Extrema of sums of heterogeneous quadratic forms

Marianna Bolla

Department of Mathematics, Technical University, Budapest

and

György Michaletzky

Department of Probability Theory and Statistics, Eötvös Loránd University, Budapest

and

Gábor Tusnády

Mathematical Institute, Hungarian Academy of Sciences, Budapest

and

Margit Ziermann

Institute of Mathematics and Computer Sciences, University of Economics, Budapest

Submitted by

655 Avenue of the Americas, New York, NY 10010 0024- 3795/93/$6.00

Extrema of sums of heterogeneous quadratic forms *

ABSTRACT

In this paper we analyze the following problem arising in various situations in multivariate statistical analysis. We are given k symmetric, positive definite $n \times n$ matrices, $\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_k$ $(k \leq n)$ and we would like to maximize the function $\sum_{i=1}^{k} \mathbf{x}_i^T \mathbf{A}_i \mathbf{x}_i$ under the constraint that $\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_k \in \mathbb{R}^n$ form an orthonormal system. Some theoretical results as well as an algorithm will be presented.

1. Introduction

It is well known that given a symmetric, positive definite $n \times n$ matrix $\mathbf{A}$ an orthonormal system of k elements in $\mathbb{R}^n$ $(k \leq n)$ for which the functional $\sum_{i=1}^{k} \mathbf{x}_i^T \mathbf{A} \mathbf{x}_i$ attains its maximum is given by a system of k orthonormal eigenvectors corresponding to the k largest eigenvalues of the matrix $\mathbf{A}$. The subspace spanned by the system is uniquely determined if there is a gap in the spectrum of $\mathbf{A}$ between the k^{th} and $(k+1)^{th}$ eigenvalues in descending order, actually any orthonormal system consisting of k vectors spanning the same subspace as the eigenvectors corresponding to the k largest eigenvector gives the same value, because

$$\sum_{i=1}^{k} \mathbf{x}_i^T \mathbf{A} \mathbf{x}_i = tr \mathbf{A} \mathbf{X} \mathbf{X}^T \ ,$$

where

$$\mathbf{X} = [\mathbf{x}_1, \ldots \mathbf{x}_k] \ ,$$

thus the functional depends only on the subspace spanned by the vectors $\mathbf{x}_1, \ldots \mathbf{x}_k$. So the functional can be considered as being defined on the Grassmannian manifold $\mathcal{G}(k, \mathbb{R}^n)$ consisting of the k-dimensional subspaces of the Euclidean space $\mathbb{R}^n$. The

*This work was supported by the Hungarian Scientific Research Foundation, OTKA Grant No. 2042 and No. T015668.

structure of this functional is analyzed in details in Byrnes and Willems [BW86]. The behaviour of the matrix power method applied to this problem is investigated in Martin and Ammar [AM86].

The question naturally arises: what can be said on the maximum if the sum of the quadratic forms is generated by different matrices. Naturally, each of the quadratic forms $\mathbf{x}_i^T \mathbf{A}_i \mathbf{x}_i$ tends to be "large" (close to the maximal eigenvalue of $\mathbf{A}_i$), but in most cases it cannot be as large as $\lambda_i^{\max}$ because the eigenvectors corresponding to the maximal eigenvalues of $\mathbf{A}_i$-s are *usually* not pairwise orthogonal.

It will be shown that any system $\mathbf{x}_1, \dots \mathbf{x}_k$ giving the extremum must satisfy the matrix equation

$$(\mathbf{A}_1\mathbf{x}_1, \dots, \mathbf{A}_k\mathbf{x}_k) = \mathbf{X}\mathbf{S} \tag{1.1}$$

where $\mathbf{S}$ is a $k \times k$ symmetrical matrix , and the $n \times k$ matrices $(\mathbf{A}_1\mathbf{x}_1, \dots, \mathbf{A}_k\mathbf{x}_k)$ and $\mathbf{X} = (\mathbf{x}_1 \dots \mathbf{x}_k)$ contain the enumerated vectors as their columns. The set of orthonormal k-tuples in $\mathbb{R}^n$ is called as Stiefel-manifold and denoted by $V_{n,k}$. Slightly abusing the notation we shall write $\mathbf{X} \in V_{n,k}$ when the set of column vectors of the $n \times k$ matrix $\mathbf{X}$ is an element of $V_{n,k}$. (cf. James [Ja76]). Obviously this is equivalent to $\mathbf{X}^T\mathbf{X} = \mathbf{I}_k$. The equation (1.1) is linear in $\mathbf{X}$ so for the corresponding matrix $\mathbf{S}$ the determinant of the $nk \times nk$ matrix $\mathbf{A} - \mathbf{I}_n \otimes \mathbf{S}$ must be zero, where the $nk \times nk$ block-matrix $\mathbf{A}$ contains the matrices $\mathbf{A}_1, \dots, \mathbf{A}_k$ in its diagonal blocks and zeros otherwise.

In this paper, an iteration is proposed and its convergence to a local maximum of the objective function is analyzed. Choosing an arbitrary initial orthonormal system $\mathbf{X}^{(0)}$ the sequence $\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, ...$ is constructed in the following way: if $\mathbf{X}^{(m)}$ is already known, the polar decomposition of $(\mathbf{A}_1\mathbf{x}_1^{(m)} \dots \mathbf{A}_k\mathbf{x}_k^{(m)})$ is performed, i.e. it is decomposed as the product of an $n \times k$ matrix with orthonormal columns and a $k \times k$ symmetrical positive semidefinite one ($m = 0, 1, 2, \dots$). (This polar decomposition is unique if $\mathbf{A}_1\mathbf{x}_1^{(m)} \dots \mathbf{A}_k\mathbf{x}_k^{(m)}$ are linearly independent. Cf. equation (3.7).) In the next step let $\mathbf{X}^{(m+1)}$ be the first factor in this decomposition, etc.

2. The optimization problem

We are given k symmetrical, positive definite $n \times n$ matrices, $\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_k$ $(k \leq n)$. Find the maximum of

$$f_{\text{quad}}(\mathbf{X}) = \sum_{i=1}^{k} \mathbf{x}_i^T \mathbf{A}_i \mathbf{x}_i \tag{2.1}$$

on the constraints

$$\mathbf{x}_i^T \mathbf{x}_j = \delta_{ij}, \quad (1 \leq i, j \leq k).$$

where $\mathbf{X} = (\mathbf{x}_1 \ldots \mathbf{x}_k)$, and δ_{ij} is the Kronecker's delta. As $V_{n,k}$ is a compact manifold and f_{quad} is continuous on $V_{n,k}$ a finite global maximum exists and it is attained at some point. Obviously this maximum is at most $\sum_{i=1}^{k} \lambda_{\max}^i$, where $\lambda_{\max}^i$ denotes the maximal eigenvalue of $\mathbf{A}_i$.

To characterize the critical points of the functional let us denote by $\mathbf{A}(\mathbf{X}) = (\mathbf{A}_1\mathbf{x}_1, \ldots, \mathbf{A}_k\mathbf{x}_k)$ and $\mathbf{X} = (\mathbf{x}_1 \ldots \mathbf{x}_k)$ the $n \times k$ matrices containing the enumerated vectors as their columns. It will be shown below that for an optimal orthonormal system $(\mathbf{x}_1, \ldots, \mathbf{x}_k) \in V_{n,k}$

$$(\mathbf{A}_1\mathbf{x}_1, \ldots, \mathbf{A}_k\mathbf{x}_k) = \mathbf{XS} \tag{2.2}$$

holds, where the multipliers are entries of the $k \times k$ symmetrical matrix $\mathbf{S}$. Together with the system of equations

$$\mathbf{X}^T\mathbf{X} = \mathbf{I}_k \tag{2.3}$$

where $\mathbf{I}_k$ is the $k \times k$ identity matrix, we obtain $nk + k(k+1)/2$ equations. As the total number of unknowns in $\mathbf{X}$ and $\mathbf{S}$ is also $nk + k(k+1)/2$, in the generic case only a finite number of solutions is expected.

Before analyzing the structural behaviour of this functional let us consider an algorithm to maximize it.

3. The algorithm

In order to construct an algorithm let us return to the equations determining the critical point of the functional.

THEOREM 3.1. $\mathbf{X} \in V_{n,k}$ *is a critical point of* f_{quad} *if and only if* $\mathbf{S} = \mathbf{A}(\mathbf{X})^T\mathbf{X}$ *is symmetric, i.e.*

$$\mathbf{A}(\mathbf{X}) = \mathbf{X}\mathbf{S} \tag{3.1}$$

holds, with a symmetric $\mathbf{S}$.

Proof: Assume that $\mathbf{X} \in V_{n,k}$ is a critical point of f_{quad}. Thus the derivatives – along $V_{n,k}$ – of f_{quad} vanish at $\mathbf{X}$. Hence considering a small perturbation $\mathbf{X}_\Delta$ of $\mathbf{X}$ which is tangential to $V_{n,k}$, i.e. it satisfies the equation

$$\mathbf{X}^T\mathbf{X}_\Delta + \mathbf{X}_\Delta^T\mathbf{X} = 0 \ ,$$

in other words $\mathbf{X}^T\mathbf{X}_\Delta$ is skew-symmetric, the difference

$$f_{\text{quad}}(\mathbf{X} + \mathbf{X}_\Delta) - f_{\text{quad}}(\mathbf{X})$$

must be zero in the first order as $\mathbf{X}_\Delta \to 0$. But

$$\begin{aligned} f_{\text{quad}}(\mathbf{X} + \mathbf{X}_\Delta) - f_{\text{quad}}(\mathbf{X}) \\ = \operatorname{tr}(\mathbf{X} + \mathbf{X}_\Delta)^T\mathbf{A}(\mathbf{X} + \mathbf{X}_\Delta) - \operatorname{tr}\mathbf{X}^T\mathbf{A}(\mathbf{X}) \\ = 2\operatorname{tr}\mathbf{A}(\mathbf{X})^T\mathbf{X}_\Delta + \operatorname{tr}\mathbf{A}(\mathbf{X}_\Delta)^T\mathbf{X}_\Delta \ . \end{aligned}$$

Consequently the equation

$$\operatorname{tr}\mathbf{A}(\mathbf{X})^T\mathbf{X}_\Delta = 0 \tag{3.2}$$

when $\mathbf{X}^T\mathbf{X}_\Delta$ is skew symmetric characterizes the critical points of f_{quad}. We have

$$\begin{aligned} \operatorname{tr}\mathbf{A}(\mathbf{X})^T\mathbf{X}_\Delta = \frac{1}{2}\operatorname{tr}[\mathbf{A}(\mathbf{X})^T\mathbf{X} - \mathbf{X}^T\mathbf{A}(\mathbf{X})]\mathbf{X}^T\mathbf{X}_\Delta \\ + \operatorname{tr}[(\mathbf{I} - \mathbf{X}\mathbf{X}^T)\mathbf{A}(\mathbf{X})]^T[(\mathbf{I} - \mathbf{X}\mathbf{X}^T)\mathbf{X}_\Delta] \ . \end{aligned}$$

Observe that $\mathbf{A}(\mathbf{X})^T\mathbf{X} - \mathbf{X}^T\mathbf{A}(\mathbf{X})$ is skew-symmetric, and

$$\mathbf{X}^T((\mathbf{I} - \mathbf{X}\mathbf{X}^T)\mathbf{A}(\mathbf{X})) = 0 \ .$$

On the other hand for any skew-symmetric $k \times k$ matrix $\mathbf{Z}$ and for any $n \times k$ matrix $\mathbf{V}$ satisfying the identity $\mathbf{X}^T\mathbf{V} = 0$ there exists obviously a perturbation $\mathbf{X}_\Delta$ for which

$$\mathbf{X}^T\mathbf{X}_\Delta = \mathbf{Z} \ ,$$

and

$$(\mathbf{I} - \mathbf{X}\mathbf{X}^T)\mathbf{X}_\Delta = \mathbf{V} \ .$$

Consequently (3.2) implies that the set of equations

$$\mathbf{A}(\mathbf{X})^T\mathbf{X} = \mathbf{X}^T\mathbf{A}(\mathbf{X}) \tag{3.3}$$

$$\mathbf{A}(\mathbf{X}) = \mathbf{X}\mathbf{X}^T\mathbf{A}(\mathbf{X}) \tag{3.4}$$

characterizes the critical points of f_{quad}, concluding the proof. ■

At an arbitrary critical point the matrix $\mathbf{S} = \mathbf{X}^T\mathbf{A}(\mathbf{X})$ is not necessarily positive semidefinite. The next lemma shows that this is necessary at the global maximum points of f_{quad}.

LEMMA 3.1. *If $\mathbf{X} \in V_{n,k}$ is a global maximum of the functional f_{quad} then the corresponding matrix*

$$\mathbf{S} = \mathbf{X}^T\mathbf{A}(\mathbf{X})$$

is positive semidefinite.

Proof: Consider the decomposition

$$\mathbf{A}(\mathbf{X}) = \mathbf{X}\mathbf{S} \ ,$$

and assume in contrary with the statement that $\mathbf{S}$ is *not* positive semidefinite. Writing

$$\mathbf{S} = \mathbf{C}\Delta\mathbf{C}^T$$

where Δ is a diagonal, $\mathbf{C}$ is an orthogonal matrix, according to our assumption there is at least one negative element in the diagonal of Δ. Choosing an appropriate diagonal matrix $\mathbf{D}$ with values $+1$ or -1 in the diagonal we can achieve that $\mathbf{D}\Delta$ has only nonnegative elements in its diagonal. Obviously

$$\operatorname{tr}\mathbf{D}\Delta > \operatorname{tr}\Delta \ .$$

With the notation

$$\mathbf{Y} = \mathbf{X}\mathbf{C}\mathbf{D}\mathbf{C}^T$$

one can write $\mathbf{A}(\mathbf{X})$ in the form

$$\mathbf{A}(\mathbf{X}) = \mathbf{Y}(\mathbf{C}\mathbf{D}\Delta\mathbf{C}^T) \ .$$

Observe that the column vectors of $\mathbf{Y}$ are orthogonal and

$$\sum_{i=1}^{k} \mathbf{x}_i^T\mathbf{A}_i\mathbf{x}_i = \operatorname{tr}\mathbf{S} = \operatorname{tr}\Delta \ ,$$

$$\sum_{i=1}^{k} \mathbf{y}_i^T\mathbf{A}_i\mathbf{x}_i = \operatorname{tr}\mathbf{D}\Delta \ .$$

Consequently

$$\sum_{i=1}^{k} \mathbf{y}_i^T \mathbf{A}_i \mathbf{x}_i > \sum_{i=1}^{k} \mathbf{x}_i^T \mathbf{A}_i \mathbf{x}_i \ . \tag{3.5}$$

The inequality

$$\sum_{i=1}^{k} (\mathbf{y}_i - \mathbf{x}_i)^T \mathbf{A}_i (\mathbf{y}_i - \mathbf{x}_i) \geq 0 \tag{3.6}$$

together with (3.5) gives that

$$f_{quad}(\mathbf{Y}) > f_{\text{quad}}(\mathbf{X}) \ ,$$

so $\mathbf{X}$ is not a global maximum. ■

Theorem 1 and Lemma 1 together yield that at global maximum points

$$\mathbf{A}(\mathbf{X}) = \mathbf{X}\mathbf{S} \ ,$$

where $\mathbf{S}$ is a positive semidefinite matrix. Let us remark that a factorization of the form $\mathbf{XS}$, where $\mathbf{X}^T\mathbf{X} = \mathbf{I}_k$ and $\mathbf{S} \geq 0$ is called polar decomposition. (See [Fuhr81]. Observe that

$$\mathbf{A}(\mathbf{X}) = \left\{ \mathbf{A}(\mathbf{X}) \left[\left(\mathbf{A}(\mathbf{X})^T \mathbf{A}(\mathbf{X}) \right)^{\frac{1}{2}} \right]^{\#} \right\} \left[\left(\mathbf{A}(\mathbf{X})^T \mathbf{A}(\mathbf{X}) \right)^{\frac{1}{2}} \right] \tag{3.7}$$

determines a polar decomposition, where $^{\#}$ denotes the generalized inverse. Consequently, $\mathbf{S}$ is always unique as the positive semidefinite square root of $\mathbf{A}(\mathbf{X})^T\mathbf{A}(\mathbf{X})$, but the decomposition itself is unique only if $\mathbf{S} > 0$.)

Let us briefly analyze the connection between the singular value decomposition and the polar decomposition of the same matrix. More generally, we consider two types of decomposition of an $n \times k$ matrix $\mathbf{B}$:

$$\mathbf{B} = \mathbf{X}\mathbf{S} \ , \tag{3.8}$$

where $\mathbf{X}$ is an $n \times k$, $\mathbf{S}$ is a $k \times k$ matrix, $\mathbf{X}^T\mathbf{X} = \mathbf{I}_k$, $\mathbf{S}$ is symmetric and

$$\mathbf{B} = \mathbf{P}\mathbf{V}\mathbf{Q}^T \ , \tag{3.9}$$

where $\mathbf{P}$ is an $n \times k$, $\mathbf{Q}, \mathbf{V}$ are $k \times k$ matrices, $\mathbf{P}^T\mathbf{P} = \mathbf{Q}^T\mathbf{Q} = \mathbf{I}_k$ and $\mathbf{V}$ is a diagonal matrix.

In the polar decomposition $\mathbf{S} \geq 0$, in the singular value decomposition $\mathbf{V} \geq 0$.

PROPOSITION 3.1. *Let* $\mathbf{B}$ *be an arbitrary* $n \times k$ *matrix. The formulae*

$$\mathbf{V}, \mathbf{Q}, \mathbf{P} \to \mathbf{S} = \mathbf{Q}\mathbf{V}\mathbf{Q}^T \ , \ \mathbf{X} = \mathbf{P}\mathbf{Q}^T$$

and

$$\mathbf{S}, \mathbf{X} \to \mathbf{S}\mathbf{Q} = \mathbf{Q}\mathbf{V} \ , \ \mathbf{P} = \mathbf{X}\mathbf{Q}$$

provide a correspondence between the decompositions of $\mathbf{B}$ *of type (3.8) and (3.9).*

Proof: If $\mathbf{B} = \mathbf{P}\mathbf{V}\mathbf{Q}^T$ is of type (3.9) then

$$\mathbf{B} = \mathbf{P}\mathbf{Q}^T \ \mathbf{Q}\mathbf{V}\mathbf{Q}^T$$

and $\mathbf{X} = \mathbf{P}\mathbf{Q}^T$ satisfies the equation $\mathbf{X}^T\mathbf{X} = \mathbf{I}$, also $\mathbf{S} = \mathbf{Q}\mathbf{V}\mathbf{Q}^T \geq 0$ so $\mathbf{B} = \mathbf{X}\mathbf{S}$ is a decomposition of type (3.8).

Conversely, if $\mathbf{B} = \mathbf{X}\mathbf{S}$ is a decomposition of type (3.8) then considering the principal axis transformation of $\mathbf{S}$ leading to the equation

$$\mathbf{S} = \mathbf{Q}\mathbf{V}\mathbf{Q}^T$$

and defining $\mathbf{P} = \mathbf{X}\mathbf{Q}$ we get that

$$\mathbf{B} = \mathbf{P}\mathbf{V}\mathbf{Q}^T$$

is a decomposition of type (3.9). ■

Remark. Obviously $\mathbf{V} \geq 0$ if and only if $\mathbf{S} \geq 0$. Also if $\mathbf{B} = \mathbf{X}\mathbf{S}$ is a decomposition of type (3.8), $\mathbf{S} = \mathbf{Q}\mathbf{V}\mathbf{Q}^T$, then

$$\mathbf{B}\mathbf{Q} = (\mathbf{X}\mathbf{Q})\mathbf{V} \ ,$$
$$\mathbf{B}^T(\mathbf{X}\mathbf{Q}) = \mathbf{Q}\mathbf{V} \ .$$

Observe that the columns of $\mathbf{Q}$ and that of $\mathbf{X}\mathbf{Q}$ are orthonormal, so the absolute value of the diagonal elements of $\mathbf{V}$, i.e. the absolute value of the eigenvalues of $\mathbf{S}$ are the singular values of $\mathbf{B}$. Consequently, if $\sigma_1(\mathbf{B}) \geq \cdots \geq \sigma_k(\mathbf{B})$ denote the singular values of $\mathbf{B}$ then

$$\sum_{i=1}^{k} \sigma_i(\mathbf{B}) \geq \operatorname{tr} \mathbf{S}$$

and we have equality if and only if $\mathbf{S} \geq 0$.

The considerations before this propositions suggest the algorithm outlined in the Introduction. From an arbitrary initial set of orthonormal k-tuples we define recursively a sequence in $V_{n,k}$ as follows: from the m^{th} element of this sequence $\mathbf{X}^{(m)} \in V_{n,k}$

the next one is obtained by a polar decomposition of the matrix $\mathbf{A}(\mathbf{X}^{(m)})$ as

$$\mathbf{A}(\mathbf{X}^{(m)}) = \mathbf{X}^{(m+1)}\mathbf{S}^{(m+1)} \ . \tag{3.10}$$

Let us consider one single step in this iteration. To ease the notation let us denote the corresponding elements in $V_{n,k}$ by $\mathbf{X}$ and $\mathbf{Y}$, i.e.

$$\mathbf{A}(\mathbf{X}) = \mathbf{YS} \tag{3.11}$$

Let us recall a theorem which was proved by Bolla [Bo82] or in a slightly more general form by Brockett [Br89].

THEOREM 3.2. *Assume that $\mathbf{X} \in V_{n,k}$ is a fixed orthonormal k-tuple. Then the solution of the minimization problem*

$$\sum_{i=1}^{k} \|\mathbf{A}_i\mathbf{x}_i - \mathbf{y}_i\|^2 \to min \ , \tag{3.12}$$

where $\mathbf{y}_1 \dots \mathbf{y}_k$ are orthonormal vectors is provided by the column vectors of $\mathbf{Y}$ in the polar decomposition of $\mathbf{A}(\mathbf{X})$:

$$\mathbf{A}(\mathbf{X}) = \mathbf{YS} \ .$$

Remark. In [Br89] the so-called *matching problem A* asks for the solution of the minimization problem

$$\sum_{i=1}^{k} \|\mathbf{z}_i - \phi(\mathbf{y}_i)\|^2 \to min \ ,$$

where $\mathbf{z}_i, \mathbf{y}_i, i = 1, \dots k$ are fixed vectors, and ϕ is an element of a Lie group acting on $\mathbb{R}^n$.

Remark. Since

$$\sum_{i=1}^{k} \|\mathbf{A}_i\mathbf{x}_i - \mathbf{y}_i\|^2 = \sum_{i=1}^{k} \|\mathbf{A}_i\mathbf{x}_i\|^2 + k - 2\sum_{i=1}^{k} \mathbf{y}_i^T\mathbf{A}_i\mathbf{x}_i$$

the problem (3.12) is equivalent to

$$f_{\text{bilin}}(\mathbf{X}, \mathbf{Y}) = \sum_{i=1}^{k} \mathbf{y}_i^T\mathbf{A}_i\mathbf{x}_i \to max \ , \tag{3.13}$$

where $\mathbf{x}_1, \dots, \mathbf{x}_k$ and $\mathbf{y}_1, \dots \mathbf{y}_k$ are orthonormal vectors.

In other words the algorithm described above is a partial optimization of the functional f_{bilin}. Similar argument to the

one which was used in the proof of Theorem 1 yields that the critical points of the functional

$$\mathbf{Y} \to f_{\text{bilin}}(\mathbf{X}, \mathbf{Y})$$

for fixed $\mathbf{X} \in V_{n,k}$ are characterized by the equations

$$\mathbf{Y}^T\mathbf{A}(\mathbf{X}) = \mathbf{A}(\mathbf{X})^T\mathbf{Y} \ , \tag{3.14}$$

$$\mathbf{A}(\mathbf{X}) = \mathbf{Y}\mathbf{Y}^T\mathbf{A}(\mathbf{X}) \ . \tag{3.15}$$

The inequality (3.6) can be now written as

$$f_{\text{quad}}(\mathbf{X}) + f_{\text{quad}}(\mathbf{Y}) \geq 2f_{\text{bilin}}(\mathbf{X}, \mathbf{Y}) \tag{3.16}$$

for any $\mathbf{X}, \mathbf{Y} \in V_{n,k}$, moreover we have equality here if and only if $\mathbf{X} = \mathbf{Y}$. We shall exploit this elementary observation several times. It might be instructive to write the inequality in form

$$f_{\text{bilin}}(\mathbf{X}, \mathbf{Y}) - f_{\text{quad}}(\mathbf{X}) \leq f_{\text{quad}}(\mathbf{Y}) - f_{\text{bilin}}(\mathbf{X}, \mathbf{Y}) \ , \tag{3.17}$$

and to read it in the following way: once the substitution of one $\mathbf{X}$ for $\mathbf{Y}$ in $f_{\text{bilin}}(\mathbf{X}, \mathbf{X})$ increases the value of the bilinear form then the use of the same substitution as a second step is also increasing.

4. Structural properties of the functionals

First we determine the Hessian form of f_{quad} and f_{bilin} at critical points.

PROPOSITION 4.1. *Consider a critical point $\mathbf{X} \in V_{n,k}$ of f_{quad}. Let ξ be the derivative at $\mathbf{X}$ of a curve lying on the surface $V_{n,k}$ and going through the critical point $\mathbf{X}$, i.e. let ξ be such that $\mathbf{X}^T\xi$ is a skew-symmetric matrix. Then the quadratic form determined by the Hessian of f_{quad} at $\mathbf{X}$ evaluated at ξ has the value*

$$tr\,\xi^T\mathbf{A}(\xi) - tr\,\xi\mathbf{X}^T\mathbf{A}(\mathbf{X})\xi^T \ . \tag{4.1}$$

Proof: Consider a curve $\mathbf{X}(t) \in V_{n,k}$ with continuous second derivative and assume that $\mathbf{X}(0) = \mathbf{X}$, $\mathbf{X}^{'}(0) = \xi$. Then

$$\frac{d^2}{dt^2} f_{\text{quad}}(\mathbf{X}(t)) = 2\frac{d}{dt}\sum_{i=1}^{k} \mathbf{x}_i(t)^T \mathbf{A}_i \mathbf{x}_i^{'}(t)$$
$$= 2\sum_{i=1}^{k} \mathbf{x}_i(t)^T \mathbf{A}_i \mathbf{x}_i^{''}(t) + 2\sum_{i=1}^{k} \mathbf{x}_i^{'}(t)^T \mathbf{A}_i \mathbf{x}_i^{'}(t)$$
$$= 2\text{tr}\, \mathbf{A}(\mathbf{X}(t))^T \mathbf{X}^{''}(t) + \text{tr}\, \mathbf{A}(\mathbf{X}^{'}(t))^T \mathbf{X}^{'}(t)$$

On the other hand taking the second derivative of the identity

$$\mathbf{X}^T(t)\mathbf{X}(t) = \mathbf{I}$$

we get that

$$\mathbf{X}^T(t)\mathbf{X}^{''}(t) + 2\mathbf{X}^{'}(t)^T\mathbf{X}^{'}(t) + \mathbf{X}^{''}(t)^T\mathbf{X}(t) = 0 \ ,$$

i.e.

$$\mathbf{X}^T(t)\mathbf{X}^{''}(t) + \mathbf{X}^{'}(t)\mathbf{X}^{'}(t) \text{ is a skew-symmetric matrix.}$$

Evaluating these derivatives at $t = 0$, using the identity

$$\mathbf{A}(\mathbf{X}) = \mathbf{X}\mathbf{X}^T\mathbf{A}(\mathbf{X})$$

we obtain that

$$\frac{d^2}{dt^2} f_{\text{quad}}(\mathbf{X}(t))_{|t=0} = 2\text{tr}\, \mathbf{A}(\mathbf{X})^T\mathbf{X}\mathbf{X}^T\mathbf{X}^{''}(0) + 2\text{tr}\, \mathbf{A}(\xi)^T\xi \ .$$

But $\mathbf{A}(\mathbf{X})^T\mathbf{X}$ is a symmetric matrix consequently

$$\text{tr}\, \mathbf{A}(\mathbf{X})^T\mathbf{X}(\mathbf{X}^T\mathbf{X}^{''}(0) + \xi^T\xi) = 0$$

giving that the Hessian at ξ is (4.1) concluding the proof. ■

PROPOSITION 4.2. *Assume that the pair $\mathbf{X}, \mathbf{Y} \in V_{n,k}$ is a critical point of f_{bilin}. Let ξ, η be the derivatives (at $\mathbf{X}, \mathbf{Y}$) of such curves in $V_{n,k}$ which go through $\mathbf{X}$ and $\mathbf{Y}$, respectively. Then the Hessian of f_{bilin} at $\mathbf{X}, \mathbf{Y}$ evaluated at ξ, η is given by*

$$tr\xi^T\mathbf{A}(\eta) - \frac{1}{2}\left[tr\xi\mathbf{X}^T\mathbf{A}(\mathbf{Y})\xi^T + tr\eta\mathbf{Y}^T\mathbf{A}(\mathbf{X})\eta^T\right] \ . \tag{4.2}$$

Proof: Since the proof is similar to that one of the previous proposition we only outline it. Let $\mathbf{X}(t), \mathbf{Y}(t) \in V_{n,k}$ be two curves with continuous second derivative for which

$$\mathbf{X}(0) = \mathbf{X} \ , \ \mathbf{Y}(0) = \mathbf{Y} \ , \ \mathbf{X}^{'}(0) = \xi \ , \ \mathbf{Y}^{'}(0) = \eta \ .$$

Then

$$\frac{d^2}{dt^2} f_{\text{bilin}}(\mathbf{X}(t), \mathbf{Y}(t))_{|t=0} =$$
$$= \operatorname{tr} \mathbf{A}(\mathbf{X})^T \mathbf{Y}^{''}(0) + \operatorname{tr} \mathbf{A}(\mathbf{Y})^T \mathbf{X}^{''}(0) + 2\operatorname{tr} \xi^T \mathbf{A}(\eta) \ .$$

Using that

$$\mathbf{A}(\mathbf{X}) = \mathbf{Y}\mathbf{Y}^T \mathbf{A}(\mathbf{X}) \ , \ \mathbf{A}(\mathbf{Y}) = \mathbf{X}\mathbf{X}^T \mathbf{A}(\mathbf{Y})$$

and

$$\mathbf{Y}^T \mathbf{A}(\mathbf{X}) \ , \ \mathbf{X}^T \mathbf{A}(\mathbf{Y}) \text{ are symmetric ,}$$
$$\mathbf{X}^T \mathbf{X}^{''}(0) + \xi^T \xi \ , \ \mathbf{Y}^T \mathbf{Y}^{''}(0) + \eta^T \eta \text{ are skew-symmetric}$$

matrices we obtain that the value of the Hessian at ξ, η is (4.2). ■

If $\mathbf{X}$ is a local maximum of f_{quad} then the Hessian is negative semidefinite, i.e.

$$\operatorname{tr} \xi \mathbf{A}(\mathbf{X})^T \mathbf{X} \xi^T \geq \operatorname{tr} \xi^T \mathbf{A}(\xi) \tag{4.3}$$

Although the inequality (4.3) reflects the local properties of f_{quad}, it can be proved that

$$f_{\text{quad}}(\mathbf{X}) \geq f_{\text{quad}}(\mathbf{Y})$$

even in the case when $\mathbf{Y}$ is not necessarily in the neighbourhood of $\mathbf{X}$.

Definition 4.1. Consider an orthonormal set of vectors $(\mathbf{x}_1, \ldots, \mathbf{x}_k)$ in $\mathbb{R}^n$. We say that $\mathbf{y}_1, \ldots, \mathbf{y}_k$ is an elementary transform of $(\mathbf{x}_1, \ldots, \mathbf{x}_k)$ if one of the following two properties holds

(i) there exist $1 \leq i < j \leq k$ such that

$$\mathbf{y}_i = \mathbf{x}_j \ , \ \mathbf{y}_j = \mathbf{x}_i \ , \ \mathbf{y}_l = \mathbf{x}_l \ , \text{if } l \neq i, j \ ,$$

(ii) there exists $1 \leq i \leq k$ such that

$$\mathbf{y}_i \text{ is a unit vector, orthogonal to} (\mathbf{x}_1, \ldots, \mathbf{x}_k) \ , \text{and}$$
$$\mathbf{y}_l = \mathbf{x}_l \ , \text{ if } l \neq i \ .$$

COROLLARY 4.1. *Assume that* $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_k)$ *defines a local maximum of* f_{quad}. *Then*

$$f_{quad}(\mathbf{X}) \geq f_{quad}(\mathbf{Y})$$

where $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_k]$ *is an elementary transform of* $[\mathbf{x}_1, \ldots, \mathbf{x}_k]$.

Proof: First consider the case when $\mathbf{y}_l = \mathbf{x}_l$ except for $l = i$ and $\mathbf{y}_i$ is orthogonal to $(\mathbf{x}_1, \ldots, \mathbf{x}_k)$. Considering the matrix ξ having zero columns except the i^{th} one which is $\mathbf{y}_i$. We can apply (4.3) because $\xi^T \mathbf{X} = 0$ which leads to the inequality

$$\operatorname{tr} \xi^T \xi \mathbf{A}(\mathbf{X})^T \mathbf{X} \geq \mathbf{y}_i^T \mathbf{A}_i \mathbf{y}_i \ .$$

But the left hand side is exactly $\mathbf{x}_i^T \mathbf{A}_i \mathbf{x}_i$. So the change $\mathbf{x}_i \to \mathbf{y}_i$ does not increase the value of f_{quad}.

Now consider the case when $\mathbf{y}_1, \ldots, \mathbf{y}_k$ is defined by interchanging the vectors $\mathbf{x}_i$ and $\mathbf{x}_j$. In this case let the i^{th} column of ξ be $\mathbf{x}_j$, the j^{th} one be $-\mathbf{x}_i$ the others be zero. Then the (i, j)-th element of $\xi^T \mathbf{X}$ is 1, the (j, i)-th element is -1, the others are zero, so it is a skew-symmetric matrix, thus inequality (4.3) should hold. But now

$$\operatorname{tr} \xi^T \mathbf{A}(\xi) = \mathbf{x}_j^T \mathbf{A}_i \mathbf{x}_j + \mathbf{x}_i^T \mathbf{A}_j \mathbf{x}_i$$

and

$$\operatorname{tr} \xi^T \xi \mathbf{A}(\mathbf{X})^T \mathbf{X} = \mathbf{x}_i^T \mathbf{A}_i \mathbf{x}_i + \mathbf{x}_j^T \mathbf{A}_j \mathbf{x}_j \ .$$

Consequently, the change $\mathbf{x}_i \to \mathbf{x}_j$ does not increase the functional f_{quad}. ■

COROLLARY 4.2. *Assume that* f_{quad} *has a local maximum at* $\mathbf{X} \in V_{n,k}$. *Then*

(i) if $k < n$ *then the matrix* $\mathbf{A}(\mathbf{X})^T \mathbf{X}$ *is positive definite, especially the vectors* $(\mathbf{A}_1 \mathbf{x}_1, \ldots, \mathbf{A}_k \mathbf{x}_k)$ *are linearly independent,*

(ii) if $k = n$ *then the matrix* $\mathbf{A}(\mathbf{X})^T \mathbf{X}$ *can have only one negative eigenvalue and its trace on the two dimensional subspaces is positive.*

Proof: Consider first the case when $k < n$. Then there exists a unit vector $\mathbf{z} \in \mathbb{R}^n$ which is orthogonal to $(\mathbf{x}_1, \ldots, \mathbf{x}_k)$. Let $\mathbf{a} = [a_1, \ldots, a_k]^T \in \mathbb{R}^k$ be an arbitrary nonzero vector. Define the matrix ξ as follows:

$$\xi = [a_1\mathbf{z}, \ldots, a_k\mathbf{z}] = \mathbf{z}\mathbf{a}^T \ .$$

Obviously $\mathbf{X}^T\xi = 0$, so we can apply (4.3). But

$$\operatorname{tr} \xi \mathbf{A}(\mathbf{X})^T \mathbf{X} \xi^T = \mathbf{a}^T \mathbf{A}(\mathbf{X})^T \mathbf{X}\mathbf{a} \ ,$$

and

$$\operatorname{tr} \xi \mathbf{A}(\xi) = \sum_{k=1}^{k} a_i^2 \mathbf{z}^T \mathbf{A}_i \mathbf{z} > 0 \ .$$

This implies that

$$\mathbf{a}^T \mathbf{A}(\mathbf{X})^T \mathbf{X}\mathbf{a} > 0$$

if $\mathbf{a} \neq 0$, i.e. $\mathbf{A}(\mathbf{X})^T\mathbf{X}$ is positive definite.

If $k = n$ then the previous method cannot be applied directly. Since in this case $\mathbf{X}\mathbf{X}^T = \mathbf{X}^T\mathbf{X} = \mathbf{I}_n$, consequently

$$\operatorname{tr} \xi \mathbf{A}(\mathbf{X})^T \mathbf{X} \xi^T = \operatorname{tr} \xi^T \mathbf{X}\mathbf{X}^T \xi \mathbf{A}(\mathbf{X})^T \mathbf{X} \ .$$

Let us recall [BK84] that any real skew symmetric matrix $\mathbf{B}$ is similar (under real orthogonal transformation) to a block diagonal matrix, where the block are of order one or two. The blocks are skew symmetric matrices, so that of order one are zero matrices. This implies that every nonzero eigenvalue of the negative semidefinite symmetric matrix $\mathbf{B}^2$ is of even multiplicity.

Conversely, if a symmetric matrix $\mathbf{C}$ has the representation

$$\mathbf{C} = -\sum_{j=1}^{l} \lambda_j^2 \mathbf{D}_j \mathbf{D}_j^T \ , \tag{4.4}$$

where

$$\mathbf{D}_j = [\mathbf{x}_j, \mathbf{y}_j] \ \ , j = 1, \ldots, l$$

and the vectors $\mathbf{x}_j, \mathbf{y}_j, j = 1, \ldots l$ are orthogonal unit vectors, then the matrix

$$\mathbf{B} = 2\sum_{j=1}^{l} \lambda_j \left(\mathbf{x}_j \mathbf{y}_j^T - \mathbf{y}_j \mathbf{x}_j^T \right)$$

is skew-symmetric and

$$\mathbf{C} = \mathbf{B}^2 \ .$$

Returning to the inequality (4.3), $\operatorname{tr} \xi^T \mathbf{A}(\xi) > 0$, if $\xi \neq 0$, thus

$$\operatorname{tr} \xi^T \mathbf{X}\mathbf{X}^T \xi \mathbf{A}(\mathbf{X})^T \mathbf{X} > 0 \text{ if } \xi \neq 0 \; . \tag{4.5}$$

Since the left-hand side is linear in $\xi^T \mathbf{X}\mathbf{X}^T \xi$, which can be written in the form of (4.4) it is enough to check it for each summand in (4.5) Consequently, (4.5) is equivalent to

$$\operatorname{tr} \; [\mathbf{x}, \mathbf{y}] \, [\mathbf{x}, \mathbf{y}]^T \, \mathbf{A}(\mathbf{X})^T \mathbf{X} > 0$$

for any pair of orthonormal unit vectors $\mathbf{x}, \mathbf{y}$. Thus

$$\mathbf{x}^T \mathbf{A}(\mathbf{X})^T \mathbf{X}\mathbf{x} + \mathbf{y}^T \mathbf{A}(\mathbf{X})^T \mathbf{X}\mathbf{y} > 0$$

proving part (ii) of the corollary. ■

Remark. If $\lambda_1 \geq \ldots \lambda_n$ denote the eigenvalues of $\mathbf{A}(\mathbf{X})^T \mathbf{X}$, then we obtained that $\lambda_{n-1} > 0$ and if $\lambda_n < 0$ then $\lambda_{n-1} > |\lambda_n|$.

COROLLARY 4.3. *Assume that f_{bilin} has a local maximum at $(\mathbf{X}, \mathbf{Y})$, where $\mathbf{X}, \mathbf{Y} \in V_{n,k}$. Then*

(i) if $k < n$, then the matrices $\mathbf{X}^T \mathbf{A}(\mathbf{Y})$ and $\mathbf{Y}^T \mathbf{A}(\mathbf{X})$ are positive semidefinite,

(ii) if $k = n$, then the trace of the matrices $\mathbf{X}^T \mathbf{A}(\mathbf{Y})$ and $\mathbf{Y}^T \mathbf{A}(\mathbf{X})$ on any two dimensional subspace are nonnegative.

Proof: The Hessian at a local maximum must be negative semidefinite, especially – choosing $\eta = 0$ – , we obtain that

$$\operatorname{tr} \xi \mathbf{X}^T \mathbf{A}(\mathbf{Y}) \xi^T \geq 0 \; ,$$

if $\xi^T \mathbf{X}$ is skew-symmetric, and similarly – if $\xi = 0$ – ,

$$\operatorname{tr} \eta \mathbf{Y}^T \mathbf{A}(\mathbf{X}) \eta^T \geq 0 \; ,$$

if $\eta^T \mathbf{Y}$ is skew-symmetric. Consequently, the proof of the previous Corollary can be repeated here. ■

THEOREM 4.1. *(i) If $\mathbf{X} \in V_{n,k}$ is a local maximum of f_{quad} then $(\mathbf{X}, \mathbf{X})$ is a local maximum of f_{bilin}.*

(ii) If $k < n$ and $(\mathbf{X}, \mathbf{Y})$ $(\mathbf{X}, \mathbf{Y} \in V_{n,k})$ is a local maximum of f_{bilin} then $\mathbf{X} = \mathbf{Y}$ and $\mathbf{X}$ is a local maximum of f_{quad}.

Proof: We start with (ii). If $(\mathbf{X}, \mathbf{Y})$ is a local maximum of f_{bilin}, then $\mathbf{Y}$ is a local maximum of the functional $\mathbf{Z} \to f_{\text{bilin}}(\mathbf{X}, \mathbf{Z})$, where $\mathbf{X}$ is fixed. Corollary 3 gives that $\mathbf{Y}^T\mathbf{A}(\mathbf{X}) \geq 0$, so according to Remark after Proposition 1 it is a global maximum. Consequently

$$\operatorname{tr} \mathbf{Z}^T\mathbf{A}(\mathbf{X}) \leq \operatorname{tr} \mathbf{Y}^T\mathbf{A}(\mathbf{X})$$

for every $n \times k$ matrix $\mathbf{Z}$ with orthonormed column vectors. In particular

$$\operatorname{tr} \mathbf{X}^T\mathbf{A}(\mathbf{X}) \leq \operatorname{tr} \mathbf{Y}^T\mathbf{A}(\mathbf{X}) \ .$$

Similarly,

$$\operatorname{tr} \mathbf{Y}^T\mathbf{A}(\mathbf{Y}) \leq \operatorname{tr} \mathbf{Y}^T\mathbf{A}(\mathbf{X}) \ .$$

Comparing this to (3.17) we obtain that $\mathbf{X} = \mathbf{Y}$.

(i) Let $\mathbf{X} \in V_{n,k}$ be a local maximum of f_{quad}. If $\mathbf{Y}$ and $\mathbf{Z}$ are in small neighbourhood of $\mathbf{X}$ then

$$f_{\text{quad}}(\mathbf{X}) \geq \max(f_{\text{quad}}(\mathbf{Y}), f_{\text{quad}}(\mathbf{Z})) \ .$$

Inequality (3.16) gives that

$$f_{\text{quad}}(\mathbf{X}) \geq f_{\text{bilin}}(\mathbf{Z}, \mathbf{Y})$$

proving part (i). ■

Now we study the set of critical point of f_{quad}. Denote

$$\mathcal{C} = \{\mathbf{X} \in V_{n,k} \mid \mathbf{X} \text{ is a critical point of } f_{\text{quad}}\} \ .$$

We have proved that the equations

$$\begin{aligned} \mathbf{A}(\mathbf{X})^T\mathbf{X} &= \mathbf{X}^T\mathbf{A}(\mathbf{X}) \\ \mathbf{A}(\mathbf{X}) &= \mathbf{X}\mathbf{X}^T\mathbf{A}(\mathbf{X}) \end{aligned}$$

characterize the critical points. Since these are polynomials in the elements of $\mathbf{X}$ the set $\mathcal{C}$ can be written as

$$\mathcal{C} = \cup_{j=1}^{N} \mathcal{C}_j \ ,$$

where $\mathcal{C}_j \subset V_{n,k}$, $j = 1, \ldots, N$ are connected submanifolds.

PROPOSITION 4.3. *Let $\mathbf{X} \in \mathcal{C}$ be a critical point. Denote by $\mathcal{H}_X$ the Hessian of f_{quad} at $\mathbf{X}$. Then the $n \times k$ matrix ξ is an eigenvector of $\mathcal{H}_X$ if and only if there exists a $\lambda \in \mathbb{R}$ such that*

$$\mathbf{A}(\xi) - \xi\mathbf{A}(\mathbf{X})^T\mathbf{X} - \mathbf{X}\mathbf{A}(\xi)^T\mathbf{X} - \mathbf{X}\mathbf{A}(\mathbf{X})^T\xi = \lambda(\mathbf{I} + \mathbf{X}\mathbf{X}^T)\xi \tag{4.6}$$

and $\mathbf{X}^T\xi$ is skew-symmetric.

Proof: According to Proposition 1 the quadratic form determined by $\mathcal{H}_X$ at ξ is given by (4.1). Consequently the bilinear form determined by this quadratic form is

$$\begin{aligned}\mathcal{H}_X(\xi,\eta) &= \operatorname{tr}\eta^T\mathbf{A}(\xi) - \operatorname{tr}\eta\mathbf{X}^T\mathbf{A}(\mathbf{X})\xi^T \\ &= \operatorname{tr}\eta^T(\mathbf{I}-\mathbf{X}\mathbf{X}^T)\mathbf{A}(\xi) + \operatorname{tr}\eta^T\mathbf{X}\mathbf{X}^T\mathbf{A}(\xi) \\ &\quad - \operatorname{tr}(\mathbf{I}-\mathbf{X}\mathbf{X}^T)\eta\mathbf{X}^T\mathbf{A}(\mathbf{X})\xi^T - \operatorname{tr}\mathbf{X}\mathbf{X}^T\eta\mathbf{X}^T\mathbf{A}(\mathbf{X})\xi^T\ .\end{aligned}$$

Using that $\eta^T\mathbf{X}$ and $\xi^T\mathbf{X}$ are skew-symmetric and

$$\mathbf{X}^T\mathbf{A}(\mathbf{X})\mathbf{X}^T = \mathbf{A}(\mathbf{X})^T$$

we obtain that

$$\begin{aligned}\mathcal{H}_X(\xi,\eta) &= \operatorname{tr}\eta^T(\mathbf{I}-\mathbf{X}\mathbf{X}^T)\left[\mathbf{A}(\xi) - \xi\mathbf{A}(\mathbf{X})^T\mathbf{X}\right] \\ &+ \frac{1}{2}\operatorname{tr}\eta^T\mathbf{X}(\mathbf{X}^T\mathbf{A}(\xi) - \mathbf{A}(\xi)^T\mathbf{X} \\ &\quad -\mathbf{X}^T\xi\mathbf{A}(\mathbf{X})^T\mathbf{X} - \mathbf{X}^T\mathbf{A}(\mathbf{X})\mathbf{X}^T\xi)\end{aligned}$$

If ξ is an eigenvector with the eigenvalue λ then

$$\mathcal{H}_X(\xi,\eta) = \lambda\operatorname{tr}\eta^T\xi = \lambda\operatorname{tr}\eta^T\mathbf{X}\mathbf{X}^T\xi + \lambda\operatorname{tr}\eta^T(\mathbf{I}-\mathbf{X}\mathbf{X}^T)\xi\ .$$

This implies that ξ is characterized by the equations

$$\begin{aligned}&(\mathbf{I}-\mathbf{X}\mathbf{X}^T)(\mathbf{A}(\xi) - \xi\mathbf{A}(\mathbf{X})^T\mathbf{X}) = \lambda(\mathbf{I}-\mathbf{X}\mathbf{X}^T)\xi\ , \\ &\mathbf{X}^T(\mathbf{A}(\xi) - \mathbf{X}\mathbf{A}(\xi)^T\mathbf{X} - \xi\mathbf{A}(\mathbf{X})^T\mathbf{X} - \mathbf{X}\mathbf{A}(\mathbf{X})^T\xi) = 2\lambda\mathbf{X}^T\xi\ .\end{aligned}$$

Multiplying the second equation by $\mathbf{X}$ and adding them together we obtain that

$$\mathbf{A}(\xi) - \xi\mathbf{A}(\mathbf{X})^T\mathbf{X} - \mathbf{X}\mathbf{A}(\xi)^T\mathbf{X} - \mathbf{X}\mathbf{A}(\mathbf{X})^T\xi = \lambda(\mathbf{I}+\mathbf{X}\mathbf{X}^T)\xi$$

which is obviously equivalent to the pair of equation above. ■.

PROPOSITION 4.4. *Let $\mathbf{X} \in \mathcal{C}_j$ be a critical point. Then ξ belongs to $Ker\mathcal{H}_\mathbf{X}$ if and only if ξ is a tangent vector of the manifold $\mathcal{C}_j$.*

Proof: Obviously any tangent vector of $\mathcal{C}_j$ at $\mathbf{X}$ lies in $Ker\mathcal{H}_\mathbf{X}$. Conversely, in view of Theorem 1 the tangent vectors of $\mathcal{C}_j$ at $\mathbf{X}$ are characterized by the equations

$$\begin{aligned}\xi^T\mathbf{X} + \mathbf{X}^T\xi &= 0\ , &(4.7)\\ \xi^T\mathbf{A}(\mathbf{X}) &= -\mathbf{X}^T\mathbf{A}(\xi) + \mathbf{A}(\xi)^T\mathbf{X} + \mathbf{A}(\mathbf{X})^T\xi\ , &(4.8)\\ \mathbf{A}(\xi) &= \xi\mathbf{A}(\mathbf{X})^T\mathbf{X} + \mathbf{X}\mathbf{A}(\xi)^T\mathbf{X} + \mathbf{X}\mathbf{A}(\mathbf{X})^T\xi\ . &(4.9)\end{aligned}$$

Now if $\xi \in Ker\mathcal{H}_{\mathbf{X}}$ then $\xi^T\mathbf{X}$ is skew-symmetric and (4.8) holds. Multiplying (4.8) by $\mathbf{X}^T$ and using that

$$\mathbf{X}^T\xi = -\xi^T\mathbf{X} \ , \ \mathbf{X}\mathbf{A}(\mathbf{X})^T\mathbf{X} = \mathbf{A}(\mathbf{X}),$$

we get that (4.9) holds, as well. I.e. ξ is a tangent vector of $\mathcal{C}_j$. ■

Remark: Without going into the details we remark that this Proposition implies that over $\mathbb{Z}_2$ the function f_{quad} is a so-called Morse-Bott function, so the Morse-Bott inequalities hold connecting the Betti-numbers of the Stiefel-manifold $V_{n,k}$ with the indices of the functional f_{quad}. (cf. Byrnes - Willems [BW86].)

5. The critical points of the quadratic functional in some special cases

In this section we study the structure of f_{quad} assuming some relations between the matrices $\mathbf{A}_1, \dots, \mathbf{A}_k$.

I. Assume that $\mathbf{A}_1 = \cdots = \mathbf{A}_k = \mathbf{A} > 0$. Then

$$f_{\text{quad}}(\mathbf{X}) = \operatorname{tr} \mathbf{X}\mathbf{X}^T\mathbf{A} \ ,$$

and $\mathbf{X}\mathbf{X}^T$ is a projection, $dimRange(\mathbf{X}\mathbf{X}^T) = k$. This functional arises also in the so-called total least squares problem and it was analyzed in details in Byrnes and Willems. In this case the value of this functional depends only on the subspace generated by the column vectors of $\mathbf{X}$. This is an element of the Grassmannian manifold $G_{n,k}$. It was proved that there is a unique global maximum (in $G_{n,k}$) if and only if $\lambda_k > \lambda_{k+1}$, where $\lambda_1 \geq \lambda_2 > \cdots \geq \lambda_n > 0$ are the eigenvalues of $\mathbf{A}$. Also f_{quad} over $\mathbb{Z}_2$ is a perfect Morse-Bott function, i.e. there are equalities in the Morse-Bott inequalities. An immediate consequence of this statement that every local maxima is also global maxima and the set of global maxima is a connected submanifold. It is a so-called Schubert subvariety of $G_{n,k}$.

II. The matrices $\mathbf{A}_1, \dots, \mathbf{A}_k$ commute, so there exists a common eigenvector system. In this case we may assume that they are diagonal matrices

$$\mathbf{A}_j = \operatorname{diag}(\lambda_1^j, \dots, \lambda_k^j) \ . \tag{5.1}$$

It is natural to look for a global maximum in the set of matrices $\mathbf{X}$ having eigenvectors as their columns. Under the previous

condition the functional f_{quad} has the form

$$f_{\text{quad}}(\mathbf{X}) = \sum_{j=1}^{k} \sum_{i=1}^{n} \lambda_i^j x_{ij}^2 \ . \tag{5.2}$$

Let us remark that the special case $k = n$ was analyzed by Brockett [Br89]. To see this let let us mention that he considered the problem of minimizing

$$\sum_{j=1}^{k} \operatorname{tr} \mathbf{X}^T \mathbf{Q}_j \mathbf{X} \mathbf{R}_j \ , \tag{5.3}$$

where $\mathbf{Q}_j, \mathbf{R}_j$ are diagonal matrices. Denote

$$\mathbf{Q}_j = \operatorname{diag}(q_1^j, \ldots, q_n^j),$$

$$\mathbf{R}_j = \operatorname{diag}(r_1^j, \ldots, r_n^j).$$

This function can also be written as

$$\sum_{j=1}^{n} \sum_{i=1}^{n} \lambda_i^j x_{ij}^2 \ ,$$

where $\lambda_i^j = \sum_{l=1}^{k} q_i^l r_j^l$. This latter identity is nothing else than a diad-decomposition of a matrix with elements

$$\left[\lambda_i^j\right]_{i,j=1}^{n} \ .$$

Since every matrix can be written in this form we see that any functional of the form (5.1) can be written as (5.3). Brockett analyzed this problem in connection with matching two sets of vectors in the n-dimensional space by permuting the coordinate axis.

(The special case of (5.3) when $k = 1$ was analyzed by von Neumann [Ne37].)

Returning to the functional (5.1) first we show using linear programming methods that the global maximum is taken on a subset of the common eigenvectors of the matrices $\mathbf{A}_1, \ldots \mathbf{A}_k$.

Definition 5.1. An $n \times k$ matrix $\Pi = [b_{ij}]$ is called permutation matrix if there exists an injection

$$\pi : \{1, \ldots, k\} \to \{1, \ldots, n\}$$

such that

$$b_{ij} = \begin{cases} 1 & \text{if } \ i = \pi(j) \\ 0 & \text{otherwise} \end{cases}$$

THEOREM 5.1. *Assume that the matrices* $\mathbf{A}_1, \ldots, \mathbf{A}_k$ *commute. Then the global maximum of* f_{quad} *is attained on a subset of the common eigenvectors.*

Proof: Diagonalizing the matrices $\mathbf{A}_1, \ldots, \mathbf{A}_k$ we can write

$$f_{\mathrm{quad}}(\mathbf{X}) = \sum_{j=1}^{k} \sum_{i=1}^{n} \lambda_i^j x_{ij}^2 \ .$$

After this transformation the common eigenvectors of the matrices $\mathbf{A}_1, \ldots, \mathbf{A}_k$ are just the unit vectors $\mathbf{e}_1, \ldots, \mathbf{e}_n$.

Introducing the new variables

$$z_{ij} = x_{ij}^2$$

we see that f_{quad} is linear in $z_{ij}, i = 1, \ldots, n, j = 1, \ldots, k$. Instead of using that the vectors $\mathbf{x}_1, \ldots, \mathbf{x}_k$ are orthogonal we relax the conditions by maximizing

$$\sum_{j=1}^{k} \sum_{i=1}^{n} \lambda_i^j z_{ij}$$

under the conditions

$$\sum_{i=1}^{n} z_{ij} = 1 \ , \ \sum_{i=1}^{k} z_{ij} \leq 1 \ , \ z_{ij} \geq 0 \ .$$

This is a special form of the linear programming problems – the so-called transportation problem.

Since the extremal points of the condition set are given by permutation matrices the global maximum is attained on a permutation matrix Π. Although we have maximized the functional over a larger (convex) set, if the matrix $[z_{ij}]_{i,j}$ is a permutation matrix its square root taken element wise

$$\mathbf{X} = \Pi \mathbf{D}$$

($\mathbf{D} = \mathrm{diag}(\alpha_1, \ldots, \alpha_k)$, $\alpha_i = \pm 1$) is a matrix with orthonormal column vectors. In other words

$$\mathbf{X} = \left[\alpha_1 \mathbf{e}_{\pi(1)}, \ldots, \alpha_k \mathbf{e}_{\pi(k)}\right] \ ,$$

proving the theorem. ■

The next proposition describes the Hessian matrix $\mathcal{H}_{\mathbf{X}}$ when $\mathbf{X} = \Pi \mathbf{D}$, Π is a permutation matrix, $\mathbf{D}$ is diagonal with ± 1 diagonal entries. As we have seen these are the elementwise square roots of the extremal points of the larger set considered in the LP problem.

PROPOSITION 5.1. *Assume that* $\mathbf{A}_1, \ldots, \mathbf{A}_k$ *are diagonal matrices,* $\mathbf{X} = \Pi\mathbf{D}$*, where* Π *is a permutation matrix determined by the injection*

$$\pi : \{1, \ldots, k\} \to \{1, \ldots, n\},$$

$$\mathbf{D} = diag(\alpha_1, \ldots, \alpha_k), \alpha_i = \pm 1 \ .$$

Denote $J = Range(\Pi)$*. Consider the following matrices*

- *if* $1 \leq l \leq k$*,* $j \notin J$*, then let the* $(j,l)^{th}$ *element of the matrix* $\xi^{j,l}$ *be equal to* 1*, the others be zero,*
- *if* $1 \leq j < l \leq k$*, then let the* $(\pi(j), l)^{th}$ *element of* $\eta^{j,l}$ *be equal to* 1*, the* $(\pi(l), j)^{th}$ *element be equal to* -1*, the others be zero.*

Then the matrices $\xi^{j,l}$, $1 \leq l \leq k$, $j \notin J$, $\eta^{j,l}$, $1 \leq j < l \leq k$ *form a complete system of orthogonal eigenvectors of* $\mathcal{H}_{\mathbf{X}}$ *with the eigenvalues*

$$\lambda^l_j - \lambda^l_{\pi(l)} \ ,$$

and

$$(\lambda^l_{\pi(j)} + \lambda^j_{\pi(l)}) - (\lambda^l_{\pi(l)} + \lambda^j_{\pi(j)}) \ ,$$

respectively.

Proof: It is an elementary calculation to show that the matrices $\xi^{j,l}$, $\eta^{j,l}$ satisfy the eigenvector equation. Since they are orthogonal and their number $(n-k)k + \frac{k(k-1)}{2}$ coincides with the size of $\mathcal{H}_{\mathbf{X}}$ they form a complete orthogonal eigenvector system, concluding the proof. ■

COROLLARY 5.1. *Assume that* $\mathbf{A}_1, \ldots, \mathbf{A}_k$ *are commuting matrices, and let* $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ *be their common eigenvector system. Consider a subset of these eigenvectors* $\mathbf{x}_{\pi(1)}, \ldots, \mathbf{x}_{\pi(k)}$ *where* $\pi : \{1, \ldots, k\} \to \{1, \ldots, n\}$ *is an injection. Then* $\mathbf{X} = [\mathbf{x}_{\pi(1)}, \ldots, \mathbf{x}_{\pi(k)}]$ *determines a local maximum of* f_{quad} *if the elementary transforms of* $\mathbf{x}_{\pi(1)}, \ldots, \mathbf{x}_{\pi(k)}$ *decrease the value of this functional.*

Proof: First diagonalize the matrices $\mathbf{A}_1, \ldots, \mathbf{A}_k$. Now observe that an elementary transformation $\mathbf{x}_{\pi(j)} \to \mathbf{x}_{\pi(l)}$ changes the value of f_{quad} by

$$(\lambda^l_{\pi(j)} + \lambda^j_{\pi(l)}) - (\lambda^l_{\pi(l)} + \lambda^j_{\pi(j)}) \ ,$$

and the change $\mathbf{x}_{\pi(l)} \to \mathbf{x}_j$, $j \notin J$ changes f_{quad} by

$$\lambda_j^l - \lambda_{\pi(l)}^l \ .$$

According to our assumption these are negative, thus the Hessian at $\mathbf{X}$, $\mathcal{H}_{\mathbf{X}}$ is negative definite proving that $\mathbf{X}$ is a strict local maximum. ■

Example. Let $n = k = 3$ and define

$$\mathbf{A}_1 = \text{diag}(9, 7, 0) \ , \ \mathbf{A}_2 = \text{diag}(0, 5, 4) \ , \ \mathbf{A}_3 = \text{diag}(3, 0, 1) \ .$$

The previous Proposition and Corollary gives that the set

$$\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$$

determines a global maximum of f_{quad} while

$$\mathbf{e}_2, \mathbf{e}_3, \mathbf{e}_1$$

provides a strict local maximum of f_{quad}, because the elementary transformations always decrease the value of f_{quad}.

$$f_{\text{quad}}(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3) = 15 \ ,$$

$$f_{\text{quad}}(\mathbf{e}_2, \mathbf{e}_3, \mathbf{e}_1) = 14 \ .$$

Example. Let $n = k = 3$ and define

$$\mathbf{A}_1 = \text{diag}(9, 7, 0) \ , \ \mathbf{A}_2 = \text{diag}(0, 5, 5.5) \ , \ \mathbf{A}_3 = \text{diag}(2.5, 0, 1) \ .$$

Then $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$ and $\mathbf{e}_2, \mathbf{e}_3, \mathbf{e}_1$ are strict global maxima of f_{quad}, so in this case the maximum is attained on two different isolated points.

6. The behaviour of the algorithm

Let us return to the algorithm defined in (3.10). Choose an arbitrary initial point $\mathbf{X}^{(0)} = (\mathbf{x}_1^{(0)}, \ldots, \mathbf{x}_k^{(0)})$.

THEOREM 6.1. *Define the sequence* $\mathbf{X}^{(m)}$, $m \geq 0$, *using the polar decomposition*

$$\mathbf{A}(\mathbf{X}^{(m)}) = \mathbf{X}^{(m+1)}\mathbf{S}^{(m+1)} \ .$$

Then $f_{quad}(\mathbf{X}^{(m)})$ *is a nondecreasing sequence and*

$$dist(\mathbf{X}^{(m)}, \mathcal{C}) \to 0 \ , \ \text{as } m \to \infty \ .$$

Proof: As we pointed out $\mathbf{X}^{(m+1)}$ is obtained maximizing $f_{\text{bilin}}(\mathbf{X}^{(m)}, \mathbf{Y})$ in $\mathbf{Y}$. In particular

$$f_{\text{bilin}}(\mathbf{X}^{(m)}, \mathbf{X}^{(m+1)}) \geq f_{\text{quad}}(\mathbf{X}^{(m)}) \ .$$

Inequality (3.16) implies that

$$f_{\text{quad}}(\mathbf{X}^{(m+1)}) \geq f_{\text{bilin}}(\mathbf{X}^{(m)}, \mathbf{X}^{(m+1)})$$

proving that $f_{\text{quad}}(\mathbf{X}^{(m)})$ is nondecreasing. Since $V_{n,k}$ is a compact set, this sequence is bounded thus it is a convergent sequence.

Since the matrices $\mathbf{A}_1, \ldots \mathbf{A}_k$ are positive definite there exists a constant c such that

$$\|\mathbf{X} - \mathbf{Y}\| \leq c \sum_{i=1}^{k} (\mathbf{y}_i - \mathbf{x}_i)^T \mathbf{A}_i (\mathbf{y}_i - \mathbf{x}_i)$$

for any systems of orthonormal vectors $\mathbf{x}_1, \ldots, \mathbf{x}_k$ and $\mathbf{y}_1, \ldots \mathbf{y}_k$. The previous inequalities show that

$$f_{\text{quad}}(\mathbf{X}^{(m+1)} - \mathbf{X}^{(m)}) \to 0 \ ,$$

thus $\|\mathbf{X}^{(m+1)} - \mathbf{X}^{(m)}\| \to 0$. Furthermore, since

$$\begin{aligned} f_{\text{quad}}(\mathbf{X}^{(m+1)} - \mathbf{X}^{(m)}) &= f_{\text{quad}}(\mathbf{X}^{(m+1)}) + f_{\text{quad}}(\mathbf{X}^{(m)}) \\ &\quad -2 f_{\text{bilin}}(\mathbf{X}^{(m+1)}, \mathbf{X}^{(m)}) \\ &\leq f_{\text{quad}}(\mathbf{X}^{(m+1)}) - f_{\text{quad}}(\mathbf{X}^{(m)}) \end{aligned}$$

we obtain that

$$\sum_{m=0}^{\infty} \|\mathbf{X}^{(m+1)} - \mathbf{X}^{(m)}\|^2 < \infty \ .$$

At the same time, if $\mathbf{X}$ is a limit point of the sequence $\mathbf{X}^{(m)}$, $m \geq 0$, then obviously

$$\mathbf{A}(\mathbf{X}) = \mathbf{X}\mathbf{S}, \ \mathbf{X}^T\mathbf{X} = \mathbf{I}_k, \ \mathbf{S} \geq 0 \ ,$$

thus $\mathbf{X}$ is a critical point. The compactness of $\mathcal{C}$ implies that

$$\text{dist}(\mathbf{X}^{(m)}, \mathcal{C}) \to 0$$

■

Remark. In the case when f_{quad} has isolated critical points the previous Theorem shows that the algorithm *converges* to one of the critical points. But, in general, when the decomposition $\mathcal{C} = \cup \mathcal{C}_j$, the submanifolds $\mathcal{C}_j$ are not zero dimensional, a more sophisticated analysis is required.

THEOREM 6.2. *Assume that $k < n$. If there exists a global maximum $\mathbf{X}$ among the limit points of the sequence $\mathbf{X}^{(m)}$, $m \geq 0$, then*

$$\mathbf{X}^{(m)} \to \mathbf{X} \ , \ \textit{as } m \to \infty \ .$$

Proof: The proof is based on the second order approximation of the algorithm (3.10). Let $\mathbf{X}$ be such a critical point of f_{quad} for which the vectors $\mathbf{A}_1\mathbf{x}_1, \ldots, \mathbf{A}_k\mathbf{x}_k$ are linearly independent. Consider two - three times differentiable - curves $\mathbf{X}(t)$, $\mathbf{Z}(t) \in V_{n,k}$ defined in a neighbourhood of zero such that $\mathbf{X}(0) = \mathbf{Z}(0) = \mathbf{X}$. Denote $\mathbf{X}'(0) = \xi$, $\mathbf{Z}'(0) = \eta$. Apply one step of the algorithm to each points on the curves $\mathbf{X}(t)$, $\mathbf{Z}(t)$. In this way we obtain the curves $\widehat{\mathbf{X}(t)}$, $\widehat{\mathbf{Z}(t)}$, i.e.

$$\begin{aligned} \mathbf{A}(\mathbf{X}(t)) &= \widehat{\mathbf{X}(t)}\mathbf{A}(\mathbf{X}(t))^T\widehat{\mathbf{X}(t)} \text{ and } \mathbf{A}(\mathbf{X}(t))^T\widehat{\mathbf{X}(t)} > 0 \ , \\ \mathbf{A}(\mathbf{Z}(t)) &= \widehat{\mathbf{Z}(t)}\mathbf{A}(\mathbf{Z}(t))^T\widehat{\mathbf{Z}(t)} \text{ and } \mathbf{A}(\mathbf{Z}(t))^T\widehat{\mathbf{Z}(t)} > 0 \ . \end{aligned}$$

Denote the derivatives of these curves at zero by

$$\begin{aligned} h(\xi) &= \widehat{\mathbf{X}(0)}' \ , \\ h(\eta) &= \widehat{\mathbf{Z}(0)}' \ . \end{aligned}$$

We show that

$$\operatorname{tr} h(\xi)^T\mathbf{A}(\eta) = \operatorname{tr} h(\eta)^T\mathbf{A}(\xi) \ , \tag{6.1}$$

and

$$\operatorname{tr} \xi^T\mathbf{A}(\eta) = \operatorname{tr} h(\xi)\mathbf{A}(\mathbf{X})^T\mathbf{X}\eta^T = \operatorname{tr} \xi\mathbf{A}(\mathbf{X})^T\mathbf{X}h(\eta)^T \ . \tag{6.2}$$

To this observe that differentiating the matrix $\mathbf{A}(\mathbf{Z}(t))^T\widehat{\mathbf{Z}(t)}$ at zero we obtain that

$$\mathbf{A}(\eta)^T\mathbf{X} + \mathbf{A}(\mathbf{X})^Th(\eta) = \mathbf{X}^T\mathbf{A}(\eta) + h(\eta)^T\mathbf{A}(\mathbf{X}) \ .$$

Since $\xi^T\mathbf{X}$ is skew-symmetric, we get that

$$\operatorname{tr} \xi^T\mathbf{X}\left[\mathbf{A}(\eta)^T + \mathbf{A}(\mathbf{X})^Th(\eta)\right] = 0 \ . \tag{6.3}$$

Differentiating the identity

$$\mathbf{A}(\mathbf{X}(t)) = \widehat{\mathbf{X}(t)}\mathbf{A}(\mathbf{X}(t))^T\widehat{\mathbf{X}(t)}$$

at zero we obtain that

$$\mathbf{A}(\xi) = h(\xi)\mathbf{A}(\mathbf{X})^T\mathbf{X} + \mathbf{X}\mathbf{A}(\xi)^T\mathbf{X} + \mathbf{X}\mathbf{A}(\mathbf{X})^Th(\xi) \ .$$

Similarly

$$\mathbf{A}(\eta) = h(\eta)\mathbf{A}(\mathbf{X})^T\mathbf{X} + \mathbf{X}\mathbf{A}(\eta)^T\mathbf{X} + \mathbf{X}\mathbf{A}(\mathbf{X})^T h(\eta) \ .$$

This gives that

$$\mathbf{X}\mathbf{A}(\eta)^T\mathbf{X} + \mathbf{X}\mathbf{A}(\mathbf{X})^T h(\eta) = \mathbf{A}(\eta) - h(\eta)\mathbf{A}(\mathbf{X})^T\mathbf{X} \ .$$

Substituting this into (6.3) we get that

$$\mathrm{tr}\, \xi^T\mathbf{A}(\eta) = \mathrm{tr}\, h(\eta)\mathbf{A}(\mathbf{X})^T\mathbf{X}\xi^T = \mathrm{tr}\, \xi\mathbf{A}(\mathbf{X})^T\mathbf{X}h(\eta)^T$$

using that $\mathbf{A}(\mathbf{X})^T\mathbf{X}$ is symmetric. This proves (6.1). Instead of ξ using $h(\xi)$ in this identity we conclude that

$$\mathrm{tr}\, h(\xi)^T\mathbf{A}(\eta) = \mathrm{tr}\, h(\eta)\mathbf{A}(\mathbf{X})^T\mathbf{X}h(\xi)^T \tag{6.4}$$

must hold, proving (6.2) by symmetry. Now, if $\xi \in Ker\mathcal{H}_{\mathbf{X}}$ then there exists a curve $\mathbf{X}(t),\ \mathbf{X}(0) = 0,\ \mathbf{X}'(0) = \xi$ such that every element $\mathbf{X}(t)$ is a critical point of f_{quad}. Consequently, $\widehat{\mathbf{X}(t)} = \mathbf{X}(t)$, i.e. $h(\xi) = \xi$. This implies that

$$\mathrm{tr}\, \eta^T\mathbf{A}(\xi) = \mathrm{tr}\, h(\xi)^T\mathbf{A}(\eta) = \mathrm{tr}\, h(\eta)^T\mathbf{A}(\xi) \ .$$

In particular, if η is orthogonal to $Ker\mathcal{H}_{\mathbf{X}}$ with respect to the scalar product defined by $< \xi, \eta >= \mathrm{tr}\, \eta^T\mathbf{A}(\xi)$, then $h(\eta)$ will be also orthogonal to $Ker\mathcal{H}_{\mathbf{X}}$. Consider now a second order approximation of the functionals

$$f_{\mathrm{quad}}(\mathbf{X}(t)),\ f_{\mathrm{bilin}}(\mathbf{X}(t), \widehat{\mathbf{X}(t)}),\ f_{\mathrm{quad}}(\widehat{\mathbf{X}(t)}) \ .$$

$$\mathbf{X}(t) = \mathbf{X} + \xi t + \frac{1}{2}\mathbf{X}''(0)t^2 + o(t^3) \ ,$$
$$\widehat{\mathbf{X}(t)} = \mathbf{X} + h(\xi)t + \frac{1}{2}\widehat{\mathbf{X}''(0)}t^2 + o(t^3) \ ,$$

$$f_{\mathrm{quad}}(\mathbf{X}(t)) = f_{\mathrm{quad}}(\mathbf{X}) + t\left[\mathrm{tr}\, \xi^T\mathbf{A}(\mathbf{X}) + \mathrm{tr}\, \mathbf{X}^T\mathbf{A}(\xi)\right]$$
$$+\frac{t^2}{2}\left[\mathrm{tr}\, \mathbf{X}''(0)^T\mathbf{A}(\mathbf{X}) + \mathrm{tr}\, \mathbf{X}^T\mathbf{A}(\mathbf{X}''(0)) + 2\mathrm{tr}\, \xi^T\mathbf{A}(\xi)\right] + o(t^3) \ .$$

Since $\mathbf{X} \in \mathcal{C}$ we have that $\mathrm{tr}\, \xi^T\mathbf{A}(\mathbf{X}) = 0$ and

$$\mathbf{A}(\mathbf{X}) = \mathbf{X}\mathbf{A}(\mathbf{X})^T\mathbf{X} \ .$$

Using that $\mathbf{X}''(0)\mathbf{X} + \xi^T\xi$ is skew-symmetric we obtain that

$$f_{\text{quad}}(\mathbf{X}(t)) = f_{\text{quad}}(\mathbf{X}) + t^2 \left[\operatorname{tr} \xi^T \mathbf{A}(\xi) - \operatorname{tr} \xi \mathbf{A}(\mathbf{X})^T \mathbf{X} \xi^T\right] + o(t^3) \ .$$

Similarly

$$\begin{aligned} f_{\text{bilin}}(\widehat{\mathbf{X}(t)},\ \mathbf{X}(t)) &= f_{\text{quad}}(\mathbf{X}) + t^2[\operatorname{tr} h(\xi)^T \mathbf{A}(\xi) \\ &\quad - \frac{1}{2}\left(\operatorname{tr} \xi \mathbf{A}(\mathbf{X})^T \mathbf{X} \xi^T + \operatorname{tr} h(\xi) \mathbf{A}(\mathbf{X})^T \mathbf{X} h(\xi)^T\right)] + o(t^3) \ , \end{aligned}$$

and

$$\begin{aligned} f_{\text{quad}}(\widehat{\mathbf{X}(t)}) &= f_{\text{quad}}(\mathbf{X}) \\ &\quad + t^2 \left[\operatorname{tr} h(\xi)^T \mathbf{A}(h(\xi)) - \operatorname{tr} h(\xi) \mathbf{A}(\mathbf{X})^T \mathbf{X} h(\xi)^T\right] + o(t^3) \ . \end{aligned}$$

But the construction gives that

$$f_{\text{quad}}(\widehat{\mathbf{X}(t)}) \geq f_{\text{bilin}}(\mathbf{X}(t), \widehat{\mathbf{X}(t)}) \geq f_{\text{quad}}(\mathbf{X}(t))$$

so the same inequalities hold for the second derivatives, i.e.

$$\begin{aligned} &\operatorname{tr} \xi^T \mathbf{A}(\xi) - \operatorname{tr} \xi \mathbf{A}(\mathbf{X})^T \mathbf{X} \xi^T \\ &\qquad \leq \frac{1}{2}\left[\operatorname{tr} h(\xi) \mathbf{A}(\mathbf{X})^T \mathbf{X} h(\xi)^T - \operatorname{tr} \xi \mathbf{A}(\mathbf{X})^T \mathbf{X} \xi^T\right] \\ &\qquad\qquad \leq \operatorname{tr} h(\xi)^T \mathbf{A}(h(\xi)) - \operatorname{tr} h(\xi) \mathbf{A}(\mathbf{X})^T \mathbf{X} h(\xi)^T \end{aligned}$$

where in the second term we have applied (6.2).

If $\mathbf{X}$ is a global maximum point then according to Corollary 2 the vectors $\mathbf{A}_1\mathbf{x}_1, \ldots, \mathbf{A}_k\mathbf{x}_k$ are linearly independent so the previous considerations can be applied. But now

$$f_{\text{quad}}(\widehat{\mathbf{X}(t)}) \leq f_{\text{quad}}(\mathbf{X}) \ ,$$

so

$$\operatorname{tr} h(\xi)^T \mathbf{A}(h(\xi)) \leq \operatorname{tr} h(\xi) \mathbf{A}(\mathbf{X})^T \mathbf{X} h(\xi)^T = \operatorname{tr} h(\xi)^T \mathbf{A}(\xi)$$

must hold. Inequality (3.16) gives that

$$\operatorname{tr} h(\xi)^T \mathbf{A}(h(\xi)) \leq \operatorname{tr} \xi^T \mathbf{A}(\xi) = \operatorname{tr} \xi \mathbf{A}(\mathbf{X})^T \mathbf{X} h(\xi)^T \ .$$

Moreover

$$\operatorname{tr} h(\xi)^T \mathbf{A}(h(\xi)) + \operatorname{tr} \xi^T \mathbf{A}(\xi) \geq 2 \operatorname{tr} h(\xi)^T \mathbf{A}(\xi)$$

implying that

$$0 \geq \operatorname{tr} h(\xi)^T \mathbf{A}(h(\xi)) - \operatorname{tr} h(\xi) \mathbf{A}(\mathbf{X})^T \mathbf{X} h(\xi)^T$$
$$\geq \frac{1}{2} \left(\operatorname{tr} h(\xi)^T \mathbf{A} h(\xi)) - \operatorname{tr} \xi^T \mathbf{A}(\xi) \right) .$$

Now, if ξ is orthogonal to $Ker\mathcal{H}_{\mathbf{X}}$ in the scalar product introduced above then $h(\xi)$ is also orthogonal, thus

$$\operatorname{tr} h(\xi)^T \mathbf{A}(h(\xi)) < \operatorname{tr} \xi^T \mathbf{A}(\xi) \ , \text{ if } \xi \perp Ker\mathcal{H}_{\mathbf{X}} \ .$$

Using the compactness of $\mathcal{C}$ we obtain that

$$c_0 := \sup_{\mathbf{X} \in \mathcal{C}} \sup_{\xi} \frac{\operatorname{tr} h(\xi)^T \mathbf{A}(h(\xi))}{\operatorname{tr} \xi^T \mathbf{A}(\xi)} < 1 \ , \tag{6.5}$$

where the second sup is taken over the $n \times k$ matrices ξ for which

$$\xi^T \mathbf{X} \text{ is skew-symmetric}$$

and

$$\operatorname{tr} \xi^T \mathbf{A}(\eta) = 0 \text{ for every } \eta \in Ker\mathcal{H}_{\mathbf{X}} \ .$$

Now, let us convert $V_{n,k}$ into a Riemannian manifold introducing the quadratic forms of the tangent spaces $T_{\mathbf{X}}$ of $V_{n,k}$ at $\mathbf{X}$ defined by

$$\operatorname{tr} \xi^T \mathbf{A}(\xi)$$

where $\xi^T \mathbf{X}$ is skew-symmetric.

Consider the sequence of k-tuples of orthonormal vectors $\mathbf{X}^{(0)}, \mathbf{X}^{(1)}, \ldots$ produced by the algorithm. Let $\mathbf{X}$ be a limit point of this sequence and assume that $\mathbf{X}$ is a global maximum of f_{quad}. Denote by $\widehat{\mathcal{C}}$ the connected submanifold of $\mathcal{C}$ containing $\mathbf{X}$. Eject a geodesic curve from every element of the sequence $\mathbf{X}^{(0)}, \mathbf{X}^{(1)}, \ldots$ to $\widehat{\mathcal{C}}$. Assume that the curves are arclength parameterized, their value at zero is $\mathbf{Z}^{(m)}$. Denote their derivatives at zero by $\xi^{(0)}$. Then

$$\mathbf{X}^{(m)} = \mathbf{Z}^{(m)} + \xi^{(m)} d_R(\mathbf{X}^{(m)}, \mathbf{Z}^{(m)}) + o(d_r(\mathbf{X}^{(m)}, \mathbf{Z}^{(m)})^2)$$

where $d_r(\mathbf{X}^{(m)}, \mathbf{Z}^{(m)})$is the Riemannian distance between $\mathbf{X}^{(m)}$ and $\mathbf{Z}^{(m)}$, the remainder term is uniform over $\widehat{\mathcal{C}}$. Because of the parameterization of these geodesic curves

$$\operatorname{tr} \xi^{(m)} \mathbf{A}(\xi^{(m)}) = 1 \ .$$

Applying one step of the algorithm to the points on this geodesic we get that

$$\mathbf{X}^{(m+1)} = \mathbf{Z}^{(m)} + h(\xi^{(m)}) d_R(\mathbf{X}^{(m)}, \mathbf{Z}^{(m)}) + o(d_r(\mathbf{X}^{(m)}, \mathbf{Z}^{(m)})^2) \ ,$$

implying that

$$d_r(\mathbf{X}^{(m+1)}, \mathbf{Z}^{(m)}) \leq c_0 d_r(\mathbf{X}^{(m)}, \mathbf{Z}^{(m)}) + K d_r(\mathbf{X}^{(m)}, \mathbf{Z}^{(m)})^2 \ ,$$

where K does not depend on the sequence $\mathbf{X}^{(0)}, \mathbf{X}^{(1)}, \dots$. This gives that there exists a $c < 1$ such that if m is large enough then

$$d_r(\mathbf{X}^{(m+1)}, \mathbf{Z}^{(m+1)}) \leq d_r(\mathbf{X}^{(m+1)}, \mathbf{Z}^{(m)}) \leq c d_r(\mathbf{X}^{(m)}, \mathbf{Z}^{(m)}) \ .$$

Thus

$$d_r(\mathbf{X}^{(m+1)}, \mathbf{Z}^{(m)}) \text{ and } d_r(\mathbf{X}^{(m+1)}, \mathbf{Z}^{(m+1)})$$

tend to zero exponentially fast, so $\mathbf{Z}^{(m)}$ and together with this $\mathbf{X}^{(m)}$ is convergent exponentially fast, as well, proving our Theorem. ■

7. Possible generalizations to convex functions restricted to the Stiefel-manifold

Instead of sums of quadratic functions we might consider convex functions $f_1, \dots f_k$ defined on $\mathbb{R}^n$ and study the problem

$$f(\mathbf{X}) = \sum_{i=1}^{k} f_i(\mathbf{x}_i) \to max$$

under the condition that $(\mathbf{x}_1, \dots, \mathbf{x}_k)$ form an orthonormal system. Let us observe that a step of the algorithm considered in Section 3 can be formulated as maximizing the linear approximation of f_{quad} at $\mathbf{X}^{(m)}$ over the Stiefel manifold $V_{n,k}$, i.e.

$$\operatorname{tr} \mathbf{Y}^T \mathbf{A}(\mathbf{X}) \to \max_{(\mathbf{Y} \in V_{n,k})} \ .$$

This idea can be applied to general convex functions, as well. Starting with the element $\mathbf{X}(0) \in V_{n,k}$ this produces a sequence $\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \dots \in V_{n,k}$. Since the supporting hyperplane of a convex function is always below the graph of the function we obtain that

$$f(\mathbf{X}^{(m)}) \leq f(\mathbf{X}^{(m+1)}) \ .$$

The compactness of $V_{n,k}$ implies again that this nondecreasing sequence is convergent.

Now we construct an example showing that in general it may happen that the sequence $\mathbf{X}^{(m)}$ is not convergent although its limit points are global maxima of the functional f. Let $n = 3$, $k = 1$, i.e. we would like to maximize a convex function $f(x)$ on the 3-dimensional sphere S^2. Every element on this surface can be given by its polar coordinates. If $\mathbf{y} \in S^2$, then $\alpha(\mathbf{y})$ denotes its longitude, $\beta(\mathbf{Y})$ denotes its latitude. We define the convex function f as the supremum of linear functions in the way that the equator – the set of points with zero latitude – will be the set of global maxima. To this aim first choose an increasing sequence of integers $k_j \geq 8$, $j \geq 1$ for which $\sum_{j=1}^{\infty} \frac{1}{\sqrt{k_j}} < \infty$. Let $l_n = \sum_{j=n}^{\infty} \frac{4\pi}{\sqrt{k_j}}$, $m_n = \sum_{j=1}^{n-1} k_j$. Define the sequence $\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \cdots \in S^2$ as follows

$$\alpha(\mathbf{x}_j) = (j - m_n)\frac{2\pi}{k_n} \quad \text{if} \quad m_n \leq j < m_{n+1}$$
$$\beta(\mathbf{x}_j) = l_n - \frac{j - m_n}{k_n}\frac{4\pi}{\sqrt{k_n}} \quad \text{if} \quad m_n \leq j < m_{n+1} \ .$$

This sequence lies on a "spiral curve" converging to the equator. Consider the function

$$f(\mathbf{x}) = \sup_j c_j \mathbf{x}_{j+1}^T \mathbf{x} \ ,$$

where $c_j \geq 0$ is an appropriate sequence. We are going to show that this sequence can be chosen in such a way that the equator is the set of global maxima of f and if the algorithm starts at $\mathbf{x}_0$ then the sequence produced by the algorithm is exactly the sequence $\mathbf{x}_j$, $j \geq 0$. Since the linear function $c_j \mathbf{x}_{j+1}^T \mathbf{x}$ has a unique global maximum on S^2 at $\mathbf{x}_{j+1}$ to the last statement it is enough to check that the linear approximation at $\mathbf{x}_j$ is equal to $c_j \mathbf{x}_{j+1}^T \mathbf{x}$ in other words this is the supporting hyperplane at $\mathbf{x}_j$. To this aim the inequalities

$$c_j \mathbf{x}_{j+1}^T \mathbf{x}_j > c_l \mathbf{x}_{l+1}^T \mathbf{x}_j \quad l \neq j \tag{7.1}$$

should be fulfilled. Taking first $l = j - 1$ we get

$$c_j > c_{j-1} \frac{1}{\mathbf{x}_{j+1}^T \mathbf{x}_j} \ .$$

But, if $m_n \leq j < m_{n+1}$ then $\mathbf{x}_{j+1}^T \mathbf{x}_j = cos\frac{2\pi}{k_n} > 1 - \frac{1}{2}\left(\frac{2\pi}{k_n}\right)^2$ so it is enough to assume that

$$\frac{c_j}{c_{j-1}} = \frac{1}{1 - \frac{1}{2}\left(\frac{2\pi}{k_n}\right)^2} \ .$$

Now comparing the points $\mathbf{x}_{m_n}, \mathbf{x}_{m_{n+1}}$ we have to check the inequality

$$c_{m_n} \mathbf{x}_{m_n+1}^T \mathbf{x}_{m_n} > c_{m_{n+1}} \mathbf{x}_{m_{n+1}+1}^T \mathbf{x}_{m_n} \ . \tag{7.2}$$

But $\mathbf{x}_{m_n+1}^T \mathbf{x}_{m_n} = cos\frac{2\pi}{k_n}$, $\mathbf{x}_{m_{n+1}+1}^T \mathbf{x}_{m_n} < \mathbf{x}_{m_{n+1}}^T \mathbf{x}_{m_n} = cos\frac{4\pi}{\sqrt{k_j}}$. Since

$$\begin{aligned}
\frac{c_{m_n}}{c_{m_{n+1}}} cos\frac{2\pi}{k_n} &> \left[1 - \frac{1}{2}\left(\frac{2\pi}{k_n}\right)^2\right]^{k_n} \left[1 - \frac{1}{2}\left(\frac{2\pi}{k_n}\right)^2\right] \\
&> exp\left[-\left(\frac{2\pi}{k_n}\right)^2 (1 + k_n)\right] > 1 - \frac{(2\pi)^2}{k_n} - \frac{(2\pi)^2}{k_n^2} \\
&> 1 - \frac{1}{2}\left(\frac{4\pi}{\sqrt{k_n}}\right)^2 + \frac{1}{4!}\left(\frac{4\pi}{\sqrt{k_n}}\right)^4 > cos\frac{4\pi}{\sqrt{k_n}} \ ,
\end{aligned}$$

we see that (7.2) holds true. Similar calculations show that (7.1) is fulfilled for every $l \neq j$. Observe that the previous calculation gives that c_n is an increasing bounded sequence. Obviously

$$f(\mathbf{x}) < \sup_n c_n$$

for any $\mathbf{x} \in S^2$. On the other hand if $\beta(\mathbf{x}) = 0$ then there exists a subsequence $\mathbf{x}_{n_k}$ converging to $\mathbf{x}$, consequently

$$c_{n_k} \mathbf{x}_{n_k+1} \mathbf{x} \to \lim_{n\to\infty} c_n$$

thus $\mathbf{x}$ is a global maximum point.

REFERENCES

[AM86] G. Ammar - C. Martin. The geometry of matrix eigenvalue methods, *Acta Applicandae Mathematica* 1986,, Vol. 5. pp. 239-279.

[Bo82] M. Bolla, *Doctoral dissertation,* 1982, pp. 67-68.

[Bo54] R. Bott, Nondegenerate critical manifolds, *Annals of Mathematics,* 1954, Vol. 60 pp. 248-261.

[Br89] R.W. Brockett, Least squares matching problems, *Linear Algebra and its Application,* 1989, Vol. 122/123/124 pp. 761-777

[BW86] C.I. Byrnes - Jan C. Willems, Least-square estimation, linear programming, and momentum: a geometric parameterization of local minima, *IMA Journal of Math. Control and Information* 1986, Vol. 3. pp. 103-118.

[Fuhr81] P. A. Fuhrmann, *Linear Operators and Systems in Hilbert space* McGraw Hill, 1981.

[Ja76] I.M. James *The Topology of Stiefel Manifolds* London Mathematical Society,, Lecture Notes Series 24, Cambridge, 1976

[Ne37] J. von Neumann, Some matrix-inequalities and metrization of matrix-spaces, *Tomsk Univ. Rev.* 1937, Vol. 1 pp. 286-300.

[BK84] V. V. Vojevogyin - Ju. A. Kuznyecov, *Matrices and its computations* Nauka, 1984 (in Russian).