The effect of interaction on understanding variable contributions on linear projections

Nick, Di, Kim

Abstract

Principle 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 portaied accurately in few prosections. More recently, `grand tours` offer animations of prosectional random walks offering many angles to view embedded spaces. A `manual tour` provide a means of controling 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 a number of high-level analysis tasks across the three factors and provide subjective rankings. Accuracy, speed, and qualitative feedback is used to compare and rank analysts ability to understand the imporance of indivual variables contribution to dinguishing 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

Participant population

A sample of convenience was taken from postgraduate students in the department of econometrics and business statistics at Monash University, based in Melbourne, Australia. Participants were required to have prior knowledge of Principal Component Analysis (PCA) and scatterplot matricies (SPLOM).

Factors

We explored performance across three factors: eigenvalue decomposition from PCA, animated random walk between prosections via grand tour, and controling individual variable's contribution to the projection by manual tour.

Each participant was randomly split into one of three even factor groups. The first group was given a biplot, or scatterplot matrix coupled with a variable maping 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 of a **n** basis random walk (about 30-second loop 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 that an indivdual variable contributs 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. Participants could freely change which dimension to manipulate.

Block treatments

TODO: continue editing here Each participant performed each of 4 block treatments in random order. The blocks consisted of determining the dimensionality of the dataset, `p`, the number of the clusters, `n`, the number of important variables, `d`, and identification of variable with significant covariance, `s`.

Randomization & replication

Participants were randomly assigned to one of 3 factors deciding which visual method they received. The blocks were performed in a random order for each participant. Within each block, participants performed 4 replications, answering the block question for each of the 4 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 shown the visual for their factor with a written description of the block and how to discern it 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 2 minutes to study/explore each repetition's projection before answering a question regarding it. Answers came in the form of a numeric input for three blocks - namely, dimensionality, clusters, and important variables (`p`, `n`, and `d` respectively). For the remaining block, covariance `s`, a checkmark box was provided for each variable. Participants we instructed to mark all variables, if any, that were highly correlated. None of the data sets contained more than a single group of highly-correlated variables.

After responses for each block were collected, participants were given a short survey of 7 subjective questions on a 9-point Likert scale. These questions covered familiarity and expertise with multivariate data, its visualization, as well as, ease of use, understandability, confidence, and likely hood to recommend their factors visualization.

Experimental results

Accompanying tool: spinifex application

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.