Chapter 8. Principal Components.

Principal component, abbreviated as P.C., analysis, was invented by Carl Pearson in 1901. The aim is to create one or a few new variables by linearly combining the existing variables in study so that the new variables, which are called principal components, explain most of the variations. By so doing, a data reduction may be achieved—the "many" variables in study are compressed into "a few" most important P.C.s with, hopefully, little loss of information. In the mean time, it is possible that one may identify the major source of variation through interpretation/expression of the P.C.s. The following example may offer more insights towards the motivation and process of P.C. analysis.

Example 8.1. Analysis of stock price data See Appendix B.

8.1. Population principal components.

Let

$$X = \begin{pmatrix} X_1 \\ \vdots \\ X_p \end{pmatrix}$$

be a random vector of p dimension with $p \times p$ variance matrix

$$\Sigma = \begin{pmatrix} \sigma_{11} & \cdots & \sigma_{1p} \\ \vdots & \vdots & \vdots \\ \sigma_{p1} & \cdots & \sigma_{pp} \end{pmatrix}, \qquad \text{where } \sigma_{kl} = cov(X_k, X_l).$$

If follows from the decomposition of positive definite matrix that

$$\Sigma = \mathbf{e}\Lambda\mathbf{e}'$$

where

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_p \end{pmatrix} \quad \text{with } \lambda_1 \geq \cdots \geq \lambda_p > 0$$

and

$$\mathbf{e} = (\mathbf{e}_1 \vdots \cdots \vdots \mathbf{e}_p) = \begin{pmatrix} e_{11} & \cdots & e_{1p} \\ \vdots & \vdots & \vdots \\ e_{p1} & \cdots & e_{pp} \end{pmatrix} \quad \text{is an orthonormal matrix,}$$

i.e., $\mathbf{e}\mathbf{e}' = I_p$. Note that $\mathbf{e}$ is a $p \times p$ matrix and $\mathbf{e}_k$ is its k-th column and therefore is a p-vector. Set

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_p \end{pmatrix} = \mathbf{e}'(X - \mu) = \begin{pmatrix} \mathbf{e}_1' \\ \vdots \\ \mathbf{e}_p' \end{pmatrix} (X - \mu).$$

Clearly, $Y_j = \mathbf{e}_j'(X - \mu)$. By a simply calculation,

$$\text{var}(Y) = \mathbf{e}'\text{var}(X)\mathbf{e} = \mathbf{e}'\Sigma\mathbf{e} = \mathbf{e}'\mathbf{e}\Lambda\mathbf{e}'\mathbf{e} = \Lambda.$$

In particular, $\text{var}(Y_j) = \lambda_j$, $j = 1, ..., p$, and $\text{cov}(Y_k, Y_l) = 0$, for $1 \leq k \neq l \leq p$. Then, Y_j is called the j-th population P.C. The interpretation of the P.C.s is presented in the following. To make it clearer, we call a linear combination of X, $b'X$ with $\|b\| = 1$ a unitary linear combination.

(1). The first P.C. Y_1 explains the most variation among all unitary linear combinations of X. Namely,

$$\text{var}(Y_1) = \lambda_1 = \max\{\text{var}(b'X) : \|b\| = 1,\ b \in R^p\}.$$

The fraction of total variation of X explained by Y_1 is

$$\frac{\text{var}(Y_1)}{\text{var}(Y_1)+\cdots \text{var}(Y_p)} = \frac{\lambda_1}{\lambda_1+\cdots \lambda_p}.$$

Note that $\lambda_1+\cdots+\lambda_p = trace(\Sigma)$ is used to measure total variation of X.

(2). The k-th P.C. Y_k explains the most variation not explained by the previous $k-1$ P.C.s $Y_1, ..., Y_{k-1}$ among all unitary linear combination. Specifically,

$$\text{var}(Y_k) = \lambda_k = \max\{\text{var}(b'X) : \|b\| = 1,\ b'X \perp Y_1,\ ...,\ b'X \perp Y_{k-1},\ b \in R^p\}$$

Here and throughout, $\perp$ means 0 correlation. The fraction of total variation of X explained by Y_k is

$$\frac{\text{var}(Y_k)}{\text{var}(Y_1)+\cdots \text{var}(Y_p)} = \frac{\lambda_k}{\lambda_1+\cdots \lambda_p}.$$

We may summarize the P.C.s in the following table.

		eigenvalue (variance)	eigenvector (combination coefficient)	percent of variation explained	P.C.s as linear combination of $X-\mu$
1st P.C.	Y_1	λ_1	$\mathbf{e}_1$	$\lambda_1/(\lambda_1+\cdots+\lambda_p)$	$Y_1 = \mathbf{e}_1'(X-\mu)$
2nd P.C.	Y_2	λ_2	$\mathbf{e}_2$	$\lambda_2/(\lambda_1+\cdots+\lambda_p)$	$Y_2 = \mathbf{e}_2'(X-\mu)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
p-th P.C.	Y_p	λ_p	$\mathbf{e}_p$	$\lambda_p/(\lambda_1+\cdots+\lambda_p)$	$Y_1 = \mathbf{e}_p'(X-\mu)$

Note that $\mathbf{e}_j = (e_{1j}, ..., e_{pj})'$ is the j-th column of $\mathbf{e}$. As the P.C.s are orthogonal to each other (0 correlated), the part of variation explained by each P.C.s are distinct or non-overlapping with each other.

The relative size of the variance of a P.C. or the percentage of total variation explained measures the importance of the P.C.. Thus the 1st P.C. is the most important, the 2nd P.C. the 2nd important, and so on.

It is often desired to reduce the number of variables, especially when the number of variables in concern are too many. But the reduction must be done without much loss of information. P.C.s provide an ideal way of such reduction. One may retain the first k P.C.s, which altogether explains

$$\frac{\lambda_1+\cdots+\lambda_k}{\lambda_1+\cdots+\lambda_p}$$

of the total variation.

8.2. Sample principal components

The population P.C.s in only theoretical, but it parallels the sample P.C.s which can be computed from data. Suppose there are n observations of p variables presented as

$$\mathbf{X} = \left(X_{(1)} \vdots X_{(2)} \vdots \cdots \vdots X_{(p)}\right) = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \vdots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}_{n\times p}.$$

Then $X_{(k)}$, an n-vector, contains all n observations of the k-th variable. Let $\mathbf{S}$ be the sample variance matrix. By decomposition,

$$\mathbf{S} = \hat{\mathbf{e}}\hat{\Lambda}\hat{\mathbf{e}}'$$

Let

$$\begin{aligned}\mathbf{Y}_{n\times p} &= \Big(Y_{(1)}\vdots Y_{(2)}\vdots\cdots\vdots Y_{(p)}\Big)\\ &= \Big(X_{(1)}-\bar{X}_1\vdots X_{(2)}-\bar{X}_2\vdots\cdots\vdots X_{(p)}-\bar{X}_p\Big)\hat{\mathbf{e}}\end{aligned}$$

where $\bar{X}_k = (1/n)\sum_{i=1}^n x_{ik}$ is the sample average of the n observations of the k-th variable. We summarize the sample P.C.s as follows.

		eigenvalue (variance)	eigenvector (combination coefficient)	percent of variation explained	P.C.s as linear combination of $X-\mu$
1st P.C.	$Y_{(1)}$	$\hat{\lambda}_1$	$\hat{\mathbf{e}}_1$	$\hat{\lambda}_1/(\hat{\lambda}_1+\cdots+\hat{\lambda}_p)$	$Y_{(1)}=\sum_{j=1}^p \hat{e}_{j1}(X_{(j)}-\bar{X}_1)$
2nd P.C.	$Y_{(2)}$	$\hat{\lambda}_2$	$\hat{\mathbf{e}}_2$	$\hat{\lambda}_1/(\hat{\lambda}_1+\cdots+\hat{\lambda}_p)$	$Y_{(2)}=\sum_{j=1}^p \hat{e}_{j2}(X_{(j)}-\bar{X}_1)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
p-th P.C.	$Y_{(p)}$	$\hat{\lambda}_p$	$\hat{\mathbf{e}}_1$	$\hat{\lambda}_p/(\hat{\lambda}_1+\cdots+\hat{\lambda}_p)$	$Y_{(p)}=\sum_{j=1}^p \hat{e}_{jp}(X_{(j)}-\bar{X}_1)$

Interpretations analogous to the population P.C.s applies to the sample P.C.s. We omit details.

8.3. Inference and standardization

Under regularity conditions,

$$\sqrt{n}\begin{pmatrix}\hat{\lambda}_1-\lambda_1\\ \vdots\\ \hat{\lambda}_p-\lambda_p\end{pmatrix}\to MN(0,2\Lambda^2)$$

In particular,

$$\sqrt{n}(\hat{\lambda}_k-\lambda_k)\to N(0,2\lambda_k^2).$$

Hence a confidence interval for λ_k at approximate confidence level $1-\alpha$ is

$$\hat{\lambda}_k \pm z(\frac{\alpha}{2})\hat{\lambda}_k\sqrt{\frac{2}{n}}.$$

And simultaneous confidence intervals for λ_j, $j=1,...,K$, at confidence level $1-\alpha$ by Bonferroni's method are:

$$\text{for } \lambda_j:\ \hat{\lambda}_j \pm z(\frac{\alpha}{2K})\hat{\lambda}_j\sqrt{\frac{2}{n}},\qquad j=1,...,K.$$

Here $z(\cdot)$ is the quantile of $N(0,1)$ such that $P(N(0,1)>z(\alpha))=\alpha$.

Remark. If the variables are of different nature or measured in different units, there numerical variance are not comparable with each other. For example, one is height of students and another the exam score of students. The variance of height being larger than that of exam score cannot be interpreted as the former has larger variation than the other. Moreover, each variance may be inflated/deflated by changing the unit of the variable. Therefore, in these cases, one often consider *standardization*: for the observation of one variable, say x_{ij}, it is standardized to $(x_{ij}-\bar{X}_j)/s_j$, by subtracting the sample mean and then dividing by sample standard deviation. After standardization, the sample means are 0 and the sample variance matrix is the identity matrix. We note that P.C.s for the original data and for the standardized data may be different.

8.4. Examples.

Two examples, Examples 8.1-8.2, are presented with principal component analysis in Appendix B.