Multivariate Statistical Analysis

Fall 2011

C. L. Williams, Ph.D.

Lecture 5 for Applied Multivariate Analysis

Outline

- Multivariate distance

- Mahalanobis Distance
  - Introduce Mahalanobis distance and some ideas about multivariate normality
  - Cover a range of measures for multivariate distance: we will use these in cluster analysis and scaling (towards the end of term)
- Similarity / dissimilarity measures
  - You are very familiar with correlation and covariance
  - What about similarities and dissimilarities between individuals (rows, units) rather than variables?

Here's a brief word about the multivariate normal distribution. The mean and covariance can be defined in a similar way to to the univariate context. And the multivariate normal distribution has a pdf that shouldn't seem too strange by now:

$$f\left(y_1, y_2, ...y_p\right) = \left(\frac{1}{2\pi}\right)^{\frac{p}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}} exp\left(-\frac{1}{2}\left(\mathbf{y}-\boldsymbol{\mu}\right)' \boldsymbol{\Sigma}^{-1}\left(\mathbf{y}-\boldsymbol{\mu}\right)\right)$$

Given that we have $E(\mathbf{y}) = \boldsymbol{\mu}$, and that $Var(\mathbf{y}) = \boldsymbol{\Sigma}$, hence we can use the notation:

$$\mathbf{y} \sim MVN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

- One method for assessing multivariate normality, quantiles of the Mahalanobis distance of $\mathbf{y}_i$, $i = 1, \ldots, n$ with respect to $\boldsymbol{\mu}$ can be plotted against quantiles of the χ^2_p distribution as an assessment of multivariate normality.
- We can also define contours as a set of points of equal probability in terms of equal Mahalanobis distance:

$$(\mathbf{y}_i - \hat{\boldsymbol{\mu}})'\hat{\boldsymbol{\Sigma}}^{-1}(\mathbf{y}_i - \hat{\boldsymbol{\mu}}) = \mathbf{z}'\mathbf{z} = c^2 \qquad (1)$$

for any constant $c > 0$.

$$d(y_1, y_2) = \frac{|y_1 - y_2|}{\sigma}$$

This is the *absolute distance* between two observations in units of standard deviation. †

- Invariant under non-degenerate linear transformations,
- e.g. consider $Z = \alpha Y + \beta$, where $\alpha \neq 0$ and β are fixed constants.
- We can transform y_1 and y_2 to $z_i = \alpha y_i + \beta, i = 1, 2$, resulting standard distance:

$$\begin{aligned}
\Delta(z_1, z_2) &= \frac{|z_1 - z_2|}{\sqrt{var(Z)}} \\
&= \frac{|\alpha(y_1 - y_2|)}{\sqrt{\alpha^2\sigma^2}} \\
&= \Delta(y_1, y_2)
\end{aligned}$$

†Note if $\sigma = 1$ then this is the Euclidean distance

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

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

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*.
Originally proposed by Mahalanobis (1936) as a measure of distance between two populations:

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

Sample analogue

This has an 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 $\boldsymbol{\Sigma}$ given by $\mathbf{S} = \left[(n_1 - 1)\mathbf{S}_1 + (n_2 - 1)\mathbf{S}_2\right]/(n_1 + n_2 - 2)$.

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

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

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

In **R**, the `mahalanobis()` function is intended to return the *squared* multivariate distance between a matrix $\mathbf{Y}$ and a mean vector $\boldsymbol{\mu}$, given a user-supplied covariance matrix $\boldsymbol{\Sigma}$, i.e. we wish to calculate:

$$d(\mathbf{y}_i, \hat{\boldsymbol{\mu}})^2 = (\mathbf{y}_i - \hat{\boldsymbol{\mu}})' \hat{\boldsymbol{\Sigma}}^{-1} (\mathbf{y}_i - \hat{\boldsymbol{\mu}})$$

Distributional properties of the Mahalanobis distance

- When $z_1, \ldots, z_p \sim N(0,1)$, if we form $y = \sum_{j=1}^{p} z_j^2$ then $y \sim \chi_p^2$
- $\therefore$ with multivariate normal data, with p variables, the squared Mahalanobis distance can be compared against a χ_p^2 distribution:

$$(\mathbf{y}_i - \hat{\boldsymbol{\mu}})'\hat{\boldsymbol{\Sigma}}^{-1}(\mathbf{y}_i - \hat{\boldsymbol{\mu}}) = \mathbf{z}'\mathbf{z} \sim \chi_p^2 \qquad (2)$$