Chapter 7: Principle Component Analysis

Richard Liu

May 10, 2020

Contents

Source

- Jie Hu, *Applied Linear Models, Xiamen University*
- Chapter 10: Ridge Regression, Principal Component Regression
- Chapter 11: Principal Component Regression, Partial Least Squared Regression

Section 1

Definition of Dimensionality Reduction

Definition

- Dimensionality Reduction is a kind of method used for transferring a dataset with p features to the one with k features, normally $k << p$.
- Information of data will not be entirely preserved after transformation. But that is just what we need in some cases.
- PCA is one of the most renowned dimensionality reduction algorithms, others such as NMF, Sparse PCA, t-SNE, umap will not be extended in this chapter, but Python Code here is easy to learn.

Section 2

PCA: Theories and Applications

Introduction

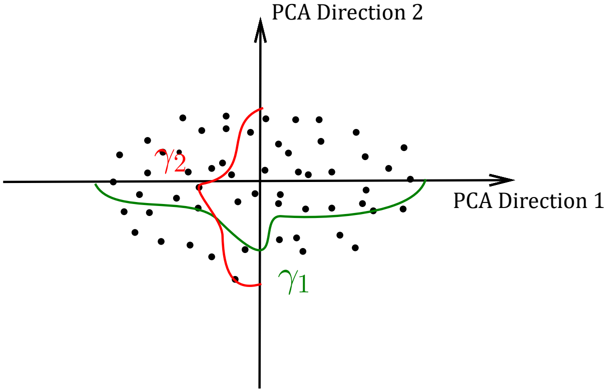
- Consider a data point $X = (X_1, X_2, \cdots, X_p)^T$, where X_i is the $i - th$ feature of X. Then for *PCA*, we want to find a linear transformation to change X to Y, which means

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_p \end{pmatrix} = \begin{pmatrix} \gamma_{11} & \gamma_{12} & \cdots & \gamma_{1p} \\ \gamma_{21} & \gamma_{22} & \cdots & \gamma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_{p1} & \gamma_{p2} & \cdots & \gamma_{pp} \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{pmatrix}$$

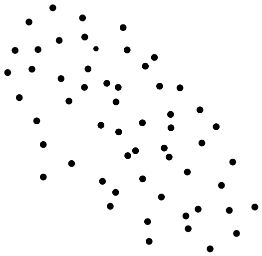
$$Y = \Gamma X$$

  Y_i is called principal component.
- If we assume $E(X) = \mu$, $\text{Cov}(X) = \Sigma$, then $\text{Cov}(Y) = \Gamma \Sigma \Gamma^T = \Lambda$.

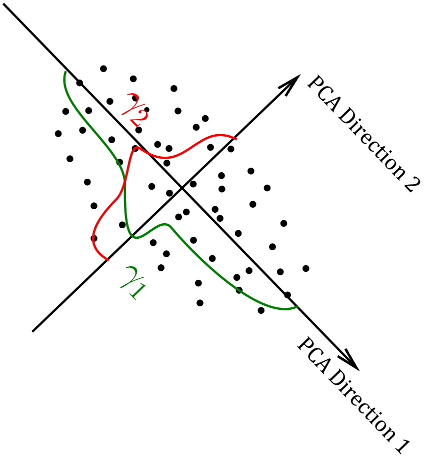
Graph



Graph



Graph



Properties of Γ

Lemma 1

If $\text{Cov}(X) = \Sigma$, then $\text{Cov}(AX) = A\Sigma A^T$.

- We want this linear transformation to complete 2 tasks:
  1. Each feature of Y is linear independent.
  2. Few dimensionality could preserve most of the information.
- This means, we want an orthogonal matrix and every linear transformation vector to "maximize" the variance. How to do?
- Note that $\text{Var}(Y_i) = \gamma_i^T \Sigma \gamma_i = \lambda_i$. $\text{Cov}(Y_i, Y_j) = \gamma_i^T \Sigma \gamma_j^T = 0$, so we could write the goal as an optimization problem

Properties of Γ

- The optimization problem for the first vector transformation is

$$\max_{\gamma_1} \gamma_1^T \Sigma \gamma_1 \quad \text{s. t. } \gamma_1^T \gamma_1 = 1$$

- For the second vector transformation, we have

$$\max_{\gamma_2} \gamma_2^T \Sigma \gamma_2 \quad \text{s. t. } \gamma_2^T \gamma_2 = 1, \gamma_2^T \gamma_1 = 0, \gamma_1^T \gamma_1 = 1$$

- To this end, we take Γ to be the eigenvectors of Σ with orthonormal transformation, this could ensure that $\mathrm{Cov}(Y)$ is a diagonal squared matrix, for information we only need to select the one with larger values of diagonal elements.

Why λ_i could be used to measure variance Σ?

- As we have mentioned, λ_i is just the variance of each new feature dimension.

Definition 1: Variance Contribution Rate (VCR)

Define $\alpha_i = \frac{\lambda_i}{\sum_{i=1}^{p} \lambda_i}$ as the VCR of the i-th principal component.

- PCA is a feature reconstruction method, how to choose an appropriate p depends on α_i.

Explanation Quality

- Problem: Where does the information in Y_k come from?
  - Also need to measure the correlation of Y_k and original data X.

Proposition 2

$$\rho\left(Y_k, X_i\right) = \frac{\sqrt{\lambda_k}}{\sqrt{\sigma_{ii}}}\gamma_{ki}$$

Proof(Part 1)

In Chapter 1 we have mentioned

$$\rho\left(Y_k, X_i\right) = \frac{\operatorname{cov}\left(Y_k, X_i\right)}{\sqrt{\operatorname{var}\left(Y_k\right)\operatorname{var}\left(X_i\right)}}$$

Explanation Quality

Proof(Part 2)

For $Y_k = \gamma_k^T X, X_k = e_i^T X$, we have

$$\rho(Y_k, X_i) = \frac{\gamma_k^T \Sigma e_i}{\sqrt{\lambda_k \sigma_{ii}}} = \frac{\lambda_k \gamma_k^T e_i}{\sqrt{\lambda_k \sigma_{ii}}} = \frac{\lambda_k \gamma_{ki}}{\sqrt{\lambda_k \sigma_{ii}}} \square$$

- In fact, it has a property similar to that of VCR.

Proposition 3

$\sum_{k=1}^{p} \rho^2 (Y_k, X_i) = 1$

Definition 2: VCR w.r.t. original variable X_i

Define $v_i = \sum_{k=1}^{m} \rho^2 \left(Y_k, X_i\right)$ as the VCR of the first m principal components w.r.t. X_i.

Section 3

Principle Component Regression

- We do need to analyze the properties of PCA even if we do not need to literally train a regression model.
    - This is because usually we need to find the correlation of variable X and response Y.
- Consider a regression model

$$Y = \beta_0 \mathbf{1} + X\beta + \epsilon$$

Explanation

We have normalized X, so in this case $E(X) = 0$.

- We will choose the canonical form of this model.

Canonical Form

Definition 3: Canonical Form

Let $R = X^T X$ be the covariance matrix of X (Why?), let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ be the eigenvalues of R, $\gamma_1, \cdots, \gamma_p$ be the orthogonalized eigenvectors of R, then the canonical form of the model is written as

$$Y = \beta_0 \mathbf{1} + Z\alpha + \epsilon$$

where $Z = X\Gamma$, $\alpha = \Gamma^T \beta$, $\Gamma = \begin{bmatrix} \gamma_1 & \gamma_2 & \cdots & \gamma_p \end{bmatrix}$

- Estimating β tantamounts to estimating α. In PCA, we want to "drop" some dimensions, that is to say we want to manually let some of the elements in α to be zero.

Estimator β^*

- Suppose that we choose the first r elements, which means $\alpha_k = 0, k = r+1, \cdots, p$.

Proposition 4

$$\beta^* = \Gamma\hat{\alpha} = \Gamma_1\hat{\alpha}_1 = \Gamma_1\Lambda_1^{-1}Z_1^T Y$$

where $\Gamma = \begin{bmatrix} \Gamma_1 & \Gamma_2 \end{bmatrix}, \alpha = \begin{bmatrix} \alpha_1 \\ 0 \end{bmatrix}$ and Λ_1, Z_1 is also the first r components (for Λ_1 and Z_1, their sizes are not the same.)

Proof

Note that $\hat{\alpha} = (Z^T Z)^{-1} Z^T Y = \Lambda^{-1} Z^T Y$ $\square$

Properties of β^*

Proposition 5

$\beta^* = \Gamma_1 \Gamma_1^T \hat{\beta}$

Proposition 6

if $r < p$, then we have $||\beta^*|| < ||\hat{\beta}||$.

Proof

$$||\beta^*|| = \left\|\Gamma_1 \Gamma_1^T \hat{\beta}\right\| = \left\|\Gamma_1^T \hat{\beta}\right\| < ||\Gamma^T \hat{\beta}|| = ||\hat{\beta}||$$

When PCA will work in real dataset?

Proposition 7

$$MSE\left(\beta^*\right) = MSE(\hat{\beta}) + \left(\sum_{i=r+1}^{p} \alpha_i^2 - \sigma^2 \sum_{i=r+1}^{p} \lambda_i^{-1}\right)$$

- When X is ill-conditioned, λ_i^{-1} will be large, then it is easy to find r such that $MSE(\beta^*) < MSE(\hat{\beta})$, which means the data will behave more robust (we have dropped more noise than useful information).
- Conversely, for a dataset with good property (nearly inter-independent), PCA will not work very well (we have dropped more useful information than noise).

Section 4

Supplementary

Subsection 1

Mathematical Foundations: Linear Algebra

Examples

Example 1

Suppose that you have known

$$\begin{bmatrix} 32 \\ 77 \\ 122 \end{bmatrix} = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix} \begin{bmatrix} 4 \\ 5 \\ 6 \end{bmatrix}$$

Then what is the answer of

$$\begin{bmatrix} 4 & 5 & 6 \end{bmatrix} \begin{bmatrix} 4 \\ 5 \\ 6 \end{bmatrix}$$

Examples

Example 1: More General

Suppose that you have known that $X = A^T B$, with $A = \begin{bmatrix} a_1 & \cdots & a_n \end{bmatrix}$, B is a column vector, then how to compute X_i?

Examples

Example 2

Suppose that you have known

$$\begin{bmatrix} 468 & 576 & 684 \\ 1062 & 1305 & 1548 \\ 1656 & 2034 & 2412 \end{bmatrix} = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}^3$$

Then what is the answer of

$$\begin{bmatrix} 4 & 5 & 6 \end{bmatrix} \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix} \begin{bmatrix} 2 \\ 5 \\ 8 \end{bmatrix}$$

Examples

Example 2: More General

Suppose that you have known $X = A^T BC$, where $A = \begin{bmatrix} a_1 & \cdots & a_n \end{bmatrix}$, $C = \begin{bmatrix} c_1 & \cdots & c_m \end{bmatrix}$, B is a matrix, then how to compute X_{ij}?

- This is the meaning of row left, column right".

A Common Extension

Tricks

For a column vector X, we have $X_i = e_i^T X$. For a row vector Y^T, we have $Y_i^T = Ye_i$. For a matrix Z, we have $Z_{ij} = e_i^T Z e_j$.

Subsection 2

Proof of Proposition 3 and 5

Proposition 3

$\sum_{k=1}^{p} \rho^2 (Y_k, X_i) = 1$

Proof(Part 1)

Firstly,

$$\sum_{k=1}^{p} \rho^2 (Y_k, X_i) = \frac{1}{\sigma_{ii}} \sum_{k=1}^{p} \lambda_k \gamma_{ki}^2$$

Consider the matrix decomposition

$$\sum_{k=1}^{p} \lambda_k \gamma_{ki}^2 = \begin{bmatrix} \lambda_1 \gamma_{1i} & \lambda_2 \gamma_{2i} & \cdots & \lambda_p \gamma_{pi} \end{bmatrix} \begin{bmatrix} \gamma_{1i} \\ \gamma_{2i} \\ \vdots \\ \gamma_{pi} \end{bmatrix}$$

Proof(Part 2)

And note that $\begin{bmatrix}\gamma_{1i}\\ \gamma_{2i}\\ \vdots\\ \gamma_{pi}\end{bmatrix} = \Gamma e_i$, $\begin{bmatrix}\lambda_1\gamma_{1i} & \lambda_2\gamma_{2i} & \cdots & \lambda_p\gamma_{pi}\end{bmatrix} = \Lambda(\Gamma e_i)^T$, with $\Gamma^T\Lambda\Gamma = \Sigma$, we have

$$\sum_{k=1}^{p} \lambda_k \gamma_{ki}^2 = \sigma_{ii} \square$$

Proposition 5

$\beta^* = \Gamma_1 \Gamma_1^T \hat{\beta}$

Proof

Note that $\beta^* = \Gamma_1 \Lambda_1^{-1} Z_1' Y = \Gamma_1 \Lambda_1^{-1} \Gamma_1' X' Y = \Gamma_1 \Lambda_1^{-1} \Gamma_1' X' X \hat{\beta}$
$= \Gamma_1 \Lambda_1^{-1} \Gamma_1' \Gamma \Lambda \Gamma' \hat{\beta}$ For

$$\Gamma_1' \Gamma = \Gamma_1' \begin{bmatrix} \Gamma_1 & \Gamma_2 \end{bmatrix} = \begin{bmatrix} I_r & 0 \end{bmatrix}$$

We have

$$\Lambda_1^{-1} \Gamma_1' \Gamma \Lambda = \Lambda_1^{-1} \begin{bmatrix} I_r & 0 \end{bmatrix} \Lambda = \begin{bmatrix} \Lambda_1^{-1} & 0 \end{bmatrix} \begin{bmatrix} \Lambda_1 & 0 \\ 0 & \Lambda_2 \end{bmatrix} = \begin{bmatrix} I_r & 0 \end{bmatrix}$$

Subsection 3

Introduction of 3 Other Dimensionality Reduction Methods

Method 1: Pearson Correlation Coefficients

- Very easy to use!
- Based on Hypothesis Test.
- See here for more information.

Method 2: Sparse PCA

- PCA has an alternative form as

$$\max_{X \in \mathbb{R}^{n \times p}} \|AX\|_F^2 \quad \text{subject to} \quad X^T X = I_p$$

Why?. Note that X is the Γ, A is the X in the previous slides for more clarity.

- Sparse PCA has an alternative form as

$$\min_{X \in \mathbb{R}^{n \times p}} -\|AX\|_F^2 + \lambda \|X\|_1 \quad \text{subject to} \quad X^T X = I_p$$

It could generate a sparse form of the linear transformation, that is to say, we could not only select few dimensions of Y, but also few dimensions of X.

Method 3: Dictionary Learning and Sparse Coding

- For example, for a dataset $\mathcal{X}$ with N elements $\{X_1, \cdots, X_N\}$ with $X_i \in \mathcal{S}_+^d, i = 1, 2, \ldots, N$ (which means each data point is a d-by-d Symmetric Positive Definite (SPD) matrix), we would like to find a dictionary $\mathbf{B} = \{B_1, B_2, \cdots, B_n\}$ with n elements and $B_i \in \mathcal{S}_+^d, i = 1, 2, \ldots, n$ and the sparse coding $\alpha_i = \{\alpha_{1i}, \cdots, \alpha_{ni}\}$ with $\alpha_{ji} \geq 0$ satisfying

$$X_i \simeq \sum_{j=1}^{n} \alpha_{ji} B_j, i = 1, 2, \cdots, N$$

Graph: An Illustration

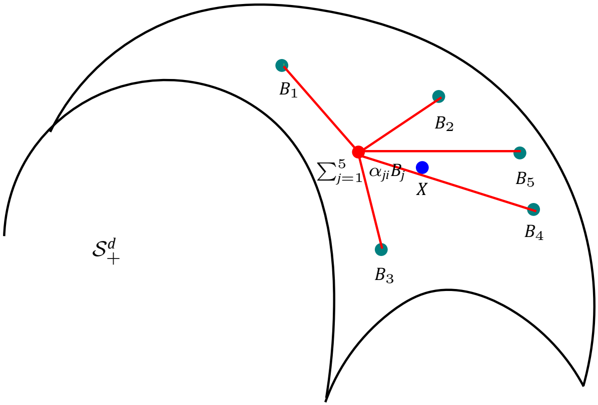


Figure: An illustration of Dictionary Learning

Objective Function

- In general, we have

$$\min_{\mathbf{B}\in\mathcal{M}_d^n, \alpha\in\mathbb{R}_+^{n\times N}} \frac{1}{2}\sum_{j=1}^{N} d_{\mathcal{R}}^2\left(X_j, \mathbf{B}\alpha_j\right) + \mathrm{Sp}\left(\alpha_j\right) + \Omega(\mathbf{B})$$

  as the minimization problem.
- This relies on Manifold Learning.
- Why adding on two sparse constraints?

Thank you!