Available online at www.sciencedirect.com
ScienceDirect
J. Math. Anal. Appl. 331 (2007) 1452–1466
Journal of MATHEMATICAL ANALYSIS AND APPLICATIONS
www.elsevier.com/locate/jmaa

Optimal ellipsoids and decomposition of positive definite matrices

Yuk J. Leung, Wenbo V. Li [1], Rakesh *

Department of Mathematical Sciences, University of Delaware, Newark, DE 19716, USA

Received 13 February 2006

Available online 3 November 2006

Submitted by William F. Ames

Abstract

Given a positive definite matrix A, we characterize the unique diagonal matrix D, $D \geqslant A$, with the smallest determinant. Equivalently, given an ellipsoid $\mathcal{A}$, we characterize the unique ellipsoid of the largest volume contained in $\mathcal{A}$, with principal axes parallel to the coordinate axes.

Keywords: Ellipsoid; Determinant maximization; John's theorem

1. Introduction

Throughout this article, all matrices are real valued, A^T represents the transpose of the matrix A, and for matrices A, B, we say $A > B$ if $A - B$ is positive definite, and $A \geqslant B$ if $A - B$ is positive semi-definite. For a positive integer n, vectors in R^n will be considered as $n \times 1$ matrices and for an $x \in R^n$ we say $x > 0$ if each component of x is positive. If $x = [x_1, \ldots, x_n]^T$ and $y = [y_1, \ldots, y_n]^T$ are vectors in R^n and α is a real number, we define $e = [1, 1, \ldots, 1]^T$, $\Pi(x) = \prod_{i=1}^n x_i$, $x \circ y = [x_1 y_1, \ldots, x_n y_n]^T$, $x^\alpha = [x_1^\alpha, \ldots, x_n^\alpha]^T$. For $p \in R^n$, $D(p)$ represents the $n \times n$ diagonal matrix whose entries are the corresponding entries of p. For a square matrix X, $\operatorname{diag}(X)$ denotes the diagonal matrix with the same diagonal as X.

* Corresponding author.
E-mail address: rakesh@math.udel.edu (Rakesh).
[1] Supported in part by NSF Grant DMS-0204513.

0022-247X/$ – see front matter
doi:10.1016/j.jmaa.2006.09.065

The following problem studied in this article is motivated by the computer simulation of multivariate Gaussian random variables by the acceptance/rejection method (see [3]) using a product of one-dimensional Gaussian distributions.

Problem 1 *(Diagonal problem). Suppose A is a positive definite matrix. Minimize* $\det(D)$ *over all diagonal matrices D for which $D \geqslant A$.*

[6] gives efficient numerical methods for finding the optimizer of Problem 1 using interior methods; our goal in this article is a theoretical analysis of the problem.

Problem 1 has an interesting geometrical interpretation. We have $D \geqslant A$ if and only if $x^T Dx \geqslant x^T Ax$ for all x, or, after normalization, that $n \geqslant x^T Ax$ for all x with $x^T Dx = n$. So $D \geqslant A$ if and only if the ellipsoid $x^T Ax \leqslant n$ contains the ellipsoid $x^T Dx = n$. The volume of the ellipsoid $x^T Dx = n$ is $n^{n/2}\omega_n/\sqrt{\det(D)}$ where ω_n is the volume of the unit ball in R^n. So Problem 1 is equivalent to the following problem:

Given an ellipsoid $\mathcal{A}$ in R^n, find the ellipsoid $\mathcal{D}$, contained in $\mathcal{A}$ with principal axes parallel to the coordinate axes, of the largest volume.

Definition. $Sym(n)$ will denote the set of real valued, symmetric matrices of order n. Given positive integers $n_1, \ldots, n_k$ and n with $n_1 + \cdots + n_k = n$, we define $\mathcal{B}(n_1, \ldots, n_k)$ to be the subset of $Sym(n)$ consisting of block matrices B of the form

$$B = \begin{bmatrix} B_1 & 0 & \cdots & 0 \\ 0 & B_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & B_k \end{bmatrix}$$

where each B_i is an $n_i \times n_i$ matrix. For a square matrix M of order n, $\text{block}(M)$ will be the unique block matrix B whose blocks B_i equal the corresponding entries of M.

One may generalize Problem 1 by conducting the search for the optimizer over block matrices of a certain type instead of diagonal matrices.

Problem 2. *Suppose A is a positive definite matrix of order n and $n_1, \ldots, n_k$ are positive integers so that $n = n_1 + \cdots + n_k$. Minimize* $\det B$ *over all block matrices $B \in \mathcal{B}(n_1, \ldots, n_k)$ for which $B \geqslant A$.*

The following theorem gives a characterization of the optimizer of Problem 2.

Theorem 1 *(Block characterization). The infimum is attained in Problem* 2 *and the minimizer is unique. Further, the block matrix $B \in \mathcal{B}(n_1, \ldots, n_k)$ is the minimizer iff the following two conditions are satisfied*:

(i) *$B \geqslant A$ and $B - A$ is singular*;
(ii) *there is an $n \times n$ matrix $X \geqslant 0$ with $(B - A)X = 0$ and* $\text{block}(X) = B^{-1}$.

A corollary of Theorem 1 characterizes the optimizer of Problem 1.

Corollary 2 *(Diagonal characterization). In Problem* 1*, the infimum is attained by a unique positive diagonal matrix. Further, a diagonal matrix D is the optimizer for Problem* 1 *if and only if the following two conditions hold*:

(i) $D \geqslant A$ *and* $D - A$ *is singular*;
(ii) *there is an* $n \times n$ *matrix* $X \geqslant 0$ *with* $(D - A)X = 0$ *and* $\operatorname{diag}(X) = D^{-1}$.

In Corollary 2, the characterization condition (ii) is the significant condition; condition (i) just determines the correct scalar multiple of a diagonal matrix. Geometrically, condition (ii) determines the lengths of the principal axes of the optimal ellipsoid up to a constant; condition (i) determines the constant.

In [4] (see [1] for a more recent exposition), John considered the question: *Given a convex set K in* R^n *with a non-empty interior, determine the ellipsoid of largest volume contained in K*; note that no restrictions are placed on the orientation of the principal axes of the ellipsoids (unlike the situation considered in our problem). John obtained necessary and sufficient conditions for an ellipsoid $\mathcal{E}$ to be the largest ellipsoid contained in the convex set K. If $\mathcal{E}$ is the largest ellipsoid in the convex set K then trivially the smallest that K can be is $\mathcal{E}$. John also considered the following question: *if* $\mathcal{E}$ *is the largest ellipsoid in K then how large can K be*? His answer was, if $\mathcal{E}$ is centered at the origin (WLOG) then K is contained in $n\mathcal{E}$ so $\mathcal{E} \subseteq K \subseteq n\mathcal{E}$, and if K is symmetric then $\mathcal{E} \subseteq K \subseteq \sqrt{n}\mathcal{E}$. Further this result is optimal in the sense that these limits are attained for certain K—see [1].

If K itself is an ellipsoid then the answer to John's largest ellipsoid question is trivial—the largest ellipsoid is K itself. Problem 1 is the determination of the largest ellipsoid, *with axes parallel to the coordinate axes*, contained in the ellipsoid K. As done for John's problem, an interpretation of Corollary 2 provides a characterization of this maximal ellipsoid. The second question posed by John suggests we ask—*if* $x^T Dx = n$ *is the largest volume ellipsoid with axes parallel to coordinate axes which is contained in* $x^T Ax \leqslant n$*, then how large is the ellipsoid* $x^T Ax \leqslant n$. For such A, the ellipsoid $x^T Ax \leqslant n$ may have points with $|x|$ as large as we wish (by changing A), but there are still some restrictions on the size of the ellipsoid $x^T Ax \leqslant n$ in certain directions as stated in the next theorem.

Theorem 3. *If A is a positive definite matrix of order n and* $B \in \mathcal{B}(n_1, \ldots, n_k)$ *is the optimizer of Problem* 2*, then* $\operatorname{block}(A) \geqslant \frac{1}{k}B$*. Further, this estimate is sharp* (*amongst estimates depending only on* $n_1, \ldots, n_k$) *at least when all the blocks are of the same size or if there are only two blocks.*

If B is the optimizer of Problem 2 then $B \geqslant A$ and in particular $B \geqslant \operatorname{block}(A)$; the theorem gives a lower bound on $\operatorname{block}(A)$ and hence places some restrictions on how large the ellipsoid $x^T Ax = n$ can be. The proof of Theorem 3 will also show that if B is the optimizer for Problem 2 then the kernel of $B - A$ has dimension at least $\max(n_1, \ldots, n_k)$.

The characterization in Corollary 2 is still far from satisfactory because we cannot verify easily whether a particular diagonal matrix D is the optimizer of Problem 1; the difficulty is the verification of condition (ii) in the characterization. However, Corollary 2, does provide a mechanism for constructing all matrices $A > 0$ for which a given diagonal matrix $D > 0$ is the solution of Problem 1—see Section 7.

The characterization in Corollary 2 may be used effectively in some special cases, covering a substantial number of matrices, for which the optimizer of Problem 1 is related to the solution of

an interesting non-linear system of algebraic equations $Ax = x^{-1}$ where $x \in R^n$. The following lemma discusses the number of solutions of this equation. This equation and its solutions also play an important role in certain other geometrical problems.

Lemma 4. *Suppose A is a positive definite matrix of order $n \times n$. The equation*

$$Ax = x^{-1} \tag{1}$$

has exactly 2^n solutions in R^n; these solutions are on the ellipsoid $x^T Ax = n$ and there is exactly one solution in each 'quadrant' of R^n.

$D - A$ is singular when D is the optimizer of Problem 1. The next three theorems study the special cases when the nullity of $D - A$ is 1 or $n - 1$ or when the entries of A are "conveniently signed."

Theorem 5 *(Kernel of $D - A$ is one-dimensional). Suppose $A > 0$ is an $n \times n$ matrix.*

(a) *If D is the optimizer of Problem 1 and the kernel of $D - A$ is one-dimensional, then $D = D(x^{-2})$ for any $x \in R^n$ which is a solution of $Ax = x^{-1}$ for which $\Pi(x^2)$ is the smallest (amongst all solutions).*
(b) *Suppose $x \in R^n$ is a solution of $Ax = x^{-1}$ for which $\Pi(x^2)$ is the smallest. Further, suppose $D(x^{-2}) \geqslant A$, then $D(x^{-2})$ is the optimizer of Problem 1.*

Section 7 contains a discussion of situations in which the kernel of $D - A$ is one-dimensional. Theorem 5 asserts that, *in certain situations*, the optimal solution of Problem 1 is obtained by examining the exactly 2^n solutions of (1).

If D is the optimizer of Problem 1 and kernel of $D - A$ is of dimension $n - 1$, then $D - A$ is a rank one matrix which is positive semi-definite, hence $D - A = uu^T$ for some vector u in R^n implying $A = D - uu^T$. The next theorem gives the solution of Problem 1 for such A.

Theorem 6 *(Kernel of $D - A$ is $(n - 1)$-dimensional). Suppose u and q are vectors in R^n with every entry of u non-zero, and $A = D(q) - uu^T > 0$. Then the optimizer for Problem 1 is either $D(q)$ or $D(x^{-2})$ where x is any solution of $Ax = x^{-1}$ with the smallest $\Pi(x^2)$ (amongst all solutions). Further, $D(q)$ is the optimizer iff*

$$2 \max_{i=1,\ldots,n} \frac{|u_i|}{\sqrt{q_i}} \leqslant \sum_{i=1}^{n} \frac{|u_i|}{\sqrt{q_i}}.$$

The next theorem gives the solution of Problem 1 for the large family of "conveniently signed matrices."

Definition. A matrix $A \geqslant 0$ is said to be *conveniently signed* if we can find a vector ϵ with ± 1 entries so that all the entries of the matrix $(a_{ij}\epsilon_i\epsilon_j)$ are non-negative.

Theorem 7 *(Conveniently signed A). For a matrix $A > 0$, A* conveniently signed *with respect to the vector ϵ, the solution of Problem 1 is $D(x^{-2})$ where x is the unique solution of $Ax = x^{-1}$ in the same quadrant as ϵ.*

Problem 1 and its generalization Problem 2 fall into a general category of problems of determinant maximization subject to linear matrix inequality constraints. [6] discusses many applications where such problems arise and also characterizes the optimizer of this family of problems. Our characterization in Theorem 1 is derived from the result in [6]. [6] also discusses a numerical algorithm to solve this family of problems. Solutions of the algebraic equation (1) in the complex domain C^n (where the behavior is quite different) were used by Ball in [2] to obtain results on the complex plank problem.

2. Proof of Theorem 1

Theorem 1 may be derived from an application of the Kuhn–Tucker conditions, or Fenchel duality, or from other results in convex optimization. We have chosen to use Theorem 3.1 in [6] because it seems the shortest route. Theorem 3.1 in [6] is a result specialized to our problem and is a consequence of results from convex optimization.

We convert our problem to the setting in [6]—see (1.1) in [6]. Since $B \geqslant A > 0$ iff $A^{-1} \geqslant B^{-1} > 0$ (see [5, p. 471]), and B is a block matrix iff B^{-1} is a block matrix (with the same structure), we have B_* is the optimizer of Problem 2 iff B_*^{-1} is the optimizer of the following problem.

Problem 3. *Given an $n \times n$ matrix $A > 0$,*

$$\begin{aligned} &\textit{minimize} && -\log\det B \\ &\textit{subject to} && B \in \mathcal{B}(n_1, \ldots, n_k),\ A^{-1} \geqslant B > 0. \end{aligned}$$

Since the map $M \mapsto \log\det M$ is strictly concave on the set of positive definite matrices [5, Theorem 7.6.7] and the constraint set $B \in \mathcal{B}(n_1, \ldots, n_k)$, $A^{-1} \geqslant B > 0$ is convex, Problem 3 has at most one solution. We now prove the existence of an optimizer for Problem 3 (hence for Problem 2) and the characterization condition in Theorem 1.

Identifying $\mathcal{B}(n_1, \ldots, n_k)$ with R^m where $m = (n_1 + \cdots + n_k) + \frac{1}{2}((n_1^2 + \cdots + n_k^2) - (n_1 + \cdots + n_k))$, Problem 3 fits into the framework of Problem (1.1) in [6] with $c = 0$, $F : R^m \to R^{n\times n}$ and $G : R^m \to R^{n\times n}$ with $B = (b_{ij})$ in $\mathcal{B}(n_1, \ldots, n_k)$ defined as follows:

$$G(B) = B = \sum_{i,j} b_{ij} E_{ij}, \qquad F(B) = A^{-1} - B = A^{-1} - \sum_{i,j} b_{ij} E_{ij},$$

where the sums range only over the "block" indices (i, j) for which $i \geqslant j$, and E_{ij} is the $n \times n$ matrix whose only non-zero entries are the (i, j) and (j, i) entries and these entries are 1. By a "block" index (i, j) we mean those indices for which the (i, j)th entry of some matrix in $\mathcal{B}(n_1, \ldots, n_k)$ is non-zero.

From (3.1) in [6] and some computations, we may show that the problem dual to Problem 3 is

$$\begin{aligned} &\text{maximize} && \log\det\operatorname{block}(W) - \operatorname{tr}\big(A^{-1}W\big) + n \\ &\text{subject to} && W \in \mathit{Sym}(n),\ W \geqslant 0,\ \operatorname{block}(W) > 0. \end{aligned}$$

Since the dual problem is strictly feasible, from Theorem 3.1 in [6], the primal optimum is achieved, that is Problem 3 has an optimizer. Further, since the primal problem is also strictly feasible, from the first paragraph of [6, p. 514], a feasible $B \in \mathcal{B}(n_1, \ldots, n_k)$ is the optimizer of Problem 3 iff there exists $Z \geqslant 0$ in $\mathit{Sym}(n)$ so that $(A^{-1} - B)Z = 0$ and $\operatorname{tr}(E_{ij}B^{-1}) + \operatorname{tr}(-E_{ij}Z) = 0$ for all $i \geqslant j$ for which (i, j) is a "block" index, that is $(A^{-1} - B)Z = 0$ and $\operatorname{block}(Z) = B^{-1}$.

So B_* is an optimizer for Problem 2 iff B_*^{-1} is an optimizer for Problem 3, that is iff there is a $Z \geqslant 0$ in $Sym(n)$ so that $(A^{-1} - B_*^{-1})Z = 0$ and $\text{block}(Z) = B_*$. Take $X = B_*^{-1} Z B_*^{-1}$, then X is symmetric, $X \geqslant 0$, and $\text{block}(X) = B_*^{-1}\text{block}(Z)B_*^{-1} = B_*^{-1}$. Further, $Z = B_* X B_*$ so

$$\begin{aligned}\left(A^{-1} - B_*^{-1}\right)Z = 0 \quad &\Longleftrightarrow \quad A^{-1}Z = B_*^{-1}Z \quad \Longleftrightarrow \quad A^{-1}B_*XB_* = B_*^{-1}B_*XB_* \\ &\Longleftrightarrow \quad B_*X = AX \quad \Longleftrightarrow \quad (B_* - A)X = 0.\end{aligned}$$

3. Proof of Theorem 3

We state a simple lemma which we use in the proof of Theorem 3. The lemma gives a condition which is equivalent to condition (ii) in Theorem 1 and Corollary 2 and is also useful in a construction discussed in Section 7. The lemma follows from simple arguments so we will not give its proof.

Lemma 8. *Suppose M and X are symmetric matrices of order n with $M \geqslant 0$. Further, let V be the kernel of M and let $u_1, \ldots, u_m$ be a basis for V. Then the following are equivalent*:

(a) $X \geqslant 0$ *and* $MX = 0$;
(b) *there exist* v_k (*some possibly zero*) *in* V, $k = 1, \ldots, m$, *so that* $X = \sum_{k=1}^m v_k v_k^T$;
(c) *there exists a positive semi-definite matrix* (α_{ij}) *of order* m *so that* $X = \sum_{i,j=1}^m \alpha_{ij} u_i u_j^T$.

3.1. Proof of the lower bound

Our proof is motivated by the proof of John's theorem in [1]. If B is the optimal matrix for Problem 2, for A, then I is the optimal matrix for Problem 2 with A replaced by $B^{-1/2}AB^{-1/2}$, and the statement of Theorem 3 also respects such a modification because $\text{block}(B^{-1/2}AB^{-1/2}) = B^{-1/2}\text{block}(A)B^{-1/2}$. Hence it is enough to prove Theorem 3 for those A for which I is the optimizer of Problem 2. To keep the notation simple, we will prove Theorem 3 only in the special case where $k = 2$, that is B consists of two blocks. The proof for general k follows from obvious modifications to our proof.

So suppose $A > 0$ and is such that I is the optimal matrix for Problem 2. Then, from Theorem 1, $I \geqslant A$ and $I - A$ is singular. So 1 is an eigenvalue of A and is the largest eigenvalue of A. If the kernel of $I - A$ is m-dimensional then A has the spectral decomposition

$$A = \sum_{i=1}^m u_i u_i^T + \sum_{i=m+1}^n \lambda_i u_i u_i^T \tag{2}$$

where $0 < \lambda_i < 1$, the vectors u_i are orthonormal, and u_i, $i = 1, \ldots, m$, are eigenvectors of A corresponding to the eigenvalue 1, and for $i = m+1, \ldots, n$, u_i is an eigenvector corresponding to λ_i. Then from Theorem 1 and Lemma 8

$$I_n = \text{block}\left(\sum_{i,j=1}^m \alpha_{ij} u_i u_j^T\right) \tag{3}$$

for some $m \times m$ matrix $\alpha = (\alpha_{ij}) \geqslant 0$.

Define the $n \times m$ matrix $U = [u_1, \ldots, u_m]$ and let $U = \left[\begin{smallmatrix} P_1 \\ P_2 \end{smallmatrix}\right]$ where the P_i are matrices of order $n_i \times m$. Then the orthonormality of the u_i implies that

$$I_m = U^T U = P_1^T P_1 + P_2^T P_2. \tag{4}$$

Further, (3) is equivalent to

$$I_n = \operatorname{block}\bigl(U\alpha U^T\bigr) = \operatorname{block}\left(\begin{bmatrix} P_1\alpha P_1^T & P_1\alpha P_2^T \\ P_2\alpha P_1^T & P_2\alpha P_2^T \end{bmatrix}\right) = \begin{bmatrix} P_1\alpha P_1^T & 0 \\ 0 & P_2\alpha P_2^T \end{bmatrix}.$$

Hence

$$P_1\alpha P_1^T = I_{n_1}, \qquad P_2\alpha P_2^T = I_{n_2}. \tag{5}$$

From (5), trivially, all the eigenvalues of $P_i\sqrt{\alpha}\sqrt{\alpha}P_i^T$ are 1; hence[2] all the non-zero eigenvalues of $\sqrt{\alpha}P_i^T P_i\sqrt{\alpha}$ are 1. This implies $\sqrt{\alpha}P_i^T P_i\sqrt{\alpha} \leqslant I_m$, for $i = 1, 2$. So multiplying (4) on both sides by $\sqrt{\alpha}$, we obtain

$$\alpha = \sqrt{\alpha}\bigl(P_1^T P_1 + P_2^T P_2\bigr)\sqrt{\alpha} = \sqrt{\alpha}P_1^T P_1\sqrt{\alpha} + \sqrt{\alpha}P_2^T P_2\sqrt{\alpha} \leqslant 2I. \tag{6}$$

Using this back in (5) we obtain $I_{n_i} = P_i\alpha P_i^T \leqslant 2P_i P_i^T$, which implies $\frac{1}{2}I_{n_i} \leqslant P_i P_i^T$ for $i = 1, 2$.

Now, from (2), since $\lambda_i > 0$,

$$A = \sum_{i=1}^{m} u_i u_i^T + \sum_{i=m+1}^{n} \lambda_i u_i u_i^T \geqslant \sum_{i=1}^{m} u_i u_i^T = UU^T = \begin{bmatrix} P_1 P_1^T & P_1 P_2^T \\ P_2 P_1^T & P_2 P_2^T \end{bmatrix}.$$

Hence

$$\operatorname{block}(A) \geqslant \begin{bmatrix} P_1 P_1^T & 0 \\ 0 & P_2 P_2^T \end{bmatrix} \geqslant \frac{1}{2}\begin{bmatrix} I_{n_1} & 0 \\ 0 & I_{n_2} \end{bmatrix} = \frac{1}{2}I_n$$

which concludes the proof of the theorem.

We also note that since P_i is an $n_i \times m$ matrix, (5) implies that $m \geqslant n_i$—this proves the remark after the statement of Theorem 3 in the introduction.

3.2. Tightness of the lower bound

We provide examples to show that the estimate is tight when all the blocks are the same size or when there are only two blocks.

Equal sized blocks. We provide an example with only two equal sized blocks—the generalization to k equal sized blocks will be obvious. Let $n = 2m$ and choose m orthonormal vectors $v_1, \ldots, v_m$ in R^m. Define the orthonormal vectors $u_1, \ldots, u_m$ in R^n as $u_i = \frac{1}{\sqrt{2}}\left[\begin{smallmatrix} v_i \\ v_i \end{smallmatrix}\right]$. Choose an additional m vectors u_i, $i = m+1, \ldots, n$, in R^n so that the u_i, $i = 1, \ldots, n$, are orthonormal; also choose a λ in the interval $(0, 1)$. Define the positive definite matrix of size n with the spectral decomposition

$$A = \sum_{i=1}^{m} u_i u_i^T + \lambda \sum_{i=m+1}^{n} u_i u_i^T. \tag{7}$$

We claim that I is the solution of Problem 2 for A, for all $\lambda < 1$, and $\operatorname{block}(A) - \frac{1}{2}I$ approaches 0 as λ approaches 0.

[2] For any two (possibly rectangular) matrices M and N for which MN and NM makes sense, MN and NM have the same non-zero eigenvalues.

Clearly $I \geqslant A$ and the kernel of $I - A$ is spanned by u_i, $i = 1, \ldots, m$. We note that

$$2\sum_{i=1}^{m} u_i u_i^T = \left[\begin{array}{c|c} \sum_{i=1}^m v_i v_i^T & \sum_{i=1}^m v_i v_i^T \\ \hline \sum_{i=1}^m v_i v_i^T & \sum_{i=1}^m v_i v_i^T \end{array}\right] = \begin{bmatrix} I_m & I_m \\ I_m & I_m \end{bmatrix},$$

which implies that $I_n = \operatorname{block}(2\sum_{i=1}^m u_i u_i^T)$. Hence, by Lemma 8, the conditions of Theorem 1 are satisfied, proving that I is the optimal matrix for A. Finally from the definition of A, we have

$$A = \sum_{i=1}^{m} u_i u_i^T + \sum_{i=m+1}^{n} u_i u_i^T = \frac{1}{2}\begin{bmatrix} I_m & I_m \\ I_m & I_m \end{bmatrix} + \lambda \sum_{i=m+1}^{n} u_i u_i^T.$$

Hence $\operatorname{block}(A) - \frac{1}{2} I_n$ approaches zero as λ approaches 0.

Two blocks. So $k = 2$ and n_1 and n_2 are two positive integers so that $n = n_1 + n_2$. Below e_i and $\bar{e}_i$ will represent vectors in R^{n_1} and R^{n_2} respectively whose ith entry is 1 and all other entries are zero. Define the vector u in R^n as $u = \left[\begin{smallmatrix} e_{n_1} \\ \bar{e}_{n_2} \end{smallmatrix}\right]$; note $\|u\| = \sqrt{2}$. Choose a number α in the interval $(0, \frac{1}{2})$ and take $A = I_n - \alpha u u^T$.

We claim that $I_n \geqslant A > 0$ and I_n is the solution of Problem 2 for A. It is clear that $I_n \geqslant A$ and $A > 0$ because the eigenvalues of A are 1 and $1 - \alpha|u|^2 = 1 - 2\alpha > 0$. Define the following vectors in R^n:

$$p_i = \begin{bmatrix} e_i \\ 0 \end{bmatrix}, \quad q = \begin{bmatrix} e_{n_1} \\ -\bar{e}_{n_2} \end{bmatrix}, \quad r_j = \begin{bmatrix} 0 \\ \bar{e}_j \end{bmatrix}, \quad i = 1, \ldots, n_1 - 1,\ j = 1, \ldots, n_2 - 1.$$

Now

$$\operatorname{block}\left(\sum_{i=1}^{n_1-1} p_i p_i^T + q q^T + \sum_{j=1}^{n_2-1} r_i r_i^T\right) = \begin{bmatrix} I_{n_1} & 0 \\ 0 & I_{n_2} \end{bmatrix} = I_n.$$

Further p_i, q, and r_j are orthogonal to u and hence reside in the kernel of $I - A$. Hence, from Theorem 1 and Lemma 8, I_n is the solution of Problem 2 for A. We also note that

$$\operatorname{block}(A) = \begin{bmatrix} I - \alpha e_{n_1} e_{n_1}^T & 0 \\ 0 & I - \alpha \bar{e}_{n_2} \bar{e}_{n_2}^T \end{bmatrix}$$

and the eigenvalues of $I - \alpha e_{n_1} e_{n_1}^T$ are 1 and $1 - \alpha$; the same is true of $I - \alpha \bar{e}_{n_2} \bar{e}_{n_2}^T$. Hence the largest multiple of I which is a lower bound for $I - \alpha e_{n_1} e_{n_1}^T$ is $(1 - \alpha) I$; the same is true for $I - \alpha \bar{e}_{n_2} \bar{e}_{n_2}^T$. Now α is chosen arbitrarily from $(0, \frac{1}{2})$ so $1 - \alpha$ comes arbitrarily close to $\frac{1}{2}$. Hence the estimate in Theorem 3 is tight, at least in the two block case.

4. Proof of Theorem 5 and Lemma 4

Define $\mathcal{H} = \{h \in R^n\colon\ h > 0,\ \Pi(h) = 1\}$ and if A is a positive definite $n \times n$ matrix then define the ellipsoids

$$\mathcal{A} = \{x \in R^n\colon\ x^T A x = n\}, \qquad \mathcal{A}' = \{x \in R^n\colon\ x^T A^{-1} x = n\}.$$

To prove Theorem 5 we restate Problem 1 as a min–max problem. Every diagonal matrix $D > 0$ has a unique representation $D = \sigma D(h^{-1})$ for a $h \in \mathcal{H}$ and a $\sigma > 0$. Since $\det(D) = \sigma^n$, Problem 1 is the same as

$$\begin{aligned} &\text{minimize} && \sigma^n \\ &\text{subject to} && h \in \mathcal{H},\ \sigma > 0,\ \sigma D\big(h^{-1}\big) \geqslant A. \end{aligned}$$

Now $\sigma D(h^{-1}) \geqslant A$ iff $\sigma A^{-1} \geqslant D(h)$ or equivalently $\sigma n \geqslant x^T D(h)x$ for all $x \in R^n$ with $x^T A^{-1} x = n$. So $\sigma D(h^{-1}) \geqslant A$ iff $n\sigma \geqslant \sup_{x\in\mathcal{A}'} x^T D(h)x$. Hence Problem 1 is equivalent to

Problem 4. *Given $A > 0$, find the optimizer of* $\min_{h\in\mathcal{H}} \max_{x\in\mathcal{A}'} x^T D(h)x$.

For future use we observe that if h is an optimizer of Problem 4 then $D = \sigma D(h^{-1})$ is the optimizer of Problem 1 where σ is the largest eigenvalue of the generalized eigenvalue problem $Ax = \sigma D(h^{-1})x$ because $D - A$ must be singular and $D \geqslant A$. On the other hand, if $D(p)$ is the optimizer of Problem 1 then $h = \Pi(p)p^{-1}$ is the optimizer of Problem 4. Finally

$$\text{optimal value of Problem 4} = n \cdot (\text{optimal value of Problem 1})^{1/n}. \tag{8}$$

Consider the max–min problem associated with the min–max problem, Problem 4.

Problem 5. *Given $A > 0$, find the optimizer of* $\max_{x\in\mathcal{A}'} \min_{h\in\mathcal{H}} x^T D(h)x$.

It may be shown easily that the optimal value of Problem 5 is bounded above by the optimal value of Problem 4—Theorem 5 deals with the special situation where these two optimal values are equal.

Proposition 9 *(Solution of Problem 5). The optimal value of Problem* 5 *is attained at a point* $(x = y^{-1}, h = \Pi(y^{-2})^{1/n} y^2)$ *where* y *is any solution of*

$$Ay = y^{-1} \tag{9}$$

for which $\Pi(y^2)$ *is the smallest* (*amongst all solutions*)*. The optimal value for Problem* 5 *is* $n\Pi(y^{-2})^{1/n}$.

If (x, h) is an optimal point for Problem 5 then $(-x, h)$ is also an optimal point. It is possible that Problem 5 has more than two optimal points; we do not know how these various optimal points are related.

Proof of Proposition 9. Firstly, for a fixed $x = [x_1, \ldots, x_n]^T$, we claim that

$$\inf_{h\in\mathcal{H}} x^T D(h)x = n\big(\Pi\big(x^2\big)\big)^{1/n}.$$

To see this, note that if $h = [h_1, \ldots, h_n]^T$ is in $\mathcal{H}$, then from the AM–GM inequality, we have

$$x^T D(h)x = \sum_{i=1}^n h_i x_i^2 \geqslant n\left(\prod_{i=1}^n h_i \prod_{i=1}^n x_i^2\right)^{1/n} = n\left(\prod_{i=1}^n x_i^2\right)^{1/n}$$

with equality occurring if $h_1 x_1^2 = \cdots = h_n x_n^2 = \lambda$ for some $\lambda \geqslant 0$. So, if all the x_i are non-zero then the minimum value is as claimed above and it occurs when $h_i = \lambda/x_i^2$; since $h_1 \cdots h_n = 1$ we have $\lambda = (x_1^2 \cdots x_n^2)^{1/n}$. Hence, if no coordinate of x is zero then the optimal h is $h = \Pi(x^2)^{1/n} x^{-2}$. If one of the x_i is zero, say $x_1 = 0$ then we can make h_1 large and

$h_2, h_3, \ldots, h_n$ as small as we wish while maintaining $h_1 \cdots h_n = 1$. So $h_1 x_1^2 + \cdots + h_n x_n^2$ may be brought as close to zero as we wish. Hence the claim is true in both cases.

So to solve Problem 5, we need to resolve the equivalent problem

$$\max_{x \in \mathcal{A}'} \Pi\left(x^2\right). \tag{10}$$

This problem clearly attains its supremum and the supremum is attained at a point x with all its components non-zero. From Lagrange multipliers, at the optimal x in $\mathcal{A}'$, we have $A^{-1}x = \mu x^{-1}$ for some μ. Since $x^T A^{-1} x = n$ we have $\mu = 1$; hence any optimizer of (10) is a solution of

$$A^{-1}x = x^{-1}. \tag{11}$$

Further, the maximum is attained at the solutions x of (11) for which $\Pi(x^2)$ is the largest. The corresponding optimal h is $h = \Pi(x^2)^{1/n} x^{-2}$. Taking $y = x^{-1}$ the optimizer of Problem 5 is (x, h) where $x = y^{-1}$, $h = \Pi(y^{-2})^{1/n} y^2$, and y is a solution of $Ay = y^{-1}$ with the smallest $\Pi(y^2)$. □

Proof of Theorem 5. (a) Suppose $D(p) > 0$ is the optimizer of Problem 1 for which the kernel of $D(p) - A$ is one-dimensional. Then, from Corollary 2 and Lemma 8, there is vector x in the kernel of $D(p) - A$ so that $D(p^{-1}) = \operatorname{diag}(xx^T)$ or equivalently $p = x^{-2}$. But $(D(p) - A)x = 0$ hence $Ax = D(p)x = D(x^{-2})x = x^{-1}$.

The optimal value of Problem 1 is $\Pi(p)$, hence from the discussion of the equivalence of Problems 1 and 4, discussed at the beginning of Section 4 (also see (8)), the optimal value of the min–max problem, Problem 4, is $n\Pi(p)^{1/n}$, that is $n\Pi(x^{-2})^{1/n}$. Let y be a solution of $Az = z^{-1}$ with the smallest $\Pi(z^2)$. Then, we have

$$\begin{aligned} n\Pi\left(x^{-2}\right)^{1/n} &= \text{optimal value of Problem 4} \geqslant \text{optimal value of Problem 5} \\ &= n\Pi\left(y^{-2}\right)^{1/n} \geqslant n\Pi\left(x^{-2}\right)^{1/n}, \end{aligned}$$

where the second last relation holds because of Proposition 9, and the last relation holds because of the definition of y. Hence equality holds throughout and $\Pi(x^2) = \Pi(y^2)$ and x is one of the solutions of $Az = z^{-1}$ for which $\Pi(z^2)$ is the smallest.

(b) Suppose x is a solution of $Az = z^{-1}$ with the smallest $\Pi(z^2)$ and also suppose that $D(x^{-2}) \geqslant A$. Then the optimal value of Problem 1 is bounded above by $\Pi(x^{-2})$. So, using (8), we have that the optimal value of Problem 4 is bounded above by $n\Pi(x^{-2})^{1/n}$. Hence

$$\begin{aligned} n\Pi\left(x^{-2}\right)^{1/n} &\geqslant \text{optimal value of Problem 4} \geqslant \text{optimal value of Problem 5} \\ &= n\Pi\left(x^{-2}\right)^{1/n}, \end{aligned}$$

where the last step follows from Proposition 9. Hence there must be equality at all stages of the above equation and we have proved (b). □

Proof of Lemma 4. By definition, any solution x of (1) will have all its coordinates non-zero and $x^T Ax = x^T x^{-1} = n$ implying x is on the ellipsoid $\mathcal{A}$. Let $\mathcal{M}$ be the subset of $\mathcal{A}$ consisting of points none of whose coordinates are zero, that is

$$\mathcal{M} = \left\{x = [x_1, \ldots, x_n]^T \in R^n\colon\ x^T Ax = n,\ x_i \neq 0 \text{ for } i = 1, \ldots, n\right\}.$$

Then $\mathcal{M}$ is a manifold of dimension $n - 1$ with 2^n connected components—the components being the intersections of the ellipsoid $\mathcal{A}$ with the 'quadrants' of R^n. We will show that (1) has

exactly one solution in each component of $\mathcal{M}$; to keep the notation simple $\mathcal{M}$ will represent just one of its components for the rest of this proof.

Define the map $\psi : \mathcal{M} \to R$ with $\psi(x) = \Pi(x^2)$; ψ has an obvious extension to $\overline{\mathcal{M}}$ (the closure of $\mathcal{M}$) and to R^n. It is clear that ψ attains its supremum on $\overline{\mathcal{M}}$ and the supremum is non-zero. Since ψ is zero on the boundary of $\mathcal{M}$ and ψ is differentiable on $\mathcal{M}$, the maximum is attained at a critical point in $\mathcal{M}$, that is at a point $x \in \mathcal{M}$ where $\nabla\psi(x) = \lambda\nabla(x^T Ax - n)$ which is equivalent to $2\psi(x)x^{-1} = \lambda Ax$. One may see that $\lambda \neq 0$ and hence $Ax = x^{-1}$ because $x^T Ax = n$ since $x \in \mathcal{M}$. Hence, the largest value of ψ on $\mathcal{M}$ occurs at a point $x \in \mathcal{M}$ which is a solution of (1), implying (1) has a solution in $\mathcal{M}$. It remains to show that this is the only solution of (1) in $\mathcal{M}$.

If x and y are two solutions of (1) in $\mathcal{M}$ then $Ax = x^{-1}$, $Ay = y^{-1}$, and $x_i/y_i > 0$ for $i = 1, \ldots, n$. Consider

$$\begin{aligned}(x-y)^T A(x-y) &= x^T Ax + y^T Ay - x^T Ay - y^T Ax \\ &= x^T x^{-1} + y^T y^{-1} - x^T y^{-1} - y^T x^{-1} \\ &= n + n - \sum_{i=1}^{n} \left(x_i y_i^{-1} + y_i x_i^{-1}\right) \leqslant 2n - 2\sum_{i=1}^{n} \sqrt{\left(x_i y_i^{-1}\right)\left(y_i x_i^{-1}\right)} \\ &= 0.\end{aligned}$$

But $A > 0$, so $x = y$. □

5. Proof of Theorem 6

For vectors u and q, $A = D(q) - uu^T$ is positive definite only if $q > 0$. So taking $q > 0$, $D(q) - uu^T > 0$ iff $I - vv^T > 0$ where $v = q^{-1/2} \circ u$. Now the eigenvalues of $I - vv^T$ are 1 and $1 - |v|^2$; hence $A = D(q) - uu^T > 0$ iff $q > 0$ and $\sum_{i=1}^{n} \frac{u_i^2}{q_i} < 1$. Since no entry of u is zero, $D(p)$ is optimal for $A = D(q) - uu^T$ iff $D(p \circ u^{-2})$ is optimal for $D(q \circ u^{-2}) - ee^T$. Further x is a solution of $(D(q) - uu^T)x = x^{-1}$ iff $y = D(u)x$ is a solution of $(D(q \circ u^{-2}) - ee^T)y = y^{-1}$. Hence we need to prove Theorem 6 only when $u = e$.

5.1. The case when $D(q)$ is not the minimizer

Suppose $D(p)$ is the solution of Problem 1 for $A = D(q) - ee^T$ and $D(p) \neq D(q)$. Then some entry of p must be strictly smaller than the corresponding entry of q otherwise $D((p+q)/2)$ would be a strictly better candidate than the optimal $D(p)$. We claim that the nullity of $D(p) - A$ is 1 and this claim combined with Theorem 5 proves Theorem 6.

To prove that the nullity of $D(p) - A$ is 1, we define $r = p - q$—note that at least one entry of $r = p - q$ is negative. Actually, exactly one entry of r is negative because if (say) $r_1 < 0$, $r_2 < 0$, then taking $x = [1, -1, 0, 0, \ldots]$, we have

$$x^T\left(D(r) + ee^T\right)x = \left(x^T e\right)^2 + x^T D(r)x = 0 + r_1 + r_2 < 0$$

which contradicts the fact that $D(r) + ee^T = D(p) - (D(q) - ee^T) \geqslant 0$.

Without loss of generality let $\bar{r} = [r_1, \ldots, r_{n-1}]^T$ be positive and $r_n < 0$. If $\bar{e}$ is the $n-1$ vector of 1's, then

$$D(p) - A = D(r) + ee^T = \begin{bmatrix} D(\bar{r}) + \bar{e}\bar{e}^T & \bar{e} \\ \bar{e}^T & 1 + r_n \end{bmatrix}. \tag{12}$$

We must have $1 + r_n > 0$ because taking $x = [\bar{e}^T, x_n]^T$ we have

$$0 \leqslant x^T \big(D(p) - A\big)x = \bar{e}^T \big(D(\bar{r}) + \bar{e}\bar{e}^T\big)\bar{e}^T + (1 + r_n)x_n^2 + 2(n-1)x_n$$

for all x_n, and the right-hand side may be made negative, if $1 + r_n \leqslant 0$, by choosing x_n to be a large negative number. Next, using the representation (12),

$$\begin{bmatrix} I & -(1+r_n)^{-1}\bar{e} \\ 0 & I \end{bmatrix} \big(D(p) - A\big) \begin{bmatrix} I & 0 \\ -(1+r_n)^{-1}\bar{e} & I \end{bmatrix} = \begin{bmatrix} D(\bar{r}) + \frac{r_n}{1+r_n}\bar{e}\bar{e}^T & 0 \\ 0 & 1 + r_n, \end{bmatrix} \tag{13}$$

so the nullity of $D(p) - A$ is the same as the nullity of $D(\bar{r}) + \frac{r_n}{1+r_n}\bar{e}\bar{e}^T$ which is the same as the nullity of $I + \frac{r_n}{1+r_n} v v^T$ where $v = \sqrt{\bar{r}^{-1}}$—note all entries of $\bar{r}$ are positive. Now the eigenvalues of the matrix $I + \sigma v v^T$ are 1 (of multiplicity $n-1$) and $1 + \sigma|v|^2$ (the eigenvectors are $v^\perp$ and v); hence the nullity of $D(p) - A$ is at most 1. But the kernel of $D(p) - A$ is not empty, hence the nullity of $D(p) - A$ is 1.

5.2. The case when $D(q)$ is the minimizer

We now determine the q for which $D(q)$ is the solution of Problem 1 for $A = D(q) - ee^T > 0$. If V is the kernel of $D(q) - A = ee^T$ then V consists of all vectors orthogonal to e. V has a basis $\{v_1, \ldots, v_{n-1}\}$ where v_i is the vector whose ith component is 1, its nth component is -1, and all other components are zero. From Corollary 2 and Lemma 8, $D(q)$ is the minimizer iff we can find an $(n-1) \times (n-1)$ matrix $(\alpha_{ij}) \geqslant 0$ so that $D(q^{-1}) = \operatorname{diag}(\sum_{i,j} \alpha_{ij} v_i v_j^T)$ or equivalently $q^{-1} = [\alpha_{11}, \ldots, \alpha_{n-1n-1}, \sum_{ij} \alpha_{ij}]^T$ for some $(n-1) \times (n-1)$ matrix $(\alpha_{ij}) \geqslant 0$. In Lemma 10 below, taking $n-1$ instead of n, and taking $d_i = 1/\sqrt{q_i}$, $i = 1, \ldots, n-1$, we have $D(q)$ is the minimizer iff $m \leqslant 1/\sqrt{q_n} \leqslant M$ where

$$M = \sum_{i=1}^{n-1} \frac{1}{\sqrt{q_i}}, \quad m = \max\left(0, 2 \max_{i=1,\ldots,n-1} \frac{1}{\sqrt{q_i}} - M\right).$$

This may be stated more succinctly as $D(q)$ is the optimizer for $D(q) - ee^T$ iff

$$2 \max_{i=1,\ldots,n} \frac{1}{\sqrt{q_i}} \leqslant \sum_{i=1}^{n} \frac{1}{\sqrt{q_i}}.$$

Lemma 10. *Given the non-negative numbers $d_1, \ldots, d_n$, let $\mathcal{P}$ denote the set of all real valued, positive semi-definite $n \times n$ matrices $P = (p_{ij})$ with $p_{ii} = d_i^2$, $i = 1, \ldots, n$. Then the range of the map from $\mathcal{P}$ to R given by $P = (p_{ij}) \mapsto \sum_{i,j} p_{ij}$ is $[m^2, M^2]$ where $M = \sum_i d_i$ and $m = \max(0, 2\max_i d_i - \sum_i d_i)$.*

Proof. Since $\mathcal{P}$ is convex and hence a connected subset of R^{n^2} and the map $P = (p_{ij}) \mapsto \sum_{ij} p_{ij}$ is continuous and real valued, the range must be an interval. We have to find the end points of this interval.

Any $P \in \mathcal{P}$ may be written as $P = (v_i^T v_j)$ for vectors $v_1, \ldots, v_n$ in R^n with $|v_i| = d_i$. Further the sum of the entries of P is $\sum_{i,j} v_i^T v_j = (\sum_i v_i^T)(\sum_j v_j) = |\sum_i v_i|^2$. Define

$$\mathcal{V} = \big\{(v_1, \ldots, v_n)\colon\ v_i \in R^n,\ |v_i| = d_i,\ i = 1, \ldots, n\big\}.$$

Then Lemma 10 follows if we determine the range of the map from $\mathcal{V}$ to R defined by

$$(v_1, \ldots, v_n) \mapsto |v_1 + \cdots + v_n|. \tag{14}$$

The range is a closed interval because the domain is compact. For $(v_1, \ldots, v_n) \in \mathcal{V}$, we have $|v_1 + \cdots + v_n| \leqslant |v_1| + \cdots + |v_n| = d_1 + \cdots + d_n$ and equality occurs iff the v_i are non-negative multiples of a fixed vector. Hence the right end point of the range interval is $d_1 + \cdots + d_n$ and is attained only at points of the form $(d_1 u, \ldots, d_n u)$ for some unit vector u in R^n.

Now we seek the minimizers of the map (14). Without loss of generality we may assume that $d_1 = \max_i d_i$ and we define $m(d_1, \ldots, d_k)$ and $M(d_1, \ldots, d_k) = d_1 + \cdots + d_k$ to be the left and right end points of the range of the map $(v_1, \ldots, v_k) \mapsto |v_1 + \cdots + v_k|$ where the v_i are restricted to vectors in R^n with $|v_i| = d_i$. Clearly $d_{k+1} \leqslant d_1 \leqslant M(d_1, \ldots, d_k)$. If $d_{k+1} \geqslant m(d_1, \ldots, d_k)$ then d_{k+1} is in the range of the map $(v_1, \ldots, v_k) \mapsto |v_1 + \cdots + v_k|$ and we can find vectors $v_1, \ldots, v_k$ so that $|v_i| = d_i$, $i = 1, \ldots, k$ and $|v_1 + \cdots + v_k| = d_{k+1}$. Let $v_{k+1} = -(v_1 + \cdots + v_k)$, note $|v_{k+1}| = d_{k+1}$. Then $|v_1 + \cdots + v_{k+1}| = 0$ and hence $m(d_1, \ldots, d_{k+1}) = 0$. On the other hand, if $d_{k+1} < m(d_1, \ldots, d_k)$ then

$$|v_1 + \cdots + v_{k+1}| \geqslant |v_1 + \cdots + v_k| - |v_{k+1}| \geqslant m(d_1, \ldots, d_k) - d_{k+1}$$

with equality occurring iff $(v_1, \ldots, v_k)$ is a minimizer giving $m(d_1, \ldots, d_k)$ and $v_{k+1} = -\alpha(v_1 + \cdots + v_k)$ where $\alpha = d_{k+1}/m(d_1, \ldots, d_k)$. In this case $m(d_1, \ldots, d_{k+1}) = m(d_1, \ldots, d_k) - d_{k+1}$. Summarizing, we have $m(d_1, \ldots, d_{k+1}) = \max(0, m(d_1, \ldots, d_k) - d_{k+1})$. From this one may observe that

$$m(d_1, \ldots, d_n) = \max(0, d_1 - d_2 - d_3 - \cdots - d_n) = \max\left(0, 2d_1 - \sum_{i=1}^{n} d_i\right)$$

which proves Lemma 10. □

6. Proof of Theorem 7

Suppose $A > 0$ is *conveniently signed* with respect to the vector ϵ. Let x be the unique solution of $Ax = x^{-1}$ in the same quadrant as ϵ. Since the entries of $(a_{ij}\epsilon_i\epsilon_j)$ have the same sign as the entries of $(a_{ij}x_i x_j)$, we may conclude that the entries of $M = D(x)AD(x)$ are non-negative. Further, $Ax = x^{-1}$ implies that $\sum_{j=1}^{n} a_{ij}x_i x_j = 1$ for each $i = 1, \ldots, n$, and hence the row sums of M are 1. So M is a symmetric matrix with non-negative entries and its row sums are 1. We will show that the solution of Problem 1 for M (instead of A) is I; hence the optimal matrix for Problem 1 for A will be $D(x^{-2})$. Note that $M > 0$ because $A > 0$.

Since the row sums of M are 1 and all entries of $M = (m_{ij})$ are non-negative, we have $1 - m_{ii} = \sum_{j=1, j\neq i}^{n} m_{ij} = \sum_{j=1, j\neq i}^{n} |m_{ij}|$ for $i = 1, \ldots, n$, which implies that $I - M$ is diagonally dominant and hence $I - M \geqslant 0$—see [5, p. 349]. Hence $I \geqslant M$. Further, since the row sums of M are 1, e is in the kernel of $I - M$ and $I = \operatorname{diag}(ee^T)$. Hence from Corollary 2 and Lemma 8, I is the optimal matrix for Problem 1 for M. This completes the proof of Theorem 7.

7. Discussion of results

Problem 1 may also be considered as the characterization of all positive definite matrices A for which a given positive diagonal matrix D is the solution of Problem 1. One may easily show

that this is equivalent to the characterization of all $A > 0$ for which I is the solution of Problem 1, or equivalently the characterization of all ellipsoids $\mathcal{A}$: $x^T Ax = 1$ for which the optimal ellipsoid is the unit sphere.

While we do not have a procedure to check whether I is the optimal matrix for Problem 1 for a given $A > 0$, the procedure below (based on Corollary 2) will construct all $A > 0$ for which I is the optimal matrix for Problem 1. Choose n unit vectors $u_1, \ldots, u_n$ in R^n and define the $n \times n$ matrix $M = [u_1, \ldots, u_n]$. Choose an orthonormal basis $\{v_1, \ldots, v_k\}$ for the range of M^T (the row space of M) and complete it to an orthonormal basis $\{v_1, \ldots, v_n\}$ for R^n. Next, choose real numbers $\lambda_1, \ldots, \lambda_n$ with $\lambda_i = 1$ for $i = 1, \ldots, k$ and $0 < \lambda_i \leqslant 1$ for $i = k+1, \ldots, n$ and take A to be the positive definite matrix with the spectral decomposition $A = \sum_{i=1}^n \lambda_i v_i v_i^T$.

From the spectral decomposition we see that $I \geqslant A$ and $(I - A)M^T = 0$ implying $(I - A)X = 0$ where $X = M^T M = (u_i^T u_j) \geqslant 0$. Further $\operatorname{diag}(X) = I = I^{-1}$ because the u_i are unit vectors. Hence Corollary 2 implies that I is the solution of Problem 1 for the A constructed above.

Conversely, suppose I is the optimal solution of Problem 1 for a given A. From Corollary 2, $I \geqslant A$, $I - A$ is singular, and there is an $X \geqslant 0$ so that $(I - A)X = 0$ and $\operatorname{diag}(X) = I$. Let $X = (u_i^T u_j)$ for some unit vectors $u_1, \ldots, u_n$ in R^n and we define $M = [u_1, \ldots, u_n]$. We claim that $(I - A)M^T = 0$; assuming this for the moment, the range of M^T is in the eigenspace of A corresponding to the eigenvalue 1. Choose an orthonormal basis $v_1, \ldots, v_k$ for the range of M^T and let $v_{k+1}, \ldots, v_n$ be the "remaining" unit length, linearly independent, eigenvectors of A. Then $A = \sum_{i=1}^n \lambda_i v_i v_i^T$ with $\lambda_i = 1$ for $i = 1, \ldots, k$ and $0 < \lambda_i \leqslant 1$ for $i = k+1, \ldots, n$ (because $I \geqslant A > 0$).

It remains to show that $(I - A)M^T = 0$. We have $M(I - A)M^T \geqslant 0$ because $I \geqslant A$; also $\operatorname{tr}(M(I - A)M^T) = \operatorname{tr}((I - A)M^T M) = \operatorname{tr}((I - A)X) = 0$. So all the eigenvalues of $M(I - A)M^T$ are zero and hence $M(I - A)M^T = 0$. Noting that $0 = M(I - A)M^T = M\sqrt{I - A}\sqrt{I - A}M^T = P^T P$ where $P = \sqrt{I - A}M^T$, we have $P = 0$. Hence $(I - A)M^T = \sqrt{I - A}P = 0$.

Theorems 5, 6 and Lemma 8 generate an explicit characterization in the $n = 3$ case of all $A > 0$ for which the optimal matrix for Problem 1 is a multiple of I (that is the optimal ellipsoid is a sphere). For a 3×3 matrix $A > 0$, the optimal ellipsoid will be a sphere iff one of the following holds:

- A is a multiple of I;
- the eigenspace corresponding to the largest eigenvalue of A is one-dimensional and $v^2 = e$ for some vector v in this eigenspace;
- the eigenspace corresponding to the largest eigenvalue of A is two-dimensional and if $u = [u_1, u_2, u_3]^T$ is the eigenvector corresponding to the other/smallest eigenvalue of A, then $2\max_i |u_i| \leqslant \sum_i |u_i|$.

References

[1] K. Ball, An Elementary Introduction to Modern Convex Geometry, Flavors of Geometry, vol. 31, MSRI Publications, 1997.

[2] K. Ball, The complex plank problem, Bull. London Math. Soc. 33 (2001) 433–442.

[3] L. Devroye, Nonuniform Random Variate Generation, Springer, New York, 1986.

[4] F. John, Extremum problems with inequalities as subsidiary conditions, in: Courant Anniversary Volume, Interscience, New York, 1948, pp. 187–204.
[5] R.A. Horn, C.R. Johnson, Matrix Analysis, Cambridge Univ. Press, 1999.
[6] L. Vandenberghe, S. Boyd, S. Wu, Determinant maximization with linear matrix inequality constraints, SIAM J. Matrix Anal. Appl. 19 (2) (1988) 499–533.