IMA Journal of Numerical Analysis (2002) **22**, 329–343

Computing the nearest correlation matrix—a problem from finance

NICHOLAS J. HIGHAM†
Department of Mathematics, University of Manchester, Manchester, M13 9PL, UK

[Received on 17 October 2000; revised on 23 July 2001]

Given a symmetric matrix, what is the nearest correlation matrix—that is, the nearest symmetric positive semidefinite matrix with unit diagonal? This problem arises in the finance industry, where the correlations are between stocks. For distance measured in two weighted Frobenius norms we characterize the solution using convex analysis. We show how the modified alternating projections method can be used to compute the solution for the more commonly used of the weighted Frobenius norms. In the finance application the original matrix has many zero or negative eigenvalues; we show that for a certain class of weights the nearest correlation matrix has correspondingly many zero eigenvalues and that this fact can be exploited in the computation.

Keywords: correlation matrix; positive semidefinite matrix; nearness problem; convex analysis; weighted Frobenius norm; alternating projections method; semidefinite programming.

1. Introduction

A correlation matrix is a symmetric positive semidefinite matrix with unit diagonal. Correlation matrices occur in several areas of numerical linear algebra, including preconditioning of linear systems and error analysis of Jacobi methods for the symmetric eigenvalue problem (see Davies & Higham (2000) for details and references). The term 'correlation matrix' comes from statistics, since a matrix whose (i, j) entry is the correlation coefficient between two random variables x_i and x_j is symmetric positive semidefinite with unit diagonal. It is a statistical application that motivates this work—one coming from the finance industry.

In stock research sample correlation matrices constructed from vectors of stock returns are used for predictive purposes. Unfortunately, on any day when an observation is made data are rarely available for all the stocks of interest. One way to deal with this problem is to compute the sample correlations of pairs of stocks using data drawn only from the days on which both stocks have data available. The resulting matrix of correlations will be only an *approximate* correlation matrix, because it has been built from inconsistent data sets. In order to justify the subsequent stock analysis it is desired to compute the nearest correlation matrix and to use that matrix in the computations. The matrices in this application are dense with dimensions in the thousands, and a particular feature is that relatively few vectors of observations are available, so that the approximate correlation matrix has low rank.

The problem we consider is, for arbitrary symmetric $A \in \mathbb{R}^{n\times n}$, to compute the

†Email: higham@ma.man.ac.uk, also at `http://www.ma.man.ac.uk/~higham/`

distance

$$\gamma(A) = \min\{\, \|A - X\| : X \text{ is a correlation matrix} \,\} \tag{1.1}$$

and a matrix achieving this minimum distance. The norm is a weighted version of the Frobenius norm, $\|A\|_F^2 = \sum_{i,j} a_{ij}^2$, the Frobenius norm being the easiest norm to work with for this problem and also being the natural choice from the statistical point of view. Two different weighted Frobenius norms are of interest. The first, and the most commonly used in numerical mathematics, is

$$\|A\|_W = \|W^{1/2} A W^{1/2}\|_F, \tag{1.2}$$

where W is a symmetric positive definite matrix. The second weighted norm is

$$\|A\|_H = \|H \circ A\|_F, \tag{1.3}$$

where H is a symmetric matrix of positive weights and $\circ$ denotes the Hadamard product: $A \circ B = (a_{ij} b_{ij})$.

The use of weights allows us to express our confidence in different elements of A: for the H-norm, if a_{ij} is known accurately (relatively to the other elements) then we can assign a large weight h_{ij}, so as to force x_{ij} to be close to a_{ij}, and conversely if a_{ij} is known relatively inaccurately then a small weight w_{ii} can be assigned. The W-norm does not allow the independent weighting of individual elements, but it is easier to work with, principally because the transformation $A \to W^{1/2} A W^{1/2}$ is a congruence, and so preserves inertia, while the transformation $A \to H \circ A$ merely preserves symmetry.

The two weighted norms coincide when $W = \mathrm{diag}(w_i)$ is diagonal and H is rank-1 with $h_{ij} = (w_i w_j)^{1/2}$, but neither norm includes the other as a special case. For the W-norm, a diagonal W is the most natural choice, but our theory and algorithms do not require W to be diagonal.

We define the sets

$$S = \{\, Y = Y^T \in \mathbb{R}^{n \times n} : Y \geqslant 0 \,\},$$
$$U = \{\, Y = Y^T \in \mathbb{R}^{n \times n} : y_{ii} = 1, i = 1\colon n \,\}.$$

Here, for a symmetric Y the notation $Y \geqslant 0$ ($\leqslant 0$) means Y is positive semidefinite (negative semidefinite). In the finance application $A \in U$ and $|a_{ij}| \leqslant 1$ for all $i \neq j$, but we will treat (1.1) with a general symmetric A.

We are looking for a matrix in the intersection of S and U that is closest to A in a weighted Frobenius norm. Since S and U are both closed convex sets, so is their intersection. It thus follows from standard results in approximation theory (for example, Luenberger 1969, p. 69) that the minimum in (1.1) is achieved and that it is achieved at a unique matrix X.

An interesting feature of the problem is that while positive definiteness is a property of the eigenvalues, and hence is basis independent, the possession of a unit diagonal is a basis-*dependent* property. This mismatch appears to preclude an explicit solution of the problem.

Some upper and lower bounds on $\gamma(A)$ are easily obtained.

LEMMA 1.1 Let $A \in \mathbb{R}^{n\times n}$ be symmetric, with eigenvalues $\lambda_1, \ldots, \lambda_n$. Then

$$\max\{\alpha_1, \alpha_2\} \leqslant \gamma(A) \leqslant \max\{\beta_1, \beta_2, \beta_3\},$$

where

$$\begin{aligned}
\alpha_1^2 &= \sum_{i=1}^{n} h_{ii}^2 (a_{ii} - 1)^2 + \sum_{\substack{|a_{ij}|>1 \\ i\neq j}} h_{ij}^2 (1 - |a_{ij}|)^2, \\
\alpha_2^2 &= \min\{\, \|A - X\| : X \in S \,\}, \\
\beta_1 &= \|A - I\|, \\
\beta_2 &= \min\{\, \|A - zz^T\| : z_i = \pm 1,\ i = 1\colon n \,\}, \\
\beta_3 &= \min_{0\leqslant\rho\leqslant 1} \|A - (\rho^{|i-j|})\|,
\end{aligned}$$

where $\|\cdot\|$ is the norm in the definition of γ. For the W-norm the lower bound α_1 is valid only for $W = \mathrm{diag}(w_i)$, in which case $h_{ij} = (w_i w_j)^{1/2}$.

Proof. Straightforward. The first lower bound follows from the fact that for any symmetric positive definite matrix A, $|a_{ij}| \leqslant \sqrt{a_{ii}a_{jj}}$. □

An explicit formula for α_2 is available for the W-norm, as we show in Section 3. When $W = I$ or $h_{ij} \equiv 1$, computing β_2 is equivalent to maximizing $z^T Az$ over all ± 1 vectors z, which is an NP-hard problem (Rohn, 1995). We are not aware of an explicit solution to the ρ minimization in the upper bound β_3, but techniques that may be helpful can be found in Suffridge & Hayden (1993, Section 3).

Two special cases in which the optimal X is known are worth noting, with the restriction that for the W-norm W is diagonal. If A is diagonal then $X = I$ (and, correspondingly, $\alpha_1 = \beta_1 = \beta_3$ in Lemma 1.1), and if A is positive semidefinite with diagonal elements less than or equal to 1 then X is obtained by replacing the diagonal elements by 1.

Finally, we note the inequality

$$|\gamma(A) - \gamma(B)| \leqslant \|A - B\|,$$

which holds for any nearness problem. The practical significance of the inequality is that if $\|A - B\|$ is sufficiently small then the nearest correlation matrix to A is a good enough approximation of the nearest correlation matrix to B.

In the next section we derive a characterization of the solution for both the W- and H-norms, and in the case of diagonal W (or rank-1 H) deduce information about the dimension of the null space of the solution. In Section 3 we show that the modified alternating projections method can be used to compute the solution, making use of a new result identifying the projection in the W-norm onto the positive semidefinite matrices. For diagonal W, we show how to exploit the low-rank property inherent in the finance application. Numerical experiments are given in Section 4 and concluding remarks in Section 5.

2. Theory

Important insight into the nearest correlation matrix problem can be obtained with the aid of optimization theory. The development in this section is inspired by the Glunt *et al.* (1990, Section 3) treatment of the nearest Euclidean distance matrix problem.

We will work with the W-norm in (1.2) and comment later on how the analysis adapts for the H-norm. We define

$$\langle A, B\rangle = \operatorname{trace}(A^T W B W),$$

which can be regarded as an inner product on $\mathbb{R}^{n\times n}$ that induces the W-norm.

The normal cone of a convex set $K \subset \mathbb{R}^{n\times n}$ at $B \in K$ is

$$\partial K(B) = \{\, Y = Y^T \in \mathbb{R}^{n\times n} : \langle Z - B, Y\rangle \leqslant 0 \text{ for all } Z \in K \,\} \tag{2.1}$$

$$= \left\{ Y = Y^T \in \mathbb{R}^{n\times n} : \langle Y, B\rangle = \sup_{Z\in K} \langle Y, Z\rangle \right\}. \tag{2.2}$$

Our starting point is the observation that the solution X to (1.1) is characterized by the condition that Luenberger (1969, p. 69)

$$\langle Z - X, A - X\rangle \leqslant 0 \quad \text{for all } Z \in S \cap U. \tag{2.3}$$

This condition can be rewritten as $A - X \in \partial(S\cap U)(X)$, the normal cone to $S\cap U$ at X. For two general convex sets K_1 and K_2, $\partial(K_1 \cap K_2)(X) = \partial K_1(X) + \partial K_2(X)$ if the relative interiors of K_1 and K_2 have a point in common (Rockafellar, 1970, Corollary 23.8.1). Any positive definite correlation matrix is in the relative interiors of both S and U, so we conclude that the solution X is characterized by

$$A - X \in \partial S(X) + \partial U(X). \tag{2.4}$$

Our task is now to determine ∂S and ∂U.

LEMMA 2.1 For $A \in U$,

$$\partial U(A) = \{\, W^{-1} \operatorname{diag}(\theta_i) W^{-1} : \theta_i \text{ arbitrary} \,\}. \tag{2.5}$$

Proof. We have

$$\partial U(A) = \left\{ Y = Y^T \in \mathbb{R}^{n\times n} : \langle Y, A\rangle = \sup_{Z\in U} \langle Y, Z\rangle \right\}$$

and the constraint can be written

$$\sum_{i,j} \widetilde{y}_{ij}\, a_{ij} = \sup_{Z\in U} \sum_{i,j} \widetilde{y}_{ij}\, z_{ij},$$

where $\widetilde{Y} = WYW$. If $\widetilde{Y}$ is not diagonal then we can choose z_{ij} arbitrarily large and of the same sign as $\widetilde{y}_{ij} \neq 0$ and thereby violate the sup condition. Therefore $\widetilde{Y}$ is diagonal, and any Y of the form $Y = W^{-1} \operatorname{diag}(\theta_i) W^{-1}$ satisfies the sup condition. □

The next two results generalize ones of Fletcher (1985).

LEMMA 2.2 For $A \in S$,

$$\partial S(A) = \{\, Y = Y^T : \langle Y, A\rangle = 0,\ Y \leqslant 0 \,\}.$$

Proof. We have

$$\partial S(A) = \left\{\, Y = Y^T : \langle Y, A\rangle = \sup_{Z\in S}\langle Y, Z\rangle \,\right\}.$$

Let $Z \in S$ have the spectral decomposition $Z = Q\Lambda Q^T$, where Q is orthogonal and $\Lambda = \mathrm{diag}(\lambda_i) \geqslant 0$. Then, with $C = Q^T W Y W Q$,

$$\begin{aligned}
\sup_{Z\in S}\langle Y, Z\rangle &= \sup_{\Lambda\geqslant 0,\ Q^TQ=I} \langle Y, Q\Lambda Q^T\rangle \\
&= \sup_{\Lambda\geqslant 0,\ Q^TQ=I} \langle C, \Lambda\rangle \\
&= \sup_{\Lambda\geqslant 0,\ Q^TQ=I} \sum_i \lambda_i c_{ii} \\
&= \begin{cases} 0, & \text{if } Y \leqslant 0, \\ \infty, & \text{otherwise.} \end{cases}
\end{aligned}$$

Thus equality holds in the sup condition for Y such that $\langle Y, A\rangle = 0$ and $Y \leqslant 0$. □

COROLLARY 2.3 For $A \in S$,

$$\begin{aligned}
\partial S(A) = \{\, Y : WYW = -VDV^T,\ &\text{where } V \in \mathbb{R}^{n\times p} \text{ has orthonormal columns} \\
&\text{spanning null}(A) \text{ and } D = \mathrm{diag}(d_i) \geqslant 0 \,\}.
\end{aligned}$$

Proof. Let A have the spectral decomposition $Q\Lambda Q^T$, where $\Lambda = \mathrm{diag}(\lambda_i)$ with $\lambda_1 \geqslant \cdots \geqslant \lambda_{n-p} > 0 = \lambda_{n-p+1} = \cdots = \lambda_n$. Write $\Lambda_1 = \mathrm{diag}(\lambda_1, \ldots, \lambda_p)$, $Q = [Q_1, Q_2]$, with $Q_1 \in \mathbb{R}^{n\times(n-p)}$, and note that the columns of Q_2 span null(A).

For $Y \in \partial S(A)$,

$$0 = \langle Y, A\rangle = \mathrm{trace}(AWYW) = \mathrm{trace}(Q_1\Lambda_1 Q_1^T WYW) = \mathrm{trace}(\Lambda_1 Q_1^T WYW Q_1),$$

and since $\Lambda_1 > 0$ and $Y \leqslant 0$ this implies that $\mathrm{diag}(Q_1^T WYW Q_1) = 0$. Now write

$$\begin{bmatrix} G & H \\ H^T & M \end{bmatrix} := Q^T(WYW)Q = \begin{bmatrix} Q_1^T(WYW)Q_1 & Q_1^T(WYW)Q_2 \\ Q_2^T(WYW)Q_1 & Q_2^T(WYW)Q_2 \end{bmatrix} \leqslant 0.$$

Since $\mathrm{diag}(G) = 0$ it follows that $G = 0$ and hence $H = 0$; furthermore, $M \leqslant 0$. Then $0 \geqslant WYW = Q_2 M Q_2^T = -VDV^T$, where we have used the spectral decomposition of M to produce $V \in \mathbb{R}^{n\times p}$ with orthonormal columns and $D \in \mathbb{R}^{p\times p}$ diagonal and positive semidefinite. □

We are now ready to state a theorem that characterizes the solution of our nearness problem.

THEOREM 2.4 The correlation matrix X solves (1.1) if and only if

$$X = A + W^{-1}\big(VDV^T + \mathrm{diag}(\theta_i)\big)W^{-1}, \tag{2.6}$$

where $V \in \mathbb{R}^{n\times p}$ has orthonormal columns spanning $\mathrm{null}(X)$, $D = \mathrm{diag}(d_i) \geqslant 0$, and the θ_i are arbitrary.

Proof. The result follows from the condition (2.4) on applying Lemma 2.1 and Corollary 2.3. □

An analogue of Theorem 2.4 holds also for the H-norm, with

$$X = A + \big(VDV^T + \mathrm{diag}(\theta_i)\big) \circ (h_{ij}^{-2}).$$

This can be proved by modifying the analysis above, or it can be deduced from a slightly more general result of Johnson *et al.* (1998, Theorem 2.2). In the case $W = I$, Theorem 2.4 can also be deduced from an expression for $\partial(S\cap U)(X)$ given by Laurent & Poljak (1995).

An immediate implication of Theorem 2.4 is that, at least when W is diagonal, X will generally be singular (and hence not positive definite). For if X is nonsingular then $V = 0$ and $X = A + W^{-1}\,\mathrm{diag}(\theta_i)W^{-1}$, which means that X is obtained simply by adjusting the diagonal elements to 1.

It is interesting to note that (2.6) implies the necessary condition for optimality that X satisfies the quadratic matrix equation $XW(X - A)W = X\,\mathrm{diag}(\theta_i)$.

In the important special case where W is diagonal and the diagonal elements of A are at least 1, we can say more.

THEOREM 2.5 Let $A = A^T$ have diagonal elements $a_{ii} \geqslant 1$ and let W be diagonal. Then, in Theorem 2.4, $\theta_i \leqslant 0$ for all i. Moreover, if A has t nonpositive eigenvalues then the nearest correlation matrix has at least t zero eigenvalues.

Proof. Since A has diagonal elements at least 1 and $W^{-1}(VDV^T)W^{-1} \geqslant 0$, the diagonal elements of $X = A + W^{-1}(VDV^T)W^{-1}$ in (2.6) are all at least 1. Therefore in order for X to have unit diagonal we need $\theta_i \leqslant 0$ for all i. Examining (2.6) we see that the perturbation $W^{-1}(VDV^T)W^{-1}$ moves the t or more nonpositive eigenvalues of $A + W^{-1}\,\mathrm{diag}(\theta_i)W^{-1}$ to become nonnegative. The perturbation VDV^T has rank at most p and so, by a standard result on the eigenvalues of a symmetric matrix subject to a low-rank perturbation (Horn & Johnson, 1985, Theorem 4.3.6), we must have $p \geqslant t$. Since p is the dimension of the null space of X, the result follows. □

That a restriction on the diagonal of A is necessary in Theorem 2.5 can be seen from that fact that if A is a diagonal matrix then, as noted at the end of Section 1, the nearest correlation matrix is I for any diagonal W, irrespective of the values of the a_{ii}.

Although they do not give a method for computing a solution, Theorems 2.4 and 2.5 can be used to verify a putative solution. To illustrate we consider the matrix

$$A = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 1 \end{bmatrix}$$

and the unweighted Frobenius norm. Since $A = ee^T - (e_1 e_3^T + e_3 e_1^T)$, where $e = [1\ 1\ 1]^T$, it is clear that A is indefinite, and in fact its eigenvalues are $1+\sqrt{2}$, 1 and $1-\sqrt{2}$. An obvious candidate for the nearest correlation matrix to A is $X = ee^T$, with $\|A - X\|_F = \sqrt{2}$. The null space of X is spanned by the columns of

$$V = \begin{bmatrix} 1 & -1 \\ -1 & 0 \\ 0 & 1 \end{bmatrix},$$

so from Theorem 2.4 (which remains true if the columns of V are not orthonormal) we must have

$$X = A + \begin{bmatrix} d_1 + d_2 & -d_1 & -d_2 \\ -d_1 & d_1 & 0 \\ -d_2 & 0 & d_2 \end{bmatrix} + \operatorname{diag}(\theta_i)$$

with $d_i \geqslant 0$. This equation implies $d_2 = -1$ and so X cannot be a solution. In fact, the solution is, to the figures shown,

$$X = \begin{bmatrix} 1{\cdot}0000 & 0{\cdot}7607 & 0{\cdot}1573 \\ 0{\cdot}7607 & 1{\cdot}0000 & 0{\cdot}7607 \\ 0{\cdot}1573 & 0{\cdot}7607 & 1{\cdot}0000 \end{bmatrix},$$

with $\|A - X\|_F = 0{\cdot}5278$. This matrix X is singular, with a one-dimensional null space spanned by $q = [-0{\cdot}4814,\ 0{\cdot}7324,\ -0{\cdot}4814]^T$. A short MATLAB computation verifies that adding a suitable positive multiple of qq^T to A reproduces the off-diagonal of X, and the θ_i can then be chosen to achieve equality in (2.6), verifying the optimality of X.

3. Computation

3.1 *Projections*

Our problem is to project from the symmetric matrices onto the correlation matrices, with respect to a weighted Frobenius form. We consider first how to project onto the sets S and U individually. We begin with U and denote by P_U the projection onto U.

THEOREM 3.1 For the W-norm,

$$P_U(A) = A - W^{-1} \operatorname{diag}(\theta_i) W^{-1},$$

where $\theta = [\theta_1, \ldots, \theta_n]^T$ is the solution of the linear system

$$(W^{-1} \circ W^{-1})\theta = \operatorname{diag}(A - I). \tag{3.1}$$

Proof. The projection $X = P_U(A)$ is characterized by $A - X \in \partial U(X)$, which, by Lemma 2.1 can be written

$$A - X = W^{-1} \operatorname{diag}(\theta_i) W^{-1}.$$

Equating diagonal elements and writing $W^{-1} = (\omega_{ij})$, we have

$$\sum_{j=1}^{n} \omega_{ij}^2 \theta_j = a_{ii} - 1.$$

These equations form the linear system (3.1). Since W is positive definite so is $W^{-1} \circ W^{-1}$, and so this linear system has a unique solution. □

In the case where W is diagonal we can write, more simply,

$$P_U(A) = (p_{ij}), \qquad p_{ij} = \begin{cases} a_{ij}, & i \neq j, \\ 1, & i = j. \end{cases} \tag{3.2}$$

It is easy to show that, for the H-norm, (3.2) is the projection onto U for all H.

Projection onto S is more difficult. No closed formula is known for the H-norm, but for the W-norm the following result, which appears to be new, provides such a formula. We need some more notation. For a symmetric $A \in \mathbb{R}^{n\times n}$ with spectral decomposition $A = QDQ^T$, where $D = \operatorname{diag}(\lambda_i)$ and Q is orthogonal, let

$$A_+ = Q \operatorname{diag}(\max(\lambda_i, 0))Q^T, \quad A_- = Q \operatorname{diag}(\min(\lambda_i, 0))Q^T.$$

Note that A_+ and A_- do not depend on the choice of spectral decomposition and $A = A_+ + A_-$, $A_+A_- = A_-A_+ = 0$.

THEOREM 3.2 For the W-norm,

$$P_S(A) = W^{-1/2}\big((W^{1/2}AW^{1/2})_+\big)W^{-1/2}. \tag{3.3}$$

Moreover,

$$\operatorname{diag}(P_S(A)) \geqslant \operatorname{diag}(A).$$

Proof. We need to show that the claimed projection X satisfies $A - X \in \partial S(X)$, that is, from Lemma 2.2, that

$$A - X \leqslant 0, \qquad \operatorname{trace}((A - X)WXW) = 0.$$

Now

$$\begin{aligned} A - X &= W^{-1/2}(W^{1/2}AW^{1/2} - (W^{1/2}AW^{1/2})_+)W^{-1/2} \\ &= W^{-1/2}(W^{1/2}AW^{1/2})_- W^{-1/2} \leqslant 0 \end{aligned}$$

and then

$$\begin{aligned} (A - X)WXW &= W^{-1/2}(W^{1/2}AW^{1/2})_- W^{-1/2} \cdot W^{1/2}\big((W^{1/2}AW^{1/2})_+\big)W^{1/2} \\ &= W^{-1/2}\,(W^{1/2}AW^{1/2})_-\,(W^{1/2}AW^{1/2})_+\,W^{1/2} = 0. \end{aligned}$$

For the last part, we have

$$(W^{1/2}AW^{1/2})_+ - W^{1/2}AW^{1/2} \geqslant 0.$$

Pre- and post-multiplying by $W^{-1/2}$ effects a congruence transformation and so preserves the inequality, and taking the diagonal parts gives the result, since the diagonal of a positive semidefinite matrix is nonnegative. □

3.2 *Alternating projections method*

To find the nearest matrix at the intersection of the sets S and U we might iteratively project by repeating the operation

$$A \leftarrow P_U(P_S(A)).$$

The idea of iteratively projecting onto subspaces was analysed in a Hilbert space setting by von Neumann, who proved convergence to the point in the intersection nearest to the starting point. See Deutsch (1992) for a survey of the large literature on von Neumann's method. Our sets are not subspaces, so von Neumann's convergence result does not apply. Indeed when the subspaces are replaced by closed convex sets the iteration can converge to non-optimal points (Han, 1988). We therefore use a modified iteration due to Dykstra (1983), which incorporates a judiciously chosen correction to each projection that can be interpreted as a normal vector to the corresponding convex set. Note that while U is not a subspace it is a translate of a subspace and, as noted in Boyle & Dykstra (1985), for a translate of a subspace the corresponding correction in the general algorithm in Dykstra (1983) can be omitted.

We restrict now to the W-norm, though the following algorithm could also be used for the H-norm if we had an efficient way of computing the projection P_S for this norm.

ALGORITHM 3.3 Given a symmetric $A \in \mathbb{R}^{n\times n}$ this algorithm computes the nearest correlation matrix to A in the W-norm.

$\Delta S_0 = 0, Y_0 = A$
for $k = 1, 2, \ldots$
  $R_k = Y_{k-1} - \Delta S_{k-1}$ % ΔS_{k-1} is Dykstra's correction.
  $X_k = P_S(R_k)$
  $\Delta S_k = X_k - R_k$
  $Y_k = P_U(X_k)$
end

General results of Boyle & Dykstra (1985, Theorem 2) and Han (1988, Theorem 4.7) show that X_k and Y_k both converge to the desired correlation matrix as $k \to \infty$. The rate of convergence of Dykstra's algorithm is linear when the sets are subspaces, the constant depending on the angle between the subspaces (Deutsch, 1983; Deutsch & Hundal, 1997), so we can expect linear convergence, at best, of Algorithm 3.3.

The next result gives some insight into the behaviour of Algorithm 3.3 for diagonal W.

THEOREM 3.4 Suppose $A = A^T$ has diagonal elements $a_{ii} \geqslant 1$ and let W be diagonal. Let $Y_k = P_U(X_k) = X_k + D_k$, where D_k is diagonal (in view of Theorem 3.1). Then in Algorithm 3.3

$$R_k = A + \Delta_k, \tag{3.4}$$

where the diagonal matrix

$$\Delta_k = \sum_{k=1}^{k-1} D_k$$

is negative semidefinite.

Proof. We have $R_1 = A$ and

$$\begin{aligned} R_{k+1} &= Y_k - \Delta S_k = P_U(X_k) - \Delta S_k \\ &= X_k + D_k - (X_k - R_k) = R_k + D_k, \end{aligned}$$

so (3.4) is proved. Furthermore,

$$\begin{aligned} I &= \operatorname{diag}(Y_k) = \operatorname{diag}(X_k + D_k) \\ &= \operatorname{diag}(P_S(R_k) + D_k) \\ &\geqslant \operatorname{diag}(R_k + D_k) \\ &= \operatorname{diag}(A + \Delta_{k+1}) \\ &\geqslant I + \Delta_{k+1}, \end{aligned}$$

so that $\Delta_{k+1} \leqslant 0$, as required. □

Theorem 3.4 shows that R_k is A minus a positive semidefinite diagonal matrix. This has an important implication for the case where A is 'highly nonpositive definite', or, in particular, highly rank-deficient.

COROLLARY 3.5 Let $A = A^T$ have diagonal elements $a_{ii} \geqslant 1$ and t nonpositive eigenvalues and let W be diagonal. Then in Algorithm 3.3 R_k has at least t nonpositive eigenvalues and X_k has at least t zero eigenvalues, for all k.

Note that by letting $k \to \infty$ in the corollary we recover the second part of Theorem 2.5.

The practical significance of the corollary is that if t is large then we can compute $P_s(R_k)$ at a much lower cost than that of computing the complete eigensystem of $W^{1/2}R_kW^{1/2}$ (recall (3.3)). It suffices to compute the largest $n - t \ll n$ eigenvalues λ_j and corresponding orthonormal eigenvectors q_j of $W^{1/2}R_kW^{1/2}$ and then take in (3.3)

$$(W^{1/2}R_kW^{1/2})_+ = \sum_{\lambda_i > 0} \lambda_i q_i q_i^T.$$

This computation can be done very efficiently by the Lanczos iteration or by orthogonally reducing R_k to tridiagonal form and applying the bisection method followed by inverse iteration (Trefethen & Bau III, 1997, pp. 227 ff.).

If A has few nonpositive eigenvalues ($t \ll n$) then it is likely that the R_k, too, will have few nonpositive eigenvalues, though an upper bound on this number is not available. Nevertheless, similar computational savings are possible in this situation by computing

$$(W^{1/2}R_kW^{1/2})_+ = W^{1/2}R_kW^{1/2} - \left(\sum_{\lambda_i \leqslant 0} \lambda_i q_i q_i^T\right),$$

where the number of nonpositive eigenvalues of $W^{1/2}R_kW^{1/2}$ is estimated from step to step, and the estimate increased if it is found to be too small.

3.3 *Semidefinite programming*

Another way to attack the nearest correlation matrix problem (1.1) is to phrase it as a semidefinite programming problem and then exploit the powerful interior-point algorithms available for semidefinite programming (see Todd (2001) and the references therein).

The positive semidefinite program in primal standard form is

$$\begin{aligned} &\text{minimize } \langle C, X\rangle \\ &\text{subject to } \langle A_i, X\rangle = b_i, \quad i = 1\colon m, \quad X = \operatorname{diag}(X_1, X_2, \ldots, X_r) \geqslant 0, \end{aligned} \tag{3.5}$$

where C and the A_i are given $n \times n$ symmetric matrices. Ignoring weights, for simplicity, our aim is to minimize

$$\|A - X\|_F^2 = \|A\|_F^2 + \langle X, X\rangle - 2\langle A, X\rangle \equiv a^T a + x^T x - 2a^T x$$

subject to X being a correlation matrix, where $x = \operatorname{vec}(X)$ and $a = \operatorname{vec}(A)$ and the vec operator stacks the columns of a matrix into one long vector. We can rephrase the problem as

$$\text{minimize } \theta \text{ subject to } Y = \begin{bmatrix} I_{n^2} & x \\ x^T & \theta + 2a^T x - a^T a \end{bmatrix} \geqslant 0, \quad X \geqslant 0, \quad \operatorname{diag}(X) = I_n.$$

Our variables are now $Z = \operatorname{diag}(X, Y, \theta) \geqslant 0$ and the equality constraints are

$$\begin{aligned} \langle e_i e_i^T, Z\rangle &= 1, \quad i = 1\colon n \quad (X \text{ has unit diagonal}), \\ \langle e_i e_j^T + e_j e_i^T, Z\rangle &= \delta_{ij}, \quad n+1 \leqslant i \leqslant j \leqslant n + n^2, \quad (Y(1\colon n^2, 1\colon n^2) = I_{n^2}), \end{aligned}$$

together with n^2 constraints relating Y to X, of the form

$$\langle e_i e_j^T + e_j e_i^T - e_p e_{n+n^2+1}^T - e_{n+n^2+1} e_p^T, Z\rangle = 0, \qquad 1 \leqslant i \leqslant j \leqslant n, \quad n+1 \leqslant p \leqslant n+n^2,$$

and a final constraint

$$\langle \operatorname{diag}(2A, -e_{n+n^2+1} e_{n+n^2+1}^T, 1), Z\rangle = a^T a.$$

In total, there are $n^4/2 + 3n^2/2 + n + 1$ constraints. Unfortunately, this number of constraints make it impractical to apply a general semidefinite programming solver—merely specifying the constraints (taking full advantage of their sparsity) requires a prohibitive amount of memory.

Another possibility is to express the problem in terms of a quadratic cone (or Lorentz cone) constraint. For example, we can write $Y = A - X$ and then minimize α subject to X being a correlation matrix and $\|\operatorname{vec}(Y)\|_2 \leqslant \alpha$. Efficient methods are available for solving problems with cone constraints but, since the number of variables is still $O(n^2)$, standard software is likely to require at least $O(n^4)$ operations per iteration, which again is impractical for large n; numerical experiments with the SeDuMi package (Sturm, 1999) confirm this conclusion (Anjos Wolkowicz, private communication).

Johnson *et al.* (1998) treat a class of positive semidefinite completion problems of which (1.1), for the H-norm (with H now allowed to have zero elements), is a special

case, and they derive two interior-point methods for solving this class of problems. Unfortunately, when applied to our problem with an H having all positive entries the methods in Johnson *et al.* (1998) are prohibitively expensive (computing the Newton direction requires $O(n^{12})$ operations) and the constraint of unit diagonal cannot be directly incorporated.

How to efficiently solve the nearest correlation matrix problem by semidefinite programming techniques (including how to take advantage of 'highly positive definite' or 'highly nonpositive definite' matrices A) therefore remains an interesting open question.

4. Numerical experiments

In our experiments we tested for convergence in Algorithm 3.3 at the end of the for loop using the condition

$$\max\left\{\frac{\|X_k - X_{k-1}\|_\infty}{\|X_k\|_\infty}, \frac{\|Y_k - Y_{k-1}\|_\infty}{\|Y_k\|_\infty}, \frac{\|Y_k - X_k\|_\infty}{\|Y_k\|_\infty}\right\} \leqslant \text{tol}, \tag{4.1}$$

where tol is a tolerance. Our experience shows that the three quantities in this test are usually of the same order of magnitude, so in practice any one of them can be used to test for convergence. Our computations were done in MATLAB 6, for which the unit roundoff $u = 2^{-53} \approx 1{\cdot}1 \times 10^{-16}$. For ease of description we use the unweighted Frobenius norm.

In our first example we take the positive definite matrix

$$A = \begin{bmatrix} 2 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 2 \end{bmatrix}.$$

With tol $= 10^{-8}$, Algorithm 3.3 converges in 19 iterations, with the relative differences in (4.1) reducing by a factor approximately 3 on each iteration. The solution is, to the figures shown,

$$X = \begin{bmatrix} 1{\cdot}0000 & -0{\cdot}8084 & 0{\cdot}1916 & 0{\cdot}1068 \\ -0{\cdot}8084 & 1{\cdot}0000 & -0{\cdot}6562 & 0{\cdot}1916 \\ 0{\cdot}1916 & -0{\cdot}6562 & 1{\cdot}0000 & -0{\cdot}8084 \\ 0{\cdot}1068 & 0{\cdot}1916 & -0{\cdot}8084 & 1{\cdot}000 \end{bmatrix}, \qquad \|A - X\|_F = 2{\cdot}13,$$

and X has rank 3. The bounds from Lemma 1.1 are shown for all our examples in Table 1, wherein we approximated β_3 by the approximate local minimum obtained with MATLAB's `fminbnd` minimizer.

For our second example we begin with a random 500×500 correlation matrix C with eigenvalues $\alpha\beta^i$, where $\beta^{10} = 10^{-8}$ and α is chosen so that the eigenvalues sum to 10; C is generated using MATLAB 6's `gallery('randcorr',...)`, which uses an algorithm described in Davies & Higham (2000). Then we set $A = C + E$, where E is a random symmetric perturbation of Frobenius norm 10^{-4} and find the nearest correlation matrix to A. This problem models the situation where a correlation matrix is corrupted by relatively large errors. With tol $= 10^{-6}$, Algorithm 3.3 converges in two iterations and $\|A - X\|_F =$

TABLE 1 *Lower and upper bounds from Lemma* 1.1 *for the three test problems. For the second and third examples* β_2 *is too expensive to compute*

			Lower bounds		Upper bounds		
Example	n	Distance	α_1	α_2	β_1	β_2	β_3
1	4	2·13	2·0	0	3·16	3·16	3·16
2	500	1·0e−5	6·21e−6	8·34e−6	6·4e+1	–	6·4e+1
3	1399	2·10e1	3·36e−1	1·32e+1	3·59e+2	–	3·58e+2

$1{\cdot}0 \times 10^{-5}$. Note from Table 1 that the lower bounds from Lemma 1.1 are good estimates in this case.

Our third example is a matrix of stock data provided to us by a fund management company. The matrix A is an approximate correlation matrix of dimension 1399; it has unit diagonal and its off-diagonal elements are bounded by 1 in magnitude, but it is not positive semidefinite. The eigenvalues of A range between $-8{\cdot}5$ and 339 (see Fig. 1), with 30 eigenvalues in the interval $[-8{\cdot}5, -10^{-3}]$, 1215 in the interval $[-10^{-11}, 10^{-11}]$, and the rest in the interval $[1{\cdot}3, 339]$. Thus A is a highly rank-deficient matrix and Theorem 2.5 tells us that the nearest correlation matrix will have at least 1245 zero eigenvalues, and therefore rank at most 154. The low rank results from the sample correlation matrix being constructed from a small number of observations and is typical in this application.

We applied Algorithm 3.3 with tol $= 10^{-4}$, since the data is accurate to 2–3 significant digits only. The algorithm converged in 67 iterations to an X with $\|A - X\|_F = 20{\cdot}96$; the spectrum of X is plotted in Fig. 1. Since $\|X\|_2 = 339$ and tol $= 10^{-4}$, all the eigenvalues of X less than about 10^{-2} (the first 1245 eigenvalues in the plot) are zeros to within the convergence tolerance (see Corollary 3.5).

When the projection P_S was computed via the full eigensystem, using MATLAB's `eig` function, the computation took 2 h 45 min on a 1Ghz Pentium III. In order to take advantage of the high rank-deficiency of A we repeated the computation by calling LAPACK's `dsyevr` (via a MEX interface) in place of `eig`; this routine reduces to tridiagonal form and (when a partial spectrum is requested) uses the bisection method and inverse iteration. We used `dsyevr` to compute just the largest $154 + 10$ eigenvalues and corresponding eigenvectors of R_k on each iteration, where the '+10' is a safety factor that enables us to check that we have obtained all the non-negligible positive eigenvalues. The number of iterations was unchanged, but the computation time dropped to 37 min—an improvement by a factor 4.5.

5. Concluding remarks

This work adds to the large literature on matrix nearness problems, a survey of which can be found in Higham (1989). Algorithm 3.3 guarantees to compute the nearest correlation matrix to A and can exploit the spectral properties of A inherent in the finance application. The main weakness of the algorithm is its linear convergence rate. We are currently investigating alternative algorithms for this problem, as well as generalizations, such as to include rank constraints on X.

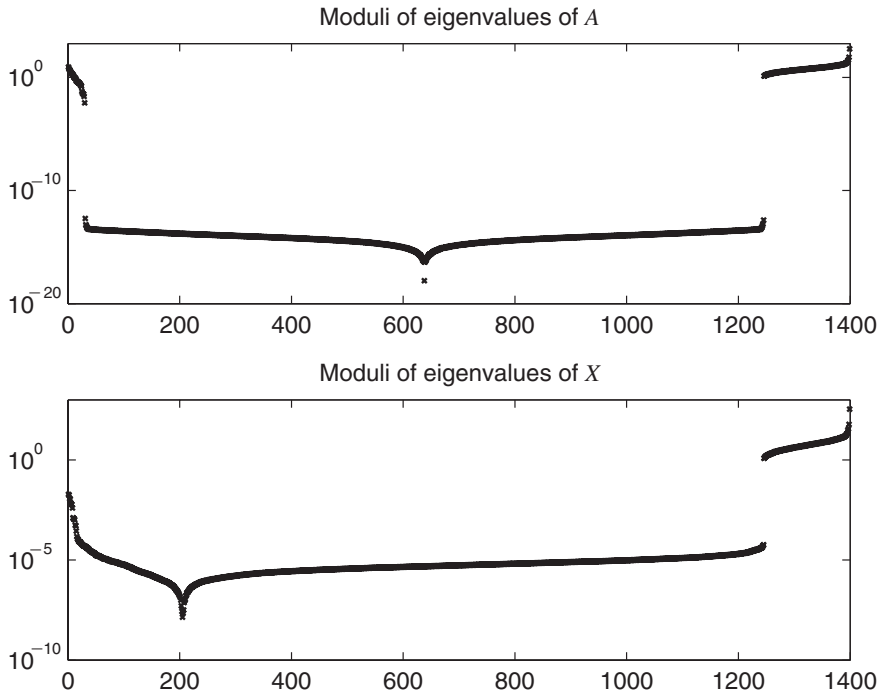

FIG. 1. Eigenvalues of A (top) and eigenvalues of the nearest correlation matrix X (bottom): the index of the eigenvalues sorted in increasing order is displayed on the x-axis and $|\lambda_i|$ on the y-axis.

Acknowledgements

I thank Craig Lucas for writing the MEX-file interface to LAPACK's `dsyevr` and Henry Wolkowicz for helpful discussions and advice on semidefinite programming. This work was supported by Engineering and Physical Sciences Research Council grants GR/L76532 and GR/R22612 and by a Royal Society Leverhulme Trust Senior Research Fellowship.

REFERENCES

BOYLE, J. P. & DYKSTRA, R. L. (1985) A method for finding projections onto the intersection of convex sets in Hilbert spaces. *Advances in Order Restricted Inference 37*, Lecture Notes in Statistics. Berlin: Springer, pp. 28–47.

DAVIES, P. I. & HIGHAM, N. J. (2000) Numerically stable generation of correlation matrices and their factors. *BIT*, **40**, 640–651.

DEUTSCH, F. (1983) Von Neumann's alternating method: The rate of convergence. *Approximation Theory IV*. (C. K. Chui, L. L. Schumaker & J. D. Ward, eds). New York: Academic, pp. 427–434.

DEUTSCH, F. (1992) The method of alternating orthogonal projections. *Approximation Theory, Spline Functions and Applications*. (S. P. Singh, ed.). Dordrecht: Kluwer, pp. 105–121.

DEUTSCH, F. & HUNDAL, H. (1997) The rate of convergence for the method of alternating projections, II. *J. Math. Anal. and Appl.*, **205**, 381–405.

DYKSTRA, R. L. (1983) An algorithm for restricted least squares regression. *J. Amer. Stat. Assoc.*, **78**, 837–842.

FLETCHER, R. (1985) Semi-definite matrix constraints in optimization. *SIAM J. Control and Optimization*, **23**, 493–513.

GLUNT, W., HAYDEN, T. L., HONG, S. & WELLS, J. (1990) An alternating projection algorithm for computing the nearest Euclidean distance matrix. *SIAM J. Matrix Anal. Appl.*, **11**, 589–600.

HAN, S.-P. (1988) A successive projection method. *Math. Prog.*, **40**, 1–14.

HIGHAM, N. J. (1989) Matrix nearness problems and applications. *Applications of Matrix Theory*. (M. J. C. Gover & S. Barnett, eds). Oxford: Oxford University Press, pp. 1–27.

HORN, R. A. & JOHNSON, C. R. (1985) *Matrix Analysis*. Cambridge: Cambridge University Press.

JOHNSON, C. R., KROSCHEL, B. & WOLKOWICZ, H. (1998) An interior-point method for approximate positive semidefinite completions. *Comput. Optim. Appl.*, **9**, 175–190.

LAURENT, M. & POLJAK, S. (1995) On a positive semidefinite relaxation of the cut polytope. *Linear Algebra Appl.*, **223/224**, 439–461.

LUENBERGER, D. G. (1969) *Optimization by Vector Space Methods*. New York: Wiley.

ROCKAFELLAR, T. R. (1970) *Convex Analysis*. Princeton, NJ: Princeton University Press.

ROHN, J. (January 1995) NP-hardness results for some linear and quadratic problems. *Technical Report No. 619*. Prague: Institute of Computer Science, Academy of Sciences of the Czech Reupublic.

STURM, J. F. (1999) Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones. *Optimization Methods and Software*, **11–12**, 625–653.

SUFFRIDGE, T. J. & HAYDEN, T. L. (1993) Approximation by a Hermitian positive semidefinite Toeplitz matrix. *SIAM J. Matrix Anal. Appl.*, **14**, 721–734.

TODD, M. J. (2001) Semidefinite optimization. *Acta Numerica*, **10**, 515–560.

TREFETHEN, L. N. & BAU III, D. (1997) *Numerical Linear Algebra*. Philadelphia, PA: Society for Industrial and Applied Mathematics.