COMPUTATIONAL & APPLIED MATHEMATICS

Volume 29, N. 2, pp. 139–171, 2010

ISSN 0101-8205
www.scielo.br/cam

Error bound for a perturbed minimization problem related with the sum of smallest eigenvalues

MARCOS VINICIO TRAVAGLIA

Departamento de Matemática, Centro de Ciências da Natureza
Universidade Federal do Piauí, 64049-550 Teresina, PI, Brazil

E-mail: mvtravaglia@ufpi.edu.br

Abstract. Let C be a $n \times n$ symmetric matrix. For each integer $1 \leq k < n$ we consider the minimization problem $m(\varepsilon) := \min_X \{\mathrm{Tr}\{CX\} + \varepsilon f(X)\}$. Here the variable X is an $n \times n$ symmetric matrix, whose eigenvalues satisfy

$$0 \leq \lambda_i(X) \leq 1 \quad \text{and} \quad \sum_{i=1}^{n} \lambda_i(X) = k,$$

the number ε is a positive (perturbation) parameter and f is a Lipchitz-continuous function (in general nonlinear). It is well known that when $\varepsilon = 0$ the minimum value, $m(0)$, is the sum of the smallest k eigenvalues of C.
Assuming that the eigenvalues of C satisfy $\lambda_1(C) \leq \cdots \leq \lambda_k(C) < \lambda_{k+1}(C) \leq \cdots \leq \lambda_n(C)$, we establish the following upper and lower bounds for the minimum value $m(\varepsilon)$:

$$\sum_{i=1}^{k} \lambda_i(C) + \varepsilon \bar{f} \geq m(\varepsilon) \geq \sum_{i=1}^{k} \lambda_i(C) + \varepsilon \bar{f} - \frac{2kL^2}{\lambda_{k+1}(C) - \lambda_k(C)} \varepsilon^2,$$

where $\bar{f}$ is the minimum value of f over the solution set of unperturbed problem and L is the Lipschitz-constant of f. The above inequality shows that the error by replacing the upper bound (or the lower bound) by the exact value is at least quadratic in the perturbation parameter. We also treat the case that $\lambda_{k+1}(C) = \lambda_k(C)$. We compare the exact solution with the upper and lower bounds for some examples.

Mathematical subject classification: 15A42, 15A18, 90C22.

Key words: matrix analysis, sum of smallest eigenvalues, minimization problem involving matrices; nonlinear perturbation; semidefinite programming.

#CAM-120/09. Received: 28/VII/09. Accepted: 07/XII/09.

1 Introduction

We denote by M_n the set of $n \times n$ real matrices. The sum of diagonal of $C \in M_n$ is denoted by $\mathrm{Tr}\{C\}$. If C is a symmetric matrix then the sum of their first (smallest) k eigenvalues, $\sum_{i=1}^{k} \lambda_i(C)$ (multiplicity included), can be obtained as the minimum value of the following minimization problem with matrix variable (see the proposition 2.3):

$$\min \left\{ \mathrm{Tr}\{CX\} \colon \mathrm{Tr}\{X\} = k,\ I - X \succeq 0 \ \text{ and } \ X \succeq 0 \right\} = \sum_{i=1}^{k} \lambda_i(C). \qquad (1)$$

The notation $X \succeq 0$ means that X is an $n \times n$ (symmetric) positive semidefinite matrix, that is, X is symmetric and its eigenvalues are nonnegative. We write $Y \succeq X$ when the difference matrix $Y - X \succeq 0$. We denote by I the identity matrix.

Note that the objective function, $\mathrm{Tr}\{CX\}$, and the restriction, $\mathrm{Tr}\{X\} = k$, of the minimization problem (1) are linear in the variable X. On the other hand, the restrictions $I - X \succeq 0$ and $X \succeq 0$ are convex but nonlinear. Due to these last two restrictions the problem (1) is, in general, not a *Linear Programming* (LP).

We denote by $\mathcal{K} := \{X \in M_n \colon \mathrm{Tr}\{X\} = k,\ I - X \succeq 0 \text{ and } X \succeq 0\}$ the domain of the objective function of the minimization problem (1). This means that the elements of $\mathcal{K}$ are symmetric matrices, whose eigenvalues satisfy: $0 \leq \lambda_i(X) \leq 1$ for $i = 1, 2, \ldots, n$ and $\sum_{i=1}^{n} \lambda_i(X) = k$.

In this work we propose to study a nonlinear perturbation of the problem (1) defined as follows:

$$m(\varepsilon) := \min \left\{ \mathrm{Tr}\{CX\} + \varepsilon f(X) \colon X \in \mathcal{K} \right\}, \qquad (2)$$

where ε (*perturbation parameter*) is a nonnegative real number and $f \colon \mathcal{K} \to \mathbb{R}$ is a function that is, in general, nonlinear. The function $\varepsilon f(X)$ is called *perturbation*. We denote by $m(\varepsilon)$ the minimum value of the problem (2). In general the solution of (2) depends on ε. Moreover, according to (1) one has $m(0) = \sum_{i=1}^{k} \lambda_i(C)$.

We close the introductory section with a motivation to study the perturbed minimization problem (2):

Semidefinite Programming: In particular case $k = 1$ the restriction $I - X \succeq 0$ can be dropped in (1). To see this, note that if $\lambda_i(X) \geq 0$ and $\sum_{i=1}^{n} \lambda_i(X) = \mathrm{Tr}\{X\} = 1$ then $0 \leq \lambda_i(X) \leq 1$. Hence, $I - X \succeq 0$. The minimization problem ($k = 1$)

$$\min \left\{ \mathrm{Tr}\{CX\} \colon \mathrm{Tr}\{X\} = 1 \text{ and } X \succeq 0 \right\} \tag{3}$$

is a special case of the following problem:

$$\min \left\{ \mathrm{Tr}\{CX\} \colon \mathrm{Tr}\{A_i X\} = b_i \text{ with } i = 1, \ldots, p \text{ and } X \succeq 0 \right\}, \tag{4}$$

which is called standard *semidefinite programming* (SDP). In (4) A_i are $n \times n$ symmetric matrices. The SDP has many applications in Combinatorial Optimization. See [5] for a survey on SDP and [6] for the relation between SDP and Eigenvalue Optimization.

A bit more general case of (3) is $\min\{\mathrm{Tr}\{CX\} \colon \mathrm{Tr}\{AX\} = b \text{ and } X \succeq 0\}$ with $A \succ 0$ (that is, A is strict positive definite) and $b > 0$. We can reduce this case to (3) by plugging $\tilde{C} := bA^{-1/2}CA^{-1/2}$ in the place of C in (3).

Strictly Convex Perturbation of SDP: It is convenient to add a strictly convex function $\varepsilon f(x)$ in order to solve the unperturbed problem (4). This makes the perturbed minimization problem strictly convex. Hence, it has only one minimizer. One hopes that this minimizer approximates to a solution of (4) as $\varepsilon \to 0$. The interior point methods [10], for solve SDPs, make use of the strictly convex function $\varepsilon f(X) = -\varepsilon \log\det(X)$, which is a log-barrier. Since this function is not Lipschtiz-continuous our error bound can not be used. On the other hand, the authors [7] arrived at an algorithm for SDPs that has several advantages over existing techniques using the perturbation functions $\varepsilon f(X) = -\varepsilon \log\det(X)$ and $\varepsilon f(X) = \varepsilon \frac{1}{2}\mathrm{Tr}\{X^2\}$. The last one is strictly convex and Lipschitz-continuous. As the main advantage, this algorithm is a first-order method, which makes it scalable.

2 Results

Since we can not obtain in general the exact value of $m(\varepsilon)$ for $\varepsilon > 0$, we propose to establish an upper $u(\varepsilon)$ and a lower bound $\ell(\varepsilon)$ for $m(\varepsilon)$. That is, $u(\varepsilon) \geq m(\varepsilon) \geq \ell(\varepsilon)$, from which follows that $0 \leq m(\varepsilon) - \ell(\varepsilon) \leq u(\varepsilon) - \ell(\varepsilon)$

and $0 \leq u(\varepsilon) - m(\varepsilon) \leq u(\varepsilon) - \ell(\varepsilon)$. Under some condition on f we prove that $u(\varepsilon) - \ell(\varepsilon) = \text{constant}\ \varepsilon^2$ (see theorem 2.8). Hence, both $m(\varepsilon) - \ell(\varepsilon)$ and $u(\varepsilon) - m(\varepsilon)$ go at least quadratic to zero as ε goes to zero. We interpret $u(\varepsilon) - \ell(\varepsilon)$ as an error bound. In contrast to the minimization problem (2), the authors [1] proved that for perturbed LP (see section 4) $u_{\mathrm{LP}}(\varepsilon) - \ell_{\mathrm{LP}}(\varepsilon) = 0$ for all $0 \leq \varepsilon \leq \varepsilon_o$ with some $\varepsilon_o > 0$. This means that for perturbed LP there is no error if we require that ε be small enough. This is, in general, not the case for the minimization problem (2) (see example 1 of section 3). The author [9] derived error bound for perturbed LP when f is strictly convex, and in [8] the perturbation results of [1] were extend to convex programmings.

As in [1], we assume that $f\colon \mathcal{K} \to \mathbb{R}$ be a Lipschitz-continuous function, that is, there is a constant $0 \leq L < \infty$ such that for all $X, Y \in \mathcal{K}$ the inequality $|f(X) - f(Y)| \leq L\|X - Y\|$ is true. Throughout this paper $\|\cdot\|$ denotes the *Frobenius-norm*, that is, for $X \in M_n$ we define

$$\|X\| := \sqrt{\mathrm{Tr}\{X^T X\}} = \sqrt{\sum_{i,j=1}^{n} X_{i,j}^2}.$$

It is well known that for a symmetric matrix X the equality $\|X\| = \sqrt{\sum_{i=1}^{n} \lambda_i^2(X)}$ holds.

The following proposition establishes that the set $\mathcal{K}$ is a bounded subset of $M_n \simeq \mathbb{R}^{n\times n}$. Hence, $\mathcal{K}$ is compact.

Proposition 2.1. *The set $\mathcal{K}$ is a subset of the ball of M_n with size $\sqrt{k}$, that is, if $X \in \mathcal{K}$ then $\|X\| \leq \sqrt{k}$.*

Proof. For $X \in \mathcal{K}$ we have $0 \leq \lambda_i(X) \leq 1$ for $i = 1, \ldots, n$. Therefore

$$\|X\|^2 = \sum_{i=1}^{n} \lambda_i^2(X) \leq \sum_{i=1}^{n} \lambda_i(X) = \mathrm{Tr}\{X\} = k \qquad \square$$

Since $\mathcal{K}$ is compact and $\mathrm{Tr}\{C\,X\}$ is continuous the minimum value of the problem (1) is attained.

In order to characterize the minimizers of the problem (1) we need the following definition:

Definition 2.2 (The values $d, r, s, \lambda_r(C)$, and $\lambda_s(C)$)**.** Consider the eigenvalues of the symmetric matrix $C \in M_n$ in increasing order, that is, $\lambda_1(C) \leq \lambda_2(C) \leq \cdots \leq \lambda_n(C)$. For each $k = 1, 2, \ldots, n$ we define:

$$d = d(C,k) := \sharp\big\{i \in \{1,\ldots,n\}\colon \lambda_i(C) = \lambda_k(C)\big\}$$

Note that d is multiplicity (degeneracy) of k-th eigenvalue. Further we define:

$$\mathcal{R}(C,k) := \big\{i \in \{1,\ldots,n\}\colon \lambda_i(C) < \lambda_k(C)\big\},$$

$$S(C,k) := \big\{i \in \{1,\ldots,n\}\colon \lambda_i(C) > \lambda_k(C)\big\},$$

$$r = r(C,k) := \begin{cases} \max\{i\colon i \in \mathcal{R}(C,k)\} & \text{if } \mathcal{R}(C,k) \neq \emptyset \\ 0 & \text{if } \mathcal{R}(C,k) = \emptyset, \end{cases} \tag{5}$$

and

$$s = s(C,k) := \begin{cases} \min\{i\colon i \in S(C,k)\} & \text{if } S(C,k) \neq \emptyset \\ n+1 & \text{if } S(C,k) = \emptyset. \end{cases} \tag{6}$$

In the case $r = 0$ we do the following convention $\lambda_r(C) = \lambda_0(C) := -\infty$ and in the case $s = n+1$ we do $\lambda_s(C) = \lambda_{n+1}(C) := +\infty$. Note that $s - r - 1 = d$ holds true.

The following example illustrates the above definition:

Example. For $n = 10$ and $k = 6$. If the eigenvalues of C are:

a) $2, 3, 4, 4, 4, 4, 4, 4, 11, 12$ then $\lambda_k = 4$, $d = 6$, $r = 2$, $s = 9$, $\lambda_r = 3$ and $\lambda_s = 11$.

b) $4, 4, 4, 4, 4, 4, 4, 4, 11, 12$ then $\lambda_k = 4$, $d = 8$, $r = 0$, $s = 9$, $\lambda_r = -\infty$ and $\lambda_s = 11$.

c) $2, 2, 2, 3, 4, 4, 4, 4, 4, 4$ then $\lambda_k = 4$, $d = 6$, $r = 4$, $s = 11$, $\lambda_r = 3$ and $\lambda_s = +\infty$.

d) $4, 4, 4, 4, 4, 4, 4, 4, 4, 4$ then $\lambda_k = 4$, $d = 10$, $r = 0$, $s = 11$, $\lambda_r = -\infty$, $\lambda_s = +\infty$. Note that, in this case, $C = 4I$.

Proposition 2.3. *The minimum value of the problem* (1) *is the sum of the smallest k eigenvalues of the symmetric matrix C. Moreover, the set of its minimizers is*

$$\left\{ V \left[\begin{array}{c|c|c} I_r & 0 & 0 \\ \hline 0 & Z & 0 \\ \hline 0 & 0 & 0 \end{array}\right] V^T : Z \in \mathcal{K}_d \right\}, \tag{7}$$

where

$$\mathcal{K}_d := \left\{ Z \in M_d \text{ with } \operatorname{Tr}\{Z\} = k - r \text{ and } I \succeq Z \succeq 0 \right\}. \tag{8}$$

Here $d = d(C,k)$ is the multiplicity of the k-th eigenvalue of C and V is any $n \times n$ orthogonal matrix whose columns are the eigenvectors (in increasing order of their eigenvalues) of the matrix C.

The matrix I_r in (7) *is the identity matrix of order r, where $r = r(C,k)$ is defined by* (5). *If $r = 0$ then this identity matrix does not appear in* (7). *The block matrix* 0 *in the diagonal of* (7) *is of order $n + 1 - s$, where $s = s(C,k)$ is defined by* (6). *If $s = n + 1$ then this block matrix* 0 *does not appear in* (7).

If $d = 1$ then there is only one minimizer, namely: $\sum_{i=1}^{k} v_i v_i^T$ (orthogonal projection with rank k), where v_i is any (normalized) eigenvector corresponding to the i-th eigenvalue of C.

Proof. See appendix A.

Remark 2.4. In order to simplify the notation we mean by $A \oplus B$ (direct sum) the block matrix $\left[\begin{array}{c|c} A & 0 \\ \hline 0 & B \end{array}\right]$. That is, $A \oplus B$ is a square matrix of order $n_a + n_b$ if A and B are square matrices of order n_a and n_b respectively.

Since the objective function $\operatorname{Tr}\{CX\} + \varepsilon f(X)$ is the sum of two continuous functions on the compact set $\mathcal{K}$ its minimum value, $m(\varepsilon)$, is attained. We denote by $\mathcal{X}_*(\varepsilon)$ the set $\operatorname{argmin}\{\operatorname{Tr}\{CX\} + \varepsilon f(X) : X \in \mathcal{K}\}$ (the set of minimizers). The proposition 2.3 states that $m(0) = \sum_{i=1}^{k} \lambda_i(C)$ and

$$\mathcal{X}_*(0) = \left\{ V(I_r \oplus Z \oplus 0) V^T : Z \in M_d \text{ with } \operatorname{Tr}\{Z\} = k - r \text{ and } I \succeq Z \succeq 0 \right\}.$$

An important value of f in order to study the relation between $m(\varepsilon)$ and $m(0)$ is $\bar{f} := \min\{f(X) : X \in \mathcal{X}_*(0)\}$.

We now establish the following upper bound for $m(\varepsilon)$:

Proposition 2.5 (Upper bound)**.** *For the minimum value $m(\varepsilon)$ of the problem* (2) *the following upper bound holds:*

$$u(\varepsilon) := m(0) + \varepsilon \bar{f} \geq m(\varepsilon)$$

for all $\varepsilon \geq 0$.

Proof. Take $Y_* \in \operatorname{argmin}\{f(X)\colon X \in \mathcal{X}_*(0)\}$. Since $Y_* \in \mathcal{K}$ we have by definition of $m(\varepsilon)$ that

$$\begin{aligned} m(\varepsilon) &\leq \operatorname{Tr}\{CY_*\} + \varepsilon f(Y_*) \\ &= m(0) + \varepsilon f(Y_*) \qquad \text{since } Y_* \in \mathcal{X}_*(0) \\ &= m(0) + \varepsilon \bar{f} \end{aligned}$$

□

Remark 2.6. In the case $d(C,k) = 1$ we have $\bar{f} = f\big(\sum_{i=1}^{k} v_i v_i^T\big)$ because

$$\mathcal{X}_*(0) = \Big\{P := \sum_{i=1}^{k} v_i v_i^T\Big\}$$

(there is only one minimizer). Hence, $u(\varepsilon) = m(0) + \varepsilon f(P)$ is an upper bound. On the other hand, in the case $d(C,k) > 1$, the function $U(\varepsilon) := m(0) + \varepsilon f(P) = \operatorname{Tr}\{CP\} + \varepsilon f(P)$ is clearly an upper bound for $m(\varepsilon)$. If $f(P) > \bar{f}$ then $U(\varepsilon) > u(\varepsilon)$ for $\varepsilon > 0$. Therefore, the upper bound $U(\varepsilon)$ is not interesting because when we are estimating an error we try, in general, to find smaller upper bounds and larger lower bounds.

We can easily obtain a crude lower bound for $m(\varepsilon)$ as the following:

Proposition 2.7 (A linear error bound)**.** *For the minimum value of* (2) *the following upper and lower bounds hold:*

$$m(0) + \varepsilon \bar{f} \geq m(\varepsilon) \geq m(0) + \varepsilon \bar{f} - 2\sqrt{k} L \varepsilon$$

for all $\varepsilon \geq 0$.

Proof. The upper bound, $u(\varepsilon) := m(0)+\varepsilon\bar{f}$, was proved in proposition 2.5. To prove the lower bound take $X_* \in \mathcal{X}_*(\varepsilon)$ and $Y_* \in \operatorname{argmin}\{f(X)\colon X \in \mathcal{X}_*(0)\}$. Hence

$$\begin{aligned}
m(\varepsilon) - u(\varepsilon) &:= m(\varepsilon) - m(0) - \varepsilon\bar{f} \\
&= \operatorname{Tr}\{CX_*\} + \varepsilon f(X_*) - \operatorname{Tr}\{CY_*\} - \varepsilon f(Y_*) \\
&= \Big(\operatorname{Tr}\{CX_*\} - \operatorname{Tr}\{CY_*\}\Big) + \varepsilon\Big(f(X_*) - f(Y_*)\Big) \\
&\geq 0 + \varepsilon\Big(f(X_*) - f(Y_*)\Big) \quad \begin{matrix}\text{since } {\scriptstyle Y_*\in\mathcal{X}_*(0)} \\ \text{and } {\scriptstyle X_*\in\mathcal{K}}\end{matrix} \\
&\geq -\varepsilon|f(X_*) - f(Y_*)| \\
&\geq -\varepsilon L\|X_* - Y_*\| \qquad \text{since } f \text{ is Lipschitz} \\
&\geq -\varepsilon L\Big(\|X_*\| + \|Y_*\|\Big) \\
&\geq -\varepsilon L(\sqrt{k} + \sqrt{k})
\end{aligned}$$

In last step we used that X_* and $Y_* \in \mathcal{K}$ and the proposition 2.1. This proves the proposition 2.7. □

The main result of this work is the following theorem, which improves the proposition 2.7:

Theorem 2.8 (A quadratic error bound)**.** *If f is a Lipschitz-continuous function with Lipschitz-constant L then the following upper and lower bound hold:*

$$m(0) + \varepsilon\bar{f} \geq m(\varepsilon) \geq m(0) + \varepsilon\bar{f} - \frac{L^2}{\alpha(C,k)}\varepsilon^2$$

for all $0 \leq \varepsilon$. Here $\alpha(C,k)$ is strict positive and given by

$$\alpha(C,k) = \frac{1}{2}\min\left\{\frac{\lambda_k(C) - \lambda_r(C)}{s-(k+1)}, \frac{\lambda_s(C) - \lambda_k(C)}{k}\right\}, \tag{9}$$

where the values $r = r(C,k)$ and $s = s(C,k)$ are given by the definition 2.2.

Example. In the last example ($n = 10$, and $k = 6$) we have for a) $\alpha(C,6) = 1/4$, b) $\alpha(C,6) = 7/12$, c) $\alpha(C,6) = 1/8$ and d) $\alpha(C,6) = +\infty$.

Remark 2.9. In the particular case, where the eigenvalues of C are ordered as $\lambda_1 \leq \lambda_2 \leq \lambda_k < \lambda_{k+1} \leq \lambda_n$, that is, $s = k+1$, we have $\alpha = \frac{\lambda_{k+1}-\lambda_k}{2k}$.

Remark 2.10. In the particular case $k = 1$ ($\Rightarrow r = 0$) we obtain:

$$\alpha(C, 1) = \frac{\lambda_{1+d(C,1)} - \lambda_1}{2},$$

where $d(C, 1)$ is the multiplicity (degeneracy) of $\lambda_1(C)$. That is, if $k = 1$, then α can be taken as the half of the gap of the first eigenvalue.

Remark 2.11. The expression (9) for the strict positive constant $\alpha(C, k)$ is obtained in lemma B.1 (see appendix B). The lemma B.1 states that $\alpha(C, k)$ given by (9) satisfies the inequality

$$\mathrm{Tr}\{CX\} - \sum_{i=1}^{k} \lambda_i(C) \geq \alpha(C, k) \min_{Y \in X_*(0)} \|X - Y\|^2 \tag{10}$$

for all $X \in \mathcal{K}$.

We comment two cases where $\alpha(C, k) = +\infty$.

Remark 2.12. If $d = n$ then one has $C = \lambda_1(C)I$ (a multiple of the identity matrix). Note that in this case we have $r = 0$ and $s = n + 1$. Hence, by definition (9) $\alpha(C, k) = +\infty$. Note that the *left hand side* (LHS) of (10) is zero since $\mathrm{Tr}\{CX\} = k\lambda_1(C)$ for $X \in \mathcal{K}$. This means that $X_*(0) = \mathcal{K}$. The RHS of (10) is also zero because in that case $X_*(0) = \mathcal{K}$. Therefore, the largest (the best) α for which the inequality (10) holds is $+\infty$. Note also that, in the case $d = n$, the equality $m(\varepsilon) = m(0) + \varepsilon \bar{f}$ holds for all $\varepsilon \geq 0$ since $X_*(0) = \mathcal{K}$. This is in agreement with the theorem 2.8: $0 \leq m(0) + \varepsilon \bar{f} - m(\varepsilon) \leq \frac{L^2}{\alpha(C)}\varepsilon^2$ since $\frac{L^2}{+\infty}\varepsilon^2 = 0$. That is, the theorem 2.8 confirms that if $d = n$ the exact minimum value $m(\varepsilon)$ coincides with the upper bound for all $\varepsilon \geq 0$.

Remark 2.13. If $k = n$ then $s = n+1$. According to definition (9) we also have $\alpha(C, k) = +\infty$. Note that, if $k = n$ then $\mathcal{K} = \{I\} = X_*(0)$ because if $X \in M_n$ with $\mathrm{Tr}\{X\} = n$ and $I \succeq X \succeq 0$ then $X = I$. Consequently both the LHS and the RHS of (10) are zero. This means that we can take $\alpha(C, n) = +\infty$. Note also that, in the case $k = n$, the equality $m(\varepsilon) = m(0) + \varepsilon \bar{f} = \mathrm{Tr}\{C\} + \varepsilon f(I)$ holds for all $\varepsilon \geq 0$ since $X_*(0) = \mathcal{K} = \{I\}$.

Proof of Theorem 2.8

Proof. The upper bound was proved in proposition 2.5. To prove the lower bound take a $X_* \in \mathcal{X}_*(\varepsilon)$, a $Y_* \in \operatorname{argmin}\{f(X)\colon X \in \mathcal{X}_*(0)\}$ and a $Y_{**} \in \operatorname{argmin}\{\|X_* - Y\|^2\colon Y \in \mathcal{X}_*(0)\}$.

Since $X_* \in \mathcal{X}_*(\varepsilon)$ and $Y_* \in \mathcal{K}$ we have:

$$0 \leq \operatorname{Tr}\{CY_*\} + \varepsilon f(Y_*) - \left(\operatorname{Tr}\{CX_*\} + \varepsilon f(X_*)\right). \tag{11}$$

Developing (11), using that $\operatorname{Tr}\{CY_*\} = \sum_{i=1}^{k} \lambda_i(C)$ and (10) we obtain:

$$\begin{aligned}
0 &\leq \operatorname{Tr}\{CY_*\} + \varepsilon f(Y_*) - \left(\operatorname{Tr}\{CX_*\} + \varepsilon f(X_*)\right) \\
&= -\left(\operatorname{Tr}\{CX_*\} - \operatorname{Tr}\{CY_*\}\right) + \varepsilon\left(f(Y_*) - f(X_*)\right) \\
&= -\left(\operatorname{Tr}\{CX_*\} - \sum_{i=1}^{k}\lambda_i(C)\right) + \varepsilon\left(f(Y_*) - f(X_*)\right) \\
&\leq -\alpha(C,k)\min_{Y\in\mathcal{X}_*(0)}\left\{\|X_* - Y\|^2\right\} + \varepsilon\left(f(Y_*) - f(X_*)\right) \quad \text{see (10)} \\
&= -\alpha(C,k)\|X_* - Y_{**}\|^2 + \varepsilon\left(f(Y_*) - f(X_*)\right) \\
&\leq -\alpha(C,k)\|X_* - Y_{**}\|^2 + \varepsilon\left(f(Y_{**}) - f(X_*)\right) \quad \text{by definition of } Y_* \\
&\leq -\alpha(C,k)\|X_* - Y_{**}\|^2 + \varepsilon L\|Y_{**} - X_*\| \quad \text{since } f \text{ is Lipschitz.}
\end{aligned}$$

We conclude two things:

$$\alpha(C,k)\|X_* - Y_{**}\|^2 \leq \varepsilon L\|X_* - Y_{**}\| \tag{12}$$

and

$$\begin{aligned}
&\operatorname{Tr}\{CY_*\} + \varepsilon f(Y_*) - \left(\operatorname{Tr}\{CX_*\} + \varepsilon f(X_*)\right) \\
&\leq -\alpha(C,k)\|X_* - Y_{**}\|^2 + \varepsilon L\|Y_{**} - X_*\|.
\end{aligned} \tag{13}$$

Note that $\operatorname{Tr}\{CY_*\} = \sum_{i=1}^{k}\lambda_i(C) = m(0)$ (because $Y_* \in \mathcal{X}_*(0)$). Further recall that

$$f(Y_*) = \min_{X\in\mathcal{X}_*(0)} f(X) =: \bar{f} \quad \text{and} \quad \operatorname{Tr}\{CX_*\} + \varepsilon f(X_*) = m(\varepsilon)$$

because $X_* \in \mathcal{X}_*(\varepsilon)$. Hence, we rewrite (13) as

$$m(0) + \varepsilon\bar{f} - m(\varepsilon) \leq -\alpha(C,k)\|X_* - Y_{**}\|^2 + \varepsilon L\|Y_{**} - X_*\|, \tag{14}$$

that is,

$$m(\varepsilon) \geq m(0) + \varepsilon \bar{f} + \alpha(C,k)\|X_* - Y_{**}\|^2 - \varepsilon L \|X_* - Y_{**}\| . \qquad (15)$$

Now, we consider two cases in (15):

Case 1: $X_* = Y_{**}$. In this case we obtain directly from (15) that $m(\varepsilon) \geq m(0) + \varepsilon \bar{f}$.

Case 2: $X_* \neq Y_{**}$. In this case follows from (12) that $\|X_* - Y_{**}\| \leq \frac{\varepsilon L}{\alpha(C,k)}$. Replacing this last inequality in (15) we obtain:

$$\begin{aligned} m(\varepsilon) &\geq m(0) + \varepsilon \bar{f} + \alpha(C,k)\|X_* - Y_{**}\|^2 - \frac{L^2}{\alpha(C,k)}\varepsilon^2 \\ &\geq m(0) + \varepsilon \bar{f} - \frac{L^2}{\alpha(C,k)}\varepsilon^2 . \end{aligned}$$

Since $m(0) + \varepsilon \bar{f} \geq m(0) + \varepsilon \bar{f} - \frac{L^2}{\alpha(C)}\varepsilon^2$ we conclude that in both cases the following lower bound for $m(\varepsilon)$ holds:

$$m(\varepsilon) \geq m(0) + \varepsilon \bar{f} - \frac{L^2}{\alpha(C,k)}\varepsilon^2 .$$

This proves the theorem 2.8. □

3 Examples and comparison between the perturbed matrix minimization problem (2) and perturbed LP

In this section we give two examples of the minimization problem (2) for $n = 2$ and $k = 1$. We present the exact minimum value and compare it with the upper and lower bounds. The first example consists in an perturbed SDP that is not a perturbed LP.

Example 1: Consider the matrix $C = \left[\begin{smallmatrix} 1 & 1 \\ 1 & 1 \end{smallmatrix}\right]$, $n = 2$, $k = 1$ and the nonlinear function $f(X) = X_{1,1}X_{1,1}$. In this example the minimization problem (2) is given by:

$$m(\varepsilon) = \min_{\substack{X_{1,1}+X_{2,2}=1\,, \\ X_{1,1}\,, X_{2,2}\geq 0\,, \\ X_{1,2}^2 \leq X_{1,1}X_{2,2}}} X_{1,1} + 2X_{1,2} + X_{2,2} + \varepsilon X_{1,1}^2 \qquad (16)$$

Remark 3.1. In (16) the constraint, $X_{1,2}^2 \leq X_{1,1}X_{2,2}$, is nonlinear and the variable $X_{1,2}$ appears in the objective function. Hence, the above perturbed matrix minimization problem is not a perturbed LP.

In this example we have:

- The eigenvalues and eigenvectors of C are: $\lambda_1(C) = 0$, $\lambda_2(C) = 2$, $v_1^T = \frac{1}{\sqrt{2}}(1, 1)$, $v_2^T = \frac{1}{\sqrt{2}}(1, -1)$;
- The solution set of linear part is $\mathcal{X}_*(0) = \left\{ v_1 v_1^T = \begin{bmatrix} 1/2 & -1/2 \\ -1/2 & 1/2 \end{bmatrix} \right\}$;
- The minimum value of f on the solution set of the linear part is $\bar{f} := \min_{X \in X_*(0)} f(X) = 1/4$;
- The Lipschitz-constant is $L = \sqrt{2}$. To see this, note that for $X, Y \in \mathcal{K}$ holds:

$$X_{2,2} = 1 - X_{1,1} \quad \text{and} \quad Y_{2,2} = 1 - Y_{1,1} \tag{17}$$

and

$$0 \leq X_{1,1}, Y_{1,1} \leq 1\,. \tag{18}$$

Taking (17) and (18) into account, we get

$$\begin{aligned}
\|X - Y\| &= \left(|X_{1,1} - Y_{1,1}|^2 + 2|X_{1,2} - Y_{1,2}|^2 + |X_{2,2} - Y_{2,2}|^2 \right)^{1/2} \\
&= \left(2|X_{1,1} - Y_{1,1}|^2 + 2|X_{1,2} - Y_{1,2}|^2 \right)^{1/2} \quad \text{due to (17)} \\
&\geq \sqrt{2}|X_{1,1} - Y_{1,1}| = \frac{\sqrt{2}}{2}\, 2\, |X_{1,1} - Y_{1,1}| \\
&\geq \frac{\sqrt{2}}{2}\, |X_{1,1} + Y_{1,1}|\, |X_{1,1} - Y_{1,1}| \quad \text{due to (18)} \\
&= \frac{1}{\sqrt{2}}\, |X_{1,1}^2 - Y_{1,1}^2|\,.
\end{aligned}$$

That is,

$$|X_{1,1}^2 - Y_{1,1}^2| \leq \sqrt{2}\, \|X - Y\| \tag{19}$$

From (19) we conclude that the Lipschitz-constant for $f(X) = X_{1,1}^2$ can be taken as $\sqrt{2}$. Moreover, $L = \sqrt{2}$ is the best (the smallest) Lipschitz-constant since $L_{\text{best}} = \sup\{|f(X) - f(Y)|/\|X - Y\| : X, Y \in$

$\mathcal{K}$ with $X \neq Y\} \geq \lim_{\delta\to 0} |f(A) - f(B_\delta)|/\|A - B_\delta\| = \sqrt{2}$ for the choice $A = \left[\begin{smallmatrix} 1 & 0 \\ 0 & 0 \end{smallmatrix}\right]$ and $B_\delta = \left[\begin{smallmatrix} 1-\delta & 0 \\ 0 & \delta \end{smallmatrix}\right]$ with $0 < \delta \leq 1$;

– The upper bound in this example (see proposition 2.5) is $u(\varepsilon) = \frac{1}{4}\varepsilon$;

– Since $r = 0$ and $s = 2$ the lower bound in this example (see theorem 2.8) is $\ell(\varepsilon) = \frac{1}{4}\varepsilon - 2\varepsilon^2$;

According to proposition C.1 the exact minimum value $m(\varepsilon)$ can be expressed as:

$$m(\varepsilon) = \max_{r\in[0,1]} \left\{ 1 + \varepsilon r - \sqrt{(\varepsilon r)^2 + 1} - \varepsilon r^2 \right\}. \tag{20}$$

For some values of $\varepsilon > 0$ the maximum value in (20) can be obtained with help of the software Maple (command maximize). We present the values of $m(\varepsilon)$ and the corresponding lower and upper bounds in the following table:

ε	$u(\varepsilon) = \varepsilon/4$	$m(\varepsilon)$	$\ell(\varepsilon) = \varepsilon/4 - 2\varepsilon^2$
1/10	0.2500000000 e-1	0.2381016622 e-1	0.0500000000 e-1
1/100	0.2500000000 e-2	0.2487562438 e-2	0.2300000000 e-2
1/1000	0.2500000000 e-3	0.2498748126 e-3	0.2480000000 e-3
1/10000	0.2500000000 e-4	0.2499949982 e-4	0.2498000000 e-4

Table 1 – Comparison among upper bound, exact minimum value and lower bound for example 1.

Table 1 shows that for $\varepsilon > 0$ the upper bound $u(\varepsilon)$ is always strict larger than the exact value $m(\varepsilon)$. We prove this fact rigorously in the proposition C.2.

Example 2: Consider the diagonal matrix $C = \left[\begin{smallmatrix} 0 & 0 \\ 0 & 2 \end{smallmatrix}\right]$, $n = 2$, $k = 1$, $r = 0$, $s = 2$ and the same nonlinear function $f(X) = X_{1,1}^2$ as in the example 1. In this case the minimization problem (2) is given by:

$$m(\varepsilon) = \min_{\substack{X_{1,1}+X_{2,2}=1\,,\\ X_{1,1}\,,\,X_{2,2}\geq 0\,,\\ X_{1,2}^2 \leq X_{1,1}X_{2,2}}} 2X_{2,2} + \varepsilon X_{1,1}^2\,.$$

Since the variable $X_{1,2}$ does not appear in the above objective function we can rewrite this minimization problem as:

$$m(\varepsilon) = \min_{0\leq X_{1,1}\leq 1} 2(1 - X_{1,1}) + \varepsilon X_{1,1}^2\,.$$

This shows that the example 2 of perturbed matrix minimization problem is not more than a perturbed LP.

In this example we have $m(0) = \lambda_1(C) = 0$, $\lambda_2(C) = 2$; $r = 0$, $s = 2$; $v_1^T = (1,0)$, $v_2^T = (0,1)$; $\bar{f} = f(v_1 v_1^T) = 1$ and $L = \sqrt{2}$ (see example 1). Therefore $u(\varepsilon) = \varepsilon$ (see proposition 2.5) and $\ell(\varepsilon) = \varepsilon - 2\varepsilon^2$ (see theorem 2.8). On the other hand, we can easily compute the exact minimum value as:

$$m(\varepsilon) = \begin{cases} \varepsilon & \text{if } 0 \leq \varepsilon \leq 1, \\ 2 - 1/\varepsilon & \text{if } 1 < \varepsilon. \end{cases} \tag{21}$$

On the contrary of example 1, note that $u(\varepsilon) = m(\varepsilon)$ for all $0 \leq \varepsilon \leq 1$. That is, there is no error by replacing the upper bound by the exact minimum value if we require that ε be small enough. This is a property of all Linear Programmings perturbed by a Lipschitz function (see section 4).

4 Note about LP perturbed by a Lipschtz function

In the context of Linear Programming (LP) the corresponding matrix minimization problem (2) is given by:

$$m_{\mathrm{LP}}(\varepsilon) := \min \left\{ \sum_{i=1}^{n} c_i x_i + \varepsilon\, f(x) : x \in \mathcal{K}_{\mathrm{LP}} \right\} \tag{22}$$

for a fixed $c \in \mathbb{R}^n$. Here the domain of the objective function is:

$$\mathcal{K}_{\mathrm{LP}} := \left\{ x \in \mathbb{R}^n : \sum_{i=1}^{n} x_i = k \text{ and } 1 \geq x_1, x_2, \ldots, x_n \geq 0 \right\}$$

and $f : \mathcal{K}_{\mathrm{LP}} \to \mathbb{R}$ is a Lipschitz function, that is, there is a $0 \leq L_{\max} < \infty$ such that:

$$|f(x) - f(y)| \leq L_{\max} \|x - y\|_{\max}, \text{ where } \|x - y\|_{\max} := \max_{i=1,\ldots n} |x_i - y_i| \tag{23}$$

is the max-norm.

We define $S(0) := \operatorname{argmin} \left\{ \sum_{i=1}^{n} c_i x_i : x \in \mathcal{K}_{\mathrm{LP}} \right\}$, that is, the preimage of the value $m_{\mathrm{LP}}(0)$ (the set of minimizers of the unperturbed problem), $S(\varepsilon) := \operatorname{argmin} \left\{ \sum_{i=1}^{n} c_i x_i + \varepsilon f(x) : x \in \mathcal{K}_{\mathrm{LP}} \right\}$, $S_f(0) := \operatorname{argmin}\{f(x) : x \in S(0)\}$ and $\bar{f} := \min\{f(x) : x \in S(0)\}$.

The authors [1] showed that for small enough ε there is no error by replacing the upper bound, $u_{\mathrm{LP}}(\varepsilon) := m_{\mathrm{LP}}(0) + \varepsilon \bar{f}$, by the exact minimum value, $m_{\mathrm{LP}}(\varepsilon)$, namely:

$$m_{\mathrm{LP}}(\varepsilon) \;=\; m_{\mathrm{LP}}(0) + \varepsilon \bar{f} \quad \text{for all } \; 0 \leq \varepsilon \leq \varepsilon_o \,, \tag{24}$$

where ε_o is strict positive and given by:

$$\varepsilon_o := \frac{\alpha_{\mathrm{LP}}}{L_{\max}} > 0 \,. \tag{25}$$

This means that $S_f(0) \subset S(\varepsilon)$ for all $0 \leq \varepsilon \leq \varepsilon_o$. In [1] α_{LP} can be taken as the largest positive constant that satisfies the inequality

$$\sum_{i=1}^{n} c_i x_i - m_{\mathrm{LP}}(0) \geq \alpha_{\mathrm{LP}} \min_{y \in S(0)} \|x - y\|_{\max} \tag{26}$$

for all $x \in \mathcal{K}_{\mathrm{LP}}$. It is interesting to compare (26) with (10) and (9). In fact, in [1] it is only proved that $\alpha_{\mathrm{LP}} > 0$. That is, in [1] there is no explicit formula for α_{LP} in terms of c.

Remark 4.1. In this section we are using the max-norm in (23) and (26) in order to follow [1]. We can assure the strict positiveness of ε_o for any choice of norm, since in $\mathbb{R}^n$ all norms are equivalent.

In order to illustrate the result (24)-(25) of [1], note that minimization problem of example 2 is as in (22). To see this, we identify $x_1 = X_{1,1}$ and $x_2 = X_{2,2}$. Hence,

$$m_{\mathrm{LP}}(\varepsilon) = \min_{\substack{x_1 + x_2 = 1 \\ x_1 \,, \, x_2 \geq 0}} 2x_2 + \varepsilon x_1^2 \,. \tag{27}$$

It is easy to show that for this example that $L_{\max} = 2$ and $\alpha_{\mathrm{LP}} = 2$. So, by (25) $\varepsilon_o = 1$. According to the exact solution, see (21), $\varepsilon_o = 1$ is the best (the largest) value for the equation (24) holds true.

Appendix A. Proof of proposition 2.3

In order to prove the proposition 2.3 we need first a result (see corollary A.2 bellow), which is a direct consequence of the following lemma:

Lemma A.1 (Cauchy-Schwarz-Bunjakowski Inequality for the entries of positive semidefinite matrices). *If $W \in M_n$ is a symmetric positive semidefinite matrix then its entries satisfy the following inequality: $\left|W_{i,j}\right|^2 \leq W_{i,i}\, W_{j,j}$ for all $i, j = 1, 2, \ldots, n$.*

Proof. see [4] page 398. □

Corollary A.2. *Consider $W \in M_n$ with $I \succeq W \succeq 0$. In this case we have the following implications:*

a) *If $W_{\ell,\ell} = 1$ for some $\ell \in \{1\,,\,\ldots\,,\,n\}$ then $W_{\ell,j} = 0$ for all $1 \leq j \leq n$ $(j \neq \ell)$ and $W_{i,\ell} = 0$ for all $1 \leq i \leq n$ $(i \neq \ell)$;*

b) *If $W_{\ell,\ell} = 0$ for some $\ell \in \{1\,,\,\ldots\,,\,n\}$ then $W_{\ell,j} = 0$ for all $1 \leq j \leq n$ and $W_{i,\ell} = 0$ for all $1 \leq i \leq n$;*

Proof. To prove a) we use that $I \succeq W$, so $I - W \succeq 0$. It follows by lemma A.1 that $\left|\delta_{i,j} - W_{i,j}\right|^2 \leq (1 - W_{i,i})(1 - W_{j,j})$. Hence, if $W_{\ell,\ell} = 1$ then $W_{\ell,j} = 0$ for $j \neq \ell$ and $W_{i,\ell} = 0$ for $i \neq \ell$. The proof of b) uses that $W \succeq 0$ and the lemma A.1. □

Proof of Proposition 2.3.

Proof. Since C is symmetric there is an orthonormal-basis of eigenvectors. We denote it by $\{v_i\}_{i=1}^n \subset \mathbb{R}^n$. Since this basis is orthonormal the trace $\mathrm{Tr}\{CX\}$ is expressed as

$$\mathrm{Tr}\{CX\} = \sum_{i=1}^{n} v_i^T C X v_i \;=\; \sum_{i=1}^{n} \lambda_i(C)\, v_i^T X v_i\,. \tag{28}$$

We define $W_{i,j}(V, X) := v_i^T X v_j$ for $i, j = 1, 2, \ldots, n$. That is, the matrix W is given by $W := V^T X V$. Recall that V is the matrix, whose columns are the eigenvectors v_i. Note that for all $X \in \mathcal{K}$ we have:

$$0 \leq W_{i,i}(V, X) \leq 1 \quad \text{and} \quad \sum_{i=1}^{n} W_{i,i}(V, X) = k, \tag{29}$$

because $I \succeq X \succeq 0$ and $\mathrm{Tr}\{X\} = k$ respectively. Combining (28) with (29) we obtain:

$$\min_{\substack{\mathrm{Tr}\{X\}=k:\\ I\succeq X\succeq 0}} \mathrm{Tr}\{CX\} \geq \min_{\mathbf{x}\in C} \sum_{i=1}^{n} \lambda_i(C)\, x_i\,, \tag{30}$$

where the set $C := \left\{\mathbf{x} = (x_1, x_2, \ldots, x_n) \in [0,1]^n \colon x_1 + x_2 + \cdots + x_n = k\right\}$ is convex and compact.

We claim that we have equality in (30). To see this, let $\mathbf{y} \in C$ be a minimizer of the *right hand side* (RHS) of (30). Further, take $X(\mathbf{y}) := VD(\mathbf{y})V^T$, where $D(\mathbf{y})$ is the diagonal matrix which the diagonal entries are the y_i's and V is the matrix which columns are the eigenvectors v_i. We will show that $X(\mathbf{y}) \in \mathcal{K}$ and $\mathrm{Tr}\{CX(\mathbf{y})\}$ is equal the minimum value of the RHS of (30). Indeed, $X(\mathbf{y}) \in \mathcal{K}$ since $\mathrm{Tr}\{VD(\mathbf{y})V^T\} = \mathrm{Tr}\{D(\mathbf{y})\} = \sum_{i=1}^{n} y_i = k$ and $I \succeq VD(\mathbf{y})V^T \succeq 0$ since $1 \geq y_i \geq 0$. Moreover, $\mathrm{Tr}\{CX(\mathbf{y})\} = \mathrm{Tr}\{CVD(\mathbf{y})V^T\} = \mathrm{Tr}\{V^TCVD(\mathbf{y})\}$ (the trace is cyclic). On the other hand, $V^TCV = D(\lambda)$ (spectral decomposition of C). Hence, $\mathrm{Tr}\{CX(\mathbf{y})\} = \mathrm{Tr}\{D(\lambda)\, D(\mathbf{y})\} = \sum_{i=1} \lambda_i(C) y_i$, which is the minimum value of the RHS of (30). Therefore,

$$\min_{\mathbf{x}\in C} \sum_{i=1}^{n} \lambda_i(C)\, x_i = \sum_{i=1} \lambda_i(C) y_i = \mathrm{Tr}\{CX(\mathbf{y})\} \geq \min_{\substack{\mathrm{Tr}\{X\}=k:\\ I\succeq X\succeq 0}} \mathrm{Tr}\{CX\}\,. \tag{31}$$

Comparing (30) with (31) we obtain the claim, that is,

$$\min_{\substack{\mathrm{Tr}\{X\}=k:\\ I\succeq X\succeq 0}} \mathrm{Tr}\{CX\} = \min_{\mathbf{x}\in C} \sum_{i=1}^{n} \lambda_i(C)\, x_i\,. \tag{32}$$

The fundamental theorem of linear optimization (FTLP, see [2]) states that

$$\min_{\mathbf{x}\in C} \sum_{i=1}^{n} \lambda_i(C)\, x_i = \min_{\mathbf{x}\in \mathrm{Ext}(C)} \sum_{i=1}^{n} \lambda_i(C)\, x_i\,. \tag{33}$$

Here $\mathrm{Ext}(C)$ is the set of the extreme points of C. It is well-known that

$$\mathrm{Ext}(C) = \left\{\mathbf{x} = (x_1, x_2, \ldots, x_n) \in \{0,1\}^n \mid x_1 + x_2 + \cdots + x_n = k\right\}. \tag{34}$$

Note that $\#\mathrm{Ext}(C) = \binom{n}{k}$ Hence:

$$\min_{\mathbf{x}\in \mathrm{Ext}(C)} \sum_{i=1}^{n} \lambda_i(C)\, x_i = \min_{1\leq i_1 < i_2 < \cdots < i_k \leq n} \lambda_{i_1}(C) + \lambda_{i_2}(C) + \cdots \lambda_{i_k}(C). \tag{35}$$

The minimization problem of the right side of (35) is easy to solve. More precisely:

$$\begin{aligned}\min_{1\le i_1<i_2<\cdots<i_k\le n} &\lambda_{i_1}(C)+\lambda_{i_2}(C)+\cdots\lambda_{i_k}(C)\\ &=\lambda_1(C)+\lambda_2(C)+\cdots+\lambda_k(C).\end{aligned} \tag{36}$$

In the last equality we used the fact that $\lambda_1(C) \le \lambda_2(C) \le \cdots \le \lambda_n(C)$ (increasing order).

Combining (32), (33), (35) and (36) follows that $\lambda_1(C)+\cdots+\lambda_k(C)$ is the minimum value of the problem (1).

Next, we will characterize the set of minimizers of the problem (1). We denote it by $\mathcal{X}_*(0)$.

According to the definitions of $r(C,k)$, $s(C,k)$ and $d(C,k)$ (see definition 2.2) the eigenvalues of C are ordered as

$$\lambda_1 \le \ldots \le \lambda_r < \lambda_{r+1} = \lambda_{r+2} = \cdots = \lambda_{r+d} < \lambda_s \le \cdots \le \lambda_n.$$

The FTLP states also that

$$\operatorname{argmin}\left\{\sum_{i=1}^{n}\lambda_i(C)x_i\colon \mathbf{x}\in C\right\} \;=\; \text{convex hull of } \operatorname{Ext}_\star(C)\,, \tag{37}$$

where

$$\operatorname{Ext}_\star(C) := \left\{\mathbf{x}=(x_1,x_2,\ldots,x_n)\in\{0,1\}^n\colon \begin{array}{c} x_1=x_2=\cdots=x_r=1,\\ x_s=x_{s+1}=\cdots=x_n=0\\ \text{and}\\ x_1+x_2+\cdots+x_n=k\end{array}\right\}. \tag{38}$$

Note that $\#\operatorname{Ext}_*(C) = \binom{d}{k-r}$. The set of the RHS of (37) is the smallest convex subset of C that contains $\operatorname{Ext}_*(C)$. Recalling (38) and the fact $n = r+d+s$, it follows that the convex hull of $\operatorname{Ext}_*(C)$ is clearly given by:

$$\left\{(1,\cdots,1,t_{r+1},t_{r+2},\cdots,t_{r+d},0,\cdots 0)\in[0,1]^n\colon \sum_{i=r+1}^{r+d}t_i = k-r\right\}. \tag{39}$$

We define the following three sets: 1) $\mathcal{K}_d := \{Z\in M_d\colon \operatorname{Tr}\{Z\}=k-r$ and $I\succeq Z\succeq 0\}$,

2)

$$\mathcal{K}_{r,d,s} := \begin{cases} \{I_r \oplus Z \oplus 0 \colon Z \in \mathcal{K}_d\} & \text{if } r \geq 1 \text{ and } s \leq n, \\ \{Z \oplus 0 \colon Z \in \mathcal{K}_d\} & \text{if } r = 0 \text{ and } s \leq n, \\ \{I_r \oplus Z \colon Z \in \mathcal{K}_d\} & \text{if } r \geq 1 \text{ and } s = n+1, \\ K_d & \text{if } r = 0 \text{ and } s = n+1, \end{cases}$$

and 3) $V\mathcal{K}_{r,d,s}V^T := \{V\,W\,V^T \colon W \in \mathcal{K}_{r,d,s}\}$. We mean by $I_r \oplus Z \oplus 0$ the $n \times n$ matrix in the block form:

$$\left[\begin{array}{c|c|c} I_r & 0 & 0 \\ \hline 0 & Z & 0 \\ \hline 0 & 0 & 0 \end{array}\right]$$

In this block, I_r is the $r \times r$ identity matrix, Z is a $d \times d$ matrix and, by 0 in the diagonal, we mean the $(n-s+1) \times (n-s+1)$ zero-matrix.

We claim that if X_* is a minimizer of problem (2) then the corresponding matrix $W_* = W(V, X_*) := V^T X_* V$ is an element of $\mathcal{K}_{r,d,s}$. To see this, note that if X_* is a minimizer then by (28) one has $\mathrm{Tr}\{CX_*\} = \sum_{i=1}^n \lambda_i(C) W_{*i,i}$ and by (37) and (39) the vector $\left(W_{*1,1}, W_{*2,2}, \ldots, W_{*n,n}\right)$ must satisfy:

$$\begin{aligned} &\text{a)}\ \ W_{*1,1} = W_{*2,2} = \cdots = W_{*r,r} = 1; \\ &\text{b)}\ \ W_{*s,s} = W_{*s+1,s+1} = \cdots = W_{*n,n} = 0; \\ &\text{c)}\ \ \textstyle\sum_{i=1}^n W_{*i,i} = k\,. \end{aligned} \tag{40}$$

Now, combining (40) a) and (40) b) with the corollary A.2 we conclude that:

a) For any $i = 1, 2, \ldots, r$ and $j = 1, 2, \ldots, n$ we have $W_{*i,j} = \delta_{i,j}$,

a') For any $j = 1, 2, \ldots, r$ and $i = 1, 2, \ldots, n$ we have $W_{*i,j} = \delta_{i,j}$,

b) For any $i = s, s+1, \ldots, n$ and $j = 1, 2, \ldots, n$ we have $W_{*i,j} = 0$ and

b') For any $j = s, s+1, \ldots, n$ and $i = 1, 2, \ldots, n$ we also have $W_{*i,j} = 0$.

This proves that the $n \times n$ matrix W_* is of the form

$$\left[\begin{array}{c|c|c} I_r & 0 & 0 \\ \hline 0 & \star & 0 \\ \hline 0 & 0 & 0 \end{array}\right].$$

Here $\star$ is a $d \times d$ (note that $d = s - r - 1$) submatrix which satisfies $I \succeq \star \succeq 0$ because $W_* = V^T X_* V$ and X_* satisfies also $I \succeq X \succeq 0$. Due to the (40) c) one has $\mathrm{Tr}\{\star\} = k$. Therefore $W_* \in \mathcal{K}_{r,d,s}$. Since $V^T = V^{-1}$ (because V is orthogonal) we obtain that $\mathcal{X}_*(0) \subset V\,\mathcal{K}_{r,d,s}\,V^T$.

On the other hand, the set $V\mathcal{K}_{r,d,s}V^T$ is a subset of $\mathcal{X}_*(0)$ because if $Z \in \mathcal{K}_d$ then

$$\begin{aligned}
&\mathrm{Tr}\left\{C\ V(I_r \oplus Z \oplus 0)V^T\right\} = \mathrm{Tr}\{\Lambda\ (I_r \oplus Z \oplus 0)\} \\
&\quad \text{since Tr is cyclic and } C = V\Lambda V^T \\
&= \sum_{i=1}^{r} \lambda_i\, 1 + \sum_{i=1}^{d} \lambda_{r+i}\, Z_{r+i,r+i} = \sum_{i=1}^{r} \lambda_i + \lambda_k \sum_{i=1}^{d} Z_{r+i,r+i} \\
&\quad \text{since } \ \lambda_{r+1} = \cdots \lambda_{r+d} = \lambda_k = \sum_{i=1}^{r} \lambda_i + (k-r)\lambda_k \\
&\quad \text{since } \mathrm{Tr}\{Z\} = k - r \ = \sum_{i=1}^{k} \lambda_i
\end{aligned}$$

Therefore $V\mathcal{K}_{r,d,s}V^T \subset \mathcal{X}_*(0)$.

In the particular case $d = 1$ ($\Rightarrow r + 1 = k$) we have $Z = [1]$. Hence, the only element of the set $\mathcal{X}_*(0)$ is $X = V(I_r \oplus [1] \oplus 0)V^T = V(I_{r+1} \oplus 0)V^T = V(I_k \oplus 0)V^T = \sum_{i=1}^{k} v_i v_i^T$. This is a orthogonal projection of rank k.

Recall that if the matrix C has degenerate eigenvalues then there are many choices for the orthogonal matrix V. We prove that the set of minimizers does not depend on the choice of V. Consider the eigenvalues of C without counting the multiplicity, that is, $\mu_1(C) < \mu_2(C) < \cdots < \mu_p(C)$. We denote by ν_1 the multiplicity of $\mu_1(C)$, by ν_2 is the multiplicity of $\mu_2(C)$ and so on. The corresponding eigenspaces $E_i \subset \mathbb{R}^n$, $i = 1, \ldots, p$ are pairwise orthogonal and $\dim E_i = \nu_i$. Note that $\sum_{i=1}^{p} \nu_i = n$. The many choices for V comes from the fact that there are many orthogonal basis for each eigenspace. However, two different orthogonal basis of an eigenspace are related to each other by an orthogonal matrix. Suppose that $\tilde{V}$ were other choice, then V and $\tilde{V}$ are related by $\tilde{V} = V(U_1 \oplus U_2 \oplus \cdots \oplus U_p)$ where U_i is a $\nu_i \times \nu_i$ orthogonal matrix. We claim that the solution sets $X_*(0) := \{V(I_r \oplus Z \oplus 0)V^T : Z \in \mathcal{K}_d\}$ and $\tilde{X}_*(0) := \{\tilde{V}(I_r \oplus Z \oplus 0)\tilde{V}^T : Z \in \mathcal{K}_d\}$ are equal. To see this, note that

$\tilde{V}(I_r \oplus Z \oplus 0)\tilde{V}^T = V(U_1 \oplus \cdots \oplus U_p)(I_r \oplus Z \oplus 0)(U_1 \oplus \cdots \oplus U_p)^T V^T = V(\tilde{I}_r \oplus \tilde{Z} \oplus 0)V^T$ where $\tilde{Z} = (U_1 \oplus \cdots \oplus U_p)Z(U_1 \oplus \cdots \oplus U_p)^T$ and $\tilde{I}_r = (U_1 \oplus \cdots \oplus U_p)I_r(U_1 \oplus \cdots \oplus U_p)^T$. But it easy to see that $\tilde{I}_r = I_r$. Therefore, $\tilde{V}(I_r \oplus Z \oplus 0)\tilde{V}^T = V(I_r \oplus \tilde{Z} \oplus 0)V^T$ and since $Z \in K_d \Leftrightarrow \tilde{Z} \in \mathcal{K}_d$ follows the claim. □

5 Proof of Lemma B.1

Lemma B.1. *Let $C \in M_n$ be a symmetric matrix. For each $k = 1, 2, \ldots, n$, the following strict positive constant $\alpha(C, k)$ defined by*

$$\alpha(C,k) := \frac{1}{2}\ \min\left\{\frac{\lambda_k(C) - \lambda_r(C)}{s-(k+1)}\ ,\ \frac{\lambda_s(C) - \lambda_k(C)}{k}\right\} \tag{41}$$

satisfies the inequality

$$\mathrm{Tr}\{CX\} - \sum_{i=1}^{k}\lambda_i(C)\ \geq\ \alpha(C,k)\ \min_{Y\in\mathcal{X}_*(0)}\left\{\|X-Y\|^2\right\} \tag{42}$$

for all $X \in \mathcal{K}$. In (41) *the values $r = r(C,k)$ and $s = s(C,k)$ are given by definition* 2.2. *Moreover, we do the convention that $\lambda_0(C) := -\infty$ in the case $r = 0$ and $\lambda_{n+1}(C) = +\infty$ in the case $s = n+1$. In* (42) *we denote by $\mathcal{X}_*(0)$ the set* $\mathrm{argmin}\{\mathrm{Tr}\{CX\}\colon X \in \mathcal{K}\}$.

The proof of the lemma B.1 is based on the three propositions below. More precisely, the propositions B.2, B.3 and B.6.

Note that the matrix function $\mathrm{Tr}\{\cdot\}$ (consequently also $\|\cdot\|^2$) is invariant by conjugation of an orthogonal matrix, that is, $\mathrm{Tr}\{SAS^T\} = \mathrm{Tr}\{A\}$ for all $A \in M_n$ and S orthogonal. Due to this fact, we can reduce the proof of (42) to the case that C is diagonal. More precisely, we have:

Proposition B.2. *Let $C = V\Lambda V^T$ be a spectral decomposition of the symmetric matrix C, with diagonal matrix $\Lambda := \mathrm{diag}(\lambda_1(C), \ldots, \lambda_n(C))$ and a corresponding orthogonal eigenvector matrix V. If $X \in \mathcal{K}$ then for $W = W(X) := V^T X V$ holds:*

$$\mathrm{Tr}\{CX\} - \sum_{i=1}^{k}\lambda_i(C)\ =\ \mathrm{Tr}\{\Lambda W\} - \sum_{i=1}^{k}\lambda_i(C) \tag{43}$$

and

$$\min_{Y\in\mathcal{X}_*(0)}\left\{\|X-Y\|^2\right\} \leq \min_{Z\in\mathcal{K}_d}\left\{\|W-I_r\oplus Z\oplus 0\|^2\right\}. \tag{44}$$

Here $\mathcal{K}_d := \left\{Z\in M_d\colon \mathrm{Tr}\{Z\}=k-r \text{ and } I\succeq Z\succeq 0\right\}$.

Proof. Since the trace is orthogonal-invariant we have

$$\mathrm{Tr}\{CX\} = \mathrm{Tr}\left\{V\Lambda V^T V W V^T\right\} = \mathrm{Tr}\{\Lambda W\}.$$

This proves (43). To prove (44) take $Z_{**}\in \mathrm{argmin}\{\|W-I_r\oplus Z\oplus 0\|^2\colon Z\in\mathcal{K}_d\}$. Due to the proposition 2.3 $Y_{**} := V(I_r\oplus Z_{**}\oplus 0)V^T \in \mathcal{X}_*(0)$. From the orthogonal invariance of $\|\cdot\|^2$ follows that the

$$\begin{aligned}\min_{Y\in\mathcal{X}_*(0)}\|X-Y\|^2 &\leq \|X-Y_{**}\|^2 = \|VWV^T - V(I_r\oplus Z_{**}\oplus 0)V^T\|^2\\ &= \|W-I_r\oplus Z_{**}\oplus 0\|^2 = \min_{Z\in\mathcal{K}_d}\|W-I_r\oplus Z\oplus 0\|^2.\end{aligned}$$

This proves (44). □

In order to prove (42) we establish a lower bound for $\mathrm{Tr}\{\Lambda W\}-\sum_{i=1}\lambda_i(C)$ (see proposition B.3) and an upper for $\min_{Z\in\mathcal{K}_d}\left\{\|W-I_r\oplus Z\oplus 0\|^2\right\}$ (see proposition B.6). Namely:

Proposition B.3. *Let be* $n\geq 2$. *For all* $W\in\mathcal{K}$ *we have*

$$\begin{aligned}\mathrm{Tr}\{\Lambda W\}-\sum_{i=1}^{k}\lambda_i(C) \geq \Big(\lambda_k(C)-\lambda_r(C)\Big)\left(r-\sum_{i=1}^{r}W_{i,i}\right)\\ +\Big(\lambda_s(C)-\lambda_k(C)\Big)\sum_{i=s}^{n}W_{i,i}.\end{aligned} \tag{45}$$

In the cases $r=0$ *and* $s=n+1$ *we have*

$$\mathrm{Tr}\{\Lambda W\}-\sum_{i=1}^{k}\lambda_i(C) \geq \Big(\lambda_s(C)-\lambda_k(C)\Big)\sum_{i=s}^{n}W_{i,i}. \tag{46}$$

and

$$\mathrm{Tr}\{\Lambda W\}-\sum_{i=1}^{k}\lambda_i(C) \geq \Big(\lambda_k(C)-\lambda_r(C)\Big)\left(r-\sum_{i=1}^{r}W_{i,i}\right) \tag{47}$$

respectively.

Proof. Recall that $W \in \mathcal{K}$ means that $\mathrm{Tr}\{W\} = k$ and $I \geq W \geq 0$. Consequently $W_{i,i} - 1 \leq 0$, $W_{i,i} \geq 0$ and $\sum_{i=1}^{n} W_{i,i} = k$. In order to simplify the notation we identify $W_{i,i}$ with w_i and $\lambda_i(C)$ with λ_i. Note that

$$\mathrm{Tr}\{\Lambda W\} - \sum_{i=1}^{k} \lambda_i = \sum_{i=1}^{n} \lambda_i w_i - \sum_{i=1}^{k} \lambda_i(C)\, 1 = \sum_{i=1}^{k} \lambda_i (w_i - 1) + \sum_{i=k+1}^{n} \lambda_i w_i$$

$$= \sum_{i=1}^{r} \lambda_i (w_i - 1) + \sum_{i=r+1}^{k} \lambda_i (w_i - 1) + \sum_{i=k+1}^{s-1} \lambda_i w_i + \sum_{i=s}^{n} \lambda_i w_i$$

From the following observations:

a) Combining $w_i - 1 \leq 0$ for $i = 1, \ldots, k$ with $\lambda_r \geq \lambda_i$ for $i = 1, \ldots, r$ and $\lambda_k \leq \lambda_j$ for $j = r+1, \ldots, k$;

b) Combining $w_i \geq 0$ for $i = s, \ldots, n$ with $\lambda_s \leq \lambda_i$ for $i = s, \ldots, n$ and $\lambda_k \geq \lambda_j$ for $j = r+1, \ldots, k$;

c) $\lambda_k = \lambda_{r+i}$ for $i = 1, \ldots, d$ and $\lambda_k = \lambda_{k+j}$ for $j = 1, \ldots, s-1$;

d) $\sum_{i=r+1}^{k} w_i + \sum_{i=k+1}^{s-1} w_i = k - \sum_{i=1}^{r} w_i - \sum_{i=s}^{n} w_i$ $\left(\text{since } \sum_{i=1}^{n} w_i = k\right)$, we conclude that

$$\mathrm{Tr}\{\Lambda W\} - \sum_{i=1}^{k} \lambda_i \geq \lambda_r \Big(-r + \sum_{i=1}^{r} w_i \Big)$$

$$+ \lambda_k \Big(-(k-r) + \sum_{i=r+1}^{k} w_i \Big) + \lambda_{k+1} \sum_{i=k+1}^{s-1} w_i + \lambda_s \sum_{i=s}^{n} w_i$$

$$= -\lambda_r \Big(r - \sum_{i=1}^{r} w_i \Big) + \lambda_k \Big(r - k + \sum_{i=r+1}^{k} w_i + \sum_{i=k+1}^{s-1} w_i \Big) + \lambda_s \sum_{i=s}^{n} w_i$$

$$= -\lambda_r \Big(r - \sum_{i=1}^{r} w_i \Big) + \lambda_k \Big(r - \sum_{i=1}^{r} w_i - \sum_{i=s}^{n} w_i \Big) + \lambda_s \sum_{i=s}^{n} w_i$$

$$= \Big(\lambda_k - \lambda_r \Big) \Big(r - \sum_{i=1}^{r} w_i \Big) + \Big(\lambda_s - \lambda_k \Big) \sum_{s=1}^{n} w_i,$$

which proves the proposition B.3. □

In order to establish an upper bound for $\min_{Z \in \mathcal{K}_d} \|W - I_r \oplus Z \oplus 0\|^2$ we need first the following two lemmas:

Lemma B.4. *For any positive semidefinite matrix $A \in M_p$ the inequality*

$$\|A\|^2 := \mathrm{Tr}\{A^2\} \leq \big(\mathrm{Tr}\{A\}\big)^2 =: \mathrm{Tr}^2\{A\} \tag{48}$$

holds.

Proof. To prove (48) note that if $A \succeq 0$ then $0 \leq \lambda_i(A)$. Hence,

$$\begin{aligned} \|A\|^2 &= \sum_{i=1}^{p} \lambda_i^2(A) \leq \sum_{i=1}^{p} \lambda_i^2(A) + 2\sum_{i<j}^{p} \lambda_i(A)\,\lambda_j(A) \\ &= \left(\sum_{i}^{p} \lambda_i(A)\right)^2 = \mathrm{Tr}^2\{A\}. \end{aligned} \tag{49}$$

□

Lemma B.5. *For any symmetric matrix $A \in M_d$ with $I \succeq A \succeq 0$ we have the following inequality*

$$\begin{aligned} \min\big\{\|A - Z\|^2 \colon Z \in M_d \ \text{ with } \ \mathrm{Tr}\{Z\} = k - r \ \text{ and } \ I \succeq Z \succeq 0\big\} \\ \leq \big((k-r) - \mathrm{Tr}\{A\}\big)^2 \end{aligned} \tag{50}$$

Proof. **Case 1:** $A \neq 0$. Since $A \succeq 0$ we have in this case that $\mathrm{Tr}\{A\} > 0$. So we can take $Z_o := \frac{k-r}{\mathrm{Tr}\{A\}}\, A$ as a trial element of the domain of the minimization problem (50). Note that Z_o satisfies $I \succeq Z \succeq 0$, since A also does, and satisfies $\mathrm{Tr}\{Z_o\} = k - r$. Hence, the Left Hand Side, LHS, of (50) has the following upper bound:

$$\begin{aligned} \mathrm{LHS} \leq \|A - Z_o\|^2 &= \left(1 - \frac{k-r}{\mathrm{Tr}\{A\}}\right)^2 \|A\|^2 \\ &= \big((k-r) - \mathrm{Tr}\{A\}\big)^2 \frac{\|A\|^2}{\mathrm{Tr}^2\{A\}}. \end{aligned} \tag{51}$$

On the other hand, since $A \succeq 0$ we have by (48) that $\|A\|^2/\mathrm{Tr}^2\{A\} \leq 1$. This closes the proof in the first case.

Case 2: $A = 0$. In this case $\mathrm{Tr}\{A\} = 0$. Using (48) again, we have $\|Z\|^2 \leq \mathrm{Tr}^2\{Z\}$. Moreover, $\mathrm{Tr}^2\{Z\} = (k-r)^2 = \big((k-r) - 0\big)^2 = \big((k-r) - \mathrm{Tr}\{A\}\big)^2$. Hence, LHS $\leq \big((k-r) - \mathrm{Tr}\{A\}\big)^2$. This proves the lemma B.5. □

Proposition B.6. *The inequality*

$$\begin{aligned}&\min_{Z\in\mathcal{K}_d} \|W - I_r \oplus Z \oplus 0\|^2 \\ &\leq 2\left\{\big(s-(k+1)\big)\left(r-\sum_{i=1}^{r} W_{i,i}\right)+k\left(\sum_{i=s}^{n} W_{i,i}\right)\right\}\end{aligned} \tag{52}$$

holds for all $W \in \mathcal{K}$. *In the cases* $r = 0$ *and* $s = n+1$ *the inequality* (52) *becomes*

$$\min_{Z\in\mathcal{K}_d} \|W - Z \oplus 0\|^2 \leq 2k\left(\sum_{i=s}^{n} W_{i,i}\right) \tag{53}$$

and

$$\min_{Z\in\mathcal{K}_d} \|W - I_r \oplus Z\|^2 \leq 2\big(s-(k+1)\big)\left(r-\sum_{i=1}^{r} W_{i,i}\right) \tag{54}$$

respectively.

Proof. We write $W \in \mathcal{K}$ in the block form:

$$W = \begin{bmatrix} R(W) & E(W) & F(W) \\ E^T(W) & A(W) & G(W) \\ F^T(W) & G^T(W) & S(W) \end{bmatrix} \tag{55}$$

Here $R(W) \in M_r$, $A(W) \in M_d$, $S(W) \in M_{n-(d+r)}$, $E(W) \in M_{r\times d}$, $F(W) \in M_{r\times(n-(r+d))}$ and $G(W) \in M_{d\times(n-(r+d))}$. In the particular case $r = 0$ it is understood that the blocks $R(W)$, $E(W)$, $E^T(W)$, $F(W)$ and $F^T(W)$ do not appear in (55). Similarly, if $s = n+1$, that is, $r+d = n$, then the blocks $S(W)$, $F(W)$, $F^T(W)$, $G(W)$and $G^T(W)$ does not appear in (55). The square of the Frobenius norm of the matrix $I_r \oplus Z \oplus 0 - W$ splits as

$$\begin{aligned}\|I_r \oplus Z \oplus 0 - W\|^2 \;=\; &\|I_r - R(W)\|^2 + \|Z - A(W)\|^2 + \|S(W)\|^2 \\ &+ 2\,\|E(W)\|^2 + 2\,\|F(W)\|^2 + 2\,\|G(W)\|^2\end{aligned} \tag{56}$$

Since $I \geq W$ we also have that $I_r \geq R(W)$, that is, $I_r - R(W) \geq 0$. Using (48) we have:

$$\|I_r - R(W)\|^2 \leq \big(r - \mathrm{Tr}\{R(W)\}\big)^2 \tag{57}$$

Since $I \succeq W \succeq 0$ we also have $I \succeq A(W) \succeq 0$. Hence, from lemma B.5 follows that

$$\min_{Z \in \mathcal{K}_d} \|A(W) - Z\|^2 \leq \big((k-r) - \mathrm{Tr}\{A(W)\} \big)^2 , \tag{58}$$

since $\mathrm{Tr}\{Z\} = k - r$.

Recall that $k = \mathrm{Tr}\{W\} = \mathrm{Tr}\{R(W)\} + \mathrm{Tr}\{A(W)\} + \mathrm{Tr}\{S(W)\}$, consequently

$$\mathrm{Tr}\{A\} = k - \mathrm{Tr}\{R\} - \mathrm{Tr}\{S\} . \tag{59}$$

Combining (58) with (59) we have

$$\min_{Z \in \mathcal{K}_d} \|A(W) - Z\|^2 \leq \big(\mathrm{Tr}\{R(W)\} + \mathrm{Tr}\{S(W)\} - r \big)^2 . \tag{60}$$

On the other hand, since $W \succeq 0$ we also have that $S(W) \succeq 0$. Combining this with (48) we have:

$$\|S\|^2 \leq \mathrm{Tr}^2\{S\}. \tag{61}$$

Now we claim that the following three inequalities hold:

$$\|E(W)\|^2 \leq \big(r - \mathrm{Tr}\{R(W)\} \big) \big(\mathrm{Tr}\{R(W)\} + \mathrm{Tr}\{S(W)\} + s - (r+k+1) \big) , \tag{62}$$

$$\|F(W)\|^2 \leq \mathrm{Tr}\{R(W)\}\, \mathrm{Tr}\{S(W)\} , \tag{63}$$

and

$$\|G(W\|^2 \leq \big(k - \mathrm{Tr}\{R(W)\} - \mathrm{Tr}\{S(W)\} \big)\, \mathrm{Tr}\{S(W)\} . \tag{64}$$

To see (62), we recall $I - W \succeq 0$ and use the lemma A.1. Hence,

$$\begin{aligned}
\|E(W)\|^2 &= \sum_{i=1}^{r} \sum_{j=r+1}^{r+d} |W_{i,j}|^2 = \sum_{i=1}^{r} \sum_{j=r+1}^{r+d} \left| \delta_{i,j} - W_{i,j} \right|^2 \\
&\leq \sum_{i=1}^{r} \sum_{j=r+1}^{r+d} (\delta_{i,i} - W_{i,i})\, (\delta_{j,j} - W_{j,j}) \\
&= \left(\sum_{i=1}^{r} (1 - W_{i,i}) \right) \left(\sum_{j=r+1}^{r+d} (1 - W_{j,j}) \right) \\
&= \big(r - \mathrm{Tr}\{R(W)\} \big) \big(d - \mathrm{Tr}\{A(W)\} \big)
\end{aligned}$$

Combing this with (59) and $d = s - (r+1)$ we prove (62).

To see (63), we recall that $W \succeq 0$ use the lemma A.1. Hence,

$$\begin{aligned}\|F(W)\|^2 &= \sum_{i=1}^{r}\sum_{j=s}^{n}|W_{i,j}|^2 \leq \sum_{i=1}^{r}\sum_{j=s}^{n} W_{i,i}\, W_{j,j} \\ &= \left(\sum_{i=1}^{r} W_{i,i}\right)\left(\sum_{j=s}^{n} W_{j,j}\right) = \operatorname{Tr}\{R(W)\}\operatorname{Tr}\{S(W)\}\,.\end{aligned}$$

The proof of (64) is analog to (63).

Now plugging (57), (60), (61), (62), (63) and (64) into (56) we obtain:

$$\begin{aligned}\min_{Z\in\mathcal{K}_d}\|W - I_r \oplus Z \oplus 0\|^2 &\leq \big(r - \operatorname{Tr}\{R\}\big)^2 + \big(\operatorname{Tr}\{R\} + \operatorname{Tr}\{S\} - r\big)^2 + \operatorname{Tr}^2\{S\} \\ &\quad + 2\big(r - \operatorname{Tr}\{R\}\big)\big(\operatorname{Tr}\{R\} + \operatorname{Tr}\{S\} + s - (r+k+1)\big) \\ &\quad + 2\operatorname{Tr}\{R\}\operatorname{Tr}\{S\} + 2\big(k - \operatorname{Tr}\{R\} - \operatorname{Tr}\{S\}\big)\operatorname{Tr}\{S\} \\ &= \big(r - \operatorname{Tr}\{R\}\big)^2 + \big(r - \operatorname{Tr}\{R\}\big)^2 - 2\operatorname{Tr}\{S\}\big(r - \operatorname{Tr}\{R\}\big) + 2\operatorname{Tr}^2\{S\} \\ &\quad - 2\big(r - \operatorname{Tr}\{R\}\big)^2 + 2\big(r - \operatorname{Tr}\{R\}\big)\big(\operatorname{Tr}\{S\} + s - (k+1)\big) \\ &\quad + 2\operatorname{Tr}\{R\}\operatorname{Tr}\{S\} + 2k\operatorname{Tr}\{S\} - 2\operatorname{Tr}\{R\}\operatorname{Tr}\{S\} - 2\operatorname{Tr}^2\{S\} \\ &= 2\big(r - \operatorname{Tr}\{R\}\big)\big(s - (k+1)\big) + 2k\operatorname{Tr}\{S\}\,,\end{aligned}$$

which proves the proposition B.6. □

Proof of Lemma B.1

Proof. Recall that the eigenvalues of C are ordered as $\lambda_1 \leq \cdots \leq \lambda_r < \lambda_{r+1} = \cdots = \lambda_k = \cdots = \lambda_{r+k} < \lambda_s \leq \cdots \leq \lambda_n$. According to the values of $r(C,k)$ and $s(C,k)$ we divide the proof in the following four cases:

Case 1: $r > 0$ and $s < n+1$. In order to treat this case, we need first the following lemma, which proof is trivial.

Lemma B.7. *Let $a > 0$, $b > 0$, $c \geq 0$ and $d > 0$ be constants. For all $x, y \geq 0$ we have the inequality*

$$a\,x + b\,y \;\geq\; \min\{ac^{-1},\, bd^{-1}\}\;(c\,x + d\,y\,). \tag{65}$$

In the case $c = 0$ *is to be understood that* $\min\{ac^{-1}\,,\ bd^{-1}\} = bd^{-1}$.

From (43) and (45) we obtain:

$$\begin{aligned}\mathrm{Tr}\{CX\} - \sum_{i=1}^{k}\lambda_i(C) \ \geq \big(\lambda_k(C) - \lambda_r(C)\big)\big(r - \mathrm{Tr}\{R(W)\}\big) \\ + \big(\lambda_s(C) - \lambda_k(C)\big)\mathrm{Tr}\{S(W)\}\end{aligned} \tag{66}$$

Now we use the lemma B.7 with $x := r - \mathrm{Tr}\{R(W)\}$, $y := \mathrm{Tr}\{S(W)\}$, $a := \lambda_k(C) - \lambda_r(C)$, $b := \lambda_s(C) - \lambda_k(C)$, $c := 2(s - (k+1))$ and $d := 2k$. Note that $\mathrm{Tr}\{R(W)\} \leq r$ since $R(W) \in M_r$ and $I \succeq R(W) \succeq 0$. Hence, $x \geq 0$. Combining the lemma B.7 with (66) we obtain:

$$\mathrm{Tr}\{CX\} - \sum_{i=1}^{k}\lambda_i(C) \geq \min\left\{a\,c^{-1}\,,\ b\,d^{-1}\right\}$$
$$\times\, 2\left[(s-(k+1))(r - \mathrm{Tr}\{R(W)\}) + k\mathrm{Tr}\{S(W)\}\right]$$

Now using (52) and (44) we have:

$$\begin{aligned}\mathrm{Tr}\{CX\} - \sum_{i=1}^{k}\lambda_i(C) &\geq \min\{a\,c^{-1}\,,\ b\,d^{-1}\}\ \min_{Z\in\mathcal{K}_d}\|W - I_r \oplus Z \oplus 0\|^2 \\ &\geq \min\{a\,c^{-1}\,,\ b\,d^{-1}\}\ \min_{Y\in\mathcal{X}_*(0)}\|X - Y\|^2\end{aligned}$$

This means that we can take $\alpha(C,k) := 1/2\ \min\left\{\frac{\lambda_k-\lambda_r}{s-(k+1)}, \frac{\lambda_s-\lambda_k}{k}\right\}$.

Case 2: $r = 0$ and $s < n+1$. This means that the block $R(W)$ does not appear. In this case we use (43), (46), (46) and (44) to obtain:

$$\begin{aligned}\mathrm{Tr}\{CX\} - \sum_{i=1}^{k}\lambda_i(C) &= \mathrm{Tr}\{\Lambda W\} - \sum_{i=1}^{k}\lambda_i(C) \\ &\geq \Big(\lambda_s(C) - \lambda_k(C)\Big)\mathrm{Tr}\{S(W)\} \\ &= \frac{1}{2k}\Big(\lambda_s(C) - \lambda_k(C)\Big)\,2k\,\mathrm{Tr}\{S(W)\} \\ &\geq \frac{1}{2k}\Big(\lambda_s(C) - \lambda_k(C)\Big)\ \min_{Z\in\mathcal{K}_d}\|W - Z \oplus 0\|^2 \\ &\geq \frac{1}{2k}\Big(\lambda_s(C) - \lambda_k(C)\Big)\ \min_{Y\in\mathcal{X}_*(0)}\|X - Y\|^2\,.\end{aligned}$$

This means that we can take $\alpha(C,k) := \frac{\lambda_s - \lambda_k}{2k}$. Note that this α satisfies

$$\alpha = \min\left\{\frac{\lambda_k - \lambda_r}{2(s-(k+1))}\,,\ \frac{\lambda_s - \lambda_k}{2k}\right\},$$

since $\frac{\lambda_k - \lambda_r}{2(s-(k+1))} = +\infty$ and $\frac{\lambda_s - \lambda_k}{2k} < +\infty$. This closes the proof in the second case.

Case 3: $s = n+1$ and $r > 0$. This means that the block $S(W)$ does not appear. We have two subcases:

Subcase 3.1: $k = n$. Note that if $X \in M_n$ with $I \geq X \geq 0$ and $\mathrm{Tr}\{X\} = n$ then $X = I_n$. Hence, $\mathcal{K} = \{I_n\}$, and so α can be take as $+\infty$ because

$$\mathrm{Tr}\{CX\} - \sum_i^n \lambda_i(C) = \mathrm{Tr}\{C\} - \sum_{i=1}^n \lambda_i(C) = 0$$

for all $X \in \mathcal{K}$ and $\min_{Y\in\mathcal{X}_*(0)} \|X-Y\|^2 = 0$ since $\mathcal{X}_*(0) = \mathcal{K} = \{I\}$. Note that this α satisfies

$$\alpha = \min\left\{\frac{\lambda_k - \lambda_r}{2(s-(k+1))}\,,\ \frac{\lambda_s - \lambda_k}{2k}\right\}.$$

To see this, note that $s = n+1 = k+1$ and $\lambda_s = +\infty$. Hence, both $\frac{\lambda_k - \lambda_r}{2(s-(k+1))}$ and $\frac{\lambda_s - \lambda_k}{2k}$ are infinity.

Subcase 3.2: $k < n$. Since $s = n+1$ we obtain $s-(k+1) > 0$. In this subcase we use (43), (47), (54) and (44) to obtain:

$$\begin{aligned}
\mathrm{Tr}\{CX\} - \sum_{i=1}^k \lambda_i(C) &= \mathrm{Tr}\{\Lambda W\} - \sum_{i=1}^k \lambda_i(C) \\
&\geq \big(\lambda_k(C) - \lambda_r(C)\big)\big(r - \mathrm{Tr}\{R(W)\}\big) \\
&= \frac{1}{2(s-(k+1))}\big(\lambda_k(C) - \lambda_r(C)\big)\,2\,(s-(k+1))\,\big(r - \mathrm{Tr}\{R(W)\}\big) \\
&\geq \frac{1}{2(s-(k+1))}\big(\lambda_k(C) - \lambda_r(C)\big)\min_{Z\in\mathcal{K}_d} \|W - I_r \oplus Z\|^2 \\
&\geq \frac{1}{2(s-(k+1))}\big(\lambda_k(C) - \lambda_r(C)\big)\min_{Y\in\mathcal{X}_*(0)} \|X-Y\|^2 .
\end{aligned}$$

This means that we can take $\alpha(C,k) := \frac{\lambda_k-\lambda_r}{2(s-(k+1))} = \frac{\lambda_k-\lambda_r}{2(n-k)}$. Note that this α satisfies

$$\alpha = \min\left\{\frac{\lambda_k-\lambda_r}{2(s-(k+1))}, \frac{\lambda_s-\lambda_k}{2k}\right\}$$

since $\lambda_s = +\infty$ and $\frac{\lambda_k-\lambda_r}{2(s-(k+1))} < +\infty$.

Case 4: $r = 0$ and $s = n+1$. This means, that $\lambda_1(C) = \cdots = \lambda_n(C)$, that is, $C = \lambda_1(C)\, I$ (a multiple of the identity). Hence, $\mathcal{X}_*(0) = \mathcal{K}$ because $\mathrm{Tr}\{CX\} - \sum_{i=1}^{k}\lambda_i(C) = \lambda_1 \mathrm{Tr}\{X\} - \lambda_1 k = \lambda_1 k - k\lambda_1(C) = 0$ for any $X \in \mathcal{K}$. Therefore, $\min_{Y\in\mathcal{X}_*(0)} \|X-Y\|^2 = 0$ for all $X \in \mathcal{K}$. This means, that we can take $\alpha = +\infty$. Hence, α satisfies

$$\alpha = \min\left\{\frac{\lambda_k-\lambda_r}{s-(k+1)}, \frac{\lambda_s-\lambda_k}{2k}\right\}$$

since $\lambda_r = -\infty$ and $\lambda_s = +\infty$. □

6 The matrix minimization problem as a counterexample

For perturbed linear programming (LP) the authors of [1] proved that there exists an $\varepsilon_o > 0$ such that $m_{\mathrm{LP}}(\varepsilon) = m_{\mathrm{LP}}(0) + \varepsilon\bar{f}$ for all $0 \le \varepsilon \le \varepsilon_o$. On the contrary, we prove that in the example 1 of matrix minimization problem (2) we have $m(0) + \varepsilon\bar{f} > m(\varepsilon)$ for all $\varepsilon > 0$. This is proved in proposition C.2. To prove this proposition we need first the following result:

Proposition C.1. *Consider the case $n = 2$ and $k = 1$. For the matrix $C = \left[\begin{smallmatrix}1 & 1\\ 1 & 1\end{smallmatrix}\right]$ and the nonlinear function $f(X) = X_{1,1}X_{1,1}$ the minimum value of $\mathrm{Tr}\{CX\} + \varepsilon f(X)$ on $\mathcal{K}$ can be expressed as:*

$$m(\varepsilon) = \max_{r\in[0,1]}\left\{1 + \varepsilon r - \sqrt{(\varepsilon r)^2+1} - \varepsilon r^2\right\}.$$

Proof. Note that $X \in \mathcal{K}$ implies $X_{1,1}, X_{2,2} \ge 0$ and $X_{1,1} + X_{2,2} = 1$. Therefore, if $X \in \mathcal{K}$ then $X_{11} \in [0,1]$. For $X \in \mathcal{K}$ we can express the nonlinear part of the objective function, $X_{1,1}X_{1,1}$, as:

$$X_{1,1}\,X_{1,1} = \left(2r\,X_{1,1} - r^2\right)\Big|_{r=X_{1,1}} = \max_{r\in[0,1]}\left\{2r\,X_{1,1} - r^2\right\}. \qquad (67)$$

With help of the identity (67) we obtain a linear expression in the variable $X \in \mathcal{K}$ (up to a maximization problem) for $\varepsilon f(X)$, namely:

$$\varepsilon\, f(X) := \varepsilon\, X_{1,1}\, X_{1,1} \;=\; \max_{r\in[0,1]} \left\{ \mathrm{Tr}\left\{ \begin{bmatrix} 2\varepsilon r & 0 \\ 0 & 0 \end{bmatrix} X \right\} - \varepsilon\, r^2 \right\}. \tag{68}$$

Now, we rewrite the minimum value $m(\varepsilon)$ as

$$\begin{aligned} m(\varepsilon) \;&:=\; \min_{X\in\mathcal{K}} \{\mathrm{Tr}\{CX\} + \varepsilon f(X)\} \\ &=\; \min_{X\in\mathcal{K}} \left\{ \max_{r\in[0,1]} \left\{ \mathrm{Tr}\left\{ \begin{bmatrix} 1+2r\varepsilon & 1 \\ 1 & 1 \end{bmatrix} X \right\} - \varepsilon\, r^2 \right\} \right\}. \end{aligned}$$

This means that $m(\varepsilon)$ is a minimax value of the two-variable function $g(X, r) := \mathrm{Tr}\left\{ \begin{bmatrix} 1+2r\varepsilon & 1 \\ 1 & 1 \end{bmatrix} X \right\} - \varepsilon\, r^2$. Note that the domains $\mathcal{K}$ and $[0, 1]$ are convex, the function g is convex in the variable X and concave in the variable r. Under these conditions we can change the order of minimization and maximization (see Corollary 37.3.2 of [3]). Therefore,

$$m(\varepsilon) = \max_{r\in[0,1]} \left\{ \min_{X\in\mathcal{K}} \left\{ \mathrm{Tr}\left\{ \begin{bmatrix} 1+2r\varepsilon & 1 \\ 1 & 1 \end{bmatrix} X \right\} - \varepsilon\, r^2 \right\} \right\}. \tag{69}$$

We use the proposition 2.3 to solve the minimization problem (in the variable X) of the above equation. The result is an expression involving the first eigenvalue of the matrix $\begin{bmatrix} 1+2r\varepsilon & 1 \\ 1 & 1 \end{bmatrix}$, namely:

$$m(\varepsilon) = \max_{r\in[0,1]} \left\{ \lambda_1 \left(\begin{bmatrix} 1+2r\varepsilon & 1 \\ 1 & 1 \end{bmatrix} \right) - \varepsilon\, r^2 \right\}. \tag{70}$$

On the other hand, it is easy to compute $\lambda_1 \left(\begin{bmatrix} 1+2r\varepsilon & 1 \\ 1 & 1 \end{bmatrix} \right)$, which is equal

$$1 + \varepsilon r - \sqrt{(\varepsilon r)^2 + 1}.$$

This proves the proposition C.1. □

Proposition C.2. *Consider the case $n = 2$ and $k = 1$. For the matrix $C = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$ and the nonlinear function $f(X) = X_{1,1}\, X_{1,1}$ the minimum value of* (2) *satisfies*

$$m(\varepsilon) < m(0) + \varepsilon\, \bar{f}$$

for all $\varepsilon > 0$.

Proof. Recall that $\bar{f} := \min_{X \in \mathcal{X}_*(0)} X_{1,1}\, X_{1,1}$. Here $\mathcal{X}_*(0) = \{v_1 v_1^T\}$ with $v_1^T = \frac{1}{\sqrt{2}}(1, -1)$, hence $\bar{f} = 1/4$. Moreover, $m(0) = \lambda_1(C) = 0$.

By proposition C.1, we only need to show that for all $\varepsilon > 0$ the strict inequality

$$m(\varepsilon) \;=\; \max_{r \in [0,1]} \big\{ a(r, \varepsilon) + b(r, \varepsilon) \big\} \;<\; \frac{1}{4}\,\varepsilon \tag{71}$$

holds. Here $a(r, \varepsilon) := 1 + \varepsilon r - \varepsilon r^2$ and $b(r, \varepsilon) := -\sqrt{1 + (\varepsilon r)^2}$. In order to prove (71) consider two cases:

Case 1: $0 \in \operatorname{argmax}\{a(r, \varepsilon) + b(r, \varepsilon) \colon r \in [0, 1]\}$. In this case, $m(\varepsilon) = a(0, \varepsilon) + b(0, \varepsilon) = 1 + (-1) = 0$ and so $0 < \frac{1}{4}\varepsilon$, since $\varepsilon > 0$. This proves (71) in the first case.

Case 2: there is an $r_*(\varepsilon) \in \operatorname{argmax}\{a(r, \varepsilon) + b(r, \varepsilon) \colon r \in [0, 1]\}$ with $r_*(\varepsilon) \neq 0$. In this case we have:

$$a\big(r_*(\varepsilon), \varepsilon\big) \leq \max_{r \in [0,1]} a(r, \varepsilon) = a(1/2, \varepsilon) \;=\; 1 + \frac{1}{4}\varepsilon\,. \tag{72}$$

Since the function $b(r, \varepsilon)$ is strict decreasing in the variable r for $\varepsilon > 0$, the following strict inequality,

$$b\big(r_*(\varepsilon), \varepsilon\big) < b(0, \varepsilon) \;=\; -1\,, \tag{73}$$

holds for $\varepsilon > 0$. Put (73) and (72) into (71) we obtain for all $\varepsilon > 0$ that

$$m(\varepsilon) = a\big(r_*(\varepsilon), \varepsilon\big) + b\big(r_*(\varepsilon), \varepsilon\big) \;<\; 1 + \frac{1}{4}\varepsilon + (-1) = \frac{1}{4}\varepsilon = m(0) + \varepsilon \bar{f}\,. \tag{74}$$

This proves (71) in the second case. □

Acknowlegments. The author thank the referees for their suggestions: a) Generalization of the case $k = 1$ to any $k = 1, 2, \ldots, n$; b) some important current references about SDP; c) Pointing out mistakes in the manuscript. The author would also like to thank João Xavier da Cruz Neto from Federal University of Piauí (UFPI) for discussions and Sissy da S. Souza from UFPI for indicating some reviews about SDP.

REFERENCES

[1] O.L. Mangasarian and R.R. Meyer, *Nonlinear perturbation of linear programs*. SIAM Journal on Control and Optimization, **17**(6) (1979), 745–752.

[2] V. Chvatal, Linear Programming. W.H. Freeman (1983).

[3] R.T. Rockefeller, Convex Analysis. Princeton University Press (1970).

[4] R.A. Horn and C.R. Johnson, Matrix Analysis. Cambridge University Press (1985).

[5] L. Vandenberghe and S. Boyd, *Semidefinite programming.* SIAM Review, **38** (1996), 49–95.

[6] A.S. Lewis and M.L. Overton, *Eigenvalue Optimization.* Acta Numerica, **5** (1996), 149–190.

[7] B. Kulis, S. Sra, S. Jegelka and I.S. Dhillon, *Scalable Semidefinite Programming using Convex Perturbations.* Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics – AISTATS, JMLR W&CP, **5** April (2009), 296–303.

[8] M.C. Ferris and O.L. Mangasarian, *Finite perturbation of convex programs.* Applied Mathematics and Optimization, **23** (1991), 221–230.

[9] P. Tseng, *Convergence and error bounds for perturbation of linear programs.* Computational Optimization and Applications, **13** (1999), 221–230.

[10] F. Alizadeh, *Interior point methods in semidefinite programming with applications to combinatorial optimization.* SIAM Journal on Optimization, **5** (1995), 13–51.