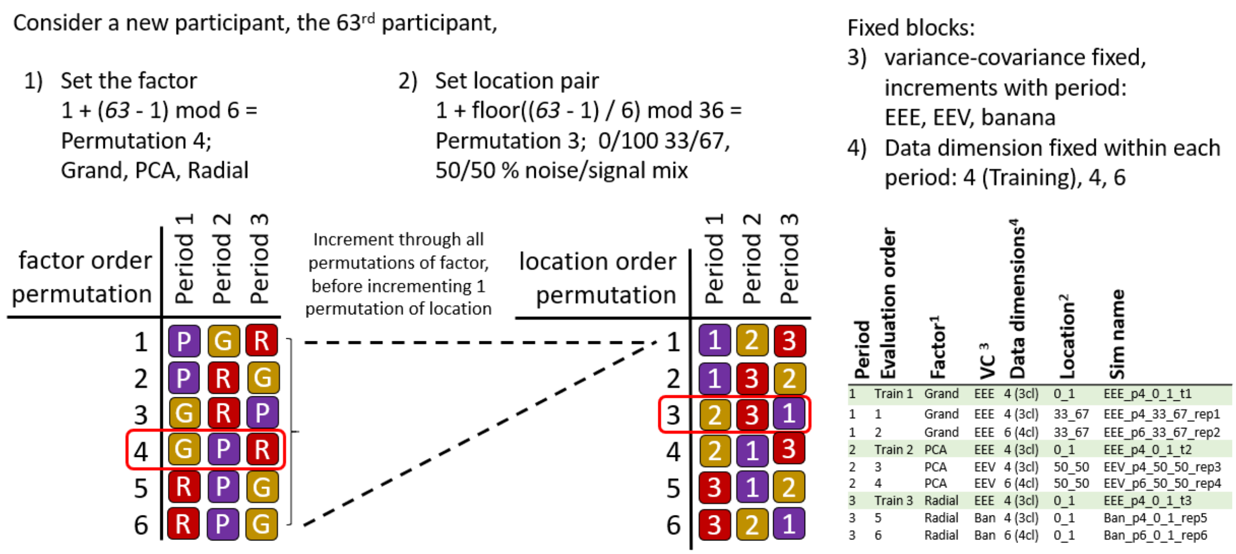

Figure 4: Example permuptaion selection. TODO XXX NEEDS CAPTION

With this setup we test the parameter space $p \in (4,\ 6)$, $shape \in (EEE,\ EEV,\ EVV{-}banana)$, $location\ \in\ (100/0\%,\ 66/33\%$ in order to evaluate the graphic display across the $factors\ \in\ (PCA,\ grand,\ radial)$. As we iterate though the possible permutations of these factors (6) and location (6) we perform an even evaluation of the full parameter space every 36 participants. While piloting these parameters we estimate that 3 even evaluations will be more than sufficient identify difference between the factors; we targeted for $n = 108$ participants for the study.

TODO: XXX CONTINUE WRITING HERE

Post study survey

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:

Demographic:

- What are your preferred pronouns? [decline to answer, he/him, she/her, they/them or other]
- 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]
- 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]

Within participant bias:
Likert scale [1-5], least agreement to most agreement.

- I understand the how to perform the task.
- I am experienced with data visualization.
- I am educated in multivariate statistical analysis.

Subjective by factor:

- I was already familiar with visualization.
- I found this visualization easy to use.
- I felt confident in my answers with this visualization.
- 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?

Results

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

Random effects regression model

In order to compare the within-participant effect we fit random effect regression model predicting the marks of non-training evaluations. We'll include most of the continuous variables, but hold out 2 that are most correlated which actually lower model AIC (Akaike's Information Criterion) when included. We also interact the number of input interactions with factor as there are no interaction for the grand tour and fundamentally different between pca and radial tours. We include a vector of coefficient against the identity function/dummy variable of our 4 factor variables (factor, dimensions, variance-covariance shape, and noise location). Lastly we include the a random effects term to account for the individual participants effect on the score. Our applied random effects model is:

$$\begin{aligned}\widehat{marks} = \ & \beta_{int} + \beta_{SecResp} * SecResp + \beta_{SecPg} * SecPg + \beta_{RespInter} * RespInter + \\ & \beta_{CntResp} * CntResp + \boldsymbol{\beta}_{InputInter*Factor} * InputInter * I(\boldsymbol{Factor}) + \boldsymbol{\beta}_{factor} * I(\boldsymbol{factor}) + \\ & \boldsymbol{\beta}_{dim} * I(\boldsymbol{dim}) + \boldsymbol{\beta}_{shape} * I(\boldsymbol{shape}) + \boldsymbol{\beta}_{location} * I(\boldsymbol{location}) + \boldsymbol{\beta}_{participant} * \boldsymbol{effect}_{participant} + \epsilon\end{aligned}$$

where,

$$\begin{aligned}& \epsilon \sim \mathcal{N}(0,\ \sigma) \\ & effect_{participant_i} \sim \mathcal{N}(0,\ \sigma_{participant_i}) \mid i \in level(participant\ ids) \\ & factor \in (pca,\ grand,\ radial) \\ & dim \in (4,\ 6) \text{ variables, with 3 \& 4 clusters respectively} \\ & shape \in (EEE,\ EEV,\ EVV\ banana) \\ & location \in (0/100,\ 33/66,\ 50/50) \text{ \% mix of a noise and signal variable respectively} \\ & I() \text{ is the indicator function, a logical value for each level of the fixed factor}\end{aligned}$$

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.

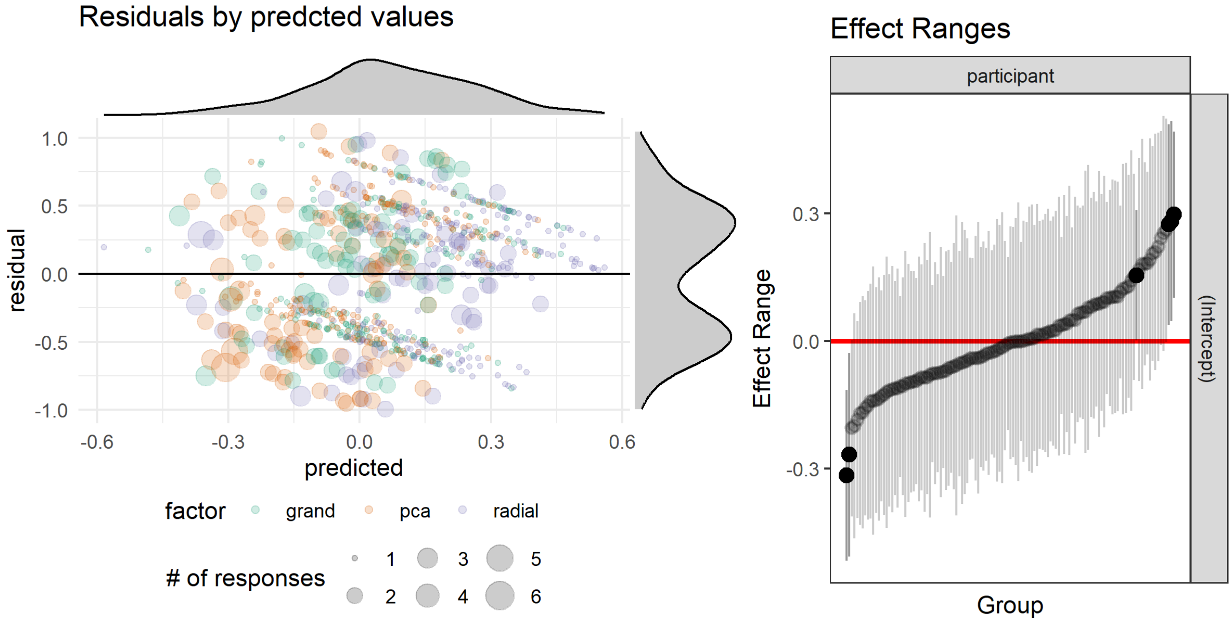

Figure 5: Model residuals and Effect ranges. TODO XXX EDIT CAPTION.

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

Supplemental material

Factor parameterizations

differences in the application and parameterization applied in the user study. Due to physically distancing from COVID-19 what was originally intended to be run in person with study invigilator had to be simplified to be understood and usable in a crowdsourcing application. We opted for precomputed images and animations in order to simplify input interactions and improve user experience.

Display of the same component on both axes simultaneously was prohibited. This results in 12 combinations of valid inputs. Half of which are homomorphic visuals, in that they mirrored on the x = y line and show no new information.

Display of the same component on both axes simultaneously was prohibited. This results in 12 combinations of valid inputs. Half of which are homomorphic visuals, in that they mirrored on the x = y line and show no new information.

Radial tours were also displayed at 5 frames per second with in interpolation step size of 0.1 radians. Users where able to swap between the 4 or 6 variables, upon which the display would change the the start of radially increasing the contribution of the selected variable till it was full, zeroed and then back to the initial. The complete animation of any 1 variable takes about 20 seconds, and is almost fully in the projection frame at around 6 second. The starting basis of each is initialized to a half-clock design, where the 4 or 6 variables were evenly distributed in half of the circle which is then orthonormalized. This is done to give no variable preference while minimizing variable interactions, as variables opposite of the manipulation variable must lose contribution as the other is rotated to full contribution (and vice versa).

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/.

Scrucca, Luca, Michael Fop, T. Brendan Murphy, and Adrian E. Raftery. 2016. "Mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models." *The R Journal* 8 (1): 289–317. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5096736/.

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.