Lecture 10: Principal Component Analysis

Hao Helen Zhang

Lecture 10: Principal Component Analysis

Hao Helen Zhang

Motivations

The principal component analysis (PCA) is concerned with explaining the variance-covariance structure of $\mathbf{X} = (X_1, \cdots, X_p)'$ through a few linear combinations of these variables.

- Main purposes:
  - data (dimension) reduction
  - interpretation
- Easy to visualize

Variance-Covariance Matrix of Random Vector

Define the random vector and its mean vector

$$\mathbf{X} = (X_1, \cdots, X_p)', \quad \boldsymbol{\mu} = E(\mathbf{X}) = (\mu_1, \cdots, \mu_p)'.$$

The variance-covariance matrix of $\mathbf{X}$ is the

$$\Sigma = \mathrm{Cov}(\mathbf{X}) = E(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})',$$

its *ij*-th entry $\sigma_{ij} = E(X_i - \mu_i)(X_j - \mu_j)$ for any $1 \leq i \leq j \leq p$.

- $\boldsymbol{\mu}$ is the *population mean*
- Σ is the *population variance-covariance* matrix.
- In practice, $\boldsymbol{\mu}$ and Σ are unknown and estimated from the data.

Sample Variance-Covariance Matrix

- Sample mean:

$$\bar{\mathbf{X}} = \frac{1}{n} X' \mathbf{1}_n,$$

  X is the design matrix, and $\mathbf{1}_n$ is the vector of 1' of length n.

- (Unbiased) Sample variance-covariance matrix

$$S_n = \frac{1}{n-1} X_c' X_c = \frac{1}{n-1} \sum_{i=1}^{n} (\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})',$$

  where X_c the centered design matrix, and $\mathbf{X}_i = (X_{i1}, \cdots, X_{ip})'$ for $i = 1, \cdots, n$.

It is easy to show that

$$S_n = \frac{1}{n-1} X' (I_n - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n') X.$$

Linear Combinations of Inputs

Consider the linear combinations

$$\begin{aligned} Z_1 &= \mathbf{v}_1'\mathbf{X} = v_{11}X_1 + v_{12}X_2 + \cdots + v_{1p}X_p, \\ Z_2 &= \mathbf{v}_2'\mathbf{X} = v_{21}X_1 + v_{22}X_2 + \cdots + v_{2p}X_p, \\ \cdots &= \cdots \\ Z_p &= \mathbf{v}_p'\mathbf{X} = v_{p1}X_1 + v_{p2}X_2 + \cdots + v_{pp}X_p. \end{aligned}$$

Then

$$Var(Z_j) = \mathbf{v}_j'\Sigma\mathbf{v}_j, \quad j = 1, \cdots, p.$$

$$Cov(Z_j, Z_k) = \mathbf{v}_j'\Sigma\mathbf{v}_k, \quad \forall j \neq k.$$

What is PCA

Principal component analysis (PCA, Pearson 1901) is a statistical procedure that

- uses an **orthogonal** transformation to convert a set of observations of correlated variables into a set of linearly uncorrelated variables (called principal components)
- finds directions with maximum variability

Principal components (PCs):

- PCs are *uncorrelated*, *orthogonal*, linear combinations $Z_1, \cdots, Z_p$ whose variances are as large as possible.
- PCs form a new coordinate system by rotating the original system constructed by $X_1, \cdots, X_p$

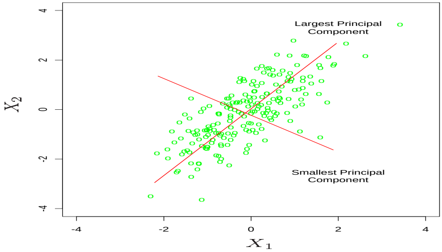


Figure 3.8: *Principal components of some input data points. The largest principal component is the direction that maximizes the variance of the projected data, and the smallest principal component minimizes that variance. Ridge regression projects* $\mathbf{y}$ *onto these components, and then shrinks the coefficients of the low-variance components more than the high-variance components.*

Mathematical Formulation

The procedure seeks the direction of high variances:

- The first PC = linear combination $Z_1 = \mathbf{v}_1'\mathbf{X}$ that maximizes $\text{Var}(\mathbf{v}_1'\mathbf{X})$ subject to $\|\mathbf{v}_1\| = 1$.
- The second PC = linear combination $Z_2 = \mathbf{v}_2'\mathbf{X}$ that maximizes $\text{Var}(\mathbf{v}_2'\mathbf{X})$ subject to $\|\mathbf{v}_2\| = 1$ and $\text{Cov}(\mathbf{v}_1'\mathbf{X}, \mathbf{v}_2'\mathbf{X}) = 0$
- The jth PC satisfies

$$\begin{aligned} \max \quad & \text{Var}(\mathbf{v}_j'\mathbf{X}) \\ \text{subject to} \quad & \|\mathbf{v}_j\| = 1, \\ & \text{Cov}(\mathbf{v}_l'\mathbf{X}, \mathbf{v}_j' X \mathbf{v}_j) = \mathbf{v}_l'\Sigma\mathbf{v}_j = 0, \\ & \text{for} \quad l = 1, ..., j-1, \end{aligned}$$

where $j = 2, \cdots, p$.

Interpretation of PCA

- $\mathbf{Z}_1 = \mathbf{v}_1'\mathbf{X}$ has the largest sample variance among all normalized linear combinations of the columns of $\mathbf{X}$.
- $\mathbf{Z}_2 = \mathbf{v}_2'\mathbf{X}$ has the highest variance among all normalized liner combinations of the columns of $\mathbf{X}$, satisfying $\mathbf{v}_2$ orthogonal to $\mathbf{v}_1$.
- ......
- The last PC $\mathbf{Z}_p = \mathbf{v}_p'\mathbf{X}$ has the minimum variance among all normalized linear combinations of the columns of $\mathbf{X}$, subject to $\mathbf{v}_p$ being orthogonal to the earlier ones.

If Σ is unknown, we use S_n as its estimator.

How to Solve PCs

There are two ways:

- eigen-decomposition of Σ
- singular value decomposition (SVD) of X_c.

Comment:

- Efficient algorithms exist to calculate SVD of X without computing $X^T X$
- Computing SVD is now the standard way to calculate PCA from a data matrix

Eigen-Decomposition of Σ

Assume Σ has p eigenvalue-eigenvector pairs $(\lambda, \mathbf{e})$ satisfying:

$$\Sigma \mathbf{e}_j = \lambda_j \mathbf{e}_j, \quad j = 1 \cdots, p,$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0$ and $\|\mathbf{e}_j\| = 1$ for all j. This gives the following *spectral* decomposition

$$\Sigma = \sum_{j=1}^{p} \lambda_j \mathbf{e}_j \mathbf{e}_j'.$$

- The jth PC is given by $Z_j = \mathbf{e}_j' \mathbf{X}$ and its variance is

$$\text{Var}(Z_j) = \mathbf{e}_j' \Sigma \mathbf{e}_j = \lambda_j.$$

- The magnitude of e_{jk} measures the importance of the kth variable to the jth PC, irrespective of the other variables.

Number of PCs

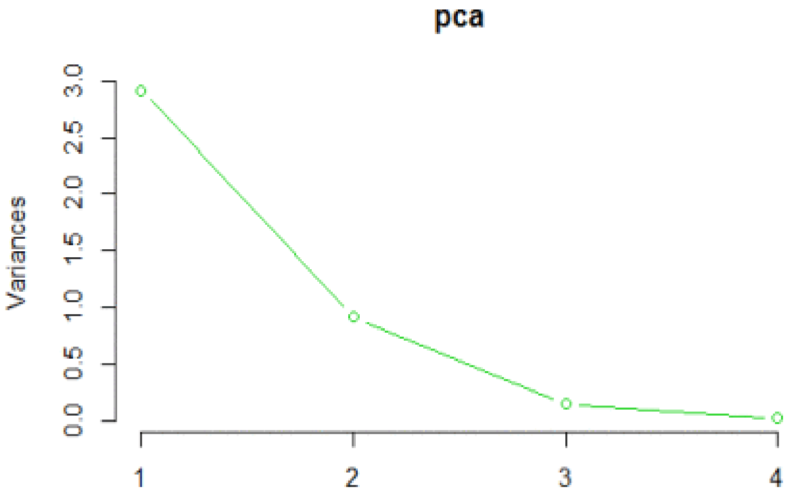
- The total (population) variance of inputs

$$\sum_{j=1}^{p} \mathrm{Var}(X_j) = \sum_{j=1}^{p} \sigma_{jj} = \sum_{j=1}^{p} \lambda_j = \sum_{j=1}^{p} \mathrm{Var}(Z_j).$$

- Proportion of total variance due to the jth PC $\frac{\lambda_j}{\sum_{k=1}^{p} \lambda_k}$.

The number of PCs are decided based on

- the amount of total sample variance explained, the variances of the sample PC, and the subject-matter interpretations
- the *scree* plot – plot the ordered eigenvalues $\lambda_1, \cdots, \lambda_p$ and look for the elbow (bend) in the plot. The number of PCs is the point where the remaining eigenvalues are relatively small and all about the same size.


pca
Variances
0.0
0.5
1.0
1.5
2.0
2.5
3.0
1
2
3
4

Wide Applications

PCA is very useful in exploratory data analysis.

- provide a simpler and more parsimonious description of the covariance structure
- dimension reduction
- visualization for high-dimensional data

Applications:

- in signal processing, called discrete KLT transform
- in linear algebra, called eigenvalue decomposition (EVD) of $X^T X$.
- Golub and Van Loan (1983), called singular value decomposition (SVD) of X.
- in noise and vibration, called spectral decomposition.

Further Remarks

Remarks:

- PCs are solely determined by the covariance matrix Σ.
- The PCA analysis does not require a multivariate normal distribution.

Concerns:

- unsupervised learning
- ignore the response