Optimality conditions and duality theory for minimizing sums of the largest eigenvalues of symmetric matrices

M. L. Overton[1] and R. S. Womersley[2]

June 1991
Submitted to *Math Programming*

Abstract: This paper gives max characterizations for the sum of the largest eigenvalues of a symmetric matrix. The elements which achieve the maximum provide a concise characterization of the generalized gradient of the eigenvalue sum in terms of a dual matrix. The dual matrix provides the information required to either verify first-order optimality conditions at a point or to generate a descent direction for the eigenvalue sum from that point, splitting a multiple eigenvalue if necessary. A model minimization algorithm is outlined, and connections with the classical literature on sums of eigenvalues are explained. Sums of the largest eigenvalues in absolute value are also addressed.

Key words: symmetric matrix, maximum eigenvalues, spectral radius, minimax problem, max characterization, generalized gradient.

Short title: Minimizing sums of eigenvalues

1. Courant Institute of Mathematical Sciences, New York University, New York, New York, 10012. The work of this author was supported by the National Science Foundation under grant CCR-88-02408.

2. School of Mathematics, University of New South Wales, P.O. Box 1, Kensington, N.S.W. 2033, Australia. The work of this author was supported in part during a visit to Argonne National Laboratory by the Applied Mathematical Sciences subprogram of the Office of Energy Research of the U.S. Department of Energy under contract W-31-109-Eng-38, and in part during a visit to the Courant Institute by the U.S. Department of Energy under Contract DEFG0288ER25053.

1 Introduction

Let A be an n by n real symmetric matrix, and let $\kappa \in \{1, \ldots, n\}$. Denote the eigenvalues of A by $\lambda_1, \ldots, \lambda_n$ and also by $\mu_1, \ldots, \mu_n$, the difference being that the former are ordered by

$$\lambda_1 \geq \cdots \geq \lambda_n, \tag{1.1}$$

while the latter are ordered by

$$|\mu_1| \geq \cdots \geq |\mu_n|. \tag{1.2}$$

Define

$$f_\kappa(A) \;=\; \sum_{i=1}^{\kappa} \lambda_i \tag{1.3}$$

and

$$g_\kappa(A) \;=\; \sum_{i=1}^{\kappa} |\mu_i|. \tag{1.4}$$

Note that $f_1(A)$ is the largest eigenvalue of A, while $g_1(A)$ is the spectral radius of A (its largest eigenvalue in absolute value).

After establishing some further notation in Section 2, the main results of the paper are given in Sections 3 and 4 for the functions $f_\kappa(A)$ and $g_\kappa(A)$ respectively. Max characterizations of these functions are established in terms of the Frobenius inner product $\langle\, A, B \,\rangle \;=\; \mathrm{tr}(AB^T)$ for $A, B \in \mathbf{R}^{m \times n}$ (where $\mathrm{tr}(A)$ denotes the trace of A), and sets of matrices defined by positive semi-definite inequalities (see Section 2). Let $\mathcal{S}_n$ denote the set of real n by n symmetric matrices,

$$\Phi_{n,\kappa} \;=\; \{\, U \in \mathcal{S}_n : 0 \leq U \leq I,\; \mathrm{tr}(U) \;=\; \kappa \,\} \tag{1.5}$$

and

$$\begin{aligned} \Psi_{n,\kappa} \;=\; & \{\, W \in \mathcal{S}_n : W = U - V, \;\text{ where } U, V \in \mathcal{S}_n, \\ & 0 \leq U \leq I,\; 0 \leq V \leq I,\; \mathrm{tr}(U) + \mathrm{tr}(V) = \kappa \,\}. \end{aligned} \tag{1.6}$$

The sets $\Phi_{n,\kappa}$ and $\Psi_{n,\kappa}$ are compact convex subsets of $\mathcal{S}_n$. It is shown that

$$f_\kappa(A) \;=\; \max_{U \in \Phi_{n,\kappa}} \langle\, A, U \,\rangle \tag{1.7}$$

and

$$g_\kappa(A) \;=\; \max_{W \in \Psi_{n,\kappa}} \langle\, A, W \,\rangle. \tag{1.8}$$

In fact, equation (1.7) implies a well known result of Fan (1949), namely

$$f_\kappa(A) = \max_{Z \in \mathbf{R}^{n \times \kappa}, Z^T Z = I} \mathrm{tr}(Z^T A Z). \tag{1.9}$$

Fan's result is widely referenced; see Wielandt (1955), Cullum, Donath and Wolfe (1975), Friedland (1981), Sameh and Wisniewski (1982), Horn and Johnson (1985) and Fletcher (1985). Both (1.7) and (1.9) show that $f_\kappa(A)$ is a convex function. The advantage of (1.7) over (1.9) is that it leads directly to a characterization of the subdifferential of f_κ which is computationally very useful, since it does not involve a convex hull operation. Equation (1.8) has a similar advantage over a corresponding analogue of (1.9).

In the case $\kappa = 1$, (1.9) reduces to the Rayleigh principle

$$f_1(A) = \max_{q^T q=1} \; q^T A(x) q \tag{1.10}$$

while (1.7) reduces to

$$f_1(A) = \max_{U \geq 0, \mathrm{tr}(U)=1} \langle\, A, U \,\rangle. \tag{1.11}$$

Note that the inequality $U \leq I$ is not required. Equation (1.11) is moderately well known; see Fletcher (1985) and Overton (1988, 1990).

Now consider the composite functions

$$f_\kappa(x) \equiv f_\kappa(A(x)), \qquad g_\kappa(x) \equiv g_\kappa(A(x)) \tag{1.12}$$

where $A(x)$ is a smooth symmetric matrix function defined on a vector of parameters $x \in \mathbf{R}^m$. The use of the same symbol f_κ for a function defined on the set of symmetric matrices and on the parameter space $\mathbf{R}^m$ is convenient, and the distinction should be clear from the context. The max characterizations (1.7) and (1.8) prove that $f_\kappa(x)$ and $g_\kappa(x)$ are locally Lipschitz, subdifferentially regular, and have generalized gradients $\partial f_\kappa(x)$ and $\partial g_\kappa(x)$ respectively, which are nonempty compact convex sets in $\mathbf{R}^m$ (see Clarke (1983)). These generalized gradients are obtained by composing the subdifferential of $f_\kappa(A)$ and $g_\kappa(A)$ with the derivative of $A(x)$, using the chain rule. In fact Clarke (1983, Proposition 2.8.8) derives an expression for the generalized gradient of the largest eigenvalue $f_1(x)$, but in a form which requires a convex hull operation. An important feature of our max characterizations is that they lead to first-order optimality conditions which are computationally verifiable, providing matrix analogues of Lagrange multipliers in constrained optimization, namely U or U and V, which we call *dual matrices*. The necessary condition $0 \in \partial f_\kappa(x)$ or $0 \in \partial g_\kappa(x)$ (see Clarke (1983)) provides systems of linear equations which can be solved to obtain the dual matrices. Inequalities of the form $0 \leq U \leq I$ determine if the current point is a stationary point, or provide information from which a descent direction can be calculated.

Equations (1.7) and (1.8) show that if $A(x)$ is an affine function then $f_\kappa(x)$ and $g_\kappa(x)$ are convex functions. They also illustrate that minimizing $f_\kappa(x)$ or $g_\kappa(x)$ are minimax problems:

$$\min_{x \in \mathbf{R}^m} f_\kappa(x) \;\equiv\; \min_{x \in \mathbf{R}^m} \max_{U \in \Phi_{n,\kappa}} \langle\, A(x), U \,\rangle, \tag{1.13}$$

and

$$\min_{x \in \mathbf{R}^m} g_\kappa(x) \equiv \min_{x \in \mathbf{R}^m} \max_{W \in \Psi_{n,\kappa}} \langle A(x), W \rangle. \tag{1.14}$$

If $A(x)$ is affine, the saddle point result

$$\min_{x \in \mathbf{R}^m} \max_{U \in \Phi_{n,\kappa}} \langle A(x), U \rangle = \max_{U \in \Phi_{n,\kappa}} \min_{x \in \mathbf{R}^m} \langle A(x), U \rangle,$$

and a similar result for (1.14) are established. This is similar to the result of Shapiro (1985) for minimizing a function of a symmetric matrix subject to positive semi-definite constraints. These saddle point results justify the dual matrix terminology.

If the eigenvalues of $A(x)$ are distinct then (1.13) and (1.14) are just minimax problems with smooth functions $\lambda_i(x) : \mathbf{R}^m \to \mathbf{R}$ for $i = 1, \ldots, n$. Let $\lambda(x) = (\lambda_1(x), \ldots, \lambda_n(x))^T$, using any ordering for the eigenvalues. Then

$$f_\kappa(x) = \max_{u \in \phi_{n,\kappa}} \lambda(x)^T u \tag{1.15}$$

where

$$\phi_{n,\kappa} = \{ u \in \mathbf{R}^n : 0 \le u_i \le 1, i = 1, \ldots, n, \quad \sum_{i=1}^n u_i = \kappa \}, \tag{1.16}$$

and

$$g_\kappa(x) = \max_{w \in \psi_{n,\kappa}} \lambda(x)^T w \tag{1.17}$$

where

$$\psi_{n,\kappa} = \{ w \in \mathbf{R}^n : -1 \le w_i \le 1, i = 1, \ldots, n, \quad \sum_{i=1}^n |w_i| = \kappa \}. \tag{1.18}$$

The additional complication in (1.13) and (1.14) arises from the possibility of multiple eigenvalues; hence the positive semi-definite constraints defining the sets $\Phi_{n,\kappa}$ and $\Psi_{n,\kappa}$.

This paper is not concerned with algorithm development. However, a brief discussion of model algorithms for minimizing $f_\kappa(x)$ and $g_\kappa(x)$ is given. These are generalizations of the algorithm presented by Overton (1988) for the case when $\kappa = 1$ and $A(x)$ is affine. It is, in fact, possible to design the model algorithms so that they have quadratic local convergence, even if the objective function is not smooth at the solution; see Overton (1988) and Overton (1990). More detail will be given by Overton and Womersley (to appear). Cullum, Donath and Wolfe (1975) gave an algorithm, related to the ϵ-subgradient methods of Lemarechal and others, for the case that the variables x are the diagonal elements of $A(x)$. The most important difference between this earlier work and our model algorithms is that the latter compute the dual matrices which demonstrate optimality. These matrices are also the key to sensitivity analysis of the solution; see Overton (1990, Section 3).

Sums of eigenvalues of symmetric matrices have been addressed in one form or another in many classical papers on matrix theory; a good overview is Bellman (1970, Chapter 8). However, the rich interconnection between this subject and the sets $\Phi_{n,\kappa}$ and $\Psi_{n,\kappa}$ appears to have been largely overlooked. Note that although all our results are given in terms of real symmetric matrices, generalization to the complex Hermitian case is straightforward.

The classical literature on sums of eigenvalues does not include much discussion of applications. However, these appear to be quite numerous, especially in connection with adjacency matrices of graphs. See in particular Cullum, Donath and Wolfe (1975), as well as Rendl and Wolkowicz (1990), and Alizadeh (1991). Another application is the "orthogonal Procrustes" problem, which refers to rotating a number of matrices towards a best least squares fit. This problem is discussed by Shapiro and Botha (1988) and has also been addressed by Watson (1990). There is a large variety of applications in the case $\kappa = 1$; see Overton(1990).

2 Notation

The following notation is used throughout this paper. Let

1. $\mathcal{S}_n$ = set of all n by n real symmetric matrices ($A^T = A$).
2. $\mathcal{K}_n$ = set of all n by n real skew-symmetric matrices ($A^T = -A$).
3. $\mathcal{D}_n$ = set of all n by n real diagonal matrices.
4. $\mathcal{O}_{m,n}$ = set of all m by n real orthogonal matrices, where $m \geq n$.
   Thus $Z^T Z = I_n$ for all $Z \in \mathcal{O}_{m,n}$.

The vector e_i is the ith coordinate vector, e is the vector of all 1s, and I is the identity matrix, with the dimensions of e_i, e and I determined by the context. A matrix $D \in \mathcal{D}_n$ is denoted by $\mathrm{diag}(\alpha_1, \ldots, \alpha_n)$, or $\mathrm{diag}(u)$ where $u \in \mathbf{R}^n$. The convex hull of a set Ω is denoted by conv Ω.

For $A \in \mathbf{R}^{n \times n}$ the eigenvalues of A are denoted by (1.1) and by (1.2). The trace of A is

$$\mathrm{tr}(A) = \sum_{i=1}^{n} a_{ii} = \sum_{i=1}^{n} \lambda_i = \sum_{i=1}^{n} \mu_i.$$

The positive semi-definite partial ordering on $\mathcal{S}_n$ is used to express matrix inequalities (see Golub and van Loan (1985) or Horn and Johnson (1985) for example). Thus $A \geq 0$ means that A is positive semi-definite (equivalently $y^T A y \geq 0 \ \forall y$, or $\lambda_i \geq 0$ for $i = 1, \ldots, n$). Hence $A \geq B$ means that $A - B$ is positive semi-definite. For example the constraints $0 \leq A \leq I$ on $A \in \mathcal{S}_n$ mean that $0 \leq \lambda_i \leq 1$ for $i = 1, \ldots, n$.

The Frobenius inner product $\langle A, B \rangle$ of two matrices $A, B \in \mathbf{R}^{m \times n}$ is

$$\langle A, B \rangle \;=\; \mathrm{tr}(AB^T) \;=\; \sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij} b_{ij}.$$

This inner product is the natural extension for real matrix variables of the standard inner product $x^T y = \sum_{i=1}^{n} x_i y_i$ on $\mathbf{R}^n$. It is widely used in problems with matrix variables, for example in the work of Bellman and Fan (1963), Arnold (1971), Craven and Mond (1981). Fletcher (1985), Overton (1988), and Overton and Womersley (1988) used the notation $A : B$ for $\langle A, B \rangle$.

Some useful properties of the Frobenius inner product are summarized below.

1. $\langle A, A \rangle \;=\; ||A||_F^2$.
2. $\langle A, I \rangle \;=\; \mathrm{tr}(A)$.
3. If $A \in \mathcal{S}_n$ and $K \in \mathcal{K}_n$ then $\langle A, K \rangle \;=\; 0$.
4. As $\mathrm{tr}(AB^T) = \mathrm{tr}(BA^T) = \mathrm{tr}(A^T B)$

$$\langle A, B \rangle \;=\; \langle B, A \rangle \;=\; \langle A^T, B^T \rangle \;=\; \langle B^T, A^T \rangle.$$

In particular

(a) For any nonsingular matrices $S \in \mathbf{R}^{m \times m}$ and $T \in \mathbf{R}^{n \times n}$

$$\langle A, B \rangle \;=\; \langle S^{-1}A, S^T B \rangle \;=\; \langle AT^{-1}, BT^T \rangle.$$

(b) If $A \in \mathbf{R}^{n \times n}, B \in \mathbf{R}^{m \times m}$ and $Z \in \mathbf{R}^{n \times m}$ then

$$\langle Z^T AZ, B \rangle \;=\; \langle A, ZBZ^T \rangle.$$

3 Sum of the largest eigenvalues

Let $A \in \mathcal{S}_n$ have eigenvalues

$$\lambda_1 \geq \cdots \geq \lambda_n,$$

and let $Q \in \mathcal{O}_{n,n}$ be a matrix whose columns are normalized eigenvectors of A, so

$$Q^T AQ \;=\; \Lambda \tag{3.1}$$

where $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$. The matrix Q will be regarded as fixed, although if A has multiple eigenvalues the choice of Q is not unique.

Let $\kappa \in \{1, \ldots, n\}$. Section 3.1 establishes a max characterization of

$$f_\kappa(A) \;=\; \sum_{i=1}^{\kappa} \lambda_i. \tag{3.2}$$

Thereafter we concentrate on the case when $A : \mathbf{R}^m \to \mathcal{S}_n$ is a smooth matrix-valued function and

$$f_\kappa(x) \;=\; f_\kappa(A(x)). \tag{3.3}$$

Section 3.2 considers the differential properties of $f_\kappa(x)$, Section 3.3 gives necessary conditions for a local minimizer, Section 3.4 establishes a saddle point result, Section 3.5 gives a formula for the directional derivative, Section 3.6 discusses the generation of descent directions by splitting multiple eigenvalues and Section 3.7 considers a model algorithm for minimizing $f_\kappa(x)$.

3.1 A max characterization

Let

$$\Phi_{n,\kappa} \;=\; \{\, U \in \mathcal{S}_n : 0 \le U \le I, \quad \mathrm{tr}(U) = \kappa \,\}, \tag{3.4}$$

and

$$\phi_{n,\kappa} = \{\, u \in \mathbf{R}^n : \; 0 \le u \le e, e^T u = \kappa \,\}. \tag{3.5}$$

Lemma 3.1 *The sets $\Phi_{n,\kappa}$ and $\phi_{n,\kappa}$ are compact and convex. Moreover $\Phi_{n,\kappa}$ is invariant under orthogonal similarity transformations (i.e. $U \in \Phi_{n,\kappa} \iff Z^T U Z \in \Phi_{n,\kappa}$ where $Z \in \mathcal{O}_{n,n}$), and $\Phi_{n,\kappa}$ and $\phi_{n,\kappa}$ are related by*

$$\begin{aligned} \Phi_{n,\kappa} \;=\; \{\, & U \in \mathcal{S}_n : U = Z D Z^T \text{ where } Z \in \mathcal{O}_{n,n}, \\ & D = \mathrm{diag}(u_1, \ldots, u_n) \text{ and } u \in \phi_{n,\kappa} \,\}, \end{aligned} \tag{3.6}$$

and

$$\phi_{n,\kappa} \;=\; \{\, u \in \mathbf{R}^n : u_i = U_{ii} \text{ for } i = 1, \ldots, n \text{ where } U \in \Phi_{n,\kappa} \,\}. \tag{3.7}$$

Proof: That $\Phi_{n,\kappa}$ and $\phi_{n,\kappa}$ are compact convex sets is immediate. A spectral decomposition of U yields (3.6) and the orthogonal invariance of $\Phi_{n,\kappa}$. The set $\phi_{n,\kappa}$ is contained in the right-hand side of (3.7) since for any $u \in \phi_{n,\kappa}$, $\mathrm{diag}(u_1, \ldots, u_n) \in \Phi_{n,\kappa}$. To obtain the reverse inclusion, let $U \in \Phi_{n,\kappa}$ and define u by $u_i = U_{ii}$. The facts that the trace is the sum of the diagonal elements and that a positive semi-definite matrix cannot have a negative diagonal element then show that $u \in \phi_{n,\kappa}$. ∎

To characterize the elements that achieve the maximum in the following results, information about the multiplicity of the eigenvalues of A is needed. Let

$$\begin{aligned} &\lambda_1 \geq \cdots \geq \lambda_r > \\ &\lambda_{r+1} = \cdots = \lambda_\kappa = \cdots = \lambda_{r+t} > \\ &\lambda_{r+t+1} \geq \cdots \geq \lambda_n, \end{aligned} \tag{3.8}$$

where $t \geq 1$ and $r \geq 0$ are integers. The multiplicity of the κth eigenvalue is t. The number of eigenvalues larger than λ_κ is r. Here r may be zero; in particular this must be the case if $\kappa = 1$. Note that by definition

$$r + 1 \leq \kappa \leq r + t \leq n,$$

so $t \geq \kappa - r$. Also, $t = 1$ implies that $\kappa = r + 1$.

First two lemmas, which depend only on the definitions of the sets $\Phi_{n,\kappa}$ and $\phi_{n,\kappa}$ and the ordering of the elements λ_i in (3.8), are established. In particular they do not require λ_i to be an eigenvalue of a matrix.

Lemma 3.2 *If the elements of $\lambda \in \mathbf{R}^n$ satisfy (3.8) then*

$$\max_{u \in \phi_{n,\kappa}} \lambda^T u = \sum_{i=1}^{\kappa} \lambda_i$$

with

$$\begin{aligned} \operatorname{argmax} \{ \lambda^T u : u \in \phi_{n,\kappa} \} = \{ u \in \mathbf{R}^n \ : \ & u_i = 1 && i = 1, \ldots, r, \\ & 0 \leq u_i \leq 1 && i = r+1, \ldots, r+t, \\ & u_i = 0 && i = r+t+1, \ldots, n \text{ and} \\ & \sum_{i=r+1}^{r+t} u_i = \kappa - r \}. \end{aligned}$$

Proof: These results follow directly from (3.5) and (3.8). ∎

Lemma 3.3 *Let $\Lambda = \operatorname{diag}(\lambda)$ where the elements of $\lambda \in \mathbf{R}^n$ satisfy (3.8). Then*

$$\max_{U \in \Phi_{n,\kappa}} \langle \Lambda, U \rangle = \sum_{i=1}^{\kappa} \lambda_i \tag{3.9}$$

with

$$\begin{aligned} \operatorname{argmax} \quad & \{ \langle \Lambda, U \rangle : U \in \Phi_{n,\kappa} \} \\ = \quad & \{ U \in \mathcal{S}_n : U = \begin{bmatrix} I & & \\ & \tilde{U} & \\ & & 0 \end{bmatrix}, \tilde{U} \in \Phi_{t,\kappa-r} \} \end{aligned} \tag{3.10}$$

where

$$\Phi_{t,\kappa-r} = \{\, \tilde{U} \in \mathcal{S}_t : 0 \le \tilde{U} \le I \ \text{ and } \ \mathrm{tr}(\tilde{U}) = \kappa - r \,\}. \tag{3.11}$$

Here the diagonal blocks of U have dimension r, t, and $n-r-t$ respectively.

Proof: $\Lambda = \mathrm{diag}(\lambda)$ so

$$\langle\, \Lambda, U \,\rangle = \sum_{i=1}^{n} \lambda_i U_{ii}. \tag{3.12}$$

Hence (3.9) follows from combining Lemmas 3.1 and 3.2.

If U^* is any element of the right hand side of (3.10) then from (3.8)

$$\langle\, \Lambda, U^* \,\rangle \ = \ \sum_{i=1}^{r} \lambda_i + \lambda_\kappa \mathrm{tr}(\tilde{U}) \ = \ \sum_{i=1}^{\kappa} \lambda_i. \tag{3.13}$$

Now suppose $U^* \in \mathrm{argmax}\,\{\, \langle\, \Lambda, U \,\rangle : U \in \Phi_{n,\kappa} \,\}$. Then $u^* = (U^*_{11}, \ldots, U^*_{nn})^T \in \phi_{n,\kappa}$ satisfies $\lambda^T u^* = \sum_{i=1}^{\kappa} \lambda_i$, so from Lemma 3.2 it follows that

$$\begin{aligned} &U^*_{ii} = 1, && i = 1, \ldots, r, \\ &0 \le U^*_{ii} \le 1, && i = r+1, \ldots, r+t \\ &U^*_{ii} = 0, && i = r+t+1, \ldots, n \\ &\sum_{i=r+1}^{r+t} U^*_{ii} \ = \ \kappa - r. \end{aligned} \tag{3.14}$$

Partition the rows and columns of U^* into blocks of r, t and $n-r-t$ elements so

$$U^* = \begin{bmatrix} C_{11} & C_{12} & C_{13} \\ C_{12}^T & C_{22} & C_{23} \\ C_{13}^T & C_{23}^T & C_{33} \end{bmatrix} \tag{3.15}$$

where $C_{11} \in \mathcal{S}_r, C_{22} \in \mathcal{S}_t, C_{33} \in \mathcal{S}_{n-r-t}$. As $U^* \in \Phi_{n,\kappa}$,

$$0 \le C_{ii} \le I, \quad i = 1,2,3 \qquad \text{and} \qquad \sum_{i=1}^{3} \mathrm{tr}(C_{ii}) = \kappa.$$

From (3.14) and (3.15) the diagonal elements of C_{11} are all 1 and the diagonal elements of C_{33} are all zero. Suppose an off-diagonal element of C_{11} is nonzero. Then $I - C_{11}$ is symmetric, has zero elements on the diagonal and a nonzero off-diagonal element, so is indefinite. This contradicts $C_{11} \le I$, so $C_{11} = I$. Similarly

$$I - \begin{bmatrix} C_{11} & C_{12} \\ C_{12}^T & C_{22} \end{bmatrix} = \begin{bmatrix} 0 & -C_{12} \\ -C_{12}^T & I - C_{22} \end{bmatrix} \ge 0$$

implies that $C_{12} = 0$. As the diagonal elements of C_{33} are all zero the existence of a nonzero off-diagonal element in C_{33} would contradict $C_{33} \ge 0$, so $C_{33} = 0$. Similarly

it follows that $C_{13} = 0$ and $C_{23} = 0$ as the existence of a nonzero element in either of these submatrices contradicts $U^* \geq 0$. The only conditions remaining then are $0 \leq C_{22} \leq I$ with $\mathrm{tr}(C_{22}) = \kappa - r$ as required. ∎

Remark: The proof uses the basic results that if $C \in \mathcal{S}_n$ then

$$C \geq 0, \quad \mathrm{tr}(C) = 0 \quad \Rightarrow \quad C = 0, \tag{3.16}$$

$$C \leq I, \quad \mathrm{tr}(C) = n \quad \Rightarrow \quad C = I, \tag{3.17}$$

which follow from the observation that if $C \geq 0$ and $C_{ii} = 0$ then $C_{ij} = 0$ and $C_{ji} = 0$ for $j = 1, \ldots, n$.

Now consider sums of the eigenvalues of A. Let $P_1 \in \mathcal{O}_{n,r}$ be the matrix consisting of the first r columns of Q (defined in equation (3.1)), and let $Q_1 \in \mathcal{O}_{n,t}$ be the matrix consisting of the next t columns of Q. By (3.8) we have

$$P_1^T A P_1 = \mathrm{diag}(\lambda_1, \ldots, \lambda_r); \qquad Q_1^T A Q_1 = \lambda_\kappa I. \tag{3.18}$$

Theorem 3.4 *Let $A \in \mathcal{S}_n$ have eigenvalues $\lambda_1 \geq \cdots \geq \lambda_n$. Then*

$$\max_{U \in \Phi_{n,\kappa}} \langle A, U \rangle \; = \; \sum_{i=1}^{\kappa} \lambda_i. \tag{3.19}$$

If the eigenvalues satisfy (3.8) then

$$\begin{aligned} \mathrm{argmax} \quad & \{ \langle A, U \rangle : U \in \Phi_{n,\kappa} \} \\ = \quad & \{ U \in \mathcal{S}_n : U = P_1 P_1^T + Q_1 \tilde{U} Q_1^T; \;\; \tilde{U} \in \Phi_{t,\kappa-r} \}, \end{aligned} \tag{3.20}$$

where P_1, Q_1 satisfy (3.18).

Proof: For any $U \in \Phi_{n,\kappa}$ equation (3.1) and the properties of the Frobenius inner product imply that

$$\langle A, U \rangle \; = \; \langle Q \Lambda Q^T, U \rangle \; = \; \langle \Lambda, Q^T U Q \rangle, \tag{3.21}$$

with $Q^T U Q \in \Phi_{n,\kappa}$ as $\Phi_{n,\kappa}$ is invariant to orthogonal transformations. Hence

$$\max_{U \in \Phi_{n,\kappa}} \langle A, U \rangle = \max_{U \in \Phi_{n,\kappa}} \langle \Lambda, U \rangle = \sum_{i=1}^{\kappa} \lambda_i,$$

where the second equality follows from Lemma 3.3.

From (3.21) and the invariance of $\Phi_{n,\kappa}$ to orthogonal transformations

$$\mathrm{argmax} \, \{ \langle A, U \rangle : U \in \Phi_{n,\kappa} \} = \{ Q U^* Q^T : U^* \in \Omega \}$$

where $\Omega = \text{argmax}\{\langle \Lambda, U \rangle : U \in \Phi_{n,\kappa}\}$. The proof is completed by applying Lemma 3.3 since

$$QU^*Q^T = Q \begin{bmatrix} I & & \\ & \tilde{U} & \\ & & 0 \end{bmatrix} Q^T = P_1 P_1^T + Q_1 \tilde{U} Q_1^T.$$

∎

Remark: The term dual matrix is used to refer to either $U \in \Phi_{n,\kappa}$ or $\tilde{U} \in \Phi_{t,\kappa-r}$. The distinction between U and $\tilde{U}$ is analogous to the question of whether or not to assign zero Lagrange multipliers to inactive constraints in nonlinear programming.

Remark: If $\kappa = r + t$ then $\tilde{U} = I$ and

$$\text{argmax}\{\langle A, U \rangle : U \in \Phi_{n,\kappa}\} \;=\; P_1 P_1^T + Q_1 Q_1^T. \tag{3.22}$$

The matrix achieving the maximum is unique if and only if $\kappa = r + t$. In particular this is the case if $t = 1$, i.e. the κth largest eigenvalue is simple.

Remark: All the freedom in the choice of Q_1 in (3.18) is absorbed into the matrix $\tilde{U}$. It also makes no difference if any of the eigenvalues $\lambda_1, \ldots, \lambda_r$ have multiplicity greater than 1, as **all** of these multiple eigenvalues are included in $f_\kappa(A)$. The corresponding columns of P_1 can be any orthonormal basis for the corresponding eigenspace, without affecting $P_1 P_1^T$.

Remark: As $A \in \mathcal{S}_n$ it follows that $\langle A, K \rangle = 0$ for any $K \in \mathcal{K}_n$ (the set of skew-symmetric n by n matrices). Hence

$$f_\kappa(A) \;=\; \langle A, U + K \rangle$$

for any U belonging to (3.20) and $K \in \mathcal{K}_n$.

Corollary 3.5 *The function $f_\kappa(A) : \mathcal{S}_n \to \mathbf{R}$ is convex and its subdifferential $\partial f_\kappa(A)$ is the nonempty compact convex set*

$$\partial f_\kappa(A) = \{ U \in \mathcal{S}_n : \exists\ \tilde{U} \in \Phi_{t,\kappa-r} \text{ with } U = P_1 P_1^T \;+\; Q_1 \tilde{U} Q_1^T \}. \tag{3.23}$$

Proof: For any $U \in \Phi_{n,\kappa}$ the inner product $\langle A, U \rangle$ is a linear function of $A \in \mathcal{S}_n$, so the convexity of $f_\kappa(A)$ follows from the max characterization in Theorem 3.4. Moreover, from Rockafellar (1970, corollary 23.5.3), the subdifferential of a function

defined as a pointwise maximum of a set of linear functions is the convex hull of the gradients of the linear functions achieving the maximum at the given point. Thus the subdifferential of $f_\kappa(A)$ is

$$\partial f_\kappa(A) \;=\; \mathrm{conv}\,\{\, U \in \mathcal{S}_n : \langle\, A, U \,\rangle = f_\kappa(A) \,\}.$$

The result follows from Theorem 3.4 as the set (3.20) is already convex. ∎

Remark: As previously noted any skew-symmetric matrix can be added to U without affecting the inner product $\langle\, A, U \,\rangle$. As $f_\kappa(A)$ is defined on $\mathcal{S}_n$ only the symmetric subdifferential is of interest.

Remark: In the case $\kappa = 1$, the condition $\tilde{U} \leq I$ is unnecessary since it is implied by $\tilde{U} \geq 0, \mathrm{tr}(\tilde{U}) = 1$. This case is moderately well known; see Fletcher (1985). Fletcher also addressed the case $\kappa > 1$ in an appendix, but his Theorem A.4 is incorrect in the case $\kappa > 1$ since the condition $\tilde{U} \leq I$ was omitted.

This section is concluded by relating our max characterization to the well known result of Fan (1949).

Lemma 3.6 *The extreme points of $\phi_{n,\kappa}$ are*

$$\{\, u : u_i = \begin{cases} 1 & \textit{for exactly } \kappa \textit{ of the indicies } 1, \ldots, n; \\ 0 & \textit{otherwise.} \end{cases} \,\}. \tag{3.24}$$

Proof: Straightforward. ∎

Lemma 3.7 *The extreme points of $\Phi_{n,\kappa}$ are the elements in $\Phi_{n,\kappa}$ with rank κ, that is the set of matrices $U \in \Phi_{n,\kappa}$ with κ eigenvalues equal to 1, and $n - \kappa$ eigenvalues equal to zero.*

Proof: Matrices in $\Phi_{n,\kappa}$ must have rank at least κ. Since $\Phi_{n,\kappa}$ and $\phi_{n,\kappa}$ are related by (3.6), it is straightforward to show that any element of $\Phi_{n,\kappa}$ with rank greater than κ is not an extreme point. The only candidates for extreme points, then, are those with rank κ. But it is clearly not possible that some rank κ elements are extreme points and others not, since the definition of $\Phi_{n,\kappa}$ does not in any way distinguish between different rank κ elements. Since a compact convex set must have extreme points, the proof is complete. ∎

Fan's theorem now follows as a corollary.

Theorem 3.8 (Fan)

$$\max_{Z \in \mathcal{O}_{n,\kappa}} \mathrm{tr}(Z^T A Z) = \sum_{i=1}^{\kappa} \lambda_i. \tag{3.25}$$

Proof: Since a linear function on a convex set must assume its maximum at an extreme point, combining Theorem 3.4 with the lemma just given shows that

$$\max_{U \in \Phi_{n,\kappa}, \operatorname{rank} U = \kappa} \langle A, U \rangle = \sum_{i=1}^{\kappa} \lambda_i. \tag{3.26}$$

Such matrices U have precisely the form ZZ^T, where $Z \in \mathcal{O}_{n,\kappa}$. The proof is completed by noting that

$$\langle A, ZZ^T \rangle = \langle Z^T A Z, I \rangle = \operatorname{tr}(Z^T A Z). \tag{3.27}$$

∎

Remark: Another expression for the subdifferential is

$$\partial f_\kappa(A) = \operatorname{conv} \{ ZZ^T \ : \ \begin{array}{l} \text{columns of } Z \text{ form an orthonormal set} \\ \text{of } \kappa \text{ eigenvectors for } \lambda_1, \ldots, \lambda_\kappa \}. \end{array} \tag{3.28}$$

Note that these are just the elements that achieve the maximum in Fan's theorem, and that the elements whose convex hull is being taken are the extreme points of (3.23). Although simpler to write than (3.23) this expression is not as useful, as the structure of the subdifferential is not apparent.

Remark: H. Woerdeman and C.-K. Li recently informed us that the equality

$$\operatorname{conv} \{ ZZ^T : \ Z \in \mathcal{O}_{n,\kappa} \} \ = \ \Phi_{n,\kappa}.$$

appeared in Fillmore and Williams (1971). This implies the equivalence of (3.19) and (3.25), although it does not by itself imply either of them; nor, strictly speaking, does it imply Lemma 3.7.

3.2 The generalized gradient

Let $A(x) : \mathbf{R}^m \to \mathcal{S}_n$ be a smooth (at least once continuously differentiable) function whose partial derivative with respect to x_k is

$$A_k(x) \ \equiv \ \frac{\partial A(x)}{\partial x_k} \ \text{ for } \ k = 1, \ldots, m.$$

This section is concerned with finding a computationally useful characterization of the generalized gradient of the function

$$f_\kappa(x) \ = \ f_\kappa(A(x)).$$

Although the eigenvalues and eigenvectors of $A(x)$ are functions of $x \in \mathbf{R}^m$, the explicit dependence on x will usually be omitted. Thus the eigenvalues of $A(x)$ are denoted by as before (3.8), with r and t now dependent on x, and with corresponding eigenvectors satisfying (3.18).

Theorem 3.9 *The function $f_\kappa(x)$ is locally Lipschitz, subdifferentially regular, and its generalized gradient is the nonempty compact convex set*

$$\begin{aligned}\partial f_\kappa(x) \;=\; \{\, u \in \mathbf{R}^m : \exists\ \tilde{U} \in \mathcal{S}_t \ \text{ with } \ 0 \le \tilde{U} \le I,\ \ \mathrm{tr}(\tilde{U}) = \kappa - r,\ \ \text{and} \\ u_k = \mathrm{tr}(P_1^T A_k(x) P_1) \;+\; \langle\, Q_1^T A_k(x) Q_1, \tilde{U} \,\rangle,\ \ k = 1, \ldots, m \,\}. \end{aligned} \tag{3.29}$$

Proof: Since $f_\kappa(A)$ is convex and $A(x)$ is smooth the chain rule (Clarke (1983), Theorem 2.3.10) implies that $f_\kappa(x)$ is locally Lipschitz, subdifferentially regular, and that

$$\begin{aligned}\partial f_\kappa(x) \;=\; \{\, u \in \mathbf{R}^m \;\;:\;\; u_k \;=\; \langle\, A_k(x), U \,\rangle, k = 1, \ldots, m \\ \text{where} \quad U \in \partial f_\kappa(A(x)) \,\}. \end{aligned} \tag{3.30}$$

Corollary 3.5 and the properties of the inner product complete the proof. ∎

Remark: Since by Theorem 3.4

$$f_\kappa(x) \;=\; \max_{U \in \Phi_{n,\kappa}} \langle\, A(x), U \,\rangle.$$

the result also follows from the characterization of generalized gradients of functions defined by a pointwise maximum in Clarke (1983, Theorem 2.8.6). The Clarke characterization shows that

$$\begin{aligned}\partial f_\kappa(x) \;=\; \mathrm{conv}\,\{\, u \in \mathbf{R}^m \;\;:\;\; u_k \;=\; \langle\, A_k(x), U \,\rangle, k = 1, \ldots, m \\ \text{where} \quad \langle\, A(x), U \,\rangle = f_\kappa(x) \,\}. \end{aligned} \tag{3.31}$$

Equation (3.29) follows from Theorem 3.4 since the maximizing set is already convex.

Remark: Bellman and Fan (1963) gave an example where the set

$$\{\, u \in \mathbf{R}^m : u_k \;=\; \langle\, A_k, U \,\rangle, \ \text{where } \ U \ge 0 \,\}$$

is not closed, and gave sufficient conditions for it to be closed. This is not a difficulty in our case because the trace condition ensures that $\Phi_{n,\kappa}$ is compact. As $f_\kappa(x) : \mathbf{R}^m \to \mathbf{R}$ is a locally Lipschitz function its generalized gradient is a nonempty compact convex set in $\mathbf{R}^m$.

Corollary 3.10 *If $\lambda_\kappa > \lambda_{\kappa+1}$ (i.e. $\kappa = r + t$) the function f_κ is differentiable at x with*

$$\frac{\partial f_\kappa(x)}{\partial x_k} \;=\; \mathrm{tr}(P_1^T A_k(x) P_1) \;+\; \mathrm{tr}(Q_1^T A_k(x) Q_1). \tag{3.32}$$

Proof: As $\kappa - r = t$ the only solution to $\mathrm{tr}(\tilde{U}) = t$ and $0 \leq \tilde{U} \leq I$ is $\tilde{U} = I$. Hence the set $\partial f_\kappa(x)$ is a singleton, $f_\kappa(x)$ is differentiable, and (3.29) reduces to (3.32). ∎

Remark: A key point here is that if λ_j is an eigenvalue of $A(x)$ with multiplicity ℓ there exists a neighbourhood of x in which the corresponding group of ℓ eigenvalues is distinct from all the other eigenvalues. The sum of **all** the eigenvalues in this group is a differentable function in this neighbourhood (see Kato(1982)). In particular, simple eigenvalues are smooth functions of x.

Remark: Cullum, Donath and Wolfe (1975) studied the case where $A(x)$ is affine and only the diagonal elements of $A(x)$ vary, so that $m = n$,

$$A(x) = A_0 + \mathrm{diag}(x_1, \ldots, x_n),$$

and $A_k(x) = e_k e_k^T$ for $k = 1, \ldots, n$. Using Fan's theorem they showed that $f_\kappa(x)$ is differentiable at x when $\lambda_\kappa > \lambda_{\kappa+1}$, and that

$$\begin{aligned} \partial f_\kappa(x) \;\; = \;\; \mathrm{conv}\, \{\, v \in \mathbf{R}^n : v_k \;\; = \;\; \mathrm{tr}(P_1^T A_k(x) P_1) \;\; + \;\; \mathrm{tr}(Z^T Q_1^T A_k(x) Q_1 Z) \\ \text{for } \; k = 1, \ldots, m \;\; \text{and} \;\; Z \in \mathcal{O}_{n,t} \,\}. \end{aligned} \tag{3.33}$$

The relationship between (3.33) and (3.29) is precisely that already explained between (3.28) and (3.23), namely the argument of the convex hull in (3.33) is the set of extreme points of (3.29).

3.3 Necessary conditions

The standard necessary condition for x to be a local minimizer of f_κ, namely $0 \in \partial f(x)$ (see Clarke (1983)), implies there exists

$$U \in \mathrm{argmax}\, \{ \langle\, A(x), U \,\rangle : U \in \Phi_{n,\kappa} \} \tag{3.34}$$

such that

$$\langle\, A_k, U \,\rangle \;\; = \;\; 0 \;\; \text{for} \;\; k = 1, \ldots, m. \tag{3.35}$$

From Theorem 3.9 equations (3.34) and (3.35) are equivalent to the existence of a $\tilde{U} \in \mathcal{S}_t$ such that

$$0 \leq \tilde{U} \leq I, \quad \mathrm{tr}(\tilde{U}) = \kappa - r, \tag{3.36}$$

and

$$\mathrm{tr}({P_1}^T A_k P_1) \;\; + \;\; \langle\, {Q_1}^T A_k Q_1, \tilde{U} \,\rangle \;\; = \;\; 0 \;\; \text{for} \;\; k = 1, \ldots, m. \tag{3.37}$$

The conditions (3.36) and (3.37) are useful computationally as one can relax the inequalities on $\tilde{U}$ and solve (3.37) together with $\mathrm{tr}(\tilde{U}) = \kappa - r$ for $\tilde{U}$. This requires solving a system of $m + 1$ linear equations for the $t(t+1)/2$ unknowns in the symmetric matrix $\tilde{U}$. If the inequalities $0 \leq \tilde{U} \leq I$ are not satisfied then a descent direction may be generated. This is discussed in Section 3.6.

If f_κ is convex (for example if $A(x)$ is affine) then equations (3.36) and (3.37) are both necessary and sufficient for x to be a minimizer of f_κ.

3.4 A saddle point result

Consider the function

$$\mathcal{L}(x,U) \;=\; \langle\, A(x), U \,\rangle. \tag{3.38}$$

This section establishes a saddle point result, based on well known results for Lagrangian functions in convex programming (Rockafellar (1970)). A point $x^* \in \mathbf{R}^m, U^* \in \Phi_{n,\kappa}$ is a **saddle point** of $\mathcal{L}(x,U)$ if and only if

$$\mathcal{L}(x^*,U) \le \mathcal{L}(x^*,U^*) \le \mathcal{L}(x,U^*) \quad \forall\, x \in \mathbf{R}^m \quad \forall\, U \in \Phi_{n,\kappa}. \tag{3.39}$$

It is well known and easy to show that

$$\min_{x\in\mathbf{R}^m} \max_{U\in\Phi_{n,\kappa}} \mathcal{L}(x,U) \;\ge\; \max_{U\in\Phi_{n,\kappa}} \min_{x\in\mathbf{R}^m} \mathcal{L}(x,U). \tag{3.40}$$

The primal problem is

$$\min_{x\in\mathbf{R}^n} f_\kappa(x) \tag{3.41}$$

where

$$f_\kappa(x) = \max_{U\in\Phi_{n,\kappa}} \mathcal{L}(x,U). \tag{3.42}$$

Define the dual problem to be

$$\max_{U\in\Phi_{n,\kappa}} h(U), \tag{3.43}$$

where

$$h(U) = \min_{x\in\mathbf{R}^m} \mathcal{L}(x,U). \tag{3.44}$$

The following saddle point result is similar to that of Shapiro (1985) for minimizing a function of a symmetric matrix subject to positive semi-definite constraints. Indeed, it follows from the general saddle point theory for convex-concave functions (Rockafellar(1970), Theorem 36.3), but we give the proof for completeness.

Theorem 3.11 *For each $U \in \Phi_{n,\kappa}$ let $\mathcal{L}(.,U)$ be a convex function, and let the primal problem attain its solution at x^*. Then the primal and dual problems have the same optimal value so*

$$\min_{x\in\mathbf{R}^m} \max_{U\in\Phi_{n,\kappa}} \mathcal{L}(x,U) = \max_{U\in\Phi_{n,\kappa}} \min_{x\in\mathbf{R}^m} \mathcal{L}(x,U), \tag{3.45}$$

and U^* satisfying (3.34) and (3.35) solves the dual problem (3.43).

Proof: From (3.42) $f_\kappa(x)$ is a convex function as it is a maximum of convex functions, so $0 \in \partial f_\kappa(x^*)$. Hence there exists a U^* satisfying equations (3.34) and

(3.35). We only have to show that (x^*, U^*) is a saddle point of $\mathcal{L}(x, U)$. From Theorem 3.4 and (3.34)

$$\mathcal{L}(x^*, U^*) = f_\kappa(x^*) \geq \mathcal{L}(x^*, U) \quad \forall \quad U \in \Phi_{n,\kappa}.$$

The function $\mathcal{L}(., U^*)$ is convex, so $0 \in \partial f_\kappa(x^*)$ implies that $\mathcal{L}(., U^*)$ attains its minimum at x^*. This establishes (3.39). ∎

Remark: As $A(x)$ is a smooth function of x, so is $\mathcal{L}(., U)$ for any U. Hence a necessary condition for x^* to be a local minimizer of $\mathcal{L}(., U)$ is that

$$\langle\, A_k(x^*), U \,\rangle \;=\; 0 \text{ for } k = 1, \ldots, m. \tag{3.46}$$

If $A(x)$ is an affine function then $\mathcal{L}(., U)$ is also an affine (and hence convex) function. In this case $\mathcal{L}(., U)$ does not have a finite minimum unless it is constant, so (3.46) holds for all $x \in \mathbf{R}^m$.

3.5 The directional derivative

As $f_\kappa(x)$ is subdifferentially regular the standard one-sided directional derivative at x in a direction $d \in \mathbf{R}^m$ exists and satisfies

$$\begin{aligned} f'_\kappa(x; d) &\equiv \lim_{\alpha \to 0+} \frac{f_\kappa(x + \alpha d) - f_\kappa(x)}{\alpha} \\ &= \max_{u \in \partial f_\kappa(x)} u^T d \\ &= \sum_{k=1}^{m} d_k \;\operatorname{tr}(P_1^T A_k P_1) + \max_{\tilde{U} \in \Phi_{t,\kappa-r}} \sum_{k=1}^{m} d_k \langle\, Q_1^T A_k Q_1, \tilde{U} \,\rangle. \end{aligned} \tag{3.47}$$

Recall that the matrices P_1 and Q_1, defined by (3.18), are evaluated at the point x, and that A_k is the partial derivative of $A(x)$ with respect to x_k evaluated at the point x. Define

$$b_k = \operatorname{tr}(P_1^T A_k P_1), \qquad B_k = Q_1^T A_k Q_1 \text{ for } k = 1, \ldots, m \tag{3.48}$$

and define $B(d) \in \mathcal{S}_t$ by

$$B(d) = \sum_{k=1}^{m} d_k B_k. \tag{3.49}$$

Note that $b^T d$ is the sum of the eigenvalues (the trace) of

$$\sum_{k=1}^{m} d_k \; P_1^T A_k P_1.$$

Let the eigenvalues of $B(d)$ be $\beta_1 \geq \cdots \geq \beta_t$. Then from (3.47) and Theorem 3.4 it follows that

$$f'_\kappa(x;d) = b^T d \; + \; \sum_{i=1}^{\kappa-r} \beta_i. \tag{3.50}$$

Hence $f'_\kappa(x;d)$ is the sum of all r eigenvalues of $\sum_{k=1}^m d_k P_1^T A_k P_1$ plus the sum of the $\kappa - r$ largest eigenvalues of $\sum_{k=1}^m d_k Q_1^T A_k Q_1$. Note that, unless $\kappa = r + t$, this is generally not the same as the sum of the κ largest eigenvalues of

$$\sum_{k=1}^m d_k \begin{bmatrix} P_1^T \\ Q_1^T \end{bmatrix} A_k [P_1 \;\; Q_1].$$

Remark: Classical eigenvalue perturbation theory (Kato(1982)) states that, in general, as $A(x)$ is perturbed by $\alpha \sum d_k A_k + o(\alpha)$, the multiple eigenvalue $\lambda_{r+1} = \cdots = \lambda_{r+t}$ splits, with the t perturbed eigenvalues having first-order changes given, respectively, by the t eigenvalues of $B(d)$. This, together with the fact that the sum of the first r eigenvalues, being separated from λ_{r+1}, is a smooth function, provides an alternative proof of (3.50). However, the proof of these classical results is by no means straightforward.

3.6 Splitting multiple eigenvalues

Given x, it is desired to either (a) generate a descent direction for f_κ, or (b) demonstrate that x satisfies the first-order condition for optimality. If $\kappa = r + t$ then $f_\kappa(x)$ is differentiable; consequently it is sufficient to examine the gradient, whose entries are given by (3.29). If the gradient is zero, the first-order optimality conditions hold; conversely if it is not zero, the negative gradient provides a descent direction. We therefore consider only the nonsmooth case $\kappa < r + t$ in the remainder of this section, although the following results actually apply in general (with a slight modification in the second case). When a descent direction exists, we are not particularly interested in obtaining the steepest descent direction, since it is generally advantageous to maintain the correct multiplicity when possible (by analogy with active set methods for constrained optimization). In the first of the next three cases a descent direction is generated keeping λ_κ of multiplicity t (to first order), while in the second case, generation of a descent direction requires splitting the group of eigenvalues corresponding to λ_κ.

Case 1. $I \in \text{Span}\{ B_1, \ldots, B_m \}$.

Solve the system

$$\delta I - \sum_{k=1}^m d_k B_k \;\; = \;\; 0 \tag{3.51}$$

$$(\kappa - r)\delta + \sum_{k=1}^{m} d_k b_k \;=\; -1. \tag{3.52}$$

This is a system of $t(t+1)/2 + 1$ linear equations in $m+1$ unknowns $\delta, d_1, \ldots, d_m$. Equation (3.51) implies that the eigenvalues of $B(d)$ defined by (3.49) are all equal to δ. The system is solvable since (3.51) is solvable for any δ by assumption, and (3.52) scales this solution. Hence, from equations (3.50) and (3.52), $f'_\kappa(x; d) = -1$, where the direction $d \in \mathbf{R}^m$ has components $d_1, \ldots, d_m$. Note that the -1 on the right-hand side of (3.52) is just a normalization constant and can be replaced by any $\eta < 0$ giving $f'_\kappa(x; d) = \eta$. To first order all the eigenvalues $\lambda_{r+1}(x), \ldots, \lambda_{r+t}(x)$ decrease at the same rate along d, and δ gives a first order estimate of the change in their common value. Case 1 holds generically if $m \geq t(t+1)/2$, that is the generic dimension of the manifold defined by

$$\lambda_{r+1}(x) = \cdots = \lambda_{r+t}(x) \tag{3.53}$$

is greater than zero (see Overton and Womersley (1988), Section 2 for more detail).

Case 2. Case 1 does not apply and the span of the $m+1$ vectors in $\mathbf{R}^{t(t+1)/2}$ associated with $I, B_1, \ldots, B_m$ has the maximum dimension $t(t+1)/2$.

Solve the linear system

$$\begin{aligned} \mathrm{tr}(\tilde{U}) &= \kappa - r & (3.54)\\ -\langle\, B_k, \tilde{U} \,\rangle &= b_k, \quad k = 1, \ldots, m & (3.55) \end{aligned}$$

for the dual matrix $\tilde{U} \in \mathcal{S}_t$. Note that the trace condition (3.54) is equivalent to $\langle\, I, \tilde{U} \,\rangle = \kappa - r$. Since the $\{B_k\}$ may not form a linearly independent set, (3.55) may be replaced by considering only a maximal independent set of $\{B_k\}$; the system cannot be inconsistent because of the related definitions of the left and right-hand sides (B_k and b_k). By the rank assumption, the resulting linear system is square and nonsingular, with order $t(t+1)/2$, and having a unique solution $\tilde{U}$. If $\tilde{U}$ satisfies $0 \leq \tilde{U} \leq I$ then $0 \in \partial f_\kappa(x)$, so x satisfies the first-order necessary conditions for a minimum. If these inequalities on $\tilde{U}$ are not satisfied then the following result shows how to generate a descent direction.

Theorem 3.12 *Suppose (3.54) and (3.55) are satisfied but $0 \notin \partial f_\kappa(x)$, so $\tilde{U}$ has an eigenvalue θ outside $[0, 1]$. Let $z \in \mathbf{R}^t$ be the corresponding normalized eigenvector of $\tilde{U}$. Choose $\beta \in \mathbf{R}$ so that $\beta < 0$ if $\theta > 1$ and $\beta > 0$ if $\theta < 0$. Solve*

$$\delta I - \sum_{k=1}^{m} d_k B_k = \beta z z^T. \tag{3.56}$$

Then $d = [d_1 \cdots, d_m]^T$ is a descent direction.

Proof: The linear system (3.56) is solvable by hypothesis, although d is unique only if the $\{B_k\}$ are independent. Note that the coefficient matrix of the left hand side (3.56) is the transpose of that for the system (3.54), (3.55). Taking the inner product of (3.56) with $\tilde{U}$ gives

$$\delta \ \mathrm{tr}(\tilde{U}) - \sum_{k=1}^{m} d_k \langle \, B_k, \tilde{U} \, \rangle = \beta \langle \, zz^T, \tilde{U} \, \rangle, \tag{3.57}$$

so from (3.54), (3.55) and the spectral decomposition of $\tilde{U}$

$$\delta(\kappa - r) + b^T d = \beta\theta. \tag{3.58}$$

From (3.56) $B(d) = \delta I - \beta zz^T$ has eigenvalues $\delta - \beta, \delta, \ldots, \delta$. If $\beta < 0$ the sum of the $\kappa - r$ largest eigenvalues of $B(d)$ is $\delta(\kappa - r) - \beta$. Hence (3.50) and (3.58) give

$$f'_\kappa(x; d) = \beta(\theta - 1). \tag{3.59}$$

Thus if $\theta > 1$ choosing $\beta < 0$ and solving (3.56) produces a descent direction. If $\beta > 0$ then, since $\kappa < r + t$, the sum of the $\kappa - r$ largest eigenvalues of $B(d)$ is $\delta(\kappa - r)$. Then (3.50) and (3.58) give

$$f'_\kappa(x; d) = \beta\theta. \tag{3.60}$$

Therefore, if $\theta < 0$, choosing $\beta > 0$ and solving (3.56) produces a descent direction. ∎

Remark: For the case $\kappa = 1$ this result was given in Overton (1988) and Overton and Womersley (1988). As noted earlier the condition $\tilde{U} \leq I$ is not required when $\kappa = 1$.

Remark: Progress is made by splitting the multiple eigenvalue while maintaining multiplicity $t - 1$ to first order, i.e., the first-order change in all the eigenvalues but one in the cluster is the same. This is analogous to moving off a single active nonlinear constraint in the context of constrained optimization. The dual matrix $\tilde{U}$ provides the information which leads to the generation of a descent direction, just as negative Lagrange multipliers provide similar information in constrained optimization. The distinction between the cases $\theta < 0$ and $\theta > 1$ is as follows: if $\theta < 0$, one eigenvalue in the group of multiplicity t is separated from the others by a **reduction**, reducing the approximate multiplicity but leaving the number of eigenvalues larger than λ_κ, to first order, unchanged. If $\theta > 1$ one eigenvalue is separated from the others by an **increase**, again reducing the approximate multiplicity but increasing the number of larger eigenvalues (to first order). In either case the theorem guarantees an overall reduction in f_κ.

Case 2 applies generically if $m+1 = t(t+1)/2$, i.e. the generic dimension of the manifold defined by (3.53) consists of a single point. It also applies if x minimizes f_κ on this manifold; see Overton (1988) for a further explanation.

Case 3. Neither of Cases 1 and 2 apply. In this case degeneracy is said to occur. Generation of a descent direction is not straightforward.

3.7 Model algorithms

Practical algorithms for minimizing $f_\kappa(x)$ based on successive linear or quadratic programming may be defined to fully exploit the structure of the generalized gradient of f_κ. Such algorithms have been described and tested extensively in the case $\kappa = 1$ by Overton (1988, 1990).

Suppose that x^* is a (local) minimizer of $f_\kappa(x)$, with corresponding values r^* and t^* defined by (3.8). The model algorithm must use estimates, say r and t, of r^* and t^*. Note that, in general, the matrix iterates generated by the algorithm will have eigenvalues which are strictly multiple only in the limit at $x = x^*$. The simplest way to estimate r and t is to use an eigenvalue separation tolerance. The basic step of a model algorithm for minimizing $f_\kappa(x)$ then becomes the solution of the following quadratic program:

$$\min_{d \in \mathbf{R}^m, \delta \in \mathbf{R}} \quad (\kappa - r)\delta + b^T d + \tfrac{1}{2} d^T H d \tag{3.61}$$

$$\text{Subject to} \quad \delta I - \sum_{k=1}^{m} d_k Q_1^T A_k Q_1 \;=\; \operatorname{diag}(\lambda_{r+1}, \ldots, \lambda_{r+t}). \tag{3.62}$$

Here H is some positive semi-definite matrix and all the quantities Q_1, A_k, λ_i and b (defined in (3.48)) are evaluated at the current point x. The next trial point is $x + d$, and δ gives an estimate of $\lambda_\kappa(x + d)$.

Briefly, equation (3.62) represents the appropriate linearization of the nonlinear system

$$\lambda_{r+1}(x + d) = \cdots = \lambda_{r+t}(x + d) = \delta; \tag{3.63}$$

since this system is not differentiable, this needs justification (Friedland, Nocedal and Overton(1987)). The first two terms in the objective function (3.61) represent a linearization of $f_\kappa(x + d)$, while the third term may be used to incorporate second derivative information. The Lagrange multipliers corresponding to the $t(t + 1)/2$ equality constraints (3.62) make up a dual matrix estimate of $\tilde{U}$.

Inequalities may be included in the quadratic program, to ensure that linearizations of the first r eigenvalues (or at least their average) are no smaller than δ, and that the linearizations of $\lambda_{r+t+1}, \ldots$, are no larger than δ. A trust region constraint may be used to ensure reduction of the objective function at the trial point. It is

possible to choose H so that local quadratic convergence takes place; alternatively, H can be set to zero for a first-order method using a linear programming subproblem. Further details are beyond the scope of this paper, but see Overton(1990) for the case $\kappa = 1$.

4 Sum of the largest eigenvalues in absolute value

We are now interested in functions of the form

$$g_\kappa(A) \;=\; \sum_{i=1}^{\kappa} |\mu_i| \tag{4.1}$$

where the eigenvalues μ_i of A are ordered by

$$|\mu_1| \geq \cdots \geq |\mu_n|. \tag{4.2}$$

Let $\zeta = |\mu_\kappa|$, the κth largest eigenvalue modulus.

One approach is to apply the results of the previous section to minimizing the sum of the κ largest eigenvalues of

$$\begin{bmatrix} A(x) & 0 \\ 0 & -A(x) \end{bmatrix}.$$

However not only is the size of the problem then doubled, but the structure of the subdifferential is lost.

The techniques in Section 4 are related to the well known idea of representing a scalar α by its positive part $\alpha_+ = \max\{0, \alpha\}$ and its negative part $\alpha_- = \max\{0, -\alpha\}$, so $\alpha = \alpha_+ - \alpha_-$ and $|\alpha| = \alpha_+ + \alpha_-$. For $w \in \mathbf{R}^n$ the vectors w_+ and w_- are defined componentwise. This is discussed further in the Appendix.

4.1 A max characterization

This section establishes a max characterization of the κ largest, in absolute value, eigenvalues of a matrix, and the elements which achieve the maximum in this characterization. Let $\kappa \in \{1, \ldots, n\}$, and define

$$\begin{aligned} \Psi_{n,\kappa} \;=\; & \{\, W \in \mathcal{S}_n : W = U - V, \text{ where } U, V \in \mathcal{S}_n, \\ & 0 \leq U \leq I,\; 0 \leq V \leq I,\; \mathrm{tr}(U) + \mathrm{tr}(V) = \kappa \,\} \end{aligned} \tag{4.3}$$

and

$$\begin{aligned} \psi_{n,\kappa} \;=\; & \{\, w \in \mathbf{R}^n :\; w = u - v, \text{ where } u, v \in \mathbf{R}^n, \\ & 0 \leq u \leq e,\; 0 \leq v \leq e \text{ for } i = 1, \ldots, n, \text{ and } e^T(u + v) = \kappa \,\}. \end{aligned} \tag{4.4}$$

Lemma 4.1 *$\Psi_{n,\kappa}$ and $\psi_{n,\kappa}$ are compact convex sets, $\Psi_{n,\kappa}$ is invariant under orthogonal similarity transformations (i.e. $W \in \Psi_{n,\kappa} \iff Z^T W Z \in \Psi_{n,\kappa}$ for any $Z \in \mathcal{O}_{n,n}$), and $\Psi_{n,\kappa}$ and $\psi_{n,\kappa}$ are related by*

$$\begin{aligned}\Psi_{n,\kappa} \;=\; \{\, W \in \mathcal{S}_n : \quad & W = U - V, \quad U = ZDZ^T, \quad V = YEY^T \\ & \text{where } Z, Y \in \mathcal{O}_{n,n}, \quad D = \operatorname{diag}(u), \quad E = \operatorname{diag}(v), \\ & 0 \le u \le e,\ 0 \le v \le e,\ i = 1, \ldots, n \text{ and } e^T(u+v) = \kappa \,\} \end{aligned} \tag{4.5}$$

and

$$\psi_{n,\kappa} \;=\; \{\, w \in \mathbf{R}^n : w_i = W_{ii} \text{ for } i = 1, \ldots, n \text{ where } W \in \Psi_{n,\kappa} \,\}. \tag{4.6}$$

Proof: It is easily verified that $\Psi_{n,\kappa}$ and $\psi_{n,\kappa}$ are compact convex sets. A spectral decomposition of the matrices $U, V \in \mathcal{S}_n$ in the definition of $\Psi_{n,\kappa}$ shows that

$$\begin{aligned}\Psi_{n,\kappa} \;=\; \{\, W \in \mathcal{S}_n : \quad & W = ZDZ^T - YEY^T \\ & \text{where } Z, Y \in \mathcal{O}_{n,n}, \quad D, E \in \mathcal{D}_n, \\ & 0 \le D \le I, \quad 0 \le E \le I \text{ and } \operatorname{tr}(D) + \operatorname{tr}(E) = \kappa \,\}, \end{aligned} \tag{4.7}$$

which establishes (4.5), and the invariance of $\Psi_{n,\kappa}$ to orthogonal similarity transformations. Let Ω denote the set on the right hand side of equation (4.6). Choosing $Z = Y = I$, $D = \operatorname{diag}(w_+)$ and $E = \operatorname{diag}(w_-)$ in (4.7) , for any $w \in \psi_{n,\kappa}$, shows that $\psi_{n,\kappa} \subseteq \Omega$. To establish the reverse inclusion let $w \in \Omega$, i.e. $w_i = W_{ii}$, where $W = U - V$ with U,V satisfying the conditions in (4.3). Define u and v by $u_i = U_{ii}$, $v_i = V_{ii}$. Then $e^T(u+v) = \operatorname{tr}(U) + \operatorname{tr}(V) = \kappa$, and $0 \le u \le e$, $0 \le v \le e$, by (4.3), i.e. $w \in \psi_{n,\kappa}$. ∎

To characterize the elements which achieve the maximum in the following results requires information about the multiplicity of the eigenvalues of A. Although the ordering (4.2) is useful to give the simple definition (4.1) for $g_\kappa(A)$, as well as $\zeta = |\mu_\kappa|$, we shall now revert to the standard ordering.

Consider the case when $\zeta > 0$ and let the eigenvalues of A be written

$$\begin{aligned} &\lambda_1 \ge \cdots \ge \lambda_{r_1} > \\ &\lambda_{r_1+1} = \cdots = \zeta = \cdots = \lambda_{r_1+t_1} > \\ &\lambda_{r_1+t_1+1} \ge \cdots \ge \lambda_{n-r_2-t_2} > \\ &\lambda_{n-r_2-t_2+1} = \cdots = -\zeta = \cdots = \lambda_{n-r_2} > \\ &\lambda_{n-r_2+1} \ge \cdots \ge \lambda_n, \end{aligned} \tag{4.8}$$

where r_1, t_1, r_2, t_2 are nonnegative integers. The eigenvalue equal to ζ has multiplicity t_1 and the eigenvalue equal to $-\zeta$ has multiplicity t_2. By assumption there is

at least one eigenvalue with modulus ζ, so $t_1 + t_2 \geq 1$. The number of eigenvalues greater than ζ is r_1, while the number of eigenvalues less than $-\zeta$ is r_2. Note that by definition

$$r_1 + r_2 + 1 \leq \kappa \leq r_1 + t_1 + r_2 + t_2. \tag{4.9}$$

Thus $r_1 = r_2 = 0$ if $\kappa = 1$. Also if $t_1 = 1$ and $t_2 = 0$, or $t_1 = 0$ and $t_2 = 1$ then $\kappa = r_1 + r_2 + 1$.

We have $\lambda_i > 0$ for $i = 1, \ldots, r_1 + t_1$ and $\lambda_i < 0$ for $i = n - r_2 - t_2 + 1, \ldots, n$. Thus

$$g_\kappa(A) = \sum_{i=1}^{\kappa} |\mu_i| = \sum_{i=1}^{r_1} \lambda_i \; - \sum_{i=n-r_2+1}^{n} \lambda_i \; + \; (\kappa - r_1 - r_2)\zeta. \tag{4.10}$$

Example 4.2 *Suppose* $n = 11$ *and* $\lambda^T = (5, 4, 4, 4, 2, -1, -4, -4, -6, -6, -7)$. *If* $\kappa = 2$ *or* 3 *then* $\zeta = 6, r_1 = 0, t_1 = 0, r_2 = 1$ *and* $t_2 = 2$. *If* $\kappa = 5, 6$, *or* 7 *then* $\zeta = 4, r_1 = 1, t_1 = 3, r_2 = 3$ *and* $t_2 = 2$.

If $\zeta = 0$ then the appropriate ordering is

$$\begin{aligned} &\lambda_1 \geq \cdots \geq \lambda_{r_1} > \\ &\lambda_{r_1+1} = \cdots = 0 = \cdots = \lambda_{r_1+t} > \\ &\lambda_{n-r_2+1} \geq \cdots \geq \lambda_n, \end{aligned} \tag{4.11}$$

and $n - r_2 = r_1 + t$. Only the situation where $\zeta > 0$ will be considered in the rest of this section. The modifications required when $\zeta = 0$ can be derived from the ordering (4.11).

We first give two lemmas which depend only on the orderings (4.2), (4.8) for a single set of n real numbers, and not on the fact that these are eigenvalues of a matrix A.

Lemma 4.3 *If the vector* $\lambda \in \mathbf{R}^n$ *is ordered as in (4.8) then*

$$\max_{w \in \psi_{n,\kappa}} \lambda^T w = \sum_{i=1}^{\kappa} |\mu_i|$$

and

$$\begin{aligned} \text{argmax}\{\, \lambda^T w : w \in \psi_{n,\kappa} \,\} \;\; = \;\; & \{\, w \in \mathbf{R}^n : \\ & w_i = 1 \;\text{ for }\; i = 1, \ldots, r_1, \\ & 0 \leq w_i \leq 1 \;\text{ for }\; i = r_1 + 1, \ldots, r_1 + t_1, \\ & w_i = 0 \;\text{ for }\; i = r_1 + t_1 + 1, \ldots, n - r_2 - t_2, \\ & -1 \leq w_i \leq 0 \;\text{ for }\; i = n - r_2 - t_2 + 1, \ldots, n - r_2, \\ & w_i = -1 \;\text{ for }\; i = n - r_2 + 1, \ldots, n \;\text{ and} \\ & \sum_{i=r_1+1}^{r_1+t_1} w_i \; - \sum_{i=n-r_2-t_2+1}^{n-r_2} w_i \;=\; \kappa - r_1 - r_2 \,\}. \end{aligned} \tag{4.12}$$

Proof: The result follows directly from (4.2), (4.8) and the definition (4.4) of $\psi_{n,\kappa}$. ∎

Lemma 4.4 *Let* $\Lambda = \mathrm{diag}(\lambda)$. *Then*

$$\max_{W \in \Psi_{n,\kappa}} \langle \Lambda, W \rangle = \sum_{i=1}^{\kappa} |\mu_i| \tag{4.13}$$

and

$$\begin{aligned}
&\mathrm{argmax} \{ \langle \Lambda, W \rangle : W \in \Psi_{n,\kappa} \} = \\
&\quad \{ W \in \mathcal{S}_n : W = U - V, \text{ where} \\
&\quad U = \mathrm{diag}(I, \tilde{U}, 0, 0, 0), \quad \tilde{U} \in \mathcal{S}_{t_1}, \quad 0 \le \tilde{U} \le I, \\
&\quad V = \mathrm{diag}(0, 0, 0, \tilde{V}, I), \quad \tilde{V} \in \mathcal{S}_{t_2}, \quad 0 \le \tilde{V} \le I, \\
&\quad \mathrm{tr}(\tilde{U}) + \mathrm{tr}(\tilde{V}) = \kappa - r_1 - r_2 \}.
\end{aligned} \tag{4.14}$$

Here the diagonal blocks of U *and* V *have dimensions* r_1, t_1, $n - r_1 - t_1 - t_2 - r_2$, t_2 *and* r_2 *respectively.*

Proof: As $\Lambda = \mathrm{diag}(\lambda)$

$$\langle \Lambda, W \rangle \;=\; \sum_{i=1}^{n} \lambda_i W_{ii}.$$

Equation (4.13) follows from Lemmas 4.1 and 4.3.

If W^* is any element of the right hand side of (4.14) then from (4.8) and (4.10)

$$\langle \Lambda, W^* \rangle \;=\; \sum_{i=1}^{r_1} \lambda_i + \zeta \mathrm{tr}(\tilde{U}) + \zeta \mathrm{tr}(\tilde{V}) - \sum_{i=n-r_2+1}^{n} \lambda_i \;=\; \sum_{i=1}^{\kappa} |\mu_i|. \tag{4.15}$$

Conversely let $W^* \in \mathrm{argmax} \{ \langle \Lambda, W \rangle : W \in \Psi_{n,\kappa} \}$. Then $w^* = (W^*_{11} \ldots W^*_{nn})^T \in \psi_{n,\kappa}$ satisfies $\lambda^T w^* = \sum_{i=1}^{\kappa} |\mu_i|$, and therefore also satisfies the properties given on the right-hand side of (4.12). Furthermore, W^* has the representation $W^* = U^* - V^*$, where $0 \le U^* \le I$, $0 \le V^* \le I$, $\mathrm{tr}(U^*) + \mathrm{tr}(V^*) = \kappa$, so

$$\begin{aligned}
&U^*_{ii} = 1, \quad V^*_{ii} = 0 \text{ for } i = 1, \ldots, r_1, \\
&0 \le U^*_{ii} \le 1, \quad V^*_{ii} = 0 \text{ for } i = r_1 + 1, \ldots, r_1 + t_1, \\
&U^*_{ii} = 0, \quad V^*_{ii} = 0 \text{ for } i = r_1 + t_1 + 1, \ldots, n - r_2 - t_2, \\
&U^*_{ii} = 0, \quad 0 \le V^*_{ii} \le 1 \text{ for } i = n - r_2 - t_2 + 1, \ldots, n - r_2, \\
&U^*_{ii} = 0, \quad V^*_{ii} = 1 \text{ for } i = n - r_2 + 1, \ldots, n \text{ and} \\
&\sum_{i=r_1+1}^{r_1+t_1} U^*_{ii} \;+ \sum_{i=n-r_2-t_2+1}^{n-r_2} V^*_{ii} = \kappa - r_1 - r_2 \}.
\end{aligned} \tag{4.16}$$

Partition the rows and columns of $W^* = U^* - V^*$ into blocks so

$$U^* = \begin{bmatrix} C_{11} & C_{12} & C_{13} & C_{14} & C_{15} \\ C_{12}^T & C_{22} & C_{23} & C_{24} & C_{25} \\ C_{13}^T & C_{23}^T & C_{33} & C_{34} & C_{35} \\ C_{14}^T & C_{24}^T & C_{34}^T & C_{44} & C_{45} \\ C_{15}^T & C_{25}^T & C_{35}^T & C_{45}^T & C_{55} \end{bmatrix}, \quad V^* = \begin{bmatrix} E_{11} & E_{12} & E_{13} & E_{14} & E_{15} \\ E_{12}^T & E_{22} & E_{23} & E_{24} & E_{25} \\ E_{13}^T & E_{23}^T & E_{33} & E_{34} & E_{35} \\ E_{14}^T & E_{24}^T & E_{34}^T & E_{44} & E_{45} \\ E_{15}^T & E_{25}^T & E_{35}^T & E_{45}^T & E_{55} \end{bmatrix}$$

where the dimensions of the symmetric matrices C_{ii} and E_{ii} for $i = 1, \ldots, 5$ are respectively r_1, t_1, $n - r_1 - t_1 - r_2 - t_2$, t_2 and r_2. As $W^* \in \Psi_{n,\kappa}$

$$\begin{aligned} &0 \le C_{ii} \le I, \quad i = 1, \ldots, 5 \\ &0 \le E_{ii} \le I, \quad i = 1, \ldots, 5 \\ &\sum_{i=1}^{5} \mathrm{tr}(C_{ii}) + \sum_{i=1}^{5} \mathrm{tr}(E_{ii}) = \kappa. \end{aligned}$$

As in Section 3 the proof is completed using the basic results (3.16) and (3.17) for positive semi-definite matrices. From (4.16) $C_{11} = I$, as $C_{11} \le I$ and the diagonal elements of C_{11} are all 1. Then as $U^* \le I$, $C_{11} = I$ implies that $C_{1j} = 0$ for all $j \ne 1$. For $i = 3, 4, 5$ $C_{ii} = 0$, as $C_{ii} \ge 0$ and the diagonal elements of C_{ii} are all zero. Then for $i = 3, 4, 5$ $U^* \ge 0$ and $C_{ii} = 0$ imply that $C_{ij} = 0$ for all $j \ne i$. Similarly $E_{ii} = 0$ for $i = 1, 2, 3$ and $E_{55} = I$. Then $0 \le V^* \le I$ implies that $E_{ij} = 0$ for all $i \ne j$. The matrices U^* and V^* now reduced to those given in (4.14). The trace condition comes directly from (4.16). ∎

Now consider sums of the eigenvalues of A, and recall that $Q \in \mathcal{O}_{n,n}$ is a matrix of eigenvectors for A satisfying

$$Q^T A Q = \Lambda. \tag{4.17}$$

Let $P_1 \in \mathcal{O}_{n,r_1}, P_2 \in \mathcal{O}_{n,r_2}$ be the matrices consisting of the first r_1 and last r_2 columns of Q respectively. Also let $Q_1 \in \mathcal{O}_{n,t_1}$ be the matrix consisting of columns $r_1 + 1, \ldots, r_1 + t_1$ of Q, and let $Q_2 \in \mathcal{O}_{n,t_2}$ be the matrix consisting of columns $n - r_2 - t_2 + 1, \ldots, n - r_2$ of Q. By (4.8) we have

$$P_1^T A P_1 = \mathrm{diag}(\lambda_1, \ldots, \lambda_{r_1}), \quad P_2^T A P_2 = \mathrm{diag}(\lambda_{n-r_2+1}, \ldots, \lambda_n), \tag{4.18}$$

and

$$Q_1^T A Q_1 = \zeta I, \quad Q_2^T A Q_2 = -\zeta I. \tag{4.19}$$

Theorem 4.5

$$\max_{W \in \Psi_{n,\kappa}} \langle A, W \rangle \;=\; g_\kappa(A), \tag{4.20}$$

with

$$\begin{aligned}
&\operatorname{argmax} \{ \langle A, W \rangle : W \in \Psi_{n,\kappa} \} = \\
&\quad \{ W \in \mathcal{S}_n : W = U - V, \text{ where } U, V \in \mathcal{S}_n \\
&\quad U = P_1 P_1^T + Q_1 \tilde{U} Q_1^T, \quad V = P_2 P_2^T + Q_2 \tilde{V} Q_2^T, \\
&\quad \tilde{U} \in \mathcal{S}_{t_1}, 0 \le \tilde{U} \le I, \quad \tilde{V} \in \mathcal{S}_{t_2}, 0 \le \tilde{V} \le I \\
&\quad \text{and } \operatorname{tr}(\tilde{U}) + \operatorname{tr}(\tilde{V}) = \kappa - r_1 - r_2 \}.
\end{aligned} \tag{4.21}$$

Proof: For any $W \in \Psi_{n,\kappa}$ equation (4.17) and the properties of the Frobenius inner product imply that

$$\langle A, W \rangle \;=\; \langle Q \Lambda Q^T, W \rangle \;=\; \langle \Lambda, Q^T W Q \rangle, \tag{4.22}$$

with $Q^T W Q \in \Psi_{n,\kappa}$ as $\Psi_{n,\kappa}$ is invariant to orthogonal similarity transformations. Hence

$$\max_{W \in \Psi_{n,\kappa}} \langle A, W \rangle = \max_{W \in \Psi_{n,\kappa}} \langle \Lambda, W \rangle = g_\kappa(A),$$

where the second equality follows from Lemma 4.4.

From (4.22) and the invariance of $\Psi_{n,\kappa}$ to orthogonal transformations

$$\operatorname{argmax} \{ \langle A, W \rangle : W \in \Psi_{n,\kappa} \} = \{ Q W^* Q^T : W^* \in \Omega \}$$

where $\Omega = \operatorname{argmax} \{ \langle \Lambda, W \rangle : W \in \Psi_{n,\kappa} \}$. The proof is completed by applying Lemma 4.4 since

$$\begin{aligned}
Q W^* Q^T &= Q \operatorname{diag}(I, \tilde{U}, 0, 0, 0)\, Q^T - Q \operatorname{diag}(0, 0, 0, \tilde{V}, I)\, Q^T \\
&= P_1 P_1^T + Q_1 \tilde{U} Q_1^T - Q_2 \tilde{V} Q_2^T - P_2 P_2^T.
\end{aligned}$$

∎

Remark: If $\kappa = r_1 + r_2 + t_1 + t_2$ then $\tilde{U} = I$, $\tilde{V} = I$ and

$$\operatorname{argmax} \{ \langle A, W \rangle : W \in \Psi_{n,\kappa} \} \;=\; P_1 P_1^T + Q_1 Q_1^T - Q_2 Q_2^T - P_2 P_2^T. \tag{4.23}$$

The matrix achieving the maximum in (4.13) is unique only in this case. It is precisely the case when $g_\kappa(A)$ is a smooth function of the elements of A.

Remark: If $t_1 = 1, t_2 = 1,$ and $\kappa = r_1 + r_2 + 1$, then $\tilde{U} = \alpha$, $\tilde{V} = \beta$ where $\alpha, \beta \in \mathbf{R}$ satisfy $\alpha, \beta \in [0, 1]$ and $\alpha + \beta = \kappa - r_1 - r_2 = 1$. Also

$$\operatorname{argmax} \{ \langle A, W \rangle : W \in \Psi_{n,\kappa} \} \;=\; P_1 P_1^T + \alpha Q_1 Q_1^T - \beta Q_2 Q_2^T - P_2 P_2^T; \tag{4.24}$$

here Q_1 and Q_2 have only one column. This is precisely the case when $g_\kappa(A)$ is nonsmooth but is the maximum of a **finite** number of smooth functions (λ_{r_1+1} and $-\lambda_{n-r_2}$).

Remark: When $t_1 > 1$ or $t_2 > 1$ all the flexibility in the choices of Q_1 and Q_2 in (4.19) is absorbed into the matrices $\tilde{U}$ and $\tilde{V}$. It makes no difference if any of the eigenvalues $\lambda_1, \ldots, \lambda_{r_1}$ or $\lambda_{n-r_2+1}, \ldots, \lambda_n$ have multiplicity greater than 1, as **all** these multiple eigenvalues are included in $g_\kappa(A)$ and the corresponding columns of P_1 or P_2 can be any orthonormal basis for the corresponding eigenspace.

Corollary 4.6 *The function $g_\kappa(A) : \mathcal{S}_n \rightarrow \mathbf{R}$ is convex and its subdifferential $\partial g_\kappa(A)$ is the nonempty compact convex set*

$$\begin{aligned} \partial g_\kappa(A) \;=\; & \{\, W \in \mathcal{S}_n : \text{there exists} \\ & \tilde{U} \in \mathcal{S}_{t_1} \;\text{ with }\; 0 \le \tilde{U} \le I, \\ & \tilde{V} \in \mathcal{S}_{t_2} \;\text{ with }\; 0 \le \tilde{V} \le I, \\ & \mathrm{tr}(\tilde{U}) + \mathrm{tr}(\tilde{V}) = \kappa - r_1 - r_2, \\ & W = P_1 P_1^T + Q_1 \tilde{U} Q_1^T - Q_2 \tilde{V} Q_2^T - P_2 P_2^T \,\}. \end{aligned} \tag{4.25}$$

Proof: For any $W \in \Psi_{n,\kappa}$ the inner product $\langle\, A, W \,\rangle$ is a linear function of $A \in \mathcal{S}_n$, so from Rockafellar (1970) $g_\kappa(A)$ is a convex function of A. Moreover the subdifferential is

$$\partial g_\kappa(A) \;=\; \mathrm{conv}\,\{\, W \in \mathcal{S}_n : \langle\, A, W \,\rangle = g_\kappa(A) \,\}.$$

The result follows from Theorem 4.5 as the set on the right hand side of (4.25) is already convex. ∎

As in Section 3 there are various equivalent forms for the argmax involving a convex hull operation, which are discussed in the Appendix. These lead to equivalent, but computationally less useful, expressions for $\partial g_\kappa(A)$.

4.2 The generalized gradient

As in Section 3.2 let $A(x) : \mathbf{R}^m \rightarrow \mathcal{S}_n$ be a continuously differentiable function with partial derivatives

$$A_k(x) \;\equiv\; \frac{\partial A(x)}{\partial x_k} \;\text{ for }\; k = 1, \ldots, m.$$

This section is concerned with finding a computationally useful characterization of the generalized gradient of the function

$$g_\kappa(x) \;=\; g_\kappa(A(x)).$$

The eigenvalues of $A(x)$ are, as before, denoted by by both (4.2) and (4.8), with r_1, t_1, r_2, t_2 now dependent on x, and with the corresponding eigenvectors satisfying (4.18), (4.19). Thus the explicit dependence of the eigensystem of $A(x)$ on x is generally omitted.

Theorem 4.7 *The function $g_\kappa(x)$ is locally Lipschitz, subdifferentially regular, and its generalized gradient is the nonempty compact convex set*

$$\begin{aligned} \partial g_\kappa(x) \;=\; & \{\, w \in \mathbf{R}^m : \exists\, \tilde{U} \in \mathcal{S}_{t_1},\ \tilde{V} \in \mathcal{S}_{t_2} \text{ with} && (4.26)\\ & 0 \le \tilde{U} \le I,\ 0 \le \tilde{V} \le I,\ \mathrm{tr}(\tilde{U}) + \mathrm{tr}(\tilde{V}) = \kappa - r_1 - r_2, \\ & \text{and for } k = 1, \ldots, m \\ & w_k \;=\; \mathrm{tr}(P_1^T A_k(x) P_1) \;+\; \langle\, Q_1^T A_k(x) Q_1, \tilde{U} \,\rangle \\ & -\; \mathrm{tr}(P_2^T A_k(x) P_2) \;-\; \langle\, Q_2^T A_k(x) Q_2, \tilde{V} \,\rangle \,\}. && (4.27) \end{aligned}$$

Proof: By the Clarke chain rule, $g_\kappa(x)$ is locally Lipschitz, subdifferentially regular, and

$$\partial g_\kappa(x) = \{\, w \in \mathbf{R}^m : w_k \;=\; \langle\, A_k(x), W \,\rangle, : W \in \partial g_\kappa(A(x)) \,\}.$$

Corollary 4.6 and the properties of the inner product complete the proof. ∎

Corollary 4.8 *If $\kappa = r_1 + t_1 + r_2 + t_2$ then the function $g_\kappa(x)$ is differentiable at x with*

$$\begin{aligned} \frac{\partial g(x)}{\partial x_k} \;=\; & \mathrm{tr}(P_1^T A_k(x) P_1) + \mathrm{tr}(Q_1^T A_k(x) Q_1) \\ & -\mathrm{tr}(Q_2^T A_k(x) Q_2) - \mathrm{tr}(P_2^T A_k(x) P_2). && (4.28) \end{aligned}$$

Proof: This is precisely the case when (4.21) and hence $\partial g_\kappa(x)$ is a singleton. Hence $g_\kappa(x)$ is differentiable, and (4.27) reduces to (4.28). ∎

4.3 Necessary conditions

If x is a local minimizer of g_κ then the standard necessary condition $0 \in \partial g_\kappa(x)$ and Theorem 4.7 imply there exist $\tilde{U} \in \mathcal{S}_{t_1}$ and $\tilde{V} \in \mathcal{S}_{t_2}$ such that

$$0 \le \tilde{U} \le I, \quad 0 \le \tilde{V} \le I, \tag{4.29}$$

$$\mathrm{tr}(\tilde{U}) + \mathrm{tr}(\tilde{V}) = \kappa - r_1 - r_2, \tag{4.30}$$

and for $k = 1, \ldots, m$

$$\begin{aligned} 0 \;\; = \;\; & \mathrm{tr}({P_1}^T A_k(x) P_1) \;+\; \langle\, {Q_1}^T A_k(x) Q_1, \tilde{U} \,\rangle \\ & - \mathrm{tr}({P_2}^T A_k(x) P_2) \;-\; \langle\, {Q_2}^T A_k(x) Q_2, \tilde{V} \,\rangle. \end{aligned} \tag{4.31}$$

These conditions are useful computationally as one can relax the inequalities on $\tilde{U}$ and $\tilde{V}$ and solve (4.30) and (4.31) for $\tilde{U}$ and $\tilde{V}$. This requires solving a system of $m+1$ linear equations for the $t_1(t_1+1)/2 + t_2(t_2+1)/2$ unknowns in the symmetric matrices $\tilde{U}$ and $\tilde{V}$. When $\kappa = 1$ these are the same conditions as those used in Overton (1988). If the inequalities $0 \leq \tilde{U} \leq I$ or $0 \leq \tilde{V} \leq I$ are not satisfied then this information can be used to generate a descent direction; see Section 4.6.

If $g_\kappa(x)$ is convex (e.g. if $A(x)$ is affine) then equations (4.29), (4.30) and (4.31) are both necessary and sufficient for x to be a minimizer of $g_\kappa(x)$.

4.4 A saddle point result

As in Section 3.4 it is possible to establish saddle point results for the Lagrangian function

$$\mathcal{L}(x, W) \;=\; \langle\, A(x), W \,\rangle, \tag{4.32}$$

where $W \in \Psi_{n,\kappa}$. The primal problem is

$$\min_{x \in \mathbf{R}^n} g_\kappa(x) \tag{4.33}$$

where

$$g_\kappa(x) = \max_{W \in \Psi_{n,\kappa}} \mathcal{L}(x, W). \tag{4.34}$$

The dual problem is

$$\max_{W \in \Psi_{n,\kappa}} h(W), \tag{4.35}$$

where

$$h(W) = \min_{x \in \mathbf{R}^m} \mathcal{L}(x, W). \tag{4.36}$$

Theorem 4.9 *For each $W \in \Psi_{n,\kappa}$ let $\mathcal{L}(., W)$ be a convex function, and let the primal problem attain its solution at x^*. Then the primal and dual problems have the same optimal value so*

$$\min_{x \in \mathbf{R}^m} \max_{W \in \Psi_{n,\kappa}} \mathcal{L}(x, W) = \max_{W \in \Psi_{n,\kappa}} \min_{x \in \mathbf{R}^m} \mathcal{L}(x, W), \tag{4.37}$$

and $W^* \in \mathrm{argmax}\{\, \langle\, A(x^*), W \,\rangle : W \in \Psi_{n,\kappa} \,\}$ satisfying $\langle\, A_k(x^*), W^* \,\rangle \;=\; 0$ for $k = 1, \ldots, m$ solves the dual problem.

Proof: Analogous to the proof of Theorem 3.11. ∎

4.5 The directional derivative

As $g_\kappa(x)$ is subdifferentially regular the standard one-sided directional derivative $g'_\kappa(x;d)$ at x in a direction $d \in \mathbf{R}^m$ exists and satisfies

$$\begin{aligned} g'_\kappa(x;d) &= \max_{w \in \partial g_\kappa(x)} w^T d \\ &= \sum_{k=1}^m d_k \ [\mathrm{tr}(P_1^T A_k P_1) - \mathrm{tr}(P_2^T A_k P_2)] + \\ &\quad \max_{\tilde{U},\tilde{V}} \sum_{k=1}^m d_k [\langle\, Q_1^T A_k Q_1, \tilde{U}\,\rangle \ - \ \langle\, Q_2^T A_k Q_2, \tilde{V}\,\rangle], \end{aligned} \tag{4.38}$$

where the maximum is over all $\tilde{U} \in \mathcal{S}_{t_1}$, $\tilde{V} \in \mathcal{S}_{t_2}$ satisfying

$$0 \le \tilde{U} \le I, \ 0 \le \tilde{V} \le I \ \text{ and } \ \mathrm{tr}(\tilde{U}) + \mathrm{tr}(\tilde{V}) = \kappa - r_1 - r_2. \tag{4.39}$$

Recall that the matrices P_1, P_2 and Q_1, Q_2 are evaluated at the point x, and that A_k is the partial derivative of $A(x)$ with respect to x_k evaluated at the point x. For $k = 1, \ldots, m$ define

$$b_k = \mathrm{tr}(P_1^T A_k P_1) \ - \ \mathrm{tr}(P_2^T A_k P_2), \tag{4.40}$$

and

$$B_k = \begin{bmatrix} Q_1^T A_k Q_1 & 0 \\ 0 & -Q_2^T A_k Q_2 \end{bmatrix}. \tag{4.41}$$

Note that $B_k \in \mathcal{S}_{t_1+t_2}$. Let

$$B(d) = \sum_{k=1}^m d_k B_k, \tag{4.42}$$

and let the eigenvalues of $B(d)$ be $\gamma_1, \ge \ldots \ge \gamma_{t_1+t_2}$. Also let

$$T = \begin{bmatrix} \tilde{U} & 0 \\ 0 & \tilde{V} \end{bmatrix}. \tag{4.43}$$

Then from Theorem 4.5 and (4.38) and (4.39) it follows that

$$g'_\kappa(x;d) = b^T d \ + \ \max_T \langle\, B(d), T\,\rangle$$

where the maximum is over all matrices T satisfying (4.39) and (4.43). Since $B(d)$ is block diagonal, the maximum may equivalently be taken over matrices $T \in \Phi_{t_1+t_2, \kappa-r_1-r_2}$. Hence the directional derivative is

$$g'_\kappa(x;d) = b^T d \ + \ \sum_{i=1}^{\kappa-r_1-r_2} \gamma_i, \tag{4.44}$$

the sum of all r_1 eigenvalues of $\sum_{k=1}^m d_k P_1^T A_k P_1$ minus the sum of all r_2 eigenvalues of $\sum_{k=1}^m d_k P_2^T A_k P_2$ plus the sum of the $\kappa - r_1 - r_2$ largest eigenvalues of $B(d)$.

4.6 Splitting multiple eigenvalues

This section discusses the generation of a descent direction from points where the first-order optimality conditions do not hold. In the first case a descent direction is generated keeping the eigenvalues corresponding to ζ of multiplicity t_1, and the eigenvalues corresponding to $-\zeta$ of multiplicity t_2 (to first order). The second case requires splitting at least one eigenvalue away from the common value $\pm\zeta$.

When $g_\kappa(x)$ is differentiable ($\kappa = r_1 + t_1 + t_2 + r_2$) and the gradient is nonzero its negative provides a descent direction. Therefore we only consider the nonsmooth case $\kappa < r_1 + t_1 + t_2 + r_2$.

Case 1. $I \in \text{Span}\{\ B_1, \ldots, B_m\ \}$.

Solve the system

$$\delta I - \sum_{k=1}^{m} d_k B_k \;=\; 0 \tag{4.45}$$

$$(\kappa - r_1 - r_2)\delta + \sum_{k=1}^{m} d_k b_k \;=\; -1. \tag{4.46}$$

Noting the structure of B_k in (4.41) this is a system of $t_1(t_1+1)/2 \;+\; t_2(t_2+1)/2$ linear equations in $m+1$ unknowns $\delta, d_1, \ldots, d_m$. Equation (4.45) implies that the eigenvalues of $B(d)$ defined by (4.42) are all equal to δ. Hence from equation (4.44) and (4.46) $g'_\kappa(x;d) = -1$ where the direction $d \in \mathbf{R}^m$ has components $d_1, \ldots, d_m$. To first order all the eigenvalues $\lambda_{r_1+1}(x), \ldots, \lambda_{r_1+t_1}(x)$ decrease at the same rate along d, and all the eigenvalues $\lambda_{n-r_2-t_2+1}(x), \ldots, \lambda_{n-r_2}(x)$ increase at the same rate along d. Thus δ gives a first order estimate of the change in ζ along d. This case holds generically if the manifold defined by

$$\lambda_{r_1+1}(x) = \cdots = \lambda_{r_1+t_1}(x) = \zeta \tag{4.47}$$

and

$$\lambda_{n-r_2-t_2+1}(x) = \cdots = \lambda_{n-r_2}(x) = -\zeta \tag{4.48}$$

has dimension greater than zero, so $m \geq t_1(t_1+1)/2 + t_2(t_2+1)/2$.

Case 2. Case 1 does not apply and $\{\ I, B_1, \ldots, B_m\ \}$ has full rank $t_1(t_1+1)/2 + t_2(t_2+1)/2$.

Solve the linear system

$$\text{tr}(\tilde{U}) + \text{tr}(\tilde{V}) = \kappa - r_1 - r_2 \tag{4.49}$$

$$-\langle\, B_k, T\,\rangle = b_k \quad k = 1, \ldots, m \tag{4.50}$$

for the dual matrices $\tilde{U} \in \mathcal{S}_{t_1}$ and $\tilde{V} \in \mathcal{S}_{t_2}$, where T is defined by (4.43). This is a system of $m+1$ equations in $t_1(t_1+1)/2 + t_2(t_2+1)/2$ unknowns. A similar

argument to that given in Section 3.6 shows that a unique solution exists. If $\tilde{U}$ satisfies $0 \leq \tilde{U} \leq I$ and $\tilde{V}$ satisfies $0 \leq \tilde{V} \leq I$ then $0 \in \partial g_\kappa(x)$, so x is a stationary point.

Theorem 4.10 *Suppose (4.49) and (4.50) are satisfied but $0 \notin \partial g_\kappa(x)$, so either $\tilde{U}$ or $\tilde{V}$ has an eigenvalue outside $[0,1]$.*

If an eigenvalue θ of $\tilde{U}$ lies outside $[0,1]$, let $z \in \mathbf{R}^{t_1}$ be the corresponding normalized eigenvector of $\tilde{U}$. Let $\beta \in \mathbf{R}$ and solve

$$\delta I - \sum_{k=1}^{m} d_k B_k = \begin{bmatrix} \beta z z^T & 0 \\ 0 & 0 \end{bmatrix}. \tag{4.51}$$

Alternatively if an eigenvalue η of $\tilde{V}$ lies outside $[0,1]$, let $y \in \mathbf{R}^{t_2}$ be the corresponding normalized eigenvector of $\tilde{V}$. Let $\beta \in \mathbf{R}$ and solve

$$\delta I - \sum_{k=1}^{m} d_k B_k = \begin{bmatrix} 0 & 0 \\ 0 & \beta y y^T \end{bmatrix}. \tag{4.52}$$

Then β can be chosen so $d = [d_1 \cdots, d_m]^T$ is a descent direction.

Proof: First consider the case when an eigenvalue θ of $\tilde{U}$ lies outside $[0,1]$. The linear system (4.51) is solvable by hypothesis, although d is not unique if $\{B_k\}, k = 1, \ldots, m$, are linearly dependent. Taking an inner product of (4.51) with T defined in (4.43) gives

$$\delta \operatorname{tr}(T) - \sum_{k=1}^{m} d_k \langle B_k, T \rangle = \beta \langle z z^T, \tilde{U} \rangle,$$

so from (4.49) and (4.50)

$$\delta(\kappa - r_1 - r_2) + b^T d = \beta\theta. \tag{4.53}$$

From (4.42) and (4.51)

$$B(d) = \delta I - \begin{bmatrix} \beta z z^T & 0 \\ 0 & 0 \end{bmatrix}$$

has $t_1 + t_2$ eigenvalues $\delta - \beta, \delta, \ldots, \delta$.

If $\beta < 0$, the sum of the $\kappa - r_1 - r_2$ largest eigenvalues of $B(d)$ is $\delta(\kappa - r_1 - r_2) - \beta$. Hence (4.44) and (4.53) give

$$g'_\kappa(x; d) = \beta(\theta - 1) \tag{4.54}$$

Thus if $\theta > 1$ choosing $\beta < 0$ and solving (4.51) produces a descent direction.

On the other hand, if $\beta > 0$, the sum of the $\kappa - r_1 - r_2$ largest eigenvalues of $B(d)$ is $\delta(\kappa - r_1 - r_2)$. Then (4.44) and (4.53) give

$$g'_\kappa(x; d) = \beta\theta. \tag{4.55}$$

Thus if $\theta < 0$ choosing $\beta > 0$ and solving (4.51) produces a descent direction.

The proof follows the same lines when an eigenvalue η of $\tilde{V}$ lies outside $[0, 1]$ and (4.52) is solved instead of (4.51). ∎

Remark: In general, the system (4.51) splits the multiple eigenvalue $\lambda_{r_1+1} = \ldots = \lambda_{r_1+t_1} = \zeta$, to first order, reducing the approximate multiplicity by one, while (4.52) splits $\lambda_{n-r_2-t_2+1} = \ldots = \lambda_{n-r_2} = -\zeta$. A special case occurs when $t_1 = t_2 = 1$ (and consequently, by assumption, $\kappa = r_1 + r_2 + 1$). In this case both eigenvalues already have multiplicity one, but the direction d splits the common value $\zeta = \lambda_{r_1+1} = -\lambda_{n-r_2}$.

Remark: For the case $\kappa = 1$ this result was given by Overton (1988). In the previous work with $\kappa = 1$ the inequalities $\tilde{U} \leq I$ and $\tilde{V} \leq I$ did not appear. This is because $\kappa = 1$ implies $r_1 = r_2 = 0$ and the conditions $\mathrm{tr}(\tilde{U}) + \mathrm{tr}(\tilde{V}) = 1$, $\tilde{U} \geq 0$ and $\tilde{V} \geq 0$ imply that $\tilde{U} \leq I$ and $\tilde{V} \leq I$.

Case 3. Neither of Cases 1 or 2 apply. In this case degeneracy is said to occur. Generation of a descent direction is not straightforward.

4.7 Model algorithms

As in Section 3.7 algorithms for minimizing $g_\kappa(x)$ may be developed based on successive linear or quadratic programming.

Suppose that x^* is a (local) minimizer of $g_\kappa(x)$, with corresponding values r_1^*, r_2^*, t_1^* and t_2^* (defined by (4.8)). The model algorithm must use estimates, say r_1, r_2, t_1 and t_2, of these quantities. Note that, in general, the matrix iterates generated by the algorithm will have eigenvalues which are strictly multiple only in the limit at $x = x^*$.

The basic step of a model algorithm for minimizing $g_\kappa(x)$ then becomes the solution of the following linear or quadratic program:

$$\min_{d \in \mathbf{R}^m, \delta \in \mathbf{R}} \quad (\kappa - r_1 - r_2)\delta + b^T d + \tfrac{1}{2} d^T H d \tag{4.56}$$

$$\delta I - \sum_{k=1}^{m} d_k Q_1^T A_k Q_1 \;=\; \mathrm{diag}(\lambda_{r_1+1}, \ldots, \lambda_{r_1+t_1}) \tag{4.57}$$

$$\delta I + \sum_{k=1}^{m} d_k Q_2^T A_k Q_2 \;=\; -\mathrm{diag}(\lambda_{n-r_2-t_2+1}, \ldots, \lambda_{n-r_2}) \tag{4.58}$$

$$\tag{4.59}$$

Here the quantities Q_1, Q_2, λ_i and b (defined by (4.40)) are evaluated at the current point x. The new trial point is $x + d$, and δ is an estimate of ζ evaluated at the point $x + d$.

Equation (4.57) represents the appropriate linearization of the nonlinear system

$$\lambda_{r_1+1}(x+d) = \cdots = \lambda_{r_1+t_1}(x+d) = \delta, \tag{4.60}$$

while (4.58) is the linearization of

$$\lambda_{n-r_2-t_2+1}(x+d) = \cdots = \lambda_{n-r_2}(x+d) = -\delta. \tag{4.61}$$

The first two terms in the objective function (4.56) represent a linearization of $g_\kappa(x+d)$, while the third term may be used to incorporate second derivative information. The Lagrange multipliers corresponding to the $t_1(t_1+1)/2$ equality constraints (4.57) make up a dual matrix estimate of $\tilde{U}$, while the Lagrange multipliers corresponding to the $t_2(t_2+1)/2$ equality constraints (4.58) make up a dual matrix estimate of $\tilde{V}$.

Inequalities may be added to the quadratic program to ensure that linearizations of the inequalities $\lambda_i(x+d) \geq \delta$ for $i = 1, \ldots, r_1$, $-\delta \leq \lambda_i(x+d) \leq \delta$ for $i = r_1+t_1+1, \ldots, n-r_2-t_2$ and $\lambda_i(x+d) \leq -\delta$ for $i = n-r+2+1, \ldots, n$ are satisfied. A trust region constraint may be added to ensure a reduction in the objective function at the trial point. See Overton (1990) for further explanation of these ideas.

A Appendix

This Appendix discusses further properties of the sets $\psi_{n,\kappa}$ and $\Psi_{n,\kappa}$ which, while not needed in the derivation of the subdifferential of $g_\kappa(A)$ given in Section 4.1, are of some independent interest. They lead to an alternative representation for the subdifferential of $g_\kappa(A)$, in much the same way that Fan's theorem was obtained in Section 3.1 as a corollary of our main results and led to (3.28), the alternative form for the subdifferential of $f_\kappa(A)$.

As already mentioned, the techniques in Section 4 are related to the idea of representing a scalar α by its positive part $\alpha_+ = \max\{0, \alpha\}$ and its negative part $\alpha_- = \max\{0, -\alpha\}$, so $\alpha = \alpha_+ - \alpha_-$ and $|\alpha| = \alpha_+ + \alpha_-$. The same definition applies to a vector componentwise, and, as will be seen, can also be extended in a natural way to matrices.

Consider the set

$$\xi_{n,\kappa} = \{\, w \in \mathbf{R}^n : \; -e \leq w \leq e, \quad \sum_{i=1}^{n} |w_i| = \kappa \,\}. \tag{A.1}$$

It is immediately apparent that $\xi_{n,\kappa} \subseteq \psi_{n,\kappa}$, since for any $w \in \xi_{n,\kappa}$, setting $u = w_+$, $v = w_-$ satisfies the requirements $w = u - v$, $0 \leq u \leq e$, $0 \leq v \leq e$, and $e^T(u+v) = \kappa$. However, the example $n = \kappa = 1$ and $u = 0.75$, $v = 0.25$ shows that the converse does not hold, since $w = 0.5 \in \psi_{n,\kappa}$, but $w \notin \xi_{n,\kappa}$. Instead, the following holds:

Lemma A.1

$$\psi_{n,\kappa} = \{\, w \in \mathbf{R}^n : \; -e \le w \le e, \quad \sum_{i=1}^{n} |w_i| \le \kappa \,\}. \tag{A.2}$$

Proof: Let $w \in \psi_{n,\kappa}$, so $w = u - v$, $0 \le u \le e$, $0 \le v \le e$, and $e^T(u+v) = \kappa$. Then

$$|w_i| = |u_i - v_i| \le \max\{u_i, v_i\} \le u_i + v_i$$

so w is an element of the right-hand side of (A.2). Conversely, if w is an element of the right-hand side, the following algorithm produces vectors u and v which demonstrate that $w = u - v \in \psi_{n,\kappa}$.

$$\begin{aligned}
&\Delta := \kappa - \sum_{i=1}^{n} |w_i| \\
&\text{For} \quad i = 1, \ldots, n \quad \text{do} \\
&\qquad \alpha := \tfrac{1}{2} \min\{1 - |w_i|, \Delta\} \\
&\qquad u_i := [w_i]_+ + \alpha \\
&\qquad v_i := [w_i]_- + \alpha \\
&\qquad \Delta := \Delta - 2\alpha.
\end{aligned}$$

Each step enforces $w_i = u_i - v_i$, $0 \le u_i \le 1$, $0 \le v_i \le 1$ but increases $e^T(u+v)$ until it equals κ. Note that Δ must be nonnegative initially and zero on termination (since $\sum_{i=1}^{n} |w_i| \le \kappa \le n$). ∎

An immediate corollary is

Corollary A.2 *The set of extreme points of the compact convex set $\psi_{n,\kappa}$ is*

$$\{\, w \in \mathbf{R}^n : \; w_i \in \{0, \pm 1\}, \;\text{ for }\; i = 1, \ldots, n, \quad \sum_{i=1}^{n} |w_i| = \kappa \,\},$$

that is the n-vectors with κ_1 elements equal to $+1$, κ_2 elements equal to -1, and all other elements equal to zero, where $\kappa_1 \ge 0$, $\kappa_2 \ge 0$ and $\kappa_1 + \kappa_2 = \kappa$.

These results are now extended to the set $\Psi_{n,\kappa}$, by consideration of the eigenvalues of $W = U - V$. The first step is the following lemma, for which it is necessary to order the eigenvalues of U, V and W so as to be able to exploit a classical result of Weyl. (The last statement in this lemma is not actually needed but is included for completeness.)

Lemma A.3 *Let $W = U - V$, where $U \geq 0$, $V \geq 0$ and $\mathrm{tr}(U) + \mathrm{tr}(V) = \kappa$. Let $\sigma_1 \geq \cdots \geq \sigma_n$, $\tau_1 \geq \cdots \geq \tau_n$, and $\omega_1 \geq \cdots \geq \omega_n$ respectively denote the eigenvalues of U, the eigenvalues of V and the eigenvalues of $W = U - V$. Then*

$$\sum_{i=1}^{n} |\omega_i| \leq \kappa. \tag{A.3}$$

Moreover, equality holds in (A.3) if and only if, for $i = 1, \ldots, n$,

$$\sigma_i \tau_{n-i+1} = 0 \tag{A.4}$$

and

$$\omega_i = \begin{cases} \sigma_i & \text{if } \sigma_i > 0; \\ 0 & \text{if } \sigma_i = \tau_{n-i+1} = 0; \\ -\tau_{n-i+1} & \text{if } \tau_{n-i+1} > 0. \end{cases} \tag{A.5}$$

Proof: Weyl's theorem (see e.g. Horn and Johnson (1985), Section 4.3) shows that for $j = 1, \ldots, n$

$$\sigma_n - \tau_{n-j+1} \leq \omega_j \leq \sigma_j - \tau_n.$$

As $U \geq 0$, $V \geq 0$ this implies that

$$-\tau_{n-j+1} \leq \omega_j \leq \sigma_j, \tag{A.6}$$

so

$$|\omega_j| \leq \max\{\sigma_j, \ \tau_{n-j+1}\} \leq \sigma_j + \tau_{n-j+1}. \tag{A.7}$$

Hence

$$\sum_{j=1}^{n} |\omega_j| \leq \sum_{j=1}^{n} \sigma_j + \sum_{j=1}^{n} \tau_{n-j+1} = \kappa, \tag{A.8}$$

which establishes (A.3).

Suppose now that (A.3) is an equality. Then (A.8) holds with equality, i.e. (A.7) holds with equality for $j = 1, \ldots, n$. This implies that for each $j = 1, \ldots, n$ at least one of σ_j and τ_{n-j+1} is zero, establishing (A.4).

If $\tau_n > 0$ equation (A.4) and the ordering of the eigenvalues imply that $U = 0$. Similarly if $\sigma_n > 0$ then $V = 0$. In both cases (A.5) is trivial. Consider now the case when $\sigma_n = \tau_n = 0$. If $\sigma_j > 0$ then $\tau_{n-j+1} = 0$ so (A.6) implies that

$$0 \leq \omega_j \leq \sigma_j.$$

Similarly if $\tau_{n-j+1} > 0$ then $\sigma_j = 0$ so

$$-\tau_{n-j+1} \leq \omega_j \leq 0.$$

If $\sigma_j = \tau_{n-j+1} = 0$ then $\omega_j = 0$. Thus if (A.3) holds with equality then

$$\begin{aligned} \kappa = \sum_{j=1}^{n} |\omega_j| &= \sum_{j:\sigma_j>0} \omega_j \quad + \quad \sum_{j:\tau_{n-j+1}>0} -\omega_j \\ &\leq \sum_{j:\sigma_j>0} \sigma_j \quad + \quad \sum_{j:\tau_{n-j+1}>0} \tau_{n-j+1} \\ &= \kappa. \end{aligned}$$

The only way this can hold with equality is if (A.5) holds.

Conversely as $U \geq 0$, $V \geq 0$ and $\mathrm{tr}(U) + \mathrm{tr}(V) = \kappa$ equations (A.4) and (A.5) directly give (A.3) with equality. ∎

The following results are stated using the positive and negative parts of a symmetric matrix, although they could also be stated directly in terms of eigenvalues. Let $W \in \mathcal{S}_n$ have eigenvalues ω_i for $i = 1, \ldots, n$ and corresponding normalized eigenvectors v_i, so

$$W = \sum_{i=1}^{n} \omega_i v_i v_i^T.$$

The positive and negative parts of the symmetric matrix W are then

$$W_+ = \sum_{i=1}^{n} [\omega_i]_+ v_i v_i^T, \tag{A.9}$$

$$W_- = \sum_{i=1}^{n} [\omega_i]_- v_i v_i^T. \tag{A.10}$$

Then $W_+ \geq 0, \quad W_- \geq 0, \quad W = W_+ - W_-,$

$$\mathrm{tr}(W) = \sum_{i=1}^{n} \omega_i = \mathrm{tr}(W_+) - \mathrm{tr}(W_-) \tag{A.11}$$

and

$$\sum_{i=1}^{n} |\omega_i| = \mathrm{tr}(W_+) + \mathrm{tr}(W_-). \tag{A.12}$$

Now consider the set

$$\Xi_{n,\kappa} = \{\, W \in \mathcal{S}_n : -I \leq W \leq I, \quad \mathrm{tr}(W_+) + \mathrm{tr}(W_-) = \kappa \,\}. \tag{A.13}$$

Note that from equations (A.1) and (A.12)

$$\Xi_{n,\kappa} = \{\, W \in \mathcal{S}_n : W = X \mathrm{diag}(w) X^T, \quad X \in \mathcal{O}_{n,n}, \quad w \in \xi_{n,\kappa} \,\}.$$

For any $W \in \Xi_{n,\kappa}$, setting $U = W_+$, $V = W_-$ shows that $W \in \Psi_{n,\kappa}$ and so $\Xi_{n,\kappa} \subseteq \Psi_{n,\kappa}$. The same example given earlier to show that $\xi_{n,\kappa}$ does not equal $\psi_{n,\kappa}$ also shows that $\Xi_{n,\kappa}$ does not equal $\Psi_{n,\kappa}$. Instead, the following holds.

Corollary A.4

$$\Psi_{n,\kappa} = \{\, W \in \mathcal{S}_n : -I \leq W \leq I, \quad \mathrm{tr}(W_+) + \mathrm{tr}(W_-) \leq \kappa \,\}. \tag{A.14}$$

Proof: Let $W \in \Psi_{n,\kappa}$. Since $W = U - V$ with $0 \leq U \leq I$, $0 \leq V \leq I$, $\mathrm{tr}(U) + \mathrm{tr}(V) = 1$, equations (A.6) and (A.8) show that $\Psi_{n,\kappa}$ is contained in the set on the right-hand side. To establish the reverse inclusion note that any W in the set on the right side has eigenvalues $w = [\omega_1, \ldots, \omega_n]^T$ satisfying the conditions on the right side of (A.2). Therefore, from Lemma A.1, there exist u, v such that $w = u - v \in \psi_{n,\kappa}$. Taking $U = X\mathrm{diag}(u)X^T$ and $V = X\mathrm{diag}(v)X^T$ shows that $W = U - V \in \Psi_{n,\kappa}$. ∎

Remark: The lack of convexity of the sets $\xi_{n,\kappa}$ and $\Xi_{n,\kappa}$ limits their use. Hence the preference for exploiting $\psi_{n,\kappa}$ and $\Psi_{n,\kappa}$, which are convex since the restricting equations are linear, involving sums rather than sums of absolute values.

Characterization of the extreme points of $\Psi_{n,\kappa}$ is now possible.

Lemma A.5 *The set*

$$\{\, YY^T - ZZ^T : Y^TZ = 0, \quad Y^TY = I_{\kappa_1}, \quad Z^TZ = I_{\kappa_2}, \text{ and } \kappa_1 + \kappa_2 = \kappa \,\}.$$

is the set of extreme points of $\Psi_{n,\kappa}$.

Proof: Combining Lemma A.1 and Corollary A.4 gives

$$\Psi_{n,\kappa} = \{\, W \in \mathcal{S}_n : W = X\mathrm{diag}(w)X^T, \quad X \in \mathcal{O}_{n,n}, \quad w \in \psi_{n,\kappa} \,\}.$$

The extreme points of $\Psi_{n,\kappa}$ therefore have the form $X\mathrm{diag}(w)X^T$ where w is an extreme point of $\psi_{n,\kappa}$. Corollary A.2 then gives the desired result. ∎

An obvious generalization of Fan's theorem (which we have nonetheless not encountered in the literature) now follows as a corollary:

Theorem A.6

$$\begin{aligned} g_\kappa(A) = \max \; & \{\mathrm{tr}(Y^TAY) - \mathrm{tr}(Z^TAZ) \; : \\ & Y^TZ = 0, \quad Y^TY = I_{\kappa_1}, \quad Z^TZ = I_{\kappa_2}, \text{ and } \kappa_1 + \kappa_2 = \kappa \,\}. \end{aligned}$$

Lemma A.5 also leads immediately to another characterization of the elements achieving the maximum in (4.21), namely

Theorem A.7

$$\begin{aligned}\operatorname{argmax}\{\langle A, W\rangle : W \in \Psi_{n,\kappa}\} \;=\; \operatorname{conv}\{\, & YY^T - ZZ^T : Y^TZ = 0 \\ & Y^TY = I_{\kappa_1}, \quad Z^TZ = I_{\kappa_2}, \\ & \kappa_1 + \kappa_2 = \kappa \text{ and} \\ & Y^TAY - Z^TAZ = g_\kappa(A)\,\} \quad \text{(A.15)}\end{aligned}$$

From the ordering (4.8) of the eigenvalues of A the requirement $Y^TAY - Z^TAZ = g_\kappa(A)$ means that the columns of Y must include r_1 orthonormal eigenvectors for all the eigenvalues $\lambda_1, \ldots, \lambda_{r_1}$ and the columns of Z must include r_2 orthonormal eigenvectors for all the eigenvalues $\lambda_{n-r_2+1}, \ldots, \lambda_n$, so $\kappa_1 \geq r_1$ and $\kappa_2 \geq r_2$. The remaining columns of Y and Z may be any orthonormal sets of eigenvectors corresponding to $\lambda_{r_1+1} = \ldots = \lambda_{r_1+t_1}$ and $\lambda_{n-r_2-t_2+1} = \ldots = \lambda_{n-r_2}$ respectively.

This leads to the following characterization of the subdifferential of $g_\kappa(A)$.

Corollary A.8

$$\begin{aligned}\partial g_\kappa(A) \;=\; \operatorname{conv}\{\, & YY^T - ZZ^T : \\ & \text{the columns of } Y \text{ form an o.n. set of } \kappa_1 \text{ eigenvectors for } \lambda_1, \ldots, \lambda_{\kappa_1} \\ & \text{the columns of } Z \text{ form an o.n. set of } \kappa_2 \text{ eigenvectors for } \lambda_{n-\kappa_2+1}, \ldots, \lambda_n \\ & \text{where } r_1 \leq \kappa_1 \leq r_1 + t_1, \quad r_2 \leq \kappa_2 \leq r_2 + t_2, \text{ and } \kappa_1 + \kappa_2 = \kappa\,\}\end{aligned}$$

The presence of the convex hull operation means that this form is not as computationally convenient as that given in Corollary 4.6, nor does it display the structure of the subdifferential revealed there.

REFERENCES

1. F. Alizadeh (1991), Private communication.

2. V. I. Arnold (1971), "On matrices depending on parameters", *Russian Mathematical Surveys* **26**,2, pp. 29-43.

3. R. Bellman (1970), *Introduction to Matrix Analysis*, McGraw-Hill, New York, 2nd edition.

4. R. Bellman and K. Fan (1963), "On systems of linear inequalities in matrix variables", in: V. L. Klee ed., *Convexity*, American Mathematical Society, Providence, R.I., pp. 1-11.

5. F. Clarke (1983), *Optimization and Nonsmooth Analysis*, John Wiley, New York. Reprinted by SIAM, Philadelphia, 1990.

6. B. D. Craven and B. Mond (1981), "Linear programming with matrix variables", *Linear Algebra and its Applications* **38**, pp. 73-80.

7. J. E. Cullum, W. E. Donath and P. Wolfe (1975), "The minimization of certain nondifferentiable sums of eigenvalues of symmetric matrices", *Mathematical Programming Study* **3**, pp. 35-55.

8. K. Fan (1949), "On a theorem of Weyl concerning the eigenvalues of linear transformations", *Proceedings of the National Academy of the Sciences of U.S.A* **35**, pp. 652-655.

9. P. A. Fillmore and J. P. Williams (1971), "Some convexity theorems for matrices", *Glasgow Mathematical Journal* **12**, pp. 110-117.

10. R. Fletcher (1985), "Semi-definite matrix constraints in optimization", *SIAM Journal on Control and Optimization* **23**, pp. 493-513.

11. S. Friedland (1981), "Convex spectral functions", *Linear and Multilinear Algebra* **9**, pp. 299-316.

12. S. Friedland, J. Nocedal and M. L. Overton (1987), "The formulation and analysis of numerical methods for inverse eigenvalue problems", *SIAM Journal on Numerical Analysis* **24**, pp. 634-667.

13. G. H. Golub and C. van Loan (1983), *Matrix Computations*, John Hopkins University Press, Baltimore.

14. R. A. Horn and C. Johnson (1985), *Matrix Analysis*, Cambridge University Press, New York.

15. T. Kato (1982), *A Short Introduction to Perturbation Theory for Linear Operators*, Springer-Verlag, New York.

16. M. L. Overton (1988), "On minimizing the maximum eigenvalue of a symmetric matrix", *SIAM Journal on Matrix Analysis and Application* **9**, pp. 256-268.

17. M. L. Overton (1990), "Large-scale optimization of eigenvalues", NYU Computer Science Dept Report 505. To appear in *SIAM Journal on Optimization*.

18. M. L. Overton and R. S. Womersley (1988), "On minimizing the spectral radius of a nonsymmetric matrix function: Optimality conditions and duality theory", *SIAM Journal on Matrix Analysis and Applications* **9**, pp. 473-498.

19. M. L. Overton and R. S. Womersley (to appear), "Second derivatives for optimizing eigenvalues of symmetric matrices". in preparation.

20. L. Qi and R. S. Womersley (to appear) "On extreme singular values", in preparation.

21. F. Rendl and H. Wolkowicz (1990), "A projection technique for partitioning the nodes of a graph", Department of Combinatorics and Optimization Report 90-20, University of Waterloo.

22. R. T. Rockafellar (1970), *Convex Analysis*, Princeton University Press, Princeton, NJ.

23. A. H. Sameh and J. A. Wisniewski (1982), "A trace minimization algorithm for the generalized eigenvalue problem", *SIAM Journal on Numerical Analysis* **19**, pp. 1243-1259.

24. A. Shapiro (1985), "Extremal problems on the set of nonnegative definite matrices", *Linear Algebra and its Applications* **67** pp. 7-18.

25. A. Shapiro and J. D. Botha (1988), "Dual algorithms for orthogonal procrustes rotations", *SIAM Journal on Matrix Analysis* **9**, pp. 378-383.

26. G. A. Watson (1990), Private communication.

27. H. Wielandt (1955), "An extremum property of sums of eigenvalues", *Proceedings of the American Mathematical Society* **6**, pp. 106-110.