Available online at www.sciencedirect.com
SCIENCE DIRECT®
Linear Algebra and its Applications 416 (2006) 815–834
LINEAR ALGEBRA AND ITS APPLICATIONS
www.elsevier.com/locate/laa

Optimization of generalized mean square error in signal processing and communication ☆

William W. Hager [a,*], Yong Liu [b], Tan F. Wong [b]

[a] *Department of Mathematics, University of Florida, P.O. Box 118105, Gainesville, FL 32611-8105, United States*
[b] *Department of Electrical and Computer Engineering, University of Florida, P.O. Box 116130, Gainesville, FL 32611-6130, United States*

Received 26 April 2005; accepted 27 December 2005
Available online 14 February 2006
Submitted by U. Rothblum

Abstract

Two matrix optimization problems are analyzed. These problems arise in signal processing and communication. In the first problem, the trace of the mean square error matrix is minimized, subject to a power constraint. The solution is the training sequence, which yields the best estimate of a communication channel. The solution is expressed in terms of the eigenvalues and eigenvectors of correlation and covariance matrices describing the communication, and an unknown permutation. Our analysis exhibits the optimal permutation when the power is either very large or very small. Based on the structure of the optimal permutation in these limiting cases, we propose a small class of permutations to focus on when computing the optimal permutation for arbitrary power. In numerical experiments, with randomly generated matrices, the optimal solution is contained in the proposed permutation class with high probability.

The second problem is connected with the optimization of the sum capacity of a communication channel. The second problem, which is obtained from the first by replacing the trace operator in the objective function by the determinant, minimizes the product of eigenvalues, while the first problem minimizes the sum of eigenvalues. For small values of the power, both problems have the same solution. As the power increases, the solutions are different, since the permutation matrix appearing in the solution of the trace problem is not

☆ This material is based upon work supported by the National Science Foundation under Grants 0203270 and ANI-0020287.

* Corresponding author. Tel.: +1 352 392 0281; fax: +1 352 392 8357.
E-mail addresses: hager@math.ufl.edu (W.W. Hager), yongliu@ufl.edu (Y. Liu), twong@ece.ufl.edu (T.F. Wong).
URL: http://www.math.ufl.edu/∼hager (W.W. Hager).

doi:10.1016/j.laa.2005.12.024

present in the solution of the determinant problem. For large power, the ordering of the eigenvectors in the solution of the trace problem is the opposite of the ordering in the determinant problem.

AMS classification: 60G35; 93E10; 94A15

Keywords: Trace; Mean square error; MSE; Capacity; Determinant; Optimization; MIMO systems; Wireless communication

1. Introduction

In this paper, we analyze two matrix optimization problems, which arise in signal processing and wireless communication. In the first problem, we minimize the trace, denoted "tr", of a matrix:

$$\min_{\mathbf{S}} \operatorname{tr}(\mathbf{DS}^*\mathbf{QSD}+\mathbf{I})^{-1} \quad \text{subject to} \quad \operatorname{tr}(\mathbf{S}^*\mathbf{S}) \leqslant P, \quad \mathbf{S} \in \mathbb{C}^{m\times n}. \tag{1.1}$$

Here $\mathbf{Q}$ and $\mathbf{D}$ are nonzero Hermitian, positive semidefinite matrices, and the positive scalar P is the power constraint associated with the signal $\mathbf{S}$. The inverse matrix

$$\mathbf{C} = (\mathbf{DS}^*\mathbf{QSD}+\mathbf{I})^{-1}$$

appearing in the cost function is the mean square error (MSE) matrix; it is the covariance of the best estimate of the matrix representing a communication channel. Application areas include joint linear transmitter–receiver design [15,17], training sequence design for channel estimation in multiple antenna communication systems [25], and spreading sequence optimization for code division multiple access (CDMA) communication systems [22].

The solution in the special case $\mathbf{D} = \mathbf{I}$, found for example in [17,25], can be expressed in terms of the eigenvalues and eigenvectors of $\mathbf{Q}$ and a Lagrange multiplier associated with the power constraint. In the applications introduced in this paper, $\mathbf{D} \neq \mathbf{I}$ and minimizing the trace of the mean square error is more difficult. We will show that (1.1) has a solution which can be expressed $\mathbf{S} = \mathbf{U\Sigma V}^*$ where $\mathbf{U}$ and $\mathbf{V}$ are orthonormal matrices of eigenvectors for $\mathbf{Q}$ and $\mathbf{D}$, respectively, and $\mathbf{\Sigma}$ is diagonal. Solving (1.1) involves computing diagonalizations of $\mathbf{Q}$ and $\mathbf{D}$, and finding an ordering for the columns of $\mathbf{U}$ and $\mathbf{V}$. We are able to evaluate the optimal ordering when either P is large or P is small. However, for intermediate values of P, the problem (1.1) has a combinatorial nature, unlike the special case $\mathbf{D} = \mathbf{I}$.

In the second matrix optimization problem, we maximize the determinant, denoted "det", of a matrix:

$$\max_{\mathbf{S}} \det(\mathbf{DS}^*\mathbf{QSD}+\mathbf{I}) \quad \text{subject to} \quad \operatorname{tr}(\mathbf{S}^*\mathbf{S}) \leqslant P, \quad \mathbf{S} \in \mathbb{C}^{m\times n}. \tag{1.2}$$

The cost function is related to the sum capacity of a CDMA communication channel. Since the determinant of the inverse of a matrix is the reciprocal of the determinant of the matrix, it follows that problem (1.2) is equivalent to replacing trace by determinant in (1.1). Hence, in the original problem (1.1), we minimize the sum of the eigenvalues of the MSE matrix $\mathbf{C}$, while in the second problem (1.2), we minimize the product of the eigenvalues of $\mathbf{C}$. In either case, we try to make the eigenvalues of $\mathbf{C}$ small, but with different metrics.

In the special case $\mathbf{D} = \mathbf{I}$, the solution of (1.2) can be found for example in [19], while for the special case $\mathbf{Q} = \mathbf{I}$, the solution of (1.2) can be found in [26]. For the more general problem (1.2), we again show that the solution can be expressed $\mathbf{S} = \mathbf{U\Sigma V}^*$, where $\mathbf{U}$ and $\mathbf{V}$ are orthonormal matrices of eigenvectors for $\mathbf{Q}$ and $\mathbf{D}$, respectively, and $\mathbf{\Sigma}$ is diagonal. Unlike the trace problem (1.1), the ordering of the columns of $\mathbf{U}$ and $\mathbf{V}$ does not depend on the power P—the columns of $\mathbf{U}$ and $\mathbf{V}$ should be ordered so that the associated eigenvalues of $\mathbf{Q}$ and $\mathbf{D}$ are in decreasing order. Also note that in the paper [2], the authors consider essentially the same optimization problem in the context of a space-time spreading scheme for correlated fading channels in the presence of interference. Based on the previous known result for the special case $\mathbf{Q} = \mathbf{I}$ [26], the authors propose to optimize $\mathbf{S}$ over the set of matrices of the form $\mathbf{S} = \mathbf{U\Pi\Sigma V}^*$; they do not show that the optimal $\mathbf{S}$ has this structure. Here we prove that the optimal solution does indeed have this structure.

The paper is organized as follows: In Section 2, we present some applications which lead to the optimization problems (1.1) and (1.2). In Section 3, we show that (1.1) has a solution of the form $\mathbf{S} = \mathbf{U\Sigma V}^*$ where the columns of $\mathbf{U}$ and $\mathbf{V}$ are eigenvectors of $\mathbf{Q}$ and $\mathbf{D}$, respectively, and $\mathbf{\Sigma}$ is diagonal. In Section 4, we obtain the optimal $\mathbf{\Sigma}$ assuming the optimal ordering of the columns of $\mathbf{U}$ and $\mathbf{V}$ is known. In Section 5, the optimal ordering of the columns is determined when P is either large or small. In Section 6, majorization theory is used to derive the optimal solution of the determinant problem (1.2). Finally, in Section 7, we compare the solutions of the trace problem and the determinant problem. Also, we propose a class of permutations in which to search for the optimal solution of the trace problem (1.1). Numerical experiments indicate that with high probability, the optimal permutation is contained in this class.

2. Applications

In this section we briefly describe a few of the many applications which lead to optimization problems of the form (1.1) and (1.2).

2.1. Channel estimation for multiple antenna communication systems

The first application arises when the channel matrix in a multiple antenna communication system is estimated. Let us consider the communication from the base station to the mobile unit of a cellular system (the downlink of the cellular system). It is assumed that the base station has multiple transmit antennas, while the mobile unit has a single antenna due to the space limit of the handset. At the beginning of each data packet, a training sequence is inserted to aid the channel estimation at the receiver. During the training period, the received baseband signal at the mobile unit satisfies

$$\mathbf{y} = \mathbf{Sh} + \mathbf{e}, \tag{2.1}$$

where $\mathbf{y}$ is an $N_r \times 1$ vector containing the N_r received symbols, $\mathbf{h}$ is a $t \times 1$ vector containing the t channel gains from the transmit antennas to the receiver, $\mathbf{S}$ is the $N_r \times t$ training symbol matrix associated with the tN_r training symbols, the ith column of $\mathbf{S}$ is the training signal transmitted by the ith antenna, and $\mathbf{e}$ represents additive Gaussian noise with zero mean and covariance $\mathbf{E}$ which models the co-channel interference from other cells. Usually, N_r is much larger than t. It is assumed that $\mathbf{S}$ is known by the receiver.

We adopt the one-ring correlated channel model [18] which models the channel gain vector $\mathbf{h}$ as follows:

$$\mathbf{h} = \mathbf{R}^{1/2}\mathbf{h}_w, \tag{2.2}$$

where $\mathbf{R}$ is the correlation between the transmit antennas and $\mathbf{h}_w$ is a $t \times 1$ vector whose elements are independent, identically distributed, zero-mean, complex circular-symmetric Gaussian random variables with unit variance. After substituting (2.2) in (2.1), the baseband signal received at the mobile unit can be expressed as

$$\mathbf{y} = \mathbf{S}\mathbf{R}^{1/2}\mathbf{h}_w + \mathbf{e}.$$

By the Bayesian Gauss–Markov Theorem [11], the minimum mean square error estimator (MMSE) for $\mathbf{h}_w$ is

$$\hat{\mathbf{h}}_w = (\mathbf{R}^{1/2}\mathbf{S}^*\mathbf{E}^{-1}\mathbf{S}\mathbf{R}^{1/2} + \mathbf{I})^{-1}\mathbf{R}^{1/2}\mathbf{S}^*\mathbf{E}^{-1}\mathbf{y}.$$

The performance of the estimator is measured by the error $\boldsymbol{\epsilon} = \mathbf{h}_w - \hat{\mathbf{h}}_w$ with mean zero and covariance

$$\mathbf{C}_\epsilon = (\mathbf{R}^{1/2}\mathbf{S}^*\mathbf{E}^{-1}\mathbf{S}\mathbf{R}^{1/2} + \mathbf{I})^{-1}.$$

The diagonal elements of the error covariance matrix $\mathbf{C}_\epsilon$ yield the minimum MSE for the estimation of the corresponding parameters of $\mathbf{h}_w$. Observe that $\mathbf{C}_\epsilon$ depends on the choice of the training symbol matrix $\mathbf{S}$. In the following optimization problem, we search for the training sequences, which achieve the best estimation performance under a transmission power constraint:

$$\begin{aligned} &\min_{\mathbf{S}} \operatorname{tr}(\mathbf{R}^{1/2}\mathbf{S}^*\mathbf{E}^{-1}\mathbf{S}\mathbf{R}^{1/2} + \mathbf{I})^{-1} \\ &\text{subject to} \quad \operatorname{tr}(\mathbf{S}^*\mathbf{S}) \leqslant P, \quad \mathbf{S} \in \mathbb{C}^{m\times n}. \end{aligned} \tag{2.3}$$

This optimization problem has the same form as (1.1) with $\mathbf{D} = \mathbf{R}^{1/2}$ and $\mathbf{Q} = \mathbf{E}^{-1}$.

2.2. *Spreading sequence optimization for CDMA systems*

The trace problem (1.1) also arises in spreading sequence optimization for code division multiple access (CDMA) systems. In cellular communication systems, multiple access schemes allow many users to share simultaneously a finite amount of radio resources. CDMA is one of the main access techniques. It is adopted in the IS-95 system and will be used in next generation cellular communication systems [23]. In a CDMA system, different users are assigned different spreading sequences so that the users can share the communication channel. We consider the uplink (communication from the mobile units to the base station) of a CDMA system where the users within a base station are symbol synchronous. The co-channel interference from the users in the neighboring cells are modeled by additive, colored Gaussian noise. The received signal at the base station is

$$\mathbf{y} = \sum_{i=1}^{K} h_i \mathbf{s}_i x_i + \mathbf{e},$$

where K is the number of signals received by the base station, x_i is the symbol transmitted from the ith user, $\mathbf{S}_i \in \mathbb{C}^{N_r}$ is the spreading sequence assigned to the ith user, h_i is the channel gain from the ith user to the base station, and $\mathbf{e} \in \mathbb{C}^{N_r}$ is the additive, colored Gaussian noise with zero mean and covariance $\mathbf{E}$. Usually the size of K and N_r are comparable. It is assumed that the symbols x_i are independent with zero mean and unit variance. The received signal can be expressed as

$$\mathbf{y} = \mathbf{SHx} + \mathbf{e}, \tag{2.4}$$

where $\mathbf{S}$, the spreading sequence matrix, has jth column $\mathbf{s}_j$, and $\mathbf{H}$ is a diagonal matrix with ith diagonal element h_i. Again, by the Bayesian Gauss–Markov Theorem [11,13], the MMSE estimator of $\mathbf{x}$ is

$$\hat{\mathbf{x}} = (\mathbf{H}^*\mathbf{S}^*\mathbf{E}^{-1}\mathbf{SH} + \mathbf{I})^{-1}\mathbf{H}^*\mathbf{S}^*\mathbf{E}^{-1}\mathbf{y}.$$

The corresponding covariance matrix of the estimation error is

$$\mathbf{C}_\epsilon = (\mathbf{H}^*\mathbf{S}^*\mathbf{E}^{-1}\mathbf{SH} + \mathbf{I})^{-1}.$$

The optimal spreading sequences for all the users which minimizes the co-channel interference to other cells, subject to a power constraint, corresponds to (1.1) with $\mathbf{Q} = \mathbf{E}^{-1}$ and $\mathbf{D} = \mathbf{H}$, a diagonal matrix.

2.3. Channel capacity for CDMA systems

For CDMA systems, a different performance measure, which arises in information theory, is the sum capacity of the channel. The sum capacity is a performance measure for coded systems. It represents the maximum sum of the rates at which users can transmit information reliably. The sum capacity of the synchronous multiple access channel (2.4) is

$$C_{\text{sum}} = \max I(x_1, \ldots, x_K; \mathbf{y}),$$

where I represents the mutual information [3] between the inputs x_i and the output vector $\mathbf{y}$. The maximization is over the independent random inputs x_i. For the Gaussian channel in (2.4), the maximum is achieved when all the random inputs are Gaussian [3]. In this case, the sum capacity [21,22] becomes

$$C_{\text{sum}} = \frac{1}{2N_r} \log\det(\mathbf{H}^*\mathbf{S}^*\mathbf{E}^{-1}\mathbf{SH} + \mathbf{I}).$$

Since log is a monotone increasing function, the maximization of the sum capacity, subject to a power constraint, corresponds to the optimization problem (1.2) with $\mathbf{Q} = \mathbf{E}^{-1}$ and $\mathbf{D} = \mathbf{H}$.

3. Solution structure for the trace problem

We begin by analyzing the structure of an optimal solution to (1.1). Let $\mathbf{U\Lambda U}^*$ and $\mathbf{V\Delta V}^*$ be diagonalizations of $\mathbf{Q}$ and $\mathbf{D}$, respectively (the columns of $\mathbf{U}$ and $\mathbf{V}$ are orthonormal eigenvectors). Let λ_j, $1 \leqslant j \leqslant m$, and δ_i, $1 \leqslant i \leqslant n$, denote the diagonal elements of $\mathbf{\Lambda}$ and $\mathbf{\Delta}$, respectively. We assume that the eigenvalues are arranged in decreasing order:

$$\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_m \quad \text{and} \quad \delta_1 \geqslant \delta_2 \geqslant \cdots \geqslant \delta_n. \tag{3.1}$$

Let us define

$$\mathbf{T} = \mathbf{U}^*\mathbf{SV}. \tag{3.2}$$

Making the substitution $\mathbf{S} = \mathbf{UTV}^*$ in (1.1) yields the following equivalent problem:

$$\begin{aligned} &\min \operatorname{tr}(\mathbf{\Delta T}^*\mathbf{\Lambda T\Delta} + \mathbf{I})^{-1} \\ &\text{subject to} \quad \operatorname{tr}(\mathbf{T}^*\mathbf{T}) \leqslant P, \quad \mathbf{T} \in \mathbb{C}^{m\times n}. \end{aligned} \tag{3.3}$$

We now show that (3.3) has a solution with at most one nonzero element in each row and column.

Theorem 3.1. *There exists a solution of* (3.3) *of the form* $\mathbf{T} = \mathbf{\Pi}_1\mathbf{\Sigma}\mathbf{\Pi}_2$ *where* $\mathbf{\Pi}_1$ *and* $\mathbf{\Pi}_2$ *are permutation matrices and* $\sigma_{ij} = 0$ *for all* $i \neq j$.

Proof. We establish the theorem under the following nondegeneracy assumption:

$$\delta_i \neq \delta_j > 0 \quad \text{and} \quad \lambda_i \neq \lambda_j > 0 \ \text{ for all } i \neq j. \tag{3.4}$$

Later we show that since any $\boldsymbol{\delta} \geqslant \mathbf{0}$ and $\boldsymbol{\lambda} \geqslant \mathbf{0}$ can be approximated arbitrarily closely by vectors $\boldsymbol{\delta}$ and $\boldsymbol{\lambda}$ satisfying (3.4), the theorem holds for arbitrary $\boldsymbol{\lambda} \geqslant \mathbf{0}$ and $\boldsymbol{\delta} \geqslant \mathbf{0}$.

There exists an optimal solution of (3.3) since the feasible set is compact and the cost function is a continuous function of $\mathbf{T}$. Since the eigenvalues of $\mathbf{\Delta}\mathbf{T}^*\mathbf{\Lambda}\mathbf{T}\mathbf{\Delta}$ are nonnegative, it follows that for any choice of $\mathbf{T}$,

$$\operatorname{tr}(\mathbf{\Delta}\mathbf{T}^*\mathbf{\Lambda}\mathbf{T}\mathbf{\Delta} + \mathbf{I})^{-1} \leqslant n,$$

with equality if and only if $\mathbf{T} = \mathbf{0}$. Hence, there exists a nonzero optimal solution of (3.3), which is denoted $\bar{\mathbf{T}}$. Since the gradient of the constraint in (3.3) does not vanish at $\bar{\mathbf{T}}$, the first-order necessary optimality conditions imply that there exists a scalar $\gamma \geqslant 0$ such that the derivative of the Lagrangian vanishes:

$$\frac{\mathrm{d}}{\mathrm{d}\mathbf{T}}\operatorname{tr}\big((\mathbf{\Delta}\mathbf{T}^*\mathbf{\Lambda}\mathbf{T}\mathbf{\Delta} + \mathbf{I})^{-1} + \gamma\mathbf{T}^*\mathbf{T}\big)_{\mathbf{T}=\bar{\mathbf{T}}} = \mathbf{0}. \tag{3.5}$$

For any invertible matrix $\mathbf{M}$, we have

$$\frac{\mathrm{d}\mathbf{M}^{-1}}{\mathrm{d}\mathbf{T}} = -\mathbf{M}^{-1}\left(\frac{\mathrm{d}\mathbf{M}}{\mathrm{d}\mathbf{T}}\right)\mathbf{M}^{-1}.$$

Equating to zero the derivative of the Lagrangian at $\bar{\mathbf{T}}$ in the direction $\boldsymbol{\delta}\mathbf{T} \in \mathbb{C}^{m\times n}$, we obtain:

$$\operatorname{tr}\big(\gamma[\bar{\mathbf{T}}^*\boldsymbol{\delta}\mathbf{T} + \boldsymbol{\delta}\mathbf{T}^*\bar{\mathbf{T}}] - \mathbf{M}^{-1}\mathbf{\Delta}[\bar{\mathbf{T}}^*\mathbf{\Lambda}\boldsymbol{\delta}\mathbf{T} + \boldsymbol{\delta}\mathbf{T}^*\mathbf{\Lambda}\bar{\mathbf{T}}]\mathbf{\Delta}\mathbf{M}^{-1}\big) = 0, \tag{3.6}$$

where

$$\mathbf{M} = \mathbf{\Delta}\bar{\mathbf{T}}^*\mathbf{\Lambda}\bar{\mathbf{T}}\mathbf{\Delta} + \mathbf{I}. \tag{3.7}$$

Let $\mathrm{Real}(z)$ denote the real part of $z \in \mathbb{C}$. Since $\operatorname{tr}(\mathbf{A} + \mathbf{A}^*) = 2(\mathrm{Real}[\operatorname{tr}(\mathbf{A})])$ and $\operatorname{tr}(\mathbf{AB}) = \operatorname{tr}(\mathbf{BA})$, it follows that

$$\mathrm{Real}\big[\operatorname{tr}(\gamma\bar{\mathbf{T}}^*\boldsymbol{\delta}\mathbf{T} - \mathbf{\Delta}\mathbf{M}^{-2}\mathbf{\Delta}\bar{\mathbf{T}}^*\mathbf{\Lambda}\boldsymbol{\delta}\mathbf{T})\big] = 0.$$

By taking $\boldsymbol{\delta}\mathbf{T}$ either pure real or pure imaginary, we conclude that

$$\operatorname{tr}\big([\gamma\bar{\mathbf{T}}^* - \mathbf{\Delta}\mathbf{M}^{-2}\mathbf{\Delta}\bar{\mathbf{T}}^*\mathbf{\Lambda}]\boldsymbol{\delta}\mathbf{T}\big) = 0$$

for all $\boldsymbol{\delta}\mathbf{T}$. Choosing $\boldsymbol{\delta}\mathbf{T} = \big([\gamma\bar{\mathbf{T}}^* - \mathbf{\Delta}\mathbf{M}^{-2}\mathbf{\Delta}\bar{\mathbf{T}}^*\mathbf{\Lambda}]\boldsymbol{\delta}\mathbf{T}\big)^*$, we deduce that

$$\gamma\bar{\mathbf{T}}^* - \mathbf{\Delta}\mathbf{M}^{-2}\mathbf{\Delta}\bar{\mathbf{T}}^*\mathbf{\Lambda} = \mathbf{0}. \tag{3.8}$$

If $\gamma = 0$, then $\bar{\mathbf{T}} = \mathbf{0}$ since both $\mathbf{\Delta}$ and $\mathbf{\Lambda}$ are invertible. Consequently, $\gamma > 0$.

By the matrix modification formula [5], we have

$$\begin{aligned}\mathbf{M}^{-1} &= (\mathbf{I} + \mathbf{\Delta}\bar{\mathbf{T}}^*\mathbf{\Lambda}\bar{\mathbf{T}}\mathbf{\Delta})^{-1} = \big(\mathbf{I} + [\mathbf{\Delta}\bar{\mathbf{T}}^*\mathbf{\Lambda}^{1/2}][\mathbf{\Lambda}^{1/2}\bar{\mathbf{T}}\mathbf{\Delta}]\big)^{-1}\\ &= \mathbf{I} - \mathbf{\Delta}\bar{\mathbf{T}}^*\mathbf{\Lambda}^{1/2}(\mathbf{I} + \mathbf{\Lambda}^{1/2}\bar{\mathbf{T}}\mathbf{\Delta}^2\bar{\mathbf{T}}^*\mathbf{\Lambda}^{1/2})^{-1}\mathbf{\Lambda}^{1/2}\bar{\mathbf{T}}\mathbf{\Delta}.\end{aligned} \tag{3.9}$$

Hence,

$$\mathbf{M}^{-1}\boldsymbol{\Delta}\bar{\mathbf{T}}^*\boldsymbol{\Lambda}^{1/2} = \boldsymbol{\Delta}\bar{\mathbf{T}}^*\boldsymbol{\Lambda}^{1/2}(\mathbf{I} + \boldsymbol{\Lambda}^{1/2}\bar{\mathbf{T}}\boldsymbol{\Delta}^2\bar{\mathbf{T}}^*\boldsymbol{\Lambda}^{1/2})^{-1}.$$

It follows that

$$\mathbf{M}^{-2}\boldsymbol{\Delta}\bar{\mathbf{T}}^*\boldsymbol{\Lambda}^{1/2} = \boldsymbol{\Delta}\bar{\mathbf{T}}^*\boldsymbol{\Lambda}^{1/2}(\mathbf{I} + \boldsymbol{\Lambda}^{1/2}\bar{\mathbf{T}}\boldsymbol{\Delta}^2\bar{\mathbf{T}}^*\boldsymbol{\Lambda}^{1/2})^{-2}.$$

Making this substitution in (3.8) gives

$$\gamma\bar{\mathbf{T}}^* - \boldsymbol{\Delta}^2\bar{\mathbf{T}}^*\boldsymbol{\Lambda}^{1/2}(\mathbf{I} + \boldsymbol{\Lambda}^{1/2}\bar{\mathbf{T}}\boldsymbol{\Delta}^2\bar{\mathbf{T}}^*\boldsymbol{\Lambda}^{1/2})^{-2}\boldsymbol{\Lambda}^{1/2} = \mathbf{0}.$$

Multiplying on the right by $\bar{\mathbf{T}}$ yields

$$\bar{\mathbf{T}}^*\bar{\mathbf{T}} = \frac{1}{\gamma}\boldsymbol{\Delta}^2\bar{\mathbf{T}}^*\boldsymbol{\Lambda}^{1/2}(\mathbf{I} + \boldsymbol{\Lambda}^{1/2}\bar{\mathbf{T}}\boldsymbol{\Delta}^2\bar{\mathbf{T}}^*\boldsymbol{\Lambda}^{1/2})^{-2}\boldsymbol{\Lambda}^{1/2}\bar{\mathbf{T}}.$$

This equation has the form

$$\bar{\mathbf{T}}^*\bar{\mathbf{T}} = \boldsymbol{\Delta}^2\mathbf{A}, \quad \mathbf{A} = \frac{1}{\gamma}\bar{\mathbf{T}}^*\boldsymbol{\Lambda}^{1/2}(\mathbf{I} + \boldsymbol{\Lambda}^{1/2}\bar{\mathbf{T}}\boldsymbol{\Delta}^2\bar{\mathbf{T}}^*\boldsymbol{\Lambda}^{1/2})^{-2}\boldsymbol{\Lambda}^{1/2}\bar{\mathbf{T}}.$$

Since $\mathbf{A}$ is Hermitian and no two diagonal elements of $\boldsymbol{\Delta}$ are equal (see the nondegeneracy condition (3.4)), we conclude that both $\bar{\mathbf{T}}^*\bar{\mathbf{T}}$ and $\mathbf{A}$ are diagonal.

Let us define

$$\mathbf{R} = \boldsymbol{\Delta}\bar{\mathbf{T}}^*\boldsymbol{\Lambda}\bar{\mathbf{T}}\boldsymbol{\Delta}. \tag{3.10}$$

By (3.7) and (3.10), $\mathbf{M} = \mathbf{R} + \mathbf{I}$. Hence, (3.8) can be written:

$$\gamma\boldsymbol{\Delta}^{-1}\bar{\mathbf{T}}^* = (\mathbf{R} + \mathbf{I})^{-2}\boldsymbol{\Delta}\bar{\mathbf{T}}^*\boldsymbol{\Lambda}.$$

Multiply on the right by $\bar{\mathbf{T}}\boldsymbol{\Delta}$ to obtain:

$$\gamma\boldsymbol{\Delta}^{-1}\bar{\mathbf{T}}^*\bar{\mathbf{T}}\boldsymbol{\Delta} = (\mathbf{R} + \mathbf{I})^{-2}\mathbf{R} = (\mathbf{I} + \mathbf{R})^{-1} - (\mathbf{I} + \mathbf{R})^{-2}. \tag{3.11}$$

The eigenvectors of $(\mathbf{I} + \mathbf{R})^{-1} - (\mathbf{I} + \mathbf{R})^{-2}$ coincide with the eigenvectors of $\mathbf{R}$. Since $\bar{\mathbf{T}}^*\bar{\mathbf{T}}$ and $\boldsymbol{\Delta}$ are both diagonal, the eigenvectors of $\gamma\boldsymbol{\Delta}^{-1}\bar{\mathbf{T}}^*\bar{\mathbf{T}}\boldsymbol{\Delta}$ coincide with the columns of the identity matrix. Hence, $\mathbf{R}$ is diagonalized by the identity matrix, which implies that $\mathbf{R}$ is diagonal.

Since $\mathbf{R}$ is diagonal, both $\mathbf{M} = \mathbf{R} + \mathbf{I}$ and $\mathbf{M}^{-1}$ are diagonal, and the factor $\boldsymbol{\Delta}\mathbf{M}^{-2}\boldsymbol{\Delta}$ in (3.8) is diagonal with real diagonal elements denoted by e_j, $1 \leqslant j \leqslant n$. By (3.8), we have

$$\gamma\bar{t}_{ij} = \lambda_i e_j \bar{t}_{ij}. \tag{3.12}$$

If $\bar{t}_{ij} \neq 0$, then (3.12) implies that

$$\lambda_i e_j = \gamma \neq 0.$$

By the nondegeneracy condition (3.4), no two diagonal elements of $\boldsymbol{\Lambda}$ are equal. If for any fixed j, $\bar{t}_{ij} \neq 0$ for $i = i_1$ and i_2, then the identity $\lambda_i e_j = \gamma$ yields a contradiction since $\gamma \neq 0$ and $\lambda_{i_1} \neq \lambda_{i_2}$. Hence, each column of $\bar{\mathbf{T}}$ has at most one nonzero. Since $\bar{\mathbf{T}}^*\bar{\mathbf{T}}$ is diagonal, two different columns cannot have their single nonzero in the same row. This shows that each column and each row of $\mathbf{T}$ have at most one nonzero. A suitable permutation of the rows and columns of $\bar{\mathbf{T}}$ yields a diagonal matrix $\boldsymbol{\Sigma}$.

So far, we have established Theorem 3.1 under the assumption that $\boldsymbol{\lambda}$ and $\boldsymbol{\delta}$ satisfy the nondegeneracy condition (3.4). Now, consider arbitrary $\hat{\boldsymbol{\delta}} \in \mathbb{R}^n$ and $\hat{\boldsymbol{\lambda}} \in \mathbb{R}^m$ with $\hat{\boldsymbol{\lambda}} \geqslant \mathbf{0}$ and $\hat{\boldsymbol{\delta}} \geqslant \mathbf{0}$. Let $\boldsymbol{\delta}_k$ and $\boldsymbol{\lambda}_k$ for $k \geqslant 0$ denote sequences which satisfy the nondegeneracy condition (3.4), and which converge to $\hat{\boldsymbol{\lambda}}$ and $\hat{\boldsymbol{\delta}}$, respectively. Let $\boldsymbol{\Delta}_k$ and $\boldsymbol{\Lambda}_k$ be diagonal matrices with $\boldsymbol{\delta}_k$ and $\boldsymbol{\lambda}_k$ on their respective diagonals. By our previous analysis, problem (3.3) with $\mathbf{\Delta} = \mathbf{\Delta}_k$ and $\mathbf{\Lambda} = \mathbf{\Lambda}_k$ has a solution $\mathbf{T}_k$ of the form $\mathbf{T}_k = \mathbf{\Pi}_1^k \mathbf{\Sigma}_k \mathbf{\Pi}_2^k$, where $\mathbf{\Pi}_1^k$ and $\mathbf{\Pi}_2^k$ are permutation matrices and $\mathbf{\Sigma}_k$ is diagonal. Since the set of permutation matrices for any fixed dimension is finite and since the $\mathbf{\Sigma}_k$ sequence lies in a compact set corresponding to the trace constraint in (3.3), there exists a diagonal matrix $\widehat{\mathbf{\Sigma}}$, fixed permutations $\mathbf{\Pi}_1$ and $\mathbf{\Pi}_2$, and a subsequence, also indexed by k for convenience, with the property that $\mathbf{\Sigma}_k$ converges to $\widehat{\mathbf{\Sigma}}$ and $(\mathbf{\Pi}_1^k, \mathbf{\Pi}_2^k) = (\mathbf{\Pi}_1, \mathbf{\Pi}_2)$, independent of k.

By the definition of $\mathbf{T}_k$, we have

$$\operatorname{tr}(\mathbf{\Delta}_k \mathbf{T}^* \mathbf{\Lambda}_k \mathbf{T} \mathbf{\Delta}_k + \mathbf{I})^{-1} \geqslant \operatorname{tr}(\mathbf{\Delta}_k \mathbf{T}_k^* \mathbf{\Lambda}_k \mathbf{T}_k \mathbf{\Delta}_k + \mathbf{I})^{-1}, \tag{3.13}$$

whenever $\mathbf{T}$ satisfies the constraint $\operatorname{tr}(\mathbf{T}^*\mathbf{T}) \leqslant P$. We let k tend to infinity in (3.13). By continuity, it follows that

$$\operatorname{tr}(\widehat{\mathbf{\Delta}} \mathbf{T}^* \widehat{\mathbf{\Lambda}} \mathbf{T} \widehat{\mathbf{\Delta}} + \mathbf{I})^{-1} \geqslant \operatorname{tr}(\widehat{\mathbf{\Delta}}\, \widehat{\mathbf{T}}^* \widehat{\mathbf{\Lambda}}\, \widehat{\mathbf{T}} \mathbf{\Delta} + \mathbf{I})^{-1}.$$

Hence, $\widehat{\mathbf{T}} = \mathbf{\Pi}_1 \widehat{\mathbf{\Sigma}} \mathbf{\Pi}_2$ is a solution of (3.3) associated with $\boldsymbol{\delta} = \widehat{\boldsymbol{\delta}}$ and $\boldsymbol{\lambda} = \widehat{\boldsymbol{\lambda}}$. □

Combining the relationship (3.2) between $\mathbf{T}$ and $\mathbf{S}$ and Theorem 3.1, we conclude that problem (1.1) has a solution of the form $\mathbf{S} = \mathbf{U}\mathbf{\Pi}_1\mathbf{\Sigma}\mathbf{\Pi}_2\mathbf{V}^*$, where $\mathbf{\Pi}_1$ and $\mathbf{\Pi}_2$ are permutation matrices. We will now show that one of these two permutation matrices can be deleted if the eigenvalues of $\mathbf{D}$ and $\mathbf{Q}$ are arranged in decreasing order.

Let N denote the minimum of m and n. Making the substitution $\mathbf{S} = \mathbf{U}\mathbf{\Pi}_1\mathbf{\Sigma}\mathbf{\Pi}_2\mathbf{V}^*$ in (1.1), we obtain the equivalent problem:

$$\begin{aligned} &\min_{\boldsymbol{\sigma}, \mathbf{\Pi}_1, \mathbf{\Pi}_2} \operatorname{tr}\left((\mathbf{\Pi}_2\mathbf{\Delta}\mathbf{\Pi}_2^*)\mathbf{\Sigma}^*(\mathbf{\Pi}_1^*\mathbf{\Lambda}\mathbf{\Pi}_1)\mathbf{\Sigma}(\mathbf{\Pi}_2\mathbf{\Delta}\mathbf{\Pi}_2^*) + \mathbf{I}\right)^{-1} \\ &\text{subject to} \quad \sum_{i=1}^{N} \sigma_i^2 \leqslant P. \end{aligned} \tag{3.14}$$

Here the minimization is over diagonal matrices $\mathbf{\Sigma}$ with $\boldsymbol{\sigma}$ on the diagonal, and permutation matrices $\mathbf{\Pi}_1$ and $\mathbf{\Pi}_2$.

The symmetric permutations $\mathbf{\Pi}_1^*\mathbf{\Lambda}\mathbf{\Pi}_1$ and $\mathbf{\Pi}_2\mathbf{\Delta}\mathbf{\Pi}_2^*$ essentially interchange diagonal elements of $\mathbf{\Lambda}$ and $\mathbf{\Delta}$. Hence, (3.14) is equivalent to

$$\begin{aligned} &\min_{\boldsymbol{\sigma}, \pi_1, \pi_2} \sum_{i=1}^{N} \frac{1}{(\delta_{\pi_2(i)}\sigma_i)^2 \lambda_{\pi_1(i)} + 1} \\ &\text{subject to} \quad \sum_{i=1}^{N} \sigma_i^2 \leqslant P, \quad \pi_1 \in \mathscr{P}_m, \quad \pi_2 \in \mathscr{P}_n, \end{aligned} \tag{3.15}$$

where $\mathscr{P}_m$ is the set of bijections of $\{1, 2, \ldots, m\}$ onto itself.

We first show that we can restrict our attention to the largest diagonal elements of $\mathbf{D}$ and $\mathbf{Q}$.

Lemma 3.2. *Let* $\mathbf{U}\mathbf{\Lambda}\mathbf{U}^*$ *and* $\mathbf{V}\mathbf{\Delta}\mathbf{V}^*$ *be diagonalizations of* $\mathbf{Q}$ *and* $\mathbf{D}$, *respectively*, *where the columns of* $\mathbf{U}$ *and* $\mathbf{V}$ *are orthonormal eigenvectors. Let* $\boldsymbol{\sigma}$, π_1, *and* π_2 *denote an optimal solution of* (3.15) *and define the sets*

$$\mathscr{N} = \{i : \sigma_i > 0\}, \quad \mathscr{Q} = \{\lambda_{\pi_1(i)} : i \in \mathscr{N}\}, \quad \textit{and} \quad \mathscr{D} = \{\delta_{\pi_2(i)} : i \in \mathscr{N}\},$$

If $\mathscr{N}$ *has* l *elements*, *then the elements of the set* $\mathscr{D}$ *and* $\mathscr{Q}$ *are all nonzero*, *and they constitute the* l *largest eigenvalues of* $\mathbf{D}$ *and* $\mathbf{Q}$, *respectively.*

Proof. Define the set

$$\mathscr{P} = \{i : \lambda_{\pi_1(i)}\sigma_i\delta_{\pi_2(i)} > 0\}.$$

If $\delta_{\pi_2(i)} = 0$ for some $i \in \mathscr{N}$, then the cost function in (3.15) is decreased by setting $\sigma_i = 0$ and by increasing σ_j for some $j \in \mathscr{P}$. This violates the optimality of $\boldsymbol{\sigma}$, π_1 and π_2; consequently, $\delta_{\pi_2(i)} > 0$. By the same reasoning, $\lambda_{\pi(i)} > 0$ for all $i \in \mathscr{N}$. Suppose $k \notin \mathscr{N}$ and $\delta_{\pi_2(k)} > \delta_{\pi_2(i)}$ for some $i \in \mathscr{N}$. Interchanging the values of $\pi_2(i)$ and $\pi_2(k)$, the new ith term is smaller than the previous ith term, which again violates the optimality of $\boldsymbol{\sigma}$ and $\boldsymbol{\pi}$. Hence, $\delta_{\pi_2(k)} \leqslant \delta_{\pi_2(i)}$. Finally, suppose that $k \notin \mathscr{N}$ and $\lambda_{\pi_1(k)} > \lambda_{\pi_1(i)}$. Interchanging the values of $\pi_1(i)$ and $\pi_1(k)$, the new ith term is smaller than the previous ith term. Hence, $\lambda_{\pi_1(k)} \leqslant \lambda_{\pi_1(i)}$. $\square$

Using Lemma 3.2, we now eliminate one of the permutations in (3.15).

Theorem 3.3. *Let* $\mathbf{U\Lambda U}^*$ *and* $\mathbf{V\Delta V}^*$ *be diagonalizations of* $\mathbf{Q}$ *and* $\mathbf{D}$, *respectively, where the columns of* $\mathbf{U}$ *and* $\mathbf{V}$ *are orthonormal eigenvectors, and the eigenvalues of* $\mathbf{Q}$ *and* $\mathbf{D}$ *are arranged in decreasing order as in* (3.1). *If* K *is the minimum of the rank of* $\mathbf{D}$ *and* $\mathbf{Q}$, *then* (3.15) *is equivalent to*

$$\min_{\boldsymbol{\sigma},\pi} \sum_{i=1}^{K} \frac{1}{(\delta_i\sigma_i)^2\lambda_{\pi(i)} + 1} \tag{3.16}$$

$$\text{subject to} \quad \sum_{i=1}^{K} \sigma_i^2 \leqslant P, \quad \pi \in \mathscr{P}_K,$$

where $\sigma_i = 0$ *for* $i > K$.

Proof. Since at most K eigenvalues of either $\mathbf{D}$ or $\mathbf{Q}$ are nonzero, it follows from Lemma 3.2 that the set $\mathscr{N}$ has at most K elements. Since the associated eigenvalue sets $\mathscr{Q}$ and $\mathscr{D}$ are the largest eigenvalues of $\mathbf{Q}$ and $\mathbf{D}$, respectively, we can assume, without loss of generality, that $\pi_1(i) \in [1, K]$ and $\pi_2(i) \in [1, K]$ for each $i \in \mathscr{N}$. Hence, we restrict the sum in (3.15) to those indices $i \in \mathscr{S}$ where

$$\mathscr{S} = \{\pi_2^{-1}(j) : 1 \leqslant j \leqslant K\}.$$

Let us define $\sigma'_j = \sigma_{\pi_2^{-1}(j)}$ and $\pi(j) = \pi_1(\pi_2^{-1}(j))$. Since $\pi(j) \in [1, K]$ for $j \in [1, K]$, it follows that $\pi \in \mathscr{P}_K$. In (3.15) we restrict the summation to $i \in \mathscr{S}$ and we replace i by $\pi_2^{-1}(j)$ to obtain

$$\sum_{i\in\mathscr{S}} \frac{1}{(\delta_{\pi_2(i)}\sigma_i)^2\lambda_{\pi_1(i)} + 1} = \sum_{j=1}^{K} \frac{1}{(\delta_j\sigma'_j)^2\lambda_{\pi(j)} + 1}, \quad \text{where } \sum_{i=1}^{K}(\sigma'_j)^2 \leqslant P.$$

This completes the proof of (3.16). $\square$

Corollary 3.4. *Problem* (1.1) *has a solution of the form* $\mathbf{S} = \mathbf{U\Pi\Sigma V}^*$ *where the columns of* $\mathbf{U}$ *and* $\mathbf{V}$ *are orthonormal eigenvectors of* $\mathbf{Q}$ *and* $\mathbf{D}$, *respectively, with the associated eigenvalues arranged in decreasing order*, $\mathbf{\Pi}$ *is a permutation matrix*, *and* $\mathbf{\Sigma}$ *is diagonal.*

Proof. Let $\boldsymbol{\sigma}$ and π be a solution of (3.16). For $i > K$, define $\pi(i) = i$ and $\sigma_i = 0$. If $\mathbf{\Pi}$ is the permutation matrix corresponding to π, then substituting $\mathbf{S} = \mathbf{U\Pi\Sigma V}^*$ in the cost function of

(1.1), we obtain the cost function in (3.16). Since (3.15) and (3.16) are equivalent by Theorem 3.3, $\mathbf{S}$ is optimal in (1.1). $\square$

4. The optimal $\boldsymbol{\Sigma}$

Assuming the permutation π in (3.16) is given, let us now consider the problem of optimizing over $\boldsymbol{\sigma}$. To simplify the indexing, let ρ_i denote $\lambda_{\pi(i)}$. Hence, for fixed π, (3.16) is equivalent to the following optimization problem:

$$\min_{\boldsymbol{\sigma}} \sum_{i=1}^{K} \frac{1}{(\delta_i \sigma_i)^2 \rho_i + 1} \quad \text{subject to} \quad \sum_{i=1}^{K} \sigma_i^2 \leqslant P. \tag{4.1}$$

The solution of (4.1) can be expressed in terms of a Lagrange multiplier associated with the constraint (this solution technique is often called "water filling" [3] in the communication literature).

Theorem 4.1. *The optimal solution of* (4.1) *is given by*

$$\sigma_i = \max\left\{\sqrt{\frac{1}{\delta_i^2 \rho_i \mu}} - \frac{1}{\delta_i^2 \rho_i}, 0\right\}^{1/2}, \tag{4.2}$$

where the parameter μ *is chosen so that*

$$\sum_{i=1}^{K} \sigma_i^2 = P. \tag{4.3}$$

Proof. Since the minimization in (4.1) takes place over a closed, bounded set, there exists a solution. The inequality constraint is active at a solution (otherwise, $\mathbf{S}$ can be multiplied by a scalar larger than 1 to reduce to the value of the cost function). Defining $s_i = \sigma_i^2$ and $p_i = \delta_i^2 \lambda_i$, the reduced problem (4.1) is equivalent to

$$\min_{\mathbf{s}} \sum_{i=1}^{K} \frac{1}{p_i s_i + 1} \quad \text{subject to} \quad \sum_{i=1}^{K} s_i = P, \quad \mathbf{s} \geqslant \mathbf{0}. \tag{4.4}$$

Due to the strict convexity of the cost function and the convexity of the constraints, (4.4) has a unique solution.

The first-order optimality conditions (KKT conditions) for an optimal solution of (4.4) are the following: There exists a scalar $\mu \geqslant 0$ and a vector $\boldsymbol{\nu} \in \mathbb{R}^K$ such that

$$-\frac{p_i}{(p_i s_i + 1)^2} + \mu - \nu_i = 0, \quad \nu_i \geqslant 0, \ \ s_i \geqslant 0, \ \ \nu_i s_i = 0, \ \ 1 \leqslant i \leqslant K. \tag{4.5}$$

Due to the convexity of the cost and the constraints, any solution of these conditions is the unique optimal solution of (4.4).

A solution to (4.5) is obtained as follows: Define the function

$$s_i(\mu) = \left(\sqrt{\frac{1}{p_i\mu}} - p_i^{-1}\right)^+. \tag{4.6}$$

Here $x^+ = \max\{x, 0\}$. This particular value for s_i is obtained by setting $\nu_i = 0$ in (4.5) and solving for s_i; when the solution is <0, we set $s_i(\mu) = 0$ (this corresponds to the + operator (4.6)). Observe that $s_i(\mu)$ is a decreasing function of μ which approaches $+\infty$ as μ approaches 0 and which approaches 0 as μ tends to $+\infty$. Hence, the equation

$$\sum_{i=1}^{K} s_i(\mu) = P \tag{4.7}$$

has a unique positive solution. Since $s_i(p_i) = 0$, we have $s_i(\mu) = 0$ for $\mu \geqslant p_i$, which implies that

$$-\frac{p_i}{(p_i s_i(\mu) + 1)^2} + \mu = -p_i + \mu > 0 \quad \text{for } \mu > p_i.$$

It follows that the KKT conditions are satisfied when μ is the positive solution of (4.7). $\square$

5. Optimal eigenvector ordering for large or small power

To solve (1.1), we need to find an optimal ordering for the eigenvalues of **D** and **Q**. In Theorems 5.1 and 5.2, we determine the optimal ordering when the power P is either large or small.

Theorem 5.1. *If the eigenvalues $\{\lambda_i\}$ and $\{\delta_i\}$ of* **Q** *and* **D**, *respectively, are arranged in decreasing order, then for P sufficiently large, an optimal permutation in* (3.16) *is*

$$\pi(i) = K + 1 - i, \quad 1 \leqslant i \leqslant K, \qquad \pi(i) = i, \quad i > K. \tag{5.1}$$

Proof. By Theorem 4.1, as P tends to infinity, μ tends to zero and the optimal σ_i tend to infinity. Choose P large enough so that the following two conditions hold for an optimal $\boldsymbol{\sigma}$ and π in (3.16):

(a) $\sqrt{\dfrac{1}{\delta_i^2\rho_i\mu}} > (\delta_i^2\rho_i)^{-1}$,

(b) $\sigma_i^2 + \sigma_j^2 > \dfrac{2}{\delta_i\delta_j\sqrt{\rho_i\rho_j}}, \quad$ when $1 \leqslant i < j \leqslant K$.

Here $\rho_i = \lambda_{\pi(i)}$ is the reordered eigenvalue of **Q**. We will show that $\rho_1 \leqslant \rho_2 \leqslant \cdots \leqslant \rho_K$. By (a), the max in (4.2) is attained by the first term, which corresponds to the solution of (4.5) associated with $\nu_i = 0$.

Suppose that there exist indices i and j such that $i < j$ and $\rho_i > \rho_j$. Since $\boldsymbol{\sigma}$ yields an optimal solution of (3.16), it follows that a solution of the following problem is $t_1 = \sigma_i^2$ and $t_2 = \sigma_j^2$:

$$\min_{t_1,t_2} \frac{1}{p_1t_1 + 1} + \frac{1}{p_2t_2 + 1} \tag{5.2}$$
$$\text{subject to} \quad t_1 + t_2 = \bar{P}, \quad t_1 \geqslant 0, t_2 \geqslant 0,$$

where $\bar{P} = \sigma_i^2 + \sigma_j^2$, $p_1 = \delta_i^2 \rho_i$, and $p_2 = \delta_j^2 \rho_j$. By Theorem 4.1, the t_i can be expressed:

$$t_i(\mu) = \sqrt{\frac{1}{p_i \mu}} - p_i^{-1}, \tag{5.3}$$

where μ is obtained from the condition $t_1 + t_2 = \bar{P}$:

$$\sqrt{\mu} = \frac{r_1 + r_2}{\bar{P} + r_1^2 + r_2^2}, \quad r_i = 1/\sqrt{p_i}. \tag{5.4}$$

Let C denote the cost function for (5.2). Combining (5.3) and (5.4) gives

$$C = \frac{1}{p_1 t_1 + 1} + \frac{1}{p_2 t_2 + 1} = \frac{(r_1 + r_2)^2}{\bar{P} + r_1^2 + r_2^2}.$$

Define the following quantities:

$$a_1 = 1/\sqrt{\rho_i}, \quad a_2 = 1/\sqrt{\rho_j}, \quad b_1 = 1/\delta_i, \quad \text{and} \quad b_2 = 1/\delta_j.$$

With these definitions, $r_i = a_i b_i$ for $i = 1, 2$, and C becomes

$$C = \frac{(a_1 b_1 + a_2 b_2)^2}{\bar{P} + (a_1 b_1)^2 + (a_2 b_2)^2}. \tag{5.5}$$

Since the eigenvalues of $\mathbf{D}$ are arranged in decreasing order, $i < j$, and $\rho_i > \rho_j$, we have

$$b_1 \leqslant b_2 \quad \text{and} \quad a_1 < a_2. \tag{5.6}$$

Now, suppose that we interchange the values of $\pi(i)$ and $\pi(j)$. This interchange has the effect of interchanging the values ρ_i and ρ_j, or equivalently, interchanging a_1 and a_2. Let C^+ denote the cost value associated with the interchange:

$$C^+ = \frac{(a_2 b_1 + a_1 b_2)^2}{\bar{P} + (a_2 b_1)^2 + (a_1 b_2)^2}.$$

After cross-multiplying the inequality $C^+ \leqslant C$, we find (after considerable algebra) that $C^+ \leqslant C$ if and only if

$$(a_1 - a_2)(a_1 + a_2)(b_1 - b_2)(b_1 + b_2)(2r_1 r_2 - \bar{P}) \leqslant 0. \tag{5.7}$$

By (b), $2r_1 r_2 - \bar{P} < 0$. Since $a_1 < a_2$ and $b_1 \leqslant b_2$, the expression (5.7) is $\leqslant 0$. If (5.7) is <0, then we contradict the optimality of $\boldsymbol{\sigma}$ and π. Hence, the expression (5.7) is zero, and by interchanging ρ_i and ρ_j, the cost function in (3.16) does not change. In summary, for each i and j with $i < j$ and $\rho_i > \rho_j$, we can interchange the values of $\pi(i)$ and $\pi(j)$ to obtain a new permutation with the same value for the cost function. After the interchange, we have $\rho_i < \rho_j$. In this way, the ρ_i are arranged in increasing order. Since the λ_i are arranged in decreasing order, we conclude that the associated optimal permutation π is (5.1).

One technical point must now be checked: We should verify that if $\rho_i > \rho_j$ with $i < j$, and if we exchange ρ_i and ρ_j, then the corresponding optimal solution of (5.2) remains positive (the formula (5.5) is based on the assumption that the optimal solution is positive). Since the original solution, before the exchange, is positive, it follows from (5.3) and (5.4) that

$$\bar{P} + r_1^2 > r_1 r_2 \quad \text{and} \quad \bar{P} + r_2^2 > r_1 r_2. \tag{5.8}$$

After the exchange, the analogous inequalities that must be satisfied to preserve nonnegativity are

$$\bar{P} + a_2^2 b_1^2 \geqslant r_1 r_2 \quad \text{and} \quad \bar{P} + a_1^2 b_2^2 \geqslant r_1 r_2.$$

These are equivalent to

$$\bar{P} + r_1^2(a_2/a_1)^2 \geqslant r_1 r_2 \quad \text{and} \quad \bar{P} + (b_2/b_1)^2 r_1^2 \geqslant r_1 r_2.$$

These follow from the first inequality in (5.8) and the fact that $a_1 < a_2$ and $b_1 \leqslant b_2$. $\square$

Theorem 5.2. *Suppose the eigenvalues* $\{\lambda_i\}$ *and* $\{\delta_i\}$ *of* $\mathbf{Q}$ *and* $\mathbf{D}$, *respectively, are arranged in decreasing order, and let L be the minimum of the multiplicities of* δ_1 *and* λ_1. *For P sufficiently small, an optimal solution of* (1.1) *is*

$$\mathbf{S} = \sqrt{\frac{P}{L}} \sum_{i=1}^{L} \mathbf{u}_i \mathbf{v}_i^*, \tag{5.9}$$

where $\mathbf{u}_i$ *and* $\mathbf{v}_i$ *are the orthonormalized eigenvectors of* $\mathbf{Q}$ *and* $\mathbf{D}$ *associated with* λ_1 *and* δ_1, *respectively.*

Proof. By Lemma 3.2, (1.1) is equivalent to (3.16). Let π and $\boldsymbol{\sigma}$ denote an optimal solution to (3.16), and let $\mathscr{N}$ denote the indices of the nonzero σ_i defined in Lemma 3.2. Again, to simplify the indexing, we define $\rho_i = \lambda_{\pi(i)}$. Let ϵ be a (positive) separation parameter defined by

$$\epsilon = \min \left\{ \left| \frac{1}{\delta_i\sqrt{\lambda_j}} - \frac{1}{\delta_k\sqrt{\lambda_l}} \right| : \delta_i^2\lambda_j \neq \delta_k^2\lambda_l,\, i, k \in [1, n],\, j, l \in [1, m] \right\}.$$

Consider any value of P small enough that

$$\epsilon > P\delta_1\sqrt{\lambda_1}. \tag{5.10}$$

We will show that if P satisfies (5.10), then $\delta_i^2\rho_i = \delta_j^2\rho_j$ for all i and $j \in \mathscr{N}$.

Suppose that i and $j \in \mathscr{N}$ and $\delta_i^2\rho_i \neq \delta_j^2\rho_j$. As in the proof of Theorem 5.1, let us consider the restricted problem (5.2) whose solution is $t_1 = \sigma_i^2$ and $t_2 = \sigma_j^2$. Later, in the proof of Theorem 5.1, we point out that positivity of σ_i and σ_j is equivalent to the pair of inequalities (5.8). Combining these inequalities, we obtain:

$$\frac{\bar{P}}{r_2} + r_2 > r_1 > r_2 - \frac{\bar{P}}{r_1}.$$

It follows that

$$\left| \frac{1}{\delta_i\sqrt{\rho_i}} - \frac{1}{\delta_j\sqrt{\rho_j}} \right| = |r_1 - r_2| \leqslant \bar{P}\max\{r_1^{-1}, r_2^{-1}\} \leqslant P\delta_1\sqrt{\lambda_1} \tag{5.11}$$

since $P \geqslant \bar{P}$,

$$1/r_1 = \delta_i\sqrt{\rho_i} \leqslant \delta_1\sqrt{\lambda_1} \quad \text{and} \quad 1/r_2 = \delta_j\sqrt{\rho_j} \leqslant \delta_1\sqrt{\lambda_1}.$$

Since $\delta_i^2\rho_i \neq \delta_j^2\rho_j$, the definition of ϵ implies that

$$0 < \epsilon \leqslant \left| \frac{1}{\delta_i\sqrt{\rho_i}} - \frac{1}{\delta_j\sqrt{\rho_j}} \right|. \tag{5.12}$$

Eqs. (5.10)–(5.12) are inconsistent. Hence, $\delta_i^2\rho_i = \delta_j^2\rho_j$ for all $i, j \in \mathscr{N}$.

Let α denote the product $\delta_i^2\rho_i$ for any $i \in \mathscr{N}$. Suppose that i and $j \in \mathscr{N}$ and again, let us consider the restricted problem (5.2), whose solution is $t_1 = \sigma_i^2$ and $t_2 = \sigma_j^2$. Since $\alpha = p_1 = p_2$, it follows from (5.3) that $t_1 = t_2 = \sigma_i^2 = \sigma_j^2$. From the constraint of (5.2), we deduce that $t_1 = t_2 = \bar{P}/2$. Hence, the cost function in (5.2) has the value

$$C = \frac{2}{1 + .5\alpha \bar{P}}.$$

Now, suppose that $\delta_i > \delta_j$. Since the product $\delta_i^2 \rho_i$ is independent of $i \in \mathcal{N}$, it follows that $\rho_i < \rho_j$. Let us interchange the indices in $\pi(i)$ and $\pi(j)$, and focus on the associated 2-variable problems:

$$\min_{t_1, t_2} \frac{1}{\beta_1 t_1 + 1} + \frac{1}{\beta_2 t_2 + 1}$$
$$\text{subject to} \quad t_1 + t_2 = \bar{P}, \quad t_1 \geqslant 0, \quad t_2 \geqslant 0,$$

where $\beta_1 = \delta_i^2 \rho_j$ and $\beta_2 = \delta_j^2 \rho_i$. Since $\delta_i > \delta_j$ and $\rho_j > \rho_i$, it follows that $\beta_1 > \alpha > \beta_2$. The choice $t_1 = \bar{P}$ and $t_2 = 0$ is feasible, and the associated cost is

$$C^+ = 1 + \frac{1}{\beta_1 \bar{P} + 1}, \quad \text{where } \beta_1 > \alpha.$$

We show that $C^+ < C$. After cross-multiplication, this inequality is equivalent to

$$\begin{aligned} 2 &> 1 + .5\alpha \bar{P} + \frac{1 + .5\alpha \bar{P}}{1 + \beta_1 \bar{P}} \\ &= 2 + \frac{\bar{P}(\alpha - \beta_1(1 - .5\alpha \bar{P}))}{1 + \beta_1 \bar{P}}. \end{aligned} \tag{5.13}$$

Since $\beta_1 > \alpha$, it follows that for P sufficiently small, $\alpha - \beta_1(1 - .5\alpha \bar{P}) < 0$. In this case, the inequality (5.13) is satisfied, which is equivalent to $C^+ < C$. This violates the optimality of $\boldsymbol{\sigma}$ and π. Hence, $\delta_i = \delta_j$ for all i and $j \in \mathcal{N}$. In a similar fashion, $\rho_i = \rho_j$ for all i and $j \in \mathcal{N}$. In particular, it follows from Lemma 3.2 that $\delta_i = \delta_1$ and $\rho_i = \lambda_1$ for all $i \in \mathcal{N}$. With these substitutions, the problem (3.16) reduces to

$$\min_{\boldsymbol{\sigma}, \pi} \sum_{i=1}^{L} \frac{1}{\lambda_1 \delta_1^2 \sigma_i^2 + 1}$$
$$\text{subject to} \quad \sum_{i=1}^{L} \sigma_i^2 \leqslant P.$$

By Eqs. (4.2) and (4.3) in Theorem 4.1, $\sigma_i^2 = P/L$ for $1 \leqslant i \leqslant L$, which yields the solution (5.9). □

6. Solution of the determinant problem

The solution to the determinant problem (1.2) can be expressed as follows:

Theorem 6.1. *Let* $\mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^*$ *and* $\mathbf{V}\boldsymbol{\Delta}\mathbf{V}^*$ *be the diagonalizations of* $\mathbf{Q}$ *and* $\mathbf{D}$, *respectively, where the columns of* $\mathbf{U}$ *and* $\mathbf{V}$ *are orthonormal eigenvectors and the corresponding eigenvalues* $\{\lambda_i\}$ *and* $\{\delta_i\}$ *are arranged in decreasing order. If* K *is the minimum of the rank of* $\mathbf{Q}$ *and* $\mathbf{D}$, *then the optimal solution of* (1.2) *is given by*

$$\mathbf{S} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^*, \tag{6.1}$$

where $\mathbf{\Sigma}$ *is diagonal with diagonal elements given by*

$$\sigma_i = \max\left\{\frac{1}{\mu} - \frac{1}{\lambda_i\delta_i^2}, 0\right\}^{1/2} \quad \textit{for } 1 \leqslant i \leqslant K, \tag{6.2}$$

and $\sigma_i = 0$ *for* $i > K$, *where the parameter* μ *is chosen so that*

$$\sum_{i=1}^{K} \sigma_i^2 = P.$$

Proof. Initially, let us assume that both $\mathbf{D}$ and $\mathbf{Q}$ are nonsingular—later we remove this restriction. Insert $\mathbf{T} = \mathbf{Q}^{1/2}\mathbf{S}$ in (1.2) and multiply the objection function on the left and right by $\det(\mathbf{D}^{-1})$ to obtain the following equivalent formulation:

$$\begin{aligned} &\max_{\mathbf{T}} \det(\mathbf{T}^*\mathbf{T} + \mathbf{D}^{-2}) \\ &\text{subject to} \quad \operatorname{tr}(\mathbf{T}\mathbf{T}^*\mathbf{Q}^{-1}) \leqslant P, \quad \mathbf{T} \in \mathbb{C}^{m\times n}. \end{aligned} \tag{6.3}$$

Let ω_i, $1 \leqslant i \leqslant n$, denote the eigenvalues of $\mathbf{T}^*\mathbf{T}$ arranged in decreasing order. By a theorem of Fiedler [4] (also see [14, Chap. 9, G.4]), the determinant of a sum $\mathbf{T}^*\mathbf{T} + \mathbf{D}^{-2}$ of Hermitian matrices is bounded by the product of the sum of the respective eigenvalues (assuming the eigenvalues of $\mathbf{T}^*\mathbf{T}$ and $\mathbf{D}$ are in decreasing order):

$$\det(\mathbf{T}^*\mathbf{T} + \mathbf{D}^{-2}) \leqslant \prod_{i=1}^{n} (\omega_i + \delta_i^{-2}). \tag{6.4}$$

Also, by a theorem of Ruhe [16] (also see [14, Chap. 9, H2]), the trace of a product $(\mathbf{T}\mathbf{T}^*)\mathbf{Q}^{-1}$ of Hermitian matrices is bounded from below by the sum of the product of respective eigenvalues (assuming the eigenvalues of $\mathbf{T}\mathbf{T}^*$ and $\mathbf{Q}$ are in decreasing order):

$$\operatorname{tr}(\mathbf{T}\mathbf{T}^*\mathbf{Q}^{-1}) \geqslant \sum_{i=1}^{N} \omega_i \lambda_i^{-1}, \quad N = \min\{m, n\}, \tag{6.5}$$

since at most N eigenvalues of $\mathbf{T}^*\mathbf{T}$ and $\mathbf{T}\mathbf{T}^*$ are nonzero.

We replace the cost function in (1.2) by the upper bound (6.4) and we replace the constraint in (1.2) by the lower bound (6.5) to obtain the problem:

$$\begin{aligned} &\max_{\boldsymbol{\omega}} \left(\prod_{i=N+1}^{n} \delta_i^{-2}\right) \prod_{i=1}^{N} (\omega_i + \delta_i^{-2}) \\ &\text{subject to} \quad \sum_{i=1}^{N} \omega_i \lambda_i^{-1} \leqslant P, \quad \omega_i \geqslant \omega_{i+1} \geqslant 0 \ \text{ for } i < N. \end{aligned} \tag{6.6}$$

If $\mathbf{T}$ is feasible in (6.3), then the first N eigenvalues ω_i, $1 \leqslant i \leqslant N$, of $\mathbf{T}^*\mathbf{T}$ are feasible in (6.6) by (6.5). And by (6.4), the value of the cost function in (6.6) is greater than or equal to the associated value (6.3). Since the feasible set for (6.6) is closed and bounded, and since the cost function is continuous, there exists a maximizing $\boldsymbol{\omega}$, and the maximum value of the cost function (6.6) is greater than or equal to the maximum value in (6.3).

Consider the matrix $\mathbf{T} = \mathbf{U}\mathbf{\Omega}^{1/2}\mathbf{V}^*$ where $\mathbf{\Omega}$ is a diagonal matrix containing the maximizing $\boldsymbol{\omega}$ on the diagonal. For this choice of $\mathbf{T}$, the inequalities (6.4) and (6.5) are both equalities. Hence, this choice for $\mathbf{T}$ attains the maximum in (6.3). The corresponding optimal solution of (1.2) is

$$\mathbf{S} = \mathbf{Q}^{-1/2}\mathbf{T} = \mathbf{U}\mathbf{\Lambda}^{-1/2}\mathbf{U}^*\mathbf{U}\mathbf{\Omega}^{1/2}\mathbf{V}^* = \mathbf{U}\mathbf{\Lambda}^{-1/2}\mathbf{\Omega}^{1/2}\mathbf{V}^*. \tag{6.7}$$

To complete the proof of the theorem, we need to explain how to compute the optimal $\boldsymbol{\omega}$ in (6.6).

At the optimal solution of (6.6), the power constraint must be an equality (otherwise, we could multiply $\boldsymbol{\omega}$ by a positive scalar and increase the cost). Let us ignore the monotonicity constraint $\omega_i \geqslant \omega_{i+1}$ (we will show that the maximizer satisfies this constraint automatically). After taking the log of the cost function, we obtain the following simplified version of (6.6):

$$\begin{aligned} &\max_{\boldsymbol{\omega}} \sum_{i=1}^{N} \log(\omega_i + \delta_i^{-2}) \\ &\text{subject to} \quad \sum_{i=1}^{N} \omega_i \lambda_i^{-1} = P, \quad \boldsymbol{\omega} \geqslant \mathbf{0}. \end{aligned} \tag{6.8}$$

Since the cost function is strictly concave, the maximizer of (6.8) is unique.

The first-order optimality conditions (KKT conditions) for an optimal solution of (6.8) are the following: There exists a scalar $\mu \geqslant 0$ and a vector $\nu \in \mathbb{R}^n$ such that

$$-\frac{1}{\omega_i + \delta_i^{-2}} + \frac{\mu}{\lambda_i} - \nu_i = 0, \quad \nu_i \geqslant 0, \;\; \omega_i \geqslant 0, \;\; \nu_i\omega_i = 0, \;\; 1 \leqslant i \leqslant N. \tag{6.9}$$

Analogous to the proof of Theorem 4.1, we define the function

$$\omega_i(\mu) = \left(\frac{\lambda_i}{\mu} - \delta_i^{-2}\right)^+. \tag{6.10}$$

This particular value for ω_i is obtained by setting $\nu_i = 0$ in (6.9), solving for ω_i; when the solution is < 0, we set $\omega_i(\mu) = 0$ (this corresponds to the + operator (6.10)). Observe that $\omega_i(\mu)$ in (6.10) is a decreasing function of μ which approaches $+\infty$ as μ approaches 0 and which approaches 0 as μ tends to $+\infty$. Hence, the equation

$$\sum_{i=1}^{n} \omega_i(\mu)\lambda_i^{-1} = P \tag{6.11}$$

has a unique positive solution. We have $\omega_i = 0$ for $\mu \geqslant \lambda_i\delta_i^2$, which implies that

$$\nu_i = -\frac{1}{\omega_i(\mu) + \delta_i^{-2}} + \frac{\mu}{\lambda_i} = -\frac{1}{\delta_i^{-2}} + \frac{\mu}{\lambda_i} \geqslant -\delta_i^2 + \delta_i^2 = 0, \quad \text{when } \mu \geqslant \lambda_i\delta_i^2.$$

It follows that the KKT conditions are satisfied when μ is the positive solution of (6.11). Since the λ_i and δ_i are both arranged in decreasing order, it follows that for any choice $\mu \geqslant 0$, the ω_i given by (6.10) are in decreasing order. Hence, the constraint $\omega_{i+1} \leqslant \omega_i$ in (6.6) is satisfied by the solution of (6.8). Combining the formula (6.10) for the solution of (6.8) with the expression (6.7) for the solution of (1.2), we obtain the solution $\mathbf{S}$ given in (6.1) and (6.2) where $\mathbf{\Sigma} = \mathbf{\Lambda}^{-1/2}\mathbf{\Omega}^{1/2}$.

Now suppose that either $\mathbf{D}$ or $\mathbf{Q}$ is singular. Let us consider a perturbed problem where we replace $\mathbf{Q}$ by $\mathbf{Q}_\epsilon = \mathbf{U}\mathbf{\Lambda}_\epsilon\mathbf{U}^*$ and $\mathbf{D}$ by $\mathbf{D}_\epsilon = \mathbf{V}\mathbf{\Delta}_\epsilon\mathbf{V}^*$:

$$\max_{\mathbf{S}} \det(\mathbf{D}_\epsilon \mathbf{S}^* \mathbf{Q}_\epsilon \mathbf{S} \mathbf{D}_\epsilon + \mathbf{I}) \quad \text{subject to} \quad \operatorname{tr}(\mathbf{S}^*\mathbf{S}) \leqslant P, \quad \mathbf{S} \in \mathbb{C}^{m \times n}. \tag{6.12}$$

Here $\mathbf{\Lambda}_\epsilon$ and $\mathbf{\Delta}_\epsilon$ are obtained from $\mathbf{\Lambda}$ and $\mathbf{\Delta}$ by setting $\delta_i = \epsilon = \lambda_j$ for i or $j > K$. Since $\mathbf{Q}_\epsilon$ and $\mathbf{D}_\epsilon$ are nonsingular, it follows from our previous analysis that the perturbed problem (6.12) has a solution of the form $\mathbf{S}_\epsilon = \mathbf{U}\mathbf{\Sigma}_\epsilon\mathbf{V}^*$ where the diagonal elements of $\mathbf{\Sigma}_\epsilon$ are given by

$$\sigma_i^\epsilon = \begin{cases} \max\left\{\dfrac{1}{\mu} - \dfrac{1}{\lambda_i \delta_i^2}, 0\right\}^{1/2} & \text{for } 1 \leqslant i \leqslant K, \\ \max\left\{\dfrac{1}{\mu} - \dfrac{1}{\epsilon^3}, 0\right\}^{1/2} & \text{for } i > K. \end{cases} \tag{6.13}$$

Let μ be chosen so that

$$\sum_{i=1}^{K} (\sigma_i^\epsilon)^2 = P.$$

Observe that when $\epsilon^3 < \mu$, we have $\sigma_i^\epsilon = 0$ for $i > K$ and

$$\sum_{i=1}^{N} (\sigma_i^\epsilon)^2 = P.$$

Hence, for each $\epsilon > 0$ with $\epsilon^3 < \mu$, the optimal solution of the perturbed problem does not depend on ϵ and the trailing diagonal elements σ_i^ϵ for $i > K$ vanish.

Let $\mathbf{S}_0$ denote the matrix $\mathbf{S}_\epsilon$ for any value of ϵ satisfying $0 < \epsilon^3 < \mu$, and let $\mathbf{D}_k$ and $\mathbf{Q}_k$ denote the sequence of matrices corresponding to $\mathbf{D}_\epsilon$ and $\mathbf{Q}_\epsilon$ with $\epsilon = 2^{-k}$. By the optimality of $\mathbf{S}_0$, we have

$$\det(\mathbf{D}_k \mathbf{T}^* \mathbf{Q}_k \mathbf{T} \mathbf{D}_k + \mathbf{I}) \leqslant \det(\mathbf{D}_k \mathbf{S}_0^* \mathbf{Q}_k \mathbf{S}_0 \mathbf{D}_k + \mathbf{I}), \tag{6.14}$$

whenever $\mathbf{T}$ satisfies the constraint $\operatorname{tr}(\mathbf{T}^*\mathbf{T}) \leqslant P$ and k is sufficiently large. We let k tend to infinity in (6.14). By continuity, it follows that

$$\det(\mathbf{D}\mathbf{T}^*\mathbf{Q}\mathbf{T}\mathbf{D} + \mathbf{I}) \leqslant \det(\mathbf{D}\mathbf{S}_0^*\mathbf{Q}\mathbf{S}_0\mathbf{D} + \mathbf{I}).$$

Consequently, the solution (6.1)–(6.2) is valid, even when either $\mathbf{Q}$ or $\mathbf{D}$ is singular. □

We note that the solution to the determinant optimization problem described in Theorem 6.1 has a standard water-filling interpretation [3]. The idea of upper bounding the cost function as in (6.4), lower bounding the constraint as in (6.5), and showing that the bounds are tight, is also used in a different context in [24,25].

7. Solution comparison and computation

Referring to Corollary 3.4, the solution to the trace problem (1.1) has the form $\mathbf{S} = \mathbf{U}\mathbf{\Pi}\mathbf{\Sigma}\mathbf{V}^*$ where the columns of $\mathbf{U}$ and $\mathbf{V}$ are orthonormal eigenvectors of $\mathbf{Q}$ and $\mathbf{D}$, respectively, whose associated eigenvalues are in decreasing order, and $\mathbf{\Pi}$ is a permutation matrix. By Theorem 6.1, the solution to the determinant problem (1.2) has the form $\mathbf{S} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*$. When the power P is small, the solution of the trace problem is given by (5.9), which implies that $\mathbf{\Pi}$ is the identity when P

is small. Hence, the solution of the trace problem coincides with the solution to the determinant problem for small P.

The equivalence of the two solutions is based on the fact that the diagonal elements σ_i of $\mathbf{\Sigma}$ all vanish, except for those elements associated with dominant eigenvalues of $\mathbf{Q}$ and $\mathbf{D}$, for small P. The power threshold where the solutions first depend on nondominant eigenvalues is different for the two problems. As P increases, the solution to the two problems become more different. In particular, by Theorem 5.1, the optimal permutation $\mathbf{\Pi}$ reverses the ordering of the first K columns of $\mathbf{U}$ for P sufficiently large, where K is the minimum rank of $\mathbf{Q}$ and $\mathbf{D}$.

The determinant problem is much easier to solve than the trace problem since there is no permutation to compute. In this paper, we do not propose an algorithm for computing the optimal permutation π in the trace problem; however, based on Theorems 5.1 and 5.2, and on Lemma 3.2, we propose the following strategy to approximate the optimal permutation: Let K denote the minimum of the ranks of $\mathbf{Q}$ and $\mathbf{D}$. In (3.16) we replace the set $\mathscr{P}_m$ of $m!$ permutations by a set $\bar{\mathscr{P}}$ of K permutations:

$$\bar{\mathscr{P}} = \{\pi_1, \pi_2, \ldots, \pi_K\},$$

where

$$\pi_j(i) = j + 1 - i \ \text{ for } 1 \leqslant i \leqslant j \quad \text{and} \quad \pi_j(i) = i \ \text{ for } i > j.$$

The motivation for focusing on the set $\bar{\mathscr{P}}$ is the following: As the power increases, the number of nonzero components of $\boldsymbol{\sigma}$ in the optimal solution to (3.16) increases; and when the power is large, the optimal permutation is given by the reverse ordering (5.1). According to Lemma 3.2, the nonzero components of an optimal $\boldsymbol{\sigma}$ are associated with the largest eigenvalues of $\mathbf{Q}$ and $\mathbf{D}$. Hence, we focus on permutations that reverse the ordering of the first j indices, $1 \leqslant j \leqslant K$.

To investigate the quality of the approximations obtained by replacing $\mathscr{P}_m$ by $\bar{\mathscr{P}}$ in (3.16), we performed the following numerical experiment: 5 choices for the power were considered: $P = 1$, 10, 100, 1000, 10,000. For each choice of P, we randomly generate 100 sets of eigenvalues δ_i and λ_i on the interval [0, 1]. We computed, using Matlab, both the exact solution of (3.16) and the approximation corresponding to the special set $\bar{\mathscr{P}}$ of permutations. To compute the exact solution of (3.16), the complete set of permutations $\mathscr{P}_K$ can be enumerated efficiently by the algorithms of Johnson [10] and Trotter [20] (see the clear exposition of Brualdi [1]). In Table 7.1, we give the results corresponding to $m = n = 5$.

Table 7.1 shows both the number of times when the exact solution of (3.16) was given by a permutation π in $\bar{\mathscr{P}}$; in those cases where the exact solution to (3.16) was outside $\bar{\mathscr{P}}$, we computed the average relative error in the approximation obtained by using $\bar{\mathscr{P}}$ in place of $\mathscr{P}_m$ in (3.16). The relative error is obtained by subtracting the exact optimal cost from the approximation and dividing by the exact optimal cost. Since there were only a few cases where the exact solution of (3.16) was not contained in $\bar{\mathscr{P}}$, the average relative error reported in Table 7.1 is a very rough approximation to the true average. In Table 7.2, we give the results corresponding to $m = n = 10$.

Table 7.1
$\bar{\mathscr{P}}$ versus $\mathscr{P}_m$ for $m = n = 5$

P	Exact cases (out of 100)	Relative error (inexact cases)
1	100	.000000
10	96	.000064
100	93	.000740
1000	99	.000649
10,000	100	.000000

Table 7.2
$\bar{\mathscr{P}}$ versus $\mathscr{P}_m$ for $m = n = 10$

P	Exact cases (out of 100)	Relative error (inexact cases)
1	100	.000000
10	82	.000007
100	92	.000416
1000	94	.000287
10,000	100	.000000

The data in Tables 7.1 and 7.2 indicates that with high probability, we can solve (3.16) by restricting our attention to permutations in $\bar{\mathscr{P}}$. Moreover, in the few cases where the exact solution is not obtained using $\bar{\mathscr{P}}$, the approximate cost agrees with the exact cost to within 3 or more digits, on average. The time to solve (4.1) is at most $O(K^2)$. Hence, the time to solve (3.16), with $\mathscr{P}_m$ replaced by $\bar{\mathscr{P}}$, is $O(K^3)$. In contrast, the time to solve the general problem (3.16) by considering all possible permutations is $K!O(K^2)$.

8. Conclusions

We analyze two matrix optimization problems. The trace problem (1.1) arises in the design of multiple-input multiple-output (MIMO) systems. The solution is the training sequence which gives the best estimate for the matrix representing the communication channel. The solution is expressed in the form $\mathbf{S} = \mathbf{U\Pi\Sigma V}^*$ where the columns of $\mathbf{U}$ and $\mathbf{V}$ are orthonormal eigenvectors of $\mathbf{Q}$ and $\mathbf{D}$, respectively, with associated eigenvalues arranged in decreasing order, $\mathbf{\Pi}$ is a permutation matrix, and $\mathbf{\Sigma}$ is diagonal. When the power constraint P is sufficiently small, all the diagonal elements of $\mathbf{\Sigma}$ are zero except for those associated with the dominant eigenvalues. The solution (5.9) is expressed in terms of the dominant eigenvectors of $\mathbf{Q}$ and $\mathbf{D}$. When P is sufficiently large, the optimal permutation (5.1) reverses the ordering of the first K columns of $\mathbf{U}$, where K is the minimum of the ranks of $\mathbf{Q}$ and $\mathbf{D}$. This column reversal result was the basis for a technique to approximate the solution to the trace problem; the complete set of permutations $\mathscr{P}_m$ is replaced by a reduced set of K permutations denoted $\bar{\mathscr{P}}$. In numerical experiments, the optimal permutation for the trace problem (1.1) was an element of $\bar{\mathscr{P}}$ with high probability.

The second matrix optimization problem is obtained by replacing the trace operation in the first problem by the determinant. The optimal solution maximizes the sum capacity of a communication channel. For small P, the solutions of the trace and determinant problems are the same. But for large P, the two solutions are different—there is no permutation in the solution to the determinant problem. The solution of the determinant problem is obtained using majorization theory, as developed in [14].

For additional optimization problems connected with the design of MIMO systems, see [6–9]. Recently in [12], we consider the following variation of the trace optimization problem (1.1):

$$\min_{\mathbf{S}} \operatorname{tr}(\mathbf{S}^*\mathbf{Q}\mathbf{S} + \mathbf{D})^{-1}$$
$$\text{subject to} \quad \operatorname{tr}(\mathbf{S}^*\mathbf{S}) \leqslant P, \quad \mathbf{S} \in \mathbb{C}^{m\times n},$$

where $\mathbf{Q}$ and $\mathbf{D}$ are Hermitian, positive definite matrices. Using techniques similar to those developed in this paper, it is shown that for any choice of the power P, the optimal solution has the structure given in Theorem 5.1. That is, $\mathbf{S} = \mathbf{U\Sigma V}^*$, where $\mathbf{U}$ and $\mathbf{V}$ are orthonormal matrices of eigenvectors for **Q** and **D**, respectively, with the eigenvalues of **Q** arranged in increasing order and the eigenvalues of **D** arranged in decreasing order.

Acknowledgments

Special thanks to the referees for their constructive comments and suggestions.

References

[1] R.A. Brualdi, Introductory Combinatorics, Prentice-Hall, Upper Saddle River, NJ, 1999.
[2] X. Cai, G.B. Giannakis, M.D. Zoltowski, Space-time spreading and block coding for correlated fading channels in the presence of interference, IEEE Trans. Commun. 53 (2005) 515–525.
[3] T.M. Cover, J.A. Thomas, Elements of Information Theory, John Wiley, New York, 1991.
[4] M. Fiedler, Bounds for the determinant of the sum of Hermitian matrices, Proc. Amer. Math. Soc. 30 (1971) 27–31.
[5] W.W. Hager, Updating the inverse of a matrix, SIAM Rev. 31 (1989) 1221–1239.
[6] Y. Jiang, W.W. Hager, J. Li, The generalized triangular decomposition, Math. Comp., Submitted for publication.
[7] Y. Jiang, W.W. Hager, J. Li, The geometric mean decomposition, Linear Algebra Appl. 396 (2005) 373–384.
[8] Y. Jiang, J. Li, W.W. Hager, Joint transceiver design for MIMO communications using geometric mean decomposition, IEEE Trans. Signal Process. 53 (2005) 3791–3803.
[9] Y. Jiang, J. Li, W.W. Hager, Uniform channel decomposition for MIMO communications, IEEE Trans. Signal Process. 53 (2005) 4283–4294.
[10] S.M. Johnson, Generation of permutations by adjacent transpositions, Math. Comput. 17 (1963) 282–285.
[11] S.M. Kay, Fundamentals of Statistical Signal Processing: Estimation Theory, Prentice-Hall, Philadelphia, 1993.
[12] Y. Liu, T.F. Wong, W.W. Hager, Training signal design for estimation of correlated MIMO channels with colored interference, IEEE Trans. Signal Process., submitted for publication.
[13] U. Madhow, M. Honig, MMSE in interference suppression for direct sequence spread-spectrum CDMA, IEEE Trans. Commun. 42 (1994) 3178–3188.
[14] A.W. Marshall, I. Olkin, Inequalities: Theory of Majorization and Its Applications, Academic Press, New York, 1979.
[15] D. Palomar, J. Cioffi, M. Lagunas, Joint *Tx–Rx* beam forming design for multicarrier MIMO channels: a unified framework for convex optimization, IEEE Trans. Signal Process. 51 (2003) 2381–2401.
[16] A. Ruhe, Perturbation bounds for means of eigenvalues and invariant subspaces, Nordisk Tidskr. Informationsbehandling (BIT) 10 (1970) 343–354.
[17] A. Scaglione, P. Stoica, S. Barbarossa, G.B. Giannakis, H. Sampath, Optimal designs for space-time linear precoders and decoders, IEEE Trans. Signal Process. 50 (2002) 1051–1064.
[18] D.S. Shiu, G.J. Foschini, M.J. Gans, J.M. Kahn, Fading correlation and its effect on the capacity of multielement antenna systems, IEEE Trans. Commun. 48 (2000) 502–513.
[19] I.E. Telatar, Capacity of multi-antenna Gaussian channels, Eur. Trans. Telecommun. 10 (1999) 585–595.
[20] H.F. Trotter, Algorithm 115, Commun. ACM 5 (1962) 434–435.
[21] S. Verdu, Capacity region of Gaussian CDMA channels: the symbol-synchronous case, in: 24th Annual Allerton Conference on Communications, Control and Computing, 1986, pp. 1025–1034.
[22] P. Viswanath, V. Anantharam, Optimal sequences for CDMA under colored noise: a schur-saddle function property, IEEE Trans. Inform. Theory 48 (2002) 1295–1318.
[23] A.J. Viterbi, CDMA-Principles of Spread Spectrum Communication, Addison-Wesley, Reading, MA, 1995.
[24] T.F. Wong, T.M. Lok, Transmitter adaptation in multicode DS-CDMA systems, IEEE J. Select Areas Commun. 19 (2001) 69–82.
[25] T.F. Wong, B. Park, Training sequence optimization in MIMO systems with colored interference, IEEE Trans. Commun. 52 (2004) 1939–1947.
[26] S. Zhou, G.B. Giannakis, Optimal transmitter eigen-beamforming and space-time block coding based on channel correlations, IEEE Trans. Inform. Theory 49 (2003) 1673–1690.