A DATA VISUALIZATION FOR PCA ANALYSIS / DATAVIZ FINAL PROJECT

Roberto Stelling *
PPGI
Universidade Federal do Rio de Janeiro
Rio de Janeiro, RJ 21941-916, Brazil
roberto@stelling.cc

ABSTRACT

Traditionally PCA Analysis for features reduction in images is performed without a systematic visual aid approach. This paper describes the implementation of a PCA Analysis tool in JavaScript that uses data visualization techniques aiming to improve the analysis process.

1 PROBLEM DESCRIPTION

There are many reasons to reduce a data set, some modern examples are: noise reduction, outlier removal, lossy image compression or even as a preliminary step in various types of data exploration and data analysis.

Principal component analysis (PCA) is well suited as a lossy image compression solution, as you can apply PCA on a set of points of a data set and reduce its dimensionality with a certain, desirably controllable, loss of precision. Of course, one want to loose as little precision as possible while compressing as much as possible. There is a clear trade off between compression and precision. With images, the main difficulty resides in this very trade off: how many dimensions can the algorithm throw away and still retain the desired level of image quality or sharpness? Is there a general rule where you can certainly decide how many dimensions will be cut off the original data set? Of course, the type and amount of data available for compression, the data set, has a big influence on the final point where the cut will be, but can the analyst decide simply on the number of dimensions or amount of variance that will be thrown away and be sure that the results will be satisfactory? We propose that using a visual helping tool during the decision process can have a positive impact on the cut off selection.

2 BRIEF INTRODUCTION TO PCA

2.1 OBJECTIVE

According to Jolliffe (1986), the central idea of principal component analysis (PCA) is to reduce the dimensionality of a data set in which there are a large number of interrelated variables, while retaining as much as possible of the variation present in the data set.

This reduction is achieved by transforming the data set to a new set of variables, the principal components, which are not correlated, and which are ordered so that the first few retain most of the variation present in all the original variables.

Principal component analysis was first described by Pearson (1901) and later developed independently by Hotelling (1933) (Jolliffe, 1986).

2.2 INTUITION

PCA can be thought of as the problem of fitting an n-dimensional ellipsoid to the data, where each axis of the ellipsoid represents a principal component. The larger the axis of a component, the larger

*PPGI/UFRJ Graduate Student, http://stelling.cc.

the variance for that component. To find the axis of the ellipsoid we must first centralize, or normalize, the data set, compute the covariance matrix of the data and finally compute the eigenvalues and eigenvectors of the covariance matrix.

To achieve that we compute the singular value decomposition of the covariance matrix.

2.3 Singular Value Decomposition

If C is the covariance matrix of our data set then:

$$C = USV^T$$

where S is a diagonal matrix with positivel real entries and U and V columns are orthonormal[1].

Suppose that we have m data points of n dimensions (EG: 2000 images with 1024 pixels, m = 2000, n = 1024) then

$$D_{m\times n} \rightarrow C_{n\times n} = cov(D)$$

For the singular value decomposition of the covariance matrix we have:

$$C_{n\times n} = U_{n\times n} S_{n\times n} V^T_{n\times n}$$

where S is a diagonal matrix.

3 How to select the number of components to retain

The problem of selecting how many components to retain is not new, Zwick & Velicer (1986), present the results of a Monte Carlo evaluation of five methods that have been proposed for determining how many factors or components to retain: Horn's parallel analysis, Velicer's minimum average partial [MAP], Cattell's scree test, Bartlett's chi-square test, and Kaiser's eigenvalue greater than 1.0 rule. The determination of the number of components or factors to retain is likely to be the most important decision a researcher will make (Zwick & Velicer, 1986).

4 Citations, figures, tables, references

These instructions apply to everyone, regardless of the formatter being used.

4.1 Citations within the text

Citations within the text should be based on the `natbib` package and include the authors' last names and year (with the "et al." construct for more than two authors). When the authors or the publication are included in the sentence, the citation should not be in parenthesis (as in "See Goodfellow et al. (2016) for more information."). Otherwise, the citation should be in parenthesis (as in "Deep learning shows promise to make progress towards AI (Goodfellow et al., 2016).").

The corresponding references are to be listed in alphabetical order of authors, in the References section. As to the format of the references themselves, any style is acceptable as long as it is used consistently.

4.2 Footnotes

Indicate footnotes with a number[2] in the text. Place the footnotes at the bottom of the page on which they appear. Precede the footnote with a horizontal rule of 2 inches (12 picas).[3]

[1]both orthogonal and normalized
[2]Sample of the first footnote
[3]Sample of the second footnote

Table 1: Sample table title

PART	**DESCRIPTION**
Dendrite	Input terminal
Axon	Output terminal
Soma	Cell body (contains cell nucleus)

4.3 Figures

All artwork must be neat, clean, and legible. Lines should be dark enough for purposes of reproduction; art work should not be hand-drawn. The figure number and caption always appear after the figure. Place one line space before the figure caption, and one line space after the figure. The figure caption is lower case (except for first word and proper nouns); figures are numbered consecutively.

Make sure the figure caption does not get separated from the figure. Leave sufficient space to avoid splitting the figure and figure caption.

You may use color figures. However, it is best for the figure captions and the paper body to make sense if the paper is printed either in black/white or in color.

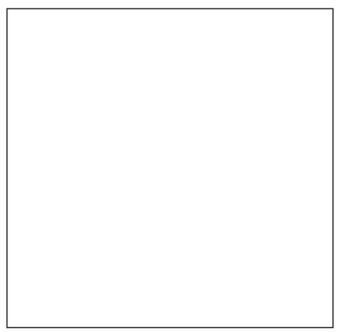

Figure 1: Sample figure caption.

4.4 Tables

All tables must be centered, neat, clean and legible. Do not use hand-drawn tables. The table number and title always appear before the table. See Table 1.

Place one line space before the table title, one line space after the table title, and one line space after the table. The table title must be lower case (except for first word and proper nouns); tables are numbered consecutively.

Acknowledgments

Use unnumbered third level headings for the acknowledgments. All acknowledgments, including those to funding agencies, go at the end of the paper.

References

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*, volume 1. MIT press Cambridge, 2016.

Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933.

Ian T Jolliffe. Principal component analysis and factor analysis. In *Principal component analysis*, pp. 115–128. Springer, 1986.

Karl Pearson. Principal components analysis. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 6(2):559, 1901.

William R Zwick and Wayne F Velicer. Comparison of five rules for determining the number of components to retain. *Psychological bulletin*, 99(3):432, 1986.