Multivariate Statistical Analysis

Lecture 15

Fudan University

luoluo@fudan.edu.cn

Outline

Outline

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

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

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

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

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

Outline

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

Principal Components Analysis

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

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

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

where λ_1 is the largest root of

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

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 $\mathbf{\Sigma}$.

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 $y_1, \dots, y_{k-1}$.

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 $\boldsymbol{\Sigma}$.

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 $\mathbf{\Sigma}$.

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

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

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

Outline

Sample Principal Components Analysis

Given observation $\mathbf{x}_1, \ldots, \mathbf{x}_N \in \mathbb{R}^p$, 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 $\mathbf{S} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}$, where $\mathbf{U} \in \mathbb{R}^{p\times p}$ is orthogonal and $\boldsymbol{\Lambda} \in \mathbb{R}^{p\times p}$ 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

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

Outline

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 corresponding kernel function (inner product)

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

Kernel Principal Component Analysis

The matrix

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

contains

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

We replace the inner product $\mathbf{x}_i^\top\mathbf{x}_j$ with

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

Kernel Principal Component Analysis

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

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

and replace $\mathbf{T} \in \mathbb{R}^{N\times N}$ with

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

The kernel PCA is achieved by spectral decomposition on $\mathbf{T}_K$.

Kernel Principal Component Analysis

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

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

and replace $\mathbf{T} \in \mathbb{R}^{N\times N}$ with

$$\mathbf{T}_K = \frac{1}{N-1}\mathbf{HKH}.$$

The kernel PCA is achieved by spectral decomposition on $\mathbf{T}_K$.

Examples of kernel functions:

1. We define the polynomial kernel as

$$K(\mathbf{x}, \mathbf{y}) = (\mathbf{x}^\top \mathbf{y} + c)^d$$

   for some $c \in \mathbb{R}$ and $d \in \mathbb{N}$.

2. We define the Gaussian kernel (radial basis function kernel) as

$$K(\mathbf{x}, \mathbf{y}) = \exp\left(-\frac{\|\mathbf{x} - \mathbf{y}\|_2^2}{2\sigma^2}\right).$$

Outline

Canonical Correlation Analysis

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

We partition $\mathbf{x}$ into q and $p-q$ components as

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}^{(1)} \\ \mathbf{x}^{(2)} \end{bmatrix}.$$

The covariance matrix is partitioned similarly as

$$\mathbf{\Sigma} = \begin{bmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22} \end{bmatrix}.$$

We shall develop $\mathbf{u}_1$ and $\mathbf{v}_1$ that maximize the correlation between

$$y^{(1)} = \mathbf{u}_1^\top \mathbf{x}^{(1)} \qquad \text{and} \qquad y^{(2)} = \mathbf{v}_1^\top \mathbf{x}^{(2)}$$

with constraints

$$\mathrm{Var}[y^{(1)}] = 1 \qquad \text{and} \qquad \mathrm{Var}[y^{(2)}] = 1.$$

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 $y_1, \dots, y_{k-1}$.

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 $\mathbf{\Sigma}$.

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 $\mathbf{\Sigma}$.