CORRELATION IN CONDITIONAL DISTRIBUTIONS:
SOME SUGGESTIONS FOR MEASURING THE RELATIONSHIPS
BETWEEN TWO SETS OF POLLUTANTS WHEN THE EFFECT OF
WEATHER HAS BEEN REMOVED

by

JUNE JURITZ¹

University of Capetown
and Stanford University

WORKING PAPER #3

This draft is intended for limited circulation

October 11, 1976

NOTICE

This report was prepared as an account of work sponsored by the United States Government. Neither the United States nor the United States Energy Research and Development Administration, nor any of their employees, nor any of their contractors, subcontractors, or their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness or usefulness of any information, apparatus, product or process disclosed, or represents that its use would not infringe privately owned rights.

There is no objection from the patent point of view to the publication or dissemination of the documents listed in this letter.

BROOKHAVEN PATENT GROUP

12/8 19 76 By WPC

¹This research was supported by Grant from SIMS

MASTER

Prepared For
THE U.S. ENERGY RESEARCH AND DEVELOPMENT ADMINISTRATION
UNDER CONTRACT NO. EY-76-S-02-2874.*000

DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency Thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.

Correlation in Conditional Distributions:

Some suggestions for measuring the relationships between two sets of pollutants when the effect of weather has been removed.

June Juritz

Introduction

In this note we review some multivariate statistical tests which could be used to measure relationships between 2 sets of measurements on air pollutants when both sets are effected by the weather. These tests are exact if the variables are normally distributed and if the observations are independent. Both these assumptions are usually violated with air pollution data sets, but we have selected tests which are robust and conservative. Moreover they are easily carried out using standard statistical packages and the results may be used to indicate if it is worth proceeding or developing more delicate procedures.

Let us assume for times $t=1,2,\dots,N$ we have measurements on a number of pollutants and weather conditions. Let the $p \times 1$ vector X_t represent the measurements taken at time t and suppose that X_t can be partitioned naturally into three subvectors $X_t^{(1)}$, $X_t^{(2)}$ and $X_t^{(3)}$ where $X_t^{(1)}$ is $q \times 1$, $X_t^{(2)}$ is $r \times 1$ and $X_t^{(3)}$ is $s \times 1$. To fix ideas let $X^{(1)}$ represent measurements on q particulate pollutants, $X^{(2)}$ measurements on r gaseous pollutants and $X^{(3)}$ measurements on s weather variables and suppose that we wish to measure the relationship between gaseous and particulate pollutants when the effect of the weather has been removed.

The "true" covariance structure between the three sets of variables can be written as, Anderson (1958) page 33,

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} & \Sigma_{13} \\ \Sigma_{21} & \Sigma_{22} & \Sigma_{23} \\ \Sigma_{31} & \Sigma_{32} & \Sigma_{33} \end{bmatrix} \quad (1)$$

and the conditional covariance matrix of $X^{(1)}$ and $X^{(2)}$ given $X^{(3)} = X^{(3)}$ is

$$\Sigma_3 = \begin{bmatrix} \Sigma_{11.3} & \Sigma_{12.3} \\ \Sigma_{21.3} & \Sigma_{22.3} \end{bmatrix}$$

where

$$\Sigma_{ij.3} = \Sigma_{ij} - \Sigma_{i3} \Sigma_{33}^{-1} \Sigma_{3j}, \quad i=1,2, \quad j=1,2. \quad (2)$$

From (1) and (2) we note that

$$\Sigma_{12} = \Sigma_{12.3} + \Sigma_{13} \Sigma_{33}^{-1} \Sigma_{32} \quad (3)$$

Now Σ_{12} measures the relationship between $X^{(1)}$ and $X^{(2)}$ and from (3) we see that this can be attributed to two sources measured by $\Sigma_{12.3}$ and $\Sigma_{13} \Sigma_{33}^{-1} \Sigma_{32}$ respectively, $\Sigma_{13} \Sigma_{33}^{-1} \Sigma_{32}$ represents the part of Σ_{12} that can be attributed to the weather and $\Sigma_{12.3}$ is the part of Σ_{12} that is unexplained by the weather. Thus it seems reasonable to conclude that when investigating any intrinsic relationship between $X^{(1)}$ and $X^{(2)}$ we should confine our attention to $\Sigma_{12.3}$ and test the hypothesis

$$H_0: \Sigma_{12.3} = 0.$$

Further we would like a measure of the relationship between the conditioned variables $(X^{(1)}|X^{(2)})$ and $(X^{(2)}|X^{(3)})$ which would tell us how much of the variation in the set $X^{(1)}$ can be explained by $X^{(2)}$ when the effect of the weather variables has been removed.

The most familiar measure of dependency in conditional distributions is of course the partial correlation coefficient which measures the dependency between two variables when the effect of a third has been removed (i.e. $q=r=s=1$, in our case). Other measures which have been defined are:

(i) the partial multiple correlation coefficient which measures the relationship between a single variable X_1 and a set of variables $X^{(2)}$ when the effect of another set $X^{(3)}$ has been removed.

(ii) The partial trace correlation coefficient, (Hooper, 1962).

(iii) Canonical partial correlation coefficients (Roy and Whittlesey, 1952).

We shall be concerned with (i) and (ii) and will briefly outline methods for testing hypotheses using these measures, which are exact under normal theory.

Let X_t , $t=1,2,\dots,N$ be a set of measurements taken on the pollution and weather data. We shall assume that, if necessary, suitable transformations have been made upon the raw data so that X_t can be regarded as being approximately normally distributed. For example, for some variables it may happen that the logarithms of the original observations are more nearly normally distributed. Each observation vector is partitioned into three sets $X_t^{(1)}$, $X_t^{(2)}$ and $X_t^{(3)}$ with

q, r and s components respectively with $q+r+s=p$ where p is the total number of variables measured at time t. We consider first the case where $X_t^{(1)}$ has a single component ($q=1$) and secondly the case where $X^{(1)}$ has several components ($q > 1$). In the first case the relationship between $X^{(1)}$ and $X^{(2)}$ will be measured by the partial multiple correlation coefficient, and in the second case by the partial trace correlation.

The Partial Multiple Correlation Coefficient.

The partial multiple correlation coefficient is equivalent to the more familiar multiple correlation coefficient which plays a large role in regression theory except that the partial multiple correlation measures a linear relationship in a conditional distribution. Since under normal theory, the conditional variables are again normally distributed, all tests and procedures using the multiple correlation coefficient apply equally to the partial multiple correlation coefficient with appropriate adjustment of the degrees of freedom.

The sampling distribution of the partial multiple correlation between a single variable and r "explanatory" variables based on a sample of size N is the same as the sampling distribution of the ordinary multiple correlation with N replaced by $N-s$ where s is the number of variables held fixed.

In the present application let $\bar{R}_{1 \cdot r | s}^2$ be the population partial multiple correlation between the single particulate pollutant X_1 and

the r gaseous pollutants when the effects of the s weather variables have been removed. Independence between the two sets is tested by

$$H_0: \bar{R}_{1.r|s}^2 = 0.$$

In testing hypotheses and estimating $\bar{R}_{1.r|s}^2$ it is not necessary to find the conditional distribution. Using any multiple regression program which allows selection of variables, we calculate two ordinary multiple correlation coefficients, $R_{1.r+s}^2$ and $R_{1.s}^2$, where

$R_{1.r+s}^2$ = the multiple correlation between X_1 and the entire set of gaseous pollutants, $X^{(2)}$, and weather variables, $X^{(3)}$.

$R_{1.s}^2$ = the multiple correlation between X_1 and the weather variables, $X^{(3)}$.

Then the partial multiple correlation coefficient is given by

$$R_{1.r|s}^2 = \frac{R_{1.r+s}^2 - R_{1.s}^2}{1 - R_{1.s}^2}. \quad (4)$$

$P = 100R_{1.r|s}^2$ % gives an estimate of the percentage of the variance of X_1 that can be explained by the set $X^{(2)}$

$$H_0: \bar{R}_{1.r|s}^2 = 0$$

can be tested using

$$\frac{R_{1.r|s}^2}{1 - R_{1.r|s}^2} \frac{N-s-r-1}{r} \quad (5)$$

which under H_0 has an F distribution with r and $N-s-r-1$ degrees of freedom. H_0 is rejected at level α if the upper α percentage point of the F distribution is exceeded.

It is easily shown that (5) can also be written as

$$\frac{R_{1.r+s}^2 - R_{1.s}^2}{1 - R_{1.r+s}^2} \cdot \frac{N-s-r-1}{r} \quad (6)$$

Note that if we had ignored the weather set $X^{(3)}$, our test statistic would have been

$$\frac{R_{1.r}^2}{1 - R_{1.r}^2} \cdot \frac{N-r-1}{r}, \quad (7)$$

which has an F distribution with r and $N-r-1$ degrees of freedom under independence.

To test the hypothesis

$$H_0: R_{1.r|s}^2 \leq \rho$$

where ρ is a specified value of the partial multiple correlation coefficient, we can use tables of the percentage points of the multiple correlation coefficient. Lee (1975). These are tabulated in terms of ρ , n_1 and $v = 60/n_2^{1/2}$ where $n_2 = N - n_1$. In the present context the tables would be used with $n_1 = r$, $n_2 = N - s - r$. These tables can also be used to construct confidence intervals for $R_{1.r|s}^2$.

The Partial Trace Correlation Coefficient

If there is more than one variable in the set $X^{(1)}$, then we cannot use the partial multiple correlation coefficient to measure conditional independence. Of the measures of correlation between sets of conditioned variables which have been defined (Juritz, 1971) we select the trace correlation coefficient because of its robustness against non-normality Olson (1944) and its easy interpretation.

Given a set of N measurements on X we calculate the Wishart matrix. (Anderson, 1958). Partitioning A according to the partitioning of Σ (see equation (1)) we have

$$A = \begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{bmatrix} \quad (8)$$

and the conditional Wishart matrix is

$$A_3 = \begin{bmatrix} A_{11.3} & A_{12.3} \\ A_{21.3} & A_{22.3} \end{bmatrix} \quad (9)$$

where

$$A_{ij} = A_{ij} - A_{i3} A_{33}^{-1} A_{j3} \quad i=1,2 \quad j=1,2$$

We then find

$$R_3 = A_{11.3}^{-1} A_{12.3} A_{22.3}^{-1} A_{21.3} \quad (10)$$

If $X^{(1)}$ has q components, the partial trace correlation (Hooper, 1962) is defined as

$$\tilde{r}^2 = \frac{1}{q} \text{tr } R_3 . \quad (11)$$

The trace of a matrix is the sum of its characteristic roots which is also equal to the sum of its diagonal elements. The characteristic roots of R_3 are the sample partial canonical correlations. Thus the partial trace correlation is the average of the canonical correlations. The partial trace correlation (and also the ordinary trace correlation) has all the desirable properties of an ordinary correlation coefficient in that it lies between 0 and 1, it is independent of the units in which the original variables are measured and $\tilde{r}^2$ gives a measure of the proportion of variation in the $X^{(1)}$ set that can be explained by the set $X^{(2)}$ when the effect of $X^{(3)}$ has been removed. Moreover it is the most robust of the tests of independence under normal theory. (Olson, 1974)

If the sets $X^{(1)}$ and $X^{(2)}$ are conditionally independent then the population trace correlation $\tilde{r}^2$ is zero. Limited tabulations percentage points of the distribution of $\tilde{r}^2$ under the hypothesis of independence are given in Pillai and Mijares (1959): These can be used for the partial trace correlation with the parameter involving the sample size, N say replaced by $N-s$, where s is the number of variables in $X^{(3)}$.

For large values of N ,

$$\frac{\tilde{r}^2}{1-\tilde{r}^2} \sim \frac{N-s-r-1}{r} \quad (12)$$

has approximately an F distribution with $q(N-r-s-1)$ and qr degrees of freedom if the two sets of variables are independent. If there is only one variable in the set $X^{(1)}$, the partial trace correlation is the same as the partial multiple correlation and the test statistic given in (12) is the same as (5).

References

1. Anderson, T.W. An Introduction to Multivariate Statistical Analysis. Wiley, New York. (1958).
2. Hooper, J.W. Partial trace correlations. *Econometrica* (1962) 30 324-331.
3. Juritz, J.M. The generalized partial correlation matrix. *South African Statistical Journal* (1971) 5 1-4.
4. Lee, Y.S. Tables of upper percentage points of the multiple correlation coefficient. *Biometrika* (1972) 59 175-189.
5. Mijares, T.A. Percentage points of the sum $V_1^{(s)}$ of s roots ($s=1-50$) Manila: Statistical Centre, University of the Philipines (1964).
6. Olson, C.L. Comparative robustness of six tests in multivariate analysis of variance. *J. Amer. Statist. Assoc.* 69 894-908. (1974)
7. Pillai, K.S. and Mijares, T.A. On the moments of the trace of a matrix and approximations to its distribution. *Annals of Mathematical Statistics* 30 1135-1140. (1959)