ED 351 364 TM 019 180

AUTHOR Becker, Betsy Jane
TITLE Missing Data and the Synthesis of Correlation Matrices.
PUB DATE Apr 92
NOTE 39p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, April 20-24, 1992).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Bayesian Statistics; College Entrance Examinations; *Correlation; Equations (Mathematics); *Estimation (Mathematics); High Schools; High School Students; *Mathematical Models; Mathematics Tests; *Matrices; Meta Analysis; Sample Size; Scores; Spatial Ability; *Synthesis; Verbal Tests
IDENTIFIERS EM Algorithm; *Missing Data; Scholastic Aptitude Test

ABSTRACT

Analyses for results of a series of studies examining intercorrelations among a set of as many as p+1 variables are presented. Several estimators of a pooled or average correlation vector and its variances are derived for cases in which some studies do not report complete correlation matrices. A test of the homogeneity (consistency) of the correlation matrices is also given. Data from a synthesis of relationships among mathematical, verbal, and spatial ability measures illustrate the procedures. These data are taken from 10 samples (sample sizes 74, 153, 48, 55, 51, 18, 27, 43, 35, and 34, respectively) from 4 studies exploring the relationship of spatial ability to Scholastic Aptitude Test scores for high school or junior high school students. The empirical Bayes procedure (based on the EM algorithm) involves no data loss and is recommended if it is reasonable to assume that the unobserved correlations are missing at random. Three tables present illustrative data. (Author/SLD)

Missing Data and the Synthesis of Correlation Matrices

Betsy Jane Becker

Michigan State University

Paper presented at the annual meeting of the
American Educational Research Association
San Francisco, CA April, 1992

Keywords: Meta-analysis, correlations, missing data, imputation

Abstract

This paper outlines analyses for results of a series of studies examining intercorrelations among a set of as many as p+1 variables. Several estimators of a pooled or average correlation vector and its variance are derived for cases in which some studies do not report complete correlation matrices. A test of the homogeneity (consistency) of the correlation matrices is also given. Data from a synthesis of relationships among mathematical, verbal, and spatial ability measures illustrate the procedures. The empirical Bayes procedure (based on the EM algorithm) involves no data loss, and is recommended if it is reasonable to assume that the unobserved correlations are missing at random.

Missing Data and the Synthesis of Correlation Matrices

Many research syntheses which examine relationships in education and the social sciences examine one relationship or at most a few different bivariate relationships. In some research domains, however, series of studies may examine similar or identical collections or sets of variables. One example is the literature on the prediction of college grade-point average from entrance-examination scores and high-school records. In such cases it may be desirable to combine the correlation matrices among the variables common to a number of studies in order to draw general conclusions about the interrelationships among thc variables.

When all the studies under consideration share a common population matrix it is sensible to estimate a common (pooled) correlation matrix. In other situations it may be useful to estimate the average of a series of correlation matrices (and its variance). One problem which arises in attempting to pool or average correlation matrices from series of studies is that some studies may not have measured every variable of interest. Consequently some of the correlations of interest may not be observed in every study.

The first section below presents notation and a model for the results of a series of studies examining intercorrelations among a set of $p+1$ variables. Several estimators of a pooled correlation matrix and its variance are derived for the case in which correlations may be unobserved in some studies. Estimators based on available-data and complete-case analyses, and on imputation of both unconditional and conditional means are described and critiqued. An empirical Bayes estimator is also provided for the case in which a random-effects model is assumed to underly the series of studies. Data from a synthesis of relationships among mathematical, verbal, and spatial ability measures (Friedman, in press) are used to illustrate the procedures.

Notation and Model

Let $\underline{Y}_1, \ldots, \underline{Y}_p$ be random variables with the multivariate normal distribution. For example, $\underline{Y}_1$ may be an outcome and $\underline{Y}_2, \ldots, \underline{Y}_p$ may be $\underline{p}$-1 predictors. Consider the situation in which each of a series of $\underline{k}$ studies has examined correlations among these same $\underline{p}$ variables or a subset of those variables. The number of measured variables in the $\underline{i}$th study will be denoted $\underline{p}_i$ and the number of nonredundant correlations reported by study $\underline{i}$ is $\underline{m}_i = \underline{p}_i(\underline{p}_i-1)/2$.

Consider first a study $\underline{i}$ which has examined the intercorrelations among all $\underline{p}$ variables (i.e., in which $\underline{p}_i = \underline{p}$). Let $\underline{r}_{ist}$ and $\underline{\rho}_{ist}$ be the sample and population correlations between $\underline{Y}_s$ and $\underline{Y}_t$ in the $\underline{i}^{th}$ study and let $\mathbf{r}_i = (\underline{r}_{i12}, \underline{r}_{i13}, \ldots, \underline{r}_{i1p}, \underline{r}_{i23}, \ldots, \underline{r}_{i(p-1)p})'$ and $\underline{\rho}_i = (\rho_{i12}, \rho_{i13}, \ldots, \rho_{i1p}, \rho_{i23}, \ldots, \rho_{i(p-1)p})'$ be the vectors of $\underline{m}_i = \underline{p}(\underline{p}-1)/2 = \underline{p}^*$ nonredundant sample and population correlations, respectively. When it is convenient to refer to the elements of the vectors $\mathbf{r}_i$ and $\underline{\rho}_i$ by sequential position, a Greek subscript α or γ will be used (e.g., $\underline{r}_{i\alpha}$ and $\underline{r}_{i\gamma}$ are elements of $\mathbf{r}_i$). Thus $\mathbf{r}_i = (\underline{r}_{i\alpha})$ where α runs over the range $\alpha = 1, \ldots, \underline{p}^*$.

Distribution of r

Olkin and Siotani (1976) showed that if all $\underline{p}^*$ correlations have been observed in study $\underline{i}$, with a sample of size $\underline{n}_i$, the asymptotic distribution of $\sqrt{\underline{n}_i}\,(\mathbf{r}_i - \underline{\rho}_i)$ is normal with mean zero and variance-covariance matrix that depends on $\underline{\rho}_i$. This implies that in large samples, $\mathbf{r}_i$ is approximately normally distributed with mean vector $\underline{\rho}_i$ and variance-covariance matrix Σ_i, where the elements of Σ_i are defined by $\sigma_{i\alpha\gamma}$, and

$$\sigma_{i\alpha\alpha} = \mathrm{Var}(\underline{r}_{i\alpha}) = (1 - \rho_{i\alpha}^2)^2/\underline{n}_i \ , \qquad (1)$$

and

$$\sigma_{i\alpha\gamma} = \mathrm{Cov}(r_{i\alpha}, r_{i\gamma}).$$

A formula for $\sigma_{i\alpha\gamma}$ is given by Olkin and Siotani (1976, p. 238). The covariance can most easily be expressed by noting that if $r_{i\alpha} = r_{ist}$, the correlation between the s^th^ and t^th^ variables in study i; $r_{i\gamma} = r_{iuv}$, and ρ_{ist} and ρ_{iuv} are the corresponding population values, then

$$\mathrm{Cov}\ (r_{ist}\ ,\ r_{iuv}\ ) = [0.5\ \rho_{ist}\ \rho_{iuv}\ (\rho_{isu}^2 + \rho_{isv}^2 + \rho_{itu}^2 + \rho_{itv}^2\ ) + \rho_{isu}\ \rho_{itv} + \rho_{isv}\ \rho_{itu} - (\rho_{ist}\ \rho_{isu}\ \rho_{isv} + \rho_{its}\ \rho_{itu}\ \rho_{itv} + \rho_{ius}\ \rho_{iut}\ \rho_{iuv} + \rho_{ivs}\ \rho_{ivt}\ \rho_{ivu}\ )]/n_i. \qquad (2)$$

Typically $\sigma_{i\alpha\alpha}$ and $\sigma_{i\alpha\gamma}$ are estimated by substituting corresponding sample estimates for the parameters in (1) and (2). These estimates are denoted below as $\hat{\sigma}_{i\alpha\alpha}$ and $\hat{\sigma}_{i\alpha\gamma}$.

Missing Correlations

When a study has measured fewer than the p variables of interest in the series of studies, p_i is less than p. The vector $\mathbf{r}_i = (r_{i\alpha})$ for $i = 1$ to m_i would then have length $m_i < p^*$. For convenience, however, we will use the subs^ript α to represent the particular relationship measured by $r_{i\alpha}$ rather than the position of $r_{i\alpha}$ in the (shortened) vector $\mathbf{r}_i$. Thus every vector $\mathbf{r}_i$ will have length p^*, but for studies in which fewer than p^* correlations have been observed $\mathbf{r}_i$ will contain m_i observed correlations and $p^* - m_i$ unobserved values. The unobserved values will be identified via an indicator vector $\mathbf{m}_i = (m_{i\alpha})$, $\alpha = 1, \ldots, p^*$, where

$$m_{i\alpha} = \begin{cases} 1, & \text{if } r_{i\alpha} \text{ is observed,} \\ 0, & \text{otherwise.} \end{cases}$$

If $m_i = m_j$ for studies i and j, we say that studies i and j have the same missing data patterns. Also note that $\Sigma_\alpha m_{i\alpha} = m_i$ and denote $\Sigma_i m_i = m$.

Missing covariances. When a study i observes $m_i < p^*$ correlations, the covariance matrix Σ_i defined by (1) and (2) will contain even fewer than $m_i(m_i-1)/2$ observed covariance values. A hidden consequence of missing correlations is that covariances between other reported correlations become impossible to compute. This results from the form of the covariance in (2). Thus, for instance, in study i the correlation r_{itv} is needed to compute the covariance between r_{ist} and r_{isv}.

The values of covariances between observed correlations are indicated by the Hadamard product (*) of the matrix $m_i\ m_i'$ with the full $p^* \times p^*$ matrix Σ_i, specifically $m_i\ m_i' * \Sigma_i$. The matrix $m_i\ m_i'$ contains zeros in the positions of covariances between unobserved correlations, and ones for covariances between observed correlations. The Hadamard product thus shows covariance values where they are observed (or imputed), and full rows and columns of zero values elsewhere. To use the matrix $m_i\ m_i' * \Sigma_i$ in computations involving matrix inversion, its dimension must be reduced from $p^* \times p^*$ to $m_i \times m_i$ by removing all columns and rows which are identically zero. This corresponds to ignoring the unobserved elements in the vector r_i and their associated variances and covariances.

The literature on missing data from experiments and sample surveys offers little specific assistance in how to deal with missing covariances. One approach is to simply ignore potential dependencies between correlations for which covariances cannot be computed. That is, the covariances could be estimated as zero. However, since results are typically intercorrelated[1] this may lead to overweighting the results of studies which have not reported full correlation matrices.

An *ad hoc* adjustment that might be made is to impute values for the missing correlations into the covariance formula (2) using one of the methods discussed below. Becker (1992) discussed two other *ad hoc* approaches to computing missing covariances. In one approach pooled correlations were substituted for subsample values in a study which had not reported complete correlation matrices for the two subsamples of interest. Another approach used patterns of correlations between tests at two times (before and after an intervention) to estimate the between-test correlations across time (e.g., the correlation of pretest A with posttest B).

Results of Series of Studies

The results of *k* independent studies, each examining as many as *p** correlations, can be expressed as the concatenation of the vectors $\mathbf{r}_1, \ldots, \mathbf{r}_k$ containing the nonredundant elements of the matrices of results of the *k* studies. Let the *kp** x 1 vectors of (observed and unobserved) sample and population correlations be denoted as

$$\mathbf{r} = \begin{bmatrix} \mathbf{r}_1 \\ \cdot \\ \cdot \\ \cdot \\ \mathbf{r}_k \end{bmatrix},$$

and

$$\underset{\sim}{\rho} = \begin{bmatrix} \underset{\sim}{\rho}_1 \\ \cdot \\ \cdot \\ \cdot \\ \underset{\sim}{\rho}_k \end{bmatrix},$$

respectively. If the sample sizes of the k independent studies tend to infinity at the same rate (formally if $N = \Sigma_{i=1}^{k} n_i$ and if the $\pi_i = n_i/N$ for $i = 1$ to k remain fixed as $N \rightarrow \infty$) then $\sqrt{N}\,(r - \rho)$ has a nondegenerate asymptotic distribution as $N \rightarrow \infty$. This leads to the large sample approximation that r is normally distributed about ρ. The large sample variance-covariance matrix of r is then Σ, where Σ is a blockwise diagonal matrix with submatrices Σ_1 through Σ_k, and Σ_i is defined above. Specifically,

$$\Sigma = \begin{bmatrix} \Sigma_1 & & & 0 \\ & \Sigma_2 & & \\ & & \cdots & \\ 0 & & & \Sigma_k \end{bmatrix} . \qquad (3)$$

When some studies have not observed all correlations, we also require the concatenated vector of zeros and ones

$$\mathbf{m} = \begin{bmatrix} \mathbf{m}_1 \\ \cdot \\ \cdot \\ \cdot \\ \mathbf{m}_k \end{bmatrix} .$$

The total number of observed correlations is $\Sigma\,\Sigma\, m_{i\alpha} = \Sigma_i\, m_i = m$. Also note that if $\Sigma_i\, m_{i\alpha} = k_\alpha = k$, then all studies have observed correlations for the αth relationship. As above, the covariances among observed rs are indicated by the matrix $\mathbf{m}\,\mathbf{m}' * \Sigma$.

Estimating the Pooled Correlation Matrix

When all of the studies share a common population correlation matrix, that is when $\rho_1 = \ldots = \rho_k$, it makes sense to pool estimates from the studies to estimate the common correlation matrix. In practice one would first test the hypothesis that all studies arise from a single population, then estimate either a pooled (common) or average correlation matrix. Procedures for estimation and testing of the pooled matrix for the case in which all correlations are observed (from Becker, in press) are repeated here for conv nience of notation. Results for the incomplete-data case are given in the next section.

Notation and Model

To estimate a common correlation vector of length p^*, the generalized least squares (GLS) model is

$$r = X\,\rho. + e,$$

where r is the vector of kp^* correlation coefficients, $\rho. = (\rho_{\cdot 1}, \ldots, \rho_{\cdot p^*})$ is the set of common correlations to be estimated, and X is a kp^* x p^* matrix created by "stacking" k identity matrices, each of dimension p^* x p^*. If $k =$ 10 and $p^* = 3$ (as in the examples which follow), X would be the 30 x 3 matrix

$$X = \left[\begin{array}{ccc} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ \hdashline 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ \hdashline & \vdots & \\ \hdashline 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{array}\right].$$

Under the assumptions of the GLS model, the error vector $e = r - X\,\rho.$ has mean zero and approximate covariance matrix Σ. The estimate of the pooled correlation matrix is then simply the usual GLS estimate of the regression coefficients, here,

$$r_{.} = (X' \Sigma^{-1} X)^{-1} X' \Sigma^{-1} r, \qquad (4)$$

with apprcximate variance-covariance matrix given by

$$V = (X' \Sigma^{-1} X)^{-1}. \qquad (5)$$

Typically both $r_.$ and V are computed using an estimated variance matrix in place of Σ. When the large-sample normality of the vector r is justified, $r_.$ can also be assumed normal, and standard inferential procedures (e.g., confidence intervals, test of significance about the elements of $\rho_.$) are possible.

Test of Homogeneity

Becker (in press) also presents a test for homogeneity of correlation matrices, similar to that derived by Hedges and Olkin (1985). The test of the hypothesis of homogeneity of correlation matrices, that is to test

$$H_0 : \rho_1 = \ldots = \rho_k,$$

uses the statistic

$$Q = r' [\Sigma^{-1} - \Sigma^{-1} X(X' \Sigma^{-1} X)^{-1} X' \Sigma^{-1}] r.$$

When H_0 is true Q has approximately a chi-square distribution with $kp^* - p^*$ degrees of freedom. Thus a test of H_0 at the 100α percent level of significance is given by rejecting H_0 if Q exceeds the $100(1-\alpha)$ percentile point of the chi-square distribution with $(k-1)p^*$ degrees of freedom.

Estimation when some Correlations are Unobserved

Complete-case Analysis

The complete-case analysis approach to missing data suggests that parameters be estimated using those cases (here, studies) which report complete data. The drawback of this approach is that the data loss can be great if more than a few studies fail to report all p^* correlations. Because exact replications are discouraged by journal editors and avoided by researchers, this problem is likely to be encountered in syntheses of most research domains in the social sciences. The complete-case analysis is not the analysis of choice in most situations. However, it may be useful in providing estimates to use, for example, in computing missing covariances.

The complete-case analysis would involve the application of GLS estimation methods to the set of results of studies reporting all correlations. Denote the number of studies which report complete correlation matrices as k_c. To estimate a common correlation vector of length p^*, the model is

$$\mathbf{r}_c = \mathbf{X}_c \boldsymbol{\rho}. + \mathbf{e}_c,$$

where $\mathbf{r}_c$ is a concatenation of the $\mathbf{r}_i$ vectors (like $\mathbf{r}$ above) but includes only the results of the k_c samples with complete matrices, $\boldsymbol{\rho}. = (\rho_{\cdot 1}, \ldots, \rho_{\cdot p^*})$ is the set of common correlations to be estimated, and $\mathbf{X}_c$ is a $k_c p^* \times p^*$ matrix created by "stacking" k_c identity matrices, each of dimension $p^* \times p^*$. The variance matrix for the vector $\mathbf{r}_c$ is denoted Σ_c, and contains the Σ_i matrices for the k_c samples with complete results.

The estimates of the pooled correlation matrix and its variance are then simply the usual GLS estimates, computed using $\mathbf{r}_c$, $\mathbf{X}_c$, and Σ_c in place of $\mathbf{r}$, $\mathbf{X}$, and Σ, respectively, in (4) and (5).

Available-cases Analysis

One relatively simple approach to handling missing data in multivariate analysis is to use "available-cases" analysis (Little & Rubin, 1987, p. 41). A number of different estimators are possible within this framework The estimate given here is essentially based on pairwise available cases.

The available-cases estimate presented here is a generalization of the pooled correlation matrix estimated via generalized least squares (GLS) shown above in (4) and (5). However, since $\underline{p}^* - \underline{m}$ of the possible correlations in r are unobserved, we omit all rows of r and X that represent unobserved correlations. We denote the reduced vector and matrix as r_0 and $\mathbf{X}_0$. Both r_0 and $\mathbf{X}_0$ then contain $\underline{m}$ rows.

Also we reduce the dimension of Σ (or $\hat{\Sigma}$) as described above (omitting rows and columns of all zeros), and denote the new $\underline{m} \times \underline{m}$ covariance matrix as Σ_0. As noted above, it is also necessary either to assume that the missing covariances between <u>reported</u> $\underline{r}$ values (which require unobserved $\underline{r}$ values to be computed) equal zero, or to estimate those covariances using other values for the missing $\underline{r}$s.

We rewrite the GLS model as

$$r_0 = X_0 \rho_{\cdot} + e_0.$$

In the new model the error vector $e_0 = r_0 - X_0 \rho_{\cdot}$ has approximate covariance matrix Σ_0, so the GLS estimate of $\rho_{\cdot}$ is

$$r_{\cdot} = (X_0' \Sigma_0^{-1} X_0)^{-1} X_0' \Sigma_0^{-1} r_0 \tag{6}$$

with approximate variance covariance matrix given by

$$V = (X_0' \Sigma_0^{-1} X_0)^{-1}. \tag{7}$$

For the available case analysis the test of the hypothesis of homogeneity of correlation matrices uses the statistic

$$Q = r_0' \, [\Sigma_0^{-1} - \Sigma_0^{-1} X_0 (X_0' \, \Sigma_0^{-1} X_0)^{-1} X_0' \; \Sigma_0^{-1} \,] \, r_0 .$$

When H_0 (given above) is true Q has approximately a chi-square distribution with $\underline{m}$ - $\underline{p}^*$ degrees of freedom. Thus a test of H_0 at the 100α percent level of significance is given by rejecting H_0 if Q exceeds the 100(1-α) percentile point of the chi-square distribution with $\underline{m}$ - $\underline{p}^*$ degrees of freedom.

Imputing Unconditional Means

Using this approach we would substitute for each unreported $\underline{r}$ value the appropriate mean estimated by the available-case analysis, then proceed (e.g., with GLS estimation) as if the data were complete. Specifically, if the value $\underline{r}_{i\alpha}$ is not reported in study $\underline{i}$, one would substitute the mean $\underline{r}_{.\alpha}$ computed using (6) above. Although one typically obtains reasonable average values using this approach, variances and covariances are systematically underestimated because the imputed values by definition lie near the center of the distribution of the observed correlations (Little & Rubin, 1987). In meta-analysis this implies that tests of homogeneity can be reduced when mean values are substituted.

Two complications which arise with this approach involve questions of homogeneity and the precision of the predicted correlations (the unconditional means). Unlike the two approaches described above, this approach involves substituting particular values for the missing correlations, and using them as if they had been reported by the studies as actual data. Thus it is important to ask whether the means that are imputed are "reasonable" values.

Because the unobserved correlations are unavailable for comparison it is impossible to really gauge whether the substituted mean values are appropriate. However, one indication of the representativeness of the means for the studies from which they are obtained is the test of homogeneity. Thus failure to reject the hypothesis of homogeneity for the complete cases suggests that the mean values are good measures of the relationships of interest in all of the studies which reported them.

The second question which arises when imputed values are used in the GLS estimation framework is how their sampling variances and covariances should be computed. The formulas (1) and (2) assume that the correlation values are all computed for the same sample. Substitution of other values into these formulas can lead to correlations between rs that are out of range and to within-study covariance matrices that are not positive definite. Standard imputation procedures for missing data in experiments suggest a number of adjustments for the general underestimation of sample variances, however, in those cases all missing values on a particular variable are homoscedastic, which is not generally the case in meta-analysis.

Ad hoc estimates of sampling variability were used in the present analyses. The variance of the imputed correlation $r_{\cdot\alpha}$ in study i was computed as

$$\hat{\sigma}_{i\alpha\alpha} + V_{\alpha\alpha}, \qquad (8)$$

where $\hat{\sigma}_{i\alpha\alpha}$ is the estimated sampling variance in (1), computed using the mean $r_{\cdot\alpha}$ and n_i, and $V_{\alpha\alpha}$ is the variance of the mean $r_{\cdot\alpha}$ from the available-case analysis (i.e., the α^{th} diagonal element in (7) above).

The rationale for (8) is based on the theory for the estimation of a value of a response variable Y^* from a predicted value $\hat{Y}^*$ in linear regression (see Seber, 1977, Sec. 5.3). In that simpler case, Seber noted that var[$\hat{Y}^*$ -

$\underline{Y}$*] = $\sigma^2(\underline{v}^* + 1)$, where σ^2 is the variance of $\underline{Y}$* and $\underline{v}^*\sigma^2$ is the variance of the predicted score based on a particular set of predictor values (**x***). The confidence interval for $\underline{Y}$* is computed using an estimate of the standard error $\sigma(\underline{v}^* + 1)^{1/2}$. The analogue in the present context is to use the estimated variance of the predicted correlation (i.e., the mean) in place of var($\hat{\underline{Y}}$*) and the sampling variance computed from (1) in place of var($\underline{Y}$*).

Covariances involving $\underline{r}_{\cdot\alpha}$ were computed as Cov($\underline{r}_{\cdot\alpha}$, $\underline{r}_{i\gamma}$), for $\alpha \neq \gamma$, using formula (2). However, the question of how best to estimate both variances and covariances involving imputed values requires further investigation.

Imputing Conditional Means

This method, proposed by Buck (1960), involves estimating the unreported correlations from a prediccion model based on the complete cases. Typically (in practice) several regression models would be estimated, one for each variable (i.e., correlation) with unreported values. The unreported values are estimated case by case for each variable.

In the multivariate meta-analysis context the predictors of a correlation with unreported values would be the other observed correlations. For instance, consider a case in which some values of $\underline{r}_{i\alpha}$ are unreported but all other correlations are completely reported. To predict missing values of the α^{th} correlation in study $\underline{j}$ one would regress the $\underline{k}_\alpha$ reported values of $\underline{r}_{i\alpha}$ on values of $\underline{r}_{i1}$, ..., $\underline{r}_{i(\alpha-1)}$, $\underline{r}_{i(\alpha+1)}$,... $\underline{r}_{ip*}$, then use the reported correlations in study $\underline{j}$ to predict $\hat{\underline{r}}_{j\alpha}$. Weighted least squares regression should be used in the meta-analysis context (see, e.g., Hedges, 1983), weighting each value of $\underline{r}_{i\alpha}$ by the inverse of its variance. The drawback of this approach is, again, that estimation may be difficult if few studies provide complete correlation matrices.

The analysis then proceeds as if the data were complete, with the predicted values in place of the unobserved correlations. Also, because the unobserved values are predicted from the other data, their variances and covariances are again not given by $\sigma_{i\alpha\gamma}$. Below another *ad hoc* variance estimate is used (following the same rationale as above). Specifically, the value

$$\hat{\sigma}_{i\alpha\alpha} + \underline{v}^*_{i\alpha\alpha}, \qquad (9)$$

is used, where $\hat{\sigma}_{i\alpha\alpha}$ is the sampling variance (1) computed using the imputed value of $\underline{r}_{i\alpha}$, $\underline{v}^*_{i\alpha\alpha} = x^*_i{}' \, V_b \, x^*_i$, and $x^*_i = (1, \underline{r}_{i1}, \underline{r}_{i\alpha-1}, \underline{r}_{i\alpha+1}, \ldots, \underline{r}_{ip^*})'$ is the vector of "predictor values" for the *i*th study. The value $\underline{v}^*_{i\alpha\alpha}$ is an approximate variance of the predicted *r* value. Note, however, that this variance does not account for the fact that the predictors are themselves random variables, measured with uncertainty. Similarly, the regression slope estimates treat the predictors as though they are known. A more appropriate but more complex analysis could treat the observed correlations as regressors measured with error (e.g., Seber, 1977, sec. 6.4). Covariances are computed as $\mathrm{Cov}(\hat{\underline{r}}_{i\alpha}, \underline{r}_{i\gamma})$ using formula (2).

Empirical Bayes Estimation

In some situations it may be more reasonable to expect the patterns of intercorrelations among a set of variables to differ between studies. Population correlations might be *expected* to vary if a variety of subject groups had been studied. Even if variation in the pattern of correlations is not expected, the test of homogeneity may suggest that the population correlation matrices differ.

When the population correlation matrices vary a random-effects model may be appropriate for the data. If we are willing to treat the distribution of the population correlations as a prior distribution for the data, we can use empirical Bayes estimation techniques to estimate the mean correlation vector and its variance.

Random-effects Model

Consider again the large sample distribution of the correlation vector r_i. The result that r_i is approximately normal with a mean $\underline{\rho}_i$ implies that we can write the vector r_i in terms of a parameter $\underline{\rho}_i$ plus a vector of errors, say e_i. That is,

$$r_i = \underline{\rho}_i + e_i, \tag{10}$$

and e_i is then distributed approximately normally with a mean of 0 and variance Σ_i, for $\underline{i} = 1, \ldots, \underline{k}$, where the elements of Σ_i are given by (1) and (2) above.

In the random-effects case we assume further that each vector of parameters $\underline{\rho}_i$ is composed of a common component $\underline{\rho}. = (\rho._\alpha)$ for $\alpha = 1, \ldots, \underline{p}^*$ plus a residual vector, say, u_i. Specifically,

$$\underline{\rho}_i = \underline{\rho}. + u_i, \tag{11}$$

for $\underline{i} = 1, \ldots, \underline{k}$. The variation represented by u_i is parameter variation, rather than sampling variation, which is represented by the error term e_i above. That is, we assume that the vectors of population correlations vary randomly about a common mean (which we wish to estimate). We denote the matrix of parameter variances as $T = (\tau_{\alpha\gamma})$ for $\alpha, \gamma = 1$ to $\underline{p}^*$.

Estimation

The estimation of $\underline{\rho}.$ and T can be accomplished via the EM algorithm

(Dempster, Laird, & Rubin, 1977; Dempster, Rubin, & Tsutakawa, 1981), an iterative procedure. Implementation of the algorithm in the present context involves, first, imputation of the conditional means of the missing data values (given the observed correlations) as described above. The observed and imputed values are then treated as complete data, and initial estimates of the mean and variance component ($\mathbf{r}_{\cdot(0)}$ and $\hat{T}_{(0)}$) are obtained.

The estimated mean and variance are next treated as a Bayesian prior for the observed (and imputed) correlations. Weighted estimates of the $\underline{\rho}_i$ vectors are then computed, as are their standard errors. The cycle begins again as these "study-parameter" estimates are used to re-estimate the means and their standard errors. The iteration between these two procedures continues until the estimates of the mean vector and the parameter variances do not change materially with added iterations (i.e., until the maximum of the likelihood function is attained).

In some situations implementation of the EM algorithm can be computationally intensive. However, the computations in the present case are relatively straightforward. The appendix gives a program written using SAS PROC MATRIX which accomplishes the computations outlined below.

Posterior distribution of $\underline{\rho}$. The estimation of the mean vector $\underline{\rho}_{\cdot}$ and the variance-covariance matrix T requires the posterior distribution of the vector of study parameters ρ_{11} through ρ_{kp*} (i.e., $\underline{\rho}$). From model (10) above and the distribution of the sample correlation in (1) and (2) we know that for large samples the within-study sampling error ($\underline{e}_{i\alpha}$) is normally distributed. Since

$$\underline{r}_{i\alpha} = \rho_{i\alpha} + \underline{e}_{i\alpha}, \qquad \text{for } \alpha = 1 \text{ to } \underline{p}^* \text{ and } \underline{i} = 1 \text{ to } \underline{k},$$

then if $\underline{N} = \Sigma_i \underline{n}_i$ and $\pi_i = \underline{n}_i/\underline{N}$ remain fixed as $\underline{N}$ approaches ∞, we can write

$$\sqrt{\underline{N}}\,(r - \varrho) \;{}^{\blacktriangle}\; N(0,\, \Sigma^*),$$

where Σ^* is defined via $\Sigma^* = \sqrt{\underline{N}}\,\Sigma$, and the elements of Σ are given by (1) and (2). Thus the approximate density of the vector r conditional on the vector of study-parameters ϱ is given by

$$f(r \mid \varrho) = |\Sigma|^{1/2}\,(2\pi)^{kp^*/2}\,\exp\{-\tfrac{1}{2}\,(r - \varrho)\Sigma^{-1}(r - \varrho)'\}.$$

The second-stage model (11) shows the population correlation vector for each study varying around the mean population correlation for the α^{th} relationship, across studies. In terms of individual correlations, we write

$$\rho_{i\alpha} = \rho_{\cdot\alpha} + \underline{u}_{i\alpha}, \qquad \text{for } \alpha = 1 \text{ to } \underline{p}^* \text{ and } \underline{i} = 1 \text{ to } \underline{k}.$$

We define $\tau_{\alpha\gamma} = \mathrm{Cov}(\rho_{i\alpha}, \rho_{i\gamma})$ for $\underline{i} = 1$ to $\underline{k}$ and α, $\gamma = 1$ to $\underline{p}^*$. If we are willing to assume that the study-parameters $\rho_{i\alpha}$ are normally distributed about the means $\rho_{\cdot\alpha}$, then we can write the density of the vector of study-parameters ϱ as

$$f(\varrho) = |T|^{1/2}\,(2\pi)^{kp^*/2}\,\exp\{-\tfrac{1}{2}\,(\varrho - \varrho_{\cdot})T^{-1}(\varrho - \varrho_{\cdot})'\},$$

where T is a $\underline{kp}^* \times \underline{kp}^*$ blockwise diagonal matrix containing $\underline{k}$ $(\underline{p}^* \times \underline{p}^*)$ blocks of $\tau_{\alpha\gamma}$ values and $\varrho_{\cdot}$ is a $\underline{kp}^* \times 1$ vector defined as

$$\boldsymbol{\rho}_{\cdot} = (\rho_1,\ \rho_2,\ \ldots,\ \rho_{p^*},\ \ldots,\ \rho_1,\ \rho_2,\ \ldots,\ \rho_{p^*}).$$

That is, $\boldsymbol{\rho}_{\cdot}$ is a concatenation of $\underline{k}$ sets of the average population correlations ρ_1 through ρ_{p^*}. This slight variation on the notation used above gives $\boldsymbol{\rho}_{\cdot}$ the same dimension as $\boldsymbol{\rho}$, the vector of study parameters.

The posterior distribution of $\boldsymbol{\rho}$ given $\mathbf{r}$ is then

$$\begin{aligned} f(\boldsymbol{\rho} \mid \mathbf{r}) \quad &\propto \quad f(\mathbf{r} \mid \boldsymbol{\rho})\ f(\boldsymbol{\rho}) \;=\; |\Sigma\ T|^{1/2}\ (2\pi)^{kp^*} \exp\{-\tfrac{1}{2}\ (\mathbf{r} - \boldsymbol{\rho})\Sigma^{-1}(\mathbf{r} - \boldsymbol{\rho})'\} \\ &\qquad\qquad\qquad\qquad\qquad \times\ \exp\{-\tfrac{1}{2}\ (\boldsymbol{\rho} - \boldsymbol{\rho}_{\cdot})T^{-1}(\boldsymbol{\rho} - \boldsymbol{\rho}_{\cdot})'\} \\ &\propto \quad \exp\{-\tfrac{1}{2}\ (\mathbf{r} - \boldsymbol{\rho})\Sigma^{-1}(\mathbf{r} - \boldsymbol{\rho})' + -\tfrac{1}{2}\ (\boldsymbol{\rho} - \boldsymbol{\rho}_{\cdot})T^{-1}(\boldsymbol{\rho} - \boldsymbol{\rho}_{\cdot})'\}. \qquad (12) \end{aligned}$$

By expanding the quadratic forms in (12) and eliminating terms that do not depend on $\boldsymbol{\rho}$ we obtain

$$\begin{aligned} f(\boldsymbol{\rho} \mid \mathbf{r}) \ &\propto\ \exp\{-\tfrac{1}{2}\ [\boldsymbol{\rho}\ \Sigma^{-1}\ \boldsymbol{\rho}' - 2\boldsymbol{\rho}\ \Sigma^{-1}\ \mathbf{r}' + \boldsymbol{\rho}\ T^{-1}\ \boldsymbol{\rho}' - 2\boldsymbol{\rho}\ T^{-1}\ \boldsymbol{\rho}_{\cdot}']\} \\ &=\ \exp\{-\tfrac{1}{2}\ [\boldsymbol{\rho}\ (\Sigma^{-1} + T^{-1})\ \boldsymbol{\rho}' - 2\boldsymbol{\rho}\ (\Sigma^{-1}\ \mathbf{r}' + T^{-1}\ \boldsymbol{\rho}_{\cdot}')]\}. \qquad (13) \end{aligned}$$

We next define the matrices

$$\Psi = (\Sigma^{-1} + T^{-1})^{-1}$$

and

$$\boldsymbol{\rho}_1' = (\Sigma^{-1} + T^{-1})^{-1}(\Sigma^{-1}\ \mathbf{r}' + T^{-1}\ \boldsymbol{\rho}_{\cdot}') = \Psi\ (\Sigma^{-1}\ \mathbf{r}' + T^{-1}\ \boldsymbol{\rho}_{\cdot}').$$

Note that although $\boldsymbol{\rho}_1$ is a one-dimensional vector, we will denote its elements as ρ_{1ij} in order to identify the study and relationship associated with each element. The elements of $\boldsymbol{\rho}_1'$ are thus arrayed as $(\rho_{111},\ \rho_{112},\ \ldots,\ \rho_{11p^*},\ \ldots,\ \rho_{1k1},\ \rho_{1k2},\ \ldots,\ \rho_{1kp^*})$.

Next multiply (13) by the term $\exp\{-\frac{1}{2}\, \boldsymbol{\rho}_1\, \Psi^{-1}\, \boldsymbol{\rho}_1'\}$, which is independent of $\boldsymbol{\rho}$. Substituting Ψ^{-1} for $(\Sigma^{-1} + \mathrm{T}^{-1})$ and $\Psi^{-1}\, \boldsymbol{\rho}_1'$ for $(\Sigma^{-1}\mathrm{r}' + \mathrm{T}^{-1}\, \boldsymbol{\rho}.')$ produces

$$\begin{aligned} f(\boldsymbol{\rho} \mid \mathrm{r}) \;\;\propto\;\; & \exp\{-\tfrac{1}{2}\, [\boldsymbol{\rho}\, (\Sigma^{-1} + \mathrm{T}^{-1})\, \boldsymbol{\rho}' - 2\boldsymbol{\rho}\, (\Sigma^{-1}\mathrm{r}' + \mathrm{T}^{-1}\, \boldsymbol{\rho}.') + \boldsymbol{\rho}_1\, \Psi^{-1}\, \boldsymbol{\rho}_1']\} \\ = \;& \exp\{-\tfrac{1}{2}\, [\boldsymbol{\rho}\, \Psi^{-1}\boldsymbol{\rho}' - 2\boldsymbol{\rho}\, \Psi^{-1}\, \boldsymbol{\rho}_1' + \boldsymbol{\rho}_1\, \Psi^{-1}\, \boldsymbol{\rho}_1']\} \\ = \;& \exp\{-\tfrac{1}{2}\, (\boldsymbol{\rho} - \boldsymbol{\rho}_1)\Psi^{-1}(\boldsymbol{\rho} - \boldsymbol{\rho}_1)'\}, \end{aligned} \qquad (14)$$

which is the kernel of the multivariate normal distribution. Thus the posterior distribution of $\boldsymbol{\rho}$ (given r) is normal with mean $\boldsymbol{\rho}_1$ and variance Ψ.

EM algorithm. The EM algorithm makes use of the distribution defined by (14) in the E or expectation step of the process. The EM approach requires initial estimates of T and ρ_1 through ρ_{p*} (the average correlations). Because these are starting values, simple estimators are typically all that is needed. The starting values $\hat{\mathrm{T}}^{(0)}$ and $\hat{\boldsymbol{\rho}}.^{(0)} = \mathrm{r}.$ are used to compute the posterior mean of $\boldsymbol{\rho}$ and its variance, that is, $\hat{\boldsymbol{\rho}}_1^{(1)}$ and $\hat{\Psi}^{(1)}$. New estimates of T and $\boldsymbol{\rho}.$ (i.e., $\hat{\mathrm{T}}^{(1)}$ and $\hat{\boldsymbol{\rho}}.^{(1)}$) are then computed based on the sufficient statistics from the $\hat{\boldsymbol{\rho}}_1^{(1)}$ values. (The specific forms of the estimates are given below.) The cycle continues until the likelihood in (14) is maximized, or practically speaking, until the differences between parameter estimates from one iteration to the next are small.

Starting values. For starting values we use weighted method-of-moments estimators $\hat{\rho}_\alpha^{(0)}$ for $\alpha = 1$ to $\underline{p}*$ and $\hat{\mathrm{T}}^{(0)} = (\hat{\tau}_{\alpha\gamma}^{(0)})$ for α, $\gamma = 1$ to $\underline{p}*$, specifically,

$$\hat{\rho}_\alpha^{(0)} = \underline{r}_{\cdot\alpha} = \Sigma_i\, \underline{w}_{i\alpha}\, \underline{r}_{i\alpha}\;,$$

and

$$\hat{\tau}_{\alpha\gamma}{}^{(0)} = \frac{\underline{S}_{\alpha\gamma} - \Sigma_i\, \underline{w}_{i\alpha\gamma}\, \hat{\sigma}_{i\alpha\gamma}(1 - \underline{w}_{i\alpha} - \underline{w}_{i\gamma}) - (\Sigma_i\, \underline{w}_{i\alpha\gamma})(\Sigma_i\, \underline{w}_{i\alpha}\, \underline{w}_{i\gamma}\, \hat{\sigma}_{i\alpha\gamma})}{\Sigma_i\, \underline{w}_{i\alpha\gamma}\, (1 - \underline{w}_{i\alpha} - \underline{w}_{i\gamma} - \Sigma_s\, \underline{w}_{s\alpha}\, \underline{w}_{s\gamma})},$$

for

$$\underline{S}_{\alpha\gamma} = \Sigma_i\, \underline{w}_{i\alpha\gamma}\, (\underline{r}_{i\alpha} - \underline{r}_{\cdot\alpha})(\underline{r}_{i\gamma} - \underline{r}_{\cdot\gamma}),$$

where $\underline{w}_{i\alpha\gamma} = (\underline{w}_{i\alpha}\, \underline{w}_{i\gamma})^{1/2}$ is a weight associated with correlations $\underline{r}_{i\alpha}$ and $\underline{r}_{i\gamma}$, $\underline{w}_{i\alpha} = [1/\hat{\sigma}_{i\alpha\alpha}]/\Sigma_s[1/\hat{\sigma}_{s\alpha\alpha}]$ (that is, $\underline{w}_{i\alpha}$ is the usual inverse-variance weight used in univariate meta-analyses), and where the values of $\hat{\sigma}_{i\alpha\gamma}$ are given by (1) and (2). These estimators are superscripted with the index zero to indicate that they are starting values.

When the amount of variation in the sample correlations for the α^{th} relationship is quite small the variance estimate $\hat{\tau}_{\alpha\gamma}{}^{(0)}$ can frequently be negative. By convention, negative values are set to zero, as would be any other covariance estimates involving the α^{th} relationship (i.e., values of $\hat{\tau}_{\alpha\gamma}{}^{(0)}$ for that value of α).

Expectation step. The posterior distribution of the values $\rho_{11}, \ldots, \rho_{1p*}, \ldots, \rho_{k1}, \ldots, \rho_{kp*}$ (given the data) is then used to obtain estimates of the study parameters. These are essentially weighted combinations of the original data (the $\underline{r}$s) and the starting values of the mean correlations ρ_1 through ρ_{p*}. We compute

$$\hat{\Psi}^{(0)} = (\Sigma^{-1} + [\hat{T}^{(0)}]^{-1})^{-1}$$

and

$$\hat{\varrho}_1{}^{(0)} = \hat{\Psi}^{(0)}\, (\Sigma^{-1}\, \mathbf{r}' + [\hat{T}^{(0)}]^{-1}\, \hat{\varrho}_{\cdot}{}^{(0)\prime}).$$

Maximization step. In this step new estimates of T (i.e., $\hat{T}^{(1)}$) and the mean correlation vector are obtained from the sufficient statistics for the study-parameter estimates. The estimates of the elements of T and $\boldsymbol{\rho}_{.}$ on iteration ($\underline{t}$ + 1) are given by

$$\hat{\tau}_{\alpha\gamma}^{(t+1)} = (1/\underline{k})[\ \Sigma_i\ (\hat{\psi}_{i\alpha\gamma}^{(t)} + \hat{\rho}_{1i\alpha}^{(t)}\ \hat{\rho}_{1i\gamma}^{(t)}) - \underline{k}\ \hat{\rho}_{.\alpha}^{(t)}\ \hat{\rho}_{.\gamma}^{(t)}],$$

and

$$\hat{\rho}_{\cdot\alpha}^{(t+1)} = [\Sigma_i\ \hat{\rho}_{1i\alpha}^{(t)}]/\underline{k}, \qquad \text{for } \alpha,\ \gamma = 1 \text{ to } \underline{p}*,$$

where $\hat{\psi}_{i\alpha\gamma}^{(t)}$ is an element of the matrix $\hat{\Psi}^{(t)}$, $\hat{\rho}_{1i\alpha}^{(t)}$ is an element of $\hat{\boldsymbol{\rho}}_{1}^{(t)}$, and

$$\hat{\boldsymbol{\rho}}_{1}^{(t)} = \hat{\Psi}^{(t)}\ (\Sigma^{-1}\ r' + [\hat{T}^{(t)}]^{-1}\ \hat{\boldsymbol{\rho}}_{.}^{(t)\prime}).$$

Iteration. The process of estimation and maximization is repeated until the likelihood function is maximized, that is, until the paramrter estimates (e.g., the estimates of T and ρ_1 through ρ_{p*}) do not change much from one iteration to the next. Note, however, that the program given in the appendix stops after iterating for a fixed number of cycles rather than stopping after a convergence criterion has been met.

Missing data. The EM algorithm can be applied when all correlations have been observed or when some correlations are missing. When data are missing at random (that is, when the reason that correlations are unobserved is unrelated to the actual values of the unobserved correlations) then it is possible to get maximum likelihood estimates by ignoring the missing-data mechanism. Little and Rubin (1987, Chapter 8) discuss this problem in detail for multivariate normal examples with unknown covariance matrices. The present case is similar, but involves normal data with a known covariance matrix.

Application of Little and Rubin's methodology for handling missing data adds only one step to the estimation process described above. Before obtaining starting values $\hat{T}^{(0)}$ and $\hat{\underline{\rho}}_{\cdot}^{(0)}$ we must impute values of the unobserved correlations.

The value imputed is the expected value of the missing correlation, conditional on the observed data. In practice this means substituting the best estimate of $\underline{r}_{i\alpha}$ available, based on the observed data. If correlation $\underline{r}_{i\alpha}$ is missing, one would use the regression method for imputing conditional means to predict a value $\hat{\underline{r}}_{i\alpha}$ for study $\underline{i}$, as described above. When some studies are missing more than one correlation, missing values would be estimated for each pattern of missing data, using an approach similar to that described in Little and Rubin's (1987) Chapter 6. The imputed values are then substituted into the data set and analysis via the EM algorithm proceeds as if the data were complete.

Example

Data

Data for the example are from ten samples in four studies which explored the relationship of spatial ability to SAT scores for high-school or junior-high students. The ten samples from these studies are drawn from a more extensive synthesis of sex differences in the relations among math, spatial, and verbal ability measures by Friedman (in press). This example considers correlations among measures of at most three variables from each sample (i.e., $\underline{p}^* = 3$), as shown in Table 1. We have omitted the correlations between SAT-M and spatial ability reported for the two samples from Rosenberg (1981) to create an example with less than complete data. Correlations and sample sizes are shown in Table 2.

Insert Tables 1 and 2 about here

In the ith study, the correlations among the three variables are represented in our notation as:

$$\begin{array}{l|ccc|}
 & \text{Math} & \text{Spatial} & \text{Verbal} \\
\text{Math} & 1.0 & r_{i1} & r_{i2} \\
\text{Spatial} & r_{i1} & 1.0 & r_{i3} \\
\text{Verbal} & r_{i2} & r_{i3} & 1.0
\end{array} .$$

Writing these correlations as a vector $\mathbf{r}_i$, the relationships represented are

$$\begin{array}{l|c|}
\text{Math-Spatial} & r_{i1} \\
\text{Math-Verbal} & r_{i2} \\
\text{Spatial-Verbal} & r_{i3}
\end{array} .$$

The $\mathbf{r}_i$ vectors for four of the ten samples in the example are

$$\mathbf{r}_1 = \begin{bmatrix} .47 \\ -.21 \\ -.15 \end{bmatrix}, \qquad \mathbf{r}_2 = \begin{bmatrix} .28 \\ .19 \\ .18 \end{bmatrix},$$

$$\mathbf{r}_5 = \begin{bmatrix} r_{51} \\ .48 \\ .23 \end{bmatrix}, \qquad \mathbf{r}_6 = \begin{bmatrix} r_{61} \\ .74 \\ .44 \end{bmatrix}.$$

Vectors $\mathbf{r}_1$ and $\mathbf{r}_2$, from Becker (1978) represent complete correlation matrices. However, in $\mathbf{r}_5$ and $\mathbf{r}_6$ from Rosenberg (1981), the elements $\underline{r}_{51}$ and $\underline{r}_{61}$ are not observed in our example data.

Each vector of correlations has an associated limiting variance-covariance matrix, computed using (1) and (2) above. The limiting variance-covariance matrices for the Becker (1978) samples are

$$\hat{\Sigma}_1 = \begin{bmatrix} .0082 & -.0010 & -.0018 \\ -.0010 & .0123 & .0058 \\ -.0018 & .0058 & .0129 \end{bmatrix} \text{ and } \hat{\Sigma}_2 = \begin{bmatrix} .0056 & .0009 & .0010 \\ .0009 & .0061 & .0016 \\ .0010 & .0016 & .0061 \end{bmatrix}.$$

The two matrices for the samples from Rosenberg (1981) are, respectively

$$\hat{\Sigma}_5 = \begin{bmatrix} \hat{\sigma}_{511} & \hat{\sigma}_{512} & \hat{\sigma}_{513} \\ \hat{\sigma}_{521} & .0116 & \hat{\sigma}_{523} \\ \hat{\sigma}_{531} & \hat{\sigma}_{532} & .0176 \end{bmatrix} \text{ and } \hat{\Sigma}_6 = \begin{bmatrix} \hat{\sigma}_{611} & \hat{\sigma}_{612} & \hat{\sigma}_{613} \\ \hat{\sigma}_{621} & .0114 & \hat{\sigma}_{623} \\ \hat{\sigma}_{631} & \hat{\sigma}_{632} & .0361 \end{bmatrix}.$$

The covariances of the reported correlations from Rosenberg (the off-diagonal elements) must also be imputed because values of $\underline{r}_{51}$ and $\underline{r}_{61}$, respectively, are needed in their computation.

The vector of all correlations to be synthesized then is

$$\mathbf{r}' = (.47\ -.21\ -.15\quad .28\ .19\ .18\quad .48\ .41\ .26\quad .37\ .40\ .27\quad \underline{r}_{51}\ .48\ .23$$
$$\underline{r}_{61}\ .74\ .44\quad .26\ .72\ .36\quad .32\ .52\ .10\quad .58\ .64\ .40\quad .34\ .28\ -.03)$$

and its variance $\hat{\Sigma}$ is the 30 x 30 blockwise diagonal matrix comprised of $\hat{\Sigma}_1$ through $\hat{\Sigma}_{10}$.

Complete-case Analysis

Complete data from eight of the ten samples (i.e., from all studies except Rosenberg (1981)) is used to estimate $\rho.$ and V. The GLS estimate of the mean correlation vector is

$$(.367,\ .202,\ .421)',$$

with variance-covariance matrix

$$V = \begin{bmatrix} .0014 & .0002 & .0005 \\ .0002 & .0014 & .0005 \\ .0005 & .0005 & .0017 \end{bmatrix}.$$

The test of homogeneity for the complete-case analysis is $Q = 62.09$, which under the null hypothesis of homogeneity is a chi-square with (8-1)3 or 21 degrees of freedom. The value of Q is larger than the upper-tail $\alpha = .05$ critical value for 21 degrees of freedom, suggesting that the eight samples do not share a single population matrix. Thus although we have used the data above to estimate a pooled correlation matrix, the interpretation of that matrix as a shared or common population matrix seems unwarranted.

Available-case Analysis

In Friedman's data $m = 28$, so two elements of r and two rows of X are eliminated. Thus the reduced matrices are

$$r_0 = \begin{bmatrix} .47 \\ -.21 \\ -.15 \\ .28 \\ .19 \\ .18 \\ .48 \\ .41 \\ .26 \\ .37 \\ .40 \\ .27 \\ \mathbf{.48} \\ \mathbf{.23} \\ \mathbf{.74} \\ \mathbf{.44} \\ .26 \\ .72 \\ .36 \\ .32 \\ .52 \\ .10 \\ .58 \\ .64 \\ .40 \\ .34 \\ .28 \\ -.03 \end{bmatrix} \quad \text{and} \quad X_0 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ \mathbf{0} & \mathbf{1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{1} \\ \mathbf{0} & \mathbf{1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{1} \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} .$$

The rows shown in bold represent the Rosenberg results. Note that every row of **X** for Rosenberg shows a zero in column one (i.e., the column for the first element of the pooled correlation matrix). In order to estimate the covariances between the two reported values for the Rosenberg samples, $\underline{r}_{\cdot 1}$ = .367 from the complete-case analysis was used as the value of $\underline{r}_1$.

The estimate of the pooled correlation matrix from the available-case analysis, using (6) above, is

$$r. = (.374, .437, .227)',$$

with variance-covariance matrix computed from formula (7) as

$$V = \begin{bmatrix} .0014 & .0001 & .0004 \\ .0001 & .0011 & .0004 \\ .0004 & .0004 & .0015 \end{bmatrix} .$$

The overall test for homogeneity is $\underline{Q} = 82.12$ ($\underline{df} = 25$, $\underline{p} < .0005$). The six samples do not seem to share a common correlation matrix. Thus again the interpretation of r. as a shared or common population matrix seems unwarranted.

Imputing Unconditional Means

In this analysis mean values were substituted for the two missing correlations. For our example, the value $\underline{r}_{.1} = .373$ was substituted for $\underline{r}_{51}$ and $\underline{r}_{61}$. The variance of the mean ($\underline{V}_{11} = 0.0014$) was added to the computed values of $\hat{\sigma}_{511}$ and $\hat{\sigma}_{611}$. The GLS estimate based on all 10 samples, including the imputed data points, was the mean vector

$$(.363,\ .463,\ .228)'$$

with variance-covariance matrix

$$\begin{bmatrix} .0012 & .0001 & .0005 \\ .0001 & .0011 & .0005 \\ .0005 & .0005 & .0015 \end{bmatrix} .$$

The estimated variance of the first element of the correlation vector has only decreased from .0014 (in the available-cases analysis) to .0012, which corresponds to a standard error which is roughly six percent smaller (i.e., .035 versus .037), which would have only a small impact on inferential procedures. The homogeneity test value of 73.17 is significant when compared to the $\alpha = .05$ upper-tail critical value of the chi-square distribution with

$(k - 1)p^* = 27$ degrees of freedom.

Imputing Conditional Means

In the imputation of conditional means we again estimate the missing values using data from the eight cases which report all three correlations. The weighted regression of r_{i1} on r_{i2} and r_{i3} (weighting by the inverse of the variance of each r_{i1} value) for the eight samples with complete data gives the weighted regression model

$$\hat{r}_{i1} = 0.488 - 0.024\, r_{i2} + 0.062\, r_{i3} .$$

This model predicts values of $r_{51} = .391$ and $r_{61} = .398$. Our example data do not illustrate the potential advantages of this procedure well because r_{i1} is essentially unrelated to r_{i2} and r_{i3} .

Because the two unobserved values have been predicted from the other data, their covariances are computed as

$$\hat{\sigma}_{511} + v^*_{511} = .0141 + .0025 = .0166,$$

and

$$\hat{\sigma}_{611} + v^*_{611} = .0393 + .0058 = .0451.$$

Covariances involving $r_{\cdot 1}$ were computed using $r_{\cdot 1}$ in place of r_{i1} in formula (2).

For our data, the estimated mean correlation vector (using GLS estimation) is (.367, .461, .224)′, with variance-covariance matrix

$$\begin{bmatrix} .0012 & .0001 & .0005 \\ .0001 & .0011 & .0004 \\ .0005 & .0004 & .0015 \end{bmatrix} .$$

The test of homogeneity for this analysis is $Q = 71.99$, which is approximately distributed as a chi-square variable with 27 degrees of freedom. As above, the hypothesis of homogeneity is rejected.

Empirical Bayes Estimates

Next the mean vector and its variance-covariance matrix were estimated via the EM algorithm. We first imputed the values $\underline{r}_{51} = .391$ and $\underline{r}_{61} = .398$ (with estimated variances .0166 and .0451, respectively), using the method of imputing conditional means described above.

The starting values for the correlation vector and its variance covariance matrix were

$$\hat{\rho}^{(0)} = (.39,\ .42,\ .21)'$$

and

$$\hat{T}^{(0)} = \begin{bmatrix} .0006 & -.0005 & -.0054 \\ -.0005 & .0723 & -.0396 \\ -.0054 & -.0396 & .0146 \end{bmatrix} .$$

After 600 iterations the values of the mean correlations and their variance estimates were changing by less than 10^{-5}. The estimated mean vector was

$$(.393,\ .424,\ .226)'$$

with variance-covariance matrix

$$\hat{T}^{(600)} = \begin{bmatrix} .0004 & .0006 & .0001 \\ .0006 & .0619 & .0323 \\ .0001 & .0323 & .0170 \end{bmatrix} .$$

The parameters representing the relationship of SAT-M with SAT-V (the ρ_{i2} values) showed the most variability, with a standard error of nearly 0.25 i.e., the square root of the diagonal element .0619). The correlations between SAT-V and spatial ability also showed considerable parameter variation, with a standard error of 0.13.

The empirical Bayes estimates of the individual study parameters after 600 iterations are shown in Table 3. These values can be compared to the original sample correlations. The minimal amount of variation in the SAT-M -- spatial ability correlations has led to very similar estimates of $\hat{\rho}_{i1}$ for the ten samples. Values of $\underline{r}_{i2}$, which showed considerable variability, produced more dispersed values of $\hat{\rho}_{i2}$.

Insert Table 3 about here

Conclusions

Missing or unreported study results are an impediment to thorough reviews of any research literature. The problem of unreported correlation values is pervasive in research reviews which attempt to synthesize results of complete correlation matrices, especially matrices which involve more than a few variables. The methods reported here, particularly the empirical Bayes estimation procedures, should enable researchers to accomplish reasonable initial analyses in situations wherein the unreported values appear to be missing at random or simply not studied. Further work is needed to explore cases in which the missing-data mechanism is more complicated (e.g., involving truncation) in which the data are unlikely to be missing at random.

References

Becker, B. J. (1978). The relationship of spatial ability to sex differences in the performance of mathematically precocious youths on the mathematical sections of the Scholastic Aptitude Test. Unpublished master's thesis, Johns Hopkins University.

Becker, B. J. (1992). Models of science achievement: Factors affecting male and female performance in school science. In T.D. Cook, H.M. Cooper, D.S. Cordray, H. Hartmann, L.V. Hedges, T. Louis, & F. Mosteller (Eds.) *Towards explanatory meta-analysis*. New York: Russell Sage Foundation.

Becker, B. J. (in press b). Using results from replicated studies to estimate linear models. *Journal of Educational Statistics*.

Berry, P. C. (1957). *An exploration of the interrelationships among some non-intellectual predictors of achievement in problems solving*. Technical Report No. 4). New Haven: Yale University, Department of Industrial Administration and Psychology.

Buck, S. F. (1960). A method of estimation of missing data in multivariate data suitable for use with an electronic computer. *Journal of the Royal Statistical Society*, *B*, *22*, 302-306.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1978). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, *B*, *39*, 1-38.

Dempster, A. P., Rubin, D. B., & Tsutakawa, R. (1981). Estimation in covariance components models. *Journal of the American Statistical Association*, *76*, 341-353.

Friedman, L. (In press). Meta-analytic contributions to the study of gender differences. *International Journal of Educational Research*.

Hedges, L. V. (1983). Combining independent estimators in research synthesis. British Journal of Educational Psychology, 36, 123-131.

Hedges, L. V., & Olkin, I. (1985). Statistical methods for meta-analysis. New York: Academic Press.

Little, R. J. A., & Rubin, D. B. (1987). Statistical analysis with missing data. New York: Wiley.

Olkin, I., & Siotani, M. (1976). Asymptotic distribution of functions of a correlation matrix. In S. Ideka (Ed.) Essays in probability and statistics. Tokyo: Sinko Tsusho.

Rao, C. R. (1971). Linear statistical inference. New York: Wiley.

Rosenberg, J. H. (1981). The ability of selected cognitive, affective, and educational variables to predict the presence of anxiety related to mathematics. Unpublished doctoral dissertation, University of Connecticut.

Searle, S. R. (1971). Linear models. New York: Wiley.

Seber, G. A. F. (1977). Linear regression analysis. New York: Wiley.

Weiner, N. C. (1984). Cognitive aptitudes, personality variables, and gender difference effects on mathematical achievement for mathematically gifted students. Unpublished doctoral dissertation, Arizona State University.

Table 1

Variables Measured in Example Studies

	Measures		
Study	Math	Verbal	Spatial ability
Becker	SAT-M	SAT-V	Differential Aptitude Tests: Space Relations
Berry	SAT-M	SAT-V	Thurstone and Jeffrey Concealed Figures Test
Rosenberg	SAT-M	SAT-V	Differential Aptitude Tests: Space Relations
Weiner	SAT-M	SAT-V	Group Embedded Figures Test

Table 2

Sample Sizes and Correlations for Example Data

			Correlations		
			SAT-M	SAT-M	SAT-V
Sample id	Sample	Sample size	Spatial	SAT-V	Spatial
1	Becker 1 (1978)	$n_1 = 74$	.47	-.21	-.15
2	Becker 2 (1978)	$n_2 = 153$	.28	.19	.18
3	Berry 1 (1957)	$n_3 = 48$	.48	.41	.26
4	Berry 2 (1957)	$n_4 = 55$	.37	.40	.27
5	Rosenberg 1 (1980)	$n_5 = 51$	r_{51}	.48	.23
6	Rosenberg 2 (1980)	$n_6 = 18$	r_{61}	.74	.44
7	Weiner 1 (1984)	$n_7 = 27$	.26	.72	.36
8	Weiner 2 (1984)	$n_8 = 43$	.32	.52	.10
9	Weiner 3 (1984)	$n_9 = 35$	.58	.64	.40
10	Weiner 4 (1984)	$n_{10} = 34$	.34	.28	-.03

Table 3

Population Correlations for Example Data Estimated using EM Algorithm

		Correlations		
		$\hat{\rho}_{i1}$	$\hat{\rho}_{i2}$	$\hat{\rho}_{i3}$
		SAT-M	SAT-M	SAT-V
Sample id	Samp'e	Spatial	SAT-V	Spatial
1	Becker 1 (1978)	.392	-.099	-.049
2	Becker 2 (1978)	.381	.244	.136
3	Berry 1 (1957)	.396	.402	.214
4	Berry 2 (1957)	.391	.418	.224
5	Rosenberg 1 (1980)	.394	.469	.249
6	Rosenberg 2 (1980)	.394	.706	.374
7	Weiner 1 (1984)	.392	.722	.383
8	Weiner 2 (1984)	.395	.493	.261
9	Weiner 3 (1984)	.402	.583	.306
10	Weiner 4 (1984)	.393	.303	.162

Note

1. For instance, Becker (1992) found correlations among *r*s ranging from small negative to large positive values in a synthesis of predictors of science achievement. Correlations were as large .82 between *r*s which represented similar relationships.