On Some Quadratic Optimization Problems Arising in Computer Vision

Jean Gallier

Department of Computer and Information Science
University of Pennsylvania
Philadelphia, PA 19104, USA
e-mail: jean@cis.upenn.edu

April 13, 2016

Abstract. The goal of this paper is to find methods for solving various quadratic optimization problems, mostly arising from computer vision (image segmentation and contour grouping). We consider mainly two problems:

Problem 1. Let A be an $n \times n$ Hermitian matrix and let $b \in \mathbb{C}^n$ be any vector,

$$\begin{array}{ll} \text{maximize} & z^*Az + z^*b + b^*z \\ \text{subject to} & z^*z = 1, \quad z \in \mathbb{C}^n. \end{array}$$

Problem 2. If A is a real $n \times n$ symmetric matrix and $b \in \mathbb{R}$ is any vector,

$$\begin{array}{ll} \text{maximize} & x^\top Ax + 2x^\top b \\ \text{subject to} & x^\top x = 1, \quad x \in \mathbb{R}^n. \end{array}$$

First, we show that Problem 1 reduces to Problem 2. We reduce Problem 2 to the problem of finding the intersection of an algebraic curve generalizing the hyperbola to $\mathbb{R}^n$ with the unit sphere. This allows us to analyze the number of solutions of Problem 2 in terms of the nature of the eigenvalues of the (symmetric) matrix A. As a consequence, we prove that the maximum of the function $f(x) = x^\top Ax + 2x^\top b$ on the unit sphere is achieved for all the critical points (x, λ) of the Lagrangian $L(x, \lambda) = x^\top Ax + 2x^\top b - \lambda(x^\top x - 1)$, such that $\lambda \geq \sigma_i$, for all eigenvalues σ_i of A. Problem 2 has been considered before, but our approach involving a simple algebraic curve sheds some new light on the problem and simplifies some proofs. We provide an extensive discussion of related works.

1 Formulation of the Optimization Problems

The goal of this paper is to find methods for solving various quadratic optimization problems, mostly arising from computer vision (image segmentation and contour grouping). For a quick overview of the problems, we suggest reading Sections 1 and 2, omitting proofs at first, and then jumping directly to Section 6 which contains a thorough discussion of related work.

We consider mainly two problems:

Problem 1. Let A be an $n \times n$ Hermitian matrix and let $b \in \mathbb{C}^n$ be any vector. Consider the following optimization problem:

$$\begin{array}{ll} \text{maximize} & z^*Az + z^*b + b^*z \\ \text{subject to} & z^*z = 1, \quad z \in \mathbb{C}^n. \end{array}$$

Because the matrix A is Hermitian, the quantity $f(z) = z^*Az + z^*b + b^*z$ is real.

Problem 2. If A is a real $n \times n$ symmetric matrix and $b \in \mathbb{R}$ is any vector,

$$\begin{array}{ll} \text{maximize} & x^\top Ax + 2x^\top b \\ \text{subject to} & x^\top x = 1, \quad x \in \mathbb{R}^n. \end{array}$$

First, we show that Problem 1 reduces to Problem 2. Since A is Hermitian, we can write $A = H + iS$, with

$$H = \frac{A + A^\top}{2}, \qquad S = \frac{A - A^\top}{2i},$$

where H is a real symmetric matrix and S is a real skew symmetric matrix ($S^\top = -S$) and if we let $z = x + iy$ and $b = b_r + ib_c$, with $x, y \in \mathbb{R}^n$ and $b_r, b_c \in \mathbb{R}^n$, then $x^\top Hy = y^\top Hx$, $x^\top Sy = -y^\top Sx$, and $x^\top Sx = y^\top Sy = 0$, so we have

$$\begin{aligned} z^*Az &= (x^\top - iy^\top)A(x + iy) \\ &= x^\top Ax + ix^\top Ay - iy^\top Ax + y^\top Ay \\ &= x^\top Hx + ix^\top Sx + ix^\top Hy - x^\top Sy - iy^\top Hx + y^\top Sx + y^\top Hy + iy^\top Sy \\ &= x^\top Hx + y^\top Hy - 2x^\top Sy \\ &= (x^\top, y^\top) \begin{pmatrix} H & -S \\ S & H \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \end{aligned}$$

and

$$\begin{aligned} z^*b + b^*z &= (x^\top - iy^\top)(b_r + ib_c) + (b_r^\top - ib_c^\top)(x + iy) \\ &= x^\top b_r + ix^\top b_c - iy^\top b_r + y^\top b_c + b_r^\top x + ib_r^\top y - ib_c^\top x + b_c^\top y \\ &= 2x^\top b_r + 2y^\top b_c \\ &= 2(x^\top, y^\top) \begin{pmatrix} b_r \\ b_c \end{pmatrix}. \end{aligned}$$

Observe that the matrix

$$\begin{pmatrix} H & -S \\ S & H \end{pmatrix}$$

is real symmetric. Therefore, our optimization problem reduces to the problem

$$\begin{aligned} \text{maximize} \quad & (x^\top, y^\top) \begin{pmatrix} H & -S \\ S & H \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + 2(x^\top, y^\top) \begin{pmatrix} b_r \\ b_c \end{pmatrix} \\ \text{subject to} \quad & (x^\top, y^\top) \begin{pmatrix} x \\ y \end{pmatrix} = 1, \quad \begin{pmatrix} x \\ y \end{pmatrix} \in \mathbb{R}^{2n} \end{aligned}$$

where the matrix involved is a real symmetric $2n \times 2n$ matrix.

Consequently, we will now focus on the following optimization problem:

Problem 2. If A is a real $n \times n$ symmetric matrix and $b \in \mathbb{R}$ is any vector,

$$\begin{aligned} \text{maximize} \quad & x^\top A x + 2x^\top b \\ \text{subject to} \quad & x^\top x = 1, \quad x \in \mathbb{R}^n. \end{aligned}$$

Observe that if $A = \mu I$, for some $\mu \in \mathbb{R}$, then on the unit sphere, $x^\top x = 1$, we have

$$f(x) = x^\top A x + 2x^\top b = \mu + 2x^\top b.$$

If $b = 0$, then f is the constant function with value μ. If $b \neq 0$, then the maximum of $f(x) = \mu + 2x^\top b$ is achieved for $x = b/\sqrt{b^\top b}$.

For the rest of this paper, we will assume that A is not of the form μI, which means that A is a symmetric matrix with at least two distinct eigenvalues.

Let $L(x, \lambda)$ be the Lagrangian of the above problem,

$$L(x, \lambda) = x^\top A x + 2x^\top b - \lambda(x^\top x - 1).$$

We know that a necessary condition for the function $f(x) = x^\top A x + 2x^\top b$ to have a local extremum on the unit sphere $x^\top x = 1$, is that $L(x, \lambda)$ has a critical point, which means that

$$\frac{\partial L}{\partial x} = 0, \qquad \frac{\partial L}{\partial \lambda} = 0.$$

Since

$$\frac{\partial L}{\partial x} = 2Ax + 2b - 2\lambda x, \qquad \frac{\partial L}{\partial \lambda} = x^\top x - 1,$$

necessary conditions for f to have a local extremum are

$$\begin{aligned} (\lambda I - A)x &= b \\ x^\top x &= 1. \end{aligned}$$

If $b = 0$, this is a standard eigenvalue problem so let us assume that $b \neq 0$. Since A is a symmetric matrix, it can be diagonalized and we can write

$$A = Q^\top \Sigma Q,$$

where Σ is a (real) diagonal matrix and Q is an orthogonal matrix. Substituting the righthand side of A into our system, we get

$$\begin{aligned} Q^\top(\lambda I - \Sigma)Qx &= b \\ x^\top x &= 1, \end{aligned}$$

which yields

$$\begin{aligned} (\lambda I - \Sigma)Qx &= Qb \\ (Qx)^\top Qx &= 1. \end{aligned}$$

If we let $c = Qb$ and $y = Qx$, the above system becomes

$$\begin{aligned} (\lambda I - \Sigma)y &= c \\ y^\top y &= 1 \end{aligned}$$

and the solutions of the original system

$$\begin{aligned} (\lambda I - A)x &= b \\ x^\top x &= 1 \end{aligned}$$

are obtained using the equation $x = Q^\top y$.

Remark: It is well-known that it is possible to "absorb" the linear term, $2x^\top b$, into the quadratic term, $x^\top Ax$, by going up one dimension, that is, by considering the unknown to be the vector $\binom{x}{t} \in \mathbb{R}^{n+1}$. Then, if we observe that

$$(x^\top, t)\begin{pmatrix} A & b \\ b^\top & 0 \end{pmatrix}\begin{pmatrix} x \\ t \end{pmatrix} = x^\top Ax + 2tx^\top b,$$

our optimization problem is clearly equivalent to

$$\begin{aligned} \text{maximize} \quad & (x^\top, t)\begin{pmatrix} A & b \\ b^\top & 0 \end{pmatrix}\begin{pmatrix} x \\ t \end{pmatrix} \\ \text{subject to} \quad & (x^\top, t)\begin{pmatrix} x \\ t \end{pmatrix} = 2, \quad \begin{pmatrix} x \\ t \end{pmatrix} \in \mathbb{R}^{n+1} \\ & t = 1. \end{aligned}$$

The constraint, $t = 1$, is linear and can be written as

$$c^\top \begin{pmatrix} x \\ t \end{pmatrix} = 1,$$

where $c^\top = (0, \ldots, 0, 1)$.

If the right-hand side of this last equation was 0, then following Golub [6] (1973), it would be possible to get rid of this constraint and reduce the problem to a standard eigenvalue problem with respect to a different matrix, namely PA', with $P = I - cc^\top$ and $A' = \begin{pmatrix} A & b \\ b^\top & 0 \end{pmatrix}$. The matrix PA' is not symmetric, but $PA'P$ is symmetric and as $P^2 = P$ (P is a projection) and since it is known that PPA' and $PA'P$ have the same eigenvalues, we would be reduced to a standard eigenvalue problem. Golub [6] also shows how to handle a more general linear constraint of the form $C^\top x = 0$, where C is a matrix (for details, see Section 6).

Unfortunately, the right-hand side of our equation, $c^\top \binom{x}{t} = 1$, is not zero and we have not made any progress. Indeed, the Lagrangian of the new formulation of our problem is

$$L'\left(\binom{x}{t}, \lambda, \mu\right) = (x^\top, t)\begin{pmatrix} A & b \\ b^\top & 0 \end{pmatrix}\binom{x}{t} - \lambda\left((x^\top, t)\binom{x}{t} - 2\right) - \mu\left(c^\top\binom{x}{t} - 1\right)$$

and necessary conditions for L' to have a critical point are

$$\begin{aligned} 2\begin{pmatrix} A & b \\ b^\top & 0 \end{pmatrix}\binom{x}{t} - 2\lambda\binom{x}{t} - \mu c &= 0 \\ (x^\top, t)\binom{x}{t} &= 2 \\ c^\top\binom{x}{t} &= 1. \end{aligned}$$

Since $c^\top = (0, \ldots, 0, 1)$, we must have $t = 1$, and then

$$\begin{aligned} Ax - \lambda x + b &= 0 \\ x^\top x &= 1 \\ \mu &= 2b^\top x - 2\lambda. \end{aligned}$$

Therefore, we are back to our original system

$$\begin{aligned} (\lambda I - A)x &= b \\ x^\top x &= 1. \end{aligned}$$

In fact, Gander, Golub and von Matt [5] (1989) have shown that eliminating a constraint of the form $N^\top x = t$ (where t is a nonzero vector) from the quadratic problem

$$\begin{array}{ll} \text{maximize} & x^\top A x \\ \text{subject to} & x^\top x = 1, \quad x \in \mathbb{R}^n \\ & N^\top x = t \end{array}$$

leads to a quadratic function of the form $x^\top C x + 2x^\top b$ (for details, see Section 6). In summary, there is really no hope of making the linear term $2x^\top b$ go away.

2 Solution in the Generic Case

Let us first assume that the eigenvalues of A are all distinct and order them in decreasing order so that $\sigma_1 > \sigma_2 > \cdots > \sigma_n$. The system

$$(\lambda I - \Sigma)y = c$$

defines a parametric curve, $C(\Sigma, c)$, in $\mathbb{R}^n$, for all $\lambda \neq \sigma_i$, $1 \leq i \leq n$, where the ith coordinate of a point on the curve is given by

$$y_i(\lambda) = \frac{c_i}{\lambda - \sigma_i}.$$

If $c_i \neq 0$, for $i = 1, \ldots, n$, then $y_i(\lambda) \longrightarrow \pm\infty$ when $\lambda \longrightarrow \sigma_i$ and note that $y \longrightarrow 0$ when $\lambda \longrightarrow \pm\infty$. In this case, the solutions of the system

$$\begin{aligned} (\lambda I - \Sigma)y &= c \\ y^\top y &= 1 \end{aligned}$$

are the points of intersection of the curve, $C(\Sigma, c)$, with the unit sphere, $y^\top y = 1$.

The (connected) branch of the curve, $C(\Sigma, c)$, for which $\lambda \in (-\infty, \sigma_n) \cup (\sigma_1, +\infty)$ always intersects the unit sphere, since it passes through the origin for $\lambda = \pm\infty$. When $\lambda \longrightarrow \sigma_n$ from $-\infty$, the line parallel to the y_n-axis for which

$$y_1 = \frac{c_1}{\sigma_n - \sigma_1}, \ldots, y_{n-1} = \frac{c_n}{\sigma_n - \sigma_{n-1}}$$

is an asymptote and when $\lambda \longrightarrow \sigma_1$ from $+\infty$, the line parallel to the y_1-axis for which

$$y_2 = \frac{c_2}{\sigma_1 - \sigma_2}, \ldots, y_n = \frac{c_{n-1}}{\sigma_1 - \sigma_n}$$

is another asymptote. Since the coordinates $y_2, \ldots, y_{n-1}$ of these two lines have different signs, this branch of the curve has a "kink" (it is not planar).

The curve, $C(\Sigma, c)$, has $n - 1$ other connected branches, one for each interval (σ_i, σ_{i-1}), where $i = n, \ldots, 2$. When $\lambda \longrightarrow \sigma_i$ from above, the line parallel to the y_i-axis for which

$$y_1 = \frac{c_1}{\sigma_i - \sigma_1}, \ldots, y_{i-1} = \frac{c_{i-1}}{\sigma_i - \sigma_{i-1}}, y_{i+1} = \frac{c_{i+1}}{\sigma_i - \sigma_{i+1}}, \ldots, y_n = \frac{c_n}{\sigma_i - \sigma_n}$$

(with y_{i+1} and y_n omitted when $i = n$) is an asymptote and when $\lambda \longrightarrow \sigma_{i-1}$ from below, the line parallel to the y_{i-1}-axis for which

$$y_1 = \frac{c_1}{\sigma_{i-1} - \sigma_1}, \ldots, y_{i-2} = \frac{c_{i-2}}{\sigma_{i-1} - \sigma_{i-2}}, y_i = \frac{c_i}{\sigma_{i-1} - \sigma_i}, \ldots, y_n = \frac{c_n}{\sigma_{i-1} - \sigma_n}$$

(with y_1 and y_{i-2} omitted when $i = 2$) is another asymptote. Since either the y_1 coordinate or the y_n coordinate of these two lines differ, these branches of the curve also have a "kink" (are not planar).

If $c_i = 0$ for some i, the situation is more subtle. Let us begin by considering the case $n = 2$.

When $n = 2$, we have the system of equations

$$\begin{aligned}(\lambda - \sigma_1)y_1 &= c_1\\ (\lambda - \sigma_2)y_2 &= c_2\\ y_1^2 + y_2^2 &= 1.\end{aligned}$$

If $c_1 = 0$, then, as $c_2 \neq 0$, the two linear equations have a solution iff $\lambda \neq \sigma_2$.

Case 1. If (y_1, y_2) is a solution of the system

$$\begin{aligned}(\lambda - \sigma_1)y_1 &= 0\\ (\lambda - \sigma_2)y_2 &= c_2\end{aligned}$$

with $y_1 = 0$, then this solution belongs to the line of equation $y_1 = 0$. This line intersects the unit circle $y_1^2 + y_2^2 = 1$ for $y_2 = \pm 1$. For these solutions, we must have

$$y_2 = \frac{c_2}{\lambda - \sigma_2} = \pm 1$$

which has the two solutions,

$$\lambda = \sigma_2 \pm c_2.$$

Since $c_2 \neq 0$, our system has the two solutions $(y_1, y_2) = (0, \pm 1)$ for $\lambda = \sigma_2 \pm c_2$.

Case 2. If (y_1, y_2) with $y_1 \neq 0$ is a solution of the system

$$\begin{aligned}(\lambda - \sigma_1)y_1 &= 0\\ (\lambda - \sigma_2)y_2 &= c_2\end{aligned}$$

then we must have $\lambda = \sigma_1$. In this case, the above system reduces to the single equation

$$(\sigma_1 - \sigma_2)y_2 = c_2$$

which defines the line of equation

$$y_2 = \frac{c_2}{\sigma_1 - \sigma_2}.$$

This line intersects the unit circle $y_1^2 + y_2^2 = 1$ iff

$$y_1^2 = 1 - \frac{c_2^2}{(\sigma_1 - \sigma_2)^2}.$$

This equation has real nonzero solutions iff

$$c_2^2 < (\sigma_1 - \sigma_2)^2$$

and if so, the solutions to our system are

$$y_1 = \pm\sqrt{1 - y_2^2}, \qquad y_2 = \frac{c_2}{\sigma_1 - \sigma_2}.$$

In summary, when $c_1 = 0$, $(y_1, y_2) = (0, \pm 1)$ are solutions and there are possibly two extra solutions if $\lambda = \sigma_1$ and $c_2^2 < (\sigma_1 - \sigma_2)^2$.

The case where $c_2 = 0$ is similar. We find that $(y_1, y_2) = (\pm 1, 0)$ are solutions and there are possibly two extra solutions if $\lambda = \sigma_2$ and $c_1^2 < (\sigma_2 - \sigma_1)^2$.

Case 3. If $c_1 \neq 0$ and $c_2 \neq 0$, by solving for λ in terms of y_1, we get

$$\lambda = \frac{c_1}{y_1} + \sigma_1$$

and by substituting in the second equation we get

$$y_2 = \frac{c_2 y_1}{c_1 + (\sigma_1 - \sigma_2) y_1}$$

the equation of a hyperbola passing through the origin and with two asymptotes parallel to the y_1 and the y_2 axes, namely,

$$y_1 = -\frac{c_1}{\sigma_1 - \sigma_2}$$

and

$$y_2 = \frac{c_2}{\sigma_1 - \sigma_2}.$$

The branch of the hyperbola passing through the origin intersects the unit circle, $y_1^2 + y_2^2 = 1$, in two points and, in general, the other branch of the hyperbola also intersects the unit circle in two points as illustrated in Figure 1.

Therefore, in general, the hyperbola intersects the unit circle in four points and always in at least two points. The corresponding values of λ are given by the equation

$$\frac{c_1^2}{(\lambda - \sigma_1)^2} + \frac{c_2^2}{(\lambda - \sigma_2)^2} = 1,$$

which yields a polynomial equation of degree 4.

In the general case, $n \geq 2$, we have the following theorem:

Theorem 2.1. *If the eigenvalues of the $n \times n$ symmetric matrix A are all distinct, then there are $2m$ values of λ, say $\lambda_1 > \lambda_2 \geq \lambda_3 > \cdots > \lambda_{2m-2} \geq \lambda_{2m-1} > \lambda_{2m}$, with $1 \leq m \leq n$, such that the system*

$$\begin{aligned} (\lambda I - A)x &= b \\ x^\top x &= 1 \end{aligned}$$

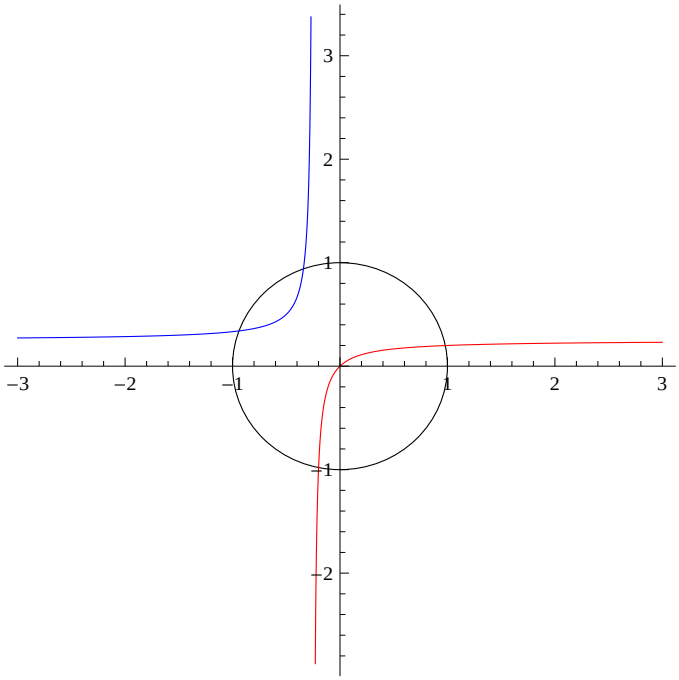

Figure 1: Intersections of $C(\Sigma, c)$ (a hyperbola) with the unit circle

(with $b \neq 0$) has a solution, (λ, x). As a consequence, the Lagrangian

$$L(x, \lambda) = x^\top Ax + 2x^\top b - \lambda(x^\top x - 1)$$

has at least two and at most $2n$ critical point (x, λ). Furthermore, the eigenvalues $\sigma_1 > \sigma_2 > \cdots > \sigma_n$ of A separate the λ's, which means that

1. $\lambda_1 \geq \sigma_1$.
2. $\lambda_{2m} \leq \sigma_n$.
3. *For every λ_i, with $2 \leq i \leq 2m-1$, either $\lambda_i = \sigma_j$ for some j with $1 \leq j \leq n$, or there is some j, with $1 \leq j \leq n-1$, so that $\sigma_j > \lambda_i > \sigma_{j+1}$.*

Proof. As we explained earlier, we first diagonalize A as $A = Q^\top \Sigma Q$ and if we let $c = Qb$ and $y = Qx$, then the above system is equivalent to the system

$$\begin{aligned} (\lambda I - \Sigma)y &= c \\ y^\top y &= 1. \end{aligned}$$

If $c_i \neq 0$ for $i = 1, \ldots, n$, then the curve, $C(\Sigma, c)$, is a kind of generalized hyperbola in $\mathbb{R}^n$, with n asymptotes corresponding the the values $\lambda = \sigma_i$. An example of this curve in shown for $n = 3$ in Figure 2.

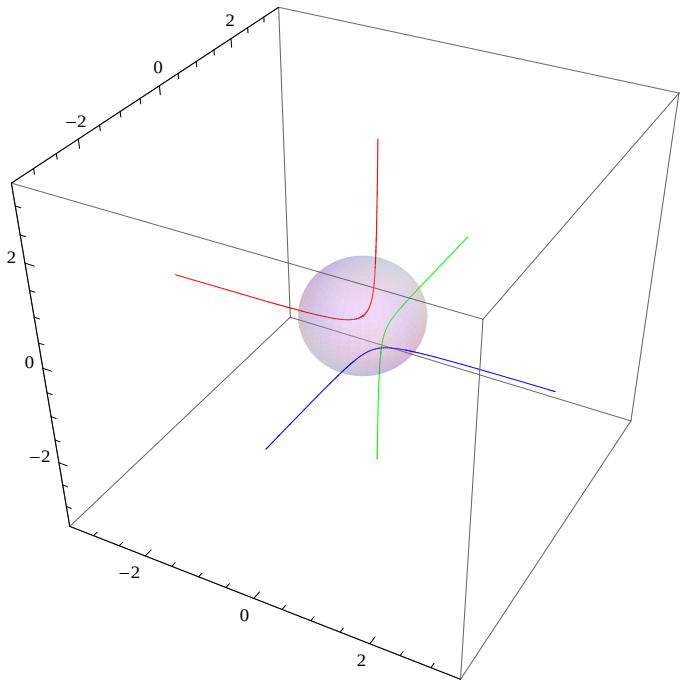

Figure 2: Intersections of $C(\Sigma, c)$ with the unit sphere ($n = 3$)

In order for some, y, on the curve $C(\Sigma, c)$ to belong to the unit sphere, the equation

$$\sum_{i=1}^{n} \frac{c_i^2}{(\lambda - \sigma_i)^2} = 1,$$

must hold, which yields a polynomial equation of degree $2n$. Observe that since every coordinate,

$$y_i = \frac{c_i}{\lambda - \sigma_i},$$

of a point on the curve, $C(\Sigma, c)$, is an injective monotonic function of λ (for $\lambda \neq \sigma_i$), the branch of the curve passing through 0 intersects the unit sphere in exactly two points:

1. One point when λ varies from $-\infty$ to σ_n, corresponding to some $\lambda_{2m} < \sigma_n$.

2. One point when λ varies from $+\infty$ to σ_1, corresponding to some $\lambda_1 > \sigma_1$.

If another branch of $C(\Sigma, c)$ corresponding to $\lambda \in (\sigma_{j+1}, \sigma_j)$ $(1 \leq j \leq n-1)$ intersects the unit sphere, then it will do so in two points corresponding to some values, λ_{2i} and λ_{2i+1}, so that $\sigma_j > \lambda_{2i} \geq \lambda_{2i+1} > \sigma_{j+1}$.

Thus, when the eigenvalues of A are all distinct and $c_i \neq 0$ for $i = 1, \ldots, n$, the system

$$\begin{aligned} (\lambda I - \Sigma)y &= c \\ y^\top y &= 1 \end{aligned}$$

has at least two and at most $2n$ solutions and the eigenvalues of A separate the solutions in λ.

Let us now consider the case where $c_i = 0$ for all $i \in Z$ in some proper subset, Z, of $\{1, \ldots, n\}$ and let $s = |Z|$. The linear system,

$$(\lambda I - \Sigma)y = c,$$

has a solution iff $\lambda \neq \sigma_i$ for all $i \notin Z$.

Case 1. There is a solution, y, of the system $(\lambda I - \Sigma)y = c$ for which $y_i = 0$ for all $i \in Z$. In this case, the system $(\lambda I - \Sigma)y = c$ defines a curve, $C(\Sigma, c)$, in the subspace of dimension $n - s$ defined by the equations $y_i = 0$, for $i \in Z$, and this curve is given parametrically by

$$y_i = \frac{c_i}{\lambda - \sigma_i}, \qquad i \notin Z.$$

This curve intersects the unit sphere, $y^\top y = 1$, iff the equation

$$\sum_{i \notin Z} \frac{c_i^2}{(\lambda - \sigma_i)^2} = 1$$

holds, which yields a polynomial equation of degree $2(n - s)$. Clearly, each of its roots, λ, must be different from σ_i, for $i \notin Z$.

If $1 \notin Z$, then the branch of the curve, $C(\Sigma, c)$, through the origin, intersects the unit sphere when λ varies from $+\infty$ to σ_1 in some point for which $\lambda > \sigma_1$. Similarly, if $n \notin Z$, then the branch of the curve, $C(\Sigma, c)$, through the origin, intersects the unit sphere when λ varies from $-\infty$ to σ_n in some point for which $\lambda < \sigma_n$.

Case 2. There is a solution, y, of the system $(\lambda I - \Sigma)y = c$ and $y_k \neq 0$ for some $k \in Z$. Then, we must have, $\lambda = \sigma_k$, in which case the system $(\sigma_k I - \Sigma)y = c$ defines a line given by the equations

$$\begin{aligned} y_i &= 0, \qquad i \in Z - \{k\} \\ y_i &= \frac{c_i}{\sigma_k - \sigma_i}, \qquad i \notin Z \end{aligned}$$

This line intersects the unit sphere, $y^\top y = 1$, iff the equation

$$y_k^2 + \sum_{i \notin Z} \frac{c_i^2}{(\sigma_k - \sigma_i)^2} = 1$$

has a solution with $y_k \neq 0$. This will be the case iff

$$\sum_{i \notin Z} \frac{c_i^2}{(\sigma_k - \sigma_i)^2} < 1,$$

and we get two solutions for y_k and thus, for y. In summary, there are up to $2s$ solutions if $\lambda = \sigma_i$ with $i \in Z$, and there are always at least two and up to $2(n-s)$ solutions with $y_i = 0$ for all $i \in Z$. Thus, in all cases, there are at least two and at most $2n$ solutions.

If $1 \in Z$, then if $\sum_{i \notin Z} \frac{c_i^2}{(\sigma_1 - \sigma_i)^2} < 1$, then $\lambda = \sigma_1$ is a solution. Consequently, the largest solution, λ, must satisfy $\lambda \geq \sigma_1$. If $\sum_{i \notin Z} \frac{c_i^2}{(\sigma_1 - \sigma_i)^2} \geq 1$, then, the point corresponding to σ_1 is not inside the unit sphere and since the point on the curve, $C(\Sigma, c)$, moves away from the origin as λ decreases from $+\infty$, the intersection with the unit sphere will occur for some $\lambda \geq \sigma_1$. It follows that $\lambda \geq \sigma_1$ for the largest solution λ.

If $n \in Z$, then if $\sum_{i \notin Z} \frac{c_i^2}{(\sigma_n - \sigma_i)^2} < 1$, then $\lambda = \sigma_n$ is a solution. Consequently, the smallest solution, λ, must satisfy $\lambda \leq \sigma_n$. If $\sum_{i \notin Z} \frac{c_i^2}{(\sigma_n - \sigma_i)^2} \geq 1$, then, the point corresponding to σ_n is not inside the unit sphere and since the point on the curve, $C(\Sigma, c)$, moves away from the origin as λ decreases from $-\infty$, the intersection with the unit sphere will occur for some $\lambda \leq \sigma_n$. Thus, $\lambda \leq \sigma_n$ for the smallest solution λ.

Suppose that λ is a solution such that $\lambda \neq \sigma_i$, for $i = 1, \ldots, n$. First, assume that λ is the smallest of the two values for which some branch of the curve, $C(\Sigma, c)$, with $\lambda \in (\sigma_j, \sigma_l)$, intersects the unit sphere. We must have $\sigma_l > \lambda > \sigma_j$. If $j > l+1$, then for every intermediate σ_k, with $l < k < j$, if

$$\sum_{i \notin Z} \frac{c_i^2}{(\sigma_k - \sigma_i)^2} < 1,$$

then, as λ increases from σ_j, we must have $\lambda < \sigma_k$.

If $\sum_{i \notin Z} \frac{c_i^2}{(\sigma_k - \sigma_i)^2} < 1$ for all k with $l < k < j$, then

$$\sigma_{j-1} > \lambda > \sigma_j.$$

Otherwise, if k, with $l < k < j$, is the largest index for which $\sum_{i \notin Z} \frac{c_i^2}{(\sigma_k - \sigma_i)^2} \geq 1$, then $\sigma_{k-1} > \lambda > \sigma_k$.

Next, assume that λ is the largest of the two values for which the branch of the curve, $C(\Sigma, c)$, with $\lambda \in (\sigma_j, \sigma_l)$, intersects the unit sphere. We must have $\sigma_l > \lambda > \sigma_j$. If $j > l+1$, then for every intermediate σ_k, with $l < k < j$, if

$$\sum_{i \notin Z} \frac{c_i^2}{(\sigma_k - \sigma_i)^2} < 1,$$

then, as λ decreases from σ_l, we must have $\lambda > \sigma_k$.

If $\sum_{i \notin Z} \frac{c_i^2}{(\sigma_k - \sigma_i)^2} < 1$ for all k with $l < k < j$, then

$$\sigma_l > \lambda > \sigma_{l-1}.$$

Otherwise, if k, with $l < k < j$, is the smallest index for which $\sum_{i \notin Z} \frac{c_i^2}{(\sigma_k - \sigma_i)^2} \geq 1$, then $\sigma_k > \lambda > \sigma_{k+1}$.

Since we have considered all possibilities, the proof is complete. □

Regarding the linear independence of the solutions, y, we have the following proposition:

Proposition 2.2. *Let m be the number of c_is such that $c_i \neq 0$. Then, any $k \leq m$ unit vectors $y^1, \ldots, y^k$ associated with distinct λ_i's such that λ_i and y^i are solutions of the system*

$$\begin{aligned} (\lambda I - \Sigma)y &= c \\ y^\top y &= 1 \end{aligned}$$

are linearly independent.

Proof. We may assume by renumbering coordinates that $c_i \neq 0$, for $i = 1, \ldots, m$, with $m \leq n$. Since the y^i are unit vectors solutions of the system

$$y^i_j = \frac{c_j}{\lambda_i - \sigma_j},$$

with $1 \leq i \leq k \leq m$ and $1 \leq j \leq n$, it is enough to prove that the determinant of the matrix

$$\begin{pmatrix} \dfrac{c_1}{\lambda_1 - \sigma_1} & \dfrac{c_2}{\lambda_1 - \sigma_2} & \cdots & \dfrac{c_k}{\lambda_1 - \sigma_k} \\ \dfrac{c_1}{\lambda_2 - \sigma_1} & \dfrac{c_2}{\lambda_2 - \sigma_2} & \cdots & \dfrac{c_k}{\lambda_2 - \sigma_k} \\ \vdots & \vdots & \vdots & \vdots \\ \dfrac{c_1}{\lambda_k - \sigma_1} & \dfrac{c_2}{\lambda_k - \sigma_2} & \cdots & \dfrac{c_k}{\lambda_k - \sigma_k} \end{pmatrix} = c_1 c_2 \cdots c_k \begin{pmatrix} \dfrac{1}{\lambda_1 - \sigma_1} & \dfrac{1}{\lambda_1 - \sigma_2} & \cdots & \dfrac{1}{\lambda_1 - \sigma_k} \\ \dfrac{1}{\lambda_2 - \sigma_1} & \dfrac{1}{\lambda_2 - \sigma_2} & \cdots & \dfrac{1}{\lambda_2 - \sigma_k} \\ \vdots & \vdots & \vdots & \vdots \\ \dfrac{1}{\lambda_k - \sigma_1} & \dfrac{1}{\lambda_k - \sigma_2} & \cdots & \dfrac{1}{\lambda_k - \sigma_k} \end{pmatrix}$$

is nonzero and since $c_1 c_2 \cdots c_k \neq 0$, this amounts to proving that

$$\det \begin{pmatrix} \dfrac{1}{\lambda_1 - \sigma_1} & \dfrac{1}{\lambda_1 - \sigma_2} & \cdots & \dfrac{1}{\lambda_1 - \sigma_k} \\ \dfrac{1}{\lambda_2 - \sigma_1} & \dfrac{1}{\lambda_2 - \sigma_2} & \cdots & \dfrac{1}{\lambda_2 - \sigma_k} \\ \vdots & \vdots & \vdots & \vdots \\ \dfrac{1}{\lambda_k - \sigma_1} & \dfrac{1}{\lambda_k - \sigma_2} & \cdots & \dfrac{1}{\lambda_k - \sigma_k} \end{pmatrix} \neq 0.$$

Now, since the λ_i are solutions of the equation

$$\sum_{j=1}^{m} \frac{c_j^2}{(\lambda_i - \sigma_j)^2} = 1,$$

we must have $\lambda_i \neq \sigma_j$ for all i, j.

The problem reduces to computing a so-called *Cauchy determinant*. A (square) *Cauchy matrix* is a matrix of the form

$$\begin{pmatrix} \dfrac{1}{\lambda_1 - \sigma_1} & \dfrac{1}{\lambda_1 - \sigma_2} & \cdots & \dfrac{1}{\lambda_1 - \sigma_n} \\ \dfrac{1}{\lambda_2 - \sigma_1} & \dfrac{1}{\lambda_2 - \sigma_2} & \cdots & \dfrac{1}{\lambda_2 - \sigma_n} \\ \vdots & \vdots & \vdots & \vdots \\ \dfrac{1}{\lambda_n - \sigma_1} & \dfrac{1}{\lambda_n - \sigma_2} & \cdots & \dfrac{1}{\lambda_n - \sigma_n} \end{pmatrix}$$

where $\lambda_i \neq \sigma_j$, for all i, j, with $1 \leq i, j \leq n$. It is known that the determinant, C_n, of a Cauchy matrix as above is given by

$$C_n = \frac{\prod_{i=2}^{n} \prod_{j=1}^{i-1} (\lambda_i - \lambda_j)(\sigma_j - \sigma_i)}{\prod_{i=1}^{n} \prod_{j=1}^{n} (\lambda_i - \sigma_j)}.$$

Here is a proof of the above formula by induction. The base case $n = 1$ is trivial. For the induction step, we perform the following row operations which preserve the determinant, C_{n+1}:

Multiply the first row by $\frac{\lambda_1 - \sigma_1}{\lambda_i - \sigma_1}$ and subtract the resulting row from the ith row, $i \geq 2$.

The effect of these linear combinations is to set all the entries of the first column of our matrix but the first to zero. More precisely, the jth entry ($j \geq 2$) of the new ith row ($i \geq 2$) is

$$\begin{aligned}
\frac{1}{\lambda_i - \sigma_j} - \frac{\lambda_1 - \sigma_1}{(\lambda_i - \sigma_1)(\lambda_1 - \sigma_j)} &= \frac{(\lambda_i - \sigma_1)(\lambda_1 - \sigma_j) - (\lambda_1 - \sigma_1)(\lambda_i - \sigma_j)}{(\lambda_i - \sigma_j)(\lambda_i - \sigma_1)(\lambda_1 - \sigma_j)} \\
&= \frac{\lambda_1\lambda_i - \lambda_i\sigma_j - \lambda_1\sigma_1 + \sigma_1\sigma_j - \lambda_1\lambda_i + \lambda_1\sigma_j + \lambda_i\sigma_1 - \sigma_1\sigma_j}{(\lambda_i - \sigma_j)(\lambda_i - \sigma_1)(\lambda_1 - \sigma_j)} \\
&= \frac{(\lambda_1 - \lambda_i)\sigma_j + (\lambda_i - \lambda_1)\sigma_1}{(\lambda_i - \sigma_j)(\lambda_i - \sigma_1)(\lambda_1 - \sigma_j)} \\
&= \frac{(\lambda_i - \lambda_1)(\sigma_1 - \sigma_j)}{(\lambda_i - \sigma_j)(\lambda_i - \sigma_1)(\lambda_1 - \sigma_j)}
\end{aligned}$$

and thus, the new ith row ($i \geq 2$) is

$$0 \quad \frac{(\lambda_i - \lambda_1)(\sigma_1 - \sigma_2)}{(\lambda_i - \sigma_2)(\lambda_i - \sigma_1)(\lambda_1 - \sigma_2)} \quad \cdots \quad \frac{(\lambda_i - \lambda_1)(\sigma_1 - \sigma_{n+1})}{(\lambda_i - \sigma_{n+1})(\lambda_i - \sigma_1)(\lambda_1 - \sigma_{n+1})}.$$

It follows that C_{n+1} is equal to the determinant

$$\begin{vmatrix}
\frac{1}{\lambda_1 - \sigma_1} & \frac{1}{\lambda_1 - \sigma_2} & \cdots & \frac{1}{\lambda_1 - \sigma_{n+1}} \\
0 & \frac{(\lambda_2 - \lambda_1)(\sigma_1 - \sigma_2)}{(\lambda_2 - \sigma_2)(\lambda_2 - \sigma_1)(\lambda_1 - \sigma_2)} & \cdots & \frac{(\lambda_2 - \lambda_1)(\sigma_1 - \sigma_{n+1})}{(\lambda_2 - \sigma_{n+1})(\lambda_2 - \sigma_1)(\lambda_1 - \sigma_{n+1})} \\
\vdots & \vdots & \vdots & \vdots \\
0 & \frac{(\lambda_n - \lambda_1)(\sigma_1 - \sigma_2)}{(\lambda_{n+1} - \sigma_2)(\lambda_{n+1} - \sigma_1)(\lambda_1 - \sigma_2)} & \cdots & \frac{(\lambda_{n+1} - \lambda_1)(\sigma_1 - \sigma_{n+1})}{(\lambda_{n+1} - \sigma_{n+1})(\lambda_{n+1} - \sigma_1)(\lambda_1 - \sigma_{n+1})}
\end{vmatrix}$$

Since the ith row ($i \geq 2$) contains the common factor

$$\frac{\lambda_i - \lambda_1}{\lambda_i - \sigma_1}$$

and the jth column ($j \geq 2$) contains the common factor

$$\frac{\sigma_1 - \sigma_j}{\lambda_1 - \sigma_j},$$

by multilinearity and by expanding the above determinant with respect to the first row, we get

$$C_{n+1} = \frac{\prod_{i=2}^{n+1}(\lambda_i - \lambda_1)(\sigma_1 - \sigma_i)}{\prod_{i=1}^{n+1}(\lambda_i - \sigma_1)\prod_{j=2}^{n+1}(\lambda_1 - \sigma_j)}\, D_n,$$

where

$$D_n = \begin{vmatrix} \dfrac{1}{\lambda_2 - \sigma_2} & \dfrac{1}{\lambda_2 - \sigma_3} & \cdots & \dfrac{1}{\lambda_2 - \sigma_{n+1}} \\ \dfrac{1}{\lambda_3 - \sigma_2} & \dfrac{1}{\lambda_3 - \sigma_3} & \cdots & \dfrac{1}{\lambda_3 - \sigma_{n+1}} \\ \vdots & \vdots & \vdots & \vdots \\ \dfrac{1}{\lambda_{n+1} - \sigma_2} & \dfrac{1}{\lambda_{n+1} - \sigma_3} & \cdots & \dfrac{1}{\lambda_{n+1} - \sigma_{n+1}} \end{vmatrix}$$

Using the induction hypothesis applied to D_n, we get the desired formula,

$$C_{n+1} = \frac{\prod_{i=2}^{n+1}\prod_{j=1}^{i-1}(\lambda_i - \lambda_j)(\sigma_j - \sigma_i)}{\prod_{i=1}^{n+1}\prod_{j=1}^{n+1}(\lambda_i - \sigma_j)}.$$

Since we assumed that the σ_j are all distinct, that the λ_i are all distinct and that $\lambda_i \neq \sigma_j$, for all i, j, we conclude that

$$\det \begin{pmatrix} \dfrac{1}{\lambda_1 - \sigma_1} & \dfrac{1}{\lambda_1 - \sigma_2} & \cdots & \dfrac{1}{\lambda_1 - \sigma_k} \\ \dfrac{1}{\lambda_2 - \sigma_1} & \dfrac{1}{\lambda_2 - \sigma_2} & \cdots & \dfrac{1}{\lambda_2 - \sigma_k} \\ \vdots & \vdots & \vdots & \vdots \\ \dfrac{1}{\lambda_k - \sigma_1} & \dfrac{1}{\lambda_k - \sigma_2} & \cdots & \dfrac{1}{\lambda_k - \sigma_k} \end{pmatrix} \neq 0,$$

which proves that $y^1, \ldots, y^k$ are linearly independent. □

Unfortunately, the y^i are generally not pairwise orthogonal.

3 Solution in the General Case (Multiple Eigenvalues)

Fortunately, when the matrix, A, has multiple eigenvalues, Theorem 3.1 can still be proved pretty much as before except for some notational complications.

Theorem 3.1. *If the $n \times n$ symmetric matrix A has p distinct eigenvalues, $\sigma_1 > \sigma_2 > \cdots > \sigma_p$, each with multiplicity $k_i \geq 1$, with $k_1 + \cdots + k_p = n$, then there are $2m$ values of λ, say $\lambda_1 > \lambda_2 \geq \lambda_3 > \cdots > \lambda_{2m-2} \geq \lambda_{2m-1} > \lambda_{2m}$, with $1 \leq m \leq p$, such that the system*

$$\begin{aligned} (\lambda I - A)x &= b \\ x^\top x &= 1 \end{aligned}$$

(with $b \neq 0$) has a solution, (λ, x). As a consequence, there are at least two and at most $2p$ values of λ for which the Lagrangian

$$L(x, \lambda) = x^\top A x + 2x^\top b - \lambda(x^\top x - 1)$$

has a critical point (x, λ), but there may be infinitely many x for which (x, λ) is a critical point. Furthermore, the distinct eigenvalues $\sigma_1 > \sigma_2 > \cdots > \sigma_p$ of A separate the λ's, which means that

1. *$\lambda_1 \geq \sigma_1$.*
2. *$\lambda_{2m} \leq \sigma_p$.*
3. *For every λ_i, with $2 \leq i \leq 2m-1$, either $\lambda_i = \sigma_j$ for some j with $1 \leq j \leq p$, or there is some j, with $1 \leq j \leq p-1$, so that $\sigma_j > \lambda_i > \sigma_{j+1}$.*

Proof. If $c_i \neq 0$ for $i = 1, \ldots, n$, then the curve, $C(\Sigma, c)$, is still a kind of generalized hyperbola in $\mathbb{R}^n$, but it only has asymptotes corresponding the distinct values of the σ_i's. To simplify notation, let
$p_1 = 1$, $q_1 = k_1$, $p_i = k_1 + \cdots + k_{i-1} + 1$, $q_i = k_1 + \cdots + k_i$, for $i = 2, \ldots, p$, and let

$$J_i = \{p_i, p_i + 1, \ldots, q_i\}, \quad i = 1, \ldots, p.$$

In order for some, y, on the curve $C(\Sigma, c)$ to belong to the unit sphere, the equation

$$\sum_{i=1}^{p} \frac{\sum_{j \in J_i} c_j^2}{(\lambda - \sigma_i)^2} = 1,$$

must hold, which yields a polynomial equation of degree $2p$. Observe that the parametric equations of the curve, $C(\Sigma, c)$, can be written as p sets of equations,

$$y_j = \frac{c_j}{\lambda - \sigma_i}, \quad j \in J_i, \quad i = 1, \ldots, p.$$

This shows that the curve, $C(\Sigma, c)$, lies in the linear subspace of dimension p (an intersection of $n - p$ hyperplanes) given by the equations

$$\frac{y_{p_i}}{c_{p_i}} = \frac{y_{k_1}}{c_{k_1}} \qquad k_1 \in J_i,\ p_i < k_1, \quad i = 1, \ldots, p.$$

For each of the p subset J_i, there are $k_i - 1$ linearly independent equations and so, a total of $n - p$ equations.

Each parametric equation of $C(\Sigma, c)$ is still an injective monotonic function of λ (for $\lambda \neq \sigma_i$), so the branch of the curve passing through 0 intersects the unit sphere in exactly two points:

1. One point when λ varies from $-\infty$ to σ_p, corresponding to some $\lambda_{2m} < \sigma_p$.

2. One point when λ varies from $+\infty$ to σ_1, corresponding to some $\lambda_1 > \sigma_1$.

If another branch of $C(\Sigma, c)$ corresponding to $\lambda \in (\sigma_{j+1}, \sigma_j)$ $(1 \leq j \leq p-1)$ intersects the unit sphere, then it will do so in two points corresponding to some values, λ_{2i} and λ_{2i+1}, so that $\sigma_j > \lambda_{2i} \geq \lambda_{2i+1} > \sigma_{j+1}$.

Thus, when $c_i \neq 0$ for $i = 1, \ldots, n$, the system

$$\begin{aligned} (\lambda I - \Sigma)y &= c \\ y^\top y &= 1 \end{aligned}$$

has at least two and at most $2p$ solutions and the eigenvalues of A separate the solutions in λ.

Let us now consider the case where $c_j = 0$, for some j. For every i, with $1 \leq i \leq p$, define the two disjoint subsets, Z_i and H_i, of J_j, by

$$\begin{aligned} Z_i &= \{j \in J_i \mid c_j = 0\} \\ H_i &= \{j \in J_i \mid c_j \neq 0\}, \end{aligned}$$

and let $s_i = |Z_i|$, $r_i = |H_i|$, $s = s_1 + \cdots + s_p$ and $r = r_1 + \cdots + r_p$. We also let q be the number of subsets, H_i, such that $H_i \neq \emptyset$. Note that some of the Z_i and H_i may empty, but not at the same time. Of course, $r + s = n$.

Again, there are two cases.

Case 1. There is a solution, y, of the system $(\lambda I - \Sigma)y = c$ for which $y_j = 0$, for all $j \in Z_i$, for $i = 1, \ldots, p$. In this case, $C(\Sigma, c)$ is a curve in the subspace of dimension $n - s - (r - q) = q$ determined by the equations,

$$\begin{aligned} y_j &= 0, \qquad j \in Z_i, \quad i = 1, \ldots, p \\ \frac{y_{k_1}}{c_{k_1}} &= \frac{y_{k_2}}{c_{k_2}} \qquad k_1 = \min(H_i), k_2 \in H_i,\ k_1 < k_2, \quad i = 1, \ldots, p. \end{aligned}$$

(For each of the q nonempty subset, H_i, there are $r_i - 1$ linearly independent equations and so, a total of $r - q$ equations of the second type). This curve intersects the unit sphere iff the equation

$$\sum_{i=1}^{q} \frac{\sum_{j \in H_i} c_j^2}{(\lambda - \sigma_i)^2} = 1$$

holds, which yields a polynomial equation of degree $2q$. The rest of the discussion is completely analogous to the corresponding discussion in the proof of Theorem 2.1. There are at least two and at most $2q$ solutions and the distinct eigenvalues of A separate the solutions in λ.

Case 2. There is a solution, y, of the system $(\lambda I - \Sigma)y = c$ such that $y_j \neq 0$ for some $j \in Z_k$ and for some k with $1 \leq k \leq p$. In this case, we must have $\lambda = \sigma_k$, a multiple solution, and also $H_k = \emptyset$. The system $(\lambda I - \Sigma)y = c$ defines a subspace of dimension $n - (s - s_k) - (r - q) = s_k + q$, defined by the equations

$$\begin{aligned} y_j &= 0, \qquad j \in Z_i, \quad i = 1, \ldots, p,\ i \neq k \\ \frac{y_{k_1}}{c_{k_1}} &= \frac{y_{k_2}}{c_{k_2}}, \qquad k_1 = \min(H_i), k_2 \in H_i,\ k_1 < k_2, \quad i = 1, \ldots, p. \end{aligned}$$

This subspace intersects the unit sphere iff the equation

$$\sum_{j \in Z_k} y_j^2 + \sum_{i=1}^{p} \frac{\sum_{j \in H_i} c_j^2}{(\lambda - \sigma_i)^2} = 1$$

has a solution with $\sum_{j \in Z_k} y_j^2 \neq 0$, which is the case iff

$$\frac{\sum_{j \in H_i} c_j^2}{(\lambda - \sigma_i)^2} < 1.$$

In general, if $s_k > 1$, there are infinitely many solutions in y.

In all cases, we proved that there are at least two solutions in λ. Since there are at most $2q$ solutions in λ in Case 1, and since for every eigenvalue σ_k which is a solution in Case 2 we must have $H_k = \emptyset$, which means that the index k corresponds to a $Z_k \neq \emptyset$, since there are $p - q$ such subsets, there are at most $p - q$ such eigenvalues (it is possible that $Z_i \neq \emptyset$ and $H_i \neq \emptyset$ for some i) so there are at most $2q + p - q = p + q \leq 2p$ solutions in λ and possibly infinitely many solutions in y. The part of the proof that shows that the distinct eigenvalues of A separate the λ's is similar to the proof in Theorem 2.1. □

Remark: The seemingly more general problem

$$\begin{aligned} &\text{maximize} && x^\top A x + 2x^\top b \\ &\text{subject to} && x^\top B x = 1, \quad x \in \mathbb{R}^n, \end{aligned}$$

where A is an arbitrary symmetric matrix and B is symmetric positive definite can be reduced to our problem. Indeed, the Lagrangian of the above problem is

$$L(x, \lambda) = x^\top A x + 2x^\top b - \lambda(x^\top B x - 1)$$

and necessary conditions for this Lagrangian to have a critical point are

$$\begin{aligned} Ax + b - \lambda Bx &= 0 \\ x^\top Bx &= 1, \end{aligned}$$

which can be written as

$$\begin{aligned} (\lambda B - A)x &= b \\ x^\top Bx &= 1. \end{aligned}$$

Since B is symmetric positive definite, both $B^{1/2}$ and $B^{-1/2}$ exist so if we make the change of variable,

$$x' = B^{1/2}x,$$

we have $x = B^{-1/2}x'$ and our system becomes

$$\begin{aligned} (\lambda B^{1/2} - AB^{-1/2})x' &= b \\ x^\top B^{1/2}B^{1/2}x &= 1, \end{aligned}$$

which, after multiplying on the left by $B^{-1/2}$, yields

$$\begin{aligned} (\lambda I - B^{-1/2}AB^{-1/2})x' &= B^{-1/2}b \\ x'^\top x' &= 1. \end{aligned}$$

Therefore, we are back to our original problem with the symmetric matrix, $A' = B^{-1/2}AB^{-1/2}$, and the vector, $b' = B^{-1/2}b$.

Observe that for computational reasons, if might be preferable to use a Cholesky decomposition, $B = CC^\top$, where C is a lower triangular matrix. In this case, the system becomes

$$\begin{aligned} (\lambda CC^\top - A)x &= b \\ x^\top CC^\top x &= 1, \end{aligned}$$

and if we let $x' = C^\top x$ and multiply on the left by C^{-1}, using the fact that $x = (C^\top)^{-1}x'$, we get

$$\begin{aligned} (\lambda I - C^{-1}A(C^\top)^{-1})x' &= C^{-1}b \\ x'^\top x' &= 1, \end{aligned}$$

which is our original problem with $A' = C^{-1}A(C^\top)^{-1} = C^{-1}A(C^{-1})^\top$, a symmetric matrix. Since C is lower triangular, C^{-1} is generally cheaper to compute than $B^{-1/2}$.

4 Local Study of the Critical Points of the Lagrangian

If (λ, x) is any critical point of the Lagrangian $L(x, \lambda)$, that is, if (λ, x) is any solution of the system

$$\begin{aligned}(\lambda I - A)x &= b \\ x^\top x &= 1,\end{aligned}$$

then $b = \lambda x - Ax$, and we have

$$\begin{aligned}f(x) &= x^\top Ax + 2x^\top b \\ &= x^\top Ax + 2x^\top(\lambda x - Ax) \\ &= 2\lambda - x^\top Ax\end{aligned}$$

and since $Ax = \lambda x - b$, we also have

$$\begin{aligned}f(x) &= 2\lambda - x^\top Ax \\ &= \lambda + x^\top b.\end{aligned}$$

Since the function $f(x) = x^\top Ax + 2x^\top b$ is continuous (and differentiable) on the unit sphere, which is compact, the function f has a minimum and a maximum and both of them are achieved. Furthermore, at a local extremum, the Lagrangian

$$L(x, \lambda) = x^\top Ax + 2x^\top b - \lambda(x^\top x - 1)$$

must have a critical point.

We can obtain more information on the critical points of the Lagrangian by computing $f(v) - f(u)$, where u is a critical point of $L(u, \lambda)$.

Proposition 4.1. *If (u, λ) is a critical point of the Lagrangian,*

$$L(x, \lambda) = x^\top Ax + 2x^\top b - \lambda(x^\top x - 1),$$

so that $Au - \lambda u + b = 0$ and $u^\top u = 1$ and if v is any other point on the unit sphere ($v^\top v = 1$), then

$$f(v) - f(u) = (v - u)^\top (A - \lambda I)(v - u).$$

Proof. We have

$$f(v) - f(u) = v^\top Av + 2v^\top b - u^\top Au - 2u^\top b$$

so we need to compute $v^\top Av - u^\top Au$. Observe that as A is symmetric, we have

$$\begin{aligned}(v - u)^\top A(v - u) &= v^\top Av - u^\top Av - v^\top Au + u^\top Au \\ &= v^\top Av - 2u^\top Av + u^\top Au,\end{aligned}$$

so we have

$$\begin{aligned} v^\top Av - u^\top Au &= (v-u)^\top A(v-u) + 2u^\top Av - 2u^\top Au \\ &= (v-u)^\top A(v-u) + 2u^\top A(v-u) \end{aligned}$$

and thus,

$$\begin{aligned} f(v) - f(u) &= v^\top Av - u^\top Au + 2(v-u)^\top b \\ &= (v-u)^\top A(v-u) + 2u^\top A(v-u) + 2(v-u)^\top b \\ &= (v-u)^\top A(v-u) + 2(Au)^\top (v-u) + 2b^\top (v-u) \\ &= (v-u)^\top A(v-u) + 2(Au+b)^\top (v-u) \end{aligned}$$

and since $Au + b = \lambda u$, we get

$$\begin{aligned} f(v) - f(u) &= (v-u)^\top A(v-u) + 2(Au+b)^\top (v-u) \\ &= (v-u)^\top A(v-u) + 2\lambda u^\top (v-u). \end{aligned}$$

However, the computation of $v^\top Av - u^\top Au$ with $A = I$ yields

$$\begin{aligned} 0 &= 1 - 1 \\ &= v^\top v - u^\top u \\ &= (v-u)^\top (v-u) + 2u^\top (v-u) \end{aligned}$$

so

$$2u^\top (v-u) = -(v-u)^\top (v-u)$$

and we finally get

$$\begin{aligned} f(v) - f(u) &= (v-u)^\top A(v-u) + 2\lambda u^\top (v-u) \\ &= (v-u)^\top A(v-u) - \lambda (v-u)^\top (v-u) \\ &= (v-u)^\top (A - \lambda I)(v-u), \end{aligned}$$

that is,

$$f(v) - f(u) = (v-u)^\top (A - \lambda I)(v-u),$$

as claimed. □

Remark: It is easy to see that the computation carried out in Proposition 4.1 applies to the more general Lagrangian,

$$L(x, \lambda) = x^\top Ax + 2x^\top b - \lambda(x^\top Bx - 1),$$

where B is any symmetric matrix, and we get

$$f(v) - f(u) = (v-u)^\top (A - \lambda B)(v-u),$$

for any critical pair, (λ, u), of L and any v such that $v^\top Bv = 1$. However, we have to assume that B is invertible in order to ensure the validity of the necessary condition for the Lagrangian to have a critical point, namely $\nabla L(u, \lambda) = 0$. We have to further assume that B is positive definite in order to reduce the problem to the situation where $x^\top x = 1$.

Using Proposition 4.1, we can find a necessary and sufficient condition for a critical point (u, λ), of the Lagrangian $L(u, \lambda)$, to correspond to a maximum of the function $f(x) = x^\top Ax + 2x^\top b$.

Proposition 4.2. *A critical point, (u, λ), of the Lagrangian,*

$$L(x, \lambda) = x^\top Ax + 2x^\top b - \lambda(x^\top x - 1)$$

corresponds to the maximum of the function $f(x) = x^\top Ax + 2x^\top b$ on the unit sphere iff $\lambda \geq \sigma_i$, for all eigenvalues σ_i of A.

Proof. Assume that the eigenvalues of the symmetric matrix, A, are $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n$, written in decreasing order and let $(e_1, \ldots, e_n)$ be an orthonormal basis of eigenvectors. For any point, v, on the unit sphere, if we write $v - u = \sum_{i=1}^n z_i e_i$, then by Proposition 4.1, we have

$$\begin{aligned} f(v) - f(u) &= (v-u)^\top (A - \lambda I)(v-u) \\ &= \left(\sum_{i=1}^n z_i e_i^\top\right)\left(\sum_{j=1}^n (\sigma_j - \lambda) z_j e_j\right) \\ &= \sum_{i=1}^n (\sigma_i - \lambda) z_i^2. \end{aligned}$$

From the above, we see that if $\lambda \geq \sigma_i$, for $i = 1, \ldots, n$, then

$$\sum_{i=1}^n (\sigma_i - \lambda) z_i^2 \leq 0$$

and $f(u)$ is indeed the maximum of f on the unit sphere.

Conversely, assume that (u, λ) corresponds to the maximum of f on the unit sphere. If $\lambda < \sigma_1$, we will prove that we can find some v on the unit sphere so that $f(v) - f(u) > 0$, and thus, $f(u)$ is not the maximum of f on the unit sphere, a contradiction.

Case 1. $u^\top e_1 \neq 0$.

If $\lambda \leq \sigma_2$, then pick $v = u + z_1 e_1 + z_2 e_2$. Since $v^\top v = u^\top u = 1$, we must have

$$z_1^2 + 2\alpha z_1 + z_2^2 + 2\beta z_2 = 0.$$

with $\alpha = u^\top e_1$ and $\beta = u^\top e_2$. As a quadratic equation in z_1, the discriminant of this equation is

$$\Delta = 4\alpha^2 - 4z_2(z_2 + 2\beta)$$

and since $\alpha \neq 0$, we can pick $z_2 \neq 0$ small enough so that $\Delta > 0$ and $z_2 + 2\beta \neq 0$, and we find a nonzero solution for z_1. For this choice of v, as $\lambda < \sigma_1$, $\lambda \leq \sigma_2$ and $z_1 \neq 0$, we have

$$f(v) - f(u) = (\sigma_1 - \lambda)z_1^2 + (\sigma_2 - \lambda)z_2^2 > 0,$$

as claimed.

If $\sigma_1 > \lambda > \sigma_2$, then pick some positive real, ρ, so that

$$\rho^2 < -\frac{\sigma_1 - \lambda}{\sigma_2 - \lambda}.$$

Then, let

$$v = u + z_1 e_1 + \rho z_1 e_2.$$

Since $v^\top v = u^\top u = 1$, we must have

$$z_1^2 + 2\alpha z_1 + \rho^2 z_1^2 + 2\rho\beta z_1 = 0,$$

with $\alpha = u^\top e_1$ and $\beta = u^\top e_2$, that is,

$$((\rho^2 + 1)z_1 + 2(\alpha + \rho\beta))z_1 = 0$$

Since $\alpha \neq 0$, we can choose $\rho > 0$ small enough so that $\alpha + \rho\beta \neq 0$ and then we can pick

$$z_1 = -\frac{2(\alpha + \rho\beta)}{\rho^2 + 1} \neq 0.$$

For such a choice of v, we get

$$\begin{aligned} f(v) - f(u) &= (\sigma_1 - \lambda)z_1^2 + (\sigma_2 - \lambda)\rho^2 z_1^2 \\ &= ((\sigma_1 - \lambda) + (\sigma_2 - \lambda)\rho^2)z_1^2 > 0, \end{aligned}$$

since $z_1 \neq 0$ and

$$\rho^2 < -\frac{\sigma_1 - \lambda}{\sigma_2 - \lambda}.$$

Case 2. $u^\top e_1 = 0$.

Let k be the smallest index so that $u^\top e_i \neq 0$ and write $\beta = u^\top e_k$. If $\lambda \leq \sigma_k$, then pick $v = u + z_1 e_1 + z_k e_k$. Since $v^\top v = u^\top u = 1$, we must have

$$z_1^2 + z_k^2 + 2\beta z_k = 0,$$

since $\alpha = u^\top e_1 = 0$. As a quadratic equation in z_k, the discriminant of this equation is

$$\Delta = 4\beta^2 - 4z_1^2$$

and since $\beta \neq 0$, we can pick $z_1 \neq 0$ small enough so that $\Delta > 0$ and we find a nonzero solution for z_k. For this choice of v, as $\lambda < \sigma_1$, $\lambda \leq \sigma_k$ and $z_1 \neq 0$, we have

$$f(v) - f(u) = (\sigma_1 - \lambda)z_1^2 + (\sigma_k - \lambda)z_k^2 > 0,$$

as claimed.

If $\sigma_1 > \lambda > \sigma_k$, then pick some positive real, ρ, so that

$$\rho^2 < -\frac{\sigma_1 - \lambda}{\sigma_k - \lambda}.$$

Then, let

$$v = u + z_1 e_1 + \rho z_1 e_k.$$

Since $v^\top v = u^\top u = 1$, we must have

$$z_1^2 + \rho^2 z_1^2 + 2\rho\beta z_1 = 0,$$

that is,

$$((\rho^2 + 1)z_1 + 2\rho\beta)z_1 = 0$$

Since $\beta \neq 0$, we can pick

$$z_1 = -\frac{2\rho\beta}{\rho^2 + 1} \neq 0.$$

For such a choice of v, we get

$$\begin{aligned} f(v) - f(u) &= (\sigma_1 - \lambda)z_1^2 + (\sigma_k - \lambda)\rho^2 z_1^2 \\ &= ((\sigma_1 - \lambda) + (\sigma_k - \lambda)\rho^2)z_1^2 > 0, \end{aligned}$$

since $z_1 \neq 0$ and

$$\rho^2 < -\frac{\sigma_1 - \lambda}{\sigma_k - \lambda}.$$

Therefore, in all cases, we proved that if $\lambda < \sigma_1$, then $f(u)$ is not the maximum of f on the unit sphere, a contradiction. □

As a corollary, since the function, $f(x) = x^\top Ax + 2x^\top b$, achieves its maximum on the unit sphere, we obtain the following theorem:

Theorem 4.3. *There are at least two and at most $2n$ scalars λ, so that (x, λ) is a critical point of the Lagrangian*

$$L(x, \lambda) = x^\top Ax + 2x^\top b - \lambda(x^\top x - 1),$$

and for the largest of these λ's, we have $\lambda \geq \sigma_i$, for all eigenvalues σ_i of A. The maximum of the function $f(x) = x^\top Ax + 2x^\top b$ on the unit sphere is achieved for all the critical points (x, λ), such that $\lambda \geq \sigma_i$, for all eigenvalues σ_i of A.

5 Finding the Intersections of the Curve $C(\Sigma, c)$, with the Unit Sphere

It remains to give an algorithm for finding the intersection points of the curve, $C(\Sigma, c)$, with the unit sphere. From the discussion in Section 5, we may assume that the entries in Σ (the eigenvalues of the matrix A) are all distinct and that $c_i \neq 0$, for $i = 1, \ldots, n$. The curve, $C(\Sigma, c)$, is given by the parametric equations

$$\begin{aligned} y_1 &= \frac{c_1}{\lambda - \sigma_1} \\ \vdots & \quad \vdots \\ y_n &= \frac{c_n}{\lambda - \sigma_n}. \end{aligned}$$

We know that this curve consists of n branches and that each of the $n-1$ branches with $\lambda \in (\sigma_{i+1}, \sigma_i)$, for $i = 1, \ldots, n-1$, intersects the unit sphere at most twice and that the branch of the curve corresponding to $\lambda \in (-\infty, \sigma_n) \cup (\sigma_1, +\infty)$ intersects the unit sphere twice.

One way to proceed is to introduce the function

$$f(\lambda) = \sum_{i=1}^{n} \frac{c_i^2}{(\lambda - \sigma_i)^2} - 1,$$

called the *secular function*, and to find the solutions of the equation

$$f(\lambda) = 0.$$

This is the approach followed by Gander, Golub and von Matt [5], who investigate several methods to solve this equation.

It is conceivable that a a more direct bissection method should exist, but we are not aware of such a method.

6 Related Work

The two earliest references that I found dealing with Problem 1 and a closely related problem are:

(1) Burrows [1] (1966), which deals with Problem 1.

(2) Forsythe and Golub [3] (1965), which deals with problem of finding the stationary values of

$$\Phi(x) = (x - b)^* A (x - b),$$

where A is an Hermitian matrix.

Burrows [1] uses exactly the method proposed in this paper, namely, to diagonalize A with respect to a unitary matrix and then, looking for the critical points of the Lagrangian, he gets a system of the form

$$\begin{aligned}(\lambda I - \Sigma)y &= c \\ y^*y &= 1,\end{aligned}$$

where $y \in \mathbb{C}^n$. Since Burrows deals with an Hermitian matrix, he need the fact that the solutions, λ, are real and for this, he refers to Forsythe and Golub [3] where this is proved. Burrows observes that the λ's are the solutions of the equation

$$\sum_{i=1}^{n} \frac{|c_i|^2}{|\lambda - \sigma_i|^2} = 1$$

(where only terms for which $c_i \neq 0$ appear) and proves that the maximum of $x^*Ax + x^*b + b^*x$ arise for the largest λ. Since the problem is cast over $\mathbb{C}$, the geometric interpretation as the intersection of a curve with the unit sphere is missed.

Forsythe and Golub [3] deals with problem of finding the local extrema of

$$\Phi(x) = (x - b)^*A(x - b),$$

where A is an Hermitian matrix. As Burrows points out,

$$(x - b)^*A(x - b) = x^*Ax - x^*Ab - b^*Ax + b^*Ab$$

so, if we let $a = Ab$, the problem is equivalent to finding the local extrema of

$$x^*Ax - x^*a - a^*x,$$

which is Problem 1. In fact, Problem 1 is more general because if A is not invertible, then a can't be expressed as $a = Ab$. Interestingly, while Burrows cites Forsythe and Golub, the converse is not true.

Forsythe and Golub also diagonalize A by picking some orthonormal basis, $(u_1, \ldots, u_n)$, of eigenvectors of A, and then express x and b over this basis. After setting the gradient of the Lagrangian to zero, they also get a system of the form

$$\begin{aligned}(A - \lambda I)x &= Ab \\ x^*x &= 1,\end{aligned}$$

which yields the system

$$\begin{aligned}(\sigma_i - \lambda)x_i &= \sigma_i b_i, \quad i = 1, \ldots, n \\ x^*x &= 1.\end{aligned}$$

The fact that the σ_i's occur on the right-hand side complicates the discussion. Essentially, the λ's are solutions of the equation

$$\sum_{i=1}^{n} \frac{\sigma_i^2 |b_i|^2}{|\lambda - \sigma_i|^2} = 1$$

Forsythe and Golub go through a thorough discussion of the two cases, (a) $\sigma_i b_i \neq 0$ for all i, and (b) $\sigma_i b_i = 0$ for some i, which culminates in the main theorem stated in Section 4. They also note that there are at least 2 and at most $2n$ solutions but they do not conduct a detailed study depending on the multiplicity of the σ_i's (as we do) and they only treat the case $n = 2$ in detail. It is worth noting that because the problem is cast over $\mathbb{C}$, it takes some work (the entire Section 8) to prove that the Lagrange multipliers corresponding to local extrema are real. However, this is a trivial consequence of the equivalence of Problem 1 and Problem 2, as we showed. Section 6 proposes a geometric interpretation of the problem but it is different from ours (the intersection of the curve $C(\Sigma, c)$ with the unit sphere).

Forsythe and Golub [3] is a bit lengthy and some complications caused by casting the problem over $\mathbb{C}$ can be avoided. In a short paper, Spjøtvoll [9] (1972) tightens up the treatment of the cases in Forsythe and Golub [3] and gives a shorter proof of their main theorem (from Section 4). Spjøtvoll [9] also shows how to solve the problem of finding the local extrema of $\Phi(x) = (x - b)^* A(x - b)$, subject to $x^* x \leq 1$.

Among other things, Golub [6] (1973) considers the following problem: Given an $n \times n$ real symmetric matrix, A, and an $n \times p$ matrix, C,

$$\begin{array}{ll} \text{minimize} & x^\top A x \\ \text{subject to} & x^\top x = 1, \quad x \in \mathbb{R}^n \\ & C^\top x = 0. \end{array}$$

Golub shows that the linear constraint, $C^\top x = 0$, can be eliminated as follows: If we use a QR decomposition of C, by permuting the columns, we may assume that

$$C = Q^\top \begin{pmatrix} R & S \\ 0 & 0 \end{pmatrix} \Pi,$$

where R is an $r \times r$ invertible upper triangular matrix and S is an $r \times (p-r)$ matrix (assuming C has rank r). Then, if we let

$$x = Q^\top \begin{pmatrix} y \\ z \end{pmatrix},$$

where $y \in \mathbb{R}^r$ and $z \in \mathbb{R}^{n-r}$, then $C^\top x = 0$ becomes

$$\Pi^\top \begin{pmatrix} R^\top & 0 \\ S^\top & 0 \end{pmatrix} Q x = \Pi^\top \begin{pmatrix} R^\top & 0 \\ S^\top & 0 \end{pmatrix} \begin{pmatrix} y \\ z \end{pmatrix} = 0,$$

which implies $y = 0$, and every solution of $C^\top x = 0$ is of the form

$$x = Q^\top \begin{pmatrix} 0 \\ z \end{pmatrix}.$$

Our original problem becomes

$$\begin{array}{ll} \text{minimize} & (y^\top, z^\top) Q A Q^\top \begin{pmatrix} y \\ z \end{pmatrix} \\ \text{subject to} & z^\top z = 1, \quad z \in \mathbb{R}^{n-r} \\ & y = 0, \quad y \in \mathbb{R}^r. \end{array}$$

Thus, the constraint $C^\top x = 0$ has been eliminated and if we write

$$QAQ^\top = \begin{pmatrix} G_{11} & G_{12} \\ G_{12}^\top & G_{22} \end{pmatrix}$$

our problem becomes

$$\begin{array}{ll} \text{minimize} & z^\top G_{22} z \\ \text{subject to} & z^\top z = 1, \quad z \in \mathbb{R}^{n-r}, \end{array}$$

a standard eigenvalue problem. Observe that if we let

$$J = \begin{pmatrix} 0 & 0 \\ 0 & I_{n-r} \end{pmatrix},$$

then

$$JQAQ^\top J = \begin{pmatrix} 0 & 0 \\ 0 & G_{22} \end{pmatrix}$$

and if we set

$$P = Q^\top J Q,$$

then

$$PAP = Q^\top J Q A Q^\top J Q.$$

Now, $Q^\top JQAQ^\top JQ$ and $JQAQ^\top J$ have the same eigenvalues, so PAP and $JQAQ^\top J$ also have the same eigenvalues. It follows that the solutions of our optimization problem are among the eigenvalues of $K = PAP$, and at least r of those are 0. Using the fact that $CC^\dagger$ is the projection onto the range of C, where $C^\dagger$ is the pseudo-inverse of C, it can also be shown that

$$P = I - CC^\dagger,$$

the projection onto the kernel of $C^\top$. In particular, when $n \geq p$ and C has full rank (the columns of C are linearly independent), then we know that $C^\dagger = (C^\top C)^{-1} C^\top$ and

$$P = I - C(C^\top C)^{-1} C^\top.$$

This fact is used by Cour and Shi [2] and implicitly by Yu and Shi [10].

The paper by Gander, Golub and von Matt [5] (1989) gives a very detailed solution of Problem 2 and is closely related to what we have done in this paper.

Gander, Golub and von Matt consider the following problem: Given an $(n+m)\times(n+m)$ real symmetric matrix, A, (with $n > 0$), an $(n+m)\times m$ matrix, N, with full rank and a nonzero vector, $t \in \mathbb{R}^m$, with $\left\|(N^\top)^\dagger t\right\| < 1$ (where $(N^\top)^\dagger$ denotes the pseudo-inverse of $N^\top$)

$$\begin{array}{ll} \text{minimize} & x^\top A x \\ \text{subject to} & x^\top x = 1, \quad x \in \mathbb{R}^{n+m} \\ & N^\top x = t. \end{array}$$

This is a generalization of the problem considered in Golub [6], since $t \neq 0$. The condition $\left\|(N^\top)^\dagger t\right\| < 1$ ensures that the problem has a solution and is not trivial. The authors begin by proving that the affine constraint, $N^\top x = t$, can be eliminated. One way to do so is to use a QR decomposition of N. If

$$N = P\begin{pmatrix} R \\ 0 \end{pmatrix}$$

where P is an orthogonal matrix and R is an $m \times m$ invertible upper triangular matrix, then if we observe that

$$\begin{aligned} x^\top A x &= x^\top P P^\top A P P^\top x \\ N^\top x &= (R^\top, 0) P^\top x = t \\ x^\top x &= x^\top P P^\top x = 1, \end{aligned}$$

if we write

$$P^\top A P = \begin{pmatrix} B & \Gamma^\top \\ \Gamma & C \end{pmatrix}$$

and

$$P^\top x = \begin{pmatrix} y \\ z \end{pmatrix},$$

then we get

$$\begin{aligned} x^\top A x &= y^\top B y + 2 z^\top \Gamma y + z^\top C z \\ R^\top y &= t \\ y^\top y + z^\top z &= 1. \end{aligned}$$

Thus,

$$y = (R^\top)^{-1} t$$

and if we write

$$s^2 = 1 - y^\top y > 0$$

and

$$b = \Gamma y,$$

we get the simplified problem

$$\begin{array}{ll} \text{minimize} & z^\top C z + 2 z^\top b \\ \text{subject to} & z^\top z = s^2, \quad z \in \mathbb{R}^m, \end{array}$$

which is equivalent to Problem 2 (since $\min z = \max -z$) except that the right hand-side of the constraint is s^2 rather than 1, but this is inessential.

Then, exactly as I did, Gander, Golub and von Matt write the necessary conditions for the Lagrangian to have a critical point and find the unescapable system

$$\begin{aligned} \lambda z - Cz &= b \\ z^\top z &= s^2. \end{aligned}$$

Then, the above system is reduced to the canonical form

$$\begin{aligned} (\lambda I - \Sigma) y &= c \\ y^\top y &= s^2 \end{aligned}$$

by diagonalizing C using an orthogonal matrix. Gander, Golub and von Matt introduce the function

$$f(\lambda) = \sum_{i=1}^{n} \frac{c_i^2}{(\lambda - \sigma_i)^2} - s^2,$$

that they call the *secular function*, and, of course, they show that the solutions of our problem are the solutions of the equation

$$f(\lambda) = 0,$$

that they call the *(explicit) secular equation*. They discuss the various cases having to do with $c_i \neq 0$ or $c_i = 0$ and they show that the minimum is achieved for any λ such that $\lambda \leq \sigma_i$ for all i.

The authors discuss the *implicit secular equation*, which is a way of solving for the smallest λ which does not require finding the eigenvalues of C. They also discuss the conditioning of the explicit secular equation

$$\sum_{i=1}^{n} \frac{c_i^2}{(\lambda - \sigma_i)^2} - s^2 = 0.$$

In Section 4, the authors present an iterative method for finding the smallest solution of the explicit secular equation.

In Section 5, the authors observe that when a solution, λ, is different from all the eigenvalues of C, then z is given by

$$z = (\lambda I - C)^{-1} b$$

and since $z^\top z = 1$, λ is a solution of the equation

$$b^\top (\lambda I - C)^{-2} b - s^2 = 0.$$

If we let $\gamma = (\lambda I - C)^{-2} b$, then it is easy to see that γ is a solution of the equation

$$(\lambda I - C)^2 \gamma = \frac{1}{s^2} bb^\top \gamma,$$

a *quadratic eigenvalue problem*. Then, the authors discuss the equivalence of the solvability of the system

$$\begin{aligned} \lambda z - Cz &= b \\ z^\top z &= s^2 \end{aligned}$$

and of the quadratic eigenvalue problem

$$(\lambda I - C)^2 \gamma = \frac{1}{s^2} bb^\top \gamma$$

and they show that this problem reduces to a standard eigenvalue problem. Indeed, if we write

$$\eta = (\lambda I - C)\gamma.$$

then $\binom{\gamma}{\delta}$ is an eigenvector associated with λ for the following matrix:

$$\begin{pmatrix} C & -I \\ \dfrac{-1}{s^2} bb^\top & C \end{pmatrix}$$

The last section of the paper is devoted to a discussion of the numerical results. It turns out that the quadratic eigenvalue problem performs very badly. The explicit and the implicit secular equations achieve the same degree of accuracy but the implicit secular equation is generally not cheaper than the explicit secular equation.

Two papers, Sorensen [8] (1982) and Moré and Sorensen [7] (1983) discuss the quadratic optimization problem of minimizing a quadratic function

$$\psi(x) = x^\top Ax + 2x^\top b$$

in which the constraint, $x^\top x = 1$, is relaxed to the convex constraint,

$$\|x\| \leq \Delta,$$

for some positive number, $\Delta > 0$. Because of the inequality constraint, which can be written as

$$x^\top x \leq \Delta^2,$$

the necessary conditions for the Lagrangian

$$L(x, \lambda) = x^\top A x + 2x^\top b - \lambda(x^\top x - \Delta^2)$$

to have a critical point are the Kuhn and Tucker conditions which read:

$$(\lambda I - A)x = b,$$

for some $\lambda \leq 0$ such that

$$\lambda(x^\top x - \Delta^2) = 0.$$

Note that in $L(x, \lambda)$ we have switched the sign of the Lagrange multiplier, λ, which traditionally has the sign $+$, and this is why we get the condition $\lambda \leq 0$ as opposed to $\lambda \geq 0$ (the two papers under discussion assume that λ has a positive sign in the Lagrangian). Our choice makes the comparison with the other optimization problems simpler.

It is easy to show that $A - \lambda I$ must be positive semidefinite, which simply means that $\lambda \leq \sigma_i$, for all eigenvalues, σ_i, of A.

The equation

$$(\lambda I - A)x = b$$

shows up again but, this time, the solutions, x, may be in the interior of the ball of radius Δ. Having obtained necessary conditions for a local extremum, Sorenson also proves the following sufficient conditions (Lemma 2.8):

Assume $\lambda \in \mathbb{R}$ and $u \in \mathbb{R}^n$ satisfy the following conditions:

$$(\lambda I - A)u = b$$

and $A - \lambda I$ is positive semidefinite. Then, the following conditions hold:

(i) If $\lambda = 0$ and $\|u\| < \Delta$, then u is a minimizer of ψ in the ball of radius Δ.

(ii) The vector, u, is a minimizer of ψ on the sphere of radius $\|u\|$. In particular, if $\|u\| = \Delta$, then u is a minimizer of ψ on the sphere of radius Δ.

(iii) If $\lambda \leq 0$ and $\|u\| = \Delta$, then u is a minimizer of ψ in the ball of radius Δ.

Furthermore, if $\lambda I - A$ is positive definite, then u is unique in all three cases.

The proof uses the fact that if $(\lambda I - A)u = b$, that is, $(A - \lambda I)u = -b$, then

$$\begin{aligned} u^\top(A - \lambda I)u + 2u^\top b &= u^\top(A - \lambda I)u - 2u^\top(A - \lambda I)u \\ &= -u^\top(A - \lambda I)u \end{aligned}$$

and so

$$\begin{aligned} x^\top(A-\lambda I)x + 2x^\top b - (u^\top(A-\lambda I)u + 2u^\top b) &= x^\top(A-\lambda I)x + 2x^\top b + u^\top(A-\lambda I)u \\ &= x^\top(A-\lambda I)x - 2x^\top(A-\lambda I)u \\ &\qquad + u^\top(A-\lambda I)u \\ &= (x-u)^\top(A-\lambda I)(x-u) \end{aligned}$$

and since $A - \lambda I$ is positive semidefinite, we get

$$x^\top(A-\lambda I)x + 2x^\top b \geq u^\top(A-\lambda I)u + 2u^\top b$$

for all $x \in \mathbb{R}^n$. The above inequality implies that

$$x^\top Ax + 2x^\top b \geq u^\top Au + 2u^\top b - \lambda(u^\top u - x^\top x)$$

for all $x \in \mathbb{R}^n$ and the above result follows immediately.

Moré and Sorensen [7] characterize when the optimization problem has no solution on the boundary of the unit ball of radius Δ:

There is no solution, x, with $\|x\| = \Delta$ iff A is positive definite and if $\|A^{-1}b\| < \Delta$.

Consequently, if our optimization problem only has solutions, x, in the interior of the ball of radius Δ (that is, $\|x\| < \Delta$), then there is a unique solution given by $\lambda = 0$ and $x = -A^{-1}b$.

Otherwise, our optimization problem has a solution on the sphere of radius Δ and *we are back to Problem 2*, as studied in Section 3, except that the solutions, λ, satisfy the conditions: $\lambda \leq 0$ and $\lambda \leq \sigma_i$, for all eigenvalues, σ_i, of A.

However, the two papers by Sorensen and Moré under discussion were written before Gander, Golub and von Matt [5] and rather than thoroughly analyzing when the system

$$\begin{aligned} (\lambda I - A)x &= b \\ x^\top x &= \Delta^2 \end{aligned}$$

has solutions, Sorensen considers the problem of solving the secular equation

$$\left\|(A-\lambda I)^{-1}b\right\| = \Delta.$$

Sorensen does observe that, by diagonalizing A, we get an equation of the form

$$\sum_{i=1}^{n} \frac{c_i^2}{(\lambda - \sigma_i)^2} = \Delta^2,$$

where the terms for which $c_i = 0$ are missing. This is a rational function with second-order poles. As we know, this equation always has a solution provided that $b \neq 0$ and $\Delta > 0$.

However, there is a difficulty with this approach when the smallest solution, λ, of this equation is greater than the smallest eigenvalue, σ_n, of A, because then, $A-\lambda I$ is not positive semidefinite. This may happen when b is orthogonal to the eigenspace, E_{σ_n}, associated with σ_n. In this case, b is the range of $\sigma_n I - A$, so the equation

$$(\sigma_n I - A)x = b$$

has solutions and since we are assuming that our optimization problem has a solution, from the results of Section 3, σ_n is the solution and so, $\sigma_n < 0$. In this case, the solution, x, is not unique.

Sorenson also proves that in this case, which corresponds to Case 2 in the proof of Theorem 3.1, it is still possible to find a solution which can be expressed as

$$x = (\sigma_n I - A)^{\dagger} b + \theta w,$$

for some eigenvector, w, of A for σ_n and where θ is chosen so that $\|x\| = \Delta$ (it is easy to see that $\left\|(\sigma_n I - A)^{\dagger} b\right\| < \Delta$ must hold). However, there are numerical difficulties. This situation if called the "hard case" by Sorenson. One of the problems in the hard case is that if $A - \lambda I$ is positive definite, then $\|u\| < \Delta$. Yet, in the hard case, we are seeking a solution on the boundary.

The rest of the paper is devoted to a modification of Newton's method using a so-called "trust region method" and to its convergence.

Moré and Sorensen [7] is more algorithmically oriented. Other algorithms based on Newton's method and using the trust region method are presented and their convergence is analyzed.

Some related work is found in Gander [4] (1981), which deals with the problem of least squares with a quadratic constraint. Given a $m \times n$ matrix, A, a $p \times n$ matrix, C, some vectors $b \in \mathbb{R}^m$ and $d \in \mathbb{R}^p$, and some positive real, α, the problem is

$$\begin{array}{ll} \text{minimize} & \|Ax - b\| \\ \text{subject to} & \|Cx - d\| = \alpha, \quad x \in \mathbb{R}^n. \end{array}$$

The conditions for the Lagrangian to have a critical point are

$$\begin{aligned} (A^{\top}A + \lambda C^{\top}C)x &= A^{\top}b + \lambda C^{\top}d \\ \|Cx - d\|^2 &= \alpha^2. \end{aligned}$$

If the matrix, $A^{\top}A + \lambda C^{\top}C$, is invertible, then we obtain the secular equation,

$$\|Cx(\lambda) - d\|^2 = \alpha^2,$$

where

$$x(\lambda) = (A^{\top}A + \lambda C^{\top}C)^{-1}(A^{\top}b + \lambda C^{\top}d).$$

Using some SVD decompositions for A and C, this equation can be simplified and yields a rational function with quadratic poles. The author gives a complete characterizations of the solutions of the system

$$\begin{aligned}(A^\top A + \lambda C^\top C)x &= A^\top b + \lambda C^\top d \\ \|Cx - d\|^2 &= \alpha^2\end{aligned}$$

and goes on to solve the least squares problem with a quadratic constraint. He also shows how to deal with the inequality constraint, $\|Cx - d\|^2 \leq \alpha^2$, and various special cases of the least squares problem with a quadratic constraint.

More recently, Cour and Shi [2] have considered the following optimization problem that arises in computer vision:

$$\begin{array}{ll}\text{maximize} & \dfrac{x^\top W x + 2x^\top V + \alpha}{x^\top x + \beta} \\ \text{subject to} & Cx = b, \quad x \in \mathbb{R}^n\end{array}$$

where $\beta > 0$, W is a symmetric matrix and $V \neq 0$. The way to proceed is to "homogenize", namely to go up one dimension as we did earlier. If we let

$$\begin{aligned}\overline{W} &= \begin{pmatrix} W & V \\ V^\top & \alpha \end{pmatrix} \\ \overline{D} &= \begin{pmatrix} I & 0 \\ 0 & \beta \end{pmatrix} \\ \overline{C} &= (C, -b)\end{aligned}$$

then we obtain the problem

$$\begin{array}{ll}\text{maximize} & (x^\top, t)\overline{W}\begin{pmatrix} x \\ t \end{pmatrix} \\ \text{subject to} & (x^\top, t)\overline{D}\begin{pmatrix} x \\ t \end{pmatrix} = 2, \quad \begin{pmatrix} x \\ t \end{pmatrix} \in \mathbb{R}^{n+1} \\ & \overline{C}\begin{pmatrix} x \\ t \end{pmatrix} = 0.\end{array}$$

It is clear that (x, t) (with $t \neq 0$) is a maximum of this last problem iff x/t is a maximum of the former problem. However, *neither problem is equivalent to our problem*, since a solution, x, with $x^\top x = 1$ is obtained iff $\beta t = \pm 1$, which is false in general.

The second formulation of Cour and Shi's problem is reduced to a more standard form by making the change of variable

$$x' = \overline{D}^{1/2}\begin{pmatrix} x \\ t \end{pmatrix},$$

(which is possible since $\beta > 0$), which leads to the symmetric matrix $W' = \overline{D}^{-1/2}\overline{W}\,\overline{D}^{-1/2}$, the matrix $C' = \overline{C}\,\overline{D}^{-1/2}$ and to the problem

$$\begin{array}{ll} \text{maximize} & x'^{\top} W' x' \\ \text{subject to} & x'^{\top} x' = 1, \quad x' \in \mathbb{R}^{n+1} \\ & C' x' = 0. \end{array}$$

This last problem reduces to a standard eigenvalue problem by eliminating the linear constraint, $C'x' = 0$, using Golub's method [6].

Finally, Yu and Shi [10] consider the following optimization problem:

$$\begin{array}{ll} \text{minimize} & \dfrac{x^{\top}(D - W)x}{x^{\top} D x} \\ \text{subject to} & V^{\top} x = 0, \quad x \in \mathbb{R}^n \end{array}$$

where D is a diagonal matrix with positive entries and V has full rank, which is equivalent to

$$\begin{array}{ll} \text{maximize} & x^{\top}(W - D)x \\ \text{subject to} & x^{\top} D x = 1, \quad x \in \mathbb{R}^n \\ & V^{\top} x = 0. \end{array}$$

This problem is also solved by making the change of variable $x' = D^{1/2}x$ and by eliminating the linear constraint, $V^{\top}x = 0$, using Golub's method [6] involving a projector.

References

[1] James W. Burrows. Maximization of a second-degree polynomial on the unit sphere. *Mathematics of Computation*, 20(95):441–444, 1966.

[2] Timothée Cour and Jianbo Shi. Solving markov random fields with spectral relaxation. In Marita Meila and Xiaotong Shen, editors, *Artifical Intelligence and Statistics*. Society for Artificial Intelligence and Statistics, 2007.

[3] George Forsythe and Gene H. Golub. On the stationary values of a second-degree polynomial on the unit sphere. *SIAM Journal on Applied Mathematics*, 13(4):1050–1068, 1965.

[4] Walter Gander. Least squares with a quadratic constraint. *Numerical Mathematics*, 36:291–307, 1981.

[5] Walter Gander, Gene H. Golub, and Urs von Matt. A constrained eigenvalue problem. *Linear Algebra and its Applications*, 114/115:815–839, 1989.

[6] Gene H. Golub. Some modified eigenvalue problems. *SIAM Review*, 15(2):318–334, 1973.

[7] Jorge Moré and D.C. Sorensen. Computing a trust step. *SIAM Journal on Scientific and Statistical Computing*, 4(3):553–572, 1983.

[8] D.C. Sorensen. Newton's method with a model trust region modification. *SIAM Journal on Numerical Analysis*, 19(2):409–426, 1982.

[9] Emil Spjøtvoll. A note on a theorem of Forsythe and Golub. *SIAM Journal on Applied Mathematics*, 23(3):307–311, 1972.

[10] Stella X. Yu and Jianbo Shi. Grouping with bias. In Thomas G. Dietterich, Sue Becker, and Zoubin Ghahramani, editors, *Neural Information Processing Systems, Vancouver, Canada, 3-8 Dec. 2001*. MIT Press, 2001.