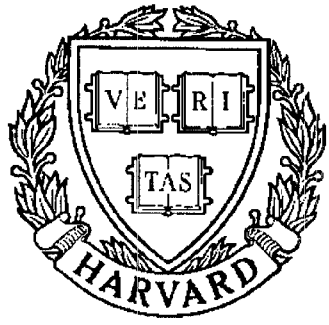
TECHNICAL RESEARCH REPORT

SYSTEMS RESEARCH CENTER



Supported by the National Science Foundation Engineering Research Center Program (NSFD CD 8803012), Industry and the University

Critical Points of Matrix Least Square Distance Functions

by U. Helmke and M.A. Shayman

TR 92-30

Critical Points of Matrix Least Square Distance Functions

Uwe Helmke* and Mark A. Shayman+

Abstract

A classical problem in matrix analysis and total least squares estimation is that of finding a best approximant of a given matrix by lower rank ones. In this paper the critical points and the local minimum of the distance function $f_A(X) = \|A - X\|^2$ on varieties of fixed rank symmetric, skew-symmetric and rectangular matrices X are determined. Our results extend earlier ones of Eckart and Young and Higham.

* Department of Mathematics, Regensburg University, 8400 Regensburg, Germany.

+ Electrical Engineering Department and Systems Research Center, University of Maryland, College Park, Maryland 20742, USA. Research partially supported by the National Science Foundation under Grant ECS-8696108.

February 1992

1. Introduction

A classical result from matrix analysis, the Eckart-Young-Mirsky Theorem, states that the best approximation of a given $M \times N$-matrix A by matrices of smaller rank is given by a truncated singular value decomposition of A. In more precise terms, an approximation $\hat{X} \in \mathbf{R}^{M\times N}$ of rank $\hat{X} \leq n$ to $A \in \mathbf{R}^{M\times N}$ is to be found that satisfies

$$\|A - \hat{X}\| = \inf \{\|A - X\| \quad | \text{ rank } X \leq n\}$$

and the theorem of Eckart-Young and Mirsky solves this problem for any unitarily invariant matrix norm; see Golub, Hoffman and Stewart (1987). Such results are of interest for Total Least Squares Approximation; see Golub and Van Loan (1980). The classical paper of Eckart and Young (1936) considers the case where $\|X\|_F = (\operatorname{tr} X'X)^{\frac{1}{2}}$ is the Frobenius norm. In this case $f_A(X) = \|A - X\|^2$ defines a smooth function in X. Since the variety $M(n, M \times N)$ of real $M \times N$-matrices X of fixed rank n is a smooth manifold (see section 4) we thus have an optimization problem for the smooth distance function

$$\begin{aligned} f_A &: M(n, M \times N) \longrightarrow \mathbf{R} \\ f_A(X) &= \quad \|A - X\|_F^2 \end{aligned} \tag{1.1}$$

defined on $M(n, M \times N)$. The fact that we have in (1.1) assumed that the rank of X is precisely n instead of being $\leq n$ is no restriction in generality, because we will later prove that any best approximant $\hat{X}$ of A of rank $\leq n$ will automatically satisfy rank $\hat{X} = n$.

In this paper we study the approximation problem of a linear operator by lower rank operators. Related questions arise in control theory where one is interested in finding best approximations of linear systems by lower order ones. The results obtained in this paper may be viewed as a prototype for an investigation one would like to pursue in model reduction theory of linear systems.

The critical points of the distance function $f_A(X) = \|A - X\|^2$ on varieties of fixed rank symmetric, skew-symmetric and rectangular matrices X are investigated. The critical points of f_A and their local stability properties are determined and the number of local minima are determined. Such results are of interest if gradient flow methods are used in order to find the best approximant. Our results contain those of Eckart and Young, as well as those of Higham (1988) and others as special cases.

2. Approximations by Symmetric Matrices

For integers $1 \leq n \leq N$ let

$$S(n,N) = \{X \in \mathbb{R}^{N\times N} \mid X' = X,\ \operatorname{rank} X = n\} \tag{2.1}$$

denote the set of real symmetric $N \times N$ matrices of rank n. Given a fixed real symmetric $N \times N$ matrix A we consider the distance function

$$\begin{aligned} f_A &: S(n,N) \longrightarrow \mathbb{R} \\ X &\longmapsto \|A - X\|^2 \end{aligned} \tag{2.2}$$

where $\|X\|^2 = \operatorname{tr}(XX')$ is the Frobenius norm. We are interested in finding the critical points and local and global minima of the function f_A, i.e. in particular the best rank n symmetric approximants of A.

The following result summarizes some well known geometric properties of the constraint set $S(n,N)$.

Proposition 2.1

(i) $S(n,N)$ is a smooth manifold of dimension $\frac{1}{2}n(2N-n+1)$ and has $n+1$ connected components

$$S(p,q;N) = \{X \in S(n,N) \mid \text{signature } X = p - q\} \tag{2.3}$$

$p, q \geq 0$, $p + q = n$. The tangent space of $S(n,N)$ at an element X is

$$T_X S(n,N) = \{\Delta X + X\Delta' \mid \Delta \in \mathbb{R}^{N\times N}\} \tag{2.4}$$

(ii) The topological closure of $S(p,q;N)$ in $\mathbb{R}^{N\times N}$ is

$$\overline{S(p,q;N)} = \bigcup_{\substack{p'\leq p \\ q'\leq q}} S(p',q';N) \tag{2.5}$$

and the topological closure of $S(n,N)$ in $\mathbb{R}^{N\times N}$ is

$$\overline{S(n,N)} = \bigcup_{i=0}^{n} S(i,N). \tag{2.6}$$

Proof

By Sylvester's inertia theorem, the function which associates to every $X \in S(n,N)$

its signature is locally constant. Thus $S(p,q;N)$ is open in $S(n,N)$ with $S(n,N) = \bigcup_{\substack{p+q=n \\ p,q\geq 0}} S(p,q;N)$. It follows that $S(n,N)$ has at least $n+1$ connected components.

Any element $X \in S(p,q;N)$ is of the form $X = \gamma Q \gamma'$ where $\gamma \in \mathrm{GL}(N,\mathbb{R})$ is invertible and $Q = \mathrm{diag}(I_p, -I_q, 0) \in S(p,q;N)$. Thus $S(p,q;N)$ is an orbit of the congruence group action $(\gamma, X) \mapsto \gamma X \gamma'$ of $\mathrm{GL}(N,\mathbb{R})$ on $S(n,N)$. Thus $S(p,q;N)$, $p,q \geq 0$, $p+q=n$, is a smooth manifold and therefore also $S(n,N)$. Let $\mathrm{GL}^+(N,\mathbb{R})$ denote the set of invertible $N \times N$ matrices γ with $\det \Gamma > 0$. Then $\mathrm{GL}^+(N,\mathbb{R})$ is a connected subset of $\mathrm{GL}(N,\mathbb{R})$ and $S(p,q;N) = \{\gamma Q \gamma' | \gamma \in \mathrm{GL}^+(N;\mathbb{R})\}$. Consider the smooth surjective map

$$\eta: \mathrm{GL}(N,\mathbb{R}) \longrightarrow S(p,q;N),\ \eta(\gamma) = \gamma Q \gamma'. \tag{2.7}$$

Then $S(p,q;N)$ is the image of the connected set $\mathrm{GL}^+(N,\mathbb{R})$ under the continuous map η and therefore $S(p,q;N)$ is connected. This completes the proof that $S(p,q;N)$ are precisely the $n+1$ connected components of $S(n,N)$. Furthermore, the derivative of η at $\gamma \in \mathrm{GL}(N,\mathbb{R})$ is the linear map $D_\eta(\gamma): T_\gamma \mathrm{GL}(N,\mathbb{R}) \to T_X S(p,q;N)$, $X = \gamma Q \gamma'$, defined by $D_\eta(\gamma)(\Delta) = \Delta X + X \Delta'$ and maps $T_\gamma \mathrm{GL}(N,\mathbb{R}) \cong \mathbb{R}^{N \times N}$ surjectively onto $T_X S(p,q;N)$. Since X and $p,q \geq 0$, $p+q=n$, are arbitrary this proves (2.4). Let

$$\Delta = \begin{bmatrix} \Delta_{11} & \Delta_{12} & \Delta_{13} \\ \Delta_{21} & \Delta_{22} & \Delta_{23} \\ \Delta_{31} & \Delta_{32} & \Delta_{33} \end{bmatrix}$$

be partitioned according to the partition $(p,q,N-n)$ of N. Then $\Delta \in \ker D_\eta(I)(\Delta)$ if and only if $\Delta_{13} = 0$, $\Delta_{23} = 0$ and the $n \times n$ submatrix

$$\begin{bmatrix} \Delta_{11} & \Delta_{12} \\ \Delta_{21} & \Delta_{22} \end{bmatrix}$$

is skew-symmetric. A simple dimension count thus yields $\dim \ker D_\eta(I) = \binom{n}{2} + N(N-n)$ and therefore

$$\dim S(p,q;N) = N^2 - \dim \ker D_\eta(I) = \frac{1}{2} n(2N - n + 1)$$

This completes the proof of (i). The proof of (ii) is left as an exercise to the reader.

□

Theorem 2.2

(i) Let $A \in \mathbb{R}^{N \times N}$ be symmetric and let

$$N_+ = \dim \mathrm{Eig}^+(A),\ N_- = \dim \mathrm{Eig}^-(A) \tag{2.8}$$

be the numbers of positive and negative eigenvalues of A, respectively. The critical points X of the distance function $f_A: S(n,N) \to \mathbb{R}$ are characterized by $AX = X^2$.

(ii) If $A = \Theta \operatorname{diag}(\lambda_1, \ldots, \lambda_N)\Theta'$, $\Theta \in O(N)$, has N distinct eigenvalues $\lambda_1 < \ldots < \lambda_N$, then the restriction of the distance function $f_A: S(p,q;N) \to \mathbb{R}$ has exactly

$$\binom{N_+}{p} \cdot \binom{N_-}{q}$$

critical points. In particular, f_A has critical points in $S(p,q;N)$ if and only if $p \leq N_+$ and $q \leq N_-$. The critical points $X \in S(p,q;N)$ of f_A, $p \leq N_+, q \leq N_-$, are characterized by

$$X = \Theta \operatorname{diag}(x_1, \ldots, x_N)\Theta' \tag{2.9a}$$

with

$$x_i = 0 \text{ or } x_i = \lambda_i,\ i = 1, \ldots, N \tag{2.9b}$$

and exactly p of the x_i are positive and q are negative.

Proof

Without loss of generality, we may assume that $A = \operatorname{diag}(\lambda_1, \ldots, \lambda_N)$ with $\lambda_1 \leq \ldots \leq \lambda_N$. A straightforward computation shows that the derivative of $f_A: S(n,N) \to \mathbb{R}$ at X is the linear map $Df_A(X): T_X S(n,N) \to \mathbb{R}$ defined by

$$\begin{aligned} Df_A(X)(\Delta X + X\Delta') &= -2\operatorname{tr}(A - X)(\Delta X + X\Delta') \\ &= -4\operatorname{tr} X(A - X)\Delta \end{aligned} \tag{2.10}$$

for all $\Delta \in \mathbb{R}^{N \times N}$. Therefore $X \in S(n,N)$ is a critical point of f_A and only if $X^2 = AX$. This proves (i). Now assume that $\lambda_1 < \ldots < \lambda_N$. From $X^2 = AX$ we have $AX = XA$ by symmetry of X. Since A has distinct eigenvalues, X must be diagonal. Thus the critical points of f_A are the diagonal matrices $X = \operatorname{diag}(x_1, \ldots, x_N)$ with $(A - X)X = 0$ and therefore $x_i = 0$ or $x_i = \lambda_i$ for $i = 1, \ldots, N$. $X \in S(p,q;N)$ if and only if exactly p of the x_i are positive and q are negative. Consequently, there are

$$\binom{N_+}{p}\binom{N_-}{q}$$

critical points of f_A in $S(p,q;N)$, characterized by (2.9). This completes the proof.

□

Theorem 2.2 has the following immediate consequence.

Corollary 2.3

Let $A = \Theta \operatorname{diag}(\lambda_1, \ldots, \lambda_N)\Theta'$ with $\Theta \in O(N)$ a real orthogonal $N \times N$ matrix and $\lambda_1 \geq \ldots \geq \lambda_N$. A minimum $\hat{X} \in S(p,q;N)$ for $f_A: S(p,q;N) \to \mathbf{R}$ exists if and only if $p \leq N_+$ and $q \leq N_-$. One such minimizing $\hat{X} \in S(p,q;N)$ is given by

$$\hat{X} = \Theta \operatorname{diag}(\lambda_1, \ldots, \lambda_p, 0, \ldots, 0, \lambda_{N-q+1}, \ldots, \lambda_N)\Theta' \tag{2.11}$$

and the minimum value of $f_A: S(p,q;N) \to \mathbf{R}$ is $\sum_{i=p+1}^{N-q} \lambda_i^2$. $\hat{X} \in S(p,q;N)$, given by (2.11) is the unique minimum of $f_A: S(p,q;N) \to \mathbf{R}$ if $\lambda_p > \lambda_{p+1}$ and $\lambda_{N-q} > \lambda_{N-q+1}$.

□

Example

Let $N = 2$, $n = 1$ and $X = \begin{bmatrix} a & b \\ b & c \end{bmatrix}$. The variety of 2×2 real symmetric matrices of rank ≤ 1 is a cone $\{(a,b,c) | ac = b^2\}$. Let A be 2×2 real symmetric nonsingular with distinct eigenvalues.

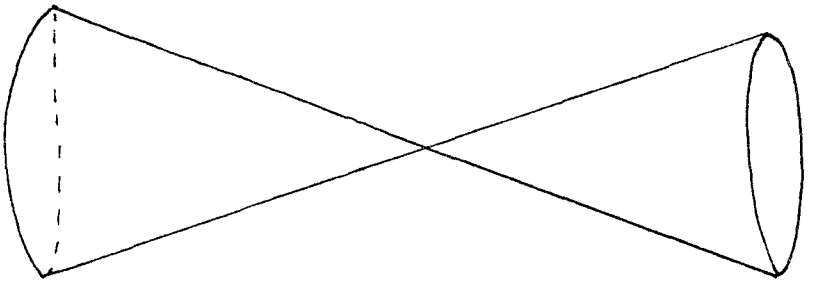


Figure: Real symmetric matrices of rank ≤ 1.

The function $f_A: S(1,2) \to \mathbf{R}$ has 2 local minima, if signature $A = 0$ and 1 local = global minimum, if $A > 0$ or $A < 0$. No other critical points exist. □

The next result shows that a best symmetric approximant of A of rank $\leq n$ necessarily has rank n. Recall that the singular values of $A \in \mathbf{R}^{N \times N}$ are the nonnegative square roots of the eigenvalues of AA'.

Proposition 2.4

Let $A \in \mathbf{R}^{N \times N}$ be symmetric with singular values $\sigma_1 \geq \ldots \geq \sigma_N$ and let $n \in \mathbf{N}$ be any integer with $n <$ rank $A = r$. There exists $\hat{X} \in S(n,N)$ which minimizes $f_A: S(n,N) \to \mathbf{R}$. If $\hat{X} \in \mathbf{R}^{N \times N}$ is any symmetric matrix which satisfies

$$\text{rank } \hat{X} \leq n \tag{2.12a}$$

$$\|A - \hat{X}\| = \inf\{\|A - X\| \mid X \in S(i,N), 0 \leq i \leq n\} \tag{2.12b}$$

then rank $\hat{X} = n$. In particular, the minimum value

$$\mu_n(A) = \min\{f_A(X) \mid X \in S(n,N)\} \tag{2.13}$$

coincides with the minimum value $\min\{f_A(X) | X \in S(i,N), 0 \leq i \leq n\}$. One has

$$\|A\|^2 = \mu_0(A) > \mu_1(A) > \ldots > \mu_r(A) = 0, \ r = \text{ rank } A \tag{2.14}$$

and for $0 \leq n \leq r$

$$\mu_n(A) = \sum_{i=n+1}^{N} \sigma_i^2 \tag{2.15}$$

Proof
By Proposition 2.1

$$\overline{S(n,N)} = \bigcup_{i=0}^{n} S(i,N)$$

is a closed subset of $\mathbf{R}^{N\times N}$. Since $f_A: \mathbf{R}^{N\times N} \to \mathbf{R}$ is proper, this implies that the restriction $f_A: \overline{S(n,N)} \to \mathbf{R}$ is proper. Since any proper continuous function $f: X \to Y$ has a closed image $f(X)$ it follows that a minimizing $\hat{X} \in \overline{S(n,N)}$ for $f_A: \overline{S(n,N)} \to \mathbf{R}$ exists:

$$\|A - \hat{X}\|^2 = \min\{f_A(X) \mid X \in \overline{S(n,N)}\}.$$

Suppose rank $\hat{X} < n$. Then for $\epsilon \in \mathbf{R}$ and $b \in \mathbf{R}^N$, $\|b\| = 1$, arbitrary we have $rk(\hat{X} + \epsilon bb') \leq n$ and thus

$$\begin{aligned}\|A - \hat{X} - \epsilon bb'\|^2 &= \|A - \hat{X}\|^2 - 2\epsilon \text{tr}\,(A - \hat{X})bb' + \epsilon^2 \\ &\geq \|A - \hat{X}\|^2.\end{aligned}$$

Thus for all $\epsilon \in \mathbf{R}$ and $b \in \mathbf{R}^N$, $\|b\| = 1$,

$$\epsilon^2 \geq 2\epsilon \,\text{tr}\,(A - \hat{X})bb'$$

and therefore $tr(A - \hat{X})bb' = 0$ for all $b \in \mathbf{R}^N$. Hence $A = \hat{X}$, contrary to assumption. Therefore $\hat{X} \in S(n,N)$ and $\mu_0(A) > \ldots > \mu_r(A) = 0$ for $r =$ rank A. (2.15) follows immediately from Corollary 2.3. □

Under a genericity assumption on A, the best symmetric rank n approximant of a symmetric matrix $A \in \mathbf{R}^{N\times N}$ in the Frobenius norm is uniquely determined. To

what extent is this also true for other critical points, such as, e.g. local minima, of $f_A: S(n,N) \to \mathbf{R}$? To answer this question we will now show that - under a suitable genericity condition on A - each critical point of $f_A: S(n,N) \to \mathbf{R}$ is nondegenerate and we will compute the *index* of each critical point, i.e. the dimension of the largest subspace on which the Hessian is negative definite.

Fix p, q and let $Q = \text{diag}\,(I_p, -I_q, 0) \in S(p,q;N)$. Define

$$\eta: \text{GL}\,(N,\mathbf{R}) \to S(p,q;N),\ \eta(\gamma) = \gamma' Q \gamma.$$

Consider the function

$$g_A: \text{GL}_+(N,\mathbf{R}) \to \mathbf{R},\ g_A = f_A \circ \eta.$$

If γ is a critical point of g_A, then a simple calculation shows that

$$g_A(\gamma+\Delta) - g_A(\gamma) = 2\text{tr}\,\{(\gamma^T Q\gamma - A)\Delta^T Q\Delta + \gamma^T Q\Delta\Delta^T Q\gamma + (\gamma^T Q\Delta)^2\} + O(\Delta^3).$$

Thus, the Hessian of g_A at γ is given by the quadratic form

$$D^2 g_A(\gamma)[\Delta] = 2\text{tr}\,\{(\gamma' Q\gamma - A)\Delta' Q\Delta + \gamma' Q\Delta\Delta' Q\gamma + (\gamma' Q\Delta)^2\}. \tag{2.16}$$

We make the generic assumption that A has distinct eigenvalues. Without loss of generality we may assume A is diagonal. Let $b_1^2 < \ldots < b_{N_+}^2$ denote the positive eigenvalues and let $-c_1^2 > \ldots > -c_{N_-}^2$ denote the negative eigenvalues. (We take $b_i > 0,\ c_i > 0,\ \forall_i$.) A may or may not have 0 as an eigenvalue. Let X be a critical point of f_A in $S(p,q;N)$. Let μ be a permuation of $\{1,\ldots,N_+\}$ such that $\mu_1 < \ldots < \mu_p$ and $\mu_{p+1} < \ldots < \mu_{N_+}$. Let ν be a permutation of $\{1,\ldots,N_-\}$ such that $\nu_1 < \ldots < \nu_q$ and $\nu_{q+1} < \ldots < \nu_{N_-}$. Without loss of generality, we may assume that for appropriate μ, ν

$$X = \text{diag}\,(b_{\mu_1}^2, \ldots, b_{\mu_p}^2, -c_{\nu_1}^2, \ldots, -c_{\nu_q}^2, 0, \ldots, 0)$$

and either

$$A = \text{diag}\,(b_{\mu_1}^2, \ldots, b_{\mu_p}^2, -c_{\nu_1}^2, \ldots, -c_{\nu_q}^2, b_{\mu_{p+1}}^2, \ldots, b_{\mu_{N_+}}^2, -c_{\nu_{q+1}}^2, \ldots, -c_{\nu_{N_-}}^2)$$

or

$$A = \text{diag}\,(b_{\mu_1}^2, \ldots, b_{\mu_p}^2, -c_{\nu_1}^2, \ldots, -c_{\nu_q}^2, b_{\mu_{p+1}}^2, \ldots, b_{\mu_{N_+}}^2, -c_{\nu_{q+1}}^2, \ldots, -c_{\nu_{N_-}}^2, 0)$$

depending on whether or not A has a 0 eigenvalue. Set

$$\gamma = \mathrm{diag}\,(b_{\mu_1}, \ldots, b_{\mu_p}, c_{\nu_1}, \ldots, c_{\nu_q}, 1, \ldots, 1).$$

Then γ is a critical point of g_A and $\eta(\gamma) = X$.

Straightforward calculations show that

$$\begin{aligned}\mathrm{tr}\,\{(\gamma' Q \gamma - A)\Delta' Q \Delta\} = &\sum_{j=n+1}^{N_+ + q} -b_{\mu_{j-q}}^2 \{\sum_{i=1}^{p} \delta_{ij}^2 - \sum_{i=p+1}^{n} \delta_{ij}^2\} \\ &+ \sum_{j=N_+ + q+1}^{N_+ + N_-} c_{\nu_{j-N_+}}^2 \{\sum_{i=1}^{p} \delta_{ij}^2 - \sum_{i=p+1}^{n} \delta_{ij}^2\}\end{aligned} \tag{2.17}$$

$$\mathrm{tr}\,\{\gamma' Q \Delta \Delta' Q \gamma\} = \sum_{i=1}^{p} b_{\mu_i}^2 \sum_{j=1}^{N} \delta_{ij}^2 + \sum_{i=p+1}^{n} c_{\nu_{i-p}}^2 \sum_{j=1}^{N} \delta_{ij}^2 \tag{2.18}$$

$$\begin{aligned}\mathrm{tr}\,(\gamma' Q \Delta)^2 = &\sum_{i=1}^{p} \sum_{j=1}^{p} b_{\mu_i} b_{\mu_j} \delta_{ij} \delta_{ji} + \sum_{i=p+1}^{n} \sum_{j=p+1}^{n} c_{\nu_{i-p}} c_{\nu_{j-p}} \delta_{ij} \delta_{ji} \\ &- \sum_{i=1}^{p} \sum_{j=p+1}^{n} b_{\mu_i} c_{\nu_{j-p}} \delta_{ij} \delta_{ji} - \sum_{i=p+1}^{n} \sum_{j=1}^{p} c_{\nu_{i-p}} b_{\mu_j} \delta_{ij} \delta_{ji}.\end{aligned} \tag{2.19}$$

From (2.16) - (2.19), it follows that

$$(1/2) D^2 g_A(\gamma)[\Delta] = \sum_{i=1}^{n} \sum_{j=1}^{N} \alpha_{ij} \delta_{ij}^2 + \sum_{1 \le i < j \le n} \beta_{ij} \delta_{ij} \delta_{ji} \tag{2.20}$$

where the coefficients α_{ij}, β_{ij} are as follows:

(i) $\alpha_{ii} = 2b_{\mu_i}^2, \quad 1 \le i \le p.$
(ii) $\alpha_{ij} = b_{\mu_i}^2, \quad 1 \le i \le p, 1 \le j \le n, i \ne j.$
(iii) $\alpha_{ii} = 2c_{\nu_{i-p}}^2, \quad p+1 \le i \le n.$
(iv) $\alpha_{ij} = c_{\nu_{i-p}}^2, \quad p+1 \le i \le n, 1 \le j \le n, i \ne j.$
(v) $\alpha_{ij} = b_{\mu_i}^2 - b_{\mu_{j-q}}^2, \quad 1 \le i \le p, n+1 \le j \le N_+ + q.$
(vi) $\alpha_{ij} = b_{\mu_i}^2 + c_{\nu_{j-N_+}}^2 \quad 1 \le i \le p, N_+ + q + 1 \le j \le N_+ + N_-.$
(vii) $\alpha_{ij} = c_{\nu_{i-p}}^2 + b_{\mu_{j-q}}^2, \quad p+1 \le i \le n, n+1 \le j \le N_+ + q.$
(viii) $\alpha_{ij} = c_{\nu_{i-p}}^2 - c_{\nu_{j-N_+}}^2, \quad p+1 \le i \le n, N_+ + q + 1 \le j \le N_+ + N_-.$
(ix) $\alpha_{ij} = b_{\mu_i}^2, \quad 1 \le i \le p, N_+ + N_- + 1 \le j \le N.$
(x) $\alpha_{ij} = c_{\nu_{i-p}}^2, \quad p+1 \le i \le n, N_+ + N_- + 1 \le j \le N.$
(xi) $\beta_{ij} = 2b_{\mu_i}, b_{\mu_j}, \quad 1 \le i < j \le p.$
(xii) $\beta_{ij} = -2b_{\mu_i} c_{\nu_{j-p}}, \quad 1 \le i \le p, p+1 \le j \le n.$

(xiii) $\beta_{ij} = 2c_{\nu_i - p}c_{\nu_j - p}, \quad p+1 \leq i < j \leq n.$

Note that cases (ix) and (x) only occur when A has 0 as an eigenvalue.

Let ω denote the symmetric bilinear form on $T_\gamma(\mathrm{GL}(N,\mathbf{R}))$ determined by the Hessian quadratic form $D^2 g_A(\gamma)$ and let Θ denote the symmetric bilinear form on $T_X(S(p,q;N))$ determined by the Hessian quadratic form $D^2 f_A(X)$. Since $X = \eta(\gamma)$ is a critical point of f_A, it follows that $\omega = \eta^* \Theta$ - i.e.,

$$\omega(v_1, v_2) = \Theta(D\eta(\gamma)[v_1], D\eta(\gamma)[v_2]), \quad \forall v_1, v_2 \in T_\gamma(\mathrm{GL}(N,\mathbf{R})). \tag{2.21}$$

Let us compute the rank of ω. It follows from (2.20) that the matrix representation for ω relative to the standard basis $\{E_{ij} | 1 \leq i,j \leq N\}$ for $T_\gamma(\mathrm{GL}(N,\mathbf{R}))$ is diagonal except for principal 2×2 blocks corresponding to each pair of indices $(i,j),(j,i), 1 \leq i < j \leq n$. If $1 \leq i < j \leq p$, the 2×2 block is

$$\begin{pmatrix} b_{\mu_i}^2 & b_{\mu_i} b_{\mu_j} \\ b_{\mu_i} b_{\mu_j} & b_{\mu_j}^2 \end{pmatrix}. \tag{2.22}$$

If $p+1 \leq i < j \leq n$, the 2×2 block is

$$\begin{pmatrix} c_{\nu_i - p}^2 & c_{\nu_i - p} c_{\nu_j - p} \\ c_{\nu_i - p} c_{\nu_j - p} & c_{\nu_j - p}^2 \end{pmatrix}. \tag{2.23}$$

If $1 \leq i \leq p, p+1 \leq j \leq n$, the 2×2 block is

$$\begin{pmatrix} b_{\mu_i}^2 & -b_{\mu_i} c_{\nu_j - p} \\ -b_{\mu_i} c_{\nu_j - p} & c_{\nu_j - p}^2 \end{pmatrix}. \tag{2.24}$$

We note that there are $\binom{n}{2}$ 2×2 submatrices of the types (2.22), (2.23), (2.24) and each has one positive eigenvalue and one zero eigenvalue. Now consider the diagonal elements in the matrix representation for ω which are not contained in submatrices of the forms (2.22), (2.23), (2.24). In other words, consider the diagonal element (i,j) where we do not have $1 \leq i \neq j \leq n$. If $n+1 \leq i \leq N$, then the $(i,j)^{\text{th}}$ diagonal element is 0. Otherwise it is nonzero. Consequently the rank of ω is $\binom{n}{2} + (nN - 2\binom{n}{2}) = (1/2)n(2N - n + 1) = \dim S(p,q;N)$. Together with (2.21), we have

$$\text{rank } \Theta \geq \text{ rank } \omega = \dim S(p,q;N)$$

which shows that the rank of Θ is equal to the dimension of $S(p,q;N)$ - i.e., Θ is nondegenerate.

Next, we determine the signature of Θ. It follows from (2.21) that the number of positive (respectively, negative) eigenvalues of Θ is at least as great as the number of

positive (respectively, negative) eigenvalues of ω. Since ω and Θ have equal rank, this implies that they have the same number of positive (respectively, negative) eigenvalues. Thus, the signature of Θ can be computed by using (2.20) to count the number of negative eigenvalues of the matrix representation of ω. Only cases (v) and (viii) can contribute negative eigenvalues. In (v), $b_{\mu_i}^2 - b_{\mu_{j-q}}^2 < 0$ if and only if $\mu_i < \mu_{j-q}$. It follows that the number of negative eigenvalues of this type is equal to

$$\text{Card } \{(i,j) | \mu_i < \mu_j, 1 \leq i \leq p, p+1 \leq j \leq N_+\}. \tag{2.25}$$

Similarly, the number of negative eigenvalues of type (viii) is equal to

$$\text{Card } \{(i,j) \mid \nu_i < \nu_j, 1 \leq i \leq q, q+1 \leq j \leq N_-\}. \tag{2.26}$$

The number given in (2.25) (respectively, (2.26)) is the dimension of the cell associated with the permutation μ (respectively, ν) in the Bruhat decomposition of the Grassmann manifold $Grass(p, \mathbb{R}^{N_+})$ (respectively, $Grass(q, \mathbb{R}^{N_-})$). Here the Grassmann manifold $Grass(k, \mathbb{R}^n)$ denotes the compact manifold of all k-dimensional linear subspaces of $\mathbb{R}^n$. Since the number of Bruhat cells of a given dimension d is equal to the d^{th} mod 2 Betti number of the Grassmann manifold, we have the following result:

Theorem 2.5
Let $A \in \mathbb{R}^{N \times N}$ be symmetric with distinct eigenvalues.

(a) Every critical point $f_A: S(n,N) \to \mathbb{R}$ is nondegenerate.

(b) The index of the critical point $X \in S(p,q;N)$ associated with the permutations μ, ν ($\mu_1 < \ldots < \mu_p$ and $\mu_{p+1} < \ldots < \mu_{N_+}$, $\nu_1 < \ldots < \nu_q$ and $\nu_{q+1} < \ldots < \nu_{N_-}$) is given by

$$\begin{aligned} &\text{Card}\{(i,j) \mid \mu_i < \mu_j, 1 \leq i \leq p,\ p+1 \leq j \leq N_+\} \\ &+ \text{ Card } \{(i,j) \mid \nu_i < \nu_j, 1 \leq i \leq q,\ q+1 \leq j \leq N_-\}. \end{aligned}$$

(c) The number of critical points in $S(p,q;N)$ which have index d is equal to the d^{th} mod 2 Betti number of the product of Grassmann manifolds

$$Grass(p, \mathbb{R}^{N_+}) \times Grass(q, \mathbb{R}^{N_-}).$$

□

Remark 2.6
The submanifold of $S(p,q;N)$ consisting of those X for which $\text{Eig}^+(X) \subset \text{Eig}^+(A)$

and $\mathrm{Eig}^-(X) \subset \mathrm{Eig}^-(A)$ is a fiber bundle with base manifold $Grass(p, \mathbf{R}^{N_+}) \times Grass(q, \mathbf{R}^{N_-})$ and Euclidean space of dimension $\frac{1}{2}p(p+1) + \frac{1}{2}q(q+1)$ as fiber. All of the critical points of f_A in $S(p, q; N)$ are contained in this submanifold.

A local minimum of $f_A: S(n, N) \to \mathbf{R}$ is a critical point of index 0. By Theorem 2.5 the number of local minima of $f_A: S(n, N) \to \mathbf{R}$ is equal to

$$\mathrm{Card}\,\{(p, q) \mid 0 \leq p \leq N_+,\ 0 \leq q \leq N_-,\ p + q = n\}.$$

In particular, for $n \leq \min(N_+, N_-)$, the number of local minima of $f_A: S(n, N) \to \mathbf{R}$ is $n + 1$. If the nonzero singular values of A are distinct, exactly one of the local minima is the global minimum.

3. Skew-Symmetric Matrices

In this section we obtain analogous results for distance functions defined on the variety of skew-symmetric matrices of fixed rank. Let N, n be positive integers with $2n \leq N$. Let

$$R(2n, N) := \{X \in \mathbf{R}^{N \times N} \mid X' = -X,\ \mathrm{rank}\ X = 2n\}$$

denote the variety of skew-symmetric $N \times N$ matrices of rank $2n$. $\mathrm{GL}(N, \mathbf{R})$ acts transitively on $R(2n, N)$ by the congruence action $(\gamma, X) \mapsto \gamma' X \gamma$.

Let $L = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$ and let

$$Q = \mathrm{diag}\,(L, \ldots, L, 0, \ldots, 0) \in R(2n, N).$$

Define

$$\eta: \mathrm{GL}(N, \mathbf{R}) \longrightarrow R(2n, N), \quad \eta(\gamma) = \gamma' Q \gamma.$$

Then $D\eta(I)[\Delta] = \Delta' Q + Q\Delta$. A straightforward calculation shows that the rank of $D\eta(I)[\Delta]$ is $n(2N-2n-1)$. Consequently, the dimension of $R(2n, N)$ is $n(2N-2n-1)$.

Let A be a fixed $N \times N$ skew-symmetric matrix, and consider the distance function

$$\begin{aligned} f_A: R(2n, N) &\longrightarrow \mathbf{R} \\ X &\longmapsto \|A - X\|^2 \end{aligned}$$

where $\|X\|^2 = \mathrm{tr}\, XX'$ is the square of the Frobenius norm. Our next result describes the critical points of f_A.

Theorem 3.1
Let $A \in \mathbb{R}^{N\times N}$ be skew-symmetric with distinct eigenvalues. Let m be such that $N = 2m$ or $N = 2m+1$ depending on whether N is even or odd. Then f_A has exactly $\binom{m}{n}$ critical points.

Proof:
Without loss of generality, we may assume that either

$$A = \operatorname{diag}(b_1^2 L, \ldots, b_m^2 L) \quad \text{or} \quad A = \operatorname{diag}(b_1^2 L, \ldots, b_m^2 L, 0), \tag{3.1}$$

depending on whether N is even or odd. We take $0 < b_1 < \ldots < b_m$. Consider the function

$$g_A : \mathrm{GL}(N, \mathbb{R}) \longrightarrow \mathbb{R}, \quad g_A = f_A \circ \eta.$$

A straightforward calculation shows that

$$dg_A(\gamma)[\Delta] = 4\operatorname{tr}(-\gamma' Q \gamma \gamma' Q + A\gamma' Q)\Delta, \quad \Delta \in \mathbb{R}^{N\times N}. \tag{3.2}$$

Thus, γ is a critical point of g_A if and only if

$$\gamma' Q \gamma \gamma' Q = A\gamma' Q. \tag{3.3}$$

Since η is a submersion, $X = \eta(\gamma)$ is a critical point of f_A if and only if γ is a critical point of g_A. Thus X is a critical point of f_A if and only if $X^2 = AX$. Consequently, if X is a critical point of f_A, then $AX = XA$. Since A has distinct eigenvalues, X must be a polynomial in A. Thus, the critical points of f_A are the rank $2n$ matrices of the form $X = \operatorname{diag}(x_1 L, \ldots, x_m L)$ or $X = \operatorname{diag}(x_1 L, \ldots, x_m L, 0)$ (depending on whether N is even or odd) with $x_i = 0$ or $x_i = b_i^2$. Consequently, there are $\binom{m}{n}$ critical points of f_A. □

Now we will show that each critical point of f_A is nondegenerate and we will compute the index of each critical point. If γ is a critical point of g_A, then a simple calculation shows that the Hessian of g_A at γ is given by the quadratic form

$$D^2 g_A(\gamma)[\Delta] = 2\operatorname{tr}\{(-\gamma^* Q\gamma + A)\Delta^* Q\Delta - \gamma^* Q\Delta\Delta^* Q\gamma - (\gamma^* Q\Delta)^2\}, \tag{3.4}$$

where we use conjugate transpose in place of transpose to allow for future complexification.

We now compute a matrix representation for the Hessian. Let A have the form (3.1). We will initially assume that N is even. Later, we indicate the minor changes when N

is odd. From the proof of Theorem 3.1, each critical point X is uniquely determined by a permutation μ of $\{1,\ldots,m\}$ such that $\mu_1 < \ldots < \mu_n$ and $\mu_{n+1} < \ldots < \mu_m$. The critical point associated with μ is $X = \mathrm{diag}\,(x_1 L,\ldots,x_m L)$ with $x_{\mu_i} = b^2_{\mu_i}$ for $i = 1,\ldots,n$ and $x_{\mu_i} = 0$ for $i = n+1,\ldots,m$. If X (and hence μ) is fixed, we may reorder to obtain

$$\begin{aligned} A &= \mathrm{diag}\,(b^2_{\mu_1}L,\ldots,b^2_{\mu_n}L,b^2_{\mu_{n+1}}L,\ldots,b^2_{\mu_m}L) \\ X &= \mathrm{diag}\,(b^2_{\mu_1}L,\ldots,b^2_{\mu_n}L,0.\ldots,0). \end{aligned} \tag{3.5}$$

Let $\gamma = \mathrm{diag}\,(b_{\mu_1},b_{\mu_1},\ldots,b_{\mu_n},b_{\mu_n},1,\ldots,1)$. Then γ is a critical point of g_A and $\eta(\gamma) = X$.

To simplify the calculation of the Hessian, we will complexify. Let $w_j = e_j + ie_{j+1}$ where e_j is the j^{th} standard basis vector for $\mathbb{R}^N$. Let U denote the $N \times N$ unitary matrix

$$U = (1/\sqrt{2})[w_1, w_3, \ldots, w_{2m-1}, \bar{w}_1, \bar{w}_3, \ldots, \bar{w}_{2m-1}], \tag{3.6}$$

where overbar denotes complex conjugation. Let $\hat{A} = U^*AU, \hat{Q} = U^*QU, \hat{\gamma} = U^*\gamma U, \hat{\Delta} = U^*\Delta U$. Since the righthand side of (3.4) is unchanged if we replace A, Q, γ, Δ by $\hat{A}, \hat{Q}, \hat{\gamma}, \hat{\Delta}$, there is no loss of generality is assuming that A, Q, γ, Δ are in fact equal to $\hat{A}, \hat{Q}, \hat{\gamma}, \hat{\Delta}$. Thus, we may assume that

$$\begin{aligned} A &= \mathrm{diag}\,(ib^2_{\mu_1},\ldots,ib^2_{\mu_m},-ib^2_{\mu_1},\ldots,-ib^2_{\mu_m}) \\ Q &= \mathrm{diag}\,(i,\ldots,i,0,\ldots,0,-i,\ldots,-i,0,\ldots,0) \\ \gamma &= \mathrm{diag}\,(b_{\mu_1},\ldots,b_{\mu_n},1,\ldots,1,b_{\mu_1},\ldots,b_{\mu_n},1,\ldots,1) \\ X &= \mathrm{diag}\,(ib^2_{\mu_1},\ldots,ib^2_{\mu_n},0,\ldots,0,-ib^2_{\mu_1},\ldots,-ib^2_{\mu_n},0,\ldots,0). \end{aligned} \tag{3.7}$$

Δ is now a complex matrix which has the partitioned form

$$\Delta = \begin{pmatrix} \Delta_{11} & \Delta_{12} \\ \bar{\Delta}_{12} & \bar{\Delta}_{11} \end{pmatrix}, \tag{3.8}$$

where each submatrix is $m \times m$. Note that Δ contains N^2 *real* parameters as before. Let $u_{kj} + iv_{kj}$ denote the kj^{th} entry of the matrix Δ.

Straightforward calculations show that

$$\mathrm{tr}\,\{(-\gamma^*Q\gamma + A)\Delta^*Q\Delta\} = 2\sum_{j=n+1}^{m} b^2_{\mu_j} \sum_{k=1}^{n} (u^2_{k,m+j} + v^2_{k,m+j} - u^2_{kj} - v^2_{kj}) \tag{3.9}$$

$$\mathrm{tr}\,\{-\gamma^*Q\Delta\Delta^*Q\gamma\} = 2\sum_{k=1}^{n} b^2_{\mu_k} \sum_{j=1}^{N} (u^2_{kj} + v^2_{kj}) \tag{3.10}$$

$$\text{tr} - (\gamma^* Q\Delta)^2 = 2\sum_{j=1}^{n}\sum_{k=1}^{n} b_{\mu_j} b_{\mu_k}(u_{jk}u_{kj} - v_{jk}v_{kj} - u_{j,m+k}u_{k,m+j} - v_{j,m+k}v_{k,m+j}). \tag{3.11}$$

We now indicate what changes are required when N is odd - i.e., $N = 2m + 1$. Modify U in (3.6) by adding "0" as the $(2m+1)^{\text{th}}$ entry in each of the $2m$ columns and appending e_{2m+1} as the $(2m+1)^{\text{th}}$ column. Append "0" as the $(2m+1)^{\text{th}}$ diagonal entry in the expressions for A, Q, X and append "1" as the $(2m+1)^{\text{th}}$ diagonal entry in the expression for γ in (3.7). The matrix Δ now has the partitioned form

$$\Delta = \begin{pmatrix} \Delta_{11} & \Delta_{12} & \Delta_{13} \\ \bar{\Delta}_{12} & \bar{\Delta}_{11} & \bar{\Delta}_{13} \\ \Delta_{31} & \bar{\Delta}_{31} & \Delta_{33} \end{pmatrix} \tag{3.8}'$$

corresponding to the partition $(m, m, 1)$ of N. In (3.8)', the scalar Δ_{33} is *real*. Thus, Δ contains N^2 real parameters. Let $u_{kj} + iv_{kj}$ denote the kj^{th} entry of the matrix Δ and note that $v_{2m+1,2m+1} = 0$. It is easily verified that the expressions (3.9), (3.10) and (3.11) are unchanged. However, note that the upper summation limit N in (3.10) is now $2m + 1$ instead of $2m$ so the Hessian contains $2n$ additional terms.

From (3.9) - (3.11), it follows that

$$\begin{aligned}(1/2)D^2 g_A(\gamma)[\Delta] = &\sum_{k=1}^{n}\sum_{j=1}^{N}(\alpha_{kj}u_{kj}^2 + \beta_{kj}v_{kj}^2) \\ &+ 4\sum_{1\le k<j\le n} b_{\mu_k}b_{\mu_j}(u_{kj}u_{jk} - u_{k.m+j}u_{j,m+k} - v_{kj}v_{jk} - v_{k,m+j}v_{j,m+k})\end{aligned} \tag{3.12}$$

where the coefficients α_{kj}, β_{kj} are as follows:

(i) $\alpha_{kk} = 4b_{\mu_k}^2, \quad \beta_{kk} = 0.$
(ii) $\alpha_{kj} = 2b_{\mu_k}^2 = \beta_{kj}, \quad 1 \le j \le n, j \ne k.$
(iii) $\alpha_{kj} = 2b_{\mu_k}^2 - 2b_{\mu_j}^2 = \beta_{kj}, \quad n+1 \le j \le m.$
(iv) $\alpha_{k,m+k} = 0 = \beta_{k,m+k}.$
(v) $\alpha_{kj} = 2b_{\mu_k}^2 = \beta_{kj}, \quad m+1 \le j \le m+n, j \ne m+k.$
(vi) $\alpha_{kj} = 2b_{\mu_k}^2 + 2b_{\mu_{j-m}}^2 = \beta_{kj}, \quad m+n+1 \le j \le 2m.$
(vii) $\alpha_{kj} = 2b_{\mu_k}^2 = \beta_{kj}, \quad 2m+1 \le j \le N.$

Note that case (vii) only occurs when N is odd.

Let ω denote the symmetric bilinear form on $T_\gamma(\text{GL}(N, \mathbb{R}))$ determined by the Hessian quadratic form $D^2 g_A(\gamma)$and let Θ denote the symmetric bilinear form on

$T_X(R(2n,N))$ determined by the Hessian quadratic form $D^2 f_A(X)$. To show the nondegeneracy of Θ, it suffices to show that the rank of ω is equal to the dimension of $R(2n,N)$, namely $n(2N-2n-1)$. This follows easily from (3.12). The number of negative eigenvalues of Θ is equal to the number of negative eigenvalues of ω. Using case (iii) of (3.12), it follows that the number of negative eigenvalues of ω is equal to

$$2 \text{ Card } \{(k,j) \mid \mu_k < \mu_j,\ 1 \le k \le n,\ n+1 \le j \le m\}. \tag{3.13}$$

The number given in (3.13) is the (real) dimension of the cell associated with the permutation μ in the Bruhat decomposition of the Grassmann manifold $Grass(n, \mathbb{C}^m)$ of n-dimensional complex linear subspaces of $\mathbb{C}^m$. Since the number of Bruhat cells of a given dimension $2d$ is equal to the $(2d)^{\text{th}}$ Betti number of the Grassmann manifold, we have the following result:

Theorem 3.2
Let $A \in \mathbb{R}^{N\times N}$ be skew-symmetric with distinct eigenvalues.

(a) Every critical point of $f_A\colon R(2n,N) \to \mathbb{R}$ is nondegenerate.

(b) The index of the critical point X associated with the permutation μ ($\mu_1 < \ldots < \mu_n$ and $\mu_{n+1} < \ldots < \mu_m$) is given by

$$2 \text{ Card } \{(k,j) \mid \mu_k < \mu_j,\ 1 \le k \le n,\ n+1 \le j \le m\}.$$

(c) The number of critical points which have index $2d$ is equal to the $(2d)^{\text{th}}$ Betti number of the Grassmann manifold $Grass(n, \mathbb{C}^m)$.

Remark 3.3
The submanifold of $R(2n,N)$ consisting of those X for which $\text{Eig}^+(iX) \subset \text{Eig}^+(iA)$ is a fiber bundle with base manifold $Grass(n, \mathbb{C}^m)$ and Euclidean space of dimension n^2 as fiber. All of the critical points of f_A are contained in this submanifold.

The above theorem implies that, for A skew-symmetric with distinct eigenvalues, there is *a unique local and hence global minimum* of $f_A\colon R(2n,N) \to \mathbb{R}$. Moreover, every critical point has an even index.

4. Approximations by Rectangular Matrices

Here we address the related issue of approximating a real rectangular matrix by matrices of lower rank.

For integers $1 \leq n \leq \min(M, N)$ let

$$M(n, M \times N) = \{X \in \mathbb{R}^{M \times N} \mid \text{rank } X = n\} \tag{4.1}$$

denote the set of real $M \times N$ matrices of rank n. Given $A \in \mathbb{R}^{M \times N}$ the approximation task is to find the minima and, more generally, the critical points of the distance function

$$\begin{aligned} F_A &: M(n, M \times N) \longrightarrow \mathbb{R} \\ F_A(X) &= \|A - X\|^2 \end{aligned} \tag{4.2}$$

for the Frobenius norm $\|X\|^2 = \text{tr}(XX')$ on $\mathbb{R}^{M \times N}$.

Proposition 4.1
$M(n, M \times N)$ is a smooth and connected manifold of dimension $n(M + N - n)$. The tangent space of $M(n, M \times N)$ at an element X is

$$T_X M(n, M \times N) = \{\Delta_1 X + X\Delta_2 \mid \Delta_1 \in \mathbb{R}^{M \times M}, \Delta_2 \in \mathbb{R}^{N \times N}\} \tag{4.3}$$

Proof
Let $Q \in M(n, M \times N)$ be defined by

$$Q = \begin{bmatrix} I_n & 0 \\ 0 & 0 \end{bmatrix} \tag{4.4}$$

Since every $X \in M(n, M \times N)$ is congruent to Q by the congruence action $((\gamma_1, \gamma_2), X) \mapsto \gamma_1 X \gamma_2^{-1}$, $\gamma_1 \in \text{GL}(M, \mathbb{R})$, $\gamma_2 \in \text{GL}(N, \mathbb{R})$, $M(n, M \times N)$ is an orbit of this smooth Lie group action of $\text{GL}(M) \times \text{GL}(N)$ on $\mathbb{R}^{M \times N}$ and therefore a smooth manifold. Here γ_1, γ_2 may be choosen to have positive determinant. Thus $M(n, M \times N)$ is the image of the connected subset $\text{GL}^+(M) \times \text{GL}^+(N)$ of the continuous (and in fact smooth) map $\pi: \text{GL}(M) \times \text{GL}(N) \to \mathbb{R}^{M \times N}$, $\pi(\gamma_1, \gamma_2) = \gamma_1 Q \gamma_2^{-1}$, and hence also connected. The derivative of π at (γ_1, γ_2) is the linear map on the tangent space $T_{(\gamma_1, \gamma_2)}(\text{GL}(M) \times \text{GL}(N)) \cong \mathbb{R}^{M \times M} \times \mathbb{R}^{N \times N}$ defined by

$$D\pi(\gamma_1, \gamma_2)[(\Delta_1, \Delta_2)] = \Delta_1(\gamma_1 Q \gamma_2^{-1}) - (\gamma_1 Q \gamma_2^{-1})\Delta_2,$$

and $T_X M(n, M \times N)$ is the image of this map. Finally the dimension result follows from a simple parameter count, since any rank n $M \times N$ matrix

$$X = \begin{bmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{bmatrix}$$

with $X_{11} \in \mathbb{R}^{n\times n}$ invertible, $X_{21} \in \mathbb{R}^{(M-n)\times n}$, $X_{12} \in \mathbb{R}^{n\times(N-n)}$ is characterized by $X_{22} = X_{21}X_{11}^{-1}X_{12}$ and thus depends on $n^2 + n(M+N-2n^2) = n(M+N-n)$ independent parameters. This completes the proof. □

Employing a standard trick from linear algebra, the general approximation problem for rectangular matrices is reduced to the approximation problem for symmetric matrices. To this end, define for $A, X \in \mathbb{R}^{M\times N}$

$$\hat{A} = \begin{bmatrix} 0 & A \\ A' & 0 \end{bmatrix}, \quad \hat{X} = \begin{bmatrix} 0 & X \\ X' & 0 \end{bmatrix}. \tag{4.5}$$

Thus $\hat{A}$ and $\hat{X}$ are $(M+N)\times(M+N)$ symmetric matrices. If $A \in \mathbb{R}^{M\times N}$ has singular values $\sigma_1, \ldots, \sigma_k$, $k = \min(M,N)$, then the eigenvalues of $\hat{A}$ are $\pm\sigma_1, \ldots, \pm\sigma_k$, and possibly 0. By (4.5) we have a smooth injective embedding

$$\begin{aligned} M(n, M\times N) &\longrightarrow S(2n, M+N) \\ X &\longmapsto \hat{X} \end{aligned} \tag{4.6}$$

with $2\|A-X\|^2 = \|\hat{A}-\hat{X}\|^2$.

It is easy to check that, for $X \in M(n, M\times N)$, $\hat{X}$ is a critical point (or a minimum) for $f_{\hat{A}}: S(2n, M+N) \to \mathbb{R}$ if and only if X is a critical point (or a minimum) for $F_A: M(n, M\times N) \to \mathbb{R}$.

Thus the results of the previous sections all carry over to results on the function $F_A: M(n, M\times N) \to \mathbb{R}$. We thus have the following results

Theorem 4.2

Let $A = \Theta_1\Sigma\Theta_2$ be the singular value decomposition of $A \in \mathbb{R}^{M\times N}$ with singular values $\sigma_1 \geq \ldots \geq \sigma_k$, $1 \leq k \leq \min(M,N)$.

(i) The critical points of $F_A: M(n, M\times N) \to \mathbb{R}$, $F_A(X) = \|A-X\|^2$, are characterized by $(A-X)X' = 0$, $X'(X-A) = 0$.

(ii) $F_A: M(n, M\times N) \to \mathbb{R}$ has a finite number of critical points if and only if $M = N$ and A has M distinct singular values.

(iii) If $\sigma_n > \sigma_{n+1}$, $n \leq k$, there exists a unique global minimum $X_{\min}$ of $F_A: M(n, M\times N) \to \mathbb{R}$ which is given by

$$X_{\min} = \Theta_1\Sigma_n\Theta_2 \tag{4.7}$$

where Σ_n is obtained from Σ by setting $\sigma_{n+1}, \ldots, \sigma_k$ all to zero. □

As an immediate corollary we obtain the following classical theorem of Eckart and Young (1936).

Corollary 4.3

Let $A = \Theta_1 \Sigma \Theta_2$ be the singular value decomposition of $X \in \mathbb{R}^{M \times N}$, $\Theta_1 \Theta_1' = I$, $\Theta_2 \Theta_2' = I$, $\Sigma = \text{diag}\,(\sigma_1, \ldots, \sigma_k)$, $\sigma_1 \geq \ldots \geq \sigma_k \geq 0$. Let $n \leq k$ and $\sigma_n > \sigma_{n+1}$.

Then with $\Sigma_n = \text{diag}\,(\sigma_1, \ldots, \sigma_n, 0, \ldots, 0)$

$$X_{\text{min}} = \Theta_1 \Sigma_n \Theta_2$$

is the unique $M \times N$ matrix of rank n which minimizes $\|A - X\|^2$ over the set of rank $\leq n$ matrices. □

References

[1] G. Eckart and G. Young, The approximation of one matrix by another of lower rank, *Psychometrika* 1 (1936), 211 - 218

[2] G.H. Golub and C. Van Loan, An analysis of the total least squares problem, SIAM J. Numer. Anal. 17 (1980), 883 - 843

[3] G.H. Golub, A. Hoffmann and G.W. Stewart, A generalization of the Eckart-Young-Mirsky matrix approximation theorem, *Linear Algebra and its Appl.* 88/89 (1987), 317 - 327

[4] N.J. Higham, Computing a nearest symmetric positive semidefinite matrix, *Linear Algebra and its Appl.* 103 (1988), 103 - 118