GENERAL PRINCIPLES FOR ESTIMATION OF COVARIANCE MATRICES

by John Parker Burg

Suppose we have N samples of a pair of random variables, (x_n, y_n), $n = 1$ to N, and that we wish to estimate their covariance matrix, i.e.,

$$\begin{bmatrix} \overline{x^2} & \overline{xy} \\ \overline{xy} & \overline{y^2} \end{bmatrix} .$$

The normal straightforward method is to estimate this matrix by

$$\frac{1}{N} \sum_{n=1}^{N} \left\{ \begin{matrix} x_n \\ y_n \end{matrix} \right\} \left\{ \begin{matrix} x_n & y_n \end{matrix} \right\} = \begin{bmatrix} \frac{1}{N} \sum_{n=1}^{N} x_n^2 & \frac{1}{N} \sum_{n=1}^{N} x_n y_n \\ \frac{1}{N} \sum_{n=1}^{N} x_n y_n & \frac{1}{N} \sum_{n=1}^{N} y_n^2 \end{bmatrix} \equiv \begin{bmatrix} A & D \\ D & B \end{bmatrix}$$

This paper agrees that this estimate of the covariance matrix of (x,y) is the "best" estimate in the absence of any other knowledge about the random variables. However, if the variance of x is known to be the same as the variance of y, then the above estimate is not acceptable since A will not be equal to B except by chance.

The general principle of covariance estimation introduced by this paper will show how to estimate the covariance matrix when there are constraints on the form of the covariance matrix. For example, if there is a'priori knowledge that the variances are equal, then the main diagonal elements of the estimated covariance matrix must be equal. It will be seen that the general principle is the same as the normal straightforward method when there is no a' priori knowledge about the form of the covariance matrix.

The new principle is based on the following observations. If we have a pair of random variables, (w,z), and we know that their variances are the same, then their covariance matrix is of the form

$$\overline{\begin{Bmatrix} w \\ z \end{Bmatrix} \{w \quad z\}} = \begin{bmatrix} a & d \\ d & a \end{bmatrix} \qquad (1)$$

Suppose we choose a number, -c, (not necessarily the optimum number), by which to multiply z for the purpose of estimating w. The mean square error in this prediction is given by

$$\{1 \quad c\} \begin{bmatrix} a & d \\ d & a \end{bmatrix} \begin{Bmatrix} 1 \\ c \end{Bmatrix} = a(1 + c^2) + 2cd \, .$$

The value of c which minimizes the mean square error is found from the prediction error filter equation

$$\begin{bmatrix} a & d \\ d & a \end{bmatrix} \begin{Bmatrix} 1 \\ c \end{Bmatrix} = \begin{Bmatrix} P \\ 0 \end{Bmatrix} ,$$

where $c = -d/a$ and $P = a(1 - c^2) = a - d^2/a$ = minimum mean square error.

If we choose a number, - q, to apply to w to predict z, the mean square error is given by

$$\{q \quad 1\}\begin{bmatrix} a & d \\ d & a \end{bmatrix}\begin{Bmatrix} q \\ 1 \end{Bmatrix} = a(1 + q^2) + 2qd.$$

The optimum value of q is given by

$$\begin{bmatrix} a & d \\ d & a \end{bmatrix}\begin{Bmatrix} q \\ 1 \end{Bmatrix} = \begin{Bmatrix} 0 \\ Q \end{Bmatrix} ,$$

or $q= -d/a$ and $Q = a(1 - q^2) = a - d^2/a$ = minimum mean square error.

We thus see that the optimum value for c is the same as the optimum value for q and that the minimum mean square error in predicting z is the same as the minimum mean square error in predicting w. Of more subtle importance, however, is the fact that

$$\overline{(w+cz)^2} = \overline{(z+cw)^2} = a(1+c^2) + 2cd$$

independent of the value of c. Thus if one is interested in estimating the error in predicting w from z, using an arbitrary value for c, one could use the sample average of $(w + cz)^2$ or one could use the sample average of $(z + cw)^2$! From the complete symmetry of the situation, using the equally weighted average of these two quantities is clearly better than either one individually. Thus, if we have N samples (w_n, z_n), $n = 1$ to N, the best estimate of the mean square error in predicting w from z or z from w, using the number $-c$, is

$$\frac{1}{2N}\sum_{n=1}^{N}[(w_n + cz_n)^2 + (z_n + cw_n)^2] . \qquad (2)$$

We now use this estimate of the mean square error for an arbitrary value of c to estimate the optimum value for c. This is done by simply choosing the value of c which minimizes (2), which is

$$c = \frac{-2 \sum_{n=1}^{N} w_n z_n}{\sum_{n=1}^{N} (w_n^2 + z_n^2)} \tag{3}$$

Turning now to the problem of estimating the covariance matrix (1), the estimate for the variance of w and z is

$$a = \frac{1}{2N} \sum_{n=1}^{N} (w_n^2 + z_n^2) \; . \tag{4}$$

Using the estimated value of c, Eq. 3, and the relation $c = d/a$, the estimate for d is simply

$$d = \frac{1}{N} \sum_{n=1}^{N} w_n z_n \; . \tag{5}$$

Thus, if the raw sample covariance matrix is

$$\frac{1}{N} \sum_{n=1}^{N} \begin{Bmatrix} w_n \\ z_n \end{Bmatrix} \{ w_n \quad z_n \} = \begin{bmatrix} \frac{1}{N} \sum_{n=1}^{N} w_n^2 & \frac{1}{N} \sum_{n=1}^{N} w_n z_n \\ \frac{1}{N} \sum_{n=1}^{N} w_n z_n & \frac{1}{N} \sum_{n=1}^{N} z_n^2 \end{bmatrix} = \begin{bmatrix} A & D \\ D & B \end{bmatrix}$$

then the final estimated covariance matrix is

$$\begin{bmatrix} \frac{A+B}{2} & D \\ D & \frac{A+B}{2} \end{bmatrix} \; . \tag{6}$$

In light of the above observations, we can now state the first general principle for estimating the covariance matrix of two random variables.

Estimate the variances of the two random variables using any á priori information. Then, if necessary, make the two variances the same and estimate the correlation coefficient, c. Using c, find the corresponding value of the covariance between the two original random variables.

The result of doing this when the variances are known to be equal is given by (6). As a further example, let us apply this principle when there is no á priori knowledge. Let the sample covariance matrix be

$$\frac{1}{N}\sum_{n=1}^{N}\begin{Bmatrix} x_n \\ y_n \end{Bmatrix}\begin{Bmatrix} x_n & y_n \end{Bmatrix} = \begin{bmatrix} A & D \\ D & B \end{bmatrix} . \qquad (7)$$

Here, A and B are the estimates of the variances of x and y. Thus $x/(A)^{1/2}$ and $y/(B)^{1/2}$ should have the same variance, namely unity. The correlation coefficient is estimated as

$$c = \frac{2\,\frac{1}{N}\sum_{n=1}^{N}\frac{x_n}{(A)^{1/2}}\,\frac{y_n}{(B)^{1/2}}}{\frac{1}{N}\sum_{n=1}^{N}\left(\frac{x_n^2}{A} + \frac{y_n^2}{B}\right)} = \left(\frac{1}{N}\sum_{n=1}^{N} x_n y_n\right)/(AB)^{1/2} = D/(AB)^{1/2}$$

Since normalizing x and y to unity produces the sample covariance matrix of

$$\begin{bmatrix} 1 & D/(AB)^{1/2} \\ D/(AB)^{1/2} & 1 \end{bmatrix} ,$$

we see that D is the estimated covariance when the general principle is used. Thus, in the absence of á priori knowledge, the general principle picks the raw sample covariance matrix, (7), as its estimate.

These two examples virtually conclude the study of a pair of random variables, so let us look at the case of three random variables (x,y,z). Let us first take on the problem where it is known that the three variances are equal. Let the sample covariance matrix be

$$\frac{1}{N}\sum_{n=1}^{N}\begin{Bmatrix} x_n \\ y_n \\ z_n \end{Bmatrix}\begin{Bmatrix} x_n & y_n & z_n \end{Bmatrix} = \begin{bmatrix} A & D & F \\ D & B & E \\ F & E & C \end{bmatrix} . \qquad (8)$$

The estimated variance, σ^2, should thus be (A + B + C)/3 and the estimated covariance matrix should be

$$\left(\frac{A+B+C}{3}\right)\begin{bmatrix} 1 & s & t \\ s & 1 & r \\ t & r & 1 \end{bmatrix} , \qquad (9)$$

where s, r and t are the correlation coefficients between x, y and z. If one blindly uses the results of the two variable theory, a reasonable final estimate might be

$$\left(\frac{A+B+C}{3}\right)\begin{bmatrix} 1 & \frac{2D}{A+B} & \frac{2F}{A+C} \\ \frac{2D}{A+B} & 1 & \frac{2E}{B+C} \\ \frac{2F}{A+C} & \frac{2E}{B+C} & 1 \end{bmatrix} . \qquad (10)$$

However, one should be nervous about (10). It should be all right to use the two variable results to estimate the correlation coefficient between x and y and between y and z, but the coefficient between x and z will then depend on those two estimates since (10) must be a non-negative definite matrix. From another point of view, the estimate of the correlation between x and z should be affected by the correlation that x and z have with y. The solution of this problem is to remove the influence of y by analyzing the residuals in the prediction of x and z from y instead of analyzing x and z directly. I.e., let us define u and v as

$$u = \frac{x-sy}{\sigma(1-s^2)^{1/2}} \qquad \text{and} \qquad v = \frac{z-ry}{\sigma(1-r^2)^{1/2}} .$$

Then u and v have unit variance and are linearly independent of y. Thus we have complete freedom in the value of the correlation coefficient, q, between u and v so long as its magnitude is not greater than unity. We estimate q as

$$q = \frac{\frac{2}{N}\Sigma u_n v_n}{\frac{1}{N}\Sigma u_n^2+v_n^2} = \frac{\frac{2}{N}\Sigma\frac{(x_n-sy_n)(z_n-ry_n)}{\sigma^2((1-s^2)(1-r^2))^{1/2}}}{\frac{1}{N}\Sigma\left[\frac{(x_n-sy_n)^2}{\sigma^2(1-s^2)}+\frac{(z_n-ry_n)^2}{\sigma^2(1-r^2)}\right]} =$$

$$\frac{\frac{2}{N}\Sigma\frac{x_nz_n-sy_nz_n-rx_ny_n+sry_n^2}{((1-s^2)(1-r^2))^{1/2}}}{\frac{1}{N}\Sigma\left[\frac{x_n^2-2sx_ny_n+s^2y_n^2}{(1-s^2)}+\frac{z_n^2-2rz_ny_n+r^2y_n^2}{(1-r^2)}\right]} =$$

$$\frac{2\,\frac{F-sE-rD+srB}{((1-s^2)(1-r^2))^{1/2}}}{\frac{A-2sD+s^2B}{(1-s^2)}+\frac{C-2rE+r^2B}{1-r^2}} .$$

This simplifies by noting that $2D = s(A+B)$ and $2E = r(B+C)$, so that $A - 2sD + s^2B = A - s^2(A+B) + s^2B = A(1-s^2)$, $C - 2rE + r^2B = C - r^2(B+C) + r^2B = C(1-r^2)$ and $2F - 2sE - 2rD + 2srB = 2F - sr(B+C) - sr(A+B) + 2srB = 2F - sr(A+C)$. Thus we have

$$q = \frac{2F-sr(A+C)}{(A+C)((1-s^2)(1-r^2))^{1/2}} = \frac{\frac{2F}{A+C} - sr}{((1-s^2)(1-r^2))^{1/2}} . \qquad (11)$$

We would now like to see what covariance between x and z is implied by this value of q. It is simpler to see what value of q is implied by (9) by calculating

$$q = \left\{1/(1-s^2)^{1/2}, - s/(1-s^2)^{1/2}, 0\right\} \begin{bmatrix} 1 & s & t \\ s & 1 & r \\ t & r & 1 \end{bmatrix} \left\{\begin{matrix} 0 \\ -r/(1-r^2)^{1/2} \\ 1/(1-r^2)^{1/2} \end{matrix}\right\}$$

$$= \left\{(1-s^2)^{1/2} \quad 0 \quad (t-sr)/(1-s^2)^{1/2}\right\} \left\{\begin{matrix} 0 \\ -r/(1-r^2)^{1/2} \\ 1/(1-r^2)^{1/2} \end{matrix}\right\} =$$

$$\frac{t-sr}{((1-s^2)(1-r^2)^{1/2}} .$$

Comparing with (11), we note that $t = 2F/(A+C)$ and we still get (10) for our covariance estimate. This happy result has the following conclusions for estimate (10).

(1) The order in which we analyze the pairs of variables does not change the estimate.

(2) Since our estimates of s, r and q have magnitudes bounded by unity, (10) is at least non-negative definite. Thus (10) is a possible

covariance matrix. This proves the matrix theorem that if (8) is non-negative definite, then so is (10).

(3) Of course, as $N \to \infty$, (10) goes to the true covariance matrix.

(4) Another property is that if (10) happens to be the raw covariance matrix, then application of the general principle leads to (10) again. Thus, if the raw covariance matrix agrees with the constraints, then there is no change in the answer.

(5) It will be proven shortly that the determinant of (10) is equal to or greater than the determinant of the raw covariance matrix, with equality occurring only if the sample variances are equal to start with.

We have just seen that our estimate of the covariance matrix (10) has internal consistency in that the analysis of (8) in terms of the residuals in predicting x and z from y gives the same covariance between x and z in (10) as does a direct analysis ignoring y. We will now show that this is true for any number of variables where their variances are known to be equal.

We will use the 4 variable case, (x_1, x_2, x_3, x_4), to illustrate the general proof. Let the raw sample covariance matrix be

$$\begin{bmatrix} A & E & H & J \\ E & B & F & I \\ H & F & C & G \\ J & I & G & D \end{bmatrix} , \qquad (12)$$

which leads to the covariance estimate of

$$\left(\frac{A+B+C+D}{4}\right)\begin{bmatrix} 1 & \frac{2\ E}{A+B} & \frac{2\ H}{A+C} & \frac{2\ J}{A+D} \\ \frac{2\ E}{A+B} & 1 & \frac{2\ F}{B+C} & \frac{2\ I}{B+D} \\ \frac{2\ H}{A+C} & \frac{2\ F}{B+C} & 1 & \frac{2\ G}{C+D} \\ \frac{2\ J}{A+D} & \frac{2\ I}{B+D} & \frac{2\ G}{C+D} & 1 \end{bmatrix} \equiv \left(\frac{A+B+C+D}{4}\right)\begin{bmatrix} 1 & p & s & w \\ p & 1 & q & t \\ s & q & 1 & r \\ w & t & r & 1 \end{bmatrix} \qquad (13)$$

From our 3 variable work, we have proven that the 3 by 3 submatrices of (13) are non-negative definite if (12) is non-negative definite. We will now first show that the estimate (13) is consistent with our general principles of covariance estimation. We do this by considering the prediction error filter for predicting x_1 from x_2 and x_3 and the prediction error filter for predicting x_4 from x_2 and x_3, i.e.

$$\begin{bmatrix} 1 & p & s \\ p & 1 & q \\ s & q & 1 \end{bmatrix}\begin{Bmatrix} 1 \\ \alpha_1 \\ \alpha_2 \end{Bmatrix} = \begin{Bmatrix} P \\ 0 \\ 0 \end{Bmatrix}, \begin{bmatrix} 1 & q & t \\ q & 1 & r \\ t & r & 1 \end{bmatrix}\begin{Bmatrix} \beta_2 \\ \beta_1 \\ 1 \end{Bmatrix} = \begin{Bmatrix} 0 \\ 0 \\ Q \end{Bmatrix}. \qquad (14)$$

We shall apply these filters to our sample covariance matrix, but for mathematical simplicity, we shall multiply (12) by two first. Then using the definitions of p, q, r, s, t and w given in (13), we have for our doubled sample covariance matrix

$$\begin{bmatrix} 2\ A & p(A+B) & s(A+C) & w(A+D) \\ p(A+B) & 2\ B & q(B+C) & t(B+D) \\ s(A+C) & q(B+C) & 2\ C & r(C+D) \\ w(A+D) & t(B+D) & r(C+D) & 2\ D \end{bmatrix}$$

$$= \begin{bmatrix} A & pB & sC & wD \\ pA & B & qC & tD \\ sA & qB & C & rD \\ wA & tB & rC & D \end{bmatrix} + \begin{bmatrix} A & pA & sA & wA \\ pB & B & qB & tB \\ sC & qC & C & rC \\ wD & tD & rD & D \end{bmatrix} . \qquad (15)$$

Post multiplying by $\begin{bmatrix} 1 & 0 \\ \alpha_1 & \beta_2 \\ \alpha_2 & \beta_1 \\ 0 & 1 \end{bmatrix}$

and using (14), we get

$$\begin{bmatrix} A & pB & sC & wD \\ pA & B & qC & tD \\ sA & qB & C & rD \\ wA & tB & rC & D \end{bmatrix} \begin{bmatrix} 1 & 0 \\ \alpha_1 & \beta_2 \\ \alpha_2 & \beta_1 \\ 0 & 1 \end{bmatrix} + \begin{bmatrix} AP & A\left[w+s\beta_1+p\beta_2\right] \\ 0 & 0 \\ 0 & 0 \\ D\left[w+t\alpha,+r\alpha_2\right] & DQ \end{bmatrix}$$

Premultiplying by $\begin{bmatrix} 1 & \alpha_1 & \alpha_2 & 0 \\ 0 & \beta_2 & \beta_1 & 1 \end{bmatrix}$, we get

$$\begin{bmatrix} AP & 0 & 0 & D\left[w+t\alpha_1+r\alpha_2\right] \\ A\left[w+s\beta_1+p\beta_2\right] & 0 & 0 & DQ \end{bmatrix} \begin{bmatrix} 1 & 0 \\ \alpha_1 & \beta_2 \\ \alpha_2 & \beta_1 \\ 0 & 1 \end{bmatrix} + \begin{bmatrix} AP & A\left[w+s\beta_1+p\beta_2\right] \\ D\left[w+t\alpha_1+r\alpha_2\right] & DQ \end{bmatrix}$$

$$= \begin{bmatrix} 2\ AP & D\left[w+t\alpha_1+r\alpha_2\right] + A\left[w+s\beta_1+p\beta_2\right] \\ A\left[w+s\beta_1+p\beta_2\right] + D\left[w+t\alpha_1+r\alpha_2\right] & 2\ DQ \end{bmatrix}, \qquad (16)$$

which is the sample covariance matrix of the residuals in predicting x_1 and x_4 from x_2 and x_3. If the prediction error filters are applied to the estimated covariance matrix (13), we get

$$\begin{bmatrix} 1 & \alpha_1 & \alpha_2 & 0 \\ 0 & \beta_2 & \beta_1 & 1 \end{bmatrix} \begin{bmatrix} 1 & p & s & w \\ p & 1 & q & t \\ s & q & 1 & r \\ w & t & r & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ \alpha_1 & \beta_2 \\ \alpha_2 & \beta_1 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} P & w+s\beta_1+p\beta_2 \\ w+t\alpha_1+r\alpha_2 & Q \end{bmatrix} \qquad (17)$$

which shows that $w+t\alpha_1+rx_2 = w+s\beta_1+p\beta_2$ since the matrix must be symmetric. Thus (16) simplifies to

$$\begin{bmatrix} 2\ AP & (A+D)\ (w+t\alpha_1+r\alpha_2) \\ (A+D)\ (w+t\alpha_1+r\alpha_2) & 2\ DQ \end{bmatrix} \qquad (18).$$

Using P and Q to make the variances of the residuals equal and using the general principle to estimate the correlation coefficients for (17) and (18), we see that the common answer is

$$\frac{2(A+D)\ (w+t\alpha_1+r\alpha_2)}{(2A+2D)\sqrt{PQ}} = \frac{w + t\alpha_1 + r\alpha_2}{\sqrt{PQ}} .$$

This shows that using 2J/(A+D) in (13) for the estimated correlation between x_1 and x_4 gives the same correlation between the x_1 and x_4 residuals as obtained from the raw sample covariance matrix.

We now prove that the determinant of (13) is equal to or greater than the determinant of (12). Let us use the prediction error filter decomposition, where the $*$ values are not needed explicitly.

$$\begin{bmatrix} 1 & p & s & w \\ p & 1 & q & t \\ s & q & 1 & r \\ w & t & r & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ \gamma_1 & 1 & 0 & 0 \\ \gamma_2 & \varepsilon_1 & 1 & 0 \\ \gamma_3 & \varepsilon_2 & \mu_1 & 1 \end{bmatrix} = \begin{bmatrix} P & * & * & * \\ 0 & Q & * & * \\ 0 & 0 & R & * \\ 0 & 0 & 0 & S \end{bmatrix} .$$

Then

$$\begin{bmatrix} 1 & \gamma_1 & \gamma_2 & \gamma_3 \\ 0 & 1 & \varepsilon_1 & \varepsilon_2 \\ 0 & 0 & 1 & \mu_1 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & p & s & w \\ p & 1 & q & t \\ s & q & 1 & r \\ w & t & r & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ \gamma_1 & 1 & 0 & 0 \\ \gamma_2 & \varepsilon_1 & 1 & 0 \\ \gamma_3 & \varepsilon_2 & \mu_1 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} P & 0 & 0 & 0 \\ 0 & Q & 0 & 0 \\ 0 & 0 & R & 0 \\ 0 & 0 & 0 & S \end{bmatrix} .$$

Thus the determinant of (13) is $(\frac{A+B+C+D}{4})^4$ PQRS.

Making the same transformation on (12), we get a non-negative definite matrix of the form

$$\begin{bmatrix} AP & * & * & * \\ * & BQ & * & * \\ * & * & CR & * \\ * & * & * & DS \end{bmatrix}$$

where the $*$'s here are not normally zero. The determinant of this matrix is equal to or less than the product of the main diagonal, the equality coming only if all the $*$ terms are zero, which will occur only if $A=B=C=D$ so that (12) = (13).

Thus we have

$$\text{determinant of (12)} \leq ABCD\ PQRS \leq \left(\frac{A+B+C+D}{4}\right)^4 PQRS$$

by the geometric inequality, with equality again occurring only if $A=B=C=D$.

We should note that (13) is uniformly less singular than (12) in the sense that every submatrix of (13) is more stable than the corresponding submatrix of (12).