THE UNIVERSITY OF CHICAGO
Booth School of Business
Business 41912, Spring Quarter 2012, Mr. Ruey S. Tsay

Lecture 6: Principal Component Analysis

1 Population Principal Components

A principal component analysis is concerned with explaining the variance-covariance structure of a set of variables through some linear combinations of these variables. Its general objectives are (a) dimension reduction and (b) interpretation.
Consider p random variables $X_1, \ldots, X_p$. Principal component analysis seeks to select a new coordinate system obtained by rotating the original system with $X_1, \ldots, X_p$ as the coordinate axes. The new axes represent the directions with maximum variability and provide a simpler and more parsimonious description of the covariance structure.
Let $\boldsymbol{X}' = (X_1, \ldots, X_p)'$. Denote the covariance matrix of $\boldsymbol{X}$ by $\boldsymbol{\Sigma}$ whose eigenvalues are $\lambda_1 \geq \lambda_2 \geq \cdots \lambda_p \geq 0$. Let $Y_i = \boldsymbol{a}_i'\boldsymbol{X} = a_{i1}X_1 + a_{i2}X_2 + \cdots + a_{ip}X_p$ be a linear combination of $\boldsymbol{X}$. Then, we have

$$\text{Var}(Y_i) = \text{Var}(\boldsymbol{a}_i'\boldsymbol{X}) = \boldsymbol{a}_i'\boldsymbol{\Sigma}\boldsymbol{a}_i, \quad i = 1, \ldots, p \tag{1}$$

and, more generally, for p such linear combinations,

$$\text{Cov}(Y_i, Y_j) = \boldsymbol{a}_i'\boldsymbol{\Sigma}\boldsymbol{a}_j, \quad i, j = 1, \ldots, p.$$

The principal components (PC) are those linear combinations $Y_1, \ldots, Y_p$ that are uncorrelated and whose variance in (1) are as large as possible. Thus,

The 1st PC = linear combination $\boldsymbol{a}_1'\boldsymbol{X}$ that maximizes $\text{Var}(\boldsymbol{a}_1'\boldsymbol{X})$ subject to $\boldsymbol{a}_1'\boldsymbol{a}_1 = 1$.

The 2nd PC = linear combination $\boldsymbol{a}_2'\boldsymbol{X}$ that maximizes $\text{Var}(\boldsymbol{a}_2'\boldsymbol{X})$ subject to $\boldsymbol{a}_2'\boldsymbol{a}_2 = 1$,
and $\text{Cov}(\boldsymbol{a}_1'\boldsymbol{X}, \boldsymbol{a}_2'\boldsymbol{X}) = \boldsymbol{a}_2'\boldsymbol{\Sigma}\boldsymbol{a}_1 = 0$.

In general,

The ith PC = linear combination $\boldsymbol{a}_i'\boldsymbol{X}$ that maximizes $\text{Var}(\boldsymbol{a}_i'\boldsymbol{X})$

subject to $\boldsymbol{a}_i'\boldsymbol{a}_i = 1$ and $\text{Cov}(\boldsymbol{a}_j'\boldsymbol{X}, \boldsymbol{a}_i'\boldsymbol{X}) = 0, \quad j = 1, \ldots, i-1$.

Result 8.1. Let $\boldsymbol{\Sigma}$ be the covariance matrix of the random vector $\boldsymbol{X}' = (X_1, \ldots, X_p)'$. Let the eignevalue-eigenvector pairs of $\boldsymbol{\Sigma}$ be $(\lambda_1, \boldsymbol{e}_1), \ldots, (\lambda_p, \boldsymbol{e}_p)$ such that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$. Then the ith PC of $\boldsymbol{X}$ is given by

$$Y_i = \boldsymbol{e}_i'\boldsymbol{X} = e_{i1}X_1 + \cdots + e_{ip}X_p, \quad i = 1, \ldots, p.$$

Furthermore,

$$\mathrm{Var}(Y_i) = \boldsymbol{e}_i'\boldsymbol{\Sigma}\boldsymbol{e}_i = \lambda_i, \quad i = 1, \ldots, p,$$
$$\mathrm{Cov}(Y_i, Y_j) = \boldsymbol{e}_i'\boldsymbol{\Sigma}\boldsymbol{e}_j = 0, \quad i \neq j.$$

If some λ_j are equal then the choices of the corresponding coefficient vectors $\boldsymbol{e}_j$ and, hence, Y_j are not unique.
Proof. The result follows Eq. (2.51) and (2.52) of the textbook. See Page 80.

Result 8.2. Let $\boldsymbol{\Sigma}$ be the covariance matrix of the random vector $\boldsymbol{X} = (X_1, \ldots, X_p)'$ and $(\lambda_1, \boldsymbol{e}_1), \ldots, (\lambda_p, \boldsymbol{e}_p)$ be the ordered (in increasing order) eigenvalue-eigenvector pairs of $\boldsymbol{\Sigma}$. Then

$$\sum_{i=1}^{p} \sigma_{ii} = \sum_{i=1}^{p} \mathrm{Var}(X_i) = \sum_{i=1}^{p} \lambda_i = \sum_{i=1}^{p} \mathrm{Var}(Y_i).$$

The proportion of total population variance explained by the ith PC is

$$\frac{\lambda_i}{\sum_{j=1}^{p} \lambda_j}.$$

Let $\rho_{X,Y}$ be the correlation coefficient between the random variables X and Y.

Result 8.3. Let $Y_i = \boldsymbol{e}_i'\boldsymbol{X}$ be the PC of the random vector $\boldsymbol{X}$ with covariance matrix $\boldsymbol{\Sigma}$. Then,

$$\rho_{Y_i, X_k} = \frac{e_{ik}\sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}}, \quad i, k = 1, 2, \ldots, p.$$

Proof: Use $X_i = \boldsymbol{a}_i'\boldsymbol{X}$, where a_i is the ith unit vector in R^p, and $\boldsymbol{\Sigma}\boldsymbol{e}_i = \lambda_i\boldsymbol{e}_i$.
Correlation matrix: Let $\boldsymbol{Z} = \boldsymbol{D}^{-1}\boldsymbol{X}$, where $\boldsymbol{D} = \mathrm{diag}\{\sqrt{\sigma_{11}}, \ldots, \sqrt{\sigma_{pp}}\}$ with σ_{ii} being the $(i.i)$th element of the covariance matrix $\boldsymbol{\Sigma}$ of $\boldsymbol{X}$. Then the ith PC of $\boldsymbol{Z}$ is

$$Y_i = \boldsymbol{e}_i'\boldsymbol{D}^{-1}(\boldsymbol{X} - \boldsymbol{\mu}),$$

where $\boldsymbol{\mu} = E(\boldsymbol{X})$ and $(\lambda_i, \boldsymbol{e}_i)$s are the ordered eigenvalue-eigenvector pairs of $\boldsymbol{\rho} = \mathrm{Cov}(\boldsymbol{Z})$. Moreover,

$$\sum_{i=1}^{p} \mathrm{Var}(Y_i) = \sum_{i=1}^{p} \mathrm{Var}(Z_i) = p$$

and

$$\rho_{Y_i, Z_k} = e_{ik}\sqrt{\Lambda_i}, \quad i, k = 1, \ldots, p.$$

Remark: Principal component analysis depends on the scales of $\boldsymbol{X}$. Thus, in some applications, correlation matrix is used in PCA to obtain scale-invariant results.

2 Sample PC

Suppose that the data $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ represent n independent draws from some p-dimensional population with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Let $\bar{\boldsymbol{x}}$, $\boldsymbol{S}$ and $\boldsymbol{R}$ be the sample mean, covariance matrix and correlation matrix, respectively. The sample PCs are the counterparts of the population PCs with $\boldsymbol{\Sigma}$ and $\boldsymbol{\rho}$ replaced by $\boldsymbol{S}$ and $\boldsymbol{R}$.

2.1 The number of PCs

The scree plot which is a time-series plot of the eigenvalues of $\boldsymbol{S}$ or $\boldsymbol{R}$ in descreasing order. That is, the scatter-plot of $(i, \hat{\lambda}_i)$, $i = 1, \ldots, p$. By looking for an elbow (bend) in the scree plot, we can determine the number of PCs. We shall mention some recent works that use information criteria to select the number of principal components later when we discuss factor models.

2.2 Large sample inferences

Asymptotic results for eigenvalues and eigenvectors are only available under the normality assumption and the condition that all eigenvalues are distinct and positive. Some limited results start to appear when the variables are not normally distributed.
Anderson (1963, *Annals of Mathematical Statistics*) derived the following large sample distribution theory for the eigenvalues $\widehat{\boldsymbol{\lambda}} = (\hat{\lambda}_1, \ldots, \hat{\lambda}_p)'$ and eigenvectors $\widehat{\boldsymbol{e}}_1, \ldots, \widehat{\boldsymbol{e}}_p$ of the sample covariance matrix $\boldsymbol{S}$:

1. Let $\boldsymbol{\Lambda}$ be the diagonal matrix of eigenvalues $\lambda_1, \ldots, \lambda_p$ of $\boldsymbol{\Sigma}$, then $\sqrt{n}(\widehat{\boldsymbol{\lambda}} - \boldsymbol{\lambda}) \sim N_p(\boldsymbol{0}, 2\boldsymbol{\Lambda}^2)$.

2. Let
$$\boldsymbol{E}_i = \lambda_i \sum_{k=1, k\neq i}^{p} \frac{\lambda_k}{(\lambda_k - \lambda_i)^2} \boldsymbol{e}_k \boldsymbol{e}_k',$$
then $\sqrt{n}(\widehat{\boldsymbol{e}}_i - \boldsymbol{e}_i) \sim N_p(\boldsymbol{0}, \boldsymbol{E}_i)$.

3. Each $\hat{\lambda}_i$ is distributed independently of the elements of the associated $\widehat{\boldsymbol{e}}_i$.

Property 1 implies that, for large n, the sample eigenvalues $\hat{\lambda}_i$ are independently distributed. Moreover, $\hat{\lambda}_i$ has an approximate $N(\lambda_i, 2\lambda_i^2/n)$ distribution. Using this result, a large sample $100(1-\alpha)\%$ confidence interval for λ_i is provided by
$$\frac{\hat{\lambda}_i}{1 + z(\alpha/2)\sqrt{2/n}} \leq \lambda_i \leq \frac{\hat{\lambda}_i}{1 - z(\alpha/2)\sqrt{2/n}},$$
where $z(\alpha/2)$ is the upper $100(\alpha/2)$ percentile of s standard normal distribution.
Peoperty 2 implies that the $\widehat{\boldsymbol{e}}_i'$s are normally distributed about the corresponding $\boldsymbol{e}_i$'s for large samples. The elements of each $\widehat{\boldsymbol{e}}_i$ are correlated, and the correlation depends to a large

extent on the separation of the eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_p$ and the sample size n. Approximate standard errors for the coefficients $\hat{e}_{i,k}$ are given by the square roots of the diagonal elements of $(1/n)\widehat{\boldsymbol{E}}_i$, where $\widehat{\boldsymbol{E}}_i$ is derived from $\boldsymbol{E}_i$ by substituting $\hat{\lambda}_i$'s for λ_i's and $\widehat{\boldsymbol{e}}_i$'s for the $\boldsymbol{e}_i$'s.

Testing for the equal correlation structure: Consider the null hypothesis

$$H_o : \boldsymbol{\rho} = \boldsymbol{\rho}_o = \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{bmatrix}$$

and the alternative hypothesis $H_1 : \boldsymbol{\rho} \neq \boldsymbol{\rho}_0$. Lawley (1963) proposed a testing procedure. Let

$$\bar{r}_k = \frac{1}{p-1} \sum_{i=1, i\neq k}^{p} r_{ik}, \quad k = 1, 2, \ldots, p; \quad \bar{r} = \frac{2}{p(p-1)} \sum\sum_{i<k} r_{ik},$$

$$\hat{\gamma} = \frac{(p-1)^2[1-(1-\bar{r})^2]}{p-(p-2)(1-\bar{r})^2},$$

where $\bar{r}_k$ is the average of the off-diagonal elements in the kth column of $\boldsymbol{R}$ and $\bar{r}$ is the overall average of the off-diagonal elements. The large sample approximate α-level test is to reject H_o in favor of $\boldsymbol{H}_a$ if

$$T = \frac{(n-1)}{(1-\bar{r})^2} \left[\sum\sum_{i<k} (r_{ik} - \bar{r})^2 - \hat{\gamma} \sum_{k=1}^{p} (\bar{r}_k - \bar{r})^2 \right] > \chi^2_{(p+1)(p-2)/2}(\alpha).$$

Remark: The R command of PC analysis is **princomp**.