Multivariate Statistical Analysis

Fall 2011

C. L. Williams, Ph.D.

Lecture 3 for Applied Multivariate Analysis

Outline

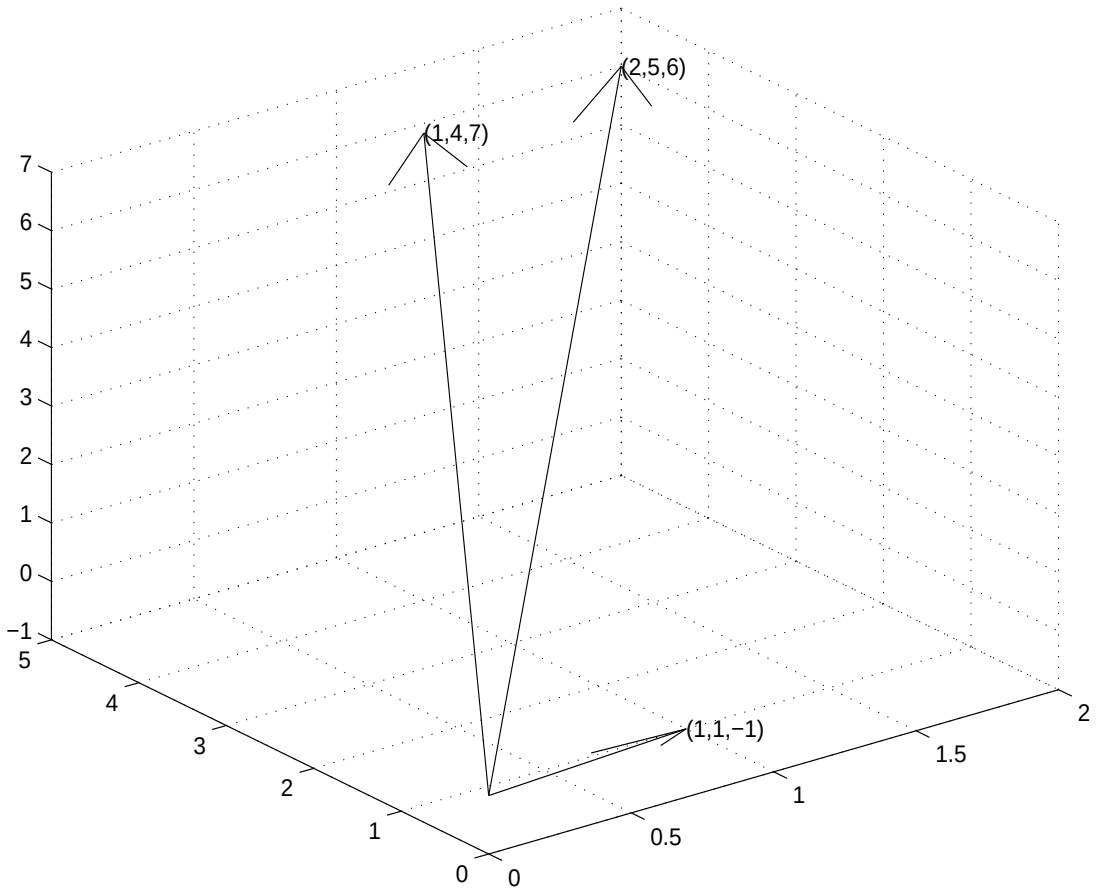
Recall

$$r_{xy} = \frac{(\mathbf{x} - \overline{x}\mathbf{j})'(\mathbf{y} - \overline{y}\mathbf{j})}{\sqrt{[(\mathbf{x} - \overline{x}\mathbf{j})'(\mathbf{x} - \overline{x}\mathbf{j})][(\mathbf{y} - \overline{y}\mathbf{j})'(\mathbf{y} - \overline{y}\mathbf{j})]}}$$

Thus if the angle θ between the two centered vectors centered as $\mathbf{x}$ - $\overline{x}\mathbf{j}$ and $\mathbf{y}$ - $\overline{y}\mathbf{j}$ is small so that cos θ is near 1, r_{xy} will be close to 1. If the two vectors are perpendicular, cos θ and r_{xy} will be zero. If the two vectors have nearly opposite directions, r_{xy} will be close to -1.

But if you now attempt to plot $\vec{x}$ and $\vec{y}$, i.e. the three dimensional points $(1, 4, 7)$ and $(2, 5, 6)$ you could attempt to measure the angle between these vectors. The cosine of this angle is the correlation.

Figure: Test vector 2-points in 3-dimensions



Unfortunately, it's a little harder to even imagine this vector for a conventional data set (with tens if not hundreds of points), but that's what you've been measuring whenever you work out the correlation coefficient. And if you think about the correlation paradox, you will appreciate that by having two modestly correlated variables (i.e. with angles in the order of 50° or more degrees), when you measure the angle between the two outermost variables it will be greater than 90° and the cosine will be negative.

$$
\begin{aligned}
cos\theta &= \frac{\mathbf{a}'\mathbf{a} + \mathbf{b}'\mathbf{b} - (\mathbf{b}-\mathbf{a})'(\mathbf{b}-\mathbf{a})}{2\sqrt{(\mathbf{a}'\mathbf{a})(\mathbf{b}'\mathbf{b})}} \\
&= \frac{\mathbf{a}'\mathbf{a} + \mathbf{b}'\mathbf{b} - (\mathbf{b}'\mathbf{b} + \mathbf{a}'\mathbf{a} - 2\mathbf{a}'\mathbf{b})}{2\sqrt{(\mathbf{a}'\mathbf{a})(\mathbf{b}'\mathbf{b})}} \\
&= \frac{\mathbf{a}'\mathbf{b}}{\sqrt{(\mathbf{a}'\mathbf{a})(\mathbf{b}'\mathbf{b})}}
\end{aligned}
$$

So for vectors say

$$
\mathbf{a} = \begin{bmatrix} 1 \\ 4 \\ 7 \end{bmatrix} \text{ and}
$$

$$
\mathbf{b} = \begin{bmatrix} 2 \\ 5 \\ 6 \end{bmatrix}
$$

...we have length of
$\mathbf{a}(L_{\mathbf{a}}) = \sqrt{\mathbf{a}'\mathbf{a}} = \sqrt{1^2 + 4^2 + 7^2} = \sqrt{66} = 8.12404$ and the length of $\mathbf{b}(L_{\mathbf{b}}) = \sqrt{\mathbf{b}'\mathbf{b}} = \sqrt{2^2 + 5^2 + 6^2} = \sqrt{65} = 8.06226$ and $(\mathbf{a}'\mathbf{b}) = \sqrt{1 * 2 + 4 * 5 + 7 * 6} = 64$ so that

$$\begin{aligned} cos\theta &= \frac{\mathbf{a}'\mathbf{b}}{\sqrt{(\mathbf{a}'\mathbf{a})(\mathbf{b}'\mathbf{b})}} \\ &= \frac{64}{\sqrt{66} \times \sqrt{65}} \\ &= \frac{64}{8.12404 \times 8.06226} = 0.977120. \end{aligned}$$

so we can determine the arccosine to determine $\theta = 0.2149$ *radians* which is 12.3129°.

Linear Dependence

A pair of vectors **a** and **b** of the same dimension is linearly dependent if there exists constants c_1 and c_2 both not zero such that

$$c_1\mathbf{a} + c_2\mathbf{b} \quad = \quad 0$$

A set of k vectors **a**, **b**, . . . , **w** of the same dimension are linearly dependent if there exists constants $c_1, c_2 \ldots, c_k$ not all zero such that

$$c_1\mathbf{a} \; + c_2\mathbf{b} \; + \; \cdots \; + \; c_k\mathbf{w} \quad = \quad 0$$

Linear dependence implies that at least one of the vectors in the set can be written as a linear combination of the other vectors. Obviously, vectors of the same dimension that are not linearly dependent are linearly independent.

Example of linearly independent vectors

Suppose we have the following three vectors

$$\mathbf{y}_1 = \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix}$$

$$\mathbf{y}_2 = \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix}$$

$$\mathbf{y}_3 = \begin{bmatrix} 1 \\ -2 \\ 1 \end{bmatrix}$$

So

$$c_1\mathbf{y}_1 + c_2\mathbf{y}_2 + c_3\mathbf{y}_3 = 0$$

so that

$$\begin{array}{rcrcrcl} c_1 & + & c_2 & + & c_3 & = & 0 \\ 2c_1 & & & + & 2c_3 & = & 0 \\ c_1 & - & c_2 & + & c_3 & = & 0 \end{array}$$

with the unique solution $c_1 = c_2 = c_3 = 0$. So the vectors are linearly independent.

```
   Y1932 Y1936 Y1940 Y1960 Y1964 Y1968
Missouri    35   38   48   50   36   45
Maryland    36   37   41   46   35 42
Kentucky    40   40   42   54   36   44
Louisiana   7    11   14   29   57 23
Mississippi 4    3    4    25   87   14
"South Carolina"  2    1    449   59 39

>votes.data<-read.table("...\\votes.dat",header=T)
>library(aplpack)
>faces(votes.data)
```

You should be very comfortable dealing with bivariate concepts. Now let's introduce some terminology that we will be using throughout in a fully *multivariate* setting. Firstly, note that conventionally, we have p variables $y_1, y_2, \ldots, y_p$ observed on n individuals

- p variable means can be collected into the *mean vector* $\bar{\mathbf{y}}^T = (\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_p)$
- Variances of the p variables collected on the diagonal and the $\frac{1}{2}p(p-1)$ covariances between every pair of variables collected in off-diagonal of the *variance covariance* matrix $\mathbf{S}$.
- Correlations between every pair of variables can be put in the off-diagonal position of the *correlation* matrix $\mathbf{R}$ (diagonal elements are all 1)
- Mean centering or standardising is conducted by carrying out the appropriate operation on each variable in turn.
- Covariance matrix of the standardised data is the same of the

Variance-Covariance Structures

We no longer have bivariate correlation coefficients (or covariance), we now have a Covariance (or even more formally the variance covariance) matrix.

$$\mathbf{S} = \begin{pmatrix} s_1 & 0 & \cdots & 0 & \cdots & 0 \\ 0 & s_2 & \cdots & 0 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ 0 & s_{j2} & \cdots & s_j & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & 0 & \cdots & s_p \end{pmatrix}$$

$$\mathbf{S} = \begin{pmatrix} s_1 & s_{12} & \cdots & s_{1j} & \cdots & s_{1p} \\ s_{21} & s_2 & \cdots & s_{2j} & \cdots & s_{2p} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ s_{j1} & s_{j2} & \cdots & s_j & \cdots & s_{jp} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pj} & \cdots & s_p \end{pmatrix}$$

$$\boldsymbol{\Sigma} = E[(\mathbf{y} - \boldsymbol{\mu})'(\mathbf{y} - \boldsymbol{\mu})]$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1j} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2j} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ \sigma_{j1} & \sigma_{j2} & \cdots & \sigma_{jj} & \cdots & \sigma_{jp} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pj} & \cdots & \sigma_{pp} \end{pmatrix}$$

Outline

We also have more subtle ways of measuring the association between variables

- For quantitative data, the correlation matrix **R** shows all pairwise correlations between variables (may be too many to interpret)
- The *partial correlation* $r_{ij,k}$ is the correlation between x_i and x_j when x_k is held at a constant value. Any number of values can be held fixed:

$$r_{ij,k} = \frac{r_{ij} - r_{ik}r_{jk}}{\sqrt{(1 - r_{ik}^2)(1 - r_{jk}^2)}}$$

$$r_{ij,kl} = \frac{r_{ij,k} - r_{il,k}r_{jl,k}}{\sqrt{(1 - r_{il,k}^2)(1 - r_{jl,k}^2)}}, \text{ etc.}$$

- A partial correlation will show whether a high correlation between two variables is caused by their mutual correlation with one of more other variables

Correlation Matrices

Relationship to Covariance Matrices

$$r_{jk} = \frac{s_{jk}}{s_{jj}s_{kk}}$$

$$\mathbf{R} = (r_{jk}) = \begin{pmatrix} 1 & r_{12} & \cdots & r_{1j} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2j} & \cdots & r_{2p} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ r_{j1} & r_{j2} & \cdots & 1 & \cdots & r_{jp} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ r_{p1} & r_{p2} & \cdots & r_{pj} & \cdots & 1 \end{pmatrix}$$

$$\mathbf{R} = \mathbf{D}_s^{-1}\mathbf{S}\mathbf{D}_s^{-1}$$

Population correlation matrix

$$\boldsymbol{\rho}_{\rho} = (\rho_{jk}) = \begin{pmatrix} 1 & \rho_{12} & \cdots & \rho_{1j} & \cdots & \rho_{1p} \\ \rho_{21} & 1 & \cdots & \rho_{2j} & \cdots & \rho_{2p} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ \rho_{j1} & \rho_{j2} & \cdots & 1 & \cdots & \rho_{jp} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ \rho_{p1} & \rho_{p2} & \cdots & \rho_{pj} & \cdots & 1 \end{pmatrix}$$

where

$$\rho_{jk} = \frac{\sigma_{jk}}{\sigma_j \sigma_k}$$

"Multivariate statistics" covers a vast and rapidly expanding discipline; developments are taking place for the analysis of gene expression data as well as psychometrics to name but two. We therefore need to narrow down our definition to make life manageable.
For the purposes of this course, our definition of multivariate is as follows:

- We don't count multiple regression as a multivariate technique.
- Relationships between variables (principal components, factor analysis, canonical correlation)
- Relationships between individuals (cluster analysis, discriminant analysis, Hotelling's T^2 test, MANOVA, principal coordinates analysis.

There are some overlaps between the techniques. Also, to make life manageable, most of the techniques we will consider involve eigenanalysis (either of the covariance or the correlation matrix, or of the ratio of between and within "covariance" matrices).

Firstly, the techniques we shall cover looking at the differences between individuals are as follows:

- Differences between Groups
  - Can we tell if a vector of means $\bar{\mathbf{y}}_1$ is different from another vector $\bar{\mathbf{y}}_2$ (Hotelling's T^2)
  - What if we have more than two groups (MANOVA)?
- Classification
  - Can we find individuals that are more alike than other individuals (Cluster Analysis)
  - If we already knew the groupings, could we investigate which variables were most important in telling the groups apart. Could we use this information to find a rule that lets us classify new observations (Discriminant analysis)

- Visualisation
  - Do we have some way of visualising the similiarities and dissimilarities between individuals (Scaling / Principal Co-ordinates Analysis)

And the techniques which are about examining the variables are as follows:

- Dimension Reduction
  - Can we represent our data in less dimensions (Principal Components Analysis)
- Relationships between variables
  - Can we model the relationships between variables (Factor analysis)
  - If we have a set of variables $\mathbf{X}$, can we find a projection that is correlated to a projection of variables $\mathbf{Y}$ (Canonical correlation)

Linear combinations

$$z_{1i} = \alpha_{11} y_{1i} + \alpha_{12} y_{2i} + \ldots + \alpha_{1p} y_{pi}$$

$$z_{2i} = \alpha_{21} y_{1i} + \alpha_{22} y_{2i} + \ldots + \alpha_{2p} y_{pi}$$

$$\vdots = \vdots$$

$$z_{pi} = \alpha_{p1} y_{1i} + \alpha_{p2} y_{2i} + \ldots + \alpha_{pp} y_{pi}$$

Many techniques are concerned with finding out useful linear combinations for a given task!