VIII Lecture

Principal Components Analysis

8.1. Introduction. The basics of the procedure

Principal components 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.

8.2. Precise mathematical formulation.

Let $\mathbf{X} \sim N_p(\mu, \Sigma)$ where p is assumed to be relatively large. To perform a reduction, we are looking for a linear combination $\alpha_1'\mathbf{X}$ with $\alpha_1 \in R^p$ suitably chosen such that it maximizes the variance of $\alpha_1'\mathbf{X}$ subject to the reasonable norming constraint $\| \alpha_1 \|^2 = \alpha_1'\alpha_1 = 1$. Since $Var(\alpha_1'\mathbf{X}) = \alpha_1'\Sigma\alpha_1$ we need to choose α_1 to maximize $\alpha_1'\Sigma\alpha_1$ subject to $\alpha_1'\alpha_1 = 1$. Since it goes about optimization with respect to a vector argument, some simple differentiation rules should be re-called.

For a vector variable $y \in R^p$, a vector of constants $a \in R^p$ and a symmetric $p \times p$ matrix A it holds

- $\frac{\partial}{\partial y}(y'Ay) = 2Ay$ (even more generally

$$\frac{\partial}{\partial y}(y'Ay) = Ay + A'y$$

  if A is not necessarily symmetric).
- $\frac{\partial}{\partial y}(y'y) = 2y$
- $\frac{\partial}{\partial y}(y'a) = a.$

Performing the optimization requires to apply Lagrange's optimization under constraint procedure:

i) construct the Lagrange function

$$Lag(\alpha_1, \lambda) = \alpha_1'\Sigma\alpha_1 + \lambda(1 - \alpha_1'\alpha_1)$$

where $\lambda \in R^1$ is the Lagrange multiplier;

ii) take the partial derivative with respect to α_1 and equate it to zero:

$$2\Sigma\alpha_1 - 2\lambda\alpha_1 = 0 \longrightarrow (\Sigma - \lambda I_p)\alpha_1 = 0 \tag{8.1}$$

From (8.1) we see that α_1 must be an eigenvector of Σ and since we know from the first lecture what the maximal value of $\frac{\alpha'\Sigma\alpha}{\alpha'\alpha}$ is , we conclude that α_1 should be the **eigenvector that corresponds to the largest eigenvalue $\bar{\lambda}_1$ of** Σ. The random variable $\alpha_1'\mathbf{X}$ is called the **first principal component**.

For the **second** principal component $\alpha_2'\mathbf{X}$ we want it to to be normed according to $\alpha_2'\alpha_2 = 1$, uncorrelated with the first component and to give maximal variance of a linear combination of the components of $\mathbf{X}$ under these constraints. To find it, we construct the Lagrange function:

$$Lag_1(\alpha_2, \lambda_1, \lambda_2) = \alpha_2'\Sigma\alpha_2 + \lambda_1(1 - \alpha_2'\alpha_2) + \lambda_2\alpha_1'\Sigma\alpha_2$$

Its partial derivative w.r. α_2 gives

$$2\Sigma\alpha_2 - 2\lambda_1\alpha_2 + \lambda_2\Sigma\alpha_1 = 0 \tag{8.2}$$

Multiplying (8.2) by α_1' from left and using the two constraints $\alpha_2'\alpha_2 = 1$ and $\alpha_2'\Sigma\alpha_1 = 0$ gives:

$$-2\lambda_1\alpha_1'\alpha_2 + \lambda_2\alpha_1'\Sigma\alpha_1 = 0 \longrightarrow \lambda_2 = 0$$

(WHY (?) Have in mind that α_1 was an eigenvector of Σ). But then (8.2) also implies that $\alpha_2 \in R^p$ must be an eigenvector of Σ (has to satisfy $(\Sigma - \lambda_1 I_p)\alpha_2 = 0$). Since it has to be different from α_1, having in mind that we aim at variance maximization, we see that α_2 has to be the normed 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 normed and should give maximal variance of a linear combination of the components of $\mathbf{X}$ under these constraints. One can easily realize then that the vector $\alpha_3 \in R^p$ in the formula $\alpha_3'\mathbf{X}$ should be the normed 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 $\sum_{i=1}^{p} Var(\alpha_i'\mathbf{X})$ will just equal the sum of all eigenvalues of Σ and hence

$$\sum_{i=1}^{p} Var(\alpha_i'\mathbf{X}) = tr(\Sigma) = \sigma_{11} + \ldots + \sigma_{pp}$$

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 $\frac{Var(\alpha_1'\mathbf{X})+\ldots Var(\alpha_k'\mathbf{X})}{\sigma_{11}+\ldots+\sigma_{pp}}.100\% = \frac{\bar{\lambda}_1+\ldots\bar{\lambda}_k}{\sigma_{11}+\ldots+\sigma_{pp}}.100\%$ of the total population variance $\sigma_{11}+\ldots+\sigma_{pp}$.

8.3. Estimation of the Principal Components

In practice, Σ is unknown and has to be estimated. The principal components are derived from the normed 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 artifact of the measurement scale used for this variable. Therefore, an alternative practice is adopted sometimes to extract principal components from the correlation matrix ρ instead of the covariance matrix Σ.

Example (Eigenvalues obtained from Covariance and Correlation Matrices- see page 437 Johnston and Wichern). It demonstrates the great effect standardization may have on the principal components. The relative magnitudes of the weights after standardization (i.e. from ρ 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 **standardized** before sample principal components are extracted. Standardization is accomplished by calculating the vectors $Z_i = \begin{pmatrix} \frac{X_{1i}-\bar{X}_1}{\sqrt{s_{11}}} \\ \frac{X_{2i}-\bar{X}_2}{\sqrt{s_{22}}} \\ \ldots \\ \frac{X_{pi}-\bar{X}_p}{\sqrt{s_{pp}}} \end{pmatrix}, i = 1, \ldots, n$. The standardized observations matrix $\mathbf{Z} = [\mathbf{Z_1}, \mathbf{Z_2}, \ldots, \mathbf{Z_n}] = \begin{pmatrix} Z_{11} & Z_{12} & \ldots & Z_{1n} \\ Z_{21} & Z_{22} & \ldots & Z_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ Z_{p1} & Z_{p2} & \ldots & Z_{pn} \end{pmatrix} \in \mathscr{M}_{p,n}$ gives the sample mean vector $\bar{\mathbf{Z}} = \frac{1}{n}\mathbf{Z1_n} = 0$ and a sample covariance matrix $\mathbf{S_z} = \frac{\mathbf{1}}{\mathbf{n-1}}\mathbf{ZZ'} = \mathbf{R}$ (the correlation matrix of the original observations). The principal components are extracted in the usual way from $\mathbf{R}$ now.

8.4. 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 $\psi_k = \frac{\bar{\lambda}_1+\ldots\bar{\lambda}_k}{\bar{\lambda}_1+\ldots\bar{\lambda}_p}$ be close to one. How could a reasonable trade-off be made? Three methods are most widely used:

- The "screeplot": 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."

- 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 please, note the arbitrariness of the choice here.

- **Kaiser's rule:** it is applied when extracting the components from the *correlation* matrix and suggests that from all p principal components only the ones should be retained whose variances 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 $\bar{\lambda}_{k+1} = \ldots = \bar{\lambda}_p = 0$ since if such a hypothesis was 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 $H_0 : \bar{\lambda}_{k+1} = \ldots = \bar{\lambda}_p$ (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 algebraic and geometric means a_0 =algebraic mean of the last $p-k$ estimated eigenvalues; g_0 =geometric mean of the last $p-k$ estimated eigenvalues, and then construct $-2log\lambda = n(p-k)\log\frac{a_0}{g_0}$. 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.

8.5. Numerical example The Crime Rates example will be discussed at the lecture. The data gives crime rates per 100000 people in seven categories for each of the 50 states in USA in 1997. Principal components are used to summarize the 7-dimensional data in 2 or 3 dimensions only and help to visualize and interpret the data.

Basically, principal components analysis can be performed in SAS by using either the PRINCOMP or the FACTOR procedures. Principal components can serve as a method for initial factor extraction in exploratory factor analysis. But one should mention here that *Principal component analysis is* **not** *Factor analysis*. The main difference is that in factor analysis (to be studied later in this course) one assumes that the covariation in the observed variables is due to the presence of one or more latent variables (factors) that exert casual influence on the observed variables. Factor analysis is being used when it is believed that certain latent factors exist and it is hoped to explore the nature and number of these factors. In contrast, in principal component analysis there is no prior assumption about an underlying casual model. The goal here is just variable reduction.

8.6. Example from finance: portfolio optimization. Many other problems in multivariate Statistics lead to formulating optimization problems that are similar in spirit to the Principal Component Analysis problem. Hereby, we shall illustrate the Efficient portfolio choice problem.

Assume that a $p-$dimensional vector X of returns of the p assets is given. Then the return of a **portfolio** that has these assets in proportions $(c_1, c_2, \ldots, c_p)$ (with $\sum_{i=1}^p c_i = 1$) is $Q = c'X$ and the mean return is $c'\mu$ (Here we assume that $EX = \mu, D(X) = \Sigma$.) The *risk* of the portfolio is $c'\Sigma c$. Further, assume that a pre-specified mean return $\bar{\mu}$ is to be achieved. The question is *how to choose the weights* c so that the risk of a portfolio that achieves the pre-specified mean return, is as small as possible.

Mathematically, this is equivalent to the requirement to find the solution of an optimization problem under two constraints. The Lagrange function is:

$$Lag(\lambda_1, \lambda_2) = c'\Sigma c + \lambda_1(\bar{\mu} - c'\mu) + \lambda_2(1 - c'\mathbf{1}_p) \tag{8.3}$$

where $\mathbf{1}$ is a p-dimensional vector of ones. Differentiating (8.3) with respect to c we get the first order conditions for a minimum:

$$2\Sigma c - \lambda_1 \mu - \lambda_2 \mathbf{1}_p = 0 \tag{8.4}$$

To simplify derivations, we shall consider the so-called case of non-existence of a riskless asset with a fixed (non-random) return. Then it makes sense to assume that Σ is positive definite and hence Σ^{-1} exists. We get from (8.4) then:

$$c = \frac{1}{2}\Sigma^{-1}(\lambda_1 \mu + \lambda_2 \mathbf{1}_p) \tag{8.5}$$

After multiplying by $\mathbf{1}_p'$ from left both sides of the equality, we get:

$$1 = \frac{1}{2}\mathbf{1}_p'\Sigma^{-1}(\lambda_1 \mu + \lambda_2 \mathbf{1}_p) \tag{8.6}$$

We can get λ_2 from (8.6) as $\lambda_2 = \frac{2-\lambda_1 \mathbf{1}_p'\Sigma^{-1}\mu}{\mathbf{1}_p'\Sigma^{-1}\mathbf{1}_p}$ and then substitute it in the formula for c to end up with:

$$c = \frac{1}{2}\lambda_1(\Sigma^{-1}\mu - \frac{\mathbf{1}_p'\Sigma^{-1}\mu}{\mathbf{1}_p'\Sigma^{-1}\mathbf{1}_p}\Sigma^{-1}\mathbf{1}_p) + \frac{\Sigma^{-1}\mathbf{1}_p}{\mathbf{1}_p'\Sigma^{-1}\mathbf{1}_p} \tag{8.7}$$

In a similar way, if we multiply both sides of (8.5) by μ' from left and use the restriction $\mu' c = \bar{\mu}$ we can get one more relationship between λ_1 and λ_2 : $\lambda_1 = \frac{2\bar{\mu}-\lambda_2 \mu'\Sigma^{-1}\mathbf{1}_p}{\mu'\Sigma^{-1}\mu}$ The linear system of 2 equations with respect to λ_1 and λ_2 can be solved then and the values substituted in (8.7) to get the final expression for c using μ, $\bar{\mu}$ and Σ. (Do it (!))

One special case is of particular interest. This is the so-called variance-efficient portfolio (as opposed to the *mean-variance efficient portfolio* considered above). For the variance-efficient portfolio *there is no pre-specified mean return, that is, there is no restriction on the mean.* It is only required to minimize the variance. Obviously, we have $\lambda_1 = 0$ then and from (8.7) we get the *optimal weights for the variance efficient portfolio:* $c_{opt} = \frac{\Sigma^{-1}\mathbf{1}_p}{\mathbf{1}_p'\Sigma^{-1}\mathbf{1}_p}$.