The effect of interaction on understanding variable contributions on linear projections

Nick, Di, Kim

Abstract

Principal Component Analysis (PCA) and related eigenvalue techniques is the traditional standard for viewing prosections of multivariate spaces. However, the full story of the data is rarely portrayed accurately in few prosections. More recently, `grand tours` offer animations of prosectional random walks offering many angles to view embedded spaces. A `manual tour` provides a means of controlling the contribution of individual variables to a projected subspace. We have developed an appliation to facilitate the exploration of multivariate data though the use of various tour methods. To explore the efficacy of this tool we performed a comparative user study. Participants in our study performed several high-level analysis tasks across the three factors and provide subjective ratings. Accuracy, speed, and qualitative feedback is used to compare and rate analysts' ability to understand the importance of individual variables' contribution to distinguishing clustering with the data. User feedback suggests that…

Introduction

Hypothesis

Does the finer control afforded by the manual tour improve the ability of the analyst to understand the importance of variables contributing to the structure?

Experimental design

Data simulations

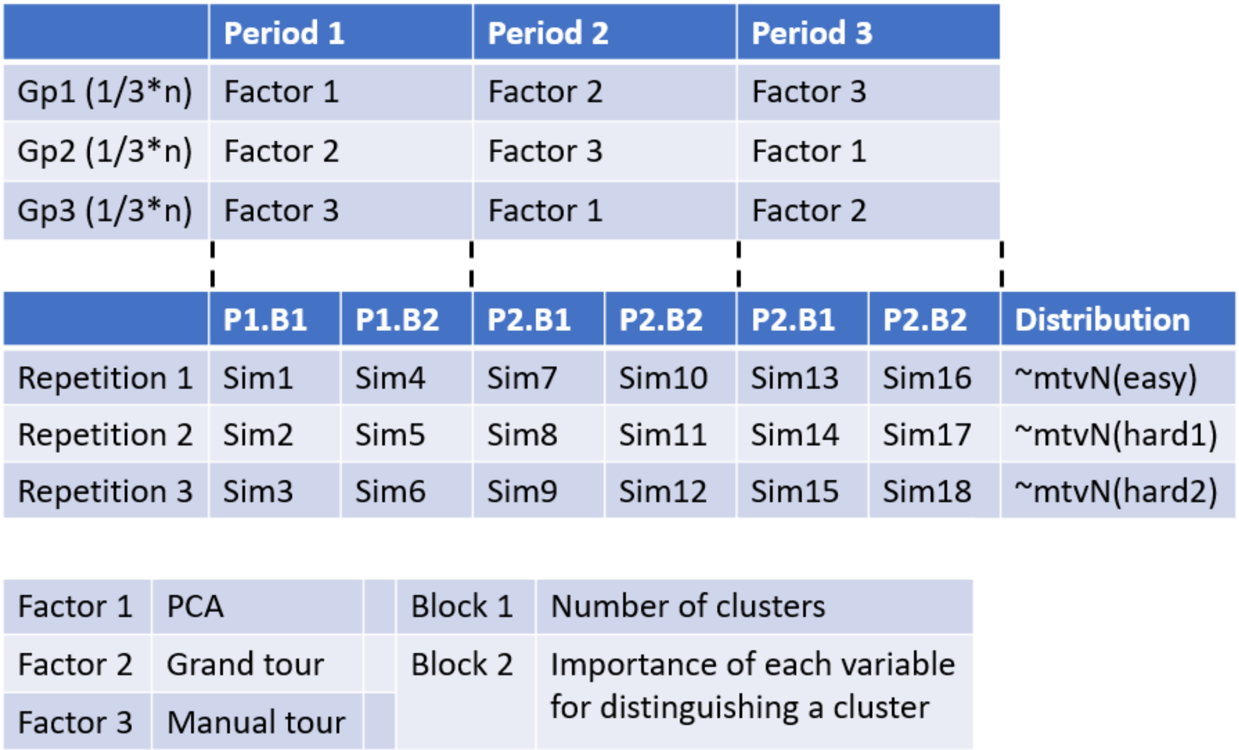
The data used for the demonstration of each block was **TODO** flea data, containing **TODO**. The remaining **3** data sets were selected from generated simulations with varing parmeters. Each set has **8 to 10** variables with **4 to 5** noise variables. Noise variables were distributed $\mathcal{N}(0,\sigma^2)$, while non-noise variables were distributed $\mathcal{N}(\mu,\sigma^2) \mid \mu \in \{-3,-2,...3\}$. In the variance-covariance matrix, non-diagonal elements of the were near -.1 to .7. Each simulation contained either 3 or 4 clusters, where each cluster recieved between 30 to 150 observations.

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 mulivariate data visualizations.

Participant groups

Each participant was randomly split into one of three even factor groups. The first group was given a biplot – a scatterplot matrix coupled with a variable mapping back to original variable space. Users were allowed to freely choise which two components to view initialized to PC1 and PC2. The second group was given the same animation, the first 30 seconds of random walk (typically spanning 6 or 7 bases interpolated into 90



	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.B1	P1.B2	P2.B1	P2.B2	P2.B1	P2.B2	Distribution
Repetition 1	Sim1	Sim4	Sim7	Sim10	Sim13	Sim16	~mtvN(easy)
Repetition 2	Sim2	Sim5	Sim8	Sim11	Sim14	Sim17	~mtvN(hard1)
Repetition 3	Sim3	Sim6	Sim9	Sim12	Sim15	Sim18	~mtvN(hard2)

Factor 1	PCA
Factor 2	Grand tour
Factor 3	Manual tour

Block 1	Number of clusters
Block 2	Importance of each variable for distinguishing a cluster

Figure 1: Experimental design

frame viewed at 3 frames per second) of a grand tour with the ability to freely control the location and speed of the animation. The third group was provided with the ability to control the magnitude of an individual variable contributes to the projection with a manual tour. Doing so performs a constrained rotation on the data object resulting in a change of the other variables to preserve orthogonality between dimensions. Participants could freely change which dimension to manipulate.

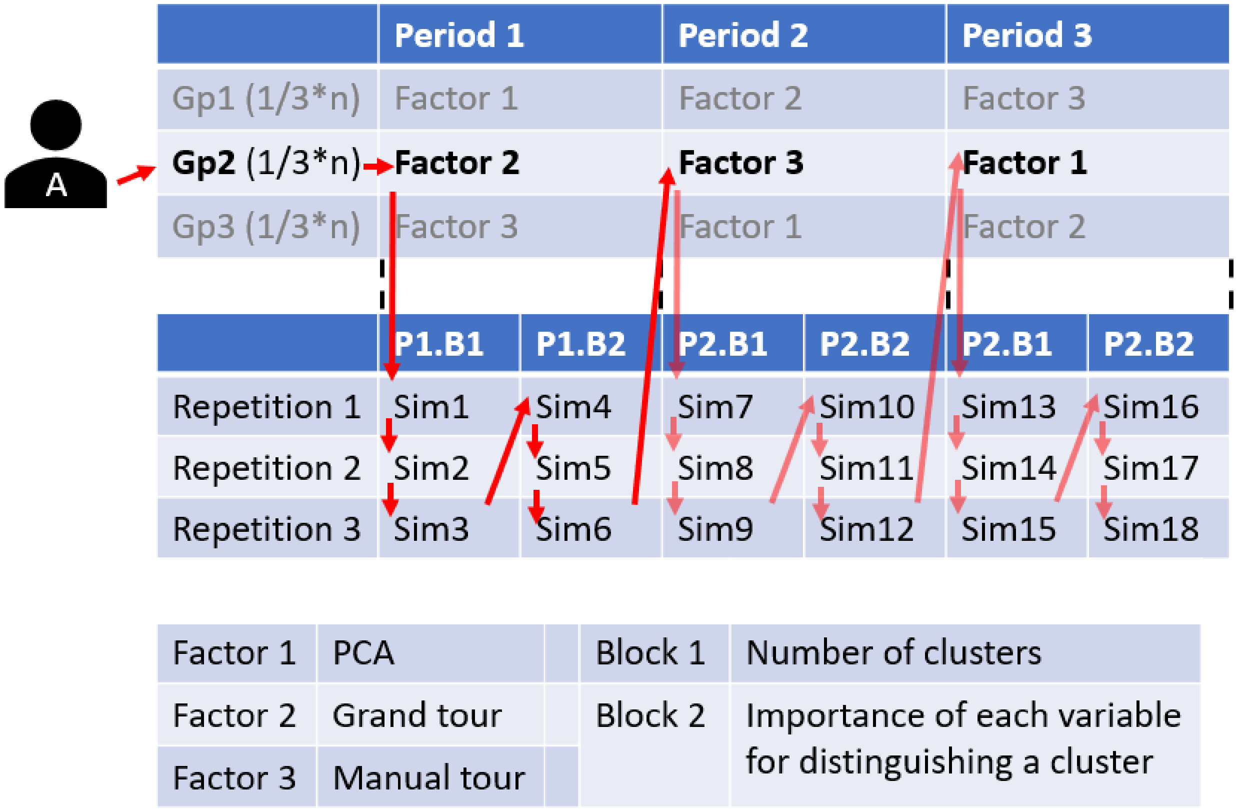
Factors

We explored performance across three factors. The first factor is eigenvalue decomposition in the form PCA. The second factor is an animated random walk between target bases via grand tour. The third factor allows for the manual control of individual variable's contribution to the projection, for performing a manual tour. Users are able to select which principal components are on the x and y-axes. Percent of full sample variation is displayed along the corrisponding axis.

Block treatments

Each participant performed both of the 2 block treatments in fixed order. One block asked participants to identify the number of clusters present in the data. In this block, clusters were unsupervised, where all observations appeared as black circles. This block also served as a control for assessing the general aptitude for this sort of high dimensional analysis as it was simpler. A second block asked participants to identify the top three variables for distinguishing clusters in order of importance. A third block asked for highly correlated variables if any. In the latter two blocks, clusters were supervised; observations were colored (with a color-blind friendly palette) and assigned a shape based on their cluster.

It was expected that the grand tour should excell in identifying the number of clusters. This is because it



	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.B1	**P1.B2**	**P2.B1**	**P2.B2**	**P2.B1**	**P2.B2**
Repetition 1	Sim1	Sim4	Sim7	Sim10	Sim13	Sim16
Repetition 2	Sim2	Sim5	Sim8	Sim11	Sim14	Sim17
Repetition 3	Sim3	Sim6	Sim9	Sim12	Sim15	Sim18

Factor 1	PCA		Block 1	Number of clusters
Factor 2	Grand tour		Block 2	Importance of each variable for distinguishing a cluster
Factor 3	Manual tour			

Figure 2: Person A

offers many and more widely varying bases are viewed in quick sucsession for a more cohesive parralax-like movement making clusters relatively easy to identify.

Randomization & replication

Participants were randomly assigned to one of **three** factors deciding which visual method they received. The blocks were performed in a random order for each participant. Within each block, participants performed **four** replications, answering the block question for each of the **four** datasets in a random order before proceeding to the next block.

Response & measures

Each block was introduced and demonstrated directly preceding each block. During this introductory segment, each participant was given a written description of the block task and instructions on how the factor visualization informed the answer, as illustrated with the same toy data set. Participants were free to ask questions and clarification from the proctor at this time. Questions were not allowed outside of the introductory segments. Participants received exactly **two** minutes to explore each repetition's projection before responding to the given task. Responses came in the form of single integer input for the block asking to identify the number of clusters. The second block collected the top 3 ordered variables that distinguish clusters. The remaining block collected `p` (number of variables in the data) inputs grouped into zero to four groups.

After responses for each block were collected, participants were given a short survey of demographic, related exiperance, and subjective evaluation of each factor on a 7-point Likert scale. These questions covered familiarity and expertise with multivariate data, its visualization, as well as, ease of use, understandability, confidence, and likelihood to recommend the participant's factor visualization.

Post-study survey

- gender [decline, F, M, Intergender/other]
- age [decline, 19 or younger, 20 to 29, 30 to 39, 40 or older]
- completed education [decline, highschool, undergraduate, honors/masters/mba, doctorate]
- experience with data vizualization [likert 1-7]
- educated in multivarate statistical analysis [likert 1-7]
- previous familiar with vizualization [likert 1-7] x3 factors
- ease [likert 1-7] x3 factors
- confidence [likert 1-7] x3 factors
- likeability [likert 1-7] x3 factors

Experimental results

Accompanying tool: spinifex application

To accomany this study we have produced a more general use tool to perform such exploratory analysis of high dimensional data. The `spinifex`{(**???**)} R package (version 0.2.0 and up) contains a free, open source version of a `shiny` {(**???**)} application. The application features traditional static visualizations including PCA, with biplots and screeplots, and scatterplot matrices. the application also implements various tours, accomidating manual tours, projection persuit, and limited versions of grand, little, and local tours. Data can be imported in .csv and .rda format, and projections can be saved as .png, .gif, and .csv formats where applicable. Run the following R code for help getting started.

```r
install.packages("spinifex")
spinifex::run_app("intro")
spinifex::run_app("primary")
```

Discussion

Acknowledgments

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

Bibliography

R Core Team. 2019. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

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.