CORRELATION, COST RISK, AND GEOMETRY

Edwin B. Dean
NASA Langley Research Center
Hampton VA 23681-0001
V: 804 864 8213
F: 804 864 8203

BACKGROUND

Parametric cost risk is a statistical phenomena. One first assumes that the cost is defined by

$$C = h(p_1, ..., p_m)$$

where h is a function of the parameters $p_1, ..., p_m$.

Second, one assumes that each of the parameters is a random variable. This applies to a single cost estimating relationship (CER) which might be in the typical Cobb-Douglas form

$$C = p_1^{q1} ... p_m^{qm}$$

where the q_i are the elasticities, or to the sum of n work breakdown structure (WBS) elements p_i in the form

$$C = p_1 + ... + p_n.$$

In a complete cost risk simulation the cost of each WBS element would be a function h_i of parameters $p_1, ..., p_m$ with the form

$$C = h_1(p) + ... + h_n(p)$$

where p is the vector

$$p = \begin{bmatrix} p_1 \\ . \\ p_m \end{bmatrix}$$

with components p_i.

Third, one must make an assumption about the dependence of the variables within the variable set. One may assume that the variables are statistically independent, that the variables are totally dependent, or that correlation exists between selected pairs of variables. If the assumption of independence is made, then the distribution function $F(C \leq \text{constant})$ of the WBS elements becomes arbitrarily narrow as more WBS elements are added. If this were the case then we could converge on a point estimate by estimating at the lowest levels of an arbitrarily deep WBS. This is not the case in real life. If the assumption of total dependence is made then the widest distribution function occurs. It has one and only one width no matter how many samples are taken. This also is not the case in real life. In real life, correlation exists between selected pairs of variables. The focus of this paper is to examine key aspects of simulating this case.

CORRELATION AND GEOMETRY

Correlation is largely perceived to be a statistical phenomena. It is. But it is also a geometric phenomena (Herr, 1980). To see this, we must first view the data in vector form (Halmos, 1974). Let x_i be the vector

$$x_i = \begin{bmatrix} p_{i1} \\ . \\ . \\ . \\ p_{in} \end{bmatrix}$$

where n is the number of data points selected for the parameter p_i. This may be viewed as a point in n-dimensional space (Kendall, 1961). Each dimension represents the particular instance of selecting a value p_{ij} for parameter p_i. The transpose of the vector x_i is denoted by

$$x_i' = \begin{bmatrix} p_{i1} \cdots p_{in} \end{bmatrix}.$$

There are m vectors x_i of dimension n. Take the mean of the n data points in each vector. Margolis (1979) shows the mean to be the orthogonal projection of the data onto the n-vector (1, ... ,1). This is displayed in 2 dimensions in Figure 1, that is, with 2 data points for the parameter p.

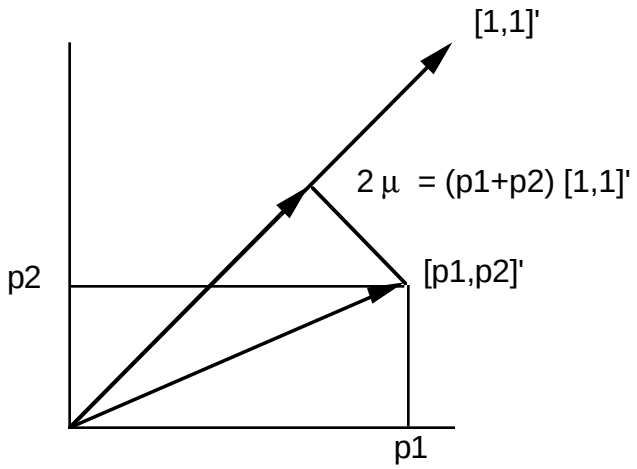

Figure 1: The Mean as an Orthogonal Projection

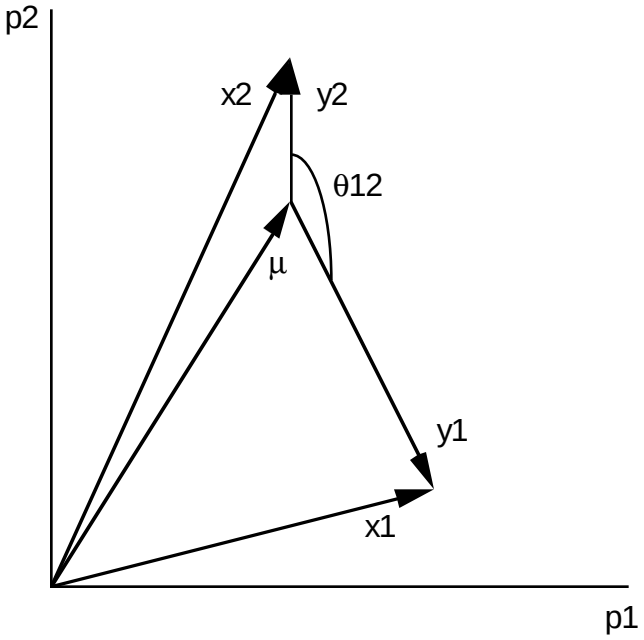

Figure 2: Data Vectors Adjusted for the Mean.

Denote the mean vector by $\mu = [\mu_1, \ldots, \mu_3]'$ where μ_i is the mean of the ith parameter p_i. Let the mean vector be the tensor operating point (O'Neill, 1966) and translate the space by the coordinate function $y_{ij} = x_{ij} - \mu_j$ so that the mean becomes the origin. We now have vectors y_1, y_2, and y_3 which originate from the new origin at μ. This is shown in two dimensions in Figure 2.

Note that the squared length $|y_i|^2 = y_{i1}^2 + \ldots + y_{in}^2$ of the vector y_i is the sum of squares of parameter i adjusted for the mean. Note further that the dot product of y_i and y_j is $y_i \bullet y_j = y_i' y_j = |y_i| |y_j| \cos \theta_{ij}$ where θ_{ij} is the angle between the vectors y_i and y_j. Consider the unit vectors $u_i = \frac{y_i}{|y_i|}$. Then $u_i \bullet u_j = \cos \theta_{ij}$.

Let $Y = [y_1, \ldots, y_3]$. Then the covariance matrix (Kendall, 1961) is

$$\Phi = Y' Y = \begin{bmatrix} y_1 \\ \cdots \\ y_m \end{bmatrix} \begin{bmatrix} y_1 & \cdots & y_m \end{bmatrix} = \begin{bmatrix} y_1 \bullet y_1 & \cdots & y_1 \bullet y_m \\ \cdots & \cdots & \cdots \\ y_m \bullet y_1 & \cdots & y_m \bullet y_m \end{bmatrix}$$

$$\Phi = \begin{bmatrix} |y_1||y_1|\cos\theta_{11} & \ldots & |y_1||y_m|\cos\theta_{1m} \\ \cdots & \cdots & \cdots \\ |y_m||y_1|\cos\theta_{m1} & \ldots & |y_m||y_m|\cos\theta_{mm} \end{bmatrix}$$

$$\Phi = \begin{bmatrix} |y_1| & 0 & 0 \\ 0 & \ldots & 0 \\ 0 & 0 & |y_m| \end{bmatrix} \begin{bmatrix} \cos\theta_{11} & \ldots & \cos\theta_{1m} \\ \cdots & \cdots & \cdots \\ \cos\theta_{m1} & \ldots & \cos\theta_{mm} \end{bmatrix} \begin{bmatrix} |y_1| & 0 & 0 \\ 0 & \ldots & 0 \\ 0 & 0 & |y_m| \end{bmatrix}$$

$$\Phi = M \Psi M$$

where

$$M = \begin{bmatrix} |y_1| & 0 & 0 \\ 0 & \ldots & 0 \\ 0 & 0 & |y_m| \end{bmatrix}$$

is the magnitude matrix and

$$\Psi = \begin{bmatrix} \cos\theta_{11} & \dots & \cos\theta_{1m} \\ \dots & \dots & \dots \\ \cos\theta_{m1} & \dots & \cos\theta_{mm} \end{bmatrix}$$

is the correlation matrix.

Since M is invertible, the correlation matrix can be found from the covariance matrix by

$$\Psi = M^{-1}\Phi M^{-1}$$

where

$$M^{-1} = \begin{bmatrix} \frac{1}{|y_1|} & 0 & 0 \\ 0 & \dots & 0 \\ 0 & 0 & \frac{1}{|y_m|} \end{bmatrix}$$

It is important to note that the covariance matrix Φ, in vector form, has the same definition as Einstein's fundamental metric tensor (Cartan, 1937) which completely determines the geometry of the data space (Einstein, 1916). Levi-Civita (1926) notes the relationship of the angles θ_{ij} of Ψ between the basis vectors and the fundamental metric tensor Φ. Although these concepts have existed for many years they have rarely been used, or even noted, by statisticians.

What does all this mean in terms of cost risk? It means that the geometric paradigm provides a way of both visualizing and implementing cost risk.

The three vectors $u_1, \dots u_m$ are orthogonal if and only if $\cos\theta_{ij} = 0$ for $i \neq j$. Thus correlated parameters have non orthogonal unit data vectors. A basis is a set of vectors which span a space or subspace such that none of them may be written as a linear combination of a subset of the other basis vectors (Halmos, 1974). This means that $U = [u_1, \dots , u_m]$ is a normal but non orthogonal basis for the data space. $Y = [y_1, \dots , y_m]$ is a non normal and non orthogonal basis for the data space. Such bases require a second tensor concept, the first being the establishment of the mean as the tensor operating point.

If w is any vector in our data space then $(w \bullet u_i)\, u_i$ is the projection of w on the unit basis vector u_i (Saville and Wood, 1991). In the terminology of tensors $(w \bullet u_i)\, u_i$ is the covariant projection of w on u_i and $w \bullet u_i = |w| \cos_{wi}$ is the covariant component (Pellionisz and Llinás, 1980). Unfortunately, as shown in Figure 3, the covariant components $u_{i\alpha}$ of non orthogonal vectors can not be used to obtain the vector sum.

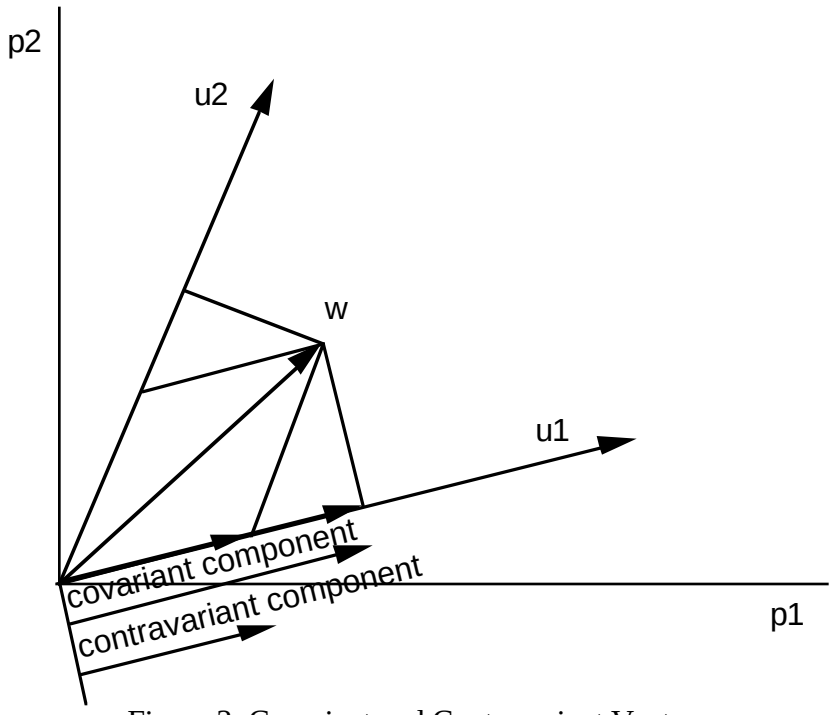

Figure 3: Covariant and Contravariant Vectors

The vector sum must use the contravariant components $u_i{}^{\beta}$. These are obtained by the transformation

$$u_i{}^{\beta} = \sum_{\alpha} \Psi^{\beta\alpha}\, u_{i\alpha}$$

where the $\Psi^{\beta\alpha}$ are the components of Ψ^{-1}. This is called "raising the index." Observe from Figure 3 that the covariant and contravariant components are identical when the basis vectors are orthogonal. This is an excellent reason for transforming the data to the basis defined by the principal components (Dean, 1988) which is an orthonormal basis. In that basis, the covariant and contravariant components are identical.

GENERATING CORRELATED RANDOM NUMBERS

In order to estimate cost risk, following Book and Young (1992), a set of random numbers p_i may be chosen from arbitrary distributions and then used to generate the estimated distribution of cost with desired correlation which defines the cost risk. It is desired that the variates p_i have the correlation Ψ and covariance Φ.

Choose the variates x independently from the desired distributions, and adjust them for the mean to obtain variates y from

$$y_{ij} = x_{ij} - \mu_j.$$

Form the covariance matrix Y' Y with magnitude M as the desired magnitude to obtain Φ. Note that

$$Y' Y = M \Psi_y M = M M$$

since the variates were chosen independently. Following Fukunaga (1990) we transform the variates by

$$Z = Y M^{-1}$$

to obtain

$$Z' Z = M^{-1,} Y' Y M^{-1} = M^{-1} M M M^{-1} = I.$$

Thus the variates z are uncorrelated and have unit magnitude. Choosing new variates v defined by

$$V = Z \Psi^{1/2}$$

we have

$$V' V = (Z \Psi^{1/2})' Z \Psi^{1/2} = \Psi^{1/2} Z' Z \Psi^{1/2} = \Psi.$$

Finally, choosing new variates u defined by

$$U = V M = Z \Psi^{1/2} M = Y M^{-1} \Psi^{1/2} M$$

we have

$$\begin{aligned} U' U &= (Y M^{-1} \Psi^{1/2} M)' Y M^{-1} \Psi^{1/2} M \\ &= M \Psi^{1/2} M^{-1} Y' Y M^{-1} \Psi^{1/2} M \\ &= M \Psi^{1/2} M^{-1} M M M^{-1} \Psi^{1/2} M \\ &= M \Psi M = \Phi. \end{aligned}$$

Thus the variates u have the desired correlation and covariance. Adjusting for the mean by

$$w_{ij} = u_{ij} + \mu_j$$

we obtain the desired variates w_{ij}.

FINDING THE SQUARE ROOT OF THE CORRELATION MATRIX

The square root of a correlation matrix is not unique. The Choleski factorization can be used (Book and Young, 1992). Another technique used by the author is as follows:

Following Dean (1988), find the principle components (Overall and Klett, 1983; Press, Teukolsky, Vetterling, and Flannery, 1992) of the correlation matrix. Thus we have the eigenvector matrix Ω and the diagonal eigenvalue matrix Λ such that

$$\Psi \Omega = \Omega \Lambda \quad \text{and } \Omega \Omega' = \Omega' \Omega = I$$

where

$$\Omega = [\omega_1 \ldots \omega_m]$$

and

$$\Lambda = \begin{bmatrix} \lambda_1 & \ldots & 0 \\ \ldots & \ldots & \ldots \\ 0 & \ldots & \lambda_m \end{bmatrix}.$$

Thus

$$\Psi = \Omega \Lambda \Omega' = \Omega \Lambda^{1/2} \Lambda^{1/2} \Omega' = \Omega \Lambda^{1/2} \Lambda^{1/2\prime} \Omega' = (\Omega \Lambda^{1/2}) (\Omega \Lambda^{1/2})'$$

Letting

$$\Psi^{1/2} = (\Omega \Lambda^{1/2})' = \Lambda^{1/2} \Omega'$$

we have

$$\Psi = \Psi^{1/2\prime} \Psi^{1/2} = \Psi^{1/2} \Psi^{1/2}$$

as desired.

Thus the desired variates w_{ij} are

$$w_{ij} = u_{ij} + \mu_j$$

where

$$U = Y M^{-1} \Lambda^{1/2} \Omega' M.$$

OBSERVATIONS

The geometric viewpoint identifies the choice of a correlation matrix for the simulation of cost risk with the pairwise choice of data vectors corresponding to the parameters used to obtain cost risk. The correlation coefficient is the cosine of the angle between the data vectors after translation to an origin at the mean and normalization for magnitude. Thus correlation is equivalent to expressing the data in terms of a non orthogonal basis. To understand the many resulting phenomena requires the use of the tensor concept of raising the index to transform the measured and observed covariant components into contravariant components before vector addition can be applied.

The geometric viewpoint also demonstrates that correlation and covariance are geometric properties, as opposed to purely statistical properties, of the variates. Thus, variates from different distributions may be correlated, as desired, after selection from independent distributions.

By determining the principal components of the correlation matrix, variates with the desired mean, magnitude, and correlation can be generated through linear transforms which include the eigenvalues and the eigenvectors of the correlation matrix.

The conversion of the data to a non orthogonal basis uses a compound linear transformation which distorts or stretches the data space. Hence, the correlated data does not have the same properties as the uncorrelated data used to generate it. This phenomena is responsible for seemingly strange observations such as the fact that the marginal distributions of the correlated data can be quite different from the distributions used to generate the data. The joint effect of statistical distributions and correlation remains a fertile area for further research.

In terms of application to cost estimating, the geometric approach demonstrates that the estimator must have data and must understand that data in order to properly choose the correlation matrix appropriate for a given estimate.

There is a general feeling by employers and managers that the field of cost requires little technical or mathematical background. Contrary to that opinion, this paper demonstrates that a background in mathematics equivalent to that needed for typical engineering and scientific disciplines at the masters or doctorate level is appropriate within the field of cost risk.

REFERENCES

Book, S. A. and P. H. Young (1992). "Applying Results of Technical-Risk Assessment to Generate a Statistical Distribution of Total System Cost," presented at the 1992 AIAA Aerospace Design Conference, Irvine CA, 3-6 February.

Cartan, E. (1937). The Theory of Spinors, The M. I. T. Press, Cambridge MA.

Dean, E. B. (1988). "Linear Least Squares for Correlated Data," Proceedings of the Tenth Annual Conference of the International Society of Parametric Analysts, Brighton, England, July 25-27.

Einstein, A. (1916). "Die Grundlage der allgemeinen Relativitätstheorie," Annalen der Physik, 49, translated as "The Foundation of the General Theory of Relativity," The Principle of Relativity, Dover Publications Inc., New York NY.

Fukunaga, K (1990). Introduction to Statistical Pattern Recognition, 2ed., Academic Press Inc., San Diego CA.

Halmos, P. R. (1974). Finite-Dimensional Vector Spaces, Springer-Verlag, New York NY.

Herr, D. G. (1980). "On the History of the Use of Geometry in the General Linear Model," The American Statistician, Feb., Vol. 34, No. 1, pp 43-47.

Kendall, M. G. (1961). A Course in the Geometry of n Dimensions, Hafner Publishing Company, New York NY.

Levi-Civita, T. (1926). The Absolute Differential Calculus, Dover Publications, New York NY, 1977.

Margolis, M. S. (1979). "Perpendicular Projections and Elementary Statistics", The American Statistician, Aug., Vol. 33, No. 3. pp. 131-135.

O'Neill, B. (1966). Elementary Differential Geometry, Academic Press, New York, NY.

Overall, J. E. and C. J. Klett (1983). Applied Multivariate Analysis, Robert E. Krieger Publishing Company, Malabar FL.

Pellionisz, A. and R. Llinás (1980). "Tensorial Approach to the Geometry of Brain Function: Cerebellar Coordination via a Metric Tensor," Neuroscience, Vol. 5, pp. 1125-1136.

Press, W. H., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery (1992). Numerical Recipes in C, Cambridge University Press, Cambridge, England.

Saville, D. J. and G. R. Wood (1991). Statistical Methods: A Geometric Approach, Springer-Verlag, New York NY.