Appendix C

Covariance and correlation

The ideas of correlation and covariance are important in the use of several methods covered elsewhere in these notes (e.g., regression, principal component analysis and linear discriminant analysis). Some of the ideas involved are summarized here to avoid repetition in the relevant chapters. Any good intermediate/advanced text that deals with the topics involved will provide a more thorough treatment of the mathematics involved.

Measures of covariance and correlation are designed to provide information about the strength of a *linear* relationship between two variables. More than one way of defining such measures exist; apart from passing mention only the 'standard' definitions are used here. This is covered in Section C.1; Section C.2 refers back to earlier chapters with additional detail added in some cases.

C.1 Definitions and notation

Suppose we have n observations $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ and $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ on two variables, and wish to measure the strength of the linear relationship between them. For the purposes of this section define

$$X_i = x_i - \bar{x} \qquad Y_i = y_i - \bar{y}$$

where $\bar{x}$ and $\bar{y}$ are the means of the variables, so X_i and Y_i are centered on zero.

The estimated *covariance* between the variables is *defined* as

$$s_{xy} = \frac{\sum_i^n X_i Y_i}{n-1}$$

and is a measure of the linear relationship between the variables. That it is a sensible measure of linearity can be seen from Figure C.1

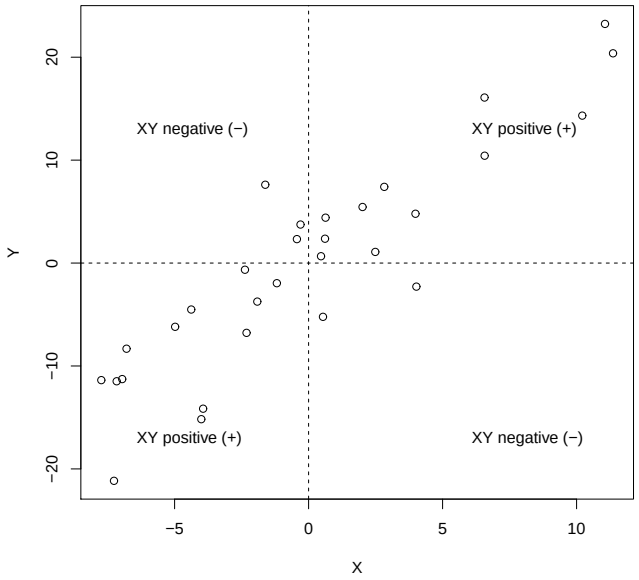

Figure C.1: *Artificial data showing a positive correlation for two variables. The population correlation coefficient from which the data are sampled is* $\rho_{xy} = 0.90$.

The plot is based on artificial data constructed to show a strong positive relationship. In the definition of s_{xy} the numerator is just the sum of terms of the form $X_i Y_i$. It can be seen from the figure that for a strong positive linear relationship these terms lie mostly in the upper-right and lower-left quadrants and are positive (remembering that the product of two negative numbers is positive) so that their sum, and hence s_{xy}, will be positive. If there is a negative relationship points will lie mostly in the upper-left and lower right quadrants, so if X_i is positive Y_i will tend to be negative (and *vice versa*) and their product will be negative. This implies that s_{xy} will be negative; if the plot is random points will be scattered around the quadrants and their effects will tend to cancel, so s_{xy} should be close to zero.

It should be obvious that the value of the numerator will be dependent on the sample size, n, and the divisor of $(n-1)$ in the definition removes this effect[1]. The other feature of a covariance, for descriptive purposes, is that the numerical result depends on the scale of the data. This means that you can't tell, just by looking

[1]The unbiased sample estimate of the population covariance is defined here, hence the divisor of $(n-1)$. Some treatments use n as the divisor which gives the definition of the population covariance. The distinction is not of great importance here except to note that the numerator is being adjusted for sample size.

at the value, whether the covariance is 'large' or 'small', and a comparison of the strength of relationship between two sets of data may not be easy. To get round this problem it is necessary to scale the covariance so it does not depend on the units of measurement.

If the covariance of a variable with itself is measured, write $s_{xx} = s_x^2$, by convention, then this is the estimated variance of x, with s_y^2 similarly defined. This gives s_x and s_y as the estimated standard deviations. The *correlation* between x and y is then defined as[2]

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

with $-1 \leq r_{xy} \leq 1$, and the correlation of a variable with itself is 1. The correlation coefficient is an estimate of the population correlation ρ_{xy}.

It is not necessary to assume that the population from which the data are drawn has a bivariate normal distribution for r (dropping the subscripts) to be useful. The assumption is necessary if formal tests of the null hypothesis $H_o : \rho = 0$ are used, but these are often not useful. If n is large then small correlations of little interest will be statistically significant. For small samples formal tests can guard against reading too much into apparently 'large' observed correlations, but usually graphical inspection is more than adequate to identify any problems.

The chapters on PCA through to that on LDA involve the analysis of multivariate data, where $p > 2$ variables are involved. All possible pairwise covariances can be calculated and summarized in the form of a $p \times p$ *covariance matrix*, $\mathbf{S}$, given by

$$\mathbf{S} = \begin{bmatrix} s_{11} & s_{12} & \dots & s_{1(p-1)} & s_{1p} \\ s_{21} & s_{22} & \dots & s_{2(p-1)} & s_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ s_{(p-1)1} & s_{(p-1)2} & \dots & s_{(p-1)(p-1)} & s_{(p-1)p} \\ s_{p1} & s_{p2} & \dots & s_{p(p-1)} & s_{pp} \end{bmatrix}$$

which is symmetric since $s_{ij} = s_{ji}$. The *correlation matrix*, $\mathbf{R}$, is similarly defined to be the $p \times p$ matrix with typical element r_{jk}. The diagonal elements of $\mathbf{R}$ are all equal to 1; otherwise the r_{ij} lie between -1 and +1.

The total variance in a data set can be defined as the sum of the individual variances $s_1^2 + s_2^2 + \ldots + s_p^2$ which is just the sum of the diagonal elements of $\mathbf{S}$. We can write this as $tr(\mathbf{S})$, where $tr(.)$, the *trace operator*, is just the sum

[2]'Correlation' as defined here is shorthand for the Pearson product-moment correlation coefficient, to give it its full name. This is sometimes needed to distinguish the coefficient from other definitions of correlation such as the non-parametric Spearman's or Kendall's rank order correlations. Usually, though, the shorthand suffices. It may be noted that if the data are ranked and the definition of r_{xy} applied to the ranks Spearman's rank-correlation coefficient is obtained.

of the diagonal elements of a square matrix. From this definition it follows that $tr(\mathbf{R}) = p$, since each of the p diagonal elements of $\mathbf{R}$ is equal to 1.

Suppose $\mathbf{X}$ is an $n \times 2$ data matrix, so that its correlation matrix, $\mathbf{R}$, is 2×2. The data set is said to be two-dimensional. If, however, X_1 and X_2 are perfectly correlated then the true dimensionality is really 1 since, given X_1, we know what X_2 is. If X_1 and X_2 are highly correlated then, in a sense, the data are 'approximately' one-dimensional. More generally, if $\mathbf{X}$ is a $p \times p$ matrix, but the variables are highly correlated, then the true dimensionality of the data will be somewhat less than p and it can be expected that low-dimensional representations of the data (which is what PCA and correspondence analysis attempt) will be quite successful.

C.2 Applications

C.2.1 Linear regression analysis

Linear regression analysis is covered in Chapter 5. The simplest practical linear regression model is

$$y = \alpha + \beta x + \varepsilon$$

where α and β are unknown parameters and ε is the error term. This is equation (5.1); interest commonly centers on obtaining $\hat{\beta}$, an estimate of β. The model can be given wider application by allowing simple transformations, such as logarithmic, of y and x as in equations (5.6) and (5.7).

Chapter 5 eschewed mathematical details in favor of proceeding by example. Introductory quantitative methods for archaeology texts intended for teaching purposes usually deal with simple linear regression (e.g., Shennan, 1997; Drennan, 2009). The texts cited do not derive the formula for $\hat{\beta}$ (which can, however, be obtained using basic calculus) but do give formulae, including computationally efficient versions. Shennan (1997: 137) gives

$$\hat{\beta} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

which in our notation is

$$\hat{\beta} = \frac{\sum X_i Y_i}{\sum X_i^2} = \frac{s_{xy}}{s_x^2}$$

and, if nothing else, is a neater way of expressing the result and makes explicit the role that covariance plays.

The correlation coefficient has been presented in various ways, among them

$$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2(y_i - \bar{y})^2}} = \frac{n\sum x_i y_i - (\sum x_i)(\sum y_i)}{\sqrt{[n\sum x_i^2 - (\sum x_i)^2][n\sum y_i^2 - (\sum y_i)^2]}} = \frac{s_{ij}}{s_i s_j}.$$

Following the first'=' the two expressions are the formula for r_{xy} usually presented and a computationally efficient equivalent version if you must do calculations 'by hand' (e.g., Shennan, 1997: 140). The final version, apart from being more compact, makes it clear that the correlation is the covariance rescaled to allow for the different 'spread' of the variables.

Software output for regression analyses typically report an R^2 value – the *coefficient of determination*. For the special case of simple linear regression R^2 is just the square of the correlation coefficient, r_{xy}^2, and can be interpreted as the amount of variation in y 'explained' by variation in x.

C.2.2 Principal component analysis

Chapter 7 dealt with PCA largely by illustration. The mathematics is dealt with in more detail in Appendix D; here only brief notice is provided of the role played by the covariance matrix.

If $\mathbf{S}$ is the covariance matrix of $\mathbf{Y}$, the data matrix used for analysis,

$$\mathbf{S} = \mathbf{V}\mathbf{D}^2\mathbf{V}'$$

where $\mathbf{V}$and $\mathbf{D}^2 = \mathbf{\Lambda}$ are $p \times p$ matrices, and the latter is diagonal with the diagonal elements, assuming they are ordered, corresponding to the variance of the PCs. The elements of the latter, λ_i, are *eigenvalues* – a term that features in some software output. The columns of $\mathbf{V}$ are *eigenvectors* and contain the coefficients of the PCs. Thus, all the ingredients for the practical analysis and interpretation of a PCA can be obtained from the covariance matrix $\mathbf{S}$ (or $\mathbf{R}$ if the data are standardized). Principal component scores are also required and these are obtained from the $n \times p$ matrix $\mathbf{YV}$.

C.2.3 Mahalanobis distance and LDA

More on MD and notation

Define $\mathbf{p}$ to be a vector with p terms, with $\mathbf{q}$ similarly defined. Mahalanobis distance (MD) can be defined as

$$\tilde{d}^2 = (\mathbf{p} - \mathbf{q})'\tilde{\mathbf{S}}^{-1}(\mathbf{p} - \mathbf{q})$$

where definition of the terms depends on the particular version of MD used. If there are G groups denote the estimated covariance matrix of group g as $\mathbf{S}_g$, using $\mathbf{S}$ for the situation where there is just one group. Assuming common population covariance matrices a weighted average can be defined as

$$\mathbf{S}_w = \sum_{g=1}^{G} \frac{(n_g - 1)\mathbf{S}_g}{(N - G)}$$

where $N = (n_1 + n_2 + \ldots + n_G)$ is the sum of the sample sizes, n_g, for each group. The term $\mathbf{S}_w$ is the *within-group* covariance matrix. It can be thought of as a measure of the 'compactness' of the groups. A particular case is when there are two groups, where

$$\mathbf{S}_w = \frac{(n_1 - 1)\mathbf{S}_1 + (n_2 - 1)\mathbf{S}_2}{(n_1 + n_2 - 2)}.$$

If $\bar{\mathbf{x}}_1$ and $\bar{\mathbf{x}}_2$ are the means for the two groups, MD as defined earlier becomes

$$d_{12}^2 = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_w^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2).$$

The other case of interest here is that of the MD of a single case, $\mathbf{w}_i$, from a group with mean $\bar{\mathbf{x}}$ and covariance matrix $\mathbf{S}$; this can be written as

$$d_i^2 = (\mathbf{w}_i - \bar{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{w}_i - \bar{\mathbf{x}}).$$

Two situations can be distinguished; the first is that $\mathbf{w}_i$ is not a member of the reference group, the second is that $\mathbf{w}_i = \mathbf{x}_i$ is a member of the group. The latter situation introduces the complication that $\mathbf{x}_i$ influences the calculation of the group statistics $\bar{\mathbf{x}}$ and $\mathbf{S}$. This has the effect of reducing the size of the MD compared to calculations that omit it. The obvious way to remedy this is to use leave-one-out (LOO) calculations (Section 11.2.2) where $d_{(i)}^2$, $\bar{\mathbf{x}}_{(i)}$) and $\mathbf{S}_{(i)}$ replace corresponding terms in the above formula, the (i) subscript indicating that case i has been omitted from calculations[3]. The implication would seem to be that to use LOO calculations n separate analyses are needed as the mean and covariance calculations change for each case. This is not necessary because it is possible to convert d_i^2 to $d_{(i)}^2$ without the need for such calculations (e.g., Section 6.4.3 of Baxter, 2003).

MD in LDA

So far MD has simply been defined; some insight into, and 'justification' for, it can be gained by considering LDA for the two-group case assuming equal population covariance matrices. With $G = 2$ there is one discriminant function that, for case i, gives rise to scores of the form $\mathbf{a}'\mathbf{x}_i$ and the task is to determine the p elements of $\mathbf{a}$ that maximize group separation on the transformed scale.

The within-group sample covariance matrix, $\mathbf{S}_w$, has been defined above for two groups. It can be thought of as an averaged 'measure' of the 'compactness' of the groups. It is also possible to define a between-groups sample covariance matrix which, in weighted form and for two groups can be written as

$$\mathbf{S}_b = n_1(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}})(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}})' + n_2(\bar{\mathbf{x}}_2 - \bar{\mathbf{x}})(\bar{\mathbf{x}}_2 - \bar{\mathbf{x}})'$$

[3]The term *jackknife* is also sometimes used for LOO calculations.

where $\bar{\mathbf{x}}$ is the mean of all the data. This can be thought of as a 'measure' of the 'separation' of the two groups.

After performing a discriminant analysis, which requires $\mathbf{a}$ to be determined, the centroids of the two groups of transformed data can be denoted as $\bar{\mathbf{z}}_1$ and $\bar{\mathbf{z}}_2$. We want the distance between these to be as large as possible, and this depends on $\mathbf{S}_b$, but this needs to be balanced against the wish to keep the groups as compact as possible which depend on $\mathbf{S}_w$. Fisher's (1936) idea, which does not depend on distributional assumptions, was to determine $\mathbf{a}$ to maximize the ratio

$$\mathbf{a}'\mathbf{S}_b\mathbf{a}/\mathbf{a}'\mathbf{S}_w\mathbf{a}.$$

Thus the initial idea of LDA, with the desire to transform the data so that the predefined groups are as distinct has possible, has been converted to a mathematical problem the solution of which is to obtain $\mathbf{a}$ from the eigenvectors of

$$\mathbf{S}_w^{-1}\mathbf{S}_b.$$

It further transpires that the Euclidean distance betwwen the transformed group means, $\bar{\mathbf{z}}_1$ and $\bar{\mathbf{z}}_2$, is given by

$$(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_w^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$$

which is just the MD as defined, for the comparison of group means, but without explanation, at the start of this section. Here we have seen how MD arises 'naturally' as a measure of distance in the context of LDA.

Constructing confidence ellipsoids

For a single group, when LOO calculations are appropriate, the set of values for which $d_i^2 = c$, where c is a constant, define ellipsoidal contours. To assign a confidence level to the contours it is necessary to assume bivariate normality for two-dimensional plots. Theory exists that shows that a suitable transformation of d_i^2 exists that is approximately distributed as an F-statistic with $(p, n-p-1)$ degrees of freedom, or a chi-squared statistic with p degrees of freedom if n is large in relation to p The choice of a value for F or chi-squared determine the confidence level, The formulae are given in Baxter (2003: 71) with references to the original theoretical derivations.

It can be noted that such calculation allow probabilities of group membership to be assigned to individual cases, so it can be judged if they are plausible group members or not. This differs from the usage in Table 11.1 where results are presented in terms of relative probabilities that assume cases must belong to one or other of the groups included in an analysis.