COMMUN. STATIST.—THEORY METH. 22(4), 965-984 (1993)

TRANSFORMATION OF NON POSITIVE SEMIDEFINITE CORRELATION MATRICES

Peter J. Rousseeuw and Geert Molenberghs

Department of Mathematics and Computer Science
Universitaire Instelling Antwerpen
Universiteitsplein 1
B-2610 Wilrijk, Belgium

Key Words and Phrases: eigenvalue method; missing data; multidimensional scaling; multivariate probit model; robust correlations; shrinking.

ABSTRACT

In multivariate statistics, estimation of the covariance or correlation matrix is of crucial importance. Computational and other arguments often lead to the use of coordinate-dependent estimators, yielding matrices that are symmetric but not positive semidefinite. We briefly discuss existing methods, based on shrinking, for transforming such matrices into positive semidefinite matrices. A simple method based on eigenvalues is also considered. Taking into account the geometric structure of correlation matrices, a new method is proposed which uses techniques similar to those of multidimensional scaling.

Copyright © 1993 by Marcel Dekker, Inc.

1. INTRODUCTION

In different areas of applied statistics, the estimation of covariance and correlation matrices is of major interest. Given a data set of column vectors $Y_k = (y_{k1}, \ldots, y_{kp})^T$ for $k = 1, \ldots, n$, the sample product-moment covariance matrix is given by

$$C = \frac{1}{n-1} \sum_{k=1}^{n} (Y_k - \overline{Y})(Y_k - \overline{Y})^T, \tag{1.1}$$

where $\overline{Y}$ denotes the sample mean vector. This matrix is always positive semidefinite (PSD). When $n > p$ it is even positive definite (PD) almost surely, assuming that the observations are i.i.d. and drawn from a continuous nondegenerate distribution. The corresponding correlation matrix $\boldsymbol{R}$ enjoys the same properties. It is defined by

$$\boldsymbol{R} = \boldsymbol{DCD}, \tag{1.2}$$

where $\boldsymbol{D}$ is the diagonal matrix with diagonal elements $1/\sqrt{c_{ii}}$ $(i = 1, \ldots, p)$. Therefore, the diagonal elements of $\boldsymbol{R}$ are all equal to 1. The off-diagonal elements r_{ij} of $\boldsymbol{R}$ reflect the correlation between the variables i and j. Note that the PSD property of $\boldsymbol{R}$ implies, among other things, that all $|r_{ij}| \leq 1$. However, for alternative estimators, the PSD property is not necessarily satisfied. Three key examples are presented.

Example 1. It is well-known that the estimators C and $\boldsymbol{R}$, defined by (1.1) and (1.2), are very sensitive to outliers. For that reason they are sometimes replaced by *robust estimators.* Devlin, Gnanadesikan and Kettenring (1975) propose several alternatives. Some techniques are known to yield a robust and PSD estimate $\boldsymbol{R}^*$ for the correlation matrix. Examples are multivariate trimming and principal component trimming. On the other hand, some techniques are coordinate-dependent in the sense that they study all pairs of variables in a bivariate way. Examples are (1) the quadrant correlation estimate, (2) a transformation of Kendall's τ and (3) a correlation estimator obtained from sums and differences of standardized observed values. For multivariate data, the pairwise correlations are patched together to form a robust estimate $\boldsymbol{R}^*$.

An important advantage of such coordinate-dependent methods is their relatively low computational cost, but the main problem is that $\boldsymbol{R}^*$ need not be PSD.

Example 2. A different motivation arises in the case of *missing values.* When some data values y_{kj} (for $k = 1, \ldots, n$ and $j = 1, \ldots, p$) are missing, there are several approaches to estimate the correlation matrix. For instance, one can attempt a complete cases analysis in which only the complete observations (i.e. the vectors without missing components) are taken into account. A disadvantage of this approach is that the number of complete observations may be small or even zero. (For a thorough discussion of mechanisms generating missing values, see Little and Rubin 1987). Another approach is to consider all vectors for which a particular pair of components is observed, i.e. the estimate r_{ij} involves all vectors Y_k of which both y_{ki} and y_{kj} are observed. In this case the estimates may fail to constitute a PSD correlation matrix.

Example 3. Another illustration is given by the *multivariate probit model* (Ashford and Sowden (1970), Pemberton (1984), Lesaffre and Molenberghs (1991)). For each subject in a trial, a vector of ordered categorical variables is recorded, together with a vector of explanatory variables. The outcome vector is considered to be a discrete realization of an underlying multivariate standard normal distribution, the location of which is determined by the covariate vector. The coefficients for the covariates and the correlation matrix of the normal distribution can be estimated by the maximum likelihood method. The major problem with this technique is the computation of the multivariate normal integral, as well as its first order derivatives. Pemberton (1984) proposed a sub-optimal method to circumvent this problem. The coefficients for the covariates are determined one at a time by means of probit regression, while the correlation coefficients are determined in a bivariate fashion. Thus, only the bivariate normal integral is needed. The problem is that the estimated correlations do not necessarily form a PSD matrix.

As in these three examples, coordinate-dependent estimators of covariance and correlation often yield what we will call *pseudo-covariance matrices* and

pseudo-correlation matrices, which are not necessarily PSD. A p-by-p matrix C will be called a pseudo-covariance matrix when

(1) C is symmetric;

(2) the diagonal elements of C satisfy $c_{ii} > 0$;

(3) the off-diagonal elements satisfy $c_{ij}^2 \leq c_{ii}c_{jj}$.

The motivation for conditions (1) and (2) is clear. Condition (3) would follow immediately if C were PSD, as can be seen from $y^T C y \geq 0$ for all vectors y consisting of zeroes except at the i-th and j-th positions. From now on, we will assume (3) to hold (for instance, we can replace c_{ij} by $\mathrm{sgn}(c_{ij})\sqrt{\min(c_{ij}^2, c_{ii}c_{jj})}$ if necessary). However, having a pseudo-covariance matrix is not enough, because most multivariate statistical methods require PSD matrices. For instance, if C is not PSD it can happen that

$$x^T C^{-1} x < 0$$

for some vector x, which makes it impossible to use Mahalanobis-type distances or statistics based on them. Therefore, our goal will be to transform C into a PSD matrix which in some sense is "close" to C.

If we multiply C as in (1.2) we obtain a pseudo-correlation matrix $\boldsymbol{R}$, that is, a square matrix satisfying the corresponding conditions:

(1) $\boldsymbol{R}$ is symmetric;

(2) all $r_{ii} = 1$;

(3) all $|r_{ij}| \leq 1$.

Even if $\boldsymbol{R}$ is built from pairwise correlation estimates, condition (3) is not necessarily satisfied. For instance, Gnanadesikan (1977, page 131) considers a robust correlation coefficient (formula (69) in his book) which does not necessarily belong to $[-1, 1]$, which motivates him to propose a version that does. (In cases where such an alternative estimate is not available, one can still re-

sort to truncating r_{ij} at $+1$ and -1.) From now on we will assume (1), (2), and (3) to hold, except for the approach in Section 4 which still works when these conditions are not satisfied.

In the remainder of this paper we will restrict attention to methods for transforming the pseudo-correlation matrix $\boldsymbol{R}$ into a true correlation matrix $\tilde{\boldsymbol{R}}$ (that is, one which is PSD). The same methods can also be used in the case of a pseudo-covariance matrix $\boldsymbol{C}$ because we can apply (1.2) yielding $\boldsymbol{R}$, transform the latter into $\tilde{\boldsymbol{R}}$, and then compute $\tilde{\boldsymbol{C}} = \boldsymbol{D}^{-1}\tilde{\boldsymbol{R}}\boldsymbol{D}^{-1}$.

Section 2 gives a brief overview of existing techniques for transforming $\boldsymbol{R}$. The *nonlinear shrinking* technique proposed by Devlin, Gnanadesikan and Kettenring (1975) uses a contracting function to transform the individual elements of $\boldsymbol{R}$. The next two sections present alternative methods that aim to take the multivariate nature of a correlation matrix into account. The *eigenvalue method* in Section 3 is based on the property that a PSD matrix has nonnegative eigenvalues. In Section 4 we propose the *scaling method* which represents correlation by means of inner products between unit vectors, thereby linearizing the problem. It is the most generally applicable among the methods considered, and allows for the use of weights if wanted.

2. SHRINKING METHODS

The first approach for transforming a pseudo-correlation matrix into a PSD matrix is due to Devlin et al. (1975). They start by mentioning the *linear shrinking* technique in which the pseudo-correlation matrix $\boldsymbol{R}$ is shrinked towards the p-dimensional identity matrix I_p according to

$$\tilde{\boldsymbol{R}} = \lambda \boldsymbol{R} + (1 - \lambda)\boldsymbol{I}_p \tag{2.1}$$

where λ is the largest value in $[0, 1]$ which makes (2.1) PSD. They note that one can also shrink $\boldsymbol{R}$ towards another PSD matrix, e.g. towards a different robust estimator $\boldsymbol{R}_1$.

In the same paper, the authors emphasize a *nonlinear shrinking* technique, which goes as follows. Each off-diagonal element r in $\boldsymbol{R}$ is replaced by $\tilde{r}$ via

the transformation procedure

$$\tilde{r} = \begin{cases} f^{-1}(f(r) + \Delta) & \text{if } r < -f^{-1}(\Delta), \\ 0 & \text{if } |r| \leq f^{-1}(\Delta), \\ f^{-1}(f(r) - \Delta) & \text{if } r > f^{-1}(\Delta). \end{cases} \tag{2.2}$$

Here Δ is a small positive constant (e.g. 0.05) and f is a monotone increasing continuous function $f : [-\infty, \infty] \to [-1, 1]$. This procedure is repeated until the resulting matrix is PSD. Devlin et al. suggest for f

$$\begin{cases} f_1(x) &= \dfrac{e^x - e^{-x}}{e^x + e^{-x}} = \tanh(x) \\ f_1^{-1}(x) &= \dfrac{1}{2} \ln\left(\dfrac{1+x}{1-x}\right) = \tanh^{-1}(x). \end{cases} \tag{2.3}$$

Note that f_1^{-1} in (2.3) is Fisher's variance-stabilizing transform for a correlation coefficient. An alternative transformation function is given by :

$$\begin{cases} f_2(x) &= \dfrac{2}{\pi} \arctan(x) \\ f_2^{-1}(x) &= \tan\left(\dfrac{\pi}{2} x\right). \end{cases} \tag{2.4}$$

We illustrate the method on an example. Consider the non-PSD matrix $\boldsymbol{R}$ below,

$$\boldsymbol{R} = \begin{pmatrix} 1 & -0.9 & -0.9 \\ . & 1 & 0.3 \\ . & . & 1 \end{pmatrix} \tag{2.5}$$

and let $\tilde{\boldsymbol{R}}_i$ be the matrix obtained by applying nonlinear shrinking using transformation f_i with $\Delta = 0.05$:

$$\tilde{\boldsymbol{R}}_1 = \begin{pmatrix} 1 & -0.719 & -0.719 \\ . & 1 & 0.194 \\ . & . & 1 \end{pmatrix} \quad \tilde{\boldsymbol{R}}_2 = \begin{pmatrix} 1 & -0.767 & -0.767 \\ . & 1 & 0.216 \\ . & . & 1 \end{pmatrix}.$$

The result of nonlinear shrinking depends on the choice of f (as seen by comparing $\tilde{\boldsymbol{R}}_1$ with $\tilde{\boldsymbol{R}}_2$) and on the value of Δ. For instance, the matrix $\tilde{\boldsymbol{R}}_1$ is obtained when $\Delta = 0.05$ or $\Delta = 0.1$, but not for $\Delta = 0.075$. In the latter case the method yields

$$\tilde{R}_3 = \begin{pmatrix} 1 & -0.760 & -0.760 \\ . & 1 & 0.220 \\ . & . & 1 \end{pmatrix}. \tag{2.6}$$

In the above example, the linear shrinking solution becomes

$$\tilde{R}_4 = \begin{pmatrix} 1 & -0.795 & -0.795 \\ . & 1 & 0.265 \\ . & . & 1 \end{pmatrix} \tag{2.7}$$

which is computed from (2.1) with $\lambda = 0.8837$. In this example, the result of linear shrinking is quite close to the original matrix.

The shrinking techniques have the advantage of being easy to apply, because (2.1) and (2.2) operate on each correlation separately. On the other hand, this has the drawback that, except when testing whether the result is PSD, the relation between the correlations is not used. Therefore, we think there is room for the consideration of alternative methods that make use of the structure of the entire correlation matrix. In the following sections we will describe two such transformation methods: the eigenvalue method and the scaling method.

3. THE EIGENVALUE METHOD

A true correlation matrix $\boldsymbol{R}$ is PSD, and hence can be written as

$$\boldsymbol{R} = \boldsymbol{P}\boldsymbol{D}\boldsymbol{P}^T \tag{3.1}$$

where $\boldsymbol{D}$ is a diagonal matrix containing the (nonnegative) eigenvalues of $\boldsymbol{R}$, and $\boldsymbol{P}$ is an orthogonal matrix (i.e. $\boldsymbol{P}\boldsymbol{P}^T = \boldsymbol{P}^T\boldsymbol{P} = \boldsymbol{I}_p$) of corresponding eigenvectors. Suppose the eigenvalues appear in decreasing order. Let $\tilde{Y}_i$ be standardized versions of the initial variables. Then an orthonormal set of eigenvectors Z_j corresponds to linear combinations of the $\tilde{Y}_i$. These linear

combinations are uncorrelated, and have decreasing variance. In case $\boldsymbol{R}$ is PD, the eigenvalues are strictly positive.

When $\boldsymbol{R}$ is symmetric but not PSD, then (3.1) still holds, but now some eigenvalues are negative. Typically, the negative eigenvalues will not have a large absolute value. A natural approach consists of replacing the negative eigenvalues by zeroes, if the result merely has to be PSD, or by a small positive number, Δ say, if PD is required. Alternatively, the negative eigenvalues could be replaced by their absolute value. In either case $\boldsymbol{D}$ is replaced by $\boldsymbol{D}'$, and we compute $\boldsymbol{R}' = \boldsymbol{P}\boldsymbol{D}'\boldsymbol{P}^T$. The diagonal elements of $\boldsymbol{R}'$ will not necessarily be equal to 1. Therefore, we transform $\boldsymbol{R}'$ to $\tilde{\boldsymbol{R}} = \boldsymbol{D}_1\boldsymbol{R}'\boldsymbol{D}_1$, where $\boldsymbol{D}_1$ is the diagonal matrix with diagonal elements $1/\sqrt{r'_{jj}}$ $(j = 1, \ldots, p)$.

The eigenvalue-eigenvector decomposition in (3.1) can easily be carried out with a standard mathematical-statistical software package (e.g. GAUSS). These packages usually give the eigenvalues and an arbitrary set of linearly independent eigenvectors. The latter then have to be normalized to unit length. Furthermore, the orthogonality of the eigenvectors has to be enforced. Eigenvectors corresponding to different eigenvalues are automatically orthogonal. Under the usual assumptions, the probability that two or more equal eigenvalues appear is zero. In practice, whenever eigenvalues are tied, the corresponding set of eigenvectors has to be replaced by an orthonormal set (using the Gram-Schmidt procedure).

The eigenvalue method can also be implemented by means of a different algorithm giving the same results. Instead of using (3.1), one can use the so-called *singular value decomposition* (see Seber 1984) :

$$\boldsymbol{R} = \boldsymbol{V}\boldsymbol{X}\boldsymbol{W}^T \tag{3.2}$$

where $\boldsymbol{X}$ is a diagonal matrix containing the singular values of $\boldsymbol{R}$. (In the case of a symmetric matrix, the latter correspond to the absolute values of the eigenvalues.) The matrices $\boldsymbol{V}$ and $\boldsymbol{W}$ contain the same column vectors, possibly up to sign. They are also eigenvectors of $\boldsymbol{R}$. The same technique as above can be applied to $\boldsymbol{X}$. Now, the singular values corresponding to eigenvectors $V_j = -W_j$ are to be replaced by either zero or $-\Delta$.

Applying the eigenvalue method to the example matrix $\boldsymbol{R}$ of (2.5) leads to

$$\tilde{\boldsymbol{R}}_5 = \begin{pmatrix} 1 & -0.781 & -0.781 \\ . & 1 & 0.327 \\ . & . & 1 \end{pmatrix}. \tag{3.3}$$

Whereas the nonlinear shrinking method shrinks each element towards zero, this is not the case for the eigenvalue method. In the example, the absolute values of r_{12} and r_{13} decrease, but that of r_{23} increases. Also note that the changes are smaller with the eigenvalue method than with nonlinear shrinking.

If two elements r_{ij} and r_{kl} coincide in the original matrix $\boldsymbol{R}$, then the nonlinear shrinking technique will by definition transform them to elements $\tilde{r}_{ij}$ and $\tilde{r}_{kl}$ that are also equal to each other. (In the above example, this happens with $r_{12} = r_{13}$.) On the other hand, the eigenvalue method may or may not yield $\tilde{r}_{ij} = \tilde{r}_{kl}$ depending on the structure of $\boldsymbol{R}$. More information on this can be found in the Appendix.

4. THE SCALING METHOD

In general we can ask the following question. Given a matrix $\boldsymbol{R}$ of estimated correlations, find a "true" correlation matrix $\tilde{\boldsymbol{R}}$ (that is, a pseudo-correlation matrix which is PSD) as close as possible to $\boldsymbol{R}$. In order to formalize "as close as possible" we need to minimize a criterion which reflects distance between matrices. A possible criterion is

$$S = \sum_{i=1}^{p} \sum_{j=1}^{p} (r_{ij} - \tilde{r}_{ij})^2 \tag{4.1}$$

which is related to the *stress function* in nonmetric multidimensional scaling (see Seber 1984). By analogy, we will refer to this approach as the scaling method.

In this setup the elements $\tilde{r}_{ij}$ of $\tilde{\boldsymbol{R}}$ are determined by minimization of (4.1), which is a (squared) euclidean distance. Therefore $\tilde{\boldsymbol{R}}$ is the closest

approximation to $\boldsymbol{R}$ in the set of PSD pseudo-correlation matrices. Because the latter is a compact and convex subset of $\mathbb{R}^{p\times p}$ it follows that $\tilde{\boldsymbol{R}}$ always exists and is unique, and that the transformation $\boldsymbol{R} \rightarrow \tilde{\boldsymbol{R}}$ is a continuous mapping.

Note that $\tilde{\boldsymbol{R}}$ is well-defined for *any* matrix $\boldsymbol{R}$ in $\mathbb{R}^{p\times p}$, which does not itself need to be a pseudo-correlation matrix (all three conditions may be violated). This makes the method much more general than those in the preceding sections. (In the special case where $\boldsymbol{R}$ is already a pseudo-correlation matrix, we can simplify (4.1) somewhat by restricting attention to the terms with $i < j$.)

From the computational viewpoint, minimizing (4.1) seems intractable at first because of the constraint that $\tilde{\boldsymbol{R}}$ be a PSD pseudo-correlation matrix, which involves all elements $\tilde{r}_{ij}$ simultaneously. However, this problem can be overcome by describing $\tilde{\boldsymbol{R}}$ in a geometric way. We can think of $\tilde{\boldsymbol{R}}$ as containing the correlation between p variables, each of which may be assumed to have unit variance. These variables can then be represented as unit vectors U_i in $\mathbb{R}^p$ $(i = 1, \ldots, p)$ in such a way that their inner products are the elements of the correlation matrix $\tilde{\boldsymbol{R}}$:

$$\tilde{r}_{ij} = U_i^T U_j. \tag{4.2}$$

If we combine the column vectors U_i in a matrix $U = (U_1, \ldots, U_p)$ then this fact is summarized by

$$\tilde{\boldsymbol{R}} = \boldsymbol{U}^T\boldsymbol{U}. \tag{4.3}$$

Such a matrix $\boldsymbol{U}$ always exists because $\tilde{\boldsymbol{R}}$ is PSD. (Note that because $\tilde{U}$ is a root of $\tilde{\boldsymbol{R}}$, this linearizes the problem of finding $\tilde{\boldsymbol{R}}$.) By Cholesky's theorem we may assume that $\boldsymbol{U}$ is an upper triangular matrix with nonnegative diagonal elements (see Appendix). This implies that only the first i components of U_i can be nonzero. Moreover, U_i contains only $i - 1$ free parameters because $||U_i||^2 = 1$. An elegant representation of these vectors is provided by using goniometric functions with the free parameters as angles. Indeed, U_i can be seen as a vector in $\mathbb{R}^i$ with unit length, and hence can be expressed in spherical coordinates using $i-1$ angles $\theta_{1i}, \theta_{2i}, \ldots, \theta_{i-1,i}$. We can combine the $i-1$ angles

of each U_i in an upper triangular matrix

$$\begin{pmatrix} 0 & \theta_{12} & \theta_{13} & \theta_{14} & \dots & \theta_{1,p-1} & \theta_{1p} \\ & 0 & \theta_{23} & \theta_{24} & \dots & \theta_{2,p-1} & \theta_{2p} \\ & & 0 & \theta_{34} & \dots & \theta_{3,p-1} & \theta_{3p} \\ & & & & \dots & & \\ & & & & & 0 & \theta_{p-1,p} \\ & & & & & & 0 \end{pmatrix} \tag{4.4}$$

where we can add dummy angles $\theta_{ii} = 0$ for convenience. Conversely, from the matrix (4.4) we can reconstruct the matrix $\boldsymbol{U}$, the i-th column of which contains the cartesian coordinates of U_i. Therefore,

$$U_{ji} = \begin{cases} \sin\theta_{i1} \sin\theta_{i2} \dots \sin\theta_{i,i-j} \cos\theta_{i,i-j+1} & \text{if } j \le i, \\ 0 & \text{if } j > i. \end{cases} \tag{4.5}$$

By minimizing S as a function of the parameters θ_{ji} in (4.4) we have incorporated the constraint that $\tilde{\boldsymbol{R}}$ be a PSD pseudo-correlation matrix. An unconstrained optimization remains. Because S is differentiable with respect to the θ_{ji} we can apply iterative numerical algorithms such as the steepest descent method, which is fast and does not require second order derivatives. Good starting values are given by $\theta_{ji} = 0$ for all i and j, corresponding to the identity correlation matrix. In our implementation we used the GAUSS library OPTMUM. We applied the algorithm to a large set of randomly generated pseudo-correlation matrices, without encountering numerical problems. On a PC with 386 processor, the solution was found in a matter of seconds for matrices with dimension $p = 3, \dots, 7$.

Applying the scaling method with criterion (4.1) to the matrix $\boldsymbol{R}$ of (2.5) leads to

$$\tilde{\boldsymbol{R}}_6 = \begin{pmatrix} 1 & -0.821 & -0.821 \\ . & 1 & 0.348 \\ . & . & 1 \end{pmatrix}. \tag{4.6}$$

The elements of $\boldsymbol{R}$ "move" in the same direction as they do when applying the eigenvalue method. In this example $\tilde{r}_{12} = \tilde{r}_{13}$ as in the original matrix $\boldsymbol{R}$. In the Appendix, general conditions on the structure of $\boldsymbol{R}$ are given under which the transformed $\tilde{\boldsymbol{R}}$ inherits such equalities.

Note that (4.1) is not the only possible minimization criterion. Suppose information is available about the precision of the entries of $\boldsymbol{R}$. For instance, when $\boldsymbol{R}$ is estimated from incomplete data, then the precision will be lower for the entries that are based on fewer observations. Assume that for each correlation a weight $w_{ij} = w(r_{ij})$ is given. (For instance, such a weight can arise as the inverse of the estimated variance on the estimated entry r_{ij}.) We can then use the weighted stress function

$$S_w = \sum_{i=1}^{p} \sum_{j=1}^{p} w_{ij} \left(r_{ij} - \tilde{r}_{ij}\right)^2 . \tag{4.7}$$

Let us apply the weighted scaling method to the matrix $\boldsymbol{R}$ of (2.5). Because $\boldsymbol{R}$ is symmetric, it is natural to consider symmetric weights ($w_{ij} = w_{ji}$). In this example, we take $w_{12} = w_{23} = 2w_{13}$ which yields

$$\tilde{\boldsymbol{R}}_7 = \begin{pmatrix} 1 & -0.840 & -0.793 \\ . & 1 & 0.335 \\ . & . & 1 \end{pmatrix} . \tag{4.8}$$

By comparing (4.6) with (4.8) we see that the elements with the highest weight change less than before, whereas the element with the lower weight changes more. Consequently, the elements $\tilde{r}_{12}$ and $\tilde{r}_{13}$ are no longer equal.

Finally, we look at an example where the original matrix is not a pseudo-correlation matrix. For instance,

$$\boldsymbol{T} = \begin{pmatrix} 0.8 & -0.9 & -0.9 \\ -1.2 & 1.1 & 0.3 \\ -0.8 & 0.4 & 0.9 \end{pmatrix} \tag{4.9}$$

is not symmetric, its diagonal elements are not exactly 1, and its off-diagonal elements are not all confined to $[-1, 1]$. Nevertheless the scaling method (4.1) still applies, yielding the PSD solution

$$\tilde{\boldsymbol{T}} = \begin{pmatrix} 1 & -0.908 & -0.759 \\ -0.908 & 1 & 0.416 \\ -0.759 & 0.416 & 1 \end{pmatrix}. \tag{4.10}$$

Had a set of (possibly asymmetric) weights been given, we could have applied criterion (4.7) also here.

5. DISCUSSION

Both the eigenvalue method and the scaling method are conceptually very simple. Furthermore, they are easy to apply and consume little computation time. For the eigenvalue method, this is because only a matrix decomposition is needed. In the case of the scaling method, we essentially work with roots of correlation matrices (corresponding to a geometric representation) thereby converting the quadratic problem to a linear one.

Whereas the shrinking methods operate on each correlation separately, the eigenvalue method and the scaling method deal with all correlations simultaneously. As a consequence, not all correlations will automatically decrease.

It is important to know which of these methods may yield a PD matrix. Linear shrinking always gives a singular PSD matrix (because the boundary of the set of PD matrices consists of PSD matrices). It can be made PD e.g. by multiplying the λ obtained from (2.1) by a constant, slightly smaller than 1. After nonlinear shrinking, the resulting $\tilde{\boldsymbol{R}}$ is PSD but not necessarily PD. One can then apply (2.2) one more time to make the result "sufficiently PD", more precisely to make it well-conditioned (i.e., with small enough condition number). When using the eigenvalue method, one can preset a small constant Δ to replace the nonpositive eigenvalues if PD is required. Applying the scaling method to non-PSD matrices again yields PSD but non-PD matrices. To obtain a PD matrix, one could first compute the scaling method solution and

then apply the eigenvalue method with given Δ to it. One might argue that in case a PD solution is required, it is easier to apply the eigenvalue method immediately. However, the scaling method has three major advantages :

- The eigenvalue method changes $\boldsymbol{R}$ only in the nonpositive eigenvalues, while the scaling method minimizes over the whole space. In general, the solution will be much closer to $\boldsymbol{R}$.

- The stress function S is quite flexible. It is possible to include extra information, such as reliability of individual correlations, by using weighting factors. This feature is shared by none of the other methods.

- The scaling method is more widely applicable because it also works when $\boldsymbol{R}$ is not a pseudo-correlation matrix. So $\boldsymbol{R}$ need not be symmetric, nor do its diagonal entries have to be precisely 1, nor must the off-diagonal elements be confined to $[-1, 1]$. An analogous result does not hold for the eigenvalue method, as asymmetric matrices can have complex eigenvalues.

APPENDIX

In this Appendix we investigate under which conditions the eigenvalue method and the scaling method transform equal correlations to equal correlations. Furthermore, it is shown how a unique Cholesky-type decomposition can be constructed for a singular PSD matrix.

Nonlinear shrinking always transforms equal correlations to the same value. We observed the same effect for the elements $r_{12} = r_{13}$ in the matrix $\boldsymbol{R}$ of (2.5) when applying the eigenvalue method and the scaling method. However, there are also counterexamples. For instance, consider the matrix $\boldsymbol{M}$ below,

$$\boldsymbol{M} = \begin{pmatrix} 1 & -0.9 & -0.9 & 0.2 \\ . & 1 & 0.3 & 0.5 \\ . & . & 1 & -0.4 \\ . & . & . & 1 \end{pmatrix} \tag{A1}$$

together with $\tilde{\boldsymbol{M}}_1$ given by the eigenvalue method and $\tilde{\boldsymbol{M}}_2$ obtained from the scaling method :

$$\tilde{\boldsymbol{M}}_1 = \begin{pmatrix} 1 & -0.765 & -0.814 & 0.158 \\ . & 1 & 0.320 & 0.456 \\ . & . & 1 & -0.409 \\ . & . & . & 1 \end{pmatrix} \quad \tilde{\boldsymbol{M}}_2 = \begin{pmatrix} 1 & -0.775 & -0.832 & 0.147 \\ . & 1 & 0.352 & 0.459 \\ . & . & 1 & -0.422 \\ . & . & . & 1 \end{pmatrix}.$$

In this example the equality $r_{12} = r_{13}$ is not preserved. This may be explained as follows. An element $\tilde{r}_{ij}$ produced by the eigenvalue method (resp. the scaling method) is not only based on the individual correlation r_{ij}, but on the available information about the variables Y_i and Y_j. For instance, the available information about Y_i includes all correlations r_{ik} for $k = 1, \ldots, p$. Thus, equal correlations $r_{ij} = r_{i'j'}$ will be transformed to equal correlations $\tilde{r}_{ij} = \tilde{r}_{i'j'}$ if the ordered pair of indices (i, j) is either interchangeable with (i', j') or with (j', i'). Assume the first situation. An example is given by a matrix $\boldsymbol{R}$ that is invariant under the following permutation on the indices :

$$\begin{aligned} i' &\to i \\ i &\to i' \\ j' &\to j \\ j &\to j' \\ k &\to k' \quad (k \neq i, i', j, j'). \end{aligned}$$

In general, assume there exists a permutation π with permutation matrix $\boldsymbol{P}_1$ such that

$$\pi(\boldsymbol{R}) = \boldsymbol{P}_1 \boldsymbol{R} \boldsymbol{P}_1^T = \boldsymbol{R}. \tag{A2}$$

A permutation matrix is a square matrix with a single 1 in every row and in every column, while the other elements are all equal to 0. Consequently, it always satisfies $P_1^{-1} = P_1^T$. The (i, j)-th entry of $\pi(\boldsymbol{R})$ is given by $r_{\pi(i)\pi(j)}$.

Let us first apply the eigenvalue method to $\boldsymbol{R}$. Application of (3.1) to $\boldsymbol{R}$ yields $\boldsymbol{PDP}^T$. Then the corresponding decomposition of $\pi(\boldsymbol{R})$ is given by

$$\pi(\boldsymbol{R}) = (\boldsymbol{P}_1\boldsymbol{P}\boldsymbol{P}_1^T)(\boldsymbol{P}_1\boldsymbol{D}\boldsymbol{P}_1^T)(\boldsymbol{P}_1\boldsymbol{P}^T\boldsymbol{P}_1^T).$$

The equality $\pi(\boldsymbol{R}) = \boldsymbol{R}$ thus becomes

$$(\boldsymbol{P}_1\boldsymbol{P}\boldsymbol{P}_1^T)(\boldsymbol{P}_1\boldsymbol{D}\boldsymbol{P}_1^T)(\boldsymbol{P}_1\boldsymbol{P}\boldsymbol{P}_1^T)^T = \boldsymbol{P}\boldsymbol{D}\boldsymbol{P}^T. \tag{A3}$$

Replacing an element d_{ii} by d'_{ii} in $\boldsymbol{D}$ on the right hand side corresponds to the same operation on $d_{\pi(i)\pi(i)}$ in $\boldsymbol{P}_1\boldsymbol{D}\boldsymbol{P}_1^T$ on the left hand side. Thus $\boldsymbol{D}$ may be replaced by $\boldsymbol{D}'$ on both sides. This yields

$$\boldsymbol{P}_1\boldsymbol{P}\boldsymbol{D}'\boldsymbol{P}^T\boldsymbol{P}_1^T = \boldsymbol{P}\boldsymbol{D}'\boldsymbol{P}^T$$

and thus

$$\boldsymbol{P}_1\boldsymbol{R}'\boldsymbol{P}_1^T = \boldsymbol{R}'.$$

Transforming the right hand side by $\boldsymbol{D}_1$, the diagonal matrix with diagonal elements $1/\sqrt{r'_{jj}}$, is equivalent to the same transformation on the left hand side with $\boldsymbol{P}_1\boldsymbol{D}_1\boldsymbol{P}_1^T$, which finally leads to

$$\boldsymbol{P}_1\tilde{\boldsymbol{R}}\boldsymbol{P}_1^T = \tilde{\boldsymbol{R}}.$$

The same result holds for the scaling method due to the following argument. If $r_{\pi(i)\pi(j)} = r_{ij}$ for all i and j, then

$$\begin{aligned} S &= \sum_{i=1}^{p}\sum_{j=1}^{p}(r_{ij} - \tilde{r}_{ij})^2 \\ &= \sum_{i=1}^{p}\sum_{j=1}^{p}(r_{\pi(i)\pi(j)} - \tilde{r}_{\pi(i)\pi(j)})^2 \\ &= \sum_{i=1}^{p}\sum_{j=1}^{p}(r_{ij} - \tilde{r}_{\pi(i)\pi(j)})^2. \end{aligned}$$

Because $\tilde{\boldsymbol{R}}$ is the unique minimizer of S, it follows that $\tilde{r}_{\pi(i)\pi(j)} = \tilde{r}_{ij}$ for all i and j.

As an example let us consider the matrix $\boldsymbol{Q}$ below,

$$\boldsymbol{Q} = \begin{pmatrix} 1 & -0.9 & 0.2 & 0.3 \\ . & 1 & 0.5 & -0.9 \\ . & . & 1 & 0.2 \\ . & . & . & 1 \end{pmatrix} \tag{A4}$$

together with $\tilde{Q}_1$ and $\tilde{Q}_2$ obtained with the eigenvalue method and the scaling method :

$$\tilde{Q}_1 = \begin{pmatrix} 1 & -0.706 & 0.133 & 0.343 \\ . & 1 & 0.361 & -0.706 \\ . & . & 1 & 0.133 \\ . & . & . & 1 \end{pmatrix} \quad \tilde{Q}_2 = \begin{pmatrix} 1 & -0.731 & 0.118 & 0.395 \\ . & 1 & 0.355 & -0.731 \\ . & . & 1 & 0.118 \\ . & . & . & 1 \end{pmatrix}.$$

Note that the equalities $q_{12} = q_{24}$ and $q_{13} = q_{34}$ are passed on to the transformed matrices, due to the fact that the original matrix $\boldsymbol{Q}$ is invariant under the permutation $\pi(1) = 4, \pi(2) = 2, \pi(3) = 3, \pi(4) = 2$ with permutation matrix

$$\boldsymbol{P}_1 = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}.$$

On the other hand, in the counterexample (A1) there exists no permutation which leaves the matrix $\boldsymbol{M}$ invariant. (Such a counterexample cannot occur for 3-by-3 matrices: whenever $r_{ij} = r_{ik}$ we can construct the permutation $\pi(i) = i, \pi(j) = k, \pi(k) = j$ in order to prove that $\tilde{r}_{ij} = \tilde{r}_{ik}$.)

Note that the above results can be extended to "signed" permutation matrices (in which some of the "1" may be replaced by "−1") in order to deal with equalities $r_{ij} = -r_{i'j'}$ in a similar way.

We now turn our attention to a Cholesky-type decomposition for a singular PSD matrix. The Cholesky decomposition for an arbitrary symmetric PD

matrix $\boldsymbol{R}$ is constructed as follows. Let $\boldsymbol{U}$ be an upper triangular matrix such that $\boldsymbol{R} = \boldsymbol{U}^T\boldsymbol{U}$, then the following equation holds (for $i \leq j$) :

$$r_{ij} = \sum_{k=1}^{i} u_{ki}u_{kj}, \tag{A5}$$

leading to

$$u_{ij}u_{ii} = r_{ij} - \sum_{k=1}^{i-1} u_{ki}u_{kj}.$$

As $\boldsymbol{R}$ is PD, every submatrix $\boldsymbol{R}_i$ formed by the first i rows and columns has a positive determinant, which equals

$$\det \boldsymbol{R}_i = \prod_{k=1}^{i} u_{kk}^2$$

and thus $u_{ii}^2 \neq 0$ holds for all i. To impose uniqueness one sets $u_{ii} > 0$, hence

$$u_{ii} = \sqrt{\det(\boldsymbol{R}_i)/\det(\boldsymbol{R}_{i-1})} = \sqrt{r_{ii} - \sum_{k=1}^{i-1} u_{ki}^2}.$$

The off-diagonal elements are then determined by

$$u_{ij} = (r_{ij} - \sum_{k=1}^{i-1} u_{ki}u_{kj})/u_{ii}.$$

In case $\boldsymbol{R}$ is a singular PSD matrix we proceed as follows. The elements u_{ij} are determined row by row. Assume $\text{rank}(R) = s$. As the eigenvalues are ordered in decreasing order, the last $p - s$ rows depend linearly on the first s rows. For the first s proceed as in the PD case. Then, set all $u_{ij} = 0$ for $i = s+1, \ldots, p; j = i, \ldots, p$. In case this procedure yields a valid solution, uniqueness is guaranteed. We will show that this choice still satisfies (A5). As row i $(i > s)$ is a linear combination of the first s rows, there exists a set of coefficients α_{it} $(t = 1, \ldots, s)$ such that

$$r_{ij} = \sum_{t=1}^{s} \alpha_{it}r_{tj} = \sum_{t=1}^{s} \alpha_{jt}r_{it} = \sum_{t=1}^{s} \alpha_{jt}r_{ti}. \tag{A6}$$

We first consider the elements r_{ij} with $i \leq s < j$. Replacing the r_{ti} by their expressions in terms of the elements of $\boldsymbol{U}$ yields

$$\sum_{k=1}^{i} u_{ki}u_{kj} = \sum_{t=1}^{s} \alpha_{jt} \sum_{k=1}^{t} u_{kt}u_{ki} = \sum_{k=1}^{s}\sum_{t=k}^{s} \alpha_{jt}u_{kt}u_{ki} = \sum_{k=1}^{i}\sum_{t=k}^{s} \alpha_{it}u_{kt}u_{ki}. \tag{A7}$$

The change from s to i in the last step simply rules out the terms u_{ki} that appear in the lower triangle of U as these are zero by definition. By considering these equations in turn for $i = 1, \ldots, s$ and using that the first s rows are independent, we arrive at

$$u_{kj} = \sum_{t=k}^{s} \alpha_{jt} u_{kt} \quad (k = 1, \ldots, s; j = s+1, \ldots, p). \tag{A8}$$

Now, for the r_{ij} with $i \geq s$ (A6) remains valid, and combining this with (A8) :

$$r_{ij} = \sum_{t=1}^{s} \alpha_{it} \sum_{k=1}^{t} u_{kt} u_{kj} = \sum_{k=1}^{s} \sum_{t=k}^{s} \alpha_{it} u_{kt} u_{kj} = \sum_{k=1}^{s} u_{ki} u_{kj}. \tag{A9}$$

On the other hand, by definition

$$r_{ij} = \sum_{k=1}^{\min(i,j)} u_{ki} u_{kj}. \tag{A10}$$

Comparing (A9) and (A10) we see that

$$\sum_{k=s+1}^{\min(i,j)} u_{ki} u_{kj} = 0,$$

showing that $u_{ij} = 0$ $(i > s; j \geq i)$ is a valid choice.

ACKNOWLEDGEMENTS

The research of Geert Molenberghs was supported by Nationaal Fonds voor Wetenschappelijk Onderzoek. The authors wish to thank Emmanuel Lesaffre for helpful comments.

BIBLIOGRAPHY

Ashford, J. R. and Sowden, R. R. (1970). "Multivariate probit analysis," *Biometrics,* 26, 535–546.

Devlin, S. J., Gnanadesikan, R. and Kettenring, J. R. (1975). "Robust estimation and outlier detection with correlation coefficients," *Biometrika,* 62(3), 531–545.

Edlefsen L. E. and Jones, S. D. GAUSS, Aptech Systems Inc., P.0. Box 6487, Kent, WA 98064, USA.

Gnanadesikan, R. (1977). *Methods for Statistical Data Analysis of Multivariate Observations,* John Wiley, New York.

Lesaffre, E. and Molenberghs, G. (1991). "Multivariate probit analysis; a neglected procedure in medical statistics." *Statistics in Medicine,* 10, 1391–1403.

Little, R. J. A. and Rubin, D. B. (1987). *Statistical Analysis with Missing Data,* John Wiley, New York.

Pemberton, J. D. (1984). "Estimation procedures for ordered categorical data," PhD thesis, University of Newcastle upon Tyne.

Seber, G. A. F. (1984). *Multivariate Observations,* John Wiley, New York.

Received October 1992; Revised November 1992