Topic 4: Principal component analysis

Introduction

Principal component analysis is applied mainly as a **variable reduction procedure**. It is usually applied in cases when data is obtained from a possibly **large number** of variables which are possibly **highly correlated**. The goal is to try to “condense” the information.

This is done by summarising the data in a (small) number of transformations of the original variables. Our motivation to do that is that we believe there is some redundancy in the presentation of the information by the original set of variables since e.g. many of these variables are measuring the same construct. In that case we try to reduce the observed variables into a smaller number of **principal components** (artificial variables) that would account for most of the variability in the observed variables.

For simplicity, these artificial new variables are presented as a **linear combinations** of the (**optimally weighted**) observed variables. If one linear combination is not enough, we can choose to construct two, three, etc. such combinations. Note also that principal components analysis may be just an intermediate step in much larger investigations. The principal components obtained can be used for example as inputs in a regression analysis or in a cluster analysis procedure. They are also a basic method in extracting factors in factor analysis.

Precise mathematical formulation

Let $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \Sigma)$ where p is assumed to be relatively large. To perform a reduction, we are looking for a linear combination $\boldsymbol{\alpha}_1^\top \boldsymbol{X}$ with $\boldsymbol{\alpha}_1 \in \mathbb{R}^p$ suitably chosen such that it maximises the variance of $\boldsymbol{\alpha}_1^\top \boldsymbol{X}$ subject to the reasonable normalising constraint $\|\boldsymbol{\alpha}_1\|^2 = \boldsymbol{\alpha}_1^\top \boldsymbol{\alpha}_1 = 1$. Since $\mathrm{Var}(\boldsymbol{\alpha}_1^\top \boldsymbol{X}) = \boldsymbol{\alpha}_1^\top \Sigma \boldsymbol{\alpha}_1$ we need to choose $\boldsymbol{\alpha}_1$ to maximise $\boldsymbol{\alpha}_1^\top \Sigma \boldsymbol{\alpha}_1$ subject to $\boldsymbol{\alpha}_1^\top \boldsymbol{\alpha}_1 = 1$.

This requires us to apply Lagrange's optimisation under constraint procedure. You will not be examined on the derivation itself, but it is helpful to examine it to understand the relationship between the principal components and the eigenvalues and the eigenvectors of Σ.

i The derivation itself (the constrained optimisation) is not examinable.

1. construct the Lagrangian function

$$\mathrm{Lag}\left(\boldsymbol{\alpha}_1, \lambda\right) = \boldsymbol{\alpha}_1^\top \Sigma \boldsymbol{\alpha}_1 + \lambda\left(1 - \boldsymbol{\alpha}_1^\top \boldsymbol{\alpha}_1\right)$$

   where $\lambda \in \mathbb{R}^1$ is the Lagrange multiplier;
2. take the partial derivative with respect to $\boldsymbol{\alpha}_1$ and equate it to zero:

$$2\Sigma\boldsymbol{\alpha}_1 - 2\lambda\boldsymbol{\alpha}_1 = \mathbf{0} \Longrightarrow \left(\Sigma - \lambda I_p\right)\boldsymbol{\alpha}_1 = 0 \tag{2.3}$$

From (2.3) we see that $\boldsymbol{\alpha}_1$ must be an eigenvector of Σ and since we know from the Example 0.1 what the maximal value of $\frac{\boldsymbol{\alpha}^\top \Sigma \boldsymbol{\alpha}}{\boldsymbol{\alpha}^\top \boldsymbol{\alpha}}$ is, we conclude that $\boldsymbol{\alpha}_1$ should be the **eigenvector that corresponds to the largest eigenvalue** $\bar{\lambda}_1$ of Σ. The random variable $\boldsymbol{\alpha}_1^\top \boldsymbol{X}$ is called the **first principal component**.

For the **second** principal component $\boldsymbol{\alpha}_2^\top \boldsymbol{X}$ we want it to be normalised according to $\boldsymbol{\alpha}_2^\top \boldsymbol{\alpha}_2 = 1$, uncorrelated with the first component and to give maximal variance of a linear combination of the components of $\boldsymbol{X}$ under these constraints. To find it, we construct the Lagrange function:

$$\mathrm{Lag}_1\left(\boldsymbol{\alpha}_2, \lambda_1, \lambda_2\right) = \boldsymbol{\alpha}_2^\top \Sigma \boldsymbol{\alpha}_2 + \lambda_1\left(1 - \boldsymbol{\alpha}_2^\top \boldsymbol{\alpha}_2\right) + \lambda_2 \boldsymbol{\alpha}_1^\top \Sigma \boldsymbol{\alpha}_2$$

Its partial derivative w.r.t. $\boldsymbol{\alpha}_2$ gives

$$2\Sigma\boldsymbol{\alpha}_2 - 2\lambda_1\boldsymbol{\alpha}_2 + \lambda_2\Sigma\boldsymbol{\alpha}_1 = \mathbf{0} \tag{2.4}$$

Multiplying (2.4) by $\boldsymbol{\alpha}_1^\top$ from left and using the two constraints $\boldsymbol{\alpha}_2^\top \boldsymbol{\alpha}_2 = 1$ and $\boldsymbol{\alpha}_2^\top \Sigma \boldsymbol{\alpha}_1 = 0$ gives:

$$-2\lambda_1 \boldsymbol{\alpha}_1^\top \boldsymbol{\alpha}_2 + \lambda_2 \boldsymbol{\alpha}_1^\top \Sigma \boldsymbol{\alpha}_1 = 0 \Longrightarrow \lambda_2 = 0$$

Have in mind that $\boldsymbol{\alpha}_1$ is an eigenvector of Σ. Note:

$$\Sigma\boldsymbol{\alpha}_1 = \bar{\lambda}_1\boldsymbol{\alpha}_1 \Longrightarrow \boldsymbol{\alpha}_1^\top\boldsymbol{\alpha}_2 = \frac{1}{\lambda_1}\boldsymbol{\alpha}_1^\top\Sigma\boldsymbol{\alpha}_2 = 0$$

But then (2.4) also implies that $\boldsymbol{\alpha}_2 \in \mathbb{R}^p$ must be an eigenvector of Σ (has to satisfy $(\Sigma - \lambda_1 I_p)\boldsymbol{\alpha}_2 = \mathbf{0}$). Since it has to be different from $\boldsymbol{\alpha}_1$, having in mind that we aim at variance maximisation, we see that $\boldsymbol{\alpha}_2$ has to be the normalised eigenvector that corresponds to the second largest eigenvalue $\bar{\lambda}_2$ of Σ. The process can be continued further. The third principal component should be uncorrelated with the first two, should be normalised and should give maximal variance of a linear combination of the components of $\boldsymbol{X}$ under these constraints. One can easily realise then that the vector $\boldsymbol{\alpha}_3 \in \mathbb{R}^p$ in the formula $\boldsymbol{\alpha}_3^\top\boldsymbol{X}$ should be the normalised eigenvector that corresponds to the third largest eigenvalue $\bar{\lambda}_3$ of the matrix Σ etc.

Note that if we extract **all possible** p principal components then $\sum_{i=1}^p \operatorname{Var}(\boldsymbol{\alpha}_i^\top\boldsymbol{X})$ will just equal the sum of all eigenvalues of Σ and hence

$$\sum_{i=1}^p \operatorname{Var}\left(\boldsymbol{\alpha}_i^\top\boldsymbol{X}\right) = \operatorname{tr}(\Sigma) = \Sigma_{11} + \cdots + \Sigma_{pp}$$

Therefore, if we only take a small number of k principal components instead of the total possible number p we can interpret their inclusion as one that explains a $\frac{\operatorname{Var}\left(\boldsymbol{\alpha}_1^\top\boldsymbol{X}\right)+\cdots+\operatorname{Var}\left(\boldsymbol{\alpha}_k^\top\boldsymbol{X}\right)}{\Sigma_{11}+\cdots+\Sigma_{pp}} \times 100\%$ $= \frac{\bar{\lambda}_1+\cdots+\bar{\lambda}_k}{\sum_{11}+\cdots+\Sigma_{pp}} \times 100\%$ of the total population variance $\Sigma_{11} + \cdots + \Sigma_{pp}$.

Estimation of the principal components

In practice, Σ is unknown and has to be estimated. The principal components are derived from the normalised eigenvectors of the estimated covariance matrix.

Note also that extracting principal components from the (estimated) covariance matrix has the drawback that it is influenced by the scale of measurement of each variable X_i, $i = 1, \ldots, p$. A variable with large variance will necessarily be a large component in the first principal component (note the goal of explaining **the bulk** of variability by using the first principal component). Yet the large variance of the variable may be just an artefact of the measurement scale used for this variable. Therefore, an alternative practice is adopted sometimes to extract principal components from the correlation matrix $\boldsymbol{\rho}$ instead of the covariance matrix Σ.

Example 2.1 (Eigenvalues obtained from Covariance and Correlation Matrices: see page 437 Johnston and Wichern)**.** It demonstrates the great effect standardisation may have on the principal components. The relative magnitudes of the weights after standardisation (i.e. from $\boldsymbol{\rho}$ may become in direct opposition to the weights attached to the same variables in the principal component obtained from Σ.

For the reasons mentioned above, variables are often **standardised** before sample principal components are extracted. Standardisation is accomplished by calculating the vectors $\boldsymbol{Z}_i = \left(\frac{X_{1i}-\bar{X}_1}{\sqrt{s_{11}}} \quad \frac{X_{2i}-\bar{X}_2}{\sqrt{s_{22}}} \quad \ldots \quad \frac{X_{pi}-\bar{X}_p}{\sqrt{s_{pp}}}\right)^\top, i = 1, \ldots, n$. The standardised observations matrix $\boldsymbol{Z} = [\boldsymbol{Z}_1, \boldsymbol{Z}_2, \ldots, \boldsymbol{Z}_n] \in \mathcal{M}_{p,n}$ gives the sample mean vector $\bar{\boldsymbol{Z}} = \frac{1}{n}\boldsymbol{Z}\boldsymbol{1}_n = 0$ and a sample covariance matrix $\boldsymbol{S}_{\boldsymbol{Z}} = \frac{1}{n-1}\boldsymbol{Z}\boldsymbol{Z}^\top = \boldsymbol{R}$ (the correlation matrix of the original observations). The principal components are extracted in the usual way from $\boldsymbol{R}$ now.

Deciding how many principal components to include

To reduce the dimensionality (which is the motivating goal), we should restrict attention to the first k principal components and ideally, k should be kept much less than p but there is a trade-off to be made here since we would also like the proportion $\psi_k = \frac{\bar{\lambda}_1 + \ldots \bar{\lambda}_k}{\lambda_1 + \ldots \lambda_p}$ be close to one. How could a reasonable trade-off be made? Three methods are most widely used:

- **The scree plot:** basically, it is a graphical method of plotting the ordered $\bar{\lambda}_k$ against k and deciding visually when the plot has flattened out. Typically, the initial part of the plot is like the side of the mountain, while the flat portion where each $\bar{\lambda}_k$ is just slightly smaller than $\bar{\lambda}_{k-1}$, is like the rough scree at the bottom. This motivates the name of the plot. The task here is to find where "the scree begins".
- **Total variance explained:** Choose an arbitrary constant $c \in (0, 1)$ and choose k to be the smallest one with the property $\psi_k \geq c$. Usually, $c = 0.9$ is used but note the arbitrariness of the choice here.
- **Kaiser's rule:** it suggests that from all p principal components only the ones should be retained whose variances (after standardisation) are greater than unity, or, equivalently, only those components which, individually, explain at least $\frac{1}{p}100\%$ of the total variance. (This is the same as excluding all principal components with eigenvalues less than the overall average). This criterion has a number of positive features that have contributed to its popularity but can not be defended on a safe theoretical ground.
- **Formal tests of significance:** Note that it actually **does not make sense** to test whether $\bar{\lambda}_{k+1} = \cdots = \bar{\lambda}_p = 0$ since if such a hypothesis were true then the population distribution would be contained **entirely** within a k-dimensional subspace and the same would be true for any **sample** from this distribution, hence we would have the **estimated** $\bar{\lambda}$ values for indices $k + 1, \ldots, p$ being also equal to zero with probability one! What seems to be reasonable to do instead, is to test $H_0 : \bar{\lambda}_{k+1} = \cdots = \bar{\lambda}_p$ (without asking the common value to be zero). This is a more quantitative variant of the scree test. A test for this hypothesis is to form the arithmetic and geometric means $a_0 =$ arithmetic mean of the last $p - k$ estimated eigenvalues; $g_0 =$ geometric mean of the last $p - k$ estimated eigenvalues, and then construct $-2 \log \lambda = n(p - k) \log \frac{a_0}{g_0}$. The asymptotic distribution of this statistic under the null hypothesis is χ^2_ν where $\nu = \frac{(p-k+2)(p-k-1)}{2}$. The interested student can find more details about this test in the monograph of Mardia, Kent and Bibby. We should note, however, that the last result holds under multivariate normality assumption and is only valid as stated for the **covariance-based** (**not** the correlation-based) version of the principal component analysis. In practice, many data analysts are reluctant to make a multivariate normality assumption at the early stage of the descriptive data analysis and hence distrust the above quantitative test but prefer the simple Kaiser criterion.

Check your understanding

Complete the below exercises to check your understanding of concepts presented so far.

1. A random vector $\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \end{pmatrix}$ is normally distributed with zero mean vector $\Sigma = \begin{pmatrix} 1 & \rho/2 & 0 \\ \rho/2 & 1 & \rho \\ 0 & \rho & 1 \end{pmatrix}$ where ρ is positive.

a) Find the coefficients of the first principal component and the variance of that component. What percentage of the overall variability does it explain?

b) Find the joint distribution of Y_1, Y_2 and $Y_1 + Y_2 + Y_3$

c) Find the conditional distribution of Y_1, Y_2 given $Y_3 = y_3$.

d) Find the multiple correlation of Y_3 with Y_1, Y_2.

Demonstration: Crime rates in the USA

Optional viewing: StatQuest: PCA in R

StatQuest with Josh Starmer. (2016). StatQuest: PCA in R. Retrieved from https://youtu.be/0Jp4gsfOLMs .

Before attempting the task, please watch the following demonstration by Dr Krivitsky.

Transcript

This demonstration can be completed using the provided RStudio or your own RStudio.

To complete this task select the 'PCA_Examples.demo.Rmd' in the 'Files' section of RStudio. Follow the demonstration contained within the RMD file.

If you choose to complete the example in your own RStudio, upload the following file:

PCA_Examples.demo.Rmd

The output of the RMD file is also displayed below:

Challenge: PCA

If you choose to complete this task in your own RStudio, upload the following file:

PCA_Examples.challenge.Rmd

Click on the 'PCA_Examples.challenge.Rmd' in the 'Files' section to begin. Enter your response to the tasks in the 'Enter your code here' section.

This activity and the solution will be discussed at the Collaborate session this week. In the meantime, share and discuss your results in the 'Tutorials' discussion forum.

The solution will also be available here on Friday of this week by clicking on the 'Solution' tab in the top right corner.

The output of the RMD file is also displayed below: