Multivariate Statistical Analysis

Fall 2011

C. L. Williams, Ph.D.

Lecture 9 for Applied Multivariate Analysis

Outline

1. Addressing ourliers
2. Chapter 5 Tests on mean vectors

Outliers in Multivariate samples

(1) For p $>$ 2 the data cannot be readily plotted to pinpoint the outliers.

(2) Multivariate data cannot be ordered as can a univariate sample, where extremes show up readily on either end.

(3) An observation vector may have a large recording error in one of its components or smaller errors in several components.

(4) A multivariate outlier may reflect slippage in mean, variance, or correlation. This is illustrated in Figure 4.8. Observation 1 causes a small shift in means and variances of both y_1 and y_2 but has little effect on the correlation. Observation 2 has little effect on means and variances, but it reduces the correlation somewhat. Observation 3 has a major effect on means, variances, and correlation.

Wilk's Statistic

$$w = max_i \frac{|(n-2)\mathbf{S}_i|}{|(n-1)\mathbf{S}|}$$

where $\mathbf{S}_i$ is the variance-covariance matrix with the i^{th} observation deleted. Then the statistic w can be expressed in terms of

$$D^2_{(n)} = max_i (\mathbf{y}_i - \bar{\mathbf{y}})' \mathbf{S}^{-1} (\mathbf{y}_i - \bar{\mathbf{y}})$$

as

$$w = 1 - \frac{nD^2_{(n)}}{(n-1)^2}$$

Yan and Lee(1987)

$$F_i = \frac{n-p-1}{p}\left(\frac{1}{1-nD^2_{(n)}/(n-1)^2}-1\right) \quad i=1,2,\ldots,n$$

$$P(max_i F_i > f) = 1 - P(allF_i \leq f) = 1-(P(F \leq f))^n$$

An alternative form to the F-test

$$max_i F_i = F_{(n)} = \frac{n-p-1}{p}\left(\frac{1}{w}-1\right)$$

Aims for the next couple of days

- Explore the relationship between univariate t-tests and analogous tests for mean vectors (Hotelling's T^2 test)
- Examine relationships between Mahalanobis distance and T^2 distribution
- Revise confidence intervals and introduce confidence ellipses

Univariate t-test revisited

Consider the following example. Let $Y_1, \ldots, Y_n$ be a random sample from a normal distribution having unknown mean μ and known variance σ^2. We consider testing the following hypothesis:

$$\begin{aligned} H_0 : \mu &= \mu_0 \\ H_0 : \mu &\neq \mu_0 \end{aligned}$$

Consider a test at a specific level α that rejects for $|\overline{Y} - \mu_0| > C$, where C is determined so that $Pr\{|\overline{Y} - \mu_0| > C\}$ if H_0 is true: $C = \sigma_{\overline{y}} \mathcal{Z}_{\frac{\alpha}{2}}$.

The test thus does not reject when:

$$|\overline{Y} - \mu_0| \quad < \quad \sigma_{\overline{y}} \mathcal{Z}_{\frac{\alpha}{2}}$$

or

$$-\sigma_{\overline{Y}} \mathcal{Z}_{\frac{\alpha}{2}} < \quad \overline{Y} - \mu_0 \quad < \sigma_{\overline{y}} \mathcal{Z}_{\frac{\alpha}{2}}$$

or

$$\overline{Y} - \sigma_{\overline{y}} \mathcal{Z}_{\frac{\alpha}{2}} < \quad \mu_0 \quad < \overline{y} + \sigma_{\overline{y}} \mathcal{Z}_{\frac{\alpha}{2}}$$

A $(1-\alpha)100\%$ confidence interval for μ_0 is

$$[\overline{Y} - \sigma_{\overline{y}} \mathcal{Z}_{\frac{\alpha}{2}}, \overline{Y} + \sigma_{\overline{y}} \mathcal{Z}_{\frac{\alpha}{2}}]$$

Duality

The general form of a confidence interval for some unknown parameter is given by:

$$\widehat{\theta} \quad \pm \quad SD_{\widehat{\theta}\frac{\alpha}{2}} SE_{\widehat{\theta}}$$

where

- $\widehat{\theta}$ is an estimator of the parameter,
- $\text{SD}_{\widehat{\theta}}$ is the sampling distribution of the estimator, and
- $\text{SE}_{\widehat{\theta}}$ is the standard error of the estimator.

That is, if we were to sample the population say, a large but finite number of times, $(1-\alpha)100\%$ of the intervals generated from the samples will contain the true population parameter.

There is a duality between confidence intervals and hypotheses tests.
Comparing the acceptance region of the test to the confidence interval, we see that μ_0 lies in the confidence interval if and only if the hypothesis tests does not reject. In other words, the confidence interval consists precisely of all those values of μ_0 for which the null hypothesis $H_0 : \mu = \mu_0$ is not rejected.

To test H_0, we use a random sample of n observation vectors $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n$ from $\text{MVN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with $\boldsymbol{\Sigma}$ known, and calculate $\bar{\mathbf{y}} = \frac{\sum_{i=1}^{n} \mathbf{y}_1}{n}$. The test statistic is

$$Z^2 = n(\bar{\mathbf{y}} - \boldsymbol{\mu}_0)' \boldsymbol{\Sigma}^{-1} (\bar{\mathbf{y}} - \boldsymbol{\mu}_0)$$

Under the null hypothesis $Z^2 \sim \chi^2_p$

Inference
Hypothesis tests

Do note that there is a general procedure for obtaining a test statistic

- Likelihood ratio tests (in any situation, multivariate or univariate)
- Union-intersection tests (designed for multivariate problems but we won't consider it any further here)

You have (implicitly at least) met Likelihood Ratio Tests. But rather conveniently, these both lead to the same test for mean vectors!!!

Likelihood Ratio Tests for the mean vector

The multivariate normal likelihood, with parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ is given by:

$$\mathscr{L}(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{np/2}|\hat{\boldsymbol{\Sigma}}|^{n/2}} e^{-\frac{1}{2}\sum_{i=1}^{n}(\mathbf{y}_i-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{y}_i-\boldsymbol{\mu})} \qquad (1)$$

which is maximised taking $\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \bar{\mathbf{y}})'$ and $\hat{\boldsymbol{\mu}} = \bar{\mathbf{y}}$.

A little algebra shows that $\sum_{i=1}^{n}(\mathbf{y}_i - \hat{\boldsymbol{\mu}})'\hat{\boldsymbol{\Sigma}}^{-1}(\mathbf{y}_i - \hat{\boldsymbol{\mu}}) = np$ so we can express the maximum as:

$$max\mathscr{L}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{np/2}|\hat{\boldsymbol{\Sigma}}|^{n/2}} e^{-np/2} \tag{2}$$

Under the hypothesis $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$

$$\mathscr{L}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{np/2}|\boldsymbol{\Sigma}|^{n/2}} e^{-\frac{1}{2}\sum_{i=1}^{n}(\mathbf{y}_i - \boldsymbol{\mu}_0)\boldsymbol{\Sigma}^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_0)'} \qquad (3)$$

so $\boldsymbol{\mu}_0$ is now fixed but $\boldsymbol{\Sigma}$ can be varied to find the most likely value. As before, this can be rearranged to give:

$$max\mathscr{L}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{np/2}|\hat{\boldsymbol{\Sigma}}_0|^{n/2}} e^{-np/2} \tag{4}$$

where $\hat{\boldsymbol{\Sigma}}_0 = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{y}_i - \boldsymbol{\mu}_0)(\mathbf{y}_i - \boldsymbol{\mu}_0)'$

To determine whether $\boldsymbol{\mu}_0$ is plausible, we wish to compare $\mathscr{L}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma})$ with $\mathscr{L}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, conventionally performed by means of the likelihood ratio statistic:

$$\Lambda = \frac{max\mathscr{L}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma})}{max\mathscr{L}(\boldsymbol{\mu}, \boldsymbol{\Sigma})} = \left(\frac{|\hat{\boldsymbol{\Sigma}}|}{|\hat{\boldsymbol{\Sigma}}_0|} \right)^{n/2} \tag{5}$$

Wilk's Lambda

In general, a likelihood ratio test of $H_0 : \mu = \mu_0$ rejects H_0 if:

$$\Lambda = \frac{max\mathscr{L}(\mu_0)}{max\mathscr{L}(\mu)} < c \tag{6}$$

for a suitable constant c.

- For each test we need to find a sampling distribution for Λ,
- For large samples, $-2\log\Lambda$ is approximated by χ^2 with degrees of freedom equivalent to the difference in dimension of the two parameter spaces
- It turns out that a little algebra turns the likelihood ratio into something much simpler when considering mean vectors

t-test

Let's take a moment to revise the (Univariate) t-test
Consider testing the univariate hypothesis $H_0 : \mu = \mu_0$:

$$t = \sqrt{n}\left(\frac{\bar{y} - \mu}{s}\right)$$

and recall that

$$\Delta(\bar{y}, \mu) = \frac{|\bar{y} - \mu|}{s}$$

So it should come as no surprise that a multivariate test on mean vectors will be based on the distance measures!

Remember Multivariate distance?

Two vectors $\mathbf{y}_1$ and $\mathbf{y}_2$, with a common covariance matrix $\mathbf{\Sigma}$ the multivariate standard distance is given by:

$$\Delta(\mathbf{y}_1, \mathbf{y}_2) = \sqrt{(\mathbf{y}_1 - \mathbf{y}_2)' \mathbf{\Sigma}^{-1} (\mathbf{y}_1 - \mathbf{y}_2)}$$

Depending on whichever textbook is consulted, this multivariate standard distance may be referred to as the *statistical distance*, the *elliptical distance* or the *Mahalanobis distance*.

Mahalanobis distance

Originally proposed by Mahalanobis:1930 as a measure of distance between two populations:

$$\Delta(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2) = \sqrt{(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)}$$

Obvious sample analogue:

$$\Delta(\bar{\mathbf{y}}_1, \bar{\mathbf{y}}_2) = \sqrt{(\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)' \mathbf{S}^{-1} (\bar{\mathbf{y}}_1 - \bar{\mathbf{y}}_2)}$$

where $\mathbf{S}$ is the pooled estimate of $\mathbf{\Sigma}$ given by $\mathbf{S} = [(n_1 - 1)\mathbf{S}_1 + (n_2 - 1)\mathbf{S}_2] / (n_1 + n_2 - 2)$.

Consider the distance between $\mathbf{x}$, a vector of random variables with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ and its mean:

$$\Delta(\mathbf{y}, \boldsymbol{\mu}) = \sqrt{(\mathbf{y} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \boldsymbol{\mu})}$$

or the sample analogue (estimating $\boldsymbol{\mu}$ by $\hat{\mathbf{y}}$ and $\boldsymbol{\Sigma}$ by $\mathbf{S} = \frac{1}{n-1}\mathbf{Y}'\mathbf{Y}$).

One sample T^2 tests

Hotelling's T^2 test for a single sample

Consider testing $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$, where $\boldsymbol{\mu}_0$ is a predetermined mean vector (such as $\mathbf{0}$). In this case, Hotelling's T^2 is given by:

$$T^2 = n\left(\bar{\mathbf{y}} - \boldsymbol{\mu}_0\right)' \mathbf{S}^{-1} \left(\bar{\mathbf{y}} - \boldsymbol{\mu}_0\right) \tag{7}$$

where n is the sample size, $\boldsymbol{\mu}_0$ is the hypothesized mean, $\bar{\mathbf{y}}$ and $\mathbf{S}$ are the sample mean and covariance matrices respectively.

The T^2 distribution

- The statistic given above follows a T^2 distribution,
- However, there is a simple relationship between the T^2 and F distribution

Theorem

If $\mathbf{y}_i$, $i = 1, \ldots n$ *represent a sample from a* p *variate normal distribution with mean* $\boldsymbol{\mu}$ *and covariance* $\boldsymbol{\Sigma}$, *provided* $\boldsymbol{\Sigma}$ *is positive definite and* $n > p$, *given sample estimators for mean and covariance* $\bar{\mathbf{y}}$ *and* $\mathbf{S}$ *respectively, then:*

$$F = \left(\frac{n}{n-1}\right)\left(\frac{n-p}{p}\right)(\bar{\mathbf{y}} - \boldsymbol{\mu}_0)' \mathbf{S}^{-1} (\bar{\mathbf{y}} - \boldsymbol{\mu}_0) \tag{8}$$

follows an F*-distribution with* p *and* $(n-p)$ *degrees of freedom.*

Footnotes to Hotelling's T^2 test

- Note the requirement that $n > p$, i.e. that $\mathbf{S}$ is non-singular which clearly limits the use of this test in bio-informatic applications
- To carry out a test on $\boldsymbol{\mu}$, we determine whether $F \leq F_{(1-\alpha),p,n-p}$, the $(1-\alpha)$ quantile of the F distribution on p and $n-p$ degrees of freedom and reject the null hypothesis if our test statistic exceeds this value.

Worked example

Consider the data matrix $\mathbf{Y} = \begin{pmatrix} 6 & 9 \\ 10 & 6 \\ 8 & 3 \end{pmatrix}$. Evaluate

$H_0 : \boldsymbol{\mu}_0' = (9, 5)$

$$\begin{aligned} \bar{\mathbf{y}} &= \begin{pmatrix} \bar{y}_1 \\ \bar{y}_2 \end{pmatrix} \\ &= \begin{pmatrix} \frac{6+10+8}{3} \\ \frac{9+6+3}{3} \end{pmatrix} \\ &= \begin{pmatrix} 8 \\ 6 \end{pmatrix} \end{aligned}$$

Find $\mathbf{S}\begin{pmatrix} 4 & -3 \\ -3 & 9 \end{pmatrix}$

$s_{11} = \frac{(6-8)^2+(10-8)^2+(8-8)^2}{2} = 4,$

$s_{12} = \frac{(6-8)(9-6)+(10-8)(6-6)+(8-8)(3-6)}{2} = -3$ and

$s_{22} = \frac{(9-6)^2+(6-6)^2+(3-6)^2}{2} = 9.$

Find $\mathbf{S}^{-1}$:

$$\begin{aligned} \mathbf{S}^{-1} &= \frac{1}{4 \times 9 - (-3) \times (-3)} \begin{pmatrix} 4 & -3 \\ -3 & 9 \end{pmatrix} \\ &= \begin{pmatrix} 1/3 & 1/9 \\ 1/9 & 4/27 \end{pmatrix} \end{aligned}$$

$$\begin{aligned} T^2 &= 3(8-9, 6-5)\begin{pmatrix} 1/3 & 1/9 \\ 1/9 & 4/27 \end{pmatrix}\begin{pmatrix} 8-9 \\ 6-5 \end{pmatrix} \\ &= 3(-1,1)\begin{pmatrix} -2/9 \\ 1/27 \end{pmatrix} \\ &= \frac{7}{9}. \end{aligned}$$

We could use table A.7 in the text and compare this to a $T^2_{3,2}$ Or we could find the corresponding value of the F distribution:

$$\begin{aligned} \frac{1}{(n-1)}\frac{(n-p)}{n}T^2 &= \left(\frac{1}{2}\right)\left(\frac{1}{3}\right)\left(\frac{7}{9}\right) \\ &= \frac{7}{54} \end{aligned}$$

Do you reject the Null Hypothesis? No

We have to consider an F distribution on $\nu_1 = p = 2$ and $\nu_2 = n - p = 1$ degrees of freedom. You should find that the critical value at $\alpha = 0.1$ is 49.5 (see table the appendix in Rencher) or that at $\alpha = 0.05$ is 199.5. Clearly the critical value is much larger than the observed value of the test statistic, and we have no evidence to reject the null hypothesis. This doesn't mean we can say that the mean vector is definitely $\boldsymbol{\mu}_0' = (9, 5)$, merely that we have insufficient evidence to rule out that possibility (we have a very small sample size here!)