Multivariate Statistical Analysis

Lecture 14

Fudan University

luoluo@fudan.edu.cn

Outline

1. Bayesian Multivariate Linear Regression
2. Principal Components Analysis
3. Principal Coordinate Analysis
4. Kernel Principal Component Analysis

Outline

1. Bayesian Multivariate Linear Regression
2. Principal Components Analysis
3. Principal Coordinate Analysis
4. Kernel Principal Component Analysis

Bayesian Multivariate Linear Regression

We can additionally suppose each b_{ij} independently follows

$$b_{ij} \sim \mathcal{N}(0, \tau^2),$$

then the posterior likelihood function is

$$\begin{aligned}
& L(\mathbf{B}, \mathbf{\Sigma}) \\
= & \prod_{i=1}^{N} \frac{1}{\sqrt{(2\pi)^p \det(\mathbf{\Sigma})}} \exp\left(-\frac{1}{2}(\mathbf{B}^\top \mathbf{x}_i - \mathbf{y}_i)^\top \mathbf{\Sigma}^{-1} (\mathbf{B}^\top \mathbf{x}_i - \mathbf{y}_i)\right) \\
& \cdot \prod_{i=1}^{p} \prod_{j=1}^{q} \frac{1}{\sqrt{2\pi\tau^2}} \exp\left(-\frac{b_{ij}^2}{2\tau^2}\right) \\
\propto & \frac{1}{(\det(\mathbf{\Sigma}))^{N/2}} \exp\left(-\frac{1}{2}\mathrm{tr}\left((\mathbf{XB} - \mathbf{Y})\mathbf{\Sigma}^{-1}(\mathbf{XB} - \mathbf{Y})^\top\right) - \frac{1}{2\tau^2}\left\|\mathbf{B}\right\|_F^2\right),
\end{aligned}$$

which leads to

$$\mathrm{vec}(\hat{\mathbf{B}}) = (\mathbf{I}_q \otimes \tau^2 \mathbf{X}^\top \mathbf{X} + \mathbf{\Sigma} \otimes \mathbf{I}_p)^{-1} \mathrm{vec}(\tau^2 \mathbf{X}^\top \mathbf{Y}).$$

Bayesian Multivariate Linear Regression

We typically suppose

$$\boldsymbol{\beta}_{(i)} \overset{\text{i.i.d}}{\sim} \mathcal{N}_q(\mathbf{0}, \tau^2\boldsymbol{\Sigma}), \qquad \text{where} \qquad \mathbf{B} = \begin{bmatrix} \boldsymbol{\beta}_{(1)}^\top \\ \vdots \\ \boldsymbol{\beta}_{(p)}^\top \end{bmatrix} \in \mathbb{R}^{p\times q},$$

then the posterior likelihood function is

$$\begin{aligned} & L(\mathbf{B}, \boldsymbol{\Sigma}) \\ = & \prod_{i=1}^{N} \frac{1}{\sqrt{(2\pi)^p \det(\boldsymbol{\Sigma})}} \exp\left(-\frac{1}{2}(\mathbf{B}^\top\mathbf{x}_i - \mathbf{y}_i)^\top \boldsymbol{\Sigma}^{-1}(\mathbf{B}^\top\mathbf{x}_i - \mathbf{y}_i)\right) \\ & \cdot \prod_{j=1}^{p} \frac{1}{\sqrt{(2\pi)^q \det(\boldsymbol{\Sigma})}} \exp\left(-\frac{1}{2\tau^2}\boldsymbol{\beta}_{(j)}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}_{(j)}\right) \\ \propto & \frac{1}{(\det(\boldsymbol{\Sigma}))^{N/2}} \exp\left(-\frac{1}{2}\mathrm{tr}\left((\mathbf{XB}-\mathbf{Y})\boldsymbol{\Sigma}^{-1}(\mathbf{XB}-\mathbf{Y})^\top\right) - \frac{1}{2\tau^2}\mathbf{B}\boldsymbol{\Sigma}^{-1}\mathbf{B}^\top\right). \end{aligned}$$

Bayesian Multivariate Linear Regression

We have

$$\hat{\mathbf{B}}_\lambda = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top \mathbf{Y},$$

and

$$\hat{\boldsymbol{\Sigma}}_\lambda = \frac{1}{N} \mathbf{Y}^\top (\mathbf{I} - \mathbf{X}(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top) \mathbf{Y},$$

where $\lambda = 1/\tau^2$.

Outline

1 Bayesian Multivariate Linear Regression

2 Principal Components Analysis

3 Principal Coordinate Analysis

4 Kernel Principal Component Analysis

Population Principal Components Analysis

Let $\mathbf{x}$ be a p-dimensional random vector with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma} \succ \mathbf{0}$.

Let $\mathbf{u}_1 \in \mathbb{R}^p$ with $\|\mathbf{u}_1\|_2 = 1$ and maximizing the variance of $\mathbf{u}_1^\top \mathbf{x}$, then

$$(\boldsymbol{\Sigma} - \lambda_1 \mathbf{I})\mathbf{u}_1 = \mathbf{0},$$

where λ_1 is the largest root of

$$\det(\boldsymbol{\Sigma} - \lambda \mathbf{I}) = 0.$$

1. We call $y_1 = \mathbf{u}_1^\top \mathbf{x}$ as the first principle component of $\mathbf{x}$.
2. The pair $\lambda_1 \in \mathbb{R}$ and $\mathbf{u}_1 \in \mathbb{R}^p$ are the largest eigenvalue and corresponding eigenvector of $\boldsymbol{\Sigma}$.

Population Principal Components Analysis

For the second principle components

$$y_2 = \mathbf{u}_2^\top \mathbf{x},$$

we determine $\mathbf{u}_2 \in \mathbb{R}^p$ by maximizing the variance of y_2 under the constraints $\|\mathbf{u}_2\|_2 = 1$ and y_2 be uncorrelated with y_1.

For the k-th principle component

$$y_k = \mathbf{u}_k^\top \mathbf{x},$$

we determine $\mathbf{u}_k$ by maximizing the variance of y_k under the constraints $\|\mathbf{u}_k\|_2 = 1$ and y_k be uncorrelated with $y_1, \dots, y_{k-1}$.

Population Principal Components Analysis

Let vector $\mathbf{u}_k \in \mathbb{R}^p$ the k-th principle component

$$y_k = \mathbf{u}_k^\top \mathbf{x}$$

holds that

$$(\mathbf{\Sigma} - \lambda_k \mathbf{I})\mathbf{u}_k = \mathbf{0},$$

where λ_k is the k-th largest root of

$$\det(\mathbf{\Sigma} - \lambda \mathbf{I}) = 0.$$

The pair $\lambda_k \in \mathbb{R}$ and $\mathbf{u}_k \in \mathbb{R}^p$ are the k-th largest eigenvalue and corresponding eigenvector of $\mathbf{\Sigma}$.

Population Principal Components Analysis

Let vector $\mathbf{u}_k \in \mathbb{R}^p$ the k-th principle component

$$y_k = \mathbf{u}_k^\top \mathbf{x}$$

holds that

$$(\boldsymbol{\Sigma} - \lambda_k \mathbf{I})\mathbf{u}_k = \mathbf{0},$$

where λ_k is the k-th largest root of

$$\det(\boldsymbol{\Sigma} - \lambda \mathbf{I}) = 0.$$

The pair $\lambda_k \in \mathbb{R}$ and $\mathbf{u}_k \in \mathbb{R}^p$ are the k-th largest eigenvalue and corresponding eigenvector of $\boldsymbol{\Sigma}$.

PCA for dimensionality Reduction

We can write

$$\mathbf{U}_k = \begin{bmatrix}\mathbf{u}_1 & \cdots & \mathbf{u}_k\end{bmatrix} \in \mathbb{R}^{p\times k} \quad \text{and} \quad \mathbf{\Lambda}_k = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_k \end{bmatrix} \in \mathbb{R}^{k\times k}$$

contains the top-k eigenvectors and eigenvalues pairs of $\mathbf{\Sigma}$, that is

$$\mathbf{\Sigma}\mathbf{U}_k = \mathbf{U}_k\mathbf{\Lambda} \qquad \text{with} \qquad \mathbf{U}_k^\top\mathbf{U}_k = \mathbf{I}.$$

PCA for dimensionality Reduction

We can keep $\mathbf{U}_k \in \mathbb{R}^{p\times k}$ and transform $\mathbf{x} \in \mathbb{R}^p$ to

$$\mathbf{U}_k^\top \mathbf{x} \in \mathbb{R}^k,$$

where $k \ll p$.

The information of $\mathbf{x}$ can be estimated by

$$\hat{\mathbf{x}} = \mathbf{U}_k(\mathbf{U}_k^\top \mathbf{x}) \in \mathbb{R}^p.$$

We have

$$\mathrm{Cov}[\hat{\mathbf{x}}] = \mathbf{U}_k \boldsymbol{\Lambda}_k \mathbf{U}_k^\top,$$

which is the best rank-k approximation of $\boldsymbol{\Sigma}$.

Outline

1. Bayesian Multivariate Linear Regression

2. Principal Components Analysis

3. Principal Coordinate Analysis

4. Kernel Principal Component Analysis

Sample Principal Components Analysis

Given observation $\mathbf{x}_1, \dots, \mathbf{x}_N \in \mathbb{R}^p$, we construct sample covariance

$$\mathbf{S} = \frac{1}{N-1}\sum_{\alpha=1}^{N}(\mathbf{x} - \bar{\mathbf{x}})(\mathbf{x} - \bar{\mathbf{x}})^\top, \qquad \text{where} \ \ \bar{\mathbf{x}} = \frac{1}{N}\sum_{\alpha=1}^{N}\mathbf{x}_\alpha.$$

Let spectral decomposition of $\mathbf{S}$ be $\mathbf{S} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}$, where $\mathbf{U} \in \mathbb{R}^{p\times p}$ is orthogonal and $\mathbf{\Lambda} \in \mathbb{R}^{p\times p}$ is diagonal.

We write

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1^\top \\ \vdots \\ \mathbf{x}_N^\top \end{bmatrix} \in \mathbb{R}^{N\times p},$$

which results the sample principle components

$$\mathbf{Y} = \begin{bmatrix} (\mathbf{x}_1 - \bar{\mathbf{x}})^\top \mathbf{U}_k \\ \vdots \\ (\mathbf{x}_N - \bar{\mathbf{x}})^\top \mathbf{U}_k \end{bmatrix} = \mathbf{H}\mathbf{X}\mathbf{U}_k \in \mathbb{R}^{N\times k}, \quad \text{where} \quad \mathbf{H} = \mathbf{I} - \frac{1}{N}\mathbf{1}_N\mathbf{1}_N^\top \in \mathbb{R}^{N\times N}.$$

Principal Coordinate Analysis

We consider the case of $p \geq N$ and define

$$\mathbf{T} = \frac{1}{N-1}\mathbf{H}\mathbf{X}\mathbf{X}^\top\mathbf{H} \in \mathbb{R}^{N\times N}$$

with spectral decomposition

$$\mathbf{T} = \mathbf{V}\boldsymbol{\Gamma}\mathbf{V}^\top,$$

where $\mathbf{V} \in \mathbb{R}^{N\times N}$ is orthogonal and $\boldsymbol{\Gamma} \in \mathbb{R}^{N\times N}$ is diagonal.

The matrix $\mathbf{Y} \in \mathbb{R}^{N\times k}$ can be written as

$$\mathbf{Y} = \mathbf{V}_k\boldsymbol{\Gamma}_k^{1/2} \in \mathbb{R}^{N\times k}.$$

Kernel Principal Component Analysis

We consider the case of $p \geq N$ and define

$$\mathbf{T} = \frac{1}{N-1}\mathbf{H}\mathbf{X}\mathbf{X}^\top\mathbf{H} \in \mathbb{R}^{N\times N}$$

with spectral decomposition

$$\mathbf{T} = \mathbf{V}\mathbf{\Gamma}\mathbf{V}^\top,$$

where $\mathbf{V} \in \mathbb{R}^{N\times N}$ is orthogonal and $\mathbf{\Gamma} \in \mathbb{R}^{N\times N}$ is diagonal.

The matrix $\mathbf{Y} \in \mathbb{R}^{N\times k}$ can be written as

$$\mathbf{Y} = \mathbf{V}_k\mathbf{\Gamma}_k^{1/2} \in \mathbb{R}^{N\times k}.$$

Outline

1. Bayesian Multivariate Linear Regression

2. Principal Components Analysis

3. Principal Coordinate Analysis

4. Kernel Principal Component Analysis

Kernel Principal Component Analysis

The matrix $\mathbf{S} \in \mathbb{R}^{p\times p}$ corresponds to outer product the centralized data, that is

$$\mathbf{X}^\top\mathbf{H}\mathbf{H}\mathbf{X} = \sum_{\alpha=1}^{N}(\mathbf{x}_\alpha\mathbf{H})(\mathbf{x}_\alpha\mathbf{H})^\top \in \mathbb{R}^{p\times p}.$$

The matrix $\mathbf{T} \in \mathbb{R}^{N\times N}$ corresponds to the inner product, that is

$$\mathbf{H}\mathbf{X}\mathbf{X}^\top\mathbf{H} = \mathbf{H}\begin{bmatrix} \mathbf{x}_1^\top\mathbf{x}_1 & \mathbf{x}_1^\top\mathbf{x}_2 & \dots & \mathbf{x}_1^\top\mathbf{x}_N \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}_N^\top\mathbf{x}_1 & \mathbf{x}_N^\top\mathbf{x}_2 & \dots & \mathbf{x}_N^\top\mathbf{x}_N \end{bmatrix}\mathbf{H} \in \mathbb{R}^{N\times N}.$$

We achieve kernel PCA by replacing $\mathbf{X}\mathbf{X}^\top$ with some kernel matrix.

Kernel Principal Component Analysis

We map the sample $\mathbf{x}_\alpha \in \mathcal{X} \subseteq \mathbb{R}^p$ to the feature space $\mathcal{H} \subseteq \mathbb{R}^d$, that is

$$\phi : \mathcal{X} \to \mathcal{H},$$

and define the kernel function (inner product)

$$K(\mathbf{x}, \mathbf{y}) \triangleq \phi(\mathbf{x})^\top \phi(\mathbf{y}).$$

We replace $\mathbf{X}\mathbf{X}^\top \in \mathbb{R}^{N\times N}$ with the kernel matrix

$$\begin{aligned}\mathbf{K} &= \begin{bmatrix} \phi(\mathbf{x}_1)^\top \\ \vdots \\ \phi(\mathbf{x}_N)^\top \end{bmatrix} \begin{bmatrix} \phi(\mathbf{x}_1) & \cdots & \phi(\mathbf{x}_N) \end{bmatrix} \\ &= \begin{bmatrix} K(\mathbf{x}_1, \mathbf{x}_1) & K(\mathbf{x}_1, \mathbf{x}_2) & \dots & K(\mathbf{x}_1, \mathbf{x}_N) \\ \vdots & \vdots & \ddots & \vdots \\ K(\mathbf{x}_N, \mathbf{x}_1) & K(\mathbf{x}_N, \mathbf{x}_2) & \dots & K(\mathbf{x}_N, \mathbf{x}_N) \end{bmatrix} \in \mathbb{R}^{N\times N},\end{aligned}$$

Kernel Principal Component Analysis

We replace

$$\mathbf{T} = \frac{1}{N-1}\mathbf{H}\mathbf{X}\mathbf{X}^\top\mathbf{H} \in \mathbb{R}^{N\times N}$$

with

$$\mathbf{T}_K = \frac{1}{N-1}\mathbf{H}\mathbf{K}\mathbf{H}$$

and achieve kernel PCA by spectral decomposition on $\mathbf{T}_K$.