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 variable dimension p to a (much) less than p dimension by linear transformations, without too much information loss, measured in terms of total variable variations, or variance. Hopefully the dimension-reduced 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 in 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 mechanism the 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 get inference from sample data to a larger population, the n observations are considered as n independent sample realizations from a p-variate random vector $(X_1, \cdots, X_p)$ of the population.

The population principal component method is derived for the p-variate random vector.

Assume that the p-variate vector $\boldsymbol{X} = (X_1, \cdots, X_p) = [X_1 \ \cdots \ X_p]'$ has covariance matrix $\Sigma_{\mathrm{X}} = \Sigma$.

The population PCA procedure is to find mutually uncorrelated linear combinations Y_i of the original variable X_j's,

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

such that the variance $var(Y_i), i = 1, \cdots, p$ is orderly maximized.

The Y_i's are called principal component variables or principal components.

The hope is that the first few principal components carry most of the total variations in the original X_i's.

By the positive definite property of the covariance matrix Σ, we may order its eigenvalues as $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$.

The method of the Lagrange Multipliers can be used sequentially to maximize $var(Y_i)$ subject to the constraints $\|\boldsymbol{a}_i\|^2 = \boldsymbol{a}_i'\boldsymbol{a}_i = \sum_{j=1}^{n} a_{ij}^2 = 1$ and $cov(Y_i, Y_k) = 0$ for all $k < i$ (thus for all $k \neq i$), $k, i = 1, \cdots, p$.

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

- Derivation of the first principal component (PC) 1

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

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

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

where λ is a parameter (the Lagrange multiplier), its value is to be determined.

Fact: The optimal $\boldsymbol{a}_1$ that maximizes $\boldsymbol{a}_1'\Sigma\boldsymbol{a}_1$ must occur at a critical point $\frac{\partial L_1}{\partial \boldsymbol{a}_1} = 0_p$.

The method of the Lagrange Multipliers is to find the values of λ at the critical points, then find $\boldsymbol{a}_1$ at the critical points, such $\boldsymbol{a}_1$ maximizes $\boldsymbol{a}_1'\Sigma\boldsymbol{a}_1 = var(\boldsymbol{a}_1'\boldsymbol{X}) = var(Y_1)$ under the constraint.

To find critical points, taking derivative of L_1 with respect to $\boldsymbol{a}_1$. According to vector calculus,

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

The last step uses the symmetry of Σ.

Note that, finding the value of λ and its corresponding vector $\boldsymbol{a}_1$ such that

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

is equivalent to finding eigenvalue value λ and corresponding eigenvectors $\boldsymbol{a}_1$ of Σ.

By the positive semi-definiteness of Σ,

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

To maximize $Var(Y) = \boldsymbol{a}_1'\Sigma\boldsymbol{a}_1 = \boldsymbol{a}_1'\lambda\boldsymbol{a}_1 = \lambda\|\boldsymbol{a}_1\|^2 = \lambda$, we choose $\boldsymbol{a}_1$ to be an eigenvector corresponding to the largest eigenvalue

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

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

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

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

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

- Derivation of PC 2

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

$$Cov(Y_2, Y_1) = Cov(\boldsymbol{a}_2'\boldsymbol{X}, \boldsymbol{a}_1'\boldsymbol{X}) = \boldsymbol{a}_2'\Sigma\boldsymbol{a}_1 = \boldsymbol{a}_2'\lambda_1\boldsymbol{a}_1 = \lambda_1\boldsymbol{a}_2'\boldsymbol{a}_1 = 0 \qquad \Longleftrightarrow \qquad \boldsymbol{a}_2'\boldsymbol{a}_1 = 0$$

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

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

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

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

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

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

$$\implies \quad \delta = 0.$$

Thus

$$\frac{\partial L_2}{\partial \boldsymbol{a}_2} = 0 \quad \Longleftrightarrow \quad (\Sigma - \lambda I)\boldsymbol{a}_2 = 0_p$$

Again, λ must be an eigenvalue of Σ, and $\boldsymbol{a}_2$ must be an eigenvector with eigenvalue λ.

$$\Sigma\boldsymbol{a}_2 = \lambda\boldsymbol{a}_2$$

This time, to find an eigenvector maximizing

$$\boldsymbol{a}_2'\Sigma\boldsymbol{a}_2 = \boldsymbol{a}_2'\lambda\boldsymbol{a}_2 = \lambda\boldsymbol{a}_2'\boldsymbol{a}_2 = \lambda$$

under the constraints, especially that $\boldsymbol{a}_2$ must be orthogonal to $\boldsymbol{a}_1$, the solution $\boldsymbol{a}_2$ must be an eigenvector with the second largest eigenvalue λ_2 from the eigenvalue sequence $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$,

$$\boldsymbol{a}_2 : \Sigma\boldsymbol{a}_2 = \lambda_2\boldsymbol{a}_2, \qquad \boldsymbol{a}_2'\boldsymbol{a}_1 = 0,\ \boldsymbol{a}_2'\boldsymbol{a}_2 = 1.$$

Again, such $\boldsymbol{a}_2$ exists, not necessarily unique, and there can be ties in the eigenvalue λ_i's.

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

  The goal is to find $Y_k = \boldsymbol{a}_k'\boldsymbol{X}$ such that $Var(Y_k) = Var(\boldsymbol{a}_k'\boldsymbol{X}) = \boldsymbol{a}_k'\Sigma\boldsymbol{a}_k$ is maximized subject to
  - the normalizing constraint $\boldsymbol{a}_k'\boldsymbol{a}_k = 1$
  - the correlation constraints $Cov(Y_k, Y_i) = 0$ for all $i = 1, \cdots, k-1$.

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

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

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

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

  To find candidates of the maximizer, take derivative L_k with respect to $\boldsymbol{a}_k$,

  $$\frac{\partial L_k}{\partial \boldsymbol{a}_k} = 2\Sigma\boldsymbol{a}_k - 2\lambda\boldsymbol{a}_k - \delta_1\boldsymbol{a}_1 - \cdots - \delta_{k-1}\boldsymbol{a}_{k-1} = 0_p$$

  To find the critical values of the parameters λ and δ_i's, take the inner product of $\boldsymbol{a}_i'$ and the vector $\frac{\partial L_k}{\partial \boldsymbol{a}_k} = 0_p$, we obtain $k-1$ equations

  $$\boldsymbol{a}_i'\frac{\partial L_k}{\partial \boldsymbol{a}_k} = \boldsymbol{a}_i'\left(2\Sigma\boldsymbol{a}_k - 2\lambda\boldsymbol{a}_k - \delta_1\boldsymbol{a}_1 - \cdots - \delta_{k-1}\boldsymbol{a}_{k-1}\right) = 0, \qquad i = 1, \cdots, k-1.$$

  Analogous to the $k = 2$ case we proved in detail, under the constraints,

  $$\boldsymbol{a}_i'\frac{\partial L_k}{\partial \boldsymbol{a}_k} = \cdots = -\delta_i\boldsymbol{a}_i'\boldsymbol{a}_i = 0 \qquad \Rightarrow \quad \delta_i = 0$$

Therefore $\delta_i = 0$ for each and all $i = 1, \cdots, k-1$ under $\frac{\partial L_k}{\partial \boldsymbol{a}_k} = 0$ and the constraints. Then

$$\frac{\partial L_k}{\partial \boldsymbol{a}_k} = 2\Sigma\boldsymbol{a}_k - 2\lambda\boldsymbol{a}_k = 2(\Sigma - \lambda I)\boldsymbol{a}_k = 0_p$$

So, λ must be an eigenvalue of Σ, $\boldsymbol{a}_k$ must be an eigenvector with some eigenvalue λ. Solve for $\boldsymbol{a}_k$ under all of the constraints. Since $\boldsymbol{a}_k$ has to be orthogonal to the previous $k-1$ eigenvectors $\boldsymbol{a}_1, \cdots, \boldsymbol{a}_{k-1}$, to maximize $\boldsymbol{a}_k\Sigma\boldsymbol{a}_k$, the eigenvector $\boldsymbol{a}_k$ must satisfy

$$\Sigma\boldsymbol{a}_k = \lambda_k\boldsymbol{a}_k.$$

Therefore we have derived the kth PC variable $Y_k = \boldsymbol{a}_k'\boldsymbol{X}$ for $k = 1, \cdots, p$.

□

Why the PCs can be constructed

From the above derivation, the principal component coefficient vectors (also called principle component directions) $\boldsymbol{a}_i$'s are orthogonal eigenvectors of Σ ordered by the magnitude of the corresponding eigenvalues.

In the steps of construction, some results from linear algebra are applied repeatedly.

Thanks to the symmetry and semi-positive-definiteness of covariance matrix Σ, the properties of vector space and matrix algebra grant that,

Eigenvectors corresponding to different eigenvalues λ_i of Σ can be chosen to be 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).

These properties guarantee the existence of the principal component variables $Y_i = \boldsymbol{a}'\boldsymbol{X}$ as constructed.

PCA properties

- Spectral decomposition of Σ

  Denote the matrices

  $$A = [\boldsymbol{a}_1\ \boldsymbol{a}_2\ \cdots\ \boldsymbol{a}_p], \qquad \Lambda = diag(\lambda_1, \cdots, \lambda_p).$$

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

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

  $$\Sigma A = A\Lambda \qquad \Longrightarrow \qquad \Sigma = A\Lambda A^{-1} = A\Lambda A'$$

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

  $$\Sigma = A\Lambda A' = \lambda_1\boldsymbol{a}_1\boldsymbol{a}_1' + \cdots + \lambda_p\boldsymbol{a}_p\boldsymbol{a}_p'$$

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

- Decomposition and attribution of variance

  From the derivation,

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

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

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

  The above derivation have shown that, Y_i's are constructed so that

  $$Var(Y_k) = Var(\boldsymbol{a}_k'\boldsymbol{X}) = \boldsymbol{a}_k'\Sigma\boldsymbol{a}_k = \boldsymbol{a}_k'\lambda_k\boldsymbol{a}_k = \lambda_k\boldsymbol{a}_k'\boldsymbol{a}_k = \lambda_k$$

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

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

  The above sum is often referred as "the total variation" or "the total variance" of the original X_i's (thus the total variation of the data), of which the proportion of the variance of the jth 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 jth 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 in terms of variance.

- 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 $X_k/\sqrt{var(X_k)}$ for all k, resulting in $var(X_k) = 1$ after scaling. PCA with variance normalized data corresponds to using $\Sigma = Corr(X)$, the correlation matrix, instead of $Cov(X)$ in the process of obtaining principal components Y_j.

  The diagonal entries in $Corr(X)$ are the variance of the variance-normalized variables, thus $\equiv 1$.

  Consequently,

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

  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 $Y_i = \boldsymbol{a}_i'\boldsymbol{X}$ and original variable X_k can be described by

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

  To derive the above correlation, first compute the covariance.

  To use our knowledge of $Cov(\boldsymbol{X})$, Write both Y_i and X_i as linear combinations of $\boldsymbol{X}$,

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

  where $\boldsymbol{e}_k$ is a p-vector with kth element $= 1$, all other elements $= 0$. From the covariance we can compute the correlation,

  $$corr(Y_i, X_k) = \frac{cov(Y_i, X_k)}{\sqrt{var(Y_i)}\sqrt{var(X_k)}} = \frac{\lambda_i a_{ik}}{\sqrt{\lambda_i}\sqrt{\sigma_k^2}} = a_{ik}\frac{\sqrt{\lambda_i}}{\sigma_k}$$

- In practice, Σ is not known, thus all of its entries need to be estimated from the data.

- In practice, PCA may be conducted with a subset of variables instead of using all variables available.

  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 $\boldsymbol{\mu} = \boldsymbol{\mu}_X$ or the covariance matrix $\Sigma = \Sigma_X$. 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: $\hat{\Sigma} = \boldsymbol{S}$; sample mean vector is used as an estimate of the population mean vector μ: $\hat{\mu} = \bar{\boldsymbol{x}}$. Consequently, the eigenvalues of $\boldsymbol{S}$ (or eigenvalues of sample correlation matrix $\boldsymbol{R}$, usually preferred), denoted as $\hat{\lambda}_i$'s, are estimates of the trur λ_i's, the population eigenvalues.

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

Sample PCA notations and terms

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

- Variable loadings

  PCs are represented as $Y_i = \boldsymbol{a}_i'\boldsymbol{X} = a_{i1}X_1 + \cdots + a_{ip}X_p$, where $\boldsymbol{a}_j, j = 1, \cdots, p$ are eigenvectors of sample covariance matrix $\boldsymbol{S}$. a_{ik}'s are "variable loadings", forming a $p \times p$ loading matrix.
  (Notation remarks: the loading matrix with $\boldsymbol{a}_i$ as the ith column has (k, j)th element a_{jk}. So another way can be denoting $\boldsymbol{a}_i = [a_{i1} \ \cdots \ a_{ip}]'$)

- PC scores

  The jth observation $(x_{j1}, \cdots, x_{jp})$ yields a value $y_{ji} = a_{i1}x_{j1} + \cdots + a_{ip}x_{jp}$ for each PC variable $Y_i, i = 1, \cdots, p$. The y_{ji}'a 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 analysis (PCA) method is closed related to the Singular Value Decomposition (SVD) method in matrix algebra. For computation efficiency, in practice, the PCA is carried out via SVD on the data matrix, rather than using eigenvalue - eigenvector decomposition in the theoretical derivation.

Let $\boldsymbol{X}$ 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 $\boldsymbol{Y}$. The relationship between Principal Component Analysis (PCA) and Singular Value Decomposition (SVD) can be described using centered data.

Let $\boldsymbol{X}_c$ be the $n \times p$ matrix of the centered data with column mean zero, that is, replacing the original data x_{ji} by the centered data $x_{ji} - \bar{x}_i$.
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 the data matrix $\boldsymbol{X}$ and the centered data matrix $\boldsymbol{X}_c$ are the same,

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

PCA and SVD

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

$$\boldsymbol{X}_c = UDV'$$

where U, V are orthogonal matrices:

$$U'U = U'U = I_n \ \in \mathbb{R}^n, \qquad V'V = V'V = I_p \ \in \mathbb{R}^p.$$

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

$$D = D_{n\times p} \text{ with upper left } \ diag\{d_1, \cdots, d_r\} \ \text{ and 0 entries otherwise, } \quad d_1 \geq \cdots \geq d_r \geq 0,$$

where

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

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

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

The equality is up to orthogonal rotation and relabeling of orthonormal eigenvector v_i's sharing the same eigenvalue, as indicated in the proof below.

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

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

with

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

The $\boldsymbol{a}_i$'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 $\boldsymbol{Y}_c$ is diagonal with descendingly ordered diagonal entries:

$$\Lambda = diag\{\lambda_1, \cdots, \lambda_p\} = \boldsymbol{S}_y = \widehat{Cov}(\boldsymbol{Y}) = \widehat{Cov}(\boldsymbol{X}_c V^*) = V^{*\prime}\widehat{Cov}(\boldsymbol{X}_c)V^* = V^{*\prime}\boldsymbol{S}V^* \tag{1}$$

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

$$\boldsymbol{X}_c = UDV',$$

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

$$\boldsymbol{X}_c'\boldsymbol{X}_c V = (VD'U')(UDV')V = VD'D.$$

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

$$\boldsymbol{X}_c\boldsymbol{X}_c'U = (UDV')(VD'U')U = UDD'.$$

Then

$$\begin{aligned} V'\boldsymbol{S}V &= V'\left(\frac{1}{n-1}\boldsymbol{X}_c'\boldsymbol{X}_c\right)V = \frac{1}{n-1}V'\left(\boldsymbol{X}_c'\boldsymbol{X}_c V\right) \\ &= \frac{1}{n-1}V'VD'D = \frac{1}{n-1}D'D \\ &= \frac{1}{n-1}diag\{d_1^2, \cdots, d_p^2\} \end{aligned} \tag{2}$$

Comparing (1) and (2):

From (1): $V^{*\prime}\boldsymbol{S}V^* = diag\{\lambda_1, \cdots, \lambda_p\}$
From (2): $V'\boldsymbol{S}V = \frac{1}{n-1}diag\{d_1^2, \cdots, d_p^2\}$

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

$$V^* = V$$

□

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 $\boldsymbol{S} = \boldsymbol{S}_x$ and the singular value d_i of the centered data matrix $\boldsymbol{X}_c$ 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 $\boldsymbol{X}_c$ is given by

$$\boldsymbol{Y}_c = \boldsymbol{X}_c V^* = \boldsymbol{X}_c V = UD \tag{3}$$

The ith row of $\boldsymbol{Y}_c$ gives the coordinates of the ith 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 kth principal component variable $\bar{y}_k = 0, k = 1, \cdots, p$, which recovers what we are familiar with:

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

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 should be adjusted accordingly.

- Principal directions

  In (3) above, write

  $$Y_c = [y_1 \ \cdots \ y_p], \qquad V = [\boldsymbol{v}_1 \ \cdots \ \boldsymbol{v}_p]$$

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

  $$y_j = \boldsymbol{X}_c \boldsymbol{v}_j$$

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

- Loadings

  The components $v_{ik}, i = 1, \cdots, p$ of principal directions $\boldsymbol{v}_k$'s (or a_k's in our derivation) are called PC loadings, which are contributions of the ith original variable to the kth 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

  $$\left[\sqrt{\lambda_1}\boldsymbol{v}_1 \ \cdots \ \sqrt{\lambda_1}\boldsymbol{v}_p\right] = V\Lambda^{1/2} = \frac{1}{\sqrt{n-1}} VD$$

- Alternative SVD

  When $n > p$, a reduced **thin SVD** for $\boldsymbol{X}_c$ 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:

  $$\boldsymbol{X}_c = U^* D^* V',$$

  then

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

  Consequently,

  $$\boldsymbol{Y}_c = \boldsymbol{X}_c V = U^* D^*,$$

- Centering

  As stated in the above, the principal component directions (loadings, $\boldsymbol{v}_i$ or $\boldsymbol{a}_i$'s) for the original $\boldsymbol{X}$ are the same ones for the centered data $\boldsymbol{X}_c$, as listed in the columns of matrix V. They are the ordered orthonormal eigenvectors of $\boldsymbol{S}$, up to orthogonal rotations and relabeling of orthonormal eigenvectors sharing the same eigenvalue.

  $$\boldsymbol{Y} = \boldsymbol{X}V = \boldsymbol{X}_c V + \bar{\boldsymbol{X}}V = \boldsymbol{Y}_c + \bar{\boldsymbol{Y}}$$

- PCA on items

  Usually, PCA is conducted with respect to variables, resulting in variable transformations. If needed and sensible, PCA can be conducted with respect to items, especially when $n \leq p$. Then the SVD will be conducted on $(\boldsymbol{X}')_c$ instead of $\boldsymbol{X}_c$, with exchanged roles and interpretations of the eigenvector matrices U and V, while the non-negative singular values stay the same.

4 Application and optimality

Dimension reduction in practice

In application, PCA is a common first step in dimension reduction when analyzing large multivariate datasets with high dimensions (i.e., many variables or features). The dimension reduction procedure 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 the interpretations for the specific application.

PCA optimal property

In our derivation, we aim for maximizing the variance of the principal components sequentially. The process also minimize the average mean-square distance between the original data and their projections onto the lower dimensional subspace formed by the first few principal components, among all linear projections to subspaces with the same dimensions. In this sense of subspace linear projection, 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 matrix.

More formally, the best subspace approximation is in the sense of minimizing the Frobenius norm of the residual matrix. Utilizing the equivalence of PCA and SVD, when the data is represented by the top k PC's, the unexplained variance can be expressed as

$$\sum_{i=k+1}^{p} \lambda_i = (n-1) \sum_{i=k+1}^{p} d_i^2 = (n-1)\|X_c - X_c(k)\|_{Frobenius}^2$$

where $X_c(k) = U_k D V_k' = \sum_{i=1}^{k} \sigma_i^2 u_i v_i'$ is the k-dimensional subspace projection by the SVD, known to be optional among all k-dimensional subspace projections in terms of minimization of the residual matrix's Frobenius norm. The $(n-1)$ can be modified to 1 if we use the SVD of $\frac{1}{\sqrt{n-1}} X_c$ instead of X_c.

PCA and linear autoencoder

In machine learning, an autoencoder is a type of neural network that learn or train efficient ways to closely copy the 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 from the lower dimensional representation back to an image, hopefully very close to the original one.

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

Some PCA related historic 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 - Outline of the Method of Lagrange multiplier

The method of Lagrange Multipliers is a commonly used procedure to maximize a function $f(x_1, \cdots, x_p)$ subject to constraints $g_i(x_1, \cdots, x_p) = c_i$, when f, g are smooth functions.

For example, under one constraint $g(x_1, \cdots, x_p) = c$, the maximum must achieve at where some contour of $f(x_1, \cdots, x_p)$ is tangent to $g(x_1, \cdots, x_p) = c$. Therefore a necessary condition at the maximum is

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

Component-wise,

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

In differential operator nabla notation,

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

In vector form,

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

Similar tangential property holds with multiple constraints, assuming all derivatives exist and are continuous.

A Lagrangian or Lagrangian function L can be defined as

$$L = f(\boldsymbol{x}) - \sum_i \lambda_i \left[g_i(\boldsymbol{x}) - c_i\right] \tag{4}$$

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

To fine the maximum under the constraints, setting

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

and solve for λ_i's yields the values of candidate maximizers λ_i's.
(For complex problems, the solutions λ_i's can be hard or impossible to obtain. Then further methods are needed.)

If we can obtain the candidate maximizer λ_i's, bring them back to (4) and obtain the corresponding $\boldsymbol{x}$ thus f. Compare and obtain the λ_i's and $\boldsymbol{x}$ at which $f(\boldsymbol{x})$ is maximized.

Notes on notations : In our derivation of finding principal components by the Lagrange multipliers, letter $\boldsymbol{a}$ is used to denote the variable vector $\boldsymbol{x}$.