Behavior Research Methods & Instrumentation
1980, Vol. 12(4), 479-480

COMOV: A program for the analysis of the relationship between two time series

R. L. RAY, H. H. EMURIAN, and R. M. WURSTER
Department of Psychiatry and Behavioral Sciences, Johns Hopkins University School of Medicine, Baltimore, Maryland 21218

One problem with the cross-correlation approach is that standard significance tests may not apply to estimates of cross-correlation (Bartlett, 1946; Holtzman, 1963). This situation arises when the repeated values of each measure are correlated among themselves or are autocorrelated. While techniques exist that filter or "whiten" the successive values so that ordinary significance tests apply, these techniques demand some knowledge of time-series analysis, as well as fairly extensive computations (Chatfield, 1975; Haugh & Box, 1977).

One simple alternative approach to the investigation of the covariation between two time series involves the use of the analysis of comovement (Goodman, 1963). Malstrom, Opton, and Lazarus (1965) used this type of analysis to investigate the covariation of heart rate and skin conductance over time in human subjects. In the analysis of comovement, each separate series of values is transformed into a series of 0s and 1s in the following fashion: If the series increases in magnitude from one sample period to the next, a 1 is recorded. If the series decreases in value from one sample period to the next, a 0 is recorded. This transformation operation is equivalent to first taking differences of the data (that is, subtracting each data point from its predecessor) and then recording a 0 when a difference is negatively signed and a 1 when a difference is positively signed. A 2 by 2 contingency table is then created from the N pairs of dichotomous scores. The cells are labeled A, B, D, and C in a clockwise fashion starting from the top left cell. The value entered into Cell A corresponds to the number of times both series were simultaneously equal to 1, the value entered in Cell B corresponds to the number of times Series 1 was equal to 1 and Series 2 was equal to 0, the value entered in Cell C corresponds to the number of times Series 1 was equal to 0 and Series 2 was equal to 1, and so on. Moore and Wallis (1943) have shown that, in the special case in which the first differences of the series of values are randomly distributed, the ordinary significance test for association in a 2 by 2 table is appropriate. This test involves computing:

$$A^* = (AD - CB)/N, \tag{1}$$

where N is the number of dichotomous pairs. The variance of A* is equal to:

$$S^2 = [(A + B)(A + C)(C + D)(B + D)]/N^3, \tag{2}$$

and for large samples A*/S is unit normal, or distributed as Z. If Z is large, this means that the two series tend to move in synchrony more often than would be expected by chance. In many cases, however, the assumption that the first differences are random is not justified. For example, if a series of raw scores is random, each successive first difference of the series is negatively related to its neighboring value. This result, a proof of which is provided by Goodman (1963), is due to the fact that each successive first difference of a series of values contains a term in common, one term being positively signed and one term being negatively signed. The negative correlation between successive pairs of differences that is produced by the differencing operation results in a negative correlation between neighboring values of the series when transformed into 0 and 1 scores. A general solution for this case (Goodman, 1963) is important, since many biological time series are very nonrandom and exhibit correlation both between immediately successive values and between values separated by one or more intervening members of the series. This extreme nonrandomness is often seen in the transformed series, albeit in an altered form. Goodman (1963) considers the situation in which each series of transformed values is M-dependent.

Given a series of values $X_1, X_2, X_3, \ldots X_n$, the series is M-dependent when any two values X_T and X_{T+J} are independent, given that J is greater than M. In other words, values separated by a lag greater than M are independent. Goodman (1963) showed that, in this case, the variance estimate for A* could be modified as follows:

$$S^{*2} = S^2 + \sum_{I=1}^{M} 2N\theta_I E_I, \tag{3}$$

where

$$\theta_I = G_I - [(A + B)^2/N^2] \tag{4}$$

and

$$E_I = H_I - [(A + C)^2/N^2]. \tag{5}$$

The values of G_I and H_I are obtained from the transformed series in the following fashion. If we call the two transformed series U_T and V_T, then θ_I is equal to the number of times that U_T and U_{T+I} are both equal to 1, and H_I is equal to the number of times that V_T and V_{T+I} are both equal to 1. For example, given the series of transformed values U_1, U_2, U_3, U_4, and U_5, G_I would be calculated in the following fashion: First, all possible pairs U_T, U_{T+I} are formed. These are U_1 and U_2, U_2 and U_3, U_3 and U_4, and U_4 and U_5. The number of these pairs of scores for which both elements are equal to 1 is the value of G_I.

0005-7878/80/040479-02$00.45/0

The modified variance estimate given by Formula 3 is the standard variance estimate given by Formula 2 plus a term proportional to the sum of the product of the two series' autocovariances. The autocovariances are estimated by Formulas 4 and 5 and refer to the covariance within a series between successive data points. In this case, M covariances are estimated for each series in which the I^{th} covariance is the covariance between values separated by $I-1$ intervening data points and I ranges from 1 to M. The choice of the proper value of M for use in Formula 3 can be approached in several different ways. Goodman (1963) recommends that Formula 3 be used with some value of $M \ll N$ that is as large or larger than the actual degree of dependence that exists in the series. An alternative approach to assigning an arbitrary value to M is to correct the variance estimate for successively larger values of M. When a value of M is found beyond which little change in the variance estimate is found, that value of M is used in the final analysis.

Description. The COMOV program is available in either a PDP-8 OS78 BASIC version or a FORTRAN IV batch version. The BASIC program accepts a data file as input that consists of a string of values separated by commas or carriage returns. The data are arranged in the following order: N, the series length; M, the largest value of M the user will examine; and then successive pairs of values from the two series, $X_1, Y_1, X_2, Y_2, \ldots$, and so on. The FORTRAN IV program accepts data from data cards in the same order as the BASIC program, but the data cards are preceded by a format card that supplies a variable-format statement enclosed by parentheses. After reading the data, the programs transform the data values into series of dichotomous values and compute and print A*, S^2, the Z value associated with A*, and the correlation coefficient (a phi coefficient) between the two series. The variance estimate is then corrected for successively larger values of M, and the corrected variance estimate and Z value are printed for each successive value of M.

Restrictions. The program ignores tied values, and series with more than a few tied values should not be analyzed in this fashion. The decision as to the actual number of tied values that are admissible should be determined by the individual user's decision as to what percent of the data he or she is prepared to ignore. Also, since the significance test requires the use of fairly long series, the series length should probably be equal to at least 35 or more observations.

Availability. A source listing, a user's handout, and sample input can be obtained free of charge from R. L. Ray, 614 Traylor, 720 Rutland Avenue, Baltimore, Maryland 21205.

REFERENCES

BARTLETT, M. S. On the theoretical specification of sampling properties of autocorrelated time-series. *Journal of the Royal Statistical Society Series B*, 1946, **8**, 27-36.

CHATFIELD, C. *The analysis of time series: Theory and practice.* London: Chapman & Hall, 1975.

GOODMAN, L. Tests based upon the movements in and the comovements between M-dependent time series. In C. Christ et al. (Eds.), *Measurement in economics.* Stanford, Calif: Stanford University Press, 1963.

HAUGH, L., & BOX, G. E. P. Identification of dynamic regression (distributed lag) models connecting two time-series. *Journal of the American Statistical Association*, 1977, **72**, 121-130.

HOLTZMAN, W. Statistical models for the study of change in the single case. In C. W. Harris (Ed.), *Problems in measuring change.* Madison: University of Wisconsin Press, 1963.

MALSTROM, E. J., OPTON, E., & LAZARUS, R. J. Heart rate measurement and the correlation of indices of arousal. *Psychosomatic Medicine*, 1965, **27**, 546-556.

MOORE, G. H., & WALLIS, W. A. Time series significance tests based on the signs of differences. *Journal of the American Statistical Association*, 1943, **43**, 153-164.

(Accepted for publication July 29, 1980.)