MH4501 Multivariate Analysis

- Midterm Revision -

Naoki Honda

March 2019

1 Multivariate Population and Sample Statistics

Population Covariance Matrix

$$\begin{aligned}\mathbf{\Sigma} = Cov(X) &= \mathbf{E}[(X-\mu)(X-\mu)^T] \\ &= \begin{pmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \dots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \dots & \sigma_{pp} \end{pmatrix}\end{aligned}$$

Population Correlation Matrix

$$\rho_{jk} = cor(X_j, X_k) = \frac{\sigma_{jk}}{\sqrt{\sigma_{jj}}\sqrt{\sigma_{kk}}}$$

$$\rho = \begin{pmatrix} 1 & \rho_{12} & \dots & \rho_{1p} \\ \rho_{21} & 1 & \dots & \rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{p1} & \rho_{p2} & \dots & 1 \end{pmatrix}$$

1.1 Population Covariance Matrix and Correlation Matrix

Let

$$V = diag(\sigma_{11}, \sigma_{22}, ..., \sigma_{pp}) = \begin{pmatrix} \sigma_{11} & 0 & \dots & 0 \\ 0 & \sigma_{22} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_{pp} \end{pmatrix}$$

$$V^{1/2} = diag(\sqrt{\sigma_{11}}, \sqrt{\sigma_{22}}, ..., \sqrt{\sigma_{pp}})$$

$$V^{-1/2} = (V^{1/2})^{-1}$$

We have:

$$\begin{aligned} \mathbf{\Sigma} &= V^{1/2}\rho V^{1/2} \\ \rho &= V^{-1/2}\mathbf{\Sigma}\, V^{-1/2} \end{aligned}$$

1.2 Linear Transformation

When another r.v. Y is defined as $Y_{q\times 1} := AX + b$,
we have:

Linear Transformation

$$\mu_Y = A\mu_X + b$$
$$\mathbf{\Sigma}_Y = A\mathbf{\Sigma}_X A^T$$

1.3 Sample Mean Vector

Sample Mean Vector

$$\bar{x} = \frac{1}{n}\begin{pmatrix} x_{11} & x_{21} & \dots & x_{n1} \\ x_{12} & x_{22} & \dots & x_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1p} & x_{2p} & \dots & x_{np} \end{pmatrix}\begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} = \frac{1}{n}X^T\mathbf{1}_{n\times 1}$$

1.4 Sample Covariance Matrix

Sample Covariance Matrix

$$
\begin{aligned}
SSCP = (n-1)S &= \begin{pmatrix} \sum_{i=1}^n (x_{i1}-\bar{x}_1)^2 & \dots & \sum_{i=1}^n (x_{i1}-\bar{x}_1)(x_{ip}-\bar{x}_p) \\ \vdots & \ddots & \vdots \\ \sum_{i=1}^n (x_{ip}-\bar{x}_p)(x_{i1}-\bar{x}_1) & \dots & \sum_{i=1}^n (x_{ip}-\bar{x}_p)^2 \end{pmatrix} \\
&= X^T(I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^T)X \\
&= X^TX - n\bar{x}\bar{x}^T
\end{aligned}
$$

2 Multivariate Normal Distribution: MVN

Probability Density Function of Multivariate Normal Distribution

$$f(x) = \frac{1}{(2\pi)^{p/2} det(\Sigma)^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$$

2.1 Properties of MVN

Linear Transformation

Linear transformation gives a new MVN

$$AX + b \sim N_q(A\mu + b, A\Sigma A^T)$$

Partition and independence

In the situation where

$$X = \begin{pmatrix} X^{(1)}_{p\times 1} \\ X^{(2)}_{q\times 1} \end{pmatrix} \sim N_{p+q}\left(\mu = \begin{pmatrix} \mu^{(1)}_{p\times 1} \\ \mu^{(2)}_{q\times 1} \end{pmatrix}, \Sigma = \begin{pmatrix} \Sigma^{(11)}_{p\times p} & \Sigma^{(12)}_{p\times q} \\ \Sigma^{(21)}_{q\times p} & \Sigma^{(22)}_{q\times q} \end{pmatrix}\right)$$

We have:

$$X^{(1)} \text{ and } X^{(2)} \text{ are independent} \Leftrightarrow cov(X^{(1)}, X^{(2)}) = \Sigma^{(12)} = \mathbf{0}_{p\times q}$$

2.2 MVN Sampling Distribution

Distribution of Mean Vector

$$\bar{x} \sim N_p\left(\mu, \frac{1}{n}\Sigma\right) \qquad \text{or equivalently,} \qquad \sqrt{n}(\bar{x} - \mu) \sim N_p\left(0_p, \Sigma\right)$$

Whishart Distribution

$SSCP = (n-1)S \sim W_p(n-1, \Sigma)$, the Wishart distribution.

- When $p = 1$ and $\Sigma_{1\times 1} = 1$, $W_1(n, 1)$ reduces to the $\chi^2(n)$
- If $W_1 \sim W_p(n_1, \Sigma)$, $W_2 \sim W_p(n_2, \Sigma)$, and W_1 and W_2 are independent, then $W_1 + W_2 \sim W_p(n_1 + n_2, \Sigma)$
- If $W \sim W_p(n, \Sigma)$, and $B_{q\times p}$ is a given matrix, then $BWB^T \sim W_q(n, B\Sigma B^T)$

Hotelling's T^2

Let $x_1, x_2, ..., x_n$ be n independent realizations from $N_p(\mu, \Sigma)$, then

- $n(\bar{x} - \mu)^T \Sigma^{-1}(\bar{x} - \mu) \sim \chi^2(p)$
- $T^2 := n(\bar{x} - \mu)^T S^{-1}(\bar{x} - \mu)$ is called **Hotelling's T^2 statistic**
- $\frac{n-p}{p(n-1)}T^2 \sim F(p, n-p)$
- When n is large (relative to q),
  - $T^2 \sim \chi^2(p)$
  - $d_i^2 = (x_i - \bar{x})^T S^{-1}(x_i - \bar{x}) \sim \chi^2(p)$

3 Multivariate Statistical Inference

3.1 Confidence Region of Mean Vector

A region $\mathcal{R} \subset \mathbb{R}^p$ is said to be a confidence region of $\mu_{p\times 1}$
at confidence level $1-\alpha$, if $Pr[\mathcal{R} \ni \mu] = 1-\alpha$

Since $T^2 = n(\bar{x}-\mu)^T S^{-1}(\bar{x}-\mu)$ and $\frac{n-p}{p(n-1)}T^2 \sim F(p, n-p)$, we have

Confidence Region

$$\mathcal{R} = \left\{ \mu : (\bar{x}-\mu)^T S^{-1}(\bar{x}-\mu) < \frac{p(n-1)}{n(n-p)} F_\alpha[p, n-p] \right\}$$

3.2 Simultaneous Confidence Interval: Bonferroni method

Bonferroni's SCI

$$\mathcal{R}_j = \left(\bar{x}_j - t_{\alpha^*/2}[n-1]\sqrt{\frac{s_{jj}}{n}}, \bar{x}_j + t_{\alpha^*/2}[n-1]\sqrt{\frac{s_{jj}}{n}} \right)$$

where $\alpha^* = \alpha/p$.

3.3 Situation when n is large

CR and SCI; when n is large

- CR:

$$\mathcal{R} = \left\{ \mu : (\bar{x}-\mu)^T S^{-1}(\bar{x}-\mu) < \frac{1}{n}\chi^2_\alpha[p] \right\}$$

- SCI:

$$\mathcal{R}_j = \left(\bar{x}_j - z_{\alpha^*/2}\sqrt{\frac{s_{jj}}{n}}, \bar{x}_j + z_{\alpha^*/2}\sqrt{\frac{s_{jj}}{n}} \right)$$

3.4 Two-Sample Comparison: Paired Observations

Consider

$$d_1 = x_{11} - x_{21}, d_2 = x_{12} - x_{22}, ..., d_n = x_{1n} - x_{2n}$$

And test $H_0 : \mu_d = \mathbf{0}_{p\times 1}$

Statistics for paired observations

If we assume d_i are from MVN, same as one-sample, we have

$$T^2 = n\bar{d}^T S_d^{-1} \bar{d}$$

$$\frac{n-p}{p(n-1)} T^2 \sim F(p, n-p)$$

The arguments for CR and SCI are also same as one for the one-sample, changing $\bar{x} \to \bar{d}$

3.5 Two-Sample Comparison: Unpaired Observations

Assumptions

- Both sample 1 and 2 are independent
- Both Population 1 and 2 are MVN
- Two populations have the same population covariance matrix: $\Sigma_1 = \Sigma_2 = \Sigma$

We use **Pooled Sample Covariance Matrix**:

Pooled Sample Covariance Matrix

$$S_{pool} = \frac{(n_1-1)S_1 + (n_2-1)S_2}{n_1+n_2-2} = \frac{SSCP_1 + SSCP_2}{n_1+n_2-2}$$

to estimate the unknown Σ

Hoteling's T^2 statistic

$$\bar{x}_1 - \bar{x}_2 \sim N_p\left(\mu_1 - \mu_2, \left(\frac{1}{n_1} + \frac{1}{n_2}\right)\Sigma\right)$$

Thus

$$[(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)]^T \left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{-1} \Sigma^{-1}[(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)] \sim \chi^2(p)$$

and

$$T^2 = [(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)]^T \left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{-1} S_{pool}^{-1}[(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)]$$

with

$$\frac{n_1+n_2-1-p}{p(n_1+n_2-2)}T^2 \sim F(p, n_1+n_2-1-p)$$

3.5.1 Hypothesis Testing under $H_0 : \mu_1 = \mu_2$

We reject H_0 if

$$\frac{n_1 + n_2 - 1 - p}{p(n_1 + n_2 - 2)} T^2 > F_\alpha[p, n_1 + n_2 - 1 - p]$$

where α is the desired significant level.

3.5.2 CR of $\delta = \mu_1 - \mu_2$ at confidence level $1 - \alpha$

CR for unpaied two-sample

$$\mathcal{R} = \left\{ \delta : (\bar{x}_1 - \bar{x}_2 - \delta)^T \left(\frac{1}{n_1} + \frac{1}{n_2} \right)^{-1} S_{pool}^{-1} (\bar{x}_1 - \bar{x}_2 - \delta) < \frac{p(n_1 + n_2 - 2)}{n_1 + n_2 - 1 - p} F_\alpha[p, n_1 + n_2 - 1 - p] \right\}$$

3.5.3 Bonferroni SCIs of μ_d at confidence level $1 - \alpha$

Bonferroni's SCI

$$\mathcal{R}_j = \left((\bar{x}_1 - \bar{x}_2)_j - t_{\alpha^*/2}[n_1 + n_2 - 2] \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2} \right)^{-1} s_{pool,jj}}, \; (\bar{x}_1 - \bar{x}_2)_j + t_{\alpha^*/2}[n_1 + n_2 - 2] \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2} \right)^{-1} s_{pool,jj}} \right)$$

where $\alpha^* = \alpha / p$

3.5.4 When n is large

Hotelling's T^2; when n is large

$$T^2 = (\bar{x}_1 - \bar{x}_2)^T \left(\frac{S_1}{n_1} + \frac{S_2}{n_2}\right)^{-1} (\bar{x}_1 - \bar{x}_2) \sim \chi^2(p), \quad \text{under } H_0 : \mu_1 = \mu_2$$

CR for unpaied two-sample; when n is large

$$\mathcal{R} = \left\{\delta : (\bar{x}_1 - \bar{x}_2 - \delta)^T \left(\frac{S_1}{n_1} + \frac{S_2}{n_2}\right)^{-1} (\bar{x}_1 - \bar{x}_2 - \delta) < \chi^2_\alpha[p]\right\}$$

Bonferroni's SCI; when n is large

$$\mathcal{R}_j = \left((\bar{x}_1 - \bar{x}_2)_j - z_{\alpha^*/2}\sqrt{\frac{s_{1,jj}}{n_1} + \frac{s_{2,jj}}{n_2}}, (\bar{x}_1 - \bar{x}_2)_j + z_{\alpha^*/2}\sqrt{\frac{s_{1,jj}}{n_1} + \frac{s_{2,jj}}{n_2}}\right)$$

where $\alpha^* = \alpha/p$

4 MANOVA: Multivariate Analysis of Variance

4.1 Step 1

We want to test the null hypothesis $\underline{H_0 : \mu_1 = \mu_2 = ... = \mu_G}$
Decompose population mean vector of each population as

Treatment Effect

$$\mu_g = \mu + (\mu_g - \mu) = \mu + \tau_g$$

where $\tau_g = \mu_g - \mu$ is the treatment effect of Population g.
Therefore, it is equivalent to test $\underline{H_0 : \tau_1 = \tau_2 = ... = \tau_g = \mathbf{0}_{p\times 1}}$

4.2 Step 2

Furthermore, let X_g be any realization/observation of Population g, then we can decompose X_g as

Residual

$$X_g = \mu + \tau_g + (X_g - \mu_g) = \mu + \tau_g + e_g$$

where $e_g = X_g - \mu_g \sim N_p(\mathbf{0}_{p\times 1}, \Sigma)$ is the term of random error.

4.3 Step 3

Now assume we have a sample from each of the populations:
For Sample g, we have n_g realizations from Population g: $N_p(\mu_g, \Sigma)$

Following the discussion on the precious page, we decompose every x_{gi} as

Realization Decomposition

$$x_{gi} = \underbrace{\bar{x}}_{\text{overall sample mean vector}} + \underbrace{(\bar{x}_g - \bar{x})}_{\text{estimated treatment effect}} + \underbrace{(x_{gi} - \bar{x}_g)}_{\text{residual}}$$

4.4 Step 4

Following the decomposition of x_{gi}, we have $x_{gi} - \bar{x} = (\bar{x}_g - \bar{x}) + (x_{gi} - \bar{x}_g)$. Therefore [1],

$$\sum_{i=1}^{n_g}(x_{gi} - \bar{x})(x_{gi} - \bar{x})^T = n_g(\bar{x}_g - \bar{x})(\bar{x}_g - \bar{x})^T + \sum_{i=1}^{n_g}(x_{gi} - \bar{x}_g)(x_{gi} - \bar{x}_g)^T$$

4.5 Step 5

If we further sum over $g = 1, 2, ..., G$, we have

SSCP Decomposition

$$\underbrace{\sum_{g=1}^{G}\sum_{i=1}^{n_g}(x_{gi} - \bar{x})(x_{gi} - \bar{x})^T}_{SSCP_{tot}\text{: total SSCP}} = \underbrace{\sum_{g=1}^{G} n_g(\bar{x}_g - \bar{x})(\bar{x}_g - \bar{x})^T}_{SSCP_{tr}\text{: treatment SSCP}} + \underbrace{\sum_{g=1}^{G}\sum_{i=1}^{n_g}(x_{gi} - \bar{x}_g)(x_{gi} - \bar{x}_g)^T}_{SSCP_{res}\text{: residual SSCP}}$$

where "SSCP" is short for "sum of squares and cross products". Now we can eventually construct the MANOVA table:

Source of variation	SSCP	df
Treatments	$B = \sum_{g=1}^{G} n_g(\bar{x}_g - \bar{x})(\bar{x}_g - \bar{x})^T$	$G - 1$
Residuals	$W = \sum_{g=1}^{G}\sum_{i=1}^{n_g}(x_{gi} - \bar{x}_g)(x_{gi} - \bar{x}_g)^T$	$\sum_{g=1}^{G} n_g - G$
Total	$B + W = \sum_{g=1}^{G}\sum_{i=1}^{n_g}(x_{gi} - \bar{x})(x_{gi} - \bar{x})^T$	$\sum_{g=1}^{G} n_g - 1$

[1] Noticing that $\sum_{i=1}^{n_g}(\bar{x}_g - \bar{x})(x_{gi} - \bar{x}_g)^T = \mathbf{0}_{p\times p}$ and $\sum_{i=1}^{n_g}(x_{gi} - \bar{x}_g)(\bar{x}_g - \bar{x})^T = \mathbf{0}_{p\times p}$. (The sums of crossing terms are zero matrix.)

4.6 Some Remarks

- $B+W = \sum_{g=1}^{G}\sum_{i=1}^{n_g}(x_{gi}-\bar{x})(x_{gi}-\bar{x})^T$ is the total SSCP, that is, if we merge all the groups together and see it as **one sample with size** $n=\sum_{g=1}^{G} n_g$, $B+W$ is just the SSCP of this sample.

- "W" is short for "within". $W = \sum_{g=1}^{G}\sum_{i=1}^{n_g}(x_{gi}-\bar{x}_g)(x_{gi}-\bar{x}_g)^T$ is the residual SSCP, it can be equivalently expressed as:

$$W = \sum_{g=1}^{G} SSCP_g = \sum_{g=1}^{G}(n_g-1)S_g$$

- "B" is short for "between". $B = \sum_{g=1}^{G} n_g(\bar{x}_g-\bar{x})(\bar{x}_g-\bar{x})^T$ is the treatment SSCP, it equals the total SSCP of an imaginary sample with size $n=\sum_{g=1}^{G} n_g$:

$$\underbrace{\bar{x}_1,\bar{x}_1,...,\bar{x}_1}_{n_1},\underbrace{\bar{x}_2,\bar{x}_2,...,\bar{x}_2}_{n_2},...,\underbrace{\bar{x}_G,\bar{x}_G,...,\bar{x}_G}_{n_G},$$

4.7 Wilk's Lambda Statistic

The test statistic in MANOVA is the Wilk's lambda statistic:

Wilk's Lambda Statistics

$$\Lambda^* = \frac{det(W)}{det(W+B)}$$

4.8 Bartlett's Approximation

If $n=\sum_{g=1}^{G} n_g$ is large, we have the Bartlett's approximation:

Bartlett's Approximation

$$-\left(n-1-\frac{p+G}{2}\right) ln\Lambda^* \sim \chi^2(p(G-1))$$

under $H_0 : \mu_1 = \mu_2 = ... = \mu_G$.