

Principal Component Analysis

Principal Component Analysis (PCA) and its variations are the most used multivariate methods.

Data: n observations of p -variate vectors.

Goal: Reduce the dimension p of the variables to a (much) less than p dimension by linear transformations, without loss of too much information in the data, measured in terms of variations, or variance. Hopefully the linear transformed variables have reasonable interpretations.

Classical PCA considers the case $n \geq p$. Newer methods have been developed to conduct PCA when $p > n$, especially the high dimensional case when p is large.

1 Population PCA construction

Often the method of Principle Component Analysis is applied to a dataset without any statistical assumption on the statistical model that data are generated from. The construction of sample principal components is thus a mathematical technique rather than a statistical model. However statistical idea is underneath the structure of principal components.

The approach

The n measurements of p -variate vectors are treated as n points in a p -dimensional space, where the p component variables form the p axes.

To derive the population principal components, the n observations are considered as n i.i.d. sample realizations from a p -variate random vector (X_1, \dots, X_p) . PCA procedure is to find mutually uncorrelated linear combinations Y_i (principal components) of the original variable X_j 's,

$$Y_i = \mathbf{a}'_i \mathbf{X} = a_{i1}X_1 + \dots + a_{ip}X_p = \sum_{j=1}^p a_{ij}X_j, \quad i = 1, \dots, p,$$

such that the variance $var(Y_i)$, $i = 1, \dots, p$ is orderly maximized. The hope is that the first few principal components carry most of variations in the data.

Assume that the p -variate vector $\mathbf{X} = (X_1, \dots, X_p) = [X_1 \dots X_p]'$ has covariance matrix $\Sigma_X = \Sigma$ with eigenvalues $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$. The method of Lagrange Multipliers is used sequentially to maximize $var(Y_i)$ subject to constraints $\sum_{j=1}^p a_{ij}^2 = \mathbf{a}'_i \mathbf{a}_i = 1$ and $cov(Y_i, Y_k) = 0$ for all $k < i$ (thus for all $k \neq i$).

Derivation. The following steps are to construct the the principal components Y_i , $i = 1, \dots, p$.

- Derivation of Principal component (PC) 1

Let $Y_1 = \mathbf{a}'_1 \mathbf{X}$. The goal is to find an optimal coefficient vector \mathbf{a}_1 such that $var(Y_1) = var(\mathbf{a}'_1 \mathbf{X}) = \mathbf{a}'_1 \Sigma \mathbf{a}_1$ is maximized subject to the constraint $\mathbf{a}'_1 \mathbf{a}_1 = 1$.

To find the optimal \mathbf{a}_1 , the method of Lagrange Multipliers sets

$$L_1 = \mathbf{a}'_1 \Sigma \mathbf{a}_1 - \lambda(\mathbf{a}'_1 \mathbf{a}_1 - 1)$$

where λ is a parameter (the Lagrange multiplier), its value is to be determined. The optimal \mathbf{a}_1 that maximizes $\mathbf{a}'_1 \Sigma \mathbf{a}_1$ must occur at a critical point $\frac{\partial L_1}{\partial \mathbf{a}_1} = 0_p$. The method of the Lagrange Multipliers is to find the values of λ at the critical points, then find \mathbf{a}_1 at the critical points.

Taking derivative of L_1 with respect to \mathbf{a}_1 . According to vector calculus,

$$\frac{\partial L_1}{\partial \mathbf{a}_1} = (\Sigma' + \Sigma)\mathbf{a}_1 - 2\lambda\mathbf{a}_1 = 2(\Sigma - \lambda I)\mathbf{a}_1$$

The last step uses the symmetry of Σ . Notice that finding the value of λ and its corresponding vector \mathbf{a}_1 such that

$$\frac{\partial L_1}{\partial \mathbf{a}_1} = (\Sigma - \lambda I)\mathbf{a}_1 = 0_p$$

is equivalent to finding eigenvalue value λ and corresponding eigenvectors \mathbf{a}_1 of Σ . By the positive semi-definiteness of Σ ,

$$|\Sigma - \lambda I| = 0 \implies \text{there are } n \text{ solutions } \lambda = \lambda_i, \quad \lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0.$$

For eigenvector \mathbf{a}_1 with eigenvalue λ , to maximize $Var(Y) = \mathbf{a}'_1 \Sigma \mathbf{a}_1 = \mathbf{a}'_1 \lambda \mathbf{a}_1 = \lambda \|\mathbf{a}_1\|^2$, we choose the largest eigenvalue

$$\lambda_1 = \max\{\lambda_i\}$$

Among all eigenvectors of Σ with eigenvalue λ_1 , we have chosen the one satisfying the normalizing condition (a.k.a. the constraint $\mathbf{a}'_1 \mathbf{a}_1 = 1$)

$$\mathbf{a}_1 : \Sigma \mathbf{a}_1 = \lambda_1 \mathbf{a}_1, \quad \mathbf{a}'_1 \mathbf{a}_1 = 1.$$

The chosen \mathbf{a}_1 maximizes $Var(Y_1) = \mathbf{a}'_1 \Sigma \mathbf{a}_1 = \mathbf{a}'_1 \lambda_1 \mathbf{a}_1 = \lambda_1$, among all $\mathbf{a}'_1 \mathbf{a}_1 = 1$.

Notice that \mathbf{a}_1 is not necessarily unique, and there can be ties in eigenvalue λ_i 's.

- Derivation of PC 2

The next step is to find \mathbf{a}_2 , $Y_2 = \mathbf{a}'_2 \mathbf{X}$ such that $Var(Y_2) = Var(\mathbf{a}'_2 \mathbf{X}) = \mathbf{a}'_2 \Sigma \mathbf{a}_2$ is maximized subject to the same normalizing constraint $\mathbf{a}'_2 \mathbf{a}_2 = 1$ and a new uncorrelatedness constraint $Cov(Y_2, Y_1) = 0$. When $\lambda_1 > 0$, the usual non-degenerate case,

$$Cov(Y_2, Y_1) = Cov(\mathbf{a}'_2 \mathbf{X}, \mathbf{a}'_1 \mathbf{X}) = \mathbf{a}'_2 \Sigma \mathbf{a}_1 = \mathbf{a}'_2 \lambda_1 \mathbf{a}_1 = \lambda_1 \mathbf{a}'_2 \mathbf{a}_1 = 0 \iff \mathbf{a}'_2 \mathbf{a}_1 = 0$$

Therefore to find an optimal \mathbf{a}_2 under the two constraints, the Lagrange method sets

$$L_2 = \mathbf{a}'_2 \Sigma \mathbf{a}_2 - \lambda(\mathbf{a}'_2 \mathbf{a}_2 - 1) - \delta(\mathbf{a}'_2 \mathbf{a}_1 - 0)$$

where λ, δ are Lagrange multipliers to be determined. Taking derivative of L_2 with respect to \mathbf{a}_2 ,

$$\frac{\partial L_2}{\partial \mathbf{a}_2} = 2\Sigma \mathbf{a}_2 - 2\lambda \mathbf{a}_2 - \delta \mathbf{a}_1 = 0_p$$

Multiply the above equation by \mathbf{a}'_1 from the left. Use the results $\Sigma \mathbf{a}_1 = \lambda_1 \mathbf{a}_1$ from the steps for PC1, and apply the constraint $\mathbf{a}'_1 \mathbf{a}_2 = 0$,

$$\begin{aligned} \mathbf{a}'_1 \frac{\partial L_2}{\partial \mathbf{a}_2} &= 2(\mathbf{a}'_1 \Sigma \mathbf{a}_2 - \mathbf{a}'_1 \lambda \mathbf{a}_2) - \delta \mathbf{a}'_1 \mathbf{a}_1 = 2((\Sigma \mathbf{a}_1)' \mathbf{a}_2 - \lambda \mathbf{a}'_1 \mathbf{a}_2) - \delta \mathbf{a}'_1 \mathbf{a}_1 = 2(\lambda_1 \mathbf{a}'_1 \mathbf{a}_2 - \lambda \mathbf{a}'_1 \mathbf{a}_2) - \delta = -\delta = 0 \\ &\implies \delta = 0. \end{aligned}$$

Thus

$$\frac{\partial L_2}{\partial \mathbf{a}_2} = 0 \iff (\Sigma - \lambda I)\mathbf{a}_2 = 0_p$$

Again, λ must be an eigenvalue of Σ , and \mathbf{a}_2 must be an eigenvector with eigenvalue λ . This time, choose the second largest eigenvalue λ_2 from the eigenvalue sequence $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$, then choose an eigenvector \mathbf{a}_2 of λ_2 with $\mathbf{a}'_2 \mathbf{a}_1 = 0$, $\mathbf{a}'_2 \mathbf{a}_2 = 1$.

- Derivation of PC k ($k = 3, \dots, p$)

The goal is to find $Y_k = \mathbf{a}'_k \mathbf{X}$ such that

- $Var(Y_k) = Var(\mathbf{a}'_k \mathbf{X}) = \mathbf{a}'_k \Sigma \mathbf{a}_k$ is maximized subject to the following constraints.
- The normalizing constraint $\mathbf{a}'_k \mathbf{a}_k = 1$
- the correlation constraint $Cov(Y_k, Y_i) = 0, i = 1, \dots, k-1$.

In the usual non-degenerate case, $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_{k-1} > 0$,

$$Cov(Y_k, Y_i) = \mathbf{a}'_k \Sigma \mathbf{a}_i = \mathbf{a}'_k \lambda_i \mathbf{a}_i = \lambda_i \mathbf{a}'_k \mathbf{a}_i = 0 \quad \Longleftrightarrow \quad \mathbf{a}'_k \mathbf{a}_i = 0, \quad i = 1, \dots, k-1.$$

Therefore to find an optimal \mathbf{a}_k under the constraints, the Lagrange method sets

$$L_k = \mathbf{a}'_k \Sigma \mathbf{a}_k - \lambda(\mathbf{a}'_k \mathbf{a}_k - 1) - \delta_1(\mathbf{a}'_k \mathbf{a}_1) - \dots - \delta_{k-1}(\mathbf{a}'_k \mathbf{a}_{k-1})$$

$$\frac{\partial L_k}{\partial \mathbf{a}_k} = 2(\Sigma - \lambda I)\mathbf{a}_k - \delta_1 \mathbf{a}_1 - \dots - \delta_{k-1} \mathbf{a}_{k-1} = 2\Sigma \mathbf{a}_k - 2\lambda \mathbf{a}_k - \delta_1 \mathbf{a}_1 - \dots - \delta_{k-1} \mathbf{a}_{k-1} = \mathbf{0}_p$$

Consecutively multiplying \mathbf{a}'_i from the left leads to $\delta_i = 0, i = 1, \dots, k-1$. Solving

$$\frac{\partial L_k}{\partial \mathbf{a}_k} = (\Sigma - \lambda I)\mathbf{a}_k = \mathbf{0}_p$$

to derive \mathbf{a}_k analogously, with $\Sigma \mathbf{a}_k = \lambda_k \mathbf{a}_k$.

□

Why PCA works?

Due to the symmetry and semi-positive-definiteness of covariance matrix Σ , linear algebra results grant that,

Eigenvectors corresponding to different eigenvalues λ_i of Σ are orthogonal.

An eigenvalue λ_i of Σ with k replications can correspond to k orthogonal eigenvectors.

Every eigenvalue λ_i is ≥ 0 by the positive-semi-definiteness of Σ .

Every eigenvalue λ_i is > 0 if Σ is positive-definite (non-degenerate).

Principal component coefficient vectors (or directions) \mathbf{a}_i 's are orthogonal eigenvectors of Σ ordered by the magnitude of the corresponding eigenvalues.

PCA properties

- Spectral decomposition of Σ

Denote the matrices

$$A = [\mathbf{a}_1 \ \mathbf{a}_2 \ \dots \ \mathbf{a}_p], \quad \Lambda = \text{diag}(\lambda_1, \dots, \lambda_p).$$

They are the eigenvector matrix and the eigenvalue matrix of Σ .

A is a $p \times p$ matrix. The constraints $\mathbf{a}'_i \mathbf{a}_j = 1_{\{i=j\}}$ means $AA' = A'A = I_p$, that is, A is an orthogonal matrix, $A' = A^{-1}$. By the PCA selection $\Sigma \mathbf{a}_k = \lambda_k \mathbf{a}_k, k = 1, \dots, p$,

$$\Sigma A = A\Lambda \quad \implies \quad \Sigma = A\Lambda A^{-1} = A\Lambda A'$$

Thus the covariance matrix Σ has an eigenvalue-eigenvector decomposition, and a spectral decomposition,

$$\Sigma = A\Lambda A' = \lambda_1 \mathbf{a}_1 \mathbf{a}'_1 + \dots + \lambda_p \mathbf{a}_p \mathbf{a}'_p$$

where the $p \times p$ matrix $\mathbf{a}_k \mathbf{a}'_k = P_{\mathbf{a}_k}$ is an orthogonal projection to one-dimensional subspace $\{c\mathbf{a}_k\}$.

- Decomposition and attribution of variance

From the derivation,

$$var(Y_i) = \lambda_i, \quad i = 1, \dots, p.$$

By the properties of trace of a matrix (the sum of the diagonal entries),

$$\text{trace}(\Sigma) = \text{trace}(A\Lambda A^{-1}) = \text{trace}(\Lambda A A^{-1}) = \text{trace}(\Lambda I) = \text{trace}(\Lambda)$$

Therefore the sum of the variance of the Y variables is the same as the sum of the variance of the X variables:

$$\sum_{i=1}^p var(Y_i) = \text{trace}(\Lambda) = \sum_{i=1}^p \lambda_i = \text{trace}(\Sigma) = \sum_{i=1}^p var(X_i)$$

The above sum is often referred as "the total variation" or "the total variance" of the data, of which the proportion of the variance of the j th principle component is

$$\frac{var(Y_j)}{\sum_{i=1}^p var(Y_i)} = \frac{\lambda_j}{\sum_{i=1}^p \lambda_i}$$

The above is often interpreted as

"the j th principal component accounts for a proportion $\frac{\lambda_j}{\sum_{i=1}^p \lambda_i}$ of the total variation in the original data."

The hope is that the first few principal components account for most variations, therefore it is possible to greatly reduce the dimensions of the original data without losing too much information.

- Scale dependence

PCA results depend on the scales of the variables. If the original data are used in PCA, the variables should be comparable, in meanings as well as in magnitude and spread. Otherwise variable transformations and scaling are needed.

- Variance-normalizing scaling

Variance normalizing the variables means replacing variable X_k by $X_k / \sqrt{V(X_k)}$ for all k , resulting in $V(X_k) = 1$ after scaling. PCA with variance normalized data corresponds to using $\Sigma = \text{Corr}(X)$ instead of $\text{Cov}(X)$ in the process of obtaining principal components Y_j . The diagonal entries in $\text{Corr}(X)$ are the variance of the variance-normalized variables, thus $\equiv 1$. Consequently,

$$\sum_{i=1}^p V(Y_i) = \text{trace}(\Lambda) = \sum_{i=1}^p \lambda_i = \text{trace}(\Sigma) = p$$

Compared with PCA on original data, PCA after variance-normalizing produces different λ_j 's and Y_j 's. The variance proportions $\lambda_j / \sum_{i=1}^p \lambda_i$'s are no longer the same. The PC loadings and PC scores are changed, too. Hence the interpretation should be adjusted accordingly.

- The correlation between the PC variable $Y_i = \mathbf{a}'_i \mathbf{X}$ and original variable X_k can be described by

$$\rho_{Y_i, X_k} = \text{corr}(Y_i, X_k) = a_{ik} \frac{\sqrt{\lambda_i}}{\sigma_k}, \quad \text{where} \quad \sigma_k^2 = \text{var}(X_k).$$

using

$$\text{cov}(Y_i, X_k) = \text{cov}(\mathbf{a}'_i \mathbf{X}, \mathbf{e}'_k \mathbf{X}) = \mathbf{a}'_i \Sigma \mathbf{e}_k = \lambda_i \mathbf{a}'_i \mathbf{e}_k = \lambda_i a_{ik}$$

where \mathbf{e}_k is a p -vector with k th element = 1, all other elements = 0. In common practice, sample variance $\hat{\sigma}_k^2 = s_{kk} = s_k^2$ is used for σ_k^2 .

- In practice, PCA may be conducted with a subset of variables instead of on all variables. The selection of included variables and adequacy of the choice depending on the purpose of the specific analysis.

2 Sample PCA

In applications, we do not know the true population mean $\boldsymbol{\mu}$ or the covariance matrix Σ . We use observed sample mean and sample covariance to approximate $\boldsymbol{\mu}$ and Σ , then carry out sample PCA as needed.

For a dataset with n observations of p -variate vectors, sample covariance matrix is used as an estimate of the population covariance matrix Σ : $\hat{\Sigma} = \mathbf{S}$; sample mean vector is used as an estimate of the population mean vector $\boldsymbol{\mu}$: $\hat{\boldsymbol{\mu}} = \bar{\mathbf{x}}$. Consequently, the eigenvalues of \mathbf{S} (or eigenvalues of sample correlation matrix \mathbf{R} , usually preferred) are estimates of the population eigenvalues: eigenvalues of $\mathbf{S} = \hat{\lambda}_i$.

Without any assumptions on the data structure, sample estimates are treated as the population parameters.

Sample PCA notations and terms

- \mathbf{X} is used to denote the $n \times p$ data matrix, its k th column consists of n measurements of the k th variable.

Variable loadings

PCs are represented as $Y_i = \mathbf{a}_i' \mathbf{X} = a_{i1}X_1 + \cdots + a_{ip}X_p$, where $\mathbf{a}_j, j = 1, \dots, p$ are eigenvectors of sample covariance matrix \mathbf{S} . a_{ik} 's are "variable loadings", forming a $p \times p$ loading matrix.

(Notation remarks: the loading matrix with \mathbf{a}_i as the i th column has (k, j) th element a_{jk} . So another way can be denoting $\mathbf{a}_i = [a_{i1} \cdots a_{ip}]'$)

PC scores

The j th observation (x_{j1}, \dots, x_{jp}) yields a value $y_{ji} = a_{i1}x_{j1} + \cdots + a_{ip}x_{jp}$ for each PC variable $Y_i, i = 1, \dots, p$. The y_{ji} 's are "PC scores", forming a $n \times p$ score matrix.

3 PCA in terms of Singular Value Decomposition*

Similar to many multivariate methods, principal component method is closed related to the Singular Value Decomposition (SVD) method in matrix algebra.

Let \mathbf{X} be the $n \times p$ data matrix denoting n observations of p -variate vectors. The corresponding p principal components can be displayed in a $n \times p$ matrix \mathbf{Y} . The relationship between Principal Component Analysis (PCA) and Singular Value Decomposition (SVD) can be described using centered data.

Let \mathbf{X}_c be the $n \times p$ matrix of the centered data with column mean zero. That is, replacing data x_{ji} by $x_{ji} - \bar{x}_i$.

Since

$$s_{ik} = \frac{1}{n-1} \sum_{j=1}^n (x_{ji} - \bar{x}_i)(x_{jk} - \bar{x}_k)$$

The sample covariance matrices for \mathbf{X} and \mathbf{X}_c are the same,

$$\mathbf{S} = \mathbf{S}_x = [s_{ik}]_{n \times n} = \mathbf{S}_{x_c} = \frac{1}{n-1} \mathbf{X}_c' \mathbf{X}_c$$

PCA and SVD

The Singular Value Decomposition (SVD) of $\mathbf{X}_c \in \mathbb{R}^{np}$ be written as

$$\mathbf{X}_c = \mathbf{U} \mathbf{D} \mathbf{V}'$$

where \mathbf{U}, \mathbf{V} are orthogonal matrices:

$$\mathbf{U}'\mathbf{U} = \mathbf{U}'\mathbf{U} = \mathbf{I}_n \in \mathbb{R}^n, \quad \mathbf{V}'\mathbf{V} = \mathbf{V}'\mathbf{V} = \mathbf{I}_p \in \mathbb{R}^p.$$

(\mathbf{U}, \mathbf{V} unitary if $\mathbf{X} \in \mathbb{C}^{np}$), and

$$\mathbf{D} = \mathbf{D}_{n \times p} \text{ with upper left } \text{diag}\{d_1, \dots, d_r\} \text{ and 0 entries otherwise, } d_1 \geq \dots \geq d_r \geq 0,$$

where

$$r = \min\{n, p\}, \quad r = p \quad \text{when } n \geq p. \quad \text{Denote } d_i = 0, \quad \text{for } i > r.$$

Then, the principal component scores for the centered data \mathbf{X}_c can be written as

$$\mathbf{Y}_c = \mathbf{X}_c \mathbf{V}$$

(The equality is up to orthogonal rotation and relabeling of orthonormal eigenvector \mathbf{v}_i 's sharing the same eigenvalue, see the proof below.)

Proof. By definition and by construction, the principal component scores for the centered data \mathbf{X}_c can be written as

$$\mathbf{Y}_c = \mathbf{X}_c [\mathbf{a}_1 \cdots \mathbf{a}_p] = \mathbf{X}_c \mathbf{V}^*$$

with

$$\mathbf{S} \mathbf{a}_i = \lambda_i \mathbf{a}_i, \quad \mathbf{a}_i' \mathbf{a}_j = \delta_{ij} = \begin{cases} 1, & i = j, \\ 0, & i \neq j, \end{cases} \quad \lambda_1 \geq \dots \geq \lambda_p \geq 0.$$

The \mathbf{a}_i 's are principal component direction vectors, or PC directions.

By construction, principal components are uncorrelated, and they are ordered by descending variance.

Therefore the sample covariance matrix of the PC scores \mathbf{Y}_c is diagonal with descendingly ordered diagonal entries:

$$\Lambda = \text{diag}\{\lambda_1, \dots, \lambda_p\} = \mathbf{S}_y = \widehat{\text{Cov}}(\mathbf{Y}) = \widehat{\text{Cov}}(\mathbf{X}_c \mathbf{V}^*) = \mathbf{V}^{*'} \widehat{\text{Cov}}(\mathbf{X}_c) \mathbf{V}^* = \mathbf{V}^{*'} \mathbf{S} \mathbf{V}^* \quad (1)$$

On the other hand, in the SVD of the centered data matrix \mathbf{X}_c ,

$$\mathbf{X}_c = \mathbf{U} \mathbf{D} \mathbf{V}',$$

by SVD construction, the columns in the $p \times p$ orthogonal matrix \mathbf{V} are eigenvectors of $\mathbf{X}_c' \mathbf{X}_c$ arranged by their eigenvalues $d_1 \geq \dots \geq d_p$ in descending order,

$$\mathbf{X}_c' \mathbf{X}_c \mathbf{V} = (\mathbf{V} \mathbf{D}' \mathbf{U}') (\mathbf{U} \mathbf{D} \mathbf{V}') \mathbf{V} = \mathbf{V} \mathbf{D}' \mathbf{D}.$$

The columns in the $n \times n$ orthogonal matrix \mathbf{U} are eigenvectors of $\mathbf{X}_c \mathbf{X}_c'$ arranged by their eigenvalues $d_1 \geq \dots \geq d_n$ in descending order,

$$\mathbf{X}_c \mathbf{X}_c' \mathbf{U} = (\mathbf{U} \mathbf{D} \mathbf{V}') (\mathbf{V} \mathbf{D}' \mathbf{U}') \mathbf{U} = \mathbf{U} \mathbf{D} \mathbf{D}'.$$

Then

$$\mathbf{V}' \mathbf{S} \mathbf{V} = \mathbf{V}' \left(\frac{1}{n-1} \mathbf{X}_c' \mathbf{X}_c \right) \mathbf{V} = \frac{1}{n-1} \mathbf{V}' (\mathbf{X}_c' \mathbf{X}_c \mathbf{V}) = \frac{1}{n-1} \mathbf{V}' \mathbf{V} \mathbf{D}' \mathbf{D} = \frac{1}{n-1} \mathbf{D}' \mathbf{D} = \frac{1}{n-1} \text{diag}\{d_1^2, \dots, d_p^2\}. \quad (2)$$

Comparing (1) and (2):

$$\text{From (1): } \mathbf{V}^{*'} \mathbf{S} \mathbf{V}^* = \text{diag}\{\lambda_1, \dots, \lambda_p\}$$

$$\text{From (2): } \mathbf{V}' \mathbf{S} \mathbf{V} = \frac{1}{n-1} \text{diag}\{d_1^2, \dots, d_p^2\}$$

both \mathbf{V} and \mathbf{V}^* are $p \times p$ orthonormal eigenvector matrices of \mathbf{S} ordered by the size of their eigenvalues, which indicates that, up to orthogonal rotation and relabeling of orthonormal eigenvectors sharing the same eigenvalue,

$$\mathbf{V}^* = \mathbf{V}$$

□

The above gives the connections between the principal components constructed from eigenvalue-eigenvectors of covariance matrix of the data and the singular value decomposition of the centered data matrix. Some properties are summarized below.

Properties and terminologies

- The eigenvalue λ_i of the sample covariance matrix $S = S_x$ and the singular value d_i of the centered data matrix X_c have the relation

$$\lambda_i = \frac{d_i^2}{n-1}$$

- Principal scores

The values of the principal component variables are called principal scores. They form a $n \times p$ matrix called score matrix. The score matrix of the principal components for the centered data X_c is given by

$$Y_c = X_c V^* = X_c V = U D \quad (3)$$

The i th row of Y_c gives the coordinates of the i th observation in the space with centered principal components as axes. U gives the score matrix standardized by column norm = 1. Thus

$$\sum_{j=1}^n y_{jk}^2 = \sum_{j=1}^n (d_k u_{jk})^2 = \sum_{j=1}^n \left(\sqrt{(n-1)\lambda_k} u_{jk} \right)^2 = (n-1)\lambda_k \sum_{j=1}^n u_{jk}^2 = (n-1)\lambda_k$$

For centered data, the mean of the k th principal component variable $\bar{y}_k = 0, k = 1, \dots, p$, which recovers what we are familiar with:

$$Var(Y_k) = \frac{1}{n-1} \sum_{j=1}^n (y_{jk} - \bar{y}_k)^2 = \frac{1}{n-1} \sum_{j=1}^n y_{jk}^2 = \lambda_k$$

Sometimes the standard scores (as in U) are standardized by the number of observations (e.g., multiplying $\sqrt{n-1}$). Then the standard score and multiplier matrix decomposition has to be adjusted accordingly.

- Principal directions

In (3) above, write

$$Y_c = [y_1 \dots y_p], \quad V = [v_1 \dots v_p]$$

Then the j th column, the set of n 'observed' values of the j th principal component is

$$y_j = X_c v_j$$

The vector $v_j \in \mathbb{R}^p$ with $\|v_j\| = 1$ gives the direction of the j th principal component in the space with the centered original variables as axes.

- Loadings

The components $v_{ik}, i = 1, \dots, p$ of principal directions v_k 's (or a_k 's in our derivation) are called PC loadings, which are contributions of the i th original variable to the k th principal component.

Loadings are often scaled depending on the need of the applications.

Plain PCA uses V as the loadings.

In some applications, e.g., factor analysis, the loadings are scaled by the standard deviation of the principal components, so the loading matrix has the form

$$[\sqrt{\lambda_1} v_1 \dots \sqrt{\lambda_p} v_p] = V \Lambda^{1/2} = \frac{1}{\sqrt{n-1}} V D$$

- Alternative SVD

When $n > p$, a reduced **thin SVD** for X_c gets rid of the null kernel basis by keeping only p of the n columns of U and keeping p of the n rows of D , while the right eigenvector matrix V stays the same:

$$X_c = U^* D^* V',$$

with

$$U^{*'} U^* = I_p, \quad V' V = V' V = I_p, \quad D = \text{diag}\{d_1, \dots, d_p\}, \quad d_1 \geq \dots \geq d_p \geq 0.$$

Then

$$Y_c = X_c V = U^* D^*,$$

- Centering

As stated in the above, the principal component directions (loadings, v_i or a_i 's) for the original X are the same ones for the centered data X_c , as listed in the columns of matrix V . They are the ordered orthonormal eigenvectors of S , up to orthogonal rotations and relabeling of orthonormal eigenvectors sharing the same eigenvalue.

$$Y = X V = X_c V + \bar{X} V = Y_c + \bar{Y}$$

- PCA on items

Just as PCA can be conducted with respect to variables as well as with respect to items, especially when $n \leq p$. Then SVD will be conducted on X'_c instead of X_c , with exchanged roles and interpretations of eigenvector matrices U and V , while the non-negative singular values stay the same.

4 Application and optimality

Dimension reduction optimal property

In application, PCA is a common first step in dimension reduction when analyzing large multivariate datasets with high dimensions (variables, features). The dimension reduction is carried out by using the first few principal components to represent the data.

The decision on how many principal components to retain is generally ad hoc. The amount of total sample variance explained or captured by the first few principal components is a major factor (as visualized in scree plots) alone with the relative sizes of the eigenvalues, and interpretations for the specific application.

In our derivation, we aim for maximizing the variance of the principal components. The process also minimize the average mean-square distance between the original data and their projections on to the principal component space

with lower dimensions, among linear projections to spaces with the same dimensions. In this linear projection sense, PCA is optimal.

Remarks: This is related to Eckart-Young theorem or Eckart-Young-Mirsky Theorem on approximation of a matrix by a lower rank one.

PCA and linear autoencoder

In machine learning, an autoencoder is a type of neural network used learn efficient ways to copy input to its output. For example, given an image of a handwritten letters or digits, an autoencoder first encodes or reduces the image into a lower dimensional representation (called latent representation), and then decodes or recovers the lower dimensional representation back to an image, hopefully close to the original one.

Thus PCA is essentially a linear encoder, the simplest one. Using fewer principal components to represent the data is the decoder part. Modern, non-linear autoencoders do a much better job than linear autoencoders.

PCA related references

- R. D. Cook (2018). Principal Components, Sufficient Dimension Reduction, and Envelopes. *Annu. Rev. Stat. Appl.* 2018. 5:533-59.
- Tipping and Bishop (1999). Probabilistic principal component analysis. *J. R. Stat. Soc. B.* 61: 611-622.
- Eckart and Young (1936). The approximation of one matrix by another of lower rank. *Psychometrika* v1: 211-218.

Note: Relevant chapter in the book by Johnson and Wichern: Chapter 8.

5 Appendix - Lagrange multipliers

The method of Lagrange Multipliers is a commonly used procedure to maximize a function $f(x_1, \dots, x_p)$ subject to constraints $g_i(x_1, \dots, x_p) = c_i$, when f, g are smooth functions. Under one constraint $g(x_1, \dots, x_p) = c$, the maximum must achieve at where some contour of $f(x_1, \dots, x_p)$ is tangent to $g(x_1, \dots, x_p) = c$. Therefore a necessary condition is

$$\left(\frac{\partial f}{\partial x_1}, \dots, \frac{\partial f}{\partial x_p} \right) = \lambda \left(\frac{\partial g}{\partial x_1}, \dots, \frac{\partial g}{\partial x_p} \right)$$

Component-wise,

$$\frac{\partial f}{\partial x_j} = \lambda \frac{\partial g}{\partial x_j}, \quad j = 1, \dots, p.$$

In other words,

$$\nabla f(x_1, \dots, x_p) = \lambda \nabla g(x_1, \dots, x_p)$$

In vector form,

$$\nabla f(\mathbf{x}) = \lambda \nabla g(\mathbf{x})$$

Similar tangential property holds with multiple constraints.

A Lagrangian L can be defined as

$$L = f(\mathbf{x}) - \sum_i \lambda_i [g_i(\mathbf{x}) - c_i] \quad (4)$$

L is a function of the variables $\mathbf{x} = (x_1, \dots, x_p)'$ and parameters λ_i 's, called Lagrange multipliers. Setting

$$\frac{\partial L}{\partial \mathbf{x}} = \mathbf{0}_p$$

and solve for λ_i 's yields candidates λ_i 's. Bring the candidate values to (4) and obtain the corresponding \mathbf{x} thus f . Compare and obtain the λ_i 's and \mathbf{x} at which $f(\mathbf{x})$ is maximized.

Notes : In the derivation of finding principal components by Lagrange multipliers, letter α is used instead of \mathbf{x} .