A BRIEF GUIDE TO PRINCIPAL COMPONENT ANALYSIS AS IMPLEMENTED IN OPENPCA

MATTEO TOMMASINI

The Principal Component Analysis (PCA) was introduced in 1933 by Harold Hotelling [1] in the context of psicometric data analysis. Since its birth, PCA has been widely applied to many fields where multivariate datasets have to be dealt with. The easier approach to introduce PCA, by also taking into consideration its numerical implementation in Matlab, is through a fully algebraic approach that focuses on the variance-covariance matrix of the dataset and its spectral decomposition. Let us introduce first the multivariate dataset matrix $\boldsymbol{X}_{ov}$, which along each row stores the results of one multivariate observation along a given number of variables (N_v). The adopted notation for the dataset matrix highlights the different role of row *vs.* column indexes. The different observations are identified in the $\boldsymbol{X}_{ov}$ matrix by the row index (o), whereas the different variables of each multivariate measurement (observation) are identified by the column index (v). In the context of spectroscopy, each row represents one spectrum, and the different variables are the wavenumbers at which the instrument has recorded a given spectral intensity (*e.g.*, Raman intensity, or absorbance). Hence, because of the adopted notation, we have the following identities:

$$\begin{aligned} \boldsymbol{X} &= \boldsymbol{X}_{ov} \\ \boldsymbol{X}_{vo} = (\boldsymbol{X}_{ov})^t &= \boldsymbol{X}^t, \end{aligned} \tag{1}$$

where ^{t} indicates matrix transposition. As described later, the variance-covariance matrix among the variables of the dataset can be straightforwardly introduced through the matrix of the centered dataset, $\boldsymbol{\chi}_{ov}$:

$$\boldsymbol{\chi}_{ov} = \boldsymbol{X}_{ov} - \boldsymbol{X}_{\langle o \rangle v}, \tag{2}$$

where $\boldsymbol{X}_{\langle o \rangle v}$ represents the row vector of the average values of the variables over the number of N_o observations, and its elements are given by:

$$X_{\langle o \rangle v} = \frac{1}{N_o} \sum_o X_{ov} \tag{3}$$

We adopt in Eq. (2) the same abuse of notation used in Matlab: by subtracting a row vector to a matrix actually one subtracts the given row vector to each row of the matrix. Hence Eq. (2) is implemented in Matlab as simply as chi = X - mean(X), because the Matlab function mean(X) gives the row vector corresponding to the average of all the rows of the X matrix – which effectively corresponds to averaging out with respect to the available observations (see above). The variance-covariance matrix among the variables

Date: April 11, 2022.

of the dataset ($\mathbf{\Sigma}_{vv}$) is defined as follows:

$$\Sigma_{v_1 v_2} = \frac{1}{N_o - 1} \sum_o \left(X_{ov_1} - X_{\langle o \rangle v_1} \right) \left(X_{ov_2} - X_{\langle o \rangle v_2} \right) = \\ = \frac{1}{N_o - 1} \sum_o \chi_{ov_1} \chi_{ov_2} \tag{4}$$

The last equality shows that, with matrix notation, the variance covariance matrix is simply given by:

$$\mathbf{\Sigma}_{vv} = \frac{1}{N_o - 1} (\boldsymbol{\chi}_{ov})^t \boldsymbol{\chi}_{ov} = \frac{1}{N_o - 1} \boldsymbol{\chi}_{vo} \boldsymbol{\chi}_{ov}. \tag{5}$$

Clearly, by definition, $\mathbf{\Sigma}$ is a symmetric matrix, and it is positive definite[1]. Therefore it admits spectral decomposition by the orthogonal matrix of its eigenvectors, and the eigenvalues are positive quantities [2]. The matrix eigenvalue problem of the variance-covariance matrix is written as:

$$\mathbf{\Sigma}_{vv} \boldsymbol{L}_{vs} = \boldsymbol{L}_{vs} \boldsymbol{\sigma}_{ss} \tag{6}$$

In Eq. (6) $\boldsymbol{\sigma}_{ss}$ is the diagonal matrix of the eigenvalues of $\mathbf{\Sigma}_{vv}$ and $\boldsymbol{L}_{vs}$ is the orthogonal matrix of the eigenvectors of $\mathbf{\Sigma}_{vv}$. The orthogonality of $\boldsymbol{L}_{vs}$ implies:

$$\begin{aligned} \boldsymbol{L}_{vs} \boldsymbol{L}_{sv} &= \mathbf{1}_{vv} \\ \boldsymbol{L}_{sv} \boldsymbol{L}_{vs} &= \mathbf{1}_{ss}. \end{aligned} \tag{7}$$

Therefore, by left-multiplying Eq. (**??**) by $\boldsymbol{L}_{sv}$, and by considering its orthonormality, one obtains the spectral decomposition of the variance-covariance matrix:

$$\boldsymbol{L}_{sv} \mathbf{\Sigma}_{vv} \boldsymbol{L}_{vs} = \boldsymbol{\sigma}_{ss} \tag{8}$$

By sustituting in the right-hand side of Eq. (8) the definition of $\mathbf{\Sigma}_{vv} = \boldsymbol{\chi}_{vo} \boldsymbol{\chi}_{ov} / (N_o - 1)$ (cfr. Eq. **??**), one obtains:

$$\boldsymbol{\sigma}_{ss} = \frac{1}{N_o - 1} \boldsymbol{L}_{sv} \boldsymbol{\chi}_{vo} \; \boldsymbol{\chi}_{ov} \boldsymbol{L}_{vs} \tag{9}$$

Similarly to the definition of a variance-covariance matrix (Eq. (5)), it is then possible to identify in the right-hand side of Eq. (**??**) a structure given by the product of a matrix (defined $\boldsymbol{S}$) by its transpose:

$$\boldsymbol{\sigma}_{ss} = \left[\frac{1}{\sqrt{N_o - 1}} \boldsymbol{L}_{sv} \boldsymbol{\chi}_{vo} \right] \left[\boldsymbol{\chi}_{ov} \boldsymbol{L}_{vs} \frac{1}{\sqrt{N_o - 1}} \right] = \boldsymbol{S}_{so} \boldsymbol{S}_{os} = \boldsymbol{S}^t \boldsymbol{S}. \tag{10}$$

The rows of such a matrix ($\boldsymbol{S}_{os}$) – named the *scores* matrix – define the observations (o label) through the so-called principal components (s label):

$$\boldsymbol{S}_{os} \equiv \left[\frac{1}{\sqrt{N_o - 1}} \boldsymbol{\chi}_{ov} \boldsymbol{L}_{vs} \right] \tag{11}$$

[1] This is straightforward to show by using the expression $\mathbf{\Sigma} = \boldsymbol{A}^t \boldsymbol{A}$, with $\boldsymbol{A} = \boldsymbol{\chi} / \sqrt{N_o - 1}$. For $\mathbf{\Sigma}$ to be positive defined it should be $\boldsymbol{x}^t \mathbf{\Sigma} \boldsymbol{x} \geq 0$ for a generic $\boldsymbol{x}$ column vector. By considering that $\mathbf{\Sigma} = \boldsymbol{A}^t \boldsymbol{A}$, such requirements becomes $0 \leq \boldsymbol{x}^t \mathbf{\Sigma} \boldsymbol{x} = \boldsymbol{x}^t \boldsymbol{A}^t \boldsymbol{A} \boldsymbol{x} = (\boldsymbol{A}\boldsymbol{x})^t (\boldsymbol{A}\boldsymbol{x}) = |\boldsymbol{A}\boldsymbol{x}|^2$. The last step clearly proves the statement, because it represents the squared modulus of the generic vector $\boldsymbol{A}\boldsymbol{x}$ (certainly a non-negative quantity).

The matrix of the eigenvectors of the variance-covariance matrix ($\boldsymbol{L}_{vs}$), which is named the *loadings* matrix, defines the linear relationship existing between each principal component and the set of variables. The PCA scatterplot *e.g.* of the first two principal components (PC1, PC2) can be obtained by plotting on the Cartesian xy plane the first column of the $\boldsymbol{S}$ matrix (x coordinates) vs. the second column of the $\boldsymbol{S}$ matrix (y coordinates). This plot immediately allows judging data clustering or the presence of outliers. Such scatterplots can be of course extended to other principal components (*i.e.*, to other columns of the $\boldsymbol{S}$ matrix). Sometimes, the scores matrix is normalized in such a way to produce an associated variance-covariance matrix (over the s variables) that is a unit matrix. This normalization is simply done as follows:

$$\boldsymbol{S}'_{os} \equiv \boldsymbol{S}_{os}\boldsymbol{\sigma}_{ss}^{-1/2} \tag{12}$$

It is then straightforward to show that the variance-covariance matrix associated to $\boldsymbol{S}'_{os}$ is a unit matrix:

$$(\boldsymbol{S}')^t\boldsymbol{S}' = \left(\boldsymbol{\sigma}_{ss}^{-1/2}\boldsymbol{S}_{so}\right)\left(\boldsymbol{S}_{os}\boldsymbol{\sigma}_{ss}^{-1/2}\right) = \boldsymbol{\sigma}_{ss}^{-1/2}\boldsymbol{\sigma}_{ss}\boldsymbol{\sigma}_{ss}^{-1/2} = \mathbf{1}_{ss}. \tag{13}$$

References

[1] H. Hotelling. Analysis of a Complex of Statistical Variables into Principal Components. *The Journal of Educational Psychology*, 24:417–441, 1933.

[2] J.R. Schott. *Matrix Analysis for Statistics*. Wiley Series in Probability and Statistics. Wiley, 2016.