MAT 653: Statistical Simulation

Instructor: Dr. Wei Li

2023-09-02

Back **Updated:** 2023-09-02 Statistical Simulation, Wei Li

PCA

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

Define the random vector and its mean vector

$$X = (X_1, \cdots, X_p)^T, \quad \boldsymbol{\mu} = E(X) = (\mu_1, \cdots, \mu_p)^T$$

The variance-covariance matrix of X is the

$$\Sigma = \operatorname{Cov}(X) = E(X - \boldsymbol{\mu})(X - \boldsymbol{\mu})^T$$

its ij -th entry $\sigma_{ij} = E\left(X_i - \mu_i\right)\left(X_j - \mu_j\right)$ for any $1 \leq i \leq j \leq p$, where $\boldsymbol{\mu}$ is the population mean and Σ is the population variance-covariance matrix. In practice, $\boldsymbol{\mu}$ and Σ are unknown and estimated from the data.

Consider the linear combinations

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

In a matrix form,

$$Z = V^T X$$

where $Z = (Z_1, \ldots, Z_p)^T$, $V = (\mathbf{v}_1, \ldots, \mathbf{v}_p)$. Then

$$\begin{aligned}
\operatorname{Var}\left(Z_j\right) &= \mathbf{v}_j^T \Sigma \mathbf{v}_j, \quad j = 1, \cdots, p \\
\operatorname{Cov}\left(Z_j, Z_k\right) &= \mathbf{v}_j^T \Sigma \mathbf{v}_k, \quad \forall j \neq k
\end{aligned}$$

Principal component analysis is a statistical procedure that finds directions $(\mathbf{v}_j, j = 1, \ldots, p)$ such that $Z_1, \cdots, Z_p$ has maximum variability and also are linearly uncorrelated.

Principal components (PCs):

- The **principal components** of X are (linearly) uncorrelated, linear combinations $Z_1, \cdots, Z_p$ whose variances are as large as possible.
- The collection $\{\mathbf{v}_1, \ldots, \mathbf{v}_p\}$ are called **principal directions** of X. Each entry in $\mathbf{v}_j$ is called a **PC loading** in the j-th direction. The magnitude of the k-entry of $\mathbf{v}_j$ measures the importance of the k th variable to the j th PC, irrespective of the other variables.

The procedure can be described as

- The first PC = linear combination $Z_1 = \mathbf{v}_1^T X$ that maximizes $\operatorname{Var}\left(\mathbf{v}_1^T X\right)$ subject to $\|\mathbf{v}_1\| = 1$
- The second PC = linear combination $\mathrm{Z}_2 = \mathbf{v}_2^T \mathrm{X}$ that maximizes $\operatorname{Var}\left(\mathbf{v}_2^T X\right)$ subject to $\|\mathbf{v}_2\| = 1$ and $\operatorname{Cov}\left(\mathbf{v}_1^T X, \mathbf{v}_2^T X\right) = 0$
- The j-th PC satisfies

$$\begin{aligned} &\max \operatorname{Var}\left(\mathbf{v}_j^T X\right) \\ \text{subject to } &\|\mathbf{v}_j\| = 1 \\ &\operatorname{Cov}\left(\mathbf{v}_l^T X, \mathbf{v}_j^T X\right) = \mathbf{v}_l^T \Sigma \mathbf{v}_j = 0 \\ &\text{for } 1 = 1, \ldots, j-1 \end{aligned}$$

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

It turns out such a matrix V can be found using eigen-decomposition of Σ: assume Σ has p eigenvalue-eigenvector pairs $(\lambda, \mathbf{v})$ satisfying:

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

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

$$\Sigma = \sum_{j=1}^p \lambda_j \mathbf{v}_j \mathbf{v}_j^T = V \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_p \end{bmatrix} V^T$$

- The j th PC is given by $Z_j = \mathbf{v}_j^T X$ and its variance is

$$\operatorname{Var}\left(Z_j\right) = \mathbf{v}_j^T \Sigma \mathbf{v}_j = \lambda_j$$

$$\operatorname{Cov}\left(Z_j, Z_k\right) = \mathbf{v}_j^T \Sigma \mathbf{v}_k = 0$$

We may view X is a function of Z. Since V is orthogonal matrix, we also have $X = VZ$, therefore,

$$X = \sum_{k=1}^p \mathbf{v}_k Z_k,$$

where each Z_k can be viewed the coordinate representation of X in the coordinate system given the columns of V. Note also that $\operatorname{Var}(X) = \operatorname{Cov}(X, Z) = V \operatorname{Var}(Z) V^T = \sum_{j=1}^p \lambda_j \mathbf{v}_j \mathbf{v}_j^T$. So $\operatorname{Cov}(X_s, X_t) = \operatorname{Cov}(X_s, Z_t) =$

$\sum_{j=1}^{p} \lambda_j v_{sj} v_{tj}$. Similarity of the PC loadings for the first few components between different X_j's can be heuristically viewed as a degree of linear association. Variables X_j's that are correlated with the first few principal components are the most important in explaining the variability in the data set.

Note $\text{Var}(X_i) = \sum_{j=1}^{p} \lambda_j v_{ij}^2$. The square of the loading value represents the variance of the variable explained by the principal component, can be heuristically understood to be the importance of the variable. The contribution of the variable X_i to the j-th component can be measured by $v_{ij}^2 / \sum_{i=1}^{p} v_{ij}^2 = v_{ij}^2$. The contribution of X_i in accounting for variation in the first l component can be measured by $(\sum_{j=1}^{l} \lambda_j v_{ij}^2)/(\sum_{i=1}^{p}\sum_{j=1}^{l} \lambda_j v_{ij}^2) = (\sum_{j=1}^{l} \lambda_j v_{ij}^2)/(\sum_{j=1}^{l} \lambda_j)$.

With data

Suppose we have $X_i = (X_{i1}, \cdots, X_{ip})^T$ for $i = 1, \cdots, n$ If Σ is unknown, we use sample vcariance S_n as its estimator.

$$\mathbf{X} = \begin{bmatrix} \boldsymbol{x}_1^T \\ \boldsymbol{x}_2^T \\ \vdots \\ \boldsymbol{x}_n^T \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1, \mathbf{x}_2 \cdots \mathbf{x}_p \end{bmatrix}$$

$$S_n = \frac{1}{n-1}\mathbf{X}_c^T\mathbf{X}_c = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(X_i - \overline{X}\right)^T$$

where $\mathbf{X}_c$ the centered design matrix of $\mathbf{X}$.

There are two ways to find $V = (\mathbf{v}_1, \ldots, \mathbf{v}_p)$:

- eigen-decomposition of $\mathbf{X}_c^T\mathbf{X}_c$: $\mathbf{X}_c^T\mathbf{X}_c = VD^2V^T$
- singular value decomposition (SVD) of $\mathbf{X}_c$: $\mathbf{X}_c = UDV^T$.
- Here $D = \text{diag(singular values of } \mathbf{X}_c) = \text{diag(eigenvalues of } \mathbf{X}_c^T\mathbf{X}_c)$

Specifically, in the following discussion, suppose we have $n \times p$ demeaned data matrix $\mathbf{X}$ with $n \geq p$ (**we assume $\mathbf{X}$ it to be in mean-deviance form $\mathbf{X}_c$**), by the full SVD (or use thin SVD):

- $V = (\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_p)$ is an $p \times p$ orthogonal matrix. $\mathbf{v}_j, j = 1, \cdots, p$, form an orthonormal basis for the space spanned by the row vectors of $\mathbf{X}$. The columns of V (i.e., $\mathbf{v}_j, j = 1, \cdots, p$) are the eigenvectors of $\mathbf{X}^T\mathbf{X}$. We call $\mathbf{v}_j$ the j-th **principal component direction** of $\mathbf{X}$.
- D is a $n \times p$ rectangular matrix with nonzero elements along the first $p \times p$ submatrix diagonal. $\text{diag}\left(d_1, d_2, \cdots, d_p\right), d_1 \geq d_2 \geq \cdots \geq d_p \geq 0$ are the singular values of $\mathbf{X}$, or square roots of the eigenvalues of $\mathbf{X}^T\mathbf{X}$.
    - the number of nonzero d_j is equal to the rank of $\mathbf{X}$.

Now, we project each row $\boldsymbol{x}_i$ of $\mathbf{X}$ onto the first PC direction $\mathbf{v}_1$ to obtain $z_{i1} = \mathbf{v}_1^T\boldsymbol{x}_i$. By collecting $\{z_{i,1}\}_{i=1}^n$, we can form a column vector $\mathbf{z}_1$ the **first principle components of $\mathbf{X}$**, We do this for the second, third, .. and the p-th PC direction:

$$\textbf{the j-th principle component:} \qquad \mathbf{z}_j = \mathbf{X}\mathbf{v}_j \qquad (n \times 1), \qquad j = 1, \cdots, p$$

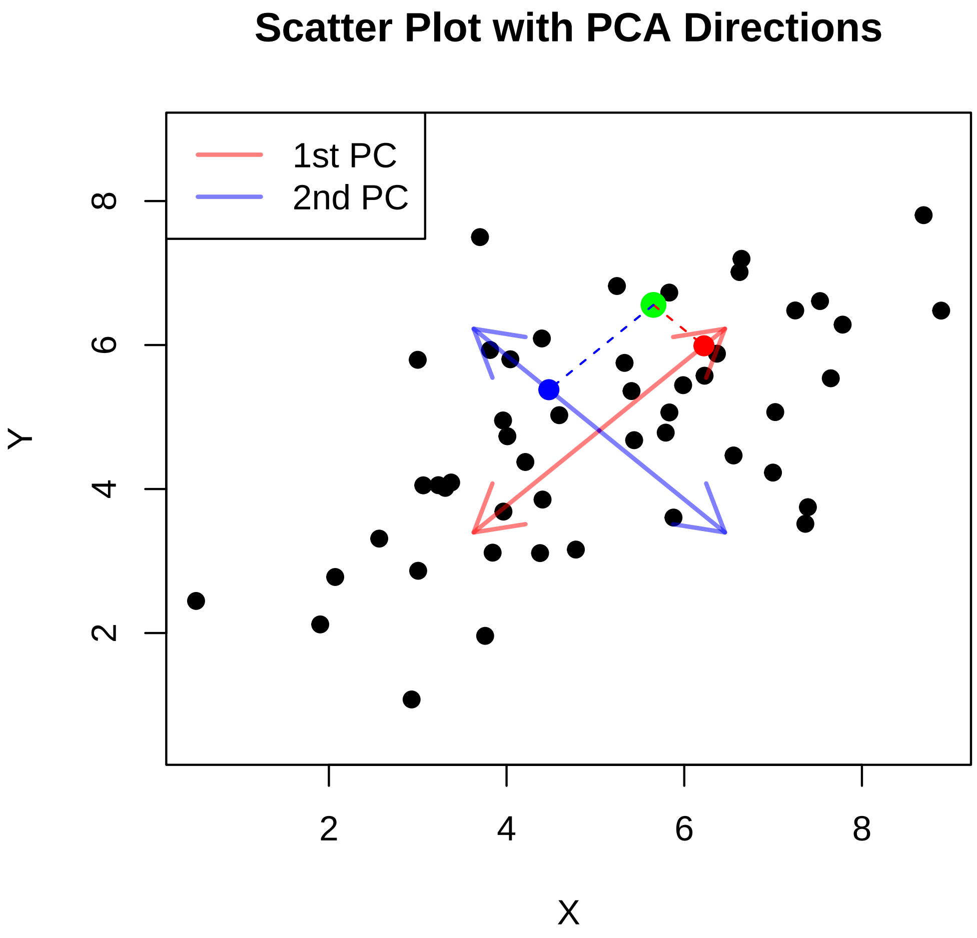
where the i-th entry $z_{ij} = \mathbf{v}_j^T\boldsymbol{x}_i$ is called the **i-th score of the j-th PC of $\mathbf{X}$**

In a data matrix form, we have $\mathbf{Z} = \mathbf{X}V$, where

$$\mathbf{Z} = \left[\begin{array}{c} \mathbf{z}_1, \mathbf{z}_2 \cdots \mathbf{z}_p \end{array}\right] = \left[\begin{array}{c} \boldsymbol{z}_1^T \\ \boldsymbol{z}_2^T \\ \vdots \\ \boldsymbol{z}_n^T \end{array}\right]$$

Note the $p \times 1$ vector $\boldsymbol{z}_i = V^T \boldsymbol{x}_i$ is the **i-th score of the principal components** of $\mathbf{X}$, $i = 1, \cdots, n$.

Geometric meaning: By V being a orthogonal matrix, $\boldsymbol{x}_i = V\boldsymbol{z}_i$, so $\boldsymbol{z}_i$ can be viewed as the new coordinate values of $\boldsymbol{x}_i$ in the p-dimensional coordinate system given by V instead of $[e_1, \ldots, e_p]$.



PCA

It is easy to show that principal components $\mathbf{z}_j$ have maximum variance $d_j^2/(n-1)$, subject to being orthogonal to the earlier ones.

- $\mathbf{z}_1 = \mathbf{X}\mathbf{v}_1$ has the largest sample variance among all normalized linear combinations of the columns of $\mathbf{X}$.
- $\mathbf{z}_2 = \mathbf{X}\mathbf{v}_2$ has the highest sample variance among all normalized liner combinations of the columns of $\mathbf{X}$,

satisfying $\mathbf{v}_2$ orthogonal to $\mathbf{v}_1$ ($\boldsymbol{z}_2$ orthogonal to $\boldsymbol{z}_1$.)

- The last PC $\mathbf{z}_p = \mathbf{X}\mathbf{v}_p$ has the minimum sample variance among all normalized linear combinations of the columns of $\mathbf{X}$, subject to $\mathbf{v}_p$ being orthogonal to the earlier ones ($\boldsymbol{z}_p$ orthogonal to earliers PC's).
- $U = (\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_n)$ is an $n \times n$ orthogonal matrix. $\mathbf{u}_j, j = 1, \cdots, n$, form an orthonormal basis for the space spanned by the column vectors of $\mathbf{X}$.
  - For $d_j \neq 0$, $\mathbf{u}_j = \mathbf{z}_j / \|\mathbf{X}\mathbf{v}_j\| = \mathbf{X}\mathbf{v}_j / d_j$–called **normalized j-th principal component**.
  - The remaining $\mathbf{u}_j$'s that correspond to $d_j = 0$ are just orthogonal extension to make U a orthogonal matrix.

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

The fraction of variance explained Note the total variance of $\mathbf{X}$ is

$$\frac{1}{n-1}\operatorname{tr}(\mathbf{X}^T\mathbf{X}) = \frac{1}{n-1}\operatorname{tr}(\mathbf{Z}^T\mathbf{Z}) = \frac{1}{n-1}\sum_{j=1}^{r} d_j^2$$

So the fraction of total varianc explained by the j-th PC $\mathbf{z}_j$ is $d_j^2 / \sum_{j=1}^{r} d_j^2$.

Dimenstion reduction One application for the principal components is the principal component regression:

Ordinary Least Square (OLS)

$\min_{\beta} \|\mathbf{X}\beta - \mathbf{y}\|^2$

Principal Components Regression (PCR)

$\min_{\theta} \|\mathbf{Z}\theta - \mathbf{y}\|^2$

PCA technique is particularly useful when there is high multi-colinearity exists between the features/variables. By construction, the first principal component will contain the most information about the data, the subsequent principal components contain less and less information about the data while being uncorrelated. The first few principal components $\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_k$ ($k < p$) can be used as predictors in lieu of the original set of all predictors in X.

One particular drawback about *PCR* is that the formation of the principal components for X does not take into account any relationship between X and Y. One remedy is the *Partial Least Squares* (PLS). Instead of focusing only on explaining the variance in $\mathbf{X}$, PLS takes into account the response $\mathbf{y}$. The PLS components are linear combinations of the original predictors, but they are chosen in a way that also explains variance in $\mathbf{y}$.

Latent factors Assume $l \leq d$. We like to approximate each $\boldsymbol{x}_i \in \mathbb{R}^p$ with some possibly lower dimensional representation $\boldsymbol{z}_i \in \mathbb{R}^l$ or **latent factor**, so that $\boldsymbol{x}_i \approx \sum_{k=1}^{l} \mathbf{v}_k z_{ik}$, where each $\mathbf{v}_k \in \mathbb{R}^l$ and orthogonal to each other. Minimize V in the following function over the restriction that V has orthonormal columns

$$\mathcal{L}(V) = \frac{1}{n}\left\|\mathbf{X} - \mathbf{Z}V^{\top}\right\|_F^2 = \frac{1}{n}\sum_{i=1}^{n}\left\|\boldsymbol{x}_i - V\boldsymbol{z}_i\right\|^2$$

here $\|\cdot\|_F$ is Frobenius form, and the data matrix of latent factors $\mathbf{Z}$ each row contains the low dimension versions of the rows of $\mathbf{X}$. One can show that the optimal solution is obtained by setting $V = V_{1:l}$, where $V_{1:l}$ contains the l eigenvectors with largest eigenvalues of the sample covariance matrix S_n. An interesting result is that this optimization problem is equivalent to the previous one we described for maximizing contrained variance. This is can be done using SVD low rank approximation.

This problem may suggest that we can view PCA as a predictive model. However, , it does not usually make sense to use a lower-dimensional representation is approximate the variable.

Note the objective function essentially is $\frac{1}{n}\left\|\mathbf{X} - \mathbf{X}VV^{\top}\right\|_F^2$ such that $V^TV = I$, so the error is just a kind of reconstruction error. The reconstruction is purely deterministic (fixing V or estimated V): given $\mathbf{z}_i$, the $\mathbf{x}_i$ is fully determined. A probabilistic PCA is a probabilistic version of PCA that allows generation of $\mathbf{x}_i$: given $\mathbf{z}_i$, $\mathbf{x}_i$ can be generated according to some specified distribution.

Back **Updated:** 2023-09-02 Statistical Simulation, Wei Li