Individual Literature Review

Topic: Principal Component Analysis

Brandy Carr (bjc57)

Fall 2022

1. A comparative dimensionality reduction study in telecom customer segmentation using deep learning and PCA

This article compares PCA with a deep learning autoencoder on telecom customer data. Goal is to segment customers based on 220 features to optimize customer satisfaction/loyalty. Raw data must be cleaned & standardized (z-score normalization) before applying PCA. A scree plot was used to visualize the number of features to keep, which was 3 that explain ~72% of data. Using the absolute values of the eigenvectors they can decide wich features contribute the most to the first 3 PC's. Overall PCA saved ~90% of total variance with just 20 features, reducing original dataset by 200. (Alkhayrat, Aljnidi, and Aljoumaa 2020)

2. Visualizing Psychological Networks: A Tutorial in R

Article aims to guide researchers on choosing the best/most interpretable visualizations of psychological networks. These networks include mental disorder symptoms (specifically OCD and depression) and the connections/correlations between them. When plotting, the symptoms are referred to as nodes and their connections as edges. The article goes on to compare 4 different plotting approaches; force-directed algorithms, multidimensional scaling, PCA, and eigenmodels. 6 different benefits were used as a checklist for all 4 approches. Benefits include: node placement is meaningful, useful for comparing replications, distances between nodes is interpretable, X/Y placement of nodes is interpretable, can be based on any network, central nodes in the center. PCA checks 3 of the 6 benefits, node placement is meaningful, useful for comparing replications, and X/Y placement of nodes is interpretable (this being the primary benefit since nodes/symptoms can be compared right to left, X, or top from bottom Y). 1 major disadvantage of PCA is that it relies on the correlation matrix and nodes/symptoms can be difficult to see when they score similarly on both PC's. (Jones, Mair, and McNally 2018)

3. SVM and PCA Based Learning Feature Classification Approaches for E-Learning System

This article aims to classify student learning attributes using PCA in order to develop a simple methodology to optimize a students dynamic learning sequence based on their individual skill, needs and preferences, and also to maximize their learning outcome in computer programming courses (java, C). The study uses 8 different learning attributes; anxiety, personality, learning style, cognitive style, grades from prev. semester, motivation, study level, and student prior knowledge. 100 students taking C programming course were used in the present study. Each student filled out a questionairre to gather information on the 8 learning attributes. They each took a 20 question midterm and also a final that scored their capabilities in 3 areas; syntax, logical, and application (each score was divided into low, med, and high categories). For this study, 3 PC's were kept & used for classification purposes and explain ~77% of the variance. From figure 5 it looks like the first PC is comprised of 3 main learning attributes prior knowledge, learning style, and personality. Second PC groups motivation, cognitive style, and grades from prev. semester. Third PC groups anxiety with study level. The article then goes on to use the 3 PC's to fit 4 classification models, a neural network, quadratic

SVM, gaussian SVM, and linear SVM. Linear SVM provided the highest accuracy, sensitivity, and specificity, outperforming the other kernal classifiers. (Khamparia and Pandey 2018)

4. Applied Multivariate Statistical Analysis and Related Topics with R

This chapter goes over PCA basic's and how and why it is useful. The overall goal of PCA is to minimize the number of predictive variables while maintaining most of the variation (70%~80% as a guideline, the book goes on to cover a few more rules of thumb for this). PCA is a good method for finding outliers since the principal components are linear combinations of the original data - plotting with lower dimensions makes them easier to spot. PCA is also useful when some of the predictor variables in a regression model are highly correlated, which if not addressed can lead to poor parameter estimates. The first PC explains the most variability in the data & each succeeding PC explains the most possible remaining variability. PCA analysis should be preformed on scaled/unit data, magnitudes must be comparable (use the correlation matrix instead of covariance OR standardize the data). (Lang and Qiu 2021)

5. The fixed effects PCA model in a common principal component environment

The article compares a fixed effects PCA model to the 2 most common approaches, descriptive algebraic and probabilistic. All three produce the same results using spectral decomposition of the sample covariance matrix but, the interpretations will differ depending on the assumptions. Graphing the low dimensional PC's (usually the first 2) is a common way to identify hidden patterns in the data such as outliers or clusters. The fixed effects model only makes assumptions about the dimensionality of the data, and incorporates knowledge about noise in the data to improve estimates. The results of the paper were that the fixed effects model incorporating CPCA (common PCA) out preformed all others. (Duras 2022)

References

Alkhayrat, Maha, Mohamad Aljnidi, and Kadan Aljoumaa. 2020. "A Comparative Dimensionality Reduction Study in Telecom Customer Segmentation Using Deep Learning and PCA." *Journal of Big Data* 7 (February): 9. https://doi.org/10.1186/s40537-020-0286-0.

Duras, Toni. 2022. "The Fixed Effects PCA Model in a Common Principal Component Environment." *Communications in Statistics-Theory and Methods* 51 (6): 1653–73.

Jones, Payton, Patrick Mair, and Richard McNally. 2018. "Visualizing Psychological Networks: A Tutorial in r." *Frontiers in Psychology* 9 (September). https://doi.org/10.3389/fpsyg.2018.01742.

Khamparia, Aditya, and Babita Pandey. 2018. "SVM and PCA Based Learning Feature Classification Approaches for e-Learning System." *International Journal of Web-Based Learning and Teaching Technologies* 13 (April): 32–45. https://doi.org/10.4018/IJWLTT.2018040103.

Lang, WU, and Jin Qiu. 2021. *Applied Multivariate Statistical Analysis and Related Topics with r.* edp sciences.