Lecture 10: Principal Component Analysis

Hao Helen Zhang
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
Define the random vector and its mean vector

\[ \mathbf{X} = (X_1, \cdots, X_p)', \quad \mu = E(\mathbf{X}) = (\mu_1, \cdots, \mu_p)' \]

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

\[ \Sigma = \text{Cov}(\mathbf{X}) = E(\mathbf{X} - \mu)(\mathbf{X} - \mu)', \]

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

- \( \mu \) is the population mean
- \( \Sigma \) is the population variance-covariance matrix.
- In practice, \( \mu \) and \( \Sigma \) are unknown and estimated from the data.
Sample Variance-Covariance Matrix

- Sample mean:
  \[ \bar{X} = \frac{1}{n} X'1_n, \]
  
  \( X \) is the design matrix, and \( 1_n \) is the vector of 1' of length \( n \).

- (Unbiased) Sample variance-covariance matrix
  \[
  S_n = \frac{1}{n-1} X_c'X_c = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})(X_i - \bar{X})',
  \]
  
  where \( X_c \) the centered design matrix, and
  \( X_i = (X_{i1}, \cdots, X_{ip})' \) for \( i = 1, \cdots, n \).

  It is easy to show that
  \[
  S_n = \frac{1}{n-1} X'(I_n - \frac{1}{n} 1_n1_n')X.
  \]
Consider the linear combinations

\[ Z_1 = v_1'X = v_{11}X_1 + v_{12}X_2 + \cdots + v_{1p}X_p, \]
\[ Z_2 = v_2'X = v_{21}X_1 + v_{22}X_2 + \cdots + v_{2p}X_p, \]
\[ \vdots \]
\[ Z_p = v_p'X = v_{p1}X_1 + v_{p2}X_2 + \cdots + v_{pp}X_p. \]

Then

\[ \text{Var}(Z_j) = v_j'\Sigma v_j, \quad j = 1, \cdots, p. \]
\[ \text{Cov}(Z_j, Z_k) = v_j'\Sigma v_k, \quad \forall j \neq k. \]
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.
The procedure seeks the direction of high variances:

- The first PC = linear combination $Z_1 = v_1'X$ that maximizes $\text{Var}(v_1'X)$ subject to $\|v_1\| = 1$.

- The second PC = linear combination $Z_2 = v_2'X$ that maximizes $\text{Var}(v_2'X)$ subject to $\|v_2\| = 1$ and $\text{Cov}(v_1'X, v_2'X) = 0$.

- The $j$th PC satisfies

  \[
  \max \quad \text{Var}(v_j'X) \\
  \text{subject to} \quad \|v_j\| = 1, \\
  \text{Cov}(v_i'X, v_j'Xv_j) = v_i'\Sigma v_j = 0, \\
  \text{for} \quad l = 1, \ldots, j - 1, \\
  \]

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

  [Hao Helen Zhang  Lecture 10: Principal Component Analysis 9/16]
Principal Component Analysis

Interpretation of PCA

- \( Z_1 = v_1'X \) has the largest sample variance among all normalized linear combinations of the columns of \( X \).
- \( Z_2 = v_2'X \) has the highest variance among all normalized linear combinations of the columns of \( X \), satisfying \( v_2 \) orthogonal to \( v_1 \).
- ......
- The last PC \( Z_p = v_p'X \) has the minimum variance among all normalized linear combinations of the columns of \( X \), subject to \( v_p \) being orthogonal to the earlier ones.

If \( \Sigma \) is unknown, we use \( S_n \) as its estimator.
How to Solve PCs

There are two ways:

- eigen-decomposition of $\Sigma$
- singular value decomposition (SVD) of $X_c$.

Comment:

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

Assume $\Sigma$ 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 $j$th 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 $k$th variable to the $j$th PC, irrespective of the other variables.
Number of PCs

- The total (population) variance of inputs
  \[ \sum_{j=1}^{p} \text{Var}(X_j) = \sum_{j=1}^{p} \sigma_{jj} = \sum_{j=1}^{p} \lambda_j = \sum_{j=1}^{p} \text{Var}(Z_j). \]

- Proportion of total variance due to the $j$th 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.
Principal Component Analysis

Variance graph showing the variance of principal components.
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 $\Sigma$.
- The PCA analysis does not require a multivariate normal distribution.

Concerns:
- unsupervised learning
- ignore the response