4.2 Multivariate random numbers

(a) and (b). We have Gaussians of given variance and a correlation coefficient; we have a prescription of §4.2 to generate (x, y) pairs which are so related. Why is this prescription right? Consider a simple application of the prescription to show why.

Take $\sigma_x = 1.0$, $\sigma_y = 1.0$, and values of $\rho = 0.1, 0.5$ and 0.9, so that we have a range of correlation from bad (0.1) to good (0.9). (Note that if we have been given ρ and the σ's, we know the covariance from $\rho = cov(x, y)/\sigma_x\sigma_y$; likewise if we have the covariance and the σ's we know ρ.)

There are four steps in the prescription:

1. Focus on the case of $\rho = 0.5$; its covariance matrix is (§4.2)

$$\begin{vmatrix} e_{1,1} & e_{1,2} \\ e_{2,1} & e_{2,2} \end{vmatrix} = \begin{vmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{vmatrix} = \begin{vmatrix} 1.0 & 0.5 \\ 0.5 & 1.0 \end{vmatrix} \tag{1}$$

2. The eigenvectors come from the identity

$$\begin{vmatrix} 1.0 & 0.5 \\ 0.5 & 1.0 \end{vmatrix} \begin{vmatrix} x \\ y \end{vmatrix} = \lambda \begin{vmatrix} x \\ y \end{vmatrix} \tag{2}$$

which gives eigenvalues of λ to construct the two eigenvectors, the vectors that are invariant under the transformation of the covariance matrix. From this equation we get

$$\begin{aligned} x + 0.5y &= \lambda x \qquad (3) \\ 0.5x + y &= \lambda y, \qquad (4) \\ \text{and thus } \; 2\lambda^2 - 2\lambda + 1.5 &= 0 \qquad (5) \end{aligned}$$

from which emerge the eigenvalues $\lambda_1 = 0.5$ and $\lambda_2 = 1.5$, and the eigenvectors (from either equation (3) or (4)), are

$$\begin{pmatrix} 1.0 \\ 1.0 \end{pmatrix} \text{ and } \begin{pmatrix} -1.0 \\ 1.0 \end{pmatrix} \tag{6}$$

These can be normalized arbitrarily; following standard practice we make them unit length.

3. This yields the transformation matrix

$$T = \begin{vmatrix} 0.707 & -0.707 \\ 0.707 & 0.707 \end{vmatrix} \tag{7}$$

4. We've done the hard part; we now simply draw random (X', Y') Gaussian pairs with variances

$$\begin{aligned} \sigma_{X'}^2 &= \lambda_1 \sigma_x^2 = 0.5 \qquad (8) \\ \sigma_{Y'}^2 &= \lambda_2 \sigma_y^2 = 1.5, \qquad (9) \end{aligned}$$

easily done with a routine such as *gasdev* from *Numerical Recipes.* The final stage is to compute the (X, Y) pairs according to

$$\begin{pmatrix} X \\ Y \end{pmatrix} = [T] \begin{pmatrix} X' \\ Y' \end{pmatrix} = \begin{vmatrix} 0.707 & -0.707 \\ 0.707 & 0.707 \end{vmatrix} \begin{pmatrix} X' \\ Y' \end{pmatrix} \tag{10}$$

Results are shown in Figure 1, centre panel.

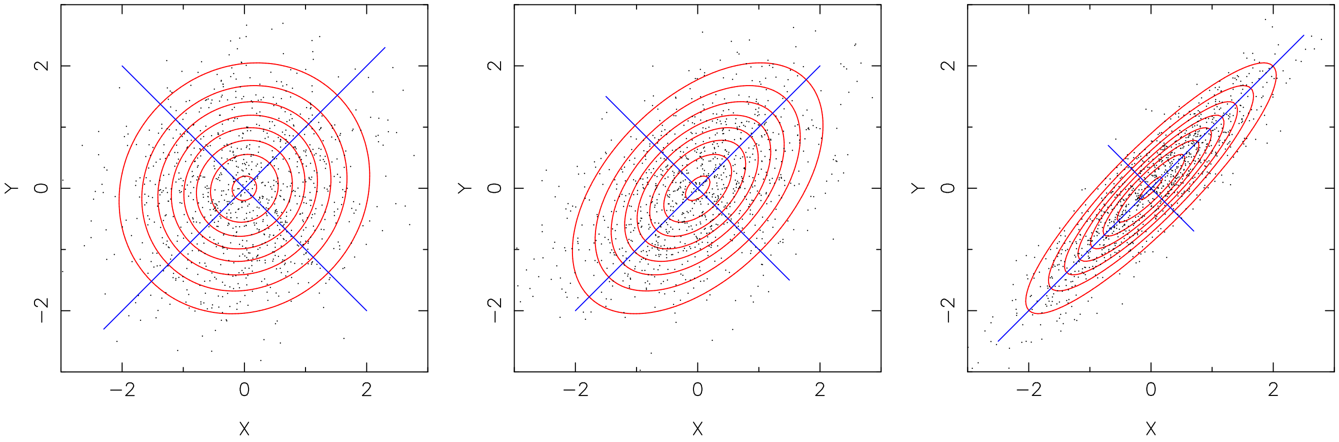

Figure 1: Bivariate random numbers. 1000 random numbers were drawn according to the bivariate Gaussian prescription in which $\sigma_x = \sigma_y = 1.0$ and $\rho = 0.1$ (left), 0.5 (centre) and 0.9 (right). The ellipses contour the 'bivariate Gaussian mountain' function described below.

Now that we have carried out the exercise and produced the diagrams, we see that our problem is identical to that described in the Principal Component Analysis (PCA) section of the book, § 4.5. In the above, we created an idealized correlation for which we formed the error matrix, calculated the eigenvalues, the eigenvectors, and hence the transpose matrix. This transpose matrix diagonalizes the error matrix, reducing its cross-terms to zero.

With the transpose matrix to hand, we then took independent random values of (X', Y'), scaled these with the eigenvalues to get the right total projection lengths, and used the matrix to reinsert the cross-terms in order to get the final (correlated and rotated) set of (X, Y).

Rather than this 'by analogy' explanation, a formal proof can follow from the formal *definition of 'jointly Gaussian'*:

Random variables $X_1, X_2, \ldots, X_n$ are said to be *jointly Gaussian* or to have a *multivariate Gaussian distribution* if their joint probability density function can be written in the following format:

$$p(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^n |C|}} \exp^{-\frac{1}{2}\mathbf{x}^T \mathbf{C}^{-1} \mathbf{x}}, \tag{11}$$

$$\text{where the vector } \ \mathbf{x} = X_1, X_2, \ldots, X_n, \tag{12}$$

$$\text{the transpose of } \mathbf{x} = \mathbf{x}^T, \tag{13}$$

$$\text{the covariance matrix } \mathbf{C} = \begin{vmatrix} < x_1^2 > & < x_1 x_2 > & \\ < x_1 x_2 > & < x_2^2 > & \\ & & \ddots \end{vmatrix} \tag{14}$$

$$\text{and the determinant of } \mathbf{C} = |C|. \tag{15}$$

The formal proof can be an exercise for the student; following the ethos of the book, we refrain from presenting it here.

For the bivariate $n = 2$ case, the exponent of equation 11 ends up as a quadratic. If we relabel x_1 as x and x_2 as y, it is in fact

$$Q = -\frac{1}{2}\frac{1}{(1-\rho^2)}\left(\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2} - \frac{2\rho xy}{\sigma_x \sigma_y}\right) \tag{16}$$

which appears as the exponent in book equation 4.1. This 2D probability function describes the bivariate Gaussian 'mountain'; and the contours describing this function are a concentric set of ellipses. It is these contours which appear in Figure 1 above.

(c) Calculations of the error matrices for the 1000 (X, Y) points of each of the three values of ρ yields

$$\begin{vmatrix} 0.971 & 0.112 \\ 0.112 & 1.028 \end{vmatrix}_{\rho=0.1} \begin{vmatrix} 1.060 & 0.478 \\ 0.478 & 0.901 \end{vmatrix}_{\rho=0.5} \begin{vmatrix} 1.023 & 0.943 \\ 0.943 & 1.053 \end{vmatrix}_{\rho=0.9} \tag{17}$$

to be compared with (see equation 1)

$$\begin{vmatrix} 1.000 & 0.100 \\ 0.100 & 1.000 \end{vmatrix}_{\rho=0.1} \begin{vmatrix} 1.000 & 0.500 \\ 0.500 & 1.000 \end{vmatrix}_{\rho=0.5} \begin{vmatrix} 1.000 & 0.900 \\ 0.900 & 1.000 \end{vmatrix}_{\rho=0.9} \tag{18}$$

We may safely assume that we have done it right; the differences are due to scatter.

Further work:

(1) Try simulations of 10^6 points to ensure that the above matrices converge to exactly the advertised values.

(2) Multiply out the matrices in the exponent of equation 11 for the bivariate case to check out equation 16.

(3) The eigenvalues are very different for each value of ρ above, but eigenvectors and the transpose matrices (equation 10) are the same. Consider why this is the case; what would make the transpose matrices differ? Generate some examples of scatter plots producing/using different transpose matrices. How does the correlation coefficient relate to the slope of the correlation?

(4) Look at the scatter about the outer contours of the rightmost member of Figure 1, that for the highest correlation. There seems to be less scatter outside the minor axis of the ellipse than the major. Why?

(5) For case in which ρ is significantly greater that zero, the scatter in X appears diminished from the $\sigma_x = 1.0$ with which it started. Why, where has it gone? Try producing the diagram with $\rho = 1$, starting with lots of scatter in X and Y. Is this result what you expected? How can the perfect straight line with zero apparent scatter be right?

Finally note that the deficiencies of the bivariate Gaussian as a model to explore correlation are exposed by these considerations yet again. (See exercise 4.1 and its solution.) If there is some scatter due to measurement error, but in fact there is a *perfect* correlation between underlying x and y, the classical correlation tests will fail to reveal it. *Only a straight line yields* $\rho = 1$. Further, it is important to emphasize that in measurement, frequently x variable is not randomly selected - and in this case the model is entirely inappropriate. The message from exercise 4.1 and this exercise is - ranking (non-parametric) tests should be favoured in looking for correlations, and probabilities can be assessed for these tests by bootstrap or jackknife procedures.