Write up - Principal component analysis

C. G. Holm (kxm508) and E. H. C. Henningsen (tzs820)
Niels Bohr Institute, University of Copenhagen
(Dated: 08-03-2023)

I. INTRODUCTION

In this write-up, the tutorial review "Principal component analysis" by Bro and Smilde is summarized, and the used statistical method is explained. The review revolves around an analysis of 44 bottles of red wine, measuring 14 different chemical properties and components for each of the bottles. Data in 14-dimensional feature space with 44 samples. The approach to these data can be of varying nature, e.g. plots of each feature vs. sample, as seen in figure 1.

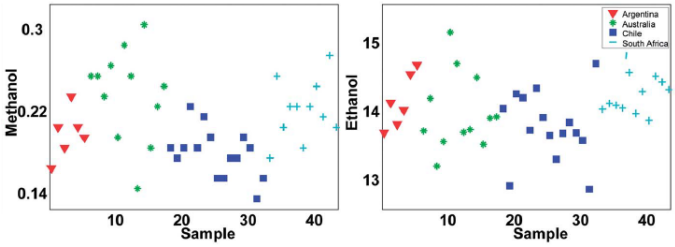

FIG. 1: Two plots of methanol and ethanol content in each of the sample wine bottles. The bottles are furthermore classified in the region of origin.

This is a univariate approach. The down-side is that important correlations between the features are prone to be lost. The question is, how is interpretable information extracted from a 14-dimensional feature space? The answer is Principal Component Analysis (PCA): Project the data from N to D dimensions preserving the most information.

II. REVIEW

To ensure that the desired outcomes are achieved when performing PCA, several steps must be taken during the actual implementation of the method.
As mentioned earlier the goal of the method is to reduce dimensionality. This is done by a change of basis i.e. rotating the data in accordance to a maximization of variance. This maximization is generalized into an eigenvalue problem where the eigenvectors make out for the new basis, which will be covered later. The number of eigenvectors equals the number of parameters in the data set, but choosing the right amount for the particular set, necessitates attentive consideration depending on the data and choice of visualisation.

Before all this some data processing is needed. It is stated in the article that when comparing data from different parameters the characteristic sizes i.e. means and variance may differ drastically.[1] The term *autoscaling* is introduced to eliminate bias towards any single parameter, and thereby promote a fair comparison that maximizes variance. Autoscaling is simply subtracting the mean from all measurements in regards to a certain parameter such that the mean over that given parameter is 0, Eq. (1). Furthermore the data is divided by the standard deviation, ensuring the new basis won't be affected by a parameter with a large variance.

$$\bar{\mathbf{y}} = \frac{1}{N}\sum_{n=1}^{N} y_n = 0 \tag{1}$$

Dividing the parameter values with the standard deviation is not always necessary and depends on the original data set. In literature, the division by σ is sometimes omitted.[2]
Subsequently, weights are introduced to weigh the data and have the length $||w|| = 1$. w will contain the eigenvalues determining in which direction the largest variances are obtained and will generally be the bridge from the autoscaled data to the new $x_n = w^T y_n$. From Eq. (2) and (3) it shows how the problem ends up being an eigenvalue problem involving the covariance matrix C. Several parts of the derivation are omitted and can be found in literature.[2]

$$\sigma_x^2 = w^T \left(\frac{1}{N}\sum_{n=1}^{N} y_n y_n^T \right) w = w^T C w \tag{2}$$

$$\Rightarrow \sigma_x^2 w = C w \tag{3}$$

After solving this eigenvalue problem the first vector e.i. the vector with the largest eigenvalue corresponds to the component of the new basis with the highest variance.

Practical aspects

As described in detail in the paper by Bro and Smilde, several considerations arise when performing PCA. The first as mentioned earlier is the number of components or length of the new basis. The article describes a lot of ways of doing this where some are illustrated in Fig. 3. Some of the ways include *Kaiser's rule*, which simplifies choosing all the components with eigenvalues above one.

Because of the autoscaling and the presumed orthogonality between variables, each should have a variance of 1. An eigenvalue greater than one should then point to the component describing variance for more than one feature. Another method, also shown in Fig. 2, is *the broken stick rule*. Using this rule, it is considered, how the eigenvalues would be placed if the data was truly random, and comparing to the eigenvalues greater than these.[1] In case of the wine samples, the first rule results in around 5 components and the latter results in 3. After the choice

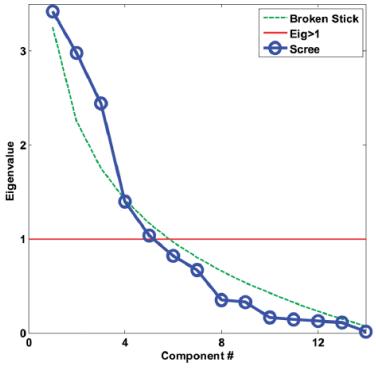

FIG. 2: Illustration of choosing component for the PCA analysis.[1]

of numbers of sub-dimensions, the next question is "how much information was lost in the compression process?"

$$X = (t \cdot p^T) + E \tag{4}$$

Eq. (4) contains the original data matrix X, the scores vector i.e. the new weighted vector t, the loading vector p and the residuals E. The loading vector is simply the transformation required to go from the new basis to the old. Lastly, E is the same shape as the original data, measuring how much information is lost in the process.[1]

Results

After the principal components of the wine data, and the number of components to include have been determined, the samples are plotted along the respective axes in figure 3:

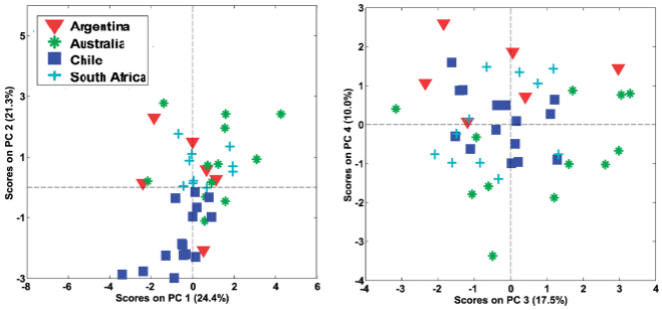

FIG. 3: The 44 samples plotted along the 4 principal components. The legend shows the classification of the samples depending on the region of origin.

It is e.g. seen that wines from Chile score exclusively negative values along principal component 2, and wines from Argentina score exclusively positive values along principal component 4. Studying the eigenvectors of eq. 3 reveals a correlation regarding the wine's region of origin and some of the chemical structure of the wine (within the 14 chosen chemical components/measurements).

III. CONCLUSION

In the review, PCA is motivated and thoroughly explained. Relevant considerations regarding preparation of data, such as autoscaling, are made, and finally the wine samples are analyzed with respect to the principal components. It is concluded that projecting data along the 3-4 greatest principal components axes is beneficial, since the eigenvalues along these component directions lie above the "broken-stick" distribution of eigenvalues of random 14-dimensional data. Projecting the samples along the 4 greatest principal component axes, a scatter plot shows the grouping of the different wine bottles in region of origin and their trends along the selected components.

[1] R. Bro and A. K. Smilde, Principal component analysis, Royal Society of Chemistry **6**, 20 (20214).
[2] S. Rogers and M. Girolami, *A FIRST COURSE IN MACHINE LEARNING* (CRC Press, 2017) pp. 235–243.