Multivariate Statistical Analysis

Fall 2011

C. L. Williams, Ph.D.

Lecture 2 for Applied Multivariate Analysis

Outline

1. Reprise-Two dimension scatter diagram
2. Correlation and Covariance
3. Bivariate Correlation
4. Points and Dimensions
5. Multivariate random variables

Some features of a scatter diagram

- The centroid of the points occurs at $(\bar{x}_1, \bar{x}_2)$(age,length), where $\bar{x}_j$ is the mean (i.e. arithmetic average $\frac{1}{n}\sum_{i=1}^{n} x_{ij}$) of the values on variable x_j. Here, this is $(\bar{x}_1, \bar{x}_2) = (66.660, 9.198)$
- The spread of the points along an axis is measured either by the variance or the standard deviation of the corresponding variable. The variance of x_j is $\frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2$ and its standard deviation is the square root of the variance. For our data, the variance of x_1 is 199.086122 and its standard deviation is $\sqrt{199.086122} = 14.10979$, while the variance of x_2 is 1.6002 and its standard deviation is $\sqrt{1.6002} = 1.26499$.
- The association between the two variables is measured by their covariance $\frac{1}{n-1}\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2)$. Here, this is -2.187429.

Correlation and Covariance

Now, we measure the linear association between two variables with the covariance. If we standardise that in terms of the variance of the two variables we have the correlation.

Since the covariance depends on the scale of measurement of x and y, it is difficult to compare covariances between different pairs of variables. For example, if we change a measurement from inches to centimeters, the covariance will change. To find a measure of linear relationship that is invariant to changes of scale, we can standardize the covariance by dividing by the standard deviations of the two variables. This standardized covariance is called a correlation. The *population correlation* of two random variables x and y is

$$\begin{aligned} \rho_{xy} &= corr(x, y) = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \\ &= \frac{E[(x - \mu_x)(y - \mu_y)]}{\sqrt{E(x - \mu_x)^2}\sqrt{E(y - \mu_y)^2}} \end{aligned}$$

$\cdots$ and the sample correlation is

$$r_{xy} = corr(x, y) = \frac{s_{xy}}{s_x s_y}$$

$$= \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \overline{x})^2 \sum_{i=1}^{n} (y_i - \overline{y})^2}}$$

The sample correlation r_{xy} is related to the cosine of the angle between two vectors. Let θ be the angle between vectors **a** and **b**. The vector from the terminal point of **a** to the terminal point of **b** can be represented as **c** = **b** - **a**. Then the law of cosines can be stated in vector form as

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

Since $\cos\left(90^0\right) = 0$, we see from (3.14) that $\mathbf{a}\cdot\mathbf{b} = 0$ when $\theta = \left(90^0\right)$. Thus $\mathbf{a}$ and $\mathbf{b}$ are perpendicular when $\mathbf{a}\cdot\mathbf{b} = 0$. By (2.99), two vectors $\mathbf{a}$ and $\mathbf{b}$, such that $\mathbf{a}\cdot\mathbf{b} = 0$, are also said to be orthogonal. Hence orthogonal vectors are perpendicular in a geometric sense.

To express the correlation in the form given in (3.14), let the n observation vectors $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ in two dimensions be represented as two vectors $\mathbf{x}' = (x_1, x_2, \ldots, x_n)$ and $\mathbf{y}' = (y_1, y_2, \ldots, y_n)$ in n dimensions, and let x and y be centered as $\mathbf{x}$ - $\overline{x}\mathbf{j}$ and $\mathbf{y}$ - $\overline{y}\mathbf{j}$. Then the cosine of the angle θ between them [see (3.14)] is equal to the sample correlation between x and y:

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

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.

Bivariate Correlation

- `cor()` in R.
- You can also get Kendall's and Spearman's coefficients by adding `method = "kendall"` and `method = "spearman"` to the function call.

There's a nice demo in R if you want to revise your ideas of what correlation is all about:

```
library(TeachingDemos)
run.cor.examp()
```

On the other hand (A correlation paradox)

- The following paradox, based on an article in "The American Statistician" by Langford et al. 2001.
- If x and y are positively correlated, and x and z are positively correlated, what will be the sign of the correlation between y and z?
- For an interesting little demonstration of this paradox, create three random variables u, v and w according to any recipe you like (e.g. `u <- rnorm(100)`). Then form x, y and z from these variables as follows:

```
> x <- u + v
> y <- u + w
> z <- v - w
> X <- cbind(x,y,z)
```

Now look at `cor(X)` and `pairs(X)`

Correlation paradox:

What's going on here:

(a) The correlation between x, y and z are positive

(b) The correlation between x and y, x and z is positive, but the correlation between y and z is negative

(c) The correlation between x and y, y and z is positive, but the correlation between x and z is negative

(d) The correlation between x and z, y and z is positive, but the correlation between y and z is negative

Correlation paradox solved?

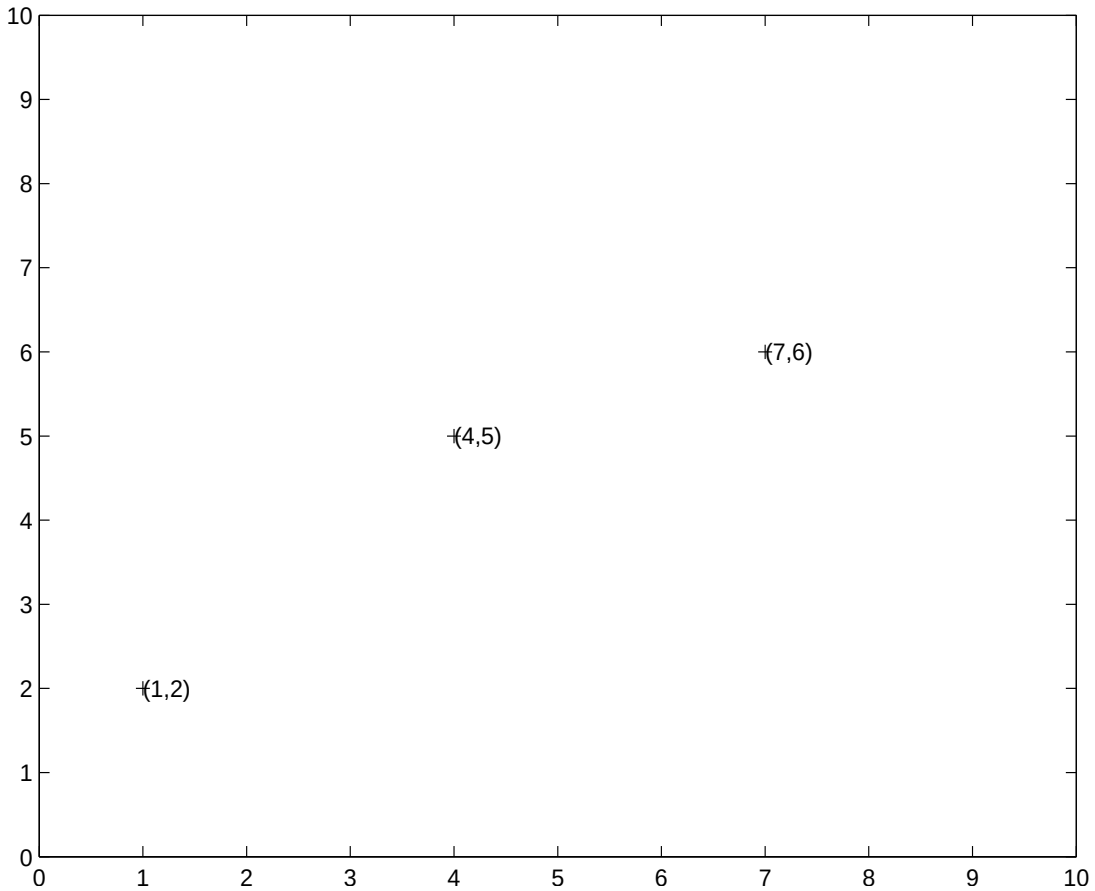
So what is going on here:

(b) The correlation between x and y, x and z is positive, but the correlation between y and z is negative.

Is this an oddity?

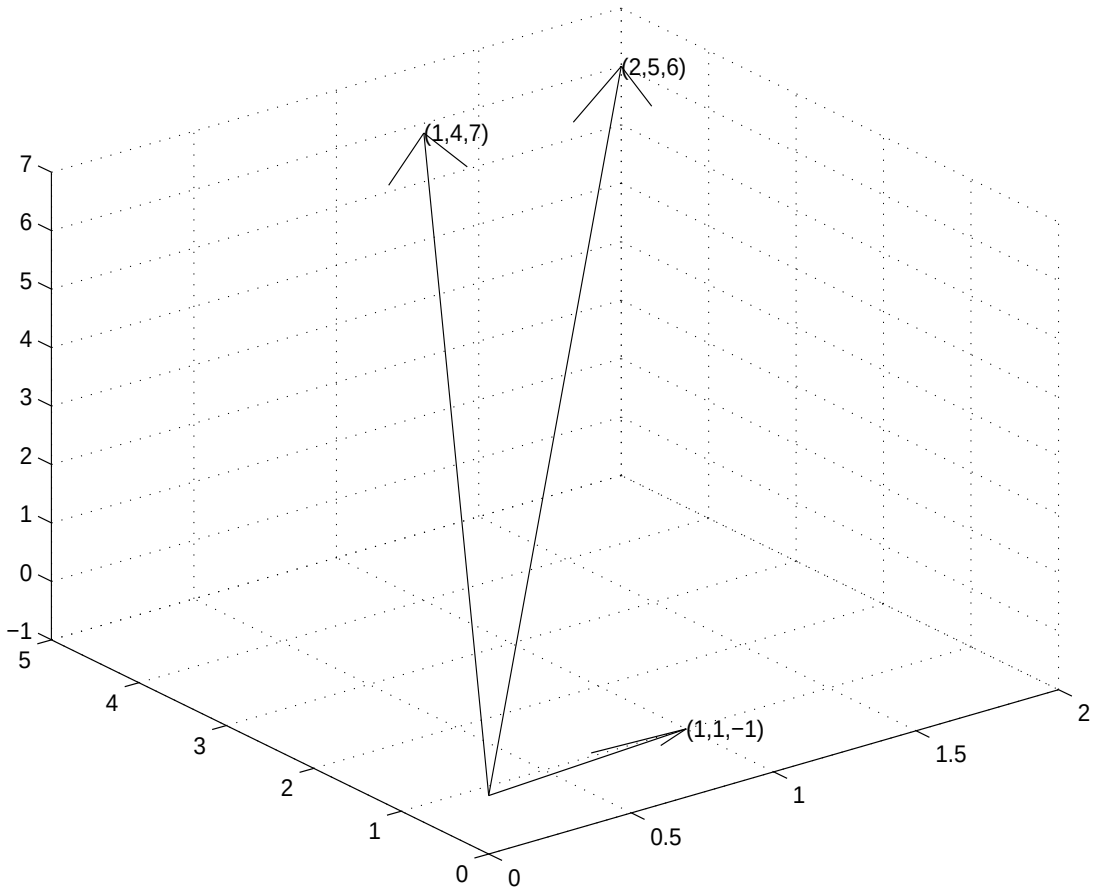
This might sound like an odd result, but it makes perfect geometric sense. Try the following exercise. Consider $x = (1, 4, 7)$ and $y = (2, 5, 6)$. If you were to plot a conventional scatterplot you would plot points at $(1, 2)$, $(4, 5)$, $(7, 6)$.

Figure: Test vector 3-points in 2-dimensions



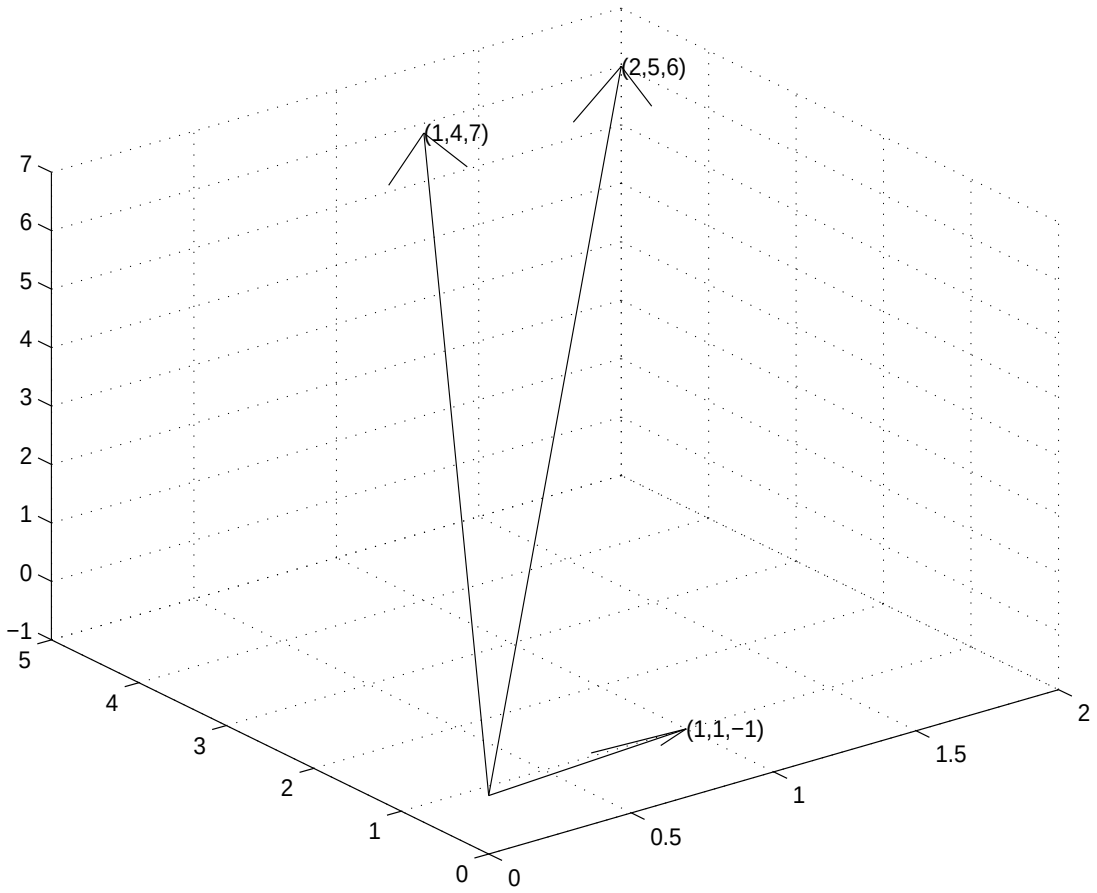
But if you now attempt to plot $\vec{x}$ and $\vec{y}$, i.e. the three dimensional points $(1, 4, 7)$ and $(2, 5, 6)$ you could attempt to measure the angle between these vectors. The cosine of this angle is the correlation.

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.

Figure: Test vector 2-points in 3-dimensions



Here's the neat part

In the univariate case we're looking at, specifically with two variables, n observations in p=2. dimensions. But the multivariate case would be synonymous to us letting the n observations be the dimensions and there being p=2 observations.

Points and Dimensions

A natural extension to a two-dimensional scatter diagram (scatterplot) where there are two variables in which a relationship is sought is the extension to p-dimensions. Let $\mathbf{y}$ represent a random vector of p variables measured on a sampling unit (subject or object). If there are n individuals in the sample, the n denoted by $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n$, where each of the $\mathbf{y}_i$'s is a vector of length p and $\mathbf{y}_{i1}, \mathbf{y}_{i2}, \ldots, \mathbf{y}_{ip}$ coordinate axes are taken to correspond to the variables so that the i^{th} point is y_{i1} units along the first axis, y_{i2} units along the second,... and y_{ip} units along the p^{th} axis. Generally, this resulting "scatterplot" will result in a distinct pattern of variability, but should also reflect any similarities and well as dissimilarities among the n observations. Clustering may also be observed.

Multivariate random variables

Uppercase boldface letters are used for matrices of random variables or constants, lowercase boldface letters represent vectors of random variables or constants, and univariate random variables or constants are usually represented by lowercase nonbolded letters.

Random vectors

The i^{th} observation

$$\mathbf{y}_i = \begin{pmatrix} y_{i1} \\ y_{i2} \\ y_{i3} \\ \vdots \\ y_{ip} \end{pmatrix}$$

Mean Vectors

$$\overline{\mathbf{y}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{y}_i = \begin{pmatrix} \overline{y}_1 \\ \overline{y}_2 \\ \overline{y}_3 \\ \vdots \\ \overline{y}_p \end{pmatrix}$$

Representing Data

The Data Matrix

$$\mathbf{Y} = \begin{pmatrix} \mathbf{y}_1' \\ \mathbf{y}_2' \\ \mathbf{y}_3' \\ \vdots \\ \mathbf{y}_n' \end{pmatrix}$$

$$= \begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1j} & \cdots & y_{1p} \\ y_{21} & y_{22} & \cdots & y_{2j} & \cdots & y_{2p} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ y_{i1} & y_{i2} & \cdots & y_{ij} & \cdots & y_{ip} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{nj} & \cdots & y_{np} \end{pmatrix}$$

The second element of the j'Y

$$(1, 1, 1 \ldots, 1) \begin{pmatrix} y_{12} \\ y_{22} \\ y_{32} \\ \vdots \\ y_{n2} \end{pmatrix} = \sum_{i=1}^{n} y_{i2}$$

$$\overline{\mathbf{y}}' = \frac{1}{n}\mathbf{j}'\mathbf{Y}$$

$$\overline{\mathbf{y}} = \frac{1}{n}\mathbf{Y}'\mathbf{j}$$

Expectation of each **y**

$$E(\mathbf{y}) = E\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_p \end{pmatrix} = E\begin{pmatrix} E(y_1) \\ E(y_2) \\ E(y_3) \\ \vdots \\ E(y_p) \end{pmatrix} = E\begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \vdots \\ \mu_p \end{pmatrix} = \boldsymbol{\mu}$$

Expectation of each $\overline{y}$

$$E(\overline{\mathbf{y}}) = E\begin{pmatrix} \overline{y}_1 \\ \overline{y}_2 \\ \overline{y}_3 \\ \vdots \\ \overline{y}_p \end{pmatrix} = E\begin{pmatrix} E(\overline{y}_1) \\ E(\overline{y}_2) \\ E(\overline{y}_3) \\ \vdots \\ E(\overline{y}_p) \end{pmatrix} = E\begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \vdots \\ \mu_p \end{pmatrix} = \boldsymbol{\mu}$$

```
   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)
```