TESTING A SET OF CORRELATION MATRICES FOR EQUALITY

Norman B. Rushforth, Division of Biometry and Developmental
Biology Center, Western Reserve University

1. Introduction

A similar type of problem to that of testing the equality of correlation matrices exists when covariance matrices are considered. The likelihood ratio test for testing the equality of covariance matrices was given by Wilks (1946). If population covariance matrices are shown to be equal then the population correlation matrices must be equal. The converse is not necessarily true, however. In many psychological studies the variables are "standardized" by dividing covariances by corresponding standard deviation elements. This is often done when the underlying variables have no natural scale of measurement and the units are somewhat arbitrary.

The problem of testing the equality of a set of correlation coefficients from one population was recently solved by Lawley (1963). For the case of m correlation matrices each based on two variables a test procedure exists (Rao, 1952). This case, which is a problem of testing the homogeneity of a set of correlation coefficients, was solved by using the z transformation (Fisher, 1924). Brander (1933) derived the likelihood ratio statistic for testing the hypothesis of equal correlation matrices for the bivariate case with two populations. He showed the statistic to be a function of the difference between z transformed correlation coefficients. Efforts by the present author to generalize Brander's results to more than two populations and more than two variables were unsuccessful.

The purpose of this paper is to present two statistics B and M to test the equality of m correlation matrices based on p variables. Test procedures based on B and M are consistent. For the m population bivariate case (p = 2), assuming equal population correlations, a logarithmic function of B is asymptotically distributed as a chi square variable with m - 1 degrees of freedom. Thus B may be used to test the homogeneity of a set of correlation coefficients. For the two population (m = 2), trivariate case (p = 3) with equal sample sizes, a similar function of B is distributed asymptotically as a quadratic form in dependent normal variables, but not as a chi-square variable (Rushforth, 1961). The complexity of this distribution led to the abandonment of B as a statistic for testing the equality of p x p correlation matrices. The statistic M is based on z transformed correlation coefficients and is analogous in structure to the likelihood ratio statistic for the multivariate analysis of variance. Under the hypothesis of equal population correlation matrices a logarithmic function of M is asymptotically chi-square with $\frac{(m-1)\,p\,(p-1)}{2}$ degrees of freedom.

2. Properties of the B Statistic

The statistic B is defined as:

$$B^{\frac{1}{N}} = \frac{\left| \sum_{j=1}^{m} k_j R_j \right|}{\prod_{j=1}^{m} \left| R_j \right|^{k_j}}$$

where R_j is the sample correlation matrix of the j^{th} population with sample size N_j. $R_j = (r_{ivj})$ i, v = 1...p, j = 1...m. The quantity $k_j = \frac{N_j}{N}$, $N = \sum_{j=1}^{m} N_j$, and therefore $\sum_{j=1}^{m} k_j = 1$.

Thus $B^{1/N}$ is seen to be a ratio of determinants. The numerator is the determinant of the average sample correlation matrix for the m populations, each correlation matrix being weighted by k_j, the proportion of the total sample size attributed to the j^{th} population. The denominator is a product of the individual determinants of the sample correlation matrices each raised to the appropriate k_j^{th} power.

It can be shown (Rushforth, 1961) that the test procedure based on rejecting the hypothesis $P_j = P$, j = 1...m, in favor of the alternative hypothesis $P_j \neq P_w$, j = 1...m, w = 1...m, for some $j \neq w$ when $B^{1/N}$ is too large is consistent. Here $P_j = (\rho_{ivj})$ is the positive definite population correlation matrix with sample estimate $R_j = (r_{ivj})$ and P the common but unknown value of the matrices under the null hypothesis.

For the case of m bivariate populations the statistic $B^{1/N}$ is equal to $\left[1 - (\Sigma k_j r_j)^2\right]\left[\prod_{j=1}^{m} (1 - r_j^2)^{k_j}\right]^{-1}$. Here the null hypothesis $P_j = P$ is equivalent to $\rho_j = \rho$, j = 1...m. It can be shown (Rushforth, 1961) that under the null hypothesis ln B =

$$(1 + \rho^2)\left[\sum_{j=1}^{m} Y_j^2 - \left(\sum_{j=1}^{m} \sqrt{k_j}\, Y_j\right)^2 + No\,(1/N)\right]$$

where Y_i is distributed asymptotically as N(0, 1) and $\lim_{N \to \infty} No\,(1/N) = 0$.

Since $\left[1 + \left(\sum_{j=1}^{m} k_j r_j\right)^2\right]$ converges stochastically to $(1 + \rho^2)$ the quantity $\left[1 + (\Sigma k_j r_j)^2\right]^{-1} \ln B$ is asymptotically distributed as

$\left[\sum_{j=1}^{m} Y_j^2 - \left(\sum_{j=1}^{m} \sqrt{k_j} Y_j \right)^2 \right]$. This latter expression may be written as $\underline{Y}'A\underline{Y}$, a quadratic form in Y_j, where $\underline{Y}'$ is a m component vector of independent N (0, 1) variables. The m x m matrix A of rank m - 1 is equal to

$$\begin{bmatrix} 1 - k_1, & -\sqrt{k_1 k_2}, \ldots & -\sqrt{k_1 k_m} \\ -\sqrt{k_1 k_2}, & 1 - k_2, \ldots & -\sqrt{k_2 k_m} \\ & \ldots & 1 - k_m \end{bmatrix}$$

Since $\underline{Y}'$ is a m component vector with covariance matrix $\underline{I}$, the identity matrix, a necessary and sufficient condition for $\underline{Y}'A\underline{Y}$ to be distributed as a chi square variable with m - 1 degrees of freedom is $AA' = A$ where A' is the transpose of A. Since A satisfies this condition then $\left[1 + \left(\sum_{j=1}^{n} k_j r_j \right)^2 \right]^{-1} \ln B$ is asymptotically distributed as a chi square variable with m - 1 degrees of freedom.

3. Properties of the M Statistic

The statistic M is based on z transformed correlation coefficients and is analogous in structure to the likelihood ratio statistic for the multivariate analysis of variance. The Fisher z transformations for a sample correlation coefficient r and population correlation coefficient ρ are defined as

$$z = \frac{1}{2} \ln \frac{(1 + r)}{(1 - r)} \quad \text{and}$$

$$\zeta = \frac{1}{2} \ln \frac{(1 + \rho)}{(1 - \rho)} .$$

Let $x_1 \ldots x_p$ be variables with the same first four moments as a standard normal variable. Let $r_{12}, r_{13} \ldots r_{1p}, \ldots, r_{p-1,p}$ and $\rho_{12}, \rho_{13} \ldots \rho_{1p}, \ldots, \rho_{p-1,p}$ be the sample and population correlation coefficients respectively for these variables. Define $z_{12}, z_{13} \ldots z_{1p}, \ldots, z_{p-1p}$ and $\zeta_{12}, \zeta_{13} \ldots \zeta_{1p}, \zeta_{p-1p}$ as the corresponding Fisher transforms of these correlation coefficients. Consider m populations defined by the parameters $\underline{\zeta}_j' = (\zeta_{12j}, \zeta_{13j}, \ldots \zeta_{1pj}, \ldots \zeta_{p-1pj})$. A test of the equality of m correlation matrices is equivalent to a test of the equality of the m vectors $\underline{\zeta}_j'$, j=1 ... m. For p underlying variables, there exists a p x p correlation matrix for each of the m populations. Since the matrix is symmetric and entries on the main diagonal are unity, $\frac{p(p-1)}{2}$ different correlational elements appear in the matrix. If Fisher transformations are made of the individual entries in the correlation matrix a $\frac{p(p-1)}{2}$ component vector $\underline{\zeta}_j'$ is obtained.

Equal population correlation matrices will give equal vectors. Thus a test of the equality of $\underline{\zeta}_j'$ vectors is equivalent to a test on the corresponding correlation matrices.

Consider the random vector $\sqrt{N}(z_{12} - \zeta_{12}, z_{13} - \zeta_{13} \ldots z_{1p} - \zeta_{1p}, \ldots, z_{p-1p} - \zeta_{p-1p}$. It can be shown (Rushforth, 1961) that $\sqrt{N}(\underline{z} - \underline{\zeta})'$ is asymptotically a multivariate normal distribution with mean vector $\underline{0}$ and covariance matrix Σ.

$$\Sigma = (u_{ijkl}) \; i, j, k, l, = 1 \ldots p, \; i < j, \; k < l,$$

where $u_{ijkl} = NE_L (z_{ij} - \zeta_{ij})(z_{kl} - \zeta_{kl}) =$

$$NE_L \left(\frac{(r_{ij} - \rho_{ij})(r_{kl} - \rho_{kl})}{(1 - \rho_{ij}^2)(1 - \rho_{kl}^2)} + o(1/N) \right)$$

The operator E_L indicates the moment of the appropriate entry of the covariance matrix of the limiting distribution. Now

$$\Sigma = NE_L(z_{ij} - \zeta_{ij})(z_{kl} - \zeta_{kl}) =$$

$$\frac{1}{2(1 - \rho_{ij}^2)(1 - \rho_{kl}^2)}$$

$$\Big((\rho_{jk} - \rho_{ij}\rho_{ik})(\rho_{il} - \rho_{ik}\rho_{lk})$$

$$+ (\rho_{il} - \rho_{ij}\rho_{jl})(\rho_{jk} - \rho_{jl}\rho_{kl})$$

$$+ (\rho_{jl} - \rho_{il}\rho_{ij})(\rho_{ik} - \rho_{il}\rho_{kl})$$

$$+ (\rho_{jl} - \rho_{jk}\rho_{kl})(\rho_{ik} - \rho_{ij}\rho_{jk}) \Big)$$

For the case k = i, l = j the above expression gives rise to the well known result (Hotelling, 1953)

$$NE_L(z_{ij} - \zeta_{ij})^2 = 1$$

Thus $\sqrt{N}(\underline{z} - \underline{\zeta})'$ is asymptotically normal with mean vector $\underline{0}$ and covariance matrix Σ, whose entries are given by $NE_L(z_{ij} - \zeta_{ij})(z_{kl} - \zeta_{kl})$.

If Σ^{-1} exists then $N(\underline{z} - \underline{\zeta})' \Sigma^{-1} (\underline{z} - \underline{\zeta})$ is asymptotically distributed as a chi square variable with $\frac{p(p-1)}{2}$ degrees of freedom.

Since Σ is the covariance matrix of the limiting distribution of $\sqrt{N}(\underline{z} - \underline{\zeta})'$ whose entries are

functions of the population correlation coefficients, then $\hat{\Sigma}$ obtained by substituting the maximum likelihood estimates (r_{ij}) for (ρ_{ij}) is the maximum likelihood estimate of Σ and $\hat{\Sigma}^{-1}$ is the maximum likelihood estimate of Σ^{-1}. Now $\hat{\Sigma}^{-1}$ is a consistent estimate of Σ^{-1}. Therefore $N(\underline{z} - \underline{\zeta})' \hat{\Sigma}^{-1} (\underline{z} - \underline{\zeta})$ is asymptotically distributed as a chi square variable with $\frac{p(p-1)}{2}$ degrees of freedom.

For m populations, $\sqrt{N_t}\,\underline{z}'_t$, $t = 1 \ldots m$ are asymptotically $N(\underline{\zeta}'_t, \Sigma_t)$ under the alternative $\underline{\zeta}_u \neq \underline{\zeta}_v$ for some $u \neq v$. The statistic M used to test $\underline{\zeta}'_t = \underline{\zeta}'$ is defined as:

$$M = \frac{\left| \hat{\Sigma}_c \right|}{\left| \hat{\Sigma}_c + \frac{B}{N} \right|} \quad \text{where } B = (b_{ij}) =$$

$$\left[\sum_{t=1}^{m} N_t (z_{it} - \bar{z}_i)(z_{jt} - \bar{z}_j) \right] \quad \text{and}$$

$$N = \sum_{t=1}^{m} N_t$$

$$\hat{\Sigma}_c = \frac{\sum_{t=1}^{m} N_t \hat{\Sigma}_t}{\sum_{t=1}^{m} N_t} \quad \text{is the sample estimate of}$$

$$\Sigma_c = \frac{\sum_{t=1}^{m} N_t \Sigma_t}{\sum_{t=1}^{m} N_t}$$

Under the null hypothesis $\underline{\zeta}'_t = \underline{\zeta}'$ $t=1 \ldots m$ the quantity $-N \ln M$ is asymptotically a chi square variable with $1/2\,(m-1)p(p-1)$ degrees of freedom. Under the alternative hypothesis $-N \ln M$ is asymptotically distributed as a linear function of non-central chi-square variables. This distribution is difficult however, and was not investigated.

It can be shown, (Rushforth, 1961) that a test procedure based on rejecting $\underline{\zeta}'_t = \underline{\zeta}'$ $t=1 \ldots m$ when M is too small in favor of $\underline{\zeta}'_u \neq \underline{\zeta}'_v$ for some $u \neq v$ $u, v = 1 \ldots m$ is consistent.

4. Application of the Statistics B and M

The statistic $\left[\ln B\right]\left[1 + \left(\sum_{j=1}^{m} k_j r_j\right)^2\right]^{-1}$ is used to test the homogeneity of a set of correlation coefficients derived from measurements in a training evaluation study (Rushforth, 1958). In this study pre- and post- training measurements were made on 15 students attending a course in conference leadership.

The table below gives correlation coefficients between the after-training Q-sorts (Stephenson, 1953) of the 15 students and their corresponding before-training Q-sorts. In this situation a test of the homogeneity of correlation coefficients has practical meaning in terms of evaluating the training course. It is of interest to determine whether or not the group was uniform with respect to the pre- and post-training measurements. Are members of the group affected to the same extent by training or do some individuals change much more than others?

Table 1. Correlation Coefficients of Student's Before and After Training Q-sorts
($N_j = 100$, $j = 1 \ldots 15$, $m = 15$)

Student (j)	Correlation Coefficient (r_j)
1	.756
2	.600
3	.770
4	.616
5	.653
6	.786
7	.762
8	.733
9	.738
10	.616
11	.751
12	.754
13	.624
14	.727
15	.733

Under the null hypothesis that the 15 correlation coefficients are from the same population $(\ln B)\left[1 + \left(\sum_{j=1}^{15} k_j r_j\right)^2\right]^{-1}$ is approximately distributed as a chi square variable with 14 degrees of freedom. A convenient method of computing the statistic is first to compute $B^{1/N}$, from which it is easy to determine $\left[N \ln B^{1/N}\right]\left[1 + \left(\sum_{j=1}^{15} k_j r_j\right)^2\right]^{-1}$. Since each of the 15 correlation coefficients r_j was based on a sample size of 100, $N = 1500$ and $k_j = 1/15$ $j = 1 \ldots 15$. In this situation

$$B^{1/N} = \left[1 - (1/15 \sum_{j=1}^{15} r_j)^2 \right] \left[\prod_{j=1}^{15} (1 - r_j^2) \right]^{-1/15}$$

The statistic $\left[1 + (1/15 \sum_{j=1}^{15} r_j)^2 \right]^{-1} \left[N \ln B^{1/N} \right]$ is 18.8, which is compared with 23.68, the 95th percentile value of a chi square distribution with 14 degrees of freedom. Thus the difference of the correlation coefficients between the before- and after- training card sorts was not significant at the 5 per cent level.

The statistic - N ln M is used to test the equality of three correlation matrices based on five variables. The data are taken from a study of the responses of industrial executives to a semantic differential questionnaire concerning job concepts (Miller, 1960). They consist of the sample correlation matrices of three rater groups assessing the jobs of executives in 5 specialities, using 25 semantic differential scales. The data are presented below in tables 2, 3, and 4.

In testing the hypothesis that these three correlation matrices are from the same population a lengthy computation is required in order to obtain the value of - N ln M. A computational form of this expression may be derived as follows.

$$M = \frac{\left| \hat{\Sigma}_c \right|}{\left| \hat{\Sigma}_c + B/N \right|} = \frac{\left| \hat{\Sigma}_c \right|}{\left| \Sigma_c \right| \left| I + \Sigma_c^{-1} B/N \right|} = \left| I + \hat{\Sigma}_c^{-1} B/N \right|^{-1}$$

Therefore $- N \ln M = N \ln \left(\left| I + \hat{\Sigma}_c^{-1} B/N \right| \right)$

Expanding $\left| I + \hat{\Sigma}_c^{-1} B/N \right|$ to order N^{-1}

$$- N \ln M = N \ln \left(1 + 1/N \sum_{i=1}^{q} \sum_{j=1}^{q} \hat{\sigma}^{ij} b_{ij} + o(1/N)\right) \quad i, j =$$

$1 \ldots q$ where $\hat{\sigma}^{ij}$ is the element of the i-th row and j-th column of $\hat{\Sigma}_c^{-1}$ and $q = \frac{p(p-1)}{2}$. Now $1/N \sum_{i=1}^{q} \sum_{j=1}^{q} \hat{\sigma}^{ij} b_{ij} + o(1/N)$ can be made arbitrarily small for sufficiently large N and therefore ln M may be expanded as a series to order N^{-1}. Thus

$$- N \ln M = \sum_{j=1}^{q} \sum_{j=1}^{q} \hat{\sigma}^{ij} b_{ij} + N\, o(1/N)$$

Now $\hat{\sigma}^{ij}$ is the element in the i-th row and j-th column of the inverse of the pooled covariance matrix $\hat{\Sigma}_c$. The value of $\hat{\Sigma}_c$ is an average of the individual sample covariance matrices. The symmetric matrix $B = (b_{ij})$ is the "between population" covariance matrix. The statistic - N ln M is equivalent therefore to the trace of the matrix obtained by pre-multiplying B by $\hat{\Sigma}_c^{-1}$. (i.e., the sum of the elements on the main diagonal).

As a first step in the test procedure, each correlation coefficient in the three matrices is transformed to a z variable. The individual sample covariance matrices $\hat{\Sigma}_t$ t=1, 2, 3 for the resulting z vectors are computed. The average covariance matrix $\hat{\Sigma}_c$ is computed from the sample matrices. Here it may be computed as an unweighted average since the sample sizes are approximately the same, $N_1 = 41$, $N_2 = 40$, $N_3 = 41$. The matrix $B = (b_{ij}) = \left[39 \sum_{t=1}^{3} (z_{it} - \bar{z}_i)(z_{jt} - \bar{z}_j) \right]$ is calculated from the z values for each matrix, $\bar{z}_i$ i = 1...q being the average z_i for the three samples. The value of 39 was approximately equal to $N_t - 3$, t = 1, 2, 3. It is used here since $(N_t - 3)^{-1}$ is a closer approximation to the variance of z_{it} than N_t^{-1} (Anderson, 1958). The inverse of $\hat{\Sigma}_c$ is post multiplied by B. Operations of determining the inverse of $\hat{\Sigma}_c$ and the subsequent multiplication by B are best effected by means of an electronic computer. The trace of the matrix resulting from this multiplication is found to be 35.1.

Thus the observed value of the statistic - N ln M is 35.1 compare with the critical value of 31.41 for the 95 percentile value of a chi square variable with 20 degrees of freedom. The hypothesis of equality of correlation matrices is rejected at the 5 per cent level of significance. Therefore no common correlation matrix is assumed to describe the semantic differential evaluations of the five job categories made by personnel, engineering and production executives.

5. Summary

To test the equality of m correlation matrices based on p variables, the two statistics B and M are proposed. A test procedure based on B is consistent. For the m population bivariate case (p = 2), assuming equal population correlations, a logarithmic function of B is asymptotically distributed as a chi square variable with m - 1 degrees of freedom.

The statistic M is based on z transformed correlation coefficients and is analogous in structure to the likelihood ratio statistic for the multivariate analysis of variance. Under the hypothesis of equal

population correlation matrices a logarithmic function of M is shown to be asymptotically chi-square with $\frac{(m-1)\,p\,(p-1)}{2}$ degrees of freedom. A test procedure based on M is consistent.

Application of the B statistic is illustrated in testing the homogeneity of a set of correlation coefficients from a training evaluation study. Use of the M statistic is demonstrated in testing the equality of correlation matrices derived in a study of the job concepts of selected industrial executive groups.

References

Anderson, T. W. (1958) An Introduction to Multivariate Statistical Analysis, New York: Wiley & Sons, 374 pp. (p. 78).

Brander, F. A. (1933) "A Test of the Significance of the Difference of the Correlation Coefficients in Normal Bivariate Samples", Biometrika: 25, pp. 102-109.

Fisher, R. A. (1924) "On a distribution yielding the error functions of several well known statistics", Proceedings of the International Mathematical Congress, Toronto, pp. 805-813.

Hotelling, H. (1953) "New Light on the Correlation Coefficient and its Transforms", Journ. Royal Stat. Soc. (Series B): 15, p. 193-232.

Lawley, D.N. (1963) "On Testing a Set of Correlation Coefficients for Equality", Ann. Math. Stat.: 34, 1. 149-151.

Miller, F. B. (1960) "Interim Report to the Social Science Research Center", Ithaca, N.Y.: Cornell University, 30 pp. (mimeographed).

Rao, C. R. (1952) Advanced Statistical Methods in Biometric Research, New York: Wiley & Sons, 390 pp. (p. 233).

Rushforth, N. B. (1958) "Evaluating Student Conference Leadership Training, A Study Utilizing Q-Technique". (M.S. Thesis, Cornell University, 1958) 131 pp.

Rushforth, N. B. (1961) "A Comparison of Sample Correlation Matrices and a Multivariate Analysis of Job Concepts of Selected Industrial Executive Groups" (Ph.D. Thesis, Cornell University, 1961) 140 pp.

Stephenson, W. (1953) The Study of Behavior; Q-Technique and its Methodology, Chicago: University of Chicago Press, 376 pp.

Wilks, S. S. (1946) "Sample Criteria for Testing Equality of Means, Equality of Variances, and Equality of Covariances in a Normal Multivariate Distribution", Ann. Math. Stat.: 17 pp. 257-281.

Table 2. Correlation Matrix of Scores for Engineering Executives (N_1 = 41) rating Five Job Categories

	Personnel	Sales	Accounting	Production	Engineering
Personnel	1.00	0.48	0.76	0.48	0.57
Sales		1.00	0.45	0.52	0.38
Accounting			1.00	0.50	0.31
Production				1.00	0.47
Engineering					1.00

Table 3. Correlation Matrix of Scores for Production Executives (N_2 = 40)

	Personnel	Sales	Accounting	Production	Engineering
Personnel	1.00	0.57	0.54	0.61	0.54
Sales		1.00	0.32	0.62	0.44
Accounting			1.00	0.35	0.49
Production				1.00	0.65
Engineering					1.00

Table 4. Correlation Matrix of Scores for Personnel Executives (N_3 = 41)

	Personnel	Sales	Accounting	Production	Engineering
Personnel	1.00	0.50	0.46	0.76	0.58
Sales		1.00	0.52	0.56	0.61
Accounting			1.00	0.51	0.50
Production				1.00	0.55
Engineering					1.00