Multivariate Analysis
Principal Component Analysis

Shaobo Jin

Department of Mathematics

Intended Learning Outcome

Through this chapter, you should be able to

- derive PCA using matrix algebra
- conduct PCA

Motivation

Extract information from data, and achieve dimension reduction as an early step of an analytical process.

- A data set may contain a long list of variables.
- We want to reduce them to a smaller set of summary indices.
- Most of the information in the original set of variables are still preserved.

Task of Principal Component Analysis (PCA)

Rough Task

Let the random vector $\boldsymbol{X}$ $(p \times 1)$ have the covariance matrix $\boldsymbol{\Sigma} \geq 0$. Find linear combinations $Y_i = \boldsymbol{a}_i^T \boldsymbol{X}$ such that

$$
\begin{aligned}
&\boldsymbol{a}_1 \text{ maximizes var}\left(\boldsymbol{a}_1^T \boldsymbol{X}\right), \\
&\boldsymbol{a}_2 \text{ maximizes var}\left(\boldsymbol{a}_2^T \boldsymbol{X}\right), \text{ and cov}\left(\boldsymbol{a}_2^T \boldsymbol{X}, \boldsymbol{a}_1^T \boldsymbol{X}\right) = 0, \\
&\boldsymbol{a}_3 \text{ maximizes var}\left(\boldsymbol{a}_3^T \boldsymbol{X}\right), \text{ and cov}\left(\boldsymbol{a}_3^T \boldsymbol{X}, \boldsymbol{a}_j^T \boldsymbol{X}\right) = 0, \quad j < 3, \\
&\qquad \vdots \\
&\boldsymbol{a}_p \text{ maximizes var}\left(\boldsymbol{a}_p^T \boldsymbol{X}\right), \text{ and cov}\left(\boldsymbol{a}_p^T \boldsymbol{X}, \boldsymbol{a}_j^T \boldsymbol{X}\right) = 0, \quad j < p.
\end{aligned}
$$

Restriction

Consider the linear combination $Y_i = \boldsymbol{a}_i^T \boldsymbol{X}$. We have

$$\begin{aligned} \operatorname{var}(Y_i) &= \boldsymbol{a}_i^T \boldsymbol{\Sigma} \boldsymbol{a}_i, \\ \operatorname{cov}(Y_i, Y_k) &= \boldsymbol{a}_i^T \boldsymbol{\Sigma} \boldsymbol{a}_k, \quad i \neq k. \end{aligned}$$

Consider the new linear combination $Z_i = cY_i = c\boldsymbol{a}_i^T \boldsymbol{X}$, for a constant c. We have

$$\begin{aligned} \operatorname{var}(Z_i) &= c^2 \boldsymbol{a}_i^T \boldsymbol{\Sigma} \boldsymbol{a}_i, \\ \operatorname{cov}(Z_i, Z_k) &= c^2 \boldsymbol{a}_i^T \boldsymbol{\Sigma} \boldsymbol{a}_k. \end{aligned}$$

Hence, we need to set the scale such as $\boldsymbol{a}_i^T \boldsymbol{a}_i = 1$.

Principal Components

Principal components are particular linear combinations of the p random variables X_1, X_2, ..., X_p.

1. First principal component is the linear combination that maximizes $\text{var}\left(\boldsymbol{a}_1^T \boldsymbol{X}\right)$ subject to $\boldsymbol{a}_1^T \boldsymbol{a}_1 = 1$.
2. Second principal component is the linear combination that maximizes $\text{var}\left(\boldsymbol{a}_2^T \boldsymbol{X}\right)$ subject to $\boldsymbol{a}_2^T \boldsymbol{a}_2 = 1$ and $\text{cov}\left(\boldsymbol{a}_2^T \boldsymbol{X}, \boldsymbol{a}_1^T \boldsymbol{X}\right) = 0$
3. ith principal component is the linear combination that maximizes $\text{var}\left(\boldsymbol{a}_i^T \boldsymbol{X}\right)$ subject to $\boldsymbol{a}_i^T \boldsymbol{a}_i = 1$ and $\text{cov}\left(\boldsymbol{a}_i^T \boldsymbol{X}, \boldsymbol{a}_k^T \boldsymbol{X}\right) = 0$ for all $k < i$.

It is not required to have a multivariate normal assumption for $\boldsymbol{X}$.

Two Useful Lemma

Lemma

Consider the function $f(\boldsymbol{x}) = \boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x}$, where $\boldsymbol{x}$ is a vector. Then,

$$\frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}} = \left(\boldsymbol{A} + \boldsymbol{A}^T\right) \boldsymbol{x}.$$

If $\boldsymbol{A}$ is also symmetric, then,

$$\frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}} = 2\boldsymbol{A}\boldsymbol{x}.$$

Lemma

Let $\boldsymbol{A}$ and $\boldsymbol{B} > 0$ be two symmetric matrices. The maximum value of $\boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x}$ subject to $\boldsymbol{x}^T \boldsymbol{B} \boldsymbol{x} = 1$ is attained when $\boldsymbol{x}$ is the eigenvector of $\boldsymbol{B}^{-1}\boldsymbol{A}$ corresponding to the largest eigenvalue of $\boldsymbol{B}^{-1}\boldsymbol{A}$. Its maximum value is the largest eigenvalue of $\boldsymbol{B}^{-1}\boldsymbol{A}$.

Find Principal Components

In order to find the first principal component, we consider

$$\max \operatorname{var}\left(\boldsymbol{a}_1^T \boldsymbol{X}\right) \qquad \text{s.t. } \boldsymbol{a}_1^T \boldsymbol{a}_1 = 1.$$

In other words, we need to optimize

$$f\left(\boldsymbol{a}_1\right) = \boldsymbol{a}_1^T \boldsymbol{\Sigma} \boldsymbol{a}_1 - \lambda\left(\boldsymbol{a}_1^T \boldsymbol{a}_1 - 1\right).$$

In order to find the first principal component, we consider

$$\max \operatorname{var}\left(\boldsymbol{a}_2^T \boldsymbol{X}\right) \qquad \text{s.t. } \boldsymbol{a}_2^T \boldsymbol{a}_2 = 1,\ \operatorname{cov}\left(\boldsymbol{a}_2^T \boldsymbol{X}, \boldsymbol{a}_1^T \boldsymbol{X}\right) = 0.$$

In other words, we need to optimize

$$f\left(\boldsymbol{a}_2\right) = \boldsymbol{a}_2^T \boldsymbol{\Sigma} \boldsymbol{a}_2 - \lambda_1\left(\boldsymbol{a}_2^T \boldsymbol{a}_2 - 1\right), \quad \boldsymbol{a}_2 \notin \operatorname{span}\left\{\boldsymbol{a}_1\right\}.$$

Principal Components

Result 8.1: Simply An Eigen Decomposition

Let $\boldsymbol{\Sigma}$ be the covariance matrix associated with the $p \times 1$ random vector $\boldsymbol{X}$. Let $\boldsymbol{\Sigma}$ have the eigenvalue-eigenvector pairs $(\lambda_1, \boldsymbol{e}_1)$, ..., $(\lambda_p, \boldsymbol{e}_p)$, where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$. Then, the ith principal component is $Y_i = \boldsymbol{e}_i^T \boldsymbol{X}$. With these choices

$$\begin{aligned} \operatorname{var}(Y_i) &= \boldsymbol{e}_i^T \boldsymbol{\Sigma} \boldsymbol{e}_i = \lambda_i, \\ \operatorname{cov}(Y_i, Y_k) &= 0. \end{aligned}$$

Total Variation Explained by Principal Components

Result 8.2

Let $\boldsymbol{X}$ have covariance matrix $\boldsymbol{\Sigma}$ with eigenvalue-eigenvector pairs $(\lambda_1, \boldsymbol{e}_1)$, ..., $(\lambda_p, \boldsymbol{e}_p)$, where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$. Let $Y_i = \boldsymbol{e}_i^T \boldsymbol{X}$, $i = 1, ..., p$, be the unique principal components. Then,

$$\sum_{i=1}^{p} \sigma_{ii} = \sum_{i=1}^{p} \lambda_i = \sum_{i=1}^{p} \operatorname{var}(Y_i).$$

This result says that the total population variance, $\sum_{i=1}^{p} \operatorname{var}(X_i)$, is the same as the total principal component variance. Hence, the proportion of total variance explained by the kth principal component is

$$\frac{\lambda_k}{\sum_{i=1}^{p} \lambda_i}.$$

Importance

Result 8.3

If $Y_i = \boldsymbol{e}_i^T \boldsymbol{X}$, $i = 1, 2, .., p$, are the principal components obtained from the covariance matrix $\boldsymbol{\Sigma}$, then

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

are the correlation coefficients between the components Y_i and the variables X_k.

The magnitude of e_{ik} measures the importance of X_k to the ith principal component Y_i.

Principal Components From Correlation Matrix

Suppose that we standardize all X_i by $Z_i = (X_i - \mu_i) / \sqrt{\sigma_{ii}}$. All our previous results apply to $\operatorname{cov}(\boldsymbol{Z}) = \operatorname{corr}(\boldsymbol{X})$.

Result 8.4

The ith principal component of the standardized variables $\boldsymbol{Z}$ with $\operatorname{cov}(\boldsymbol{Z}) = \boldsymbol{\rho}$ is given by

$$Y_i = \boldsymbol{e}_i^T \boldsymbol{Z} = \boldsymbol{e}_i^T \boldsymbol{V}^{-1/2} (\boldsymbol{X} - \boldsymbol{\mu}).$$

Moreover, $\sum_{i=1}^p \operatorname{var}(Y_i) = \sum_{i=1}^p \operatorname{var}(Z_i) = p$, and $\rho_{Y_i, Z_k} = e_{ik}\sqrt{\lambda_i}$. In this case, $(\lambda_1, \boldsymbol{e}_1)$, ..., $(\lambda_p, \boldsymbol{e}_p)$ are the eigenvalue-eigenvector pairs for $\boldsymbol{\rho}$, with $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$.

Standardization Matters

The PC derived from $\boldsymbol{\Sigma}$ are different from those derived from $\boldsymbol{\rho}$. One set of PC is not a simple function of the other.

```r
A <- matrix(c(2.0, 0.5, 0.4,
              0.5, 1.5, 0.3,
              0.4, 0.3, 1.0), 3, 3, byrow = TRUE)
eigen(A)

## eigen() decomposition
## $values
## [1] 2.477083 1.195800 0.827117
##
## $vectors
##           [,1]       [,2]       [,3]
## [1,] 0.8000667  0.5626808 -0.2080470
## [2,] 0.5075924 -0.8197795 -0.2651631
## [3,] 0.3197549 -0.1065451  0.9414908
```

```r
D <- diag(1.0 / sqrt(c(2.0, 1.5, 1)))
D %*% A %*% D

##           [,1]      [,2]      [,3]
## [1,] 1.0000000 0.2886751 0.2828427
## [2,] 0.2886751 1.0000000 0.2449490
## [3,] 0.2828427 0.2449490 1.0000000

eigen(D %*% A %*% D)

## eigen() decomposition
## $values
## [1] 1.5447573 0.7552427 0.7000000
##
## $vectors
##            [,1]       [,2]       [,3]
## [1,] -0.5958111 -0.0463897  0.8017837
## [2,] -0.5700908 -0.6787568 -0.4629100
## [3,] -0.5656904  0.7328965 -0.3779645
```

Sample Principal Components

Suppose the data $\boldsymbol{x}_1$, ..., $\boldsymbol{x}_n$ represent n independent drawings from some p-dimensional population with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, not necessarily a multivariate population.

1. First sample principal component is the linear combination that maximizes the sample variance of $\boldsymbol{a}_1^T \boldsymbol{x}$ subject to $\boldsymbol{a}_1^T \boldsymbol{a}_1 = 1$.
2. Second sample principal component is the linear combination that maximizes the sample variance of $\boldsymbol{a}_2^T \boldsymbol{x}$ subject to $\boldsymbol{a}_2^T \boldsymbol{a}_2 = 1$ and zero sample covariance between $\boldsymbol{a}_2^T \boldsymbol{x}$ and $\boldsymbol{a}_1^T \boldsymbol{x}$.
3. ith sample principal component is the linear combination that maximizes the sample variance of $\boldsymbol{a}_i^T \boldsymbol{x}$ subject to $\boldsymbol{a}_i^T \boldsymbol{a}_i = 1$ and zero sample covariance between $\boldsymbol{a}_i^T \boldsymbol{x}$ and $\boldsymbol{a}_k^T \boldsymbol{x}$ for all $k < i$.

Apply Sample Covariance

By Result 2.5,

1. the linear combination $\boldsymbol{a}_i^T \boldsymbol{x}$ has sample mean $\boldsymbol{a}_i^T \bar{\boldsymbol{x}}$, and sample variance $\boldsymbol{a}_i^T \boldsymbol{S} \boldsymbol{a}_i$,
2. the sample covariance between $\boldsymbol{a}_i^T \boldsymbol{x}$ and $\boldsymbol{a}_k^T \boldsymbol{x}$ is $\boldsymbol{a}_i^T \boldsymbol{S} \boldsymbol{a}_k$.

Hence,

1. First sample principal component is the linear combination that maximizes $\boldsymbol{a}_1^T \boldsymbol{S} \boldsymbol{a}_1$ subject to $\boldsymbol{a}_1^T \boldsymbol{a}_1 = 1$.
2. Second sample principal component is the linear combination that maximizes $\boldsymbol{a}_2^T \boldsymbol{S} \boldsymbol{a}_2$ subject to $\boldsymbol{a}_2^T \boldsymbol{a}_2 = 1$ and zero sample covariance between $\boldsymbol{a}_2^T \boldsymbol{x}$ and $\boldsymbol{a}_1^T \boldsymbol{x}$.
3. ith sample principal component is the linear combination that maximizes $\boldsymbol{a}_i^T \boldsymbol{S} \boldsymbol{a}_i$ subject to $\boldsymbol{a}_i^T \boldsymbol{a}_i = 1$ and zero sample covariance between $\boldsymbol{a}_i^T \boldsymbol{x}$ and $\boldsymbol{a}_k^T \boldsymbol{x}$ for all $k < i$.

PCA

Result: Simply An Eigen Decomposition

If $\boldsymbol{S}$ is a $p \times p$ sample covarinace matrix with eigenvalue-eigenvector pairs $\left(\hat{\lambda}_i, \hat{\boldsymbol{e}}_i\right)$, $i = 1, ..., p$, the ith sample principal component is

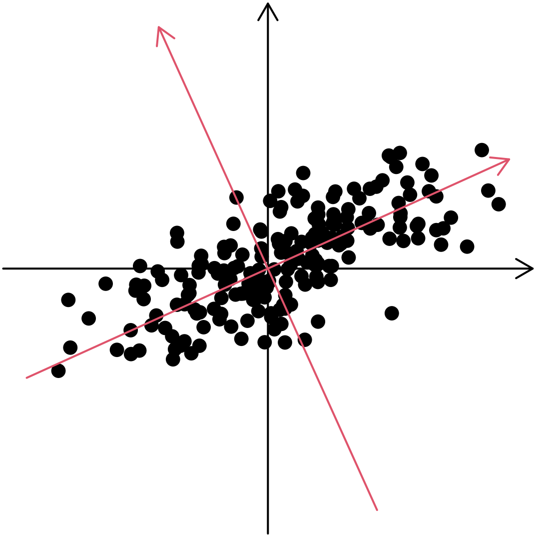
$$\hat{y}_i = \hat{\boldsymbol{e}}_i^T \boldsymbol{x},$$

where $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_p \geq 0$. The values $\begin{pmatrix} \hat{y}_1 & \cdots & \hat{y}_p \end{pmatrix}$ are the principal component scores. The sample variance of $\hat{y}_i$ is $\hat{\lambda}_k$, and the sample covariance between $\hat{y}_i$ and $\hat{y}_k$ is 0. In addition, the total sample variance satisfies $\sum_{i=1}^p s_{ii} = \sum_{i=1}^p \hat{\lambda}_i$, and

$$r_{\hat{y}_i, x_k} = \frac{\hat{e}_{ik}\sqrt{\hat{\lambda}_i}}{\sqrt{s_{kk}}}.$$

Further, the PCs from $\boldsymbol{S}$ and the sample correlation matrix are not the same.

Rotation of Axes



Eigen Decomposition and Singular Value Decomposition

PCA is simply eigen decomposition of $\boldsymbol{S}$ as $\boldsymbol{S} = \boldsymbol{E}\boldsymbol{\Lambda}\boldsymbol{E}^T$.

- The eigen decomposition is only applicable to a symmetric matrix.

Every matrix has a singular value decomposition (SVD). For any $m \times n$ matrix $\boldsymbol{A}$ of rank r, there exist an $m \times m$ orthogonal matrix $\boldsymbol{U}$ and an $n \times n$ orthogonal matrix $\boldsymbol{V}$ such that

$$\boldsymbol{A}_{m\times n} = \boldsymbol{U}_{m\times m} \begin{pmatrix} \boldsymbol{D}_1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{pmatrix}_{m\times n} \boldsymbol{V}_{n\times n}^T,$$

where $\boldsymbol{D}_1$ is an $r \times r$ diagonal matrix with diagonal elements that are positive.

- Diagonal elements in $\begin{pmatrix} \boldsymbol{D}_1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{pmatrix}$ are called singular values.

Find Singular Value Decomposition

1. The diagonal elements in $\boldsymbol{D}_1$ are the positive square root of the nonzero eigenvalues (not necessarily distinct) of $\boldsymbol{A}^T\boldsymbol{A}$ or $\boldsymbol{A}\boldsymbol{A}^T$.
2. Let the columns of $\boldsymbol{V}_1$ be the eigenvectors corresponding to the nonzero eigenvalues of $\boldsymbol{A}^T\boldsymbol{A}$, and the columns of $\boldsymbol{V}_2$ be the eigenvectors corresponding to the 0 eigenvalues. Then $\boldsymbol{V} = \begin{bmatrix} \boldsymbol{V}_1 & \boldsymbol{V}_2 \end{bmatrix}$.
3. Let the columns of $\boldsymbol{U}_1$ be the eigenvectors corresponding to the nonzero eigenvalues of $\boldsymbol{A}\boldsymbol{A}^T$, and the columns of $\boldsymbol{U}_2$ be the eigenvectors corresponding to the 0 eigenvalues. Then $\boldsymbol{U} = \begin{bmatrix} \boldsymbol{U}_1 & \boldsymbol{U}_2 \end{bmatrix}$.

The SVD is

$$\boldsymbol{A} = \boldsymbol{U} \begin{pmatrix} \boldsymbol{D}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \boldsymbol{V}^T = \boldsymbol{U}_1 \boldsymbol{D}_1 \boldsymbol{V}_1^T.$$

And $\boldsymbol{U}_1$ actually must satisfy $\boldsymbol{U}_1 = \boldsymbol{A}\boldsymbol{V}_1\boldsymbol{D}_1^{-1}$.

SVD in R

```
A <- matrix(c(2, 0, 1, 0, 1, 2), 3, 2)
edV <- eigen(t(A) %*% A)
D1 <- diag(sqrt(edV$values))
D <- rbind(D1, matrix(0, 1, 2))
V <- edV$vectors
edU <- eigen(A %*% t(A))
U <- edU$vectors
round(cbind(U %*% D %*% t(V), NA, U %*% D %*% t(-1.0 * V), NA,
            (A %*% V %*% solve(D1)) %*% D1 %*% t(V)), 6)
```

```
##      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8]
## [1,]   -2    0   NA    2    0   NA    2    0
## [2,]    0   -1   NA    0    1   NA    0    1
## [3,]   -1   -2   NA    1    2   NA    1    2
```

SVD in R

```
A <- matrix(c(2, 0, 1, 0, 1, 2), 3, 2)
svd(A)

## $d
## [1] 2.645751 1.732051
##
## $u
##            [,1]       [,2]
## [1,] -0.5345225  0.8164966
## [2,] -0.2672612 -0.4082483
## [3,] -0.8017837 -0.4082483
##
## $v
##            [,1]       [,2]
## [1,] -0.7071068  0.7071068
## [2,] -0.7071068 -0.7071068
```

PCA and SVD

Suppose that we have a data matrix $\boldsymbol{X}$ that has been demeaned. $\boldsymbol{X}$ can be SVD decomposed as

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T \Rightarrow \boldsymbol{X}\boldsymbol{V} = \boldsymbol{U}\boldsymbol{D},$$

where the m columns of $\boldsymbol{U}$ are orthonormal eigenvectors of $\boldsymbol{X}\boldsymbol{X}^T$. Then, we should have

$$\boldsymbol{X}^T\boldsymbol{X}/(n-1) = \boldsymbol{V}\boldsymbol{D}\boldsymbol{U}^T\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T/(n-1) = \boldsymbol{V}\left[\boldsymbol{D}^2/(n-1)\right]\boldsymbol{V}^T.$$

- the PCA loading matrix: $\boldsymbol{V}$.
- the principal component scores: $\boldsymbol{U}\boldsymbol{D}$.
- the variances of principal components: $\boldsymbol{D}^2/(n-1)$.

SVD and PCA

```
dx <- x - matrix(colMeans(x), nrow = N, ncol = 2, byrow = TRUE)
sqrt( (svd(dx)$d^2) / (N - 1) )

## [1] 0.7600400 0.3254688

svd(dx)$v

##            [,1]       [,2]
## [1,] -0.8377819 -0.5460050
## [2,] -0.5460050  0.8377819

prcomp(x)

## Standard deviations (1, .., p=2):
## [1] 0.7600400 0.3254688
##
## Rotation (n x k) = (2 x 2):
##             PC1        PC2
## [1,] -0.8377819 -0.5460050
## [2,] -0.5460050  0.8377819
```

How Many Components?

There is no definite answer on how many PCs we should choose. Some popular methods are

- Scree plot: the point (elbow) before where the curve flattens.
- Choose the number of PCs such that a specified percentage of variance been explained.