Multivariate Statistical Analysis

Lecture 08

Fudan University

luoluo@fudan.edu.cn

Outline

Outline

1. Asymptotic Normality
2. Bayesian Estimation
3. James–Stein Estimator

Asymptotic Normality

Let $x_1, \dots, x_n$ be independent and identically distributed random variables with the same arbitrary distribution, mean μ, and variance σ^2.

Let $\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$, then the random variable

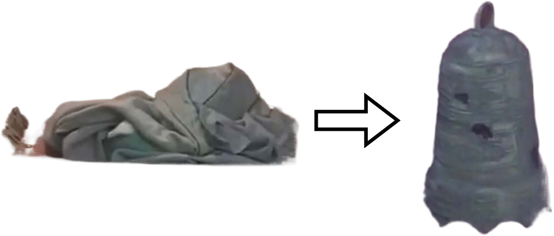
$$z = \lim_{n\to\infty} \sqrt{n}\left(\frac{\bar{x}_n - \mu}{\sigma}\right)$$

is a standard normal distribution.

What about multivariate case?

Asymptotic Normality

$$\lim_{n\to+\infty}\frac{1}{n}\sum_{i=1}^{n}$$



Multivariate Central Limit Theorem

Theorem

Let p-component vectors $\mathbf{y}_1, \mathbf{y}_2, \dots$ *be i.i.d with means* $\mathbb{E}[\mathbf{y}_\alpha] = \boldsymbol{\nu}$ *and covariance matrices* $\mathbb{E}[(\mathbf{y}_\alpha - \boldsymbol{\nu})(\mathbf{y}_\alpha - \boldsymbol{\nu})^\top] = \mathbf{T}$*. Then the limiting distribution of*

$$\frac{1}{\sqrt{n}} \sum_{\alpha=1}^{n} (\mathbf{y}_\alpha - \boldsymbol{\nu})$$

as $n \to +\infty$ *is* $\mathcal{N}(\mathbf{0}, \mathbf{T})$.

Characteristic Function and Probability

Theorem

Let $\{F_j(\mathbf{x})\}$ be a sequence of cdfs, and let $\{\phi_j(\mathbf{t})\}$ be the sequence of corresponding characteristic functions. A necessary and sufficient condition for $F_j(\mathbf{x})$ to converge to a cdf $F(\mathbf{x})$ is that, for every $\mathbf{t}$, $\phi_j(\mathbf{t})$ converges to a limit $\phi(\mathbf{t})$ that is continuous at $\mathbf{t} = \mathbf{0}$. When this condition is satisfied, the limit $\phi(\mathbf{t})$ is identical with the characteristic function of the limiting distribution $F(\mathbf{x})$.

Outline

Revisiting Linear Regression

Given dataset $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$, where $\mathbf{x}_i \in \mathbb{R}^p$ and $y_i \in \mathbb{R}$ are the feature and the corresponding label of the i-th data.

We suppose

$$y_i = \boldsymbol{\beta}^\top \mathbf{x}_i + \epsilon_i$$

with

$$\boldsymbol{\beta} \in \mathbb{R}^p \qquad \text{and} \qquad \epsilon_i \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2)$$

for $i = 1, \dots, N$, where $\sigma > 0$.

Revisiting Linear Regression

Maximizing the likelihood function leads to optimization problem

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \frac{1}{2}\left\|\mathbf{X}\boldsymbol{\beta}-\mathbf{y}\right\|_2^2.$$

Suppose $\mathbf{A}^\top\mathbf{A}$ is non-singular, then

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y},$$

which has distribution

$$\hat{\boldsymbol{\beta}} \sim \mathcal{N}_p\big(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}\big).$$

Revisiting Linear Regression

We define the sample error as

$$\hat{\boldsymbol{\epsilon}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}},$$

which is uncorrelated to $\hat{\boldsymbol{\beta}}$.

Ridge Regression

In Bayesian statistics, we regard the parameters as a random variable with prior distribution.

For linear regression, we additionally suppose the parameter has a prior distribution

$$\boldsymbol{\beta} \sim \mathcal{N}_p(\mathbf{0}, \tau^2 \mathbf{I}),$$

which leads to optimization problem

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \frac{1}{2} \|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|_2^2 + \frac{\sigma^2}{2\tau^2} \|\boldsymbol{\beta}\|_2^2 .$$

Bayesian Estimation

Theorem

If $\mathbf{x}_1, \ldots, \mathbf{x}_N$ are independently distributed and each $\mathbf{x}_\alpha$ has distribution $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and if $\boldsymbol{\mu}$ has an a prior distribution $\mathcal{N}(\boldsymbol{\nu}, \boldsymbol{\Phi})$, then the a posterior distribution of $\boldsymbol{\mu}$ given $\mathbf{x}_1, \ldots, \mathbf{x}_N$ is normal with mean

$$\boldsymbol{\Phi}\left(\boldsymbol{\Phi} + \frac{1}{N}\boldsymbol{\Sigma}\right)^{-1}\bar{\mathbf{x}} + \frac{1}{N}\boldsymbol{\Sigma}\left(\boldsymbol{\Phi} + \frac{1}{N}\boldsymbol{\Sigma}\right)^{-1}\boldsymbol{\nu}$$

and covariance matrix

$$\boldsymbol{\Phi} - \boldsymbol{\Phi}\left(\boldsymbol{\Phi} + \frac{1}{N}\boldsymbol{\Sigma}\right)^{-1}\boldsymbol{\Phi}.$$

Outline

The Biased Estimator

The sample mean $\bar{\mathbf{x}}$ seems the natural estimator of the population mean $\boldsymbol{\mu}$.

However, Stein (1956) showed $\bar{\mathbf{x}}$ is not admissible with respect to the mean squared loss when $p \geq 3$.

James–Stein Estimator

Consider the loss function

$$L(\boldsymbol{\mu}, \mathbf{m}) = \|\mathbf{m} - \boldsymbol{\mu}\|_2^2 \,,$$

where $\mathbf{m}$ is an estimator of the mean $\boldsymbol{\mu}$.

The estimator proposed by James and Stein is

$$\mathbf{m}(\bar{\mathbf{x}}) = \left(1 - \frac{p-2}{\|\bar{\mathbf{x}} - \boldsymbol{\nu}\|_2^2}\right)(\bar{\mathbf{x}} - \boldsymbol{\nu}) + \boldsymbol{\nu},$$

where $\boldsymbol{\nu} \in \mathbb{R}^p$ is an arbitrary fixed vector and $p \geq 3$.

Bayesian Estimation View

Consider $\mathbf{x}_\alpha \sim \mathcal{N}(\boldsymbol{\mu}, N\mathbf{I})$ for $\alpha = 1, \dots, N$, we additionally suppose

$$\boldsymbol{\mu} \sim \mathcal{N}(\boldsymbol{\nu}, \tau^2\mathbf{I}).$$

Then the posterior distribution of $\boldsymbol{\mu}$ given $\mathbf{x}_1, \dots, \mathbf{x}_N$ has mean

$$\left(1 - \mathbb{E}\left[\frac{p-2}{\|\bar{\mathbf{x}} - \boldsymbol{\nu}\|_2^2}\right]\right)(\bar{\mathbf{x}} - \boldsymbol{\nu}) + \boldsymbol{\nu}.$$

James–Stein Estimator

Interestingly, we have

$$\mathbb{E}\left[\|\mathbf{m}(\bar{\mathbf{x}}) - \boldsymbol{\mu}\|_2^2\right] < \mathbb{E}\left[\|\bar{\mathbf{x}} - \boldsymbol{\mu}\|_2^2\right]$$

by only suppose $\mathbf{x}_\alpha \sim \mathcal{N}(\boldsymbol{\mu}, N\mathbf{I})$ without prior on $\boldsymbol{\mu}$, where

$$\mathbf{m}(\bar{\mathbf{x}}) = \left(1 - \frac{p-2}{\|\bar{\mathbf{x}} - \boldsymbol{\nu}\|_2^2}\right)(\bar{\mathbf{x}} - \boldsymbol{\nu}) + \boldsymbol{\nu}.$$

Improved Biased Estimator

The James–Stein estimator is

$$\mathbf{m}(\bar{\mathbf{x}}) = \left(1 - \frac{p-2}{\|\bar{\mathbf{x}} - \boldsymbol{\nu}\|_2^2}\right)(\bar{\mathbf{x}} - \boldsymbol{\nu}) + \boldsymbol{\nu}.$$

For small values of $\|\bar{\mathbf{x}} - \boldsymbol{\nu}\|_2$, the multiplier of $(\bar{\mathbf{x}} - \boldsymbol{\nu})$ is negative; that is, the estimator $\mathbf{m}(\bar{\mathbf{x}})$ is in the direction from $\boldsymbol{\nu}$ opposite to that of $\bar{\mathbf{x}}$.

We can improve $\mathbf{m}(\bar{\mathbf{x}})$ by using

$$\tilde{\mathbf{m}}(\bar{\mathbf{x}}) = \left(1 - \frac{p-2}{\|\bar{\mathbf{x}} - \boldsymbol{\nu}\|_2^2}\right)^+ (\bar{\mathbf{x}} - \boldsymbol{\nu}) + \boldsymbol{\nu},$$

which holds that $\mathbb{E}\left[\|\tilde{\mathbf{m}}(\bar{\mathbf{x}}) - \boldsymbol{\mu}\|_2^2\right] \leq \mathbb{E}\left[\|\mathbf{m}(\bar{\mathbf{x}}) - \boldsymbol{\mu}\|_2^2\right]$.