Multivariate Statistical Analysis

Lecture 15

Fudan University

luoluo@fudan.edu.cn

Outline

Outline

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

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 $\boldsymbol{\Sigma}$.

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 $\mathbf{\Sigma}$.

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

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

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

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

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

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)

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

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 $\mathbf{T}_K$.

Kernel Principal Component Analysis

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

Factor Analysis

Let the observable vector $\mathbf{y} \in \mathbb{R}^p$ be written as

$$\mathbf{y} = \mathbf{W}\mathbf{x} + \boldsymbol{\mu} + \epsilon,$$

where

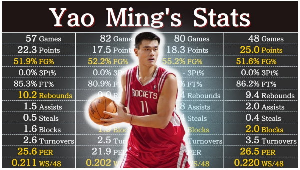
1. $\mathbf{W} \in \mathbb{R}^{p \times q}$ is the loading matrix (parameter),
2. $\mathbf{x} \in \mathbb{R}^q$ is the common factor (parameter/random vector),
3. $\boldsymbol{\mu} \in \mathbb{R}^p$ is the mean vector (parameter),
4. $\epsilon \in \mathbb{R}^p$ is the specific factor (random vector).

The model is similar to regression, but $\mathbf{x}$ is unobserved.

Example of sports games:

$$\mathbf{y} = \mathbf{W}\mathbf{x} + \boldsymbol{\mu} + \epsilon.$$

1. $\mathbf{y}$: performance in real-world
2. $\mathbf{W}$: system of the game
3. $\mathbf{x}$: attributes in the game
4. $\boldsymbol{\mu}$: average attributes
5. ϵ: noise/exception



Outline

Probabilistic Principle Component Analysis

Let $\mathbf{y}_1, \ldots, \mathbf{y}_N \in \mathbb{R}^p$ be N independent observations and we have

$$\mathbf{y}_\alpha = \mathbf{W}\mathbf{x}_\alpha + \boldsymbol{\mu} + \epsilon_\alpha,$$

where

$$\mathbf{x}_\alpha \sim \mathcal{N}_q(\mathbf{0}, \mathbf{I}) \qquad \text{and} \qquad \epsilon_\alpha \sim \mathcal{N}_p(\mathbf{0}, \sigma^2\mathbf{I})$$

are independent for some $\sigma^2 > 0$ and $q < \min\{N, p\}$.

We target to estimate parameters

$$\mathbf{W} \in \mathbb{R}^{p \times q}, \quad \boldsymbol{\mu} \in \mathbb{R}^p \quad \text{and} \quad \sigma \in (0, +\infty)$$

by maximum likelihood estimation for given $\mathbf{y}_1, \ldots, \mathbf{y}_N$.

Consider that

$$\mathbf{y}_\alpha \sim \mathcal{N}_p(\boldsymbol{\mu}, \mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I}).$$

We construct the likelihood function

$$\begin{aligned} & L(\boldsymbol{\mu}, \mathbf{W}, \sigma^2) \\ = & \prod_{\alpha=1}^{N} \frac{1}{\sqrt{(2\pi)^p \det(\boldsymbol{\Sigma})}} \exp\left(-\frac{1}{2}(\mathbf{y}_\alpha - \boldsymbol{\mu})^\top (\mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I})^{-1}(\mathbf{y}_\alpha - \boldsymbol{\mu})\right), \end{aligned}$$

then we have

$$\begin{aligned} & \ln L(\boldsymbol{\mu}, \mathbf{W}, \sigma^2) \\ \propto & -\frac{N}{2}\ln\det(\mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I}) - \frac{1}{2}\sum_{\alpha=1}^{N}(\mathbf{y}_\alpha - \boldsymbol{\mu})^\top(\mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I})^{-1}(\mathbf{y}_\alpha - \boldsymbol{\mu}). \end{aligned}$$

The Maximum Likelihood Estimators

The maximum likelihood estimators of $\boldsymbol{\mu}$, $\mathbf{W}$ and σ^2 are

$$\hat{\boldsymbol{\mu}} = \bar{\mathbf{y}} = \frac{1}{N}\sum_{\alpha=1}^{N}\mathbf{y}_\alpha, \quad \hat{\mathbf{W}} = \mathbf{U}_q(\boldsymbol{\Lambda}_q - \hat{\sigma}^2\mathbf{I})\mathbf{R} \quad \text{and} \quad \hat{\sigma}^2 = \frac{1}{p-q}\sum_{j=q+1}^{p}\lambda_j,$$

where

1. $\boldsymbol{\Lambda}_q \in \mathbb{R}^{q\times q}$ is diagonal with the largest q eigenvalues $\lambda_1, \dots, \lambda_q$ of

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{N}\sum_{\alpha=1}^{N}(\mathbf{y}_\alpha - \bar{\mathbf{y}})(\mathbf{y}_\alpha - \bar{\mathbf{y}})^\top;$$

2. $\mathbf{U}_q \in \mathbb{R}^{p\times q}$ is orthogonal column consisting of the eigenvectors associate with $\lambda_1, \dots, \lambda_q$;
3. $\mathbf{R} \in \mathbb{R}^{q\times q}$ is any orthogonal matrix.

The Maximum Likelihood Estimators

The maximum likelihood estimators also minimize the error with respect to Frobenius norm

$$(\hat{\mathbf{W}},\ \hat{\sigma}^2) = \underset{\mathbf{W}\in\mathbb{R}^{p\times q},\sigma^2\in\mathbb{R}^+}{\arg\min} \left\|\hat{\mathbf{\Sigma}} - \left(\mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I}\right)\right\|_F.$$

The Expectation-Maximization Algorithm

For the model

$$\mathbf{y} = \mathbf{W}\mathbf{x} + \boldsymbol{\mu} + \epsilon,$$

where $\mathbf{x} \sim \mathcal{N}_q(\mathbf{0}, \mathbf{I})$ and $\epsilon \sim \mathcal{N}_p(\mathbf{0}, \sigma^2\mathbf{I})$ are independent.

We regard $\{\mathbf{x}_\alpha\}_{\alpha=1}^N$ as missing data and $\{\mathbf{x}_\alpha, \mathbf{y}_\alpha\}_{\alpha=1}^N$ as the complete data, then we can achieve

$$\mathbf{y}_\alpha \,|\, \mathbf{x}_\alpha \sim \mathcal{N}_p(\mathbf{W}\mathbf{x}_\alpha + \boldsymbol{\mu}, \sigma^2\mathbf{I})$$

and

$$\mathbf{x}_\alpha \,|\, \mathbf{y}_\alpha \sim \mathcal{N}_q(\mathbf{M}^{-1}\mathbf{W}^\top(\mathbf{y}_\alpha - \boldsymbol{\mu}), \sigma^2\mathbf{M}^{-1}),$$

where $\mathbf{M} = \mathbf{W}^\top\mathbf{W} + \sigma^2\mathbf{I}$.

The Expectation-Maximization Algorithm

The update of the EM algorithm

1. In E-step, we take the expectation

$$l_C = \mathbb{E}\left[\ln\left(\prod_{\alpha=1}^{N} f(\mathbf{x}_\alpha \,|\, \mathbf{y}_\alpha)\right)\right].$$

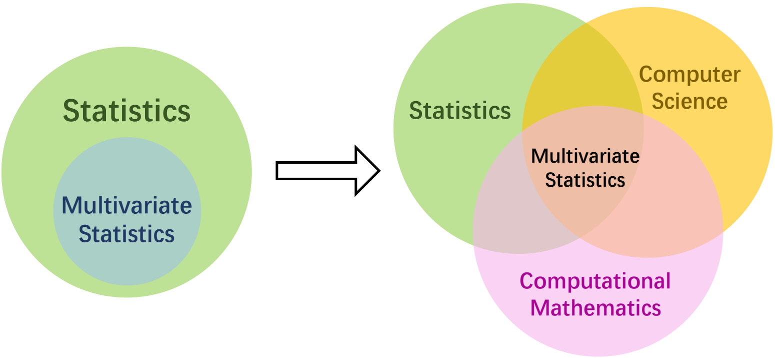
2. In the M-step, we maximized l_C with respect to $\mathbf{W}$ and σ^2:

$$\begin{aligned}\mathbf{W}_+ =& \hat{\mathbf{\Sigma}}\mathbf{W}(\sigma^2\mathbf{I} + \mathbf{M}^{-1}\mathbf{W}^\top\hat{\mathbf{\Sigma}}\mathbf{W})^{-1}, \\ \sigma_+^2 =& \frac{1}{p}\mathrm{tr}\left(\hat{\mathbf{\Sigma}} - \hat{\mathbf{\Sigma}}\mathbf{W}\mathbf{M}^{-1}\mathbf{W}_+^\top\right).\end{aligned}$$

Note that the computational complexity of EM is $\mathcal{O}(Npq)$, while the spectral decomposition in MLE requires $\mathcal{O}(Np^2 + p^3)$.

Outline

Multivariate Statistics



Good Luck on Finals!