Multivariate Statistical Analysis

Fall 2011

C. L. Williams, Ph.D.

Lecture 9 for Applied Multivariate Analysis

Outline

1. Two sample T^2 test
   - T^2 distribution in the two sample case
   - Wilk's Lambda
2. Confidence ellipses

Analogous to the univariate context, we wish to determine whether the mean vectors are comparable, more formally:

$$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 \tag{1}$$

Suppose we let $\mathbf{y}_{1i}$, $i = 1, \dots n_1$ and $\mathbf{y}_{2i}$, $i = 1, \dots n_2$ represent independent samples from two p-variate normal distribution with mean vectors $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ but with common covariance matrix $\boldsymbol{\Sigma}$ unknown, provided $\boldsymbol{\Sigma}$ is positive definite and $n > p$, given sample estimators for mean and covariance $\bar{\mathbf{y}}$ and $\mathbf{S}$ respectively.

We can then define

$$\mathbf{W}_1 = (n_1 - 1)\,\mathbf{S}_1 = \sum_{i=1}^{n_1} (\mathbf{y}_{1i} - \bar{\mathbf{y}}_1)(\mathbf{y}_{1i} - \bar{\mathbf{y}}_1)'$$
$$\mathbf{W}_2 = (n_2 - 1)\,\mathbf{S}_2 = \sum_{i=1}^{n_2} (\mathbf{y}_{2i} - \bar{\mathbf{y}}_2)(\mathbf{y}_{2i} - \bar{\mathbf{y}}_2)'$$

since each are unbiased estimators of the common covariance matrix, ie. $\mathrm{E}[(n_1 - 1)\,\mathbf{S}_1] = (n_1 - 1)\,\boldsymbol{\Sigma}$ and $\mathrm{E}[(n_2 - 1)\,\mathbf{S}_2] = (n_2 - 1)\,\boldsymbol{\Sigma}$

The T^2 statistic can be calculated as:

$$T^2 = \left(\frac{n_1 n_2}{n_1 + n_2}\right)\left(\frac{n_1 n_2}{n_1 + n_2}\right)(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)' \mathbf{S}^{-1} (\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2) \qquad (2)$$

where $\mathbf{S}^{-1}$ is the inverse of the pooled correlation matrix given by:

$$\begin{aligned} \mathbf{S} &= \frac{(n_1 - 1)\mathbf{S_1} + (n_2 - 1)\mathbf{S_2}}{n_1 + n_2 - 2} \\ &= \frac{1}{n_1 + n_2 - 2}(\mathbf{W}_1 + \mathbf{W}_2) \end{aligned}$$

given the sample estimates for covariance, $\mathbf{S_1}$ and $\mathbf{S_2}$ in the two samples.

Outline

Again, there is a simple relationship between the test statistic, T^2, and the F distribution:

Theorem

If $\mathbf{y}_{1i}$, $i = 1, \ldots n_1$ and $\mathbf{y}_{2i}$, $i = 1, \ldots n_2$ represent independent samples from two p variate normal distribution with mean vectors $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ but with common covariance matrix $\boldsymbol{\Sigma}$, provided $\boldsymbol{\Sigma}$ is positive definite and $n > p$, given sample estimators for mean and covariance $\bar{\mathbf{y}}$ and $\mathbf{S}$ respectively, then:

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

has an F distribution on p and $(n_1 + n_2 - p - 1)$ degrees of freedom.

- Essentially, we compute the test statistic, and see whether it falls within the $(1 - \alpha)$ quantile of the F distribution on those degrees of freedom.
- Note again that to ensure non-singularity of **S**, we require that $n_1 + n_2 > p$.

Characteristic form

$$T^2 = (\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)' \left[\left(\frac{1}{n_1} \frac{1}{n_2} \right) \mathbf{S}_{pl} \right]^{-1} (\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2) \tag{3}$$

Outline

1. Two sample T^2 test
   - T^2 distribution in the two sample case
   - Wilk's Lambda

2. Confidence ellipses

Wilk's Lambda

What was all that stuff about likelihood ratio's about? It turns out that it is possible to show that:

$$\Lambda^{2/n} = \left(\frac{|\hat{\boldsymbol{\Sigma}}|}{|\hat{\boldsymbol{\Sigma}}_0|} \right) = \left(1 + \frac{T^2}{n-1} \right)^{-1} \tag{4}$$

It is also possible to obtain the T^2 via union intersection methods. This is nice because it tells us a lot about the properties of the test!

Confidence ellipses

Essentially, we wish to find a region of squared Mahalanobis distance such that:

$$Pr\left((\bar{\mathbf{y}}-\boldsymbol{\mu})'\mathbf{S}^{-1}(\bar{\mathbf{y}}-\boldsymbol{\mu})\right) \leq c^2$$

and we can find c^2 as follows:

$$c^2 = \left(\frac{n-1}{n}\right)\left(\frac{p}{n-p}\right)F_{(1-\alpha),p,(n-p)}$$

where $F_{(1-\alpha),p,(n-p)}$ is the $(1-\alpha)$ quantile of the F distribution with p and $n-p$ degrees of freedom, p represents the number of variables and n the sample size.

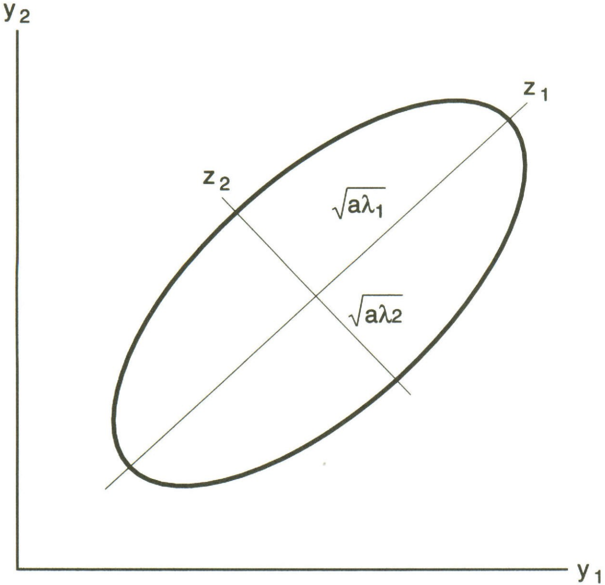
- The centroid of the ellipse is at $\bar{\mathbf{y}}$
- The half length of the semi-major axis is given by:

$$\sqrt{\lambda_1}\sqrt{\frac{p(n-1)}{n(n-p)}F_{p,n-p}(\alpha)}$$

  where λ_1 is the first eigenvalue of $\mathbf{S}$
- The half length of the semi-minor axis is given by:

$$\sqrt{\lambda_2}\sqrt{\frac{p(n-1)}{n(n-p)}F_{p,n-p}(\alpha)}$$

  where λ_2 is the second eigenvalue of $\mathbf{S}$
- The ratio of these two eigenvalues gives you some idea of the elongation of the ellipse


y2
z1
z2
√aλ1
√aλ2
y1

- In addition to the (joint) confidence ellipse, it is possible to consider *simultaneous* confidence intervals - univariate confidence intervals based on a linear combination which could be considered as shadows of the confidence ellipse
- It is also possible to carry out Bonferroni adjustments of these simultaneous intervals

- T^2 test is based upon Mahalanobis distance and can be used for inference on mean vectors - this test can be derived via a variety of routes
- Difference between univariate and multivariate inference, especially when considering confidence ellipses
- Having determined that there is a significant difference between mean vectors, you may wish to conduct a number of follow up investigations and even carry out discriminant analysis