LEAST-SQUARES APPROXIMATION OF AN IMPROPER CORRELATION MATRIX BY A PROPER ONE

DIRK L. KNOL
UNIVERSITY OF TWENTE

JOS M. F. TEN BERGE
UNIVERSITY OF GRONINGEN

An algorithm is presented for the best least-squares fitting correlation matrix approximating a given missing value or improper correlation matrix. The proposed algorithm is based upon a solution for Mosier's oblique Procrustes rotation problem offered by ten Berge and Nevels. A necessary and sufficient condition is given for a solution to yield the unique global minimum of the least-squares function. Empirical verification of the condition indicates that the occurrence of non-optimal solutions with the proposed algorithm is very unlikely. A possible drawback of the optimal solution is that it is a singular matrix of necessity. In cases where singularity is undesirable, one may impose the additional nonsingularity constraint that the smallest eigenvalue of the solution be $\delta$, where $\delta$ is an arbitrary small positive constant. Finally, it may be desirable to weight the squared errors of estimation differentially. A generalized solution is derived which satisfies the additional nonsingularity constraint and also allows for weighting. The generalized solution can readily be obtained from the standard "unweighted singular" solution by transforming the observed improper correlation matrix in a suitable way.

Key words: missing value correlation, tetrachoric correlation, indefinite correlation matrix, constrained least-squares approximation.

When product-moment correlations of a set of $n$ variables are computed by any of the missing value correlation methods described by Frane (1978), then it may happen that the resulting missing value correlation matrix is indefinite, and hence improper. This can be a serious problem in various multivariate data analysis techniques, for example, in regression and factor analysis.

One possible approach to this problem consists of avoiding an (indefinite) improper correlation matrix entirely by estimating the missing data themselves. Missing data can be estimated by maximum likelihood estimation from incomplete data (Beale & Little, 1975; Dempster, Laird & Rubin, 1977; Orchard & Woodbury, 1972) and by pragmatic procedures (Frane, 1976, 1978; Gleason & Staelin, 1975; Timm, 1970).

Another possible approach to the problem is to render the improper correlation matrix nonnegative definite by some smoothing procedure (Devlin, Gnanadesikan & Kettenring, 1975, p. 543; Dong, 1985; Frane, 1978).

The purpose of the present paper is to offer a least-squares smoothing procedure. That is, one may seek the best fitting (in the sense of least-squares) symmetric, unit-diagonal, nonnegative definite matrix $G$ to the given improper correlation matrix $R$. Specifically, the function

$$e(G) \equiv \frac{1}{2} \operatorname{tr} (G - R)^2 \tag{1}$$

Requests for reprints should be sent to Dirk Knol, University of Twente, Department of Education, PO Box 217, 7500 AE Enschede, THE NETHERLANDS or to Jos ten Berge, University of Groningen, Vakgroep Psychologie, Grote Markt 31-32, 9712 HV Groningen, THE NETHERLANDS.
can be minimized subject to the constraints \( G = G' \), \( \text{Diag}(G) = I_n \) and \( G \geq 0 \). For convenience we write \( Y \geq 0 \) and \( Y > 0 \) to denote that a symmetric matrix \( Y \) is nonnegative definite and positive definite, respectively.

The minimization problem (1) can be generalized in three ways. Firstly, the problem can be applied to any improper correlation matrix, for example, an indefinite tetrachoric correlation matrix or a correlation matrix obtained by element-wise robust estimation (Devlin, Gnanadesikan, & Kettenring, 1975, 1981; Gnanadesikan & Kettenring, 1972). Secondly, the problem can be generalized to handle indefinite matrices with fixed diagonal elements not necessary equal to one. For example, the scope of the problem can be extended to missing value covariance matrices with known variances or to product-moment correlation matrices with known communalities. Thirdly, it is possible to exclude those product-moment correlations or covariances which are computed between complete variables (no missing values) from the minimization procedure. That is, the excluded elements of \( R \) can be held constant in (1). Without loss of generality these elements can be collected in the \( n_1 \times n_1 \) (\( 0 \leq n_1 < n \)) submatrix \( R_{11} > 0 \) of \( R \), where \( R \) is partitioned as

\[
R = \begin{bmatrix}
R_{11} & R_{12} \\
R_{21} & R_{22}
\end{bmatrix}
\]

In order to incorporate these three generalizations, we shall address the generalized problem of minimizing (1) subject to the constraints

\[
G = G',
\]

\[
G \geq 0,
\]

\[
G_{11} = R_{11} \geq 0
\]

and

\[
\text{Diag}(G_{22}) = \text{Diag}(R_{22}) \geq 0,
\]

where \( G \) is partitioned as

\[
G = \begin{bmatrix}
G_{11} & G_{12} \\
G_{21} & G_{22}
\end{bmatrix},
\]

and \( G_{11} \) is of order \( n_1 \times n_1 \). Note that the constraints (4) and (5) for the problems with \( n_1 = 0 \) and \( n_1 = 1 \) are equivalent. In the next section a computational solution will be offered for the generalized problem of minimizing (1) subject to the constraints (2) through (5).

**An Algorithm**

The constraints \( G = G' \) (2) and \( G \geq 0 \) (3) can equivalently be expressed by the constraint

\[
G = AA',
\]

for some \( n \times m \) (\( n_1 \leq m \leq n \)) matrix \( A \). Consider the partitioning

\[
A = \begin{bmatrix}
A_1 \\
A_2
\end{bmatrix} = \begin{bmatrix}
A_{11} & A_{12} \\
A_{21} & A_{22}
\end{bmatrix},
\]