STAT 3690 Lecture 08

zhiyanggeezhou.github.io

Zhiyang Zhou (zhiyang.zhou@umanitoba.ca)

Feb 9, 2022

Assumptions

- Model: $\mathbf{X}_1, \ldots, \mathbf{X}_n \overset{\text{iid}}{\sim} MVN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $n > p$
- Parameter space: $\Theta = \{(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \mid \boldsymbol{\mu} \in \mathbb{R}^p, \boldsymbol{\Sigma} \in \mathbb{R}^{p\times p}, \boldsymbol{\Sigma} > 0\}$

Method of moments (MM) estimators for $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

- No requirement on normality
- Steps
  1. Equate raw moments to their sample counterparts:
  $$\begin{cases} \mathrm{E}(\mathbf{X}) = \bar{\mathbf{X}} \\ \mathrm{E}(\mathbf{X}\mathbf{X}^\top) = n^{-1}\sum_i \mathbf{X}_i\mathbf{X}_i^\top \end{cases} \Leftrightarrow \begin{cases} \boldsymbol{\mu} = \bar{\mathbf{X}} \\ \boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^\top = n^{-1}\sum_i \mathbf{X}_i\mathbf{X}_i^\top \end{cases}$$
  2. Solve the above equations w.r.t. $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ and obtain estimators
  $$\begin{cases} \hat{\boldsymbol{\mu}}_{\mathrm{MM}} = \bar{\mathbf{X}} \\ \widehat{\boldsymbol{\Sigma}}_{\mathrm{MM}} = n^{-1}\sum_i \mathbf{X}_i\mathbf{X}_i^\top - \bar{\mathbf{X}}\bar{\mathbf{X}}^\top = n^{-1}(n-1)\mathbf{S}, \end{cases}$$
  where $\mathbf{S} = (n-1)^{-1}\sum_{i=1}^n (\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})^\top$

Maximum likelihood (ML) estimation for parameters of MVN (J&W Sec 4.3)

- Likelihood function
$$\begin{aligned} L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) =& \prod_{i=1}^n \left[\frac{1}{\sqrt{(2\pi)^p \det(\boldsymbol{\Sigma})}} \exp\Big\{ -\frac{1}{2}(\mathbf{X}_i - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{X}_i - \boldsymbol{\mu}) \Big\} \right] \\ =& \frac{1}{\sqrt{(2\pi)^{np}\{\det(\boldsymbol{\Sigma})\}^n}} \exp\Big\{ -\frac{1}{2}\sum_{i=1}^n (\mathbf{X}_i - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{X}_i - \boldsymbol{\mu}) \Big\} \end{aligned}$$
- Log likelihood
$$\ell(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \ln\{L(\boldsymbol{\mu}, \boldsymbol{\Sigma})\} = -\frac{np}{2}\ln(2\pi) - \frac{n}{2}\ln\{\det(\boldsymbol{\Sigma})\} - \frac{1}{2}\sum_{i=1}^n (\mathbf{X}_i - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{X}_i - \boldsymbol{\mu})$$

```r
options(digits = 4)
set.seed(1)
n = 1e3L
Mu = matrix(c(3, 6), ncol = 1, nrow = 2)
Sigma = matrix(c(10, 2 ,2, 5), ncol = 2, nrow = 2)
X = MASS::mvrnorm(n, Mu, Sigma)

loglik <- function(Mu, Sigma, data = X) {
  n = nrow(data)
  p = length(Mu)
  X_center = sweep(X, 2, Mu)
  return(
    (-n*p/2)*log(2*pi)+
      (-n/2)*log(det(Sigma))+
      (-1/2)*sum(diag(X_center %*% solve(Sigma) %*% t(X_center)))
  )
}

grid_xy <- expand.grid(
  seq(Mu[1]-2*Sigma[1,1]^.5, Mu[1]+2*Sigma[1,1]^.5, length.out = 32),
  seq(Mu[2]-2*Sigma[2,2]^.5, Mu[2]+2*Sigma[2,2]^.5, length.out = 32))

contours <- purrr::map_df(
  seq_len(nrow(grid_xy)),
  function(i) {
    # Where we will evaluate loglik
    mu_obs <- as.numeric(grid_xy[i,])
    # Evaluate at the pop covariance
    z <- loglik(mu_obs, Sigma, X)
    # Output data.frame
    data.frame(x = mu_obs[1],
               y = mu_obs[2],
               z = z)
  })

# Contour plot
library(tidyverse)
library(ggrepel)
data_means <- data.frame(
  x = c(Mu[1], mean(X[,1])),
  y = c(Mu[2], mean(X[,2])),
  label = c("Mu", "Sample Mean"))

contours %>%
  ggplot(aes(x, y)) +
  geom_contour(aes(z = z)) +
  geom_point(data = data_means, aes(color = label)) +
  geom_label_repel(data = data_means, aes(label = label))

# 3d scatter plot
library(scatterplot3d)
with(contours, scatterplot3d(x, y, z))
```

- ML estimator
$$(\hat{\boldsymbol{\mu}}_{\mathrm{ML}}, \widehat{\boldsymbol{\Sigma}}_{\mathrm{ML}}) = \arg\max_{(\boldsymbol{\mu},\boldsymbol{\Sigma})\in\Theta} \ell(\boldsymbol{\mu},\boldsymbol{\Sigma}) = (\bar{\mathbf{X}}, n^{-1}(n-1)\mathbf{S})$$

Derive $(\hat{\mu}_{ML}, \hat{\Sigma}_{ML})$

$$l(\mu, \Sigma) = const - \frac{n}{2}\ln\{\det(\Sigma)\} - \frac{1}{2}\underbrace{\sum_{i=1}^{n}(X_i-\mu)^T\Sigma^{-1}(X_i-\mu)}_{①}$$

$$\begin{aligned}
① &= \sum_{i=1}^{n} tr\{(X_i-\mu)^T\Sigma^{-1}(X_i-\mu)\} \\
&= \sum_{i=1}^{n} tr\{\Sigma^{-1}(X_i-\mu)(X_i-\mu)^T\} \qquad (\because tr(ABC) = tr(BCA)) \\
&= tr\Big\{\sum_{i=1}^{n}\Sigma^{-1}(X_i-\bar{X}+\bar{X}-\mu)(X_i-\bar{X}+\bar{X}-\mu)^T\Big\} \\
&= tr\Big[\Sigma^{-1}\Big\{\sum_{i=1}^{n}(X_i-\bar{X})(X_i-\bar{X})^T + (\bar{X}-\mu)\underbrace{\sum_{i=1}^{n}(X_i-\bar{X})^T}_{0} + \underbrace{\sum_{i=1}^{n}(X_i-\bar{X})}_{0}(\bar{X}-\mu)^T + n(\bar{X}-\mu)(\bar{X}-\mu)^T\Big\}\Big] \\
&= tr\big[\Sigma^{-1}\{(n-1)S + n(\bar{X}-\mu)(\bar{X}-\mu)^T\}\big] \\
&= (n-1)\,tr(\Sigma^{-1}S) + n\underbrace{(\bar{X}-\mu)^T\Sigma^{-1}(\bar{X}-\mu)}_{②} \qquad (\because tr(ABC) = tr(BCA))
\end{aligned}$$

So, no matter what $\Sigma > 0$ is, $l(\mu, \Sigma)$ is maximized at $\hat{\mu}_{ML} = \bar{X}$. We then have

$$l(\hat{\mu}_{ML}, \Sigma) = const - \frac{n}{2}\ln\{\det(\Sigma)\} - \frac{n-1}{2}tr(\Sigma^{-1}S)$$

$$\begin{aligned}
\Rightarrow \frac{\partial l(\hat{\mu}_{ML}, \Sigma)}{\partial \Sigma} &= -\frac{n}{2}\{\det(\Sigma)\}^{-1}\det(\Sigma)(\Sigma^{-1})^T \qquad (\because \partial \det(A)/\partial A = \det(A)(A^{-1})^T) \\
&\quad - \frac{n-1}{2}(-S^T\Sigma^{-2}) \qquad (\because \partial tr(AB)/\partial A = B^T,\ \frac{\partial A^{-1}}{\partial A} = -A^{-2}) \\
&= -\frac{n}{2}\Sigma^{-1} + \frac{n-1}{2}S\Sigma^{-2}
\end{aligned}$$

Let $\partial l(\hat{\mu}_{ML}, \Sigma)/\partial\Sigma = 0$. Then $\hat{\Sigma}_{ML} = \frac{n-1}{n}S$

NOT trivial to verify that $(\hat{\mu}_{M}, \hat{\Sigma}_{ML})$ is the only maximizer.

- Properties of $(\hat{\boldsymbol{\mu}}_{\mathrm{ML}}, \widehat{\boldsymbol{\Sigma}}_{\mathrm{ML}})$
  - Consistency: $(\hat{\boldsymbol{\mu}}_{\mathrm{ML}}, \widehat{\boldsymbol{\Sigma}}_{\mathrm{ML}}) \xrightarrow{P} (\boldsymbol{\mu}, \boldsymbol{\Sigma})$.
  - Efficiency: As $n \to \infty$, the covariance of $(\hat{\boldsymbol{\mu}}_{\mathrm{ML}}, \widehat{\boldsymbol{\Sigma}}_{\mathrm{ML}})$ achieves the Cramer-Rao lower bound.
  - Invariance: For any function g, the ML estimator of $g(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is $g(\hat{\boldsymbol{\mu}}_{\mathrm{ML}}, \widehat{\boldsymbol{\Sigma}}_{\mathrm{ML}})$.