

# Restoring Rank and Consistency by Orthogonal Projection

A. van der Sluis

*Institute of Mathematics*

*University of Utrecht*

*Utrecht-Uithof, The Netherlands*

and

G. W. Veltkamp

*Department of Mathematics*

*Technological University*

*Eindhoven, The Netherlands*

Dedicated to Alston S. Householder  
on the occasion of his seventy-fifth birthday.

Submitted by G. W. Stewart

---

## ABSTRACT

Let  $A_0x = b_0$  be a consistent (but possibly unknown) linear algebraic system of  $m$  equations in  $n$  unknowns, with  $\text{rank}(A_0) = k$  (likewise possibly unknown). Let  $Ax = b$  be a known (but possibly inconsistent) nearby system. Procedures for "solving"  $Ax = b$  usually replace it (at least in principle) by a nearby consistent system  $\tilde{A}x = \tilde{b}$  of (hopefully) rank  $k$ , and solve that one instead. We consider consistent systems  $\tilde{A}x = \tilde{b}$  with  $\text{rank}(\tilde{A}) = k$  such that  $(\tilde{A}, \tilde{b})$  is the orthogonal projection of  $(A, b)$  on  $\text{span}(\tilde{A})$  [i.e., the columns of  $(\tilde{A} \mid \tilde{b})$  are the projections of those of  $(A \mid b)$ ]. Under suitable circumstances the rank  $k$  pair  $(\tilde{A}, \tilde{b})$  nearest to  $(A, b)$  in the sense of minimizing  $\|( \tilde{A} \mid \tilde{b} ) - (A \mid b) \|_F$  belongs to this class, as well as the pair delivered by the ordinary least squares method (if  $k = n < m$ ), or, more generally, the pair delivered by a well-known algorithm (viz. HFTI). If  $\varepsilon := \|(A \mid b) - (A_0 \mid b_0)\|_F$ , then the minimum length solutions of all systems  $\tilde{A}x = \tilde{b}$  so related to  $(A, b)$  are shown to differ mutually only by  $O(\varepsilon^2)$ . This means, e.g., that it will usually not pay to compute the solution of the nearest system.

---

## 1. INTRODUCTION

### 1.1. Least Squares Solutions of Inconsistent Systems

If  $A$  and  $b$  are a given  $m \times n$  matrix and  $m$ -vector and the linear algebraic system  $Ax = b$  is inconsistent, then one often resorts to the corresponding *least squares problem*  $Ax \simeq b$ , i.e., one determines the vector  $x$

which minimizes the functional

$$x \mapsto \|Ax - b\|_2. \quad (1.1)$$

This vector is uniquely determined if  $\text{rank}(A) = n$ ; it is not if  $\text{rank}(A) < n$ , but then the one of minimum length is uniquely determined.

As is well known, the vector  $x$  then satisfies  $Ax = \tilde{b}$ , where  $\tilde{b}$  is the orthogonal projection of  $b$  on  $\text{span}(A)$ , the linear space spanned by the columns of  $A$ . Thus the least squares approach amounts to replacing the pair  $(A, b)$  by a nearby *consistent pair*  $(A, \tilde{b})$ , i.e. a pair for which  $\tilde{b} \in \text{span}(A)$ .

### 1.2. Uncertainties in $A$ and $b$

Although this least squares approach may be rather reasonable if the given pair  $(A, b)$  is inconsistent only due to uncertainties in the elements of  $b$ , it is much less so if the elements of  $A$ , too, are subject to uncertainty. In the latter case it would be more reasonable to replace  $(A, b)$  by a nearby consistent pair  $(\tilde{A}, \tilde{b})$  in which the matrix is adjusted as well, and take  $\tilde{x}$  such that  $\tilde{A}\tilde{x} = \tilde{b}$ .

The case for adjusting  $A$  as well as  $b$  is even stronger if for the given pair  $(A, b)$  we have

$$(A, b) = (A_0, b_0) + (\delta A, \delta b),$$

where  $(A_0, b_0)$  is a fixed consistent pair with  $\text{rank}(A_0) = k < n$  and  $(\delta A, \delta b)$  is due to noise in the data. Then the least squares solution of  $Ax \simeq b$  depends *discontinuously* on  $\delta A$  and  $\delta b$ , and the simplest sensible thing to do is to find a rank  $k$  matrix  $\tilde{A}$  near to  $A$  and solve the least squares problem  $\tilde{A}x \simeq b$ . This again amounts to replacing  $(A, b)$  by the consistent pair  $(\tilde{A}, \tilde{b})$ ,  $\tilde{b}$  the orthogonal projection of  $b$  on  $\text{span}(\tilde{A})$ .

This is indeed what several numerical procedures are aiming at: using some tolerance parameter  $\eta$  they determine the so-called numerical or pseudo rank of  $A$ , which, computing accuracy permitting, will equal  $k$  if  $\delta A$  is small enough and  $\eta$  is well chosen, and then deliver a nearby matrix  $\tilde{A}$  of this pseudo rank.

### 1.3. The Generalized Least Squares Problem

When one wishes to replace  $(A, b)$  by a nearby consistent pair  $(\tilde{A}, \tilde{b})$  with  $\text{rank}(\tilde{A}) = k$ , the idea naturally arises to take this pair nearest to  $(A, b)$  in some sense.

One obvious measure for the distance between  $(\tilde{A}, \tilde{b})$  and  $(A, b)$  is  $\|(\tilde{A} \begin{smallmatrix} \vdots \\ \tilde{b} \end{smallmatrix}) - (A \begin{smallmatrix} \vdots \\ b \end{smallmatrix})\|_F$ , where  $(A \begin{smallmatrix} \vdots \\ b \end{smallmatrix})$  denotes the  $m \times (n+1)$  matrix obtained by adding  $b$  as  $(n+1)$ st column to  $A$ , and  $\|\cdot\|_F$  denotes the Frobenius norm (cf. Sec. 3.1). Or, more generally, one might take  $\|(\tilde{A} \begin{smallmatrix} \vdots \\ \tilde{b} \end{smallmatrix}) - (A \begin{smallmatrix} \vdots \\ b \end{smallmatrix})\|_F T^{-1}$ , where  $T$  is an  $(n+1) \times (n+1)$  diagonal matrix whose  $i$ th diagonal element  $t_{ii}$  is related to the magnitude of the errors in the  $i$ th column of  $(A \begin{smallmatrix} \vdots \\ b \end{smallmatrix})$  (e.g.,  $t_{ii}^2$  equals the variance of the errors in the  $i$ th column).

The problem to find for given  $A, b, k$  and  $T$  the vector  $\hat{x}$  of minimum length satisfying  $\hat{A}\hat{x} = \hat{b}$ , where  $(\hat{A}, \hat{b})$  denotes the consistent pair with  $\text{rank}(\hat{A}) = k$  nearest to  $(A, b)$  in the given sense, is called the *generalized least squares problem*. It is an essentially nonlinear problem which actually amounts to solving (at least partially) a matrix eigenvalue problem (cf. e.g., [4] and [6, §29.25]).

#### 1.4. The Purpose of This Paper

In this paper we consider consistent pairs  $(\tilde{A}, \tilde{b})$  with  $\text{rank}(\tilde{A}) = k$  such that  $(\tilde{A}, \tilde{b})$  is the orthogonal projection of  $(A, b)$  on  $\text{span}(\tilde{A})$  [i.e., the columns of  $(\tilde{A} \begin{smallmatrix} \vdots \\ \tilde{b} \end{smallmatrix})$  are the projections of those of  $(A \begin{smallmatrix} \vdots \\ b \end{smallmatrix})$ ]. Such pairs are obtained e.g.

- (1) for  $k = n$  by the ordinary least squares method, provided that  $\text{rank}(A) = n$ , and then  $\tilde{A} = A$  (cf. Sec. 1.1);
- (2) for  $k \leq n$  by the HFTI algorithm in [7, p. 81], provided that  $A$  is close enough to a rank  $k$  matrix;
- (3) for  $k \leq n$  by the generalized least squares method (cf. Sec. 1.3), provided that  $(A, b)$  is close enough to a consistent pair  $(A_0, b_0)$  with  $\text{rank}(A_0) = k$ .

Now suppose that  $(A, b)$  is only slightly inconsistent, say only  $O(\epsilon)$  away from a fixed consistent pair  $(A_0, b_0)$  with  $\text{rank}(A_0) = k$ , and consider consistent pairs  $(\tilde{A}, \tilde{b})$  with  $\tilde{A} - A = O(\epsilon)$  and  $\text{rank}(\tilde{A}) = k$ , such that  $(\tilde{A}, \tilde{b})$  is the orthogonal projection of  $(A, b)$  on  $\text{span}(\tilde{A})$ . Then we show that the minimum length solutions  $\tilde{x}$  of all systems  $\tilde{A}\tilde{x} = \tilde{b}$  so related to  $(A, b)$  mutually differ only by  $O(\epsilon^2)$  (cf. Theorem 4.11). This means that in many cases it will not pay to find the true generalized least squares solution, but that the solution  $\tilde{x}$  of  $\tilde{A}\tilde{x} = \tilde{b}$ , where  $(\tilde{A}, \tilde{b})$  is obtained by a suitable projection, will already be adequate.

Moreover, if  $k = n$ , then the sole knowledge of such a solution  $\tilde{x}$  and the corresponding residue  $b - A\tilde{x}$  enables us to find a first order (in  $\epsilon$ ) approximation for  $(\hat{A}, \hat{b})$ , the consistent rank  $k$  pair nearest to  $(A, b)$ . If  $k < n$ , we can only approximate  $\hat{b}$  (cf. Corollary 5.12).

As an additional result, we have gained a factor  $\sqrt{2}$  in a well-known bound for the difference of the pseudoinverses of two matrices of equal rank (cf. Theorem 3.14).

2. INTUITIVE APPROACH FOR THE CASE  $\text{rank}(A) = n$ 

If  $\text{rank}(A) = n$ , it is rather simple to demonstrate in an intuitive geometric way how it comes about that the ordinary and generalized least squares solutions differ only by  $O(\epsilon^2)$ , and this is what we are going to do now.

Let us consider least squares problems  $Ax \simeq b$  with  $(A, b)$  in an  $\epsilon$ -neighborhood of a consistent pair  $(A_0, b_0)$  (i.e.  $\|(A \mid b) - (A_0 \mid b_0)\|_F \leq \epsilon$ ), with  $\text{rank}(A_0) = n$ . Let  $(\hat{A}, \hat{b})$  be the consistent pair nearest to  $(A, b)$  in the sense of Sec. 1.3 with  $T = I$  (for existence cf. Sec. 5). Then  $\|\hat{A} - A\|_F \leq \epsilon$ , and for  $\epsilon$  small enough we will have  $\text{rank}(\hat{A}) = n$ . Furthermore  $(\hat{A}, \hat{b})$  will be the orthogonal projection of  $(A, b)$  on  $\text{span}(\hat{A})$  because of the simple geometrical fact that otherwise this projection would be a consistent pair nearer to  $(A, b)$ . Thus  $\hat{A}\hat{x} = \hat{b}$  is the projection of  $b$  on  $\text{span}(\hat{A})$ , and  $\hat{A}x$  is the projection of  $Ax$  (cf. Fig. 1). Hence  $\hat{A}\hat{x} - \hat{A}x$  is the projection of the vector  $r = b - Ax$ .

Since  $A - \hat{A} = O(\epsilon)$  and  $r \perp \text{span}(A)$ , we expect  $\text{angle}(r, \text{span}(\hat{A})) = \pi/2 + O(\epsilon)$ , and hence the projection of  $r$  on  $\text{span}(\hat{A})$  has length  $O(\epsilon)\|r\|$ . Since  $r$  itself is  $O(\epsilon)$ , we have  $\hat{A}(\hat{x} - x) = O(\epsilon^2)$ . Since  $\text{glb}(\hat{A})$  is bounded away from 0 in the neighborhood of  $A_0$ , we have  $\hat{x} - x = O(\epsilon^2)$ , which is the desired property.

Perhaps the only part in this discussion which would require further clarification is the statement about  $\text{angle}(r, \text{span}(\hat{A}))$ . This could be resolved in a number of ways, e.g., by considering angles between subspaces, or by constructing suitable orthonormal bases which differ elementwise  $O(\epsilon)$ .

The simplest way, probably, is to use generalized inverses (for some properties cf. Sec. 3.3). Actually,  $\hat{A}\hat{A}^+$  is the orthogonal projector on  $\text{span}(\hat{A})$ , and  $AA^+$  is the orthogonal projector on  $\text{span}(A)$ . Since  $\hat{A}\hat{A}^+ - AA^+ = O(\epsilon)$  (cf. [11, Theorem 4.1]) and  $AA^+r = 0$ , we have  $\|\hat{A}\hat{A}^+r\| = O(\epsilon)\|r\|$ , which at once establishes the statement about the angle and about the length of the projection of  $r$  on  $\text{span}(\hat{A})$ .

Although this discussion was called "intuitive," it will be clear that it actually constitutes a proof of the  $O(\epsilon^2)$  property, though perhaps not a formally complete one.

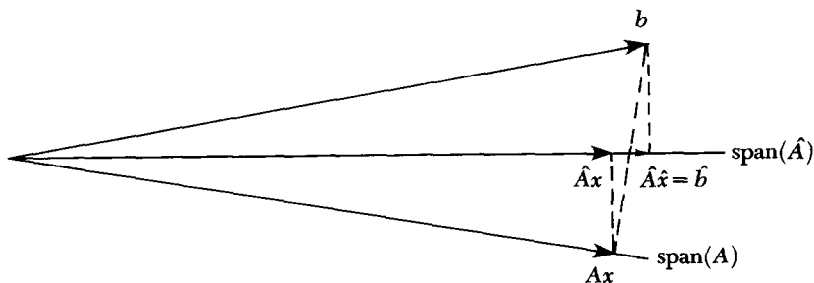


FIG. 1.

### 3. SOME DEFINITIONS AND PROPERTIES

#### 3.1. Some Definitions

By  $\|\cdot\|$  we will always designate the euclidean vector norm and its associated matrix norm.

By  $\|\cdot\|_F$  we designate the Frobenius matrix norm:  $\|(a_{ij})\|_F = (\sum |a_{ij}|^2)^{1/2}$ .

The singular values of an  $m \times n$  matrix  $P$  will be denoted in decreasing order by  $\sigma_1(P), \sigma_2(P), \dots$ . By  $\tilde{\sigma}_k(P)$  we shall denote  $[\sum_{i \geq k} \sigma_i^2(P)]^{1/2}$ . Then  $\|P\| = \sigma_1(P)$ ,  $\|P\|_F = \tilde{\sigma}_1(P)$ .

A matrix-vector pair  $(P, q)$  will be said to have rank  $k$  if the matrix  $(P \mid q)$  has rank  $k$ , and to be *consistent* if  $q \in \text{span}(P)$ . Hence  $(P, q)$  is a consistent rank  $k$  pair iff  $\text{rank}(P) = \text{rank}((P \mid q)) = k$ . Occasionally we shall identify a pair  $(P, q)$  with its matrix  $(P \mid q)$ .

#### 3.2. Near and Nearest Matrices of Given Rank

The following theorem is well known (cf. [3] and [8; Theorem 6.7]):

**THEOREM 3.1.** *Let  $P$  be a matrix of rank  $p$  whose nonzero singular values are  $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_p$ . Then*

(a) *For given  $q$ ,  $0 \leq q \leq p$ , there exists a matrix  $Q$  of rank  $q$  which minimizes the functional  $Q \mapsto \|Q - P\|_F$ ; for any such  $Q$  the properties (b), (c), (d), (e) hold:*

(b)  *$Q^H(Q - P) = 0$ ,  $(Q - P)Q^H = 0$ , or in geometrical terms: the columns of  $Q$  are the orthogonal projections of the columns of  $P$  on  $\text{span}(Q)$ , and likewise for the rows.*

(c)  *$\|Q - P\|_F = \tilde{\sigma}_{q+1}(P)$ .*

(d) *The nonzero singular values of  $Q$  are  $\sigma_1, \dots, \sigma_q$ .*

(e)  *$P$  and  $Q$  have a simultaneous singular value decomposition, i.e., there exist unitary matrices  $U$  and  $V$  such that*

$$P = U \cdot \text{diag}(\sigma_1, \dots, \sigma_p, 0, \dots) \cdot V^H, \quad Q = U \cdot \text{diag}(\sigma_1, \dots, \sigma_q, 0, \dots) \cdot V^H.$$

(f)  *$Q$  is uniquely determined iff  $q = 0$  or  $q = p$  or  $\sigma_q > \sigma_{q+1}$ .*

**THEOREM 3.2.** *Let  $P$  and  $Q$  be two matrices of the same dimensions. Then*

(a)  *$|\sigma_k(P) - \sigma_k(Q)| \leq \|P - Q\|$ ;*

(b)  *$|\tilde{\sigma}_k(P) - \tilde{\sigma}_k(Q)| \leq \|P - Q\|_F$ ;*

(c) *if  $\|P - Q\| < \sigma_k(P)$ , then  $\text{rank}(Q) \geq k$ ;*

(d) *if  $\|P - Q\|_F < \tilde{\sigma}_k(P)$ , then  $\text{rank}(Q) \geq k$ .*

*Proof.* For (a) cf. [7, Theorem (5.7)]. For (b) define for the moment  $s(P)$  as the vector  $(\sigma_k(P), \sigma_{k+1}(P), \dots)^T$ ,  $s(Q) := (\sigma_k(Q), \sigma_{k+1}(Q), \dots)^T$ . Then

$$\begin{aligned} |\tilde{\sigma}_k(P) - \tilde{\sigma}_k(Q)| &= ||s(P) - s(Q)|| \leq \|s(P) - s(Q)\| \\ &= \left\{ \sum_{i \geq k} [\sigma_i(P) - \sigma_i(Q)]^2 \right\}^{1/2} \leq \|P - Q\|_F \end{aligned}$$

(for the latter inequality cf. [7, Theorem (5.10)]).

Properties (c) and (d) follow right away from (a) and (b), respectively. ■

### 3.3. Generalized Inverses

For any  $m \times n$  matrix  $P$  we denote by  $P^+$  its generalized inverse (cf., e.g., [2]) and list a few of its properties which we are going to use:

PROPERTY 3.3.  $PP^+P = P$ ,  $P^+PP^+ = P^+$ .

PROPERTY 3.4.  $(PP^+)^H = PP^+$ ,  $(P^+P)^H = P^+P$ ,  $(P^H)^+ = (P^+)^H =: P^{+H}$  for short.

PROPERTY 3.5.  $\text{span}(P^+) = \text{span}(P^H)$ ,  $\ker(P^+) = \ker(P^H)$ .

PROPERTY 3.6.  $PP^+$  = orthogonal projector on  $\text{span}(P)$ .

PROPERTY 3.7.  $P^+P$  = orthogonal projector on  $\text{span}(P^H)$ .

PROPERTY 3.8.  $\|PP^+\|, \|P^+P\|, \|I - PP^+\|, \|I - P^+P\| \leq 1$ .

PROPERTY 3.9.  $P^+P = I$  iff  $\text{rank}(P) = n$ .

PROPERTY 3.10. For any vector  $q$ ,  $P^+q$  is the minimum length least squares solution of  $Px \simeq q$ ; denoting this solution by  $x$ , then  $P^+Px = x$ .

PROPERTY 3.11. If  $\text{rank}(P) = k$ , then  $\|P^+\| = \sigma_k(P)^{-1}$ .

PROPERTY 3.12. If  $\|P^+\| \|Q - P\| < 1$ , then  $\text{rank}(Q) \geq \text{rank}(P)$ .

PROPERTY 3.13. If  $\text{rank}(Q) \leq \text{rank}(P)$  and  $\|P^+\| \|Q - P\| < 1$ , then  $\text{rank}(Q) = \text{rank}(P)$  and  $\|Q^+\| \leq \|P^+\| / (1 - \|P^+\| \|Q - P\|)$ .

### 3.4. A Perturbation Theorem

Let  $P$  and  $Q$  be  $m \times n$ -matrices with the same rank. Then  $P^+$  and  $Q^+$  are near when  $P$  and  $Q$  are near. We want to estimate  $\|Q^+ - P^+\|_F$  in terms of  $\|Q - P\|_F$ . Such an estimate is contained in Wedin [11] (and quoted by Stewart [10]), along with results for more general norms.

Using a slightly different analysis of this problem (more symmetric in rows and columns), we found we could gain a factor  $\sqrt{2}$  and get a best possible bound. This is shown below. We note, however, that the same result could also be obtained by just a slight change in Wedin's proof, viz. by replacing, in his formula (4.15), the factor  $\|T\|_E$  by  $\|TA^+A\|_E$ , which is allowed, considering where this factor comes from.

**THEOREM 3.14.** *Let  $P$  and  $Q$  have the same dimensions and rank. Then*

$$\|Q^+ - P^+\|_F \leq \|P^+\| \|Q^+\| \|Q - P\|_F. \quad (3.1)$$

*Proof.* Let

$$E_{11} := QQ^+(Q - P)P^+P = -Q(Q^+ - P^+)P,$$

$$E_{12} := QQ^+(Q - P)(I - P^+P) = Q(I - P^+P),$$

$$E_{21} := (I - QQ^+)(Q - P)P^+P = -(I - QQ^+)P,$$

$$E_{22} := (I - QQ^+)(Q - P)(I - P^+P) = 0.$$

Then, using repeatedly that  $\|U + V\|_F^2 = \|U\|_F^2 + \|V\|_F^2$  if  $U^H V = 0$  or  $VU^H = 0$ , we have

$$Q - P = \sum_{i,j} E_{ij}, \quad \|Q - P\|_F^2 = \sum_{i,j} \|E_{ij}\|_F^2. \quad (3.2)$$

Similarly, let

$$F_{11} := Q^+Q(Q^+ - P^+)PP^+ = -Q^+(Q - P)P^+,$$

$$F_{12} := Q^+Q(Q^+ - P^+)(I - PP^+) = Q^+(I - PP^+),$$

$$F_{21} := (I - Q^+Q)(Q^+ - P^+)PP^+ = -(I - Q^+Q)P^+,$$

$$F_{22} := (I - Q^+Q)(Q^+ - P^+)(I - PP^+) = 0.$$

Then

$$Q^+ - P^+ = \sum_{i,j} F_{ij}, \quad \|Q^+ - P^+\|_F^2 = \sum_{i,j} \|F_{ij}\|_F^2. \quad (3.3)$$

We now estimate the  $F_{ij}$  in terms of the  $E_{ij}$ . Since  $F_{11} = -Q^+ E_{11} P^+$ , we have (using that  $\|UV\|_F \leq \|U\| \|V\|_F$ )

$$\|F_{11}\|_F \leq \|P^+\| \|Q^+\| \|E_{11}\|_F.$$

Using the fact that for orthogonal projectors  $U$  and  $V$  of the same rank there holds  $\|(I-U)V\|_F = \|U(I-V)\|_F$  (cf. [11, Lemma 4.1]<sup>1</sup>), we have

$$\|QF_{12}\|_F = \|QQ^+(I-PP^+)\|_F = \|(I-QQ^+)PP^+\|_F = \|E_{21}P^+\|_F,$$

$$\|F_{21}P\|_F = \|(I-Q^+Q)P^+P\|_F = \|Q^+Q(I-P^+P)\|_F = \|Q^+E_{12}\|_F;$$

hence

$$\|F_{12}\|_F = \|Q^+QF_{12}\|_F \leq \|P^+\| \|Q^+\| \|E_{21}\|_F,$$

$$\|F_{21}\|_F = \|F_{21}PP^+\|_F \leq \|P^+\| \|Q^+\| \|E_{12}\|_F.$$

Using these estimates, (3.1) follows from (3.3) and (3.2). ■

**REMARK 3.15.** Simple examples like

$$P = \begin{pmatrix} \alpha & 0 \\ 0 & 0 \end{pmatrix} \quad \text{and} \quad Q = \begin{pmatrix} 0 & 0 \\ 0 & \beta \end{pmatrix}$$

show that the bound (3.1) is best possible and involves  $\|P^+\|$  and  $\|Q^+\|$  in the correct combination.

**REMARK 3.16.** The estimate for  $\|Q^+\|$  in Property 3.13 enables us to express the bound in (3.1) in terms of  $\|P^+\|$  and  $\|P-Q\|_F$  only.

<sup>1</sup>In [11] it is stated that  $\|(I-U)V\| = \|U(I-V)\|$  for arbitrary unitarily invariant norms. For the  $F$ -norm the result is contained in Afriat [1]; its proof then simply consists in expressing the  $F$ -norms in terms of traces, and using some properties of traces.



## 4. ENFORCING CONSISTENCY BY ORTHOGONAL PROJECTION

## 4.1. Definition and Examples

Let  $(A, b)$  be a pair with dimensions  $m \times (n, 1)$ . Let  $E \in \mathbb{C}^{m \times m}$  be an orthogonal projector onto a  $k$ -dimensional subspace  $V \subset \mathbb{C}^m$ . Then we shall say that the pair  $(\tilde{A}, \tilde{b}) := E(A, b)$  is obtained from  $(A, b)$  by orthogonal projection.

Obviously  $(\tilde{A}, \tilde{b})$  has rank  $\leq k$ , but it is not necessarily consistent. If  $\text{rank}(\tilde{A}) = k$ , however, the pair is consistent, and  $\text{span}(\tilde{A}) = V = \text{span}(E)$ . Moreover, since then  $E = \tilde{A}\tilde{A}^+$ , we have  $\tilde{b} = \tilde{A}\tilde{A}^+b$ ; hence  $\tilde{A}^+\tilde{b} = \tilde{A}^+b$  (cf. Property 3.3), and the solution set of the consistent system  $\tilde{A}x = \tilde{b}$  coincides with the set of least squares solutions of  $\tilde{A}x \simeq b$ .

Projections leading to consistent rank  $k$  pairs can be characterized in the following way.

**THEOREM 4.1.** *Let  $(A, b)$  and  $(\tilde{A}, \tilde{b})$  be pairs with the same dimensions. Then the following statements are equivalent:*

- (1) *There exists an orthogonal projector  $E$  with rank  $k$  such that*
  - (a)  $\text{span}(E) \cap (\text{span}(A))^\perp = \{0\}$ ;
  - (b)  $(\tilde{A}, \tilde{b}) = E(A, b)$
- (2) *(a)  $(\tilde{A}, \tilde{b})$  is a consistent rank  $k$  pair;*
  - (b)  $\tilde{A}^H(A - \tilde{A}, b - \tilde{b}) = 0$ .

*Proof.* We first observe that, if  $E$  is an orthogonal projector such that  $\tilde{A} = EA$ , then

$$\begin{aligned} (1a) &\Leftrightarrow \text{rank}(A^HE) = \text{rank}(E) \Leftrightarrow \text{rank}(\tilde{A}) = \text{rank}(E) \\ &\Leftrightarrow \text{span}(\tilde{A}) = \text{span}(E) \Leftrightarrow E = \tilde{A}\tilde{A}^+. \end{aligned}$$

Now let (1) hold. Then from (1a)  $\text{span}(\tilde{A}) = \text{span}(E)$ , and since  $\tilde{b} \in \text{span}(E)$ , (2a) follows. Since  $(A - \tilde{A}, b - \tilde{b}) = (I - E)(A, b)$  and  $\tilde{A}^H(I - E) = \tilde{A}^H(I - \tilde{A}\tilde{A}^+) = 0$ , (2b) also follows.

If, conversely, (2) holds, we take  $E := \tilde{A}\tilde{A}^+$ . Then  $E$  is an orthogonal projector with rank  $k$ , and from (2b) it follows readily that (1b) holds. The first sentence of this proof now shows the validity of (1a). ■

If  $A$  is close enough to a rank  $k$  matrix  $A_0$ , then orthogonal projection of a pair  $(A, b)$  on a  $k$ -dimensional subspace close enough to  $\text{span}(A_0)$  yields a consistent pair  $(\tilde{A}, \tilde{b})$  for which  $\tilde{A}$  is close to  $A$ . This is the substance of the following theorem:

**THEOREM 4.2.** *Let  $A_0$  have rank  $k$ ,  $E$  be an orthogonal projector,  $(A, b)$  be a pair with  $A = A_0 + \delta A$  and  $(\tilde{A}, \tilde{b}) := E(A, b)$ . If  $E$  and  $\delta A$  satisfy*

$$\|E - A_0 A_0^+\| \|A_0\| + \|\delta A\| \leq \varepsilon < \sigma_k(A_0), \quad (4.1)$$

*then the pair  $(\tilde{A}, \tilde{b})$  is a consistent rank  $k$  pair and  $\|\tilde{A} - A\| \leq \varepsilon$ .*

*Proof.* From (4.1) we have  $\|E - A_0 A_0^+\| < \sigma_k(A_0)/\|A_0\| \leq 1$ ; hence (cf. [10, Theorem 2.3])  $\text{rank}(E) = \text{rank}(A_0 A_0^+) = k$  and therefore  $\text{rank}(\tilde{A}) \leq k$ . Furthermore  $\|\tilde{A} - A_0\| = \|E(A_0 + \delta A) - A_0\| = \|(E - A_0 A_0^+)A_0 + E\delta A\| \leq \varepsilon < \sigma_k(A_0)$ , implying  $\text{rank}(\tilde{A}) \geq k$ . Hence  $\text{rank}(\tilde{A}) = k$  and therefore  $\text{span}(\tilde{A}) = \text{span}(E)$ . Since  $\tilde{b} \in \text{span}(E)$ ,  $(\tilde{A}, \tilde{b})$  is consistent. Finally,  $\|\tilde{A} - A\| = \|(E - I)(A_0 + \delta A)\| = \|(E - A_0 A_0^+)A_0 - (I - E)\delta A\| \leq \varepsilon$ . ■

In the following two examples we discuss projectors  $E$  of practical importance.

**EXAMPLE 4.3.** If  $\text{rank}(A) = n$  and we want a consistent rank  $n$  pair, then we may take  $E = AA^+$ , in which case  $\tilde{A} = A$ ,  $\tilde{b} = AA^+b$ , i.e.,  $\tilde{b}$  is just the projection of  $b$  on  $\text{span}(A)$ . This is the full rank least squares case.

**EXAMPLE 4.4.** For any given matrix  $A$  we may write

$$A\Pi = QR = \begin{pmatrix} Q_1 & Q_2 \end{pmatrix} \begin{pmatrix} R_{11} & R_{12} \\ 0 & R_{22} \end{pmatrix}, \quad (4.2)$$

where  $\Pi \in \mathbb{C}^{n \times n}$  is a permutation matrix,  $Q \in \mathbb{C}^{m \times n}$  has orthonormal columns,  $R \in \mathbb{C}^{n \times n}$  is upper triangular,  $Q_1 \in \mathbb{C}^{m \times k}$  and  $R_{11} \in \mathbb{C}^{k \times k}$ .

If  $\text{rank}(A) \geq k$ , then we may choose  $\Pi$  so that  $R_{11}$  is regular (Golub and Businger [5]). Then

$$\tilde{A} := Q_1 Q_1^H A = Q_1 \begin{pmatrix} R_{11} & R_{12} \end{pmatrix} \Pi^T \quad (4.3)$$

is a rank  $k$  approximation to  $A$ , and for any  $b \in \mathbb{C}^m$ ,  $(\tilde{A}, Q_1 Q_1^H b)$  is a consistent rank  $k$  pair obtained from  $(A, b)$  by orthogonal projection onto the span of the first  $k$  columns of  $A\Pi$ .

If, in particular,  $A = A_0 + \delta A$ , where  $A_0$  has rank  $k$  and  $\delta A$  is small [ $\|\delta A\| \ll \sigma_k(A_0)$ ], then by a judicious choice of  $\Pi$  we may obtain an  $\tilde{A}$  of rank  $k$  that is near to  $A$ . Lawson and Hanson have implemented this idea in their procedure HFTI (cf. [7, p.81]). Using their Theorem (6.31) (ascribed to Faddeev, Kublanovskaya and Faddeeva) it can be shown that

$$\|A - \tilde{A}\|_F \leq C(n, k) \sigma_{k+1}(A) \leq C(n, k) \|\delta A\|, \quad (4.4)$$

where  $C(n, k) := 3^{-1}(n - k)^{1/2}(4^{k+1} + 6k + 5)^{1/2}$ .

Hence this algorithm (which is noniterative, involving only rational operations and square roots) affords a rigorous *a priori* bound for  $\|A - \tilde{A}\|/\|\delta A\|$  (with a disappointingly large numerical factor  $C(n, k)$ , however).

Concerning this algorithm we note that in [7, Theorem (6.31)], the assumption that all column vectors of  $A$  have unit euclidean length is superfluous. This is of importance because with respect to the column interchange strategy in HFTI it stands to reason to scale the columns in such a way that they all have about the same (absolute) uncertainty (i.e. take  $AT^{-1}$ , where  $T^{-1}$  is a matrix as mentioned in Sec. 1.3). If that is done, then (4.4) applies to the scaled matrices, not to the original ones. One should remember, however, in the case  $k < n$  to undo this scaling after the column interchanges if the minimum length solution is sought.

#### 4.2. Perturbation Results

Let  $(A_0, b_0)$  be a consistent rank  $k$  pair, and let  $x_0 := A_0^+ b_0$ , so  $b_0 = A_0 x_0$ . Let

$$(A, b) = (A_0, b_0) + (\delta A, \delta b). \quad (4.5)$$

Let  $E$  be an orthogonal projector with rank  $k$  such that  $E(A, b)$  is a consistent rank  $k$  pair, and let

$$(\tilde{A}, \tilde{b}) := E(A, b), \quad \Delta \tilde{A} := A - \tilde{A}, \quad \tilde{x} := \tilde{A}^+ \tilde{b}. \quad (4.6)$$

Then we have  $E = \tilde{A} \tilde{A}^+$ ,  $\tilde{A}^H \Delta \tilde{A} = 0$  and  $\tilde{x} = \tilde{A}^+ b$  (cf. Theorem 4.1).

LEMMA 4.5. *With the notation introduced above and*

$$r_0 := b - Ax_0 = \delta b - \delta A x_0, \quad (4.7)$$

we have

$$\tilde{x} = x_0 + \tilde{A}^+ r_0 - (I - \tilde{A}^+ \tilde{A}) x_0 \quad (4.8)$$

$$= x_0 + \tilde{A}^+ r_0 + (I - \tilde{A}^+ \tilde{A}) [A_0^+ (\delta A - \Delta \tilde{A})]^H x_0, \quad (4.9)$$

$$b - A\tilde{x} = (I - \tilde{A} \tilde{A}^+) r_0 - \Delta \tilde{A} (\tilde{x} - x_0). \quad (4.10)$$

*Proof.* From  $Eb = \tilde{b} = \tilde{A}\tilde{x}$  and  $EA = \tilde{A}$  we have

$$Er_0 = E(b - Ax_0) = \tilde{A}(\tilde{x} - x_0).$$

Since  $\tilde{A}^+ E = \tilde{A}^+$  and  $\tilde{A}^+ \tilde{A}\tilde{x} = \tilde{x}$ , it follows that

$$\tilde{A}^+ r_0 = \tilde{x} - \tilde{A}^+ \tilde{A} x_0,$$

which is equivalent to (4.8).

In a similar way we have

$$(I - E)r_0 = (I - E)[b - A\tilde{x} + A(\tilde{x} - x_0)] = b - A\tilde{x} + \Delta \tilde{A}(\tilde{x} - x_0),$$

which is (4.10).

Since  $x_0 \in \text{span}(A_0^H)$  (cf. Properties 3.5 and 3.10) and  $\text{span}(\tilde{A}^H) = \ker(I - \tilde{A}^+ \tilde{A})$ , the last term in (4.8) must involve a factor  $(\tilde{A} - A_0)^H$ . Indeed, using Properties 3.10 and 3.4,

$$\begin{aligned} (I - \tilde{A}^+ \tilde{A}) x_0 &= (I - \tilde{A}^+ \tilde{A}) A_0^H A_0^{+H} x_0 = (I - \tilde{A}^+ \tilde{A}) (A_0^H - \tilde{A}^H) A_0^{+H} x_0 \\ &= -(I - \tilde{A}^+ \tilde{A}) [A_0^+ (\delta A - \Delta \tilde{A})]^H x_0. \end{aligned}$$

This proves (4.9). ■

**REMARK 4.6.** If  $k = n$ , then  $\tilde{A}^+ \tilde{A} = I$ , which simplifies the expressions for  $\tilde{x}$ .

**REMARK 4.7.** If  $k = n$ , we may choose  $E = AA^+$ , in which case we obtain  $\tilde{A} = A$ ,  $\Delta \tilde{A} = 0$ ,  $\tilde{x} = A^+ b = x_0 + A^+ r_0$ ,  $b - A\tilde{x} = (I - AA^+) r_0$ . This is the ordinary least squares approach for the full rank situation.

REMARK 4.8. Since it follows from (4.6) and Property 3.5 that  $\tilde{A}^+ \Delta \tilde{A} = 0$ , we may expect that the term  $A_0^+ \Delta \tilde{A}$  in (4.9) is small of second order, and this implies that the first order terms in  $\tilde{x} - x_0$  only depend on  $\delta b$  and  $\delta A$ , not on  $\Delta \tilde{A}$ . This is elucidated in the following theorem.

THEOREM 4.9. Let  $(\tilde{A}, \tilde{b})$  and  $(\tilde{A}', \tilde{b}')$  be consistent rank  $k$  pairs obtained from  $(A, b)$  by orthogonal projection, and define  $\tilde{x} = \tilde{A}^+ \tilde{b}$ ,  $\tilde{x}' = \tilde{A}'^+ \tilde{b}'$ . Then, with the notation introduced above, and  $\Delta \tilde{A}' := A - \tilde{A}'$ , we have

$$\tilde{x} - \tilde{x}' = (\tilde{A}^+ - \tilde{A}'^+) r_0 + (\tilde{A}^+ \tilde{A} - \tilde{A}'^+ \tilde{A}') x_0 \quad (4.11)$$

and

$$\|\tilde{x} - \tilde{x}'\| \leq \alpha^2 \|\Delta \tilde{A} - \Delta \tilde{A}'\|_F [\|r_0\| + \min(\|\Delta \tilde{A}\|, \|\Delta \tilde{A}'\|) \|x_0\|], \quad (4.12)$$

with

$$\begin{aligned} \alpha &:= \max(\|\tilde{A}^+\|, \|\tilde{A}'^+\|) \\ &\leq \frac{\|A_0^+\|}{1 - \|A_0^+\| \max(\|\Delta \tilde{A} - \delta A\|, \|\Delta \tilde{A}' - \delta A\|)}, \end{aligned}$$

whenever the denominator of the last expression is positive.

*Proof.* The relation (4.11) follows trivially from (4.8). We already know how to estimate  $\|\tilde{A}^+ - \tilde{A}'^+\|$  (cf. Theorem 3.14). Now we wish to estimate  $\|\tilde{A}^+ \tilde{A} - \tilde{A}'^+ \tilde{A}'\|$ . We note that  $\tilde{A}^+ \Delta \tilde{A} = \tilde{A}'^+ \Delta \tilde{A}' = 0$  [cf. (4.6) and Property 3.5]. Now starting off with a well-known relation for orthogonal projectors of the same rank (cf. [10, Theorem 2.3]), we have

$$\begin{aligned} \|\tilde{A}^+ \tilde{A} - \tilde{A}'^+ \tilde{A}'\| &= \|\tilde{A}^+ \tilde{A} (I - \tilde{A}'^+ \tilde{A}')\| = \|\tilde{A}^+ (\tilde{A} - \tilde{A}') (I - \tilde{A}'^+ \tilde{A}')\| \\ &\leq \|\tilde{A}^+ (\tilde{A} - \tilde{A}')\| = \|\tilde{A}^+ (\Delta \tilde{A}' - \Delta \tilde{A})\| = \|\tilde{A}^+ \Delta \tilde{A}'\| \\ &= \|(\tilde{A}^+ - \tilde{A}'^+) \Delta \tilde{A}'\| \leq \|\tilde{A}^+ - \tilde{A}'^+\| \|\Delta \tilde{A}'\|. \end{aligned}$$

Since  $\tilde{A}$  and  $\tilde{A}'$  occur symmetrically we have, finally,

$$\|\tilde{A}^+ \tilde{A} - \tilde{A}'^+ \tilde{A}'\| \leq \|\tilde{A}^+ - \tilde{A}'^+\| \min(\|\Delta \tilde{A}\|, \|\Delta \tilde{A}'\|). \quad (4.13)$$

Hence we have from (4.11)

$$\|\tilde{x} - \tilde{x}'\| \leq \|\tilde{A}^+ - \tilde{A}'^+\| [\|r_0\| + \min(\|\Delta\tilde{A}\|, \|\Delta\tilde{A}'\|)\|x_0\|], \quad (4.14)$$

from which (4.12) immediately follows, using (3.1).

The bound for  $\alpha$  follows from Property 3.13. ■

**REMARK 4.10.** The formula (4.13) contains the remarkable result that  $\|\tilde{A}^+ \tilde{A} - \tilde{A}'^+ \tilde{A}'\| = O(\epsilon^2)$  if  $\delta A$ ,  $\Delta\tilde{A}$  and  $\Delta\tilde{A}'$  are  $O(\epsilon)$ . This means in fact that the angle between the nullspaces of  $\tilde{A}$  and  $\tilde{A}'$  is  $O(\epsilon^2)$ .

From Theorem 4.9 we have [using (4.7)] the result mentioned in Sec. 1.4:

**THEOREM 4.11.** *Let  $\|(\delta A : \delta b)\|_F \leq \epsilon$ ,  $\|\Delta\tilde{A}\|_F \leq \epsilon$ ,  $\|\Delta\tilde{A}'\|_F \leq \epsilon$  with  $\epsilon < \frac{1}{2}\|A_0^+\|^{-1}$ . Then  $\tilde{x} - \tilde{x}' = O(\epsilon^2)$ . More specifically,*

$$\|\tilde{x} - \tilde{x}'\| \leq 2\epsilon^2 \alpha^2 (\sqrt{1 + \|x_0\|^2} + \|x_0\|). \quad (4.15)$$

**REMARK 4.12.** The bounds in (4.12) and (4.15) have an *a priori* character in view of their containing the quantities  $x_0$  and  $r_0$ . However, looking at their proofs and at (4.8), we note that  $x_0$  may be an arbitrary vector, provided we take  $r_0 = b - Ax_0$ . Thus we may as well use (4.11) with  $\tilde{x}$  instead of  $x_0$  and  $\tilde{r} := b - A\tilde{x}$  instead of  $r_0$ :

$$\tilde{x} - \tilde{x}' = (\tilde{A}^+ - \tilde{A}'^+) \tilde{r} + (\tilde{A}^+ \tilde{A} - \tilde{A}'^+ \tilde{A}') \tilde{x}. \quad (4.16)$$

Hence in estimates like (4.12) we may replace  $r_0$  and  $x_0$  by  $\tilde{r}$  and  $\tilde{x}$ . Since also the quantity  $\alpha$  may be bounded in terms of  $\tilde{A}^+$ ,  $\Delta\tilde{A}$  and  $\Delta\tilde{A}'$ , we obtain instead of (4.12) a bound for  $\|\tilde{x} - \tilde{x}'\|$  which involves only factors that can be computed from the least squares system  $\tilde{A}x = b$ , provided bounds for  $\Delta\tilde{A}$  and  $\Delta\tilde{A}'$  are known.

We now want to derive a first order perturbation formula for  $\tilde{x}$ . Therefore we consider the projector  $E_0 := A_0 A_0^+$ . Then  $\tilde{A}_0 := E_0 A = A_0(I + A_0^+ \delta A)$ . If we assume that  $\|\delta A\| < \|A_0^+\|^{-1}$ , then  $I + A_0^+ \delta A$  is regular; hence  $\text{span}(\tilde{A}_0) = \text{span}(A_0)$ ,  $E_0 = A_0 A_0^+ = \tilde{A}_0 \tilde{A}_0^+$ ,  $\tilde{A}_0 - A_0 = E_0 \delta A$  and  $\Delta\tilde{A}_0 := A - \tilde{A}_0 = (I - E_0) \delta A$ , implying that  $A_0^+ \Delta\tilde{A}_0 = 0$ . Thus we find from (4.9) for  $\tilde{x}_0 := \tilde{A}_0^+ \tilde{b}_0$ :

$$\tilde{x}_0 = x_0 + \tilde{A}_0^+ r_0 + (I - \tilde{A}_0^+ \tilde{A}_0)(A_0^+ \delta A)^H x_0. \quad (4.17)$$

Then we have

**THEOREM 4.13.** *Let  $\|\delta A\| < \|A_0^+\|^{-1}$ . Then with the above notation,*

$$\|\tilde{x}_0 - x_0\| \leq \alpha_0(\|r_0\| + \|\delta A\| \|x_0\|), \quad (4.18)$$

$$\|\tilde{x}_0 - x_0 - A_0^+ r_0 - (I - A_0^+ A_0)(A_0^+ \delta A)^H x_0\| \leq \alpha_0^2 \|\delta A\|_F (\|r_0\| + \|\delta A\| \|x_0\|) \quad (4.19)$$

with

$$\alpha_0 := \max(\|A_0^+\|, \|\tilde{A}_0^+\|) \leq \frac{\|A_0^+\|}{1 - \|A_0^+\| \|\delta A\|}.$$

*Proof.* The estimate (4.18) follows trivially from (4.17). In order to prove (4.19) we rewrite (4.17) as

$$\begin{aligned} \tilde{x}_0 - x_0 - A_0^+ r_0 - (I - A_0^+ A_0)(A_0^+ \delta A)^H x_0 \\ = (\tilde{A}_0^+ - A_0^+) r_0 - (\tilde{A}_0^+ \tilde{A}_0 - A_0^+ A_0)(A_0^+ \delta A)^H x_0. \end{aligned}$$

In order to estimate  $\tilde{A}_0^+ \tilde{A}_0 - A_0^+ A_0$  we proceed as in the proof of Theorem 4.9 and have

$$\begin{aligned} \|\tilde{A}_0^+ \tilde{A}_0 - A_0^+ A_0\| &= \|\tilde{A}_0^+ \tilde{A}_0 (I - A_0^+ A_0)\| \\ &= \|\tilde{A}_0^+ (\tilde{A}_0 - A_0)(I - A_0^+ A_0)\| \leq \|\tilde{A}_0^+\| \|\delta A\|. \end{aligned}$$

Using this and (3.1) now yields (4.19). ■

**REMARK 4.14.** If  $(\tilde{A}, \tilde{b})$  is a consistent rank  $k$  pair, obtained from  $(A, b)$  by orthogonal projection, and  $\tilde{x} := \tilde{A}^+ \tilde{b}$ , then  $\tilde{x} - x_0$  may be estimated by combination of (4.15) and (4.18), and  $\tilde{x} - x_0 - A_0^+ r_0 - (I - A_0^+ A_0)(A_0^+ \delta A)^H x_0$  may be estimated by (4.15) and (4.19). Notably we have: if  $\|(\delta A \ ; \ \delta b)\|_F \leq \varepsilon$  and  $\|\Delta \tilde{A}\|_F \leq \varepsilon$ ,  $\varepsilon < \frac{1}{2} \|A_0^+\|^{-1}$ , then

$$\|\tilde{x} - x_0 - A_0^+ r_0 - (I - A_0^+ A_0)(A_0^+ \delta A)^H x_0\| \leq 3\varepsilon^2 \tilde{\alpha}^2 \left( \sqrt{1 + \|x_0\|^2} + \|x_0\| \right), \quad (4.20)$$

where now  $\tilde{\alpha} := \max(\|A_0^+\|, \|\tilde{A}_0^+\|, \|\tilde{A}^+\|) \leq \|A_0^+\| / (1 - 2\varepsilon \|A_0^+\|)$ .

This result shows most clearly that, to first order,  $\tilde{x}$  does not depend on the choice of the  $k$ -dimensional subspace  $\text{span}(\tilde{A})$ , provided  $\Delta\tilde{A}=0(\epsilon)$ .

Similarly we have

$$b - A\tilde{x} = (I - A_0 A_0^+) r_0 + \text{second order.} \quad (4.21)$$

## 5. NEAREST CONSISTENT PAIRS OF GIVEN RANK

### 5.1. Existence of Nearest Consistent Pairs

If  $(A, b)$  is a given pair of rank  $p$ , then for any  $k$ ,  $0 \leq k \leq p$ , there exists a rank  $k$  pair  $(\hat{A}, \hat{b})$  which is nearest to  $(A, b)$  in the sense of the Frobenius distance  $\|(A \vdash b) - (\hat{A} \vdash \hat{b})\|_F$ . However, this nearest pair need not be consistent, and indeed, a nearest consistent pair of rank  $k$  need not exist. This is rather obvious for  $k=p$ , but that it may also happen for  $k < p$  is shown in the following example.

EXAMPLE 5.1 (cf. also Golub [4, p. 329]). Let

$$(A \vdash b) = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}.$$

Then 2 and 1 are the singular values of this matrix, and thus the nearest rank 1 matrix  $Q$  is uniquely determined [cf. Theorem 3.1(f)], and

$$Q = \begin{pmatrix} 0 & 0 \\ 0 & 2 \end{pmatrix}.$$

Thus the rank 1 pair  $(\hat{A}, \hat{b})$  nearest to  $(A, b)$  is uniquely determined and is not consistent. But rank 1 matrices like  $\begin{pmatrix} 0 & 0 \\ \epsilon & 2 \end{pmatrix}$  show that there are consistent rank 1 pairs in every neighborhood of  $(\hat{A}, \hat{b})$ , and hence there can be no nearest consistent rank 1 pair; thus, on the set of all consistent rank 1 pairs the distance from  $(A, b)$  has only an infimum, not a minimum.

This example can obviously be generalized to arbitrary dimension and rank. Also, it is generally true that (as is the case in this example) if no nearest consistent rank  $k$  pair exists, there are consistent rank  $k$  pairs in any neighborhood of the nearest rank  $k$  pair; this is the subject of the following theorem.

THEOREM 5.2. Let  $(A, b)$  have rank  $p$ . For given  $k$ ,  $0 \leq k \leq p$ , let  $S$  denote the set of all consistent rank  $k$  pairs. Let  $(\hat{A}, \hat{b})$  be the rank  $k$  pair nearest to  $(A, b)$ . Then  $\inf_{(A', b') \in S} \|(A' \vdash b') - (\hat{A} \vdash \hat{b})\|_F = 0$ .



*Proof.* We need only look at the case that  $\text{rank}(\hat{A}) = k - 1$ . Now take  $u \in \ker(\hat{A})$ ,  $\|u\| = 1$ , and consider  $A_\varepsilon := \hat{A} + \varepsilon \hat{b}u^H$ . Then  $\hat{A} = A_\varepsilon(I - uu^H)$ ; hence  $\text{span}(\hat{A}) \subset \text{span}(A_\varepsilon)$ . Since  $\hat{b} \notin \text{span}(\hat{A})$  and  $\hat{b} \in \text{span}(A_\varepsilon)$ , we have  $\text{rank}(A_\varepsilon) = k$  and  $(A_\varepsilon, \hat{b}) \in S$  for  $\varepsilon \neq 0$ . ■

**COROLLARY 5.3.** *If  $\bar{S}$  denotes the set of all rank  $k$  pairs, then*

$$\inf_{(A', b') \in S} \|(A' \vdash b') - (A \vdash b)\|_F = \inf_{(A', b') \in \bar{S}} \|(A' \vdash b') - (A \vdash b)\|_F.$$

However, under suitable conditions nearest consistent pairs of given rank do exist, as is seen from the following theorem. Here  $\Omega_\varepsilon(P)$  denotes the set of matrices  $Q$  with  $\|Q - P\|_F \leq \varepsilon$ .

**THEOREM 5.4.** *Let  $(A_0, b_0)$  be a consistent rank  $k$  pair. Let  $(A, b) \in \Omega_\varepsilon(A_0, b_0)$ ,  $\varepsilon < \frac{1}{2} \tilde{\sigma}_k(A_0)$ . Then there exists a consistent rank  $k$  pair  $(\hat{A}, \hat{b})$  nearest to  $(A, b)$ , and there holds  $(\hat{A}, \hat{b}) \in \Omega_\varepsilon(A, b)$ .*

*Proof.* Obviously a rank  $k$  pair  $(\hat{A}, \hat{b})$  nearest to  $(A, b)$  belongs to  $\Omega_\varepsilon(A, b)$ , since  $(A_0, b_0) \in \Omega_\varepsilon(A, b)$ . Hence  $\hat{A} \in \Omega_{2\varepsilon}(A_0)$ , and thus  $\text{rank}(\hat{A}) = k$  (cf. Theorem 3.2(d)).

**THEOREM 5.5.** *Let  $(A, b)$  have rank  $\geq k$ , let  $A'$  have rank  $k$ , and let  $d$  denote the distance between  $b$  and  $\text{span}(A')$ . If one of the conditions*

$$\|A' - A\|_F^2 + d^2 < \tilde{\sigma}_k(A)^2, \quad (5.1)$$

$$\tilde{\sigma}_k(A')^2 - 2\tilde{\sigma}_k(A')\|A' - A\|_F - d^2 > 0 \quad (5.2)$$

*is satisfied, then there exists a consistent rank  $k$  pair  $(\hat{A}, \hat{b})$  nearest to  $(A, b)$ .*

*Proof.* Assume (5.1). If  $b'$  is the projection of  $b$  on  $\text{span}(A')$ , then  $(A', b')$  is a consistent pair, and  $(A', b') \in \Omega_\varepsilon(A, b)$  with  $\varepsilon^2 = \|A' - A\|_F^2 + d^2 < \tilde{\sigma}_k(A)^2$ . Hence the nearest rank  $k$  pair  $(\hat{A}, \hat{b})$  must belong to  $\Omega_\varepsilon(A, b)$ , and therefore  $\hat{A} \in \Omega_\varepsilon(A)$ . Thus  $\hat{A}$  has rank  $k$  (cf. Theorem 3.2). The condition (5.2) implies (5.1): from Theorem 3.2(b), we have  $\tilde{\sigma}_k(A) \geq \tilde{\sigma}_k(A') - \|A' - A\|_F > 0$  [cf. (5.2)]. Squaring this inequality and using (5.2), we have  $\|A' - A\|_F^2 + d^2 < (\tilde{\sigma}_k(A') - \|A' - A\|_F)^2 \leq \tilde{\sigma}_k(A)^2$ . ■

**REMARK 5.6.** In the context of this paper Theorem 5.5, in particular with the condition (5.2), is more practicable than Theorem 5.4, since it uses only quantities which are known in principle.

REMARK 5.7. Theorems 5.4 and 5.5 remain true if all Frobenius norms are replaced by 2-norms (also in the definition of  $\Omega_\epsilon$ ) and  $\tilde{\sigma}_k$  is replaced by  $\sigma_k$ .

### 5.2. Results for Nearest Consistent Pairs

Let  $(\hat{A}, \hat{b})$  be a rank  $k$  pair nearest to  $(A, b)$  in the sense of the Frobenius distance, and write

$$(A, b) = (\hat{A}, \hat{b}) + (\Delta\hat{A}, \Delta\hat{b}). \quad (5.3)$$

Then from Theorem 3.1 we have

$$\hat{A}^H \Delta\hat{A} = 0, \quad \hat{A}^H \Delta\hat{b} = 0, \quad (5.4)$$

$$\Delta\hat{A} \hat{A}^H + \Delta\hat{b} \hat{b}^H = 0. \quad (5.5)$$

Obviously, (5.4) is equivalent to

LEMMA 5.8.  $(\hat{A}, \hat{b})$  is obtained from  $(A, b)$  by orthogonal projection.

In view of Lemma 5.8 we may apply the results of Sec. 4.2 with  $(\hat{A}, \hat{b})$  as  $(\tilde{A}, \tilde{b})$ . As an application of Theorem 4.11 we have

THEOREM 5.9. Let the assumptions about  $(A_0, b_0)$ ,  $(A, b)$ ,  $(\tilde{A}, \tilde{b})$  made at the beginning of Sec. 4.2 hold. Let  $\|(\delta A \ ; \ \delta b)\|_F \leq \epsilon$ ,  $\|\Delta\tilde{A}\|_F \leq \epsilon$  with  $\epsilon < \frac{1}{2}\|A_0^+\|^{-1}$ . Then  $(\hat{A}, \hat{b})$  is consistent, and  $\hat{x} = \hat{A}^+ \hat{b}$  satisfies

$$\|\hat{x} - \tilde{x}\| \leq 2\epsilon^2 \alpha^2 (\sqrt{1 + \|x_0\|^2} + \|x_0\|) \quad (5.6)$$

with  $\alpha = \|A_0^+\|/(1 - 2\epsilon\|A_0^+\|)$ .

*Proof.* Since  $(A_0, b_0)$  has Frobenius distance at most  $\epsilon$  from  $(A, b)$  and is a consistent rank  $k$  pair, we certainly have  $\|(\Delta\tilde{A} \ ; \ \Delta\tilde{b})\|_F \leq \epsilon$ . Consequently  $\|A_0 - \hat{A}\|_F \leq 2\epsilon < \sigma_k(A_0)$ , and thus  $\text{rank}(\hat{A}) = k$ , implying that  $(\hat{A}, \hat{b})$  is consistent. Now (5.6) follows right away from (4.15). ■

Thus, by solving a consistent rank  $k$  pair  $(\tilde{A}, \tilde{b})$  near (but not nearest) to  $(A, b)$  we can obtain an  $O(\epsilon^2)$  approximation to  $\hat{x}$ , and hence also to  $\hat{r} = b - A\hat{x}$ .

The next theorem shows that (5.5) implies a close relationship between  $\Delta\hat{b}$ ,  $\Delta\hat{A}$ ,  $\hat{x}$  and  $\hat{r}$ , from which it will be possible to obtain  $O(\epsilon^2)$  approximations to  $\Delta\hat{b}$  and, if  $k = n$ , to  $\Delta\hat{A}$ .

**THEOREM 5.10.** *If  $(\hat{A}, \hat{b})$  is a consistent rank  $k$  pair nearest to a given pair  $(A, b)$  and  $\hat{x} = \hat{A}^+ \hat{b}$ ,  $\hat{r} = b - A\hat{x}$ , then*

$$\Delta \hat{b} = \frac{\hat{r}}{1 + \|\hat{x}\|^2}, \quad \Delta \hat{A} = -\Delta \hat{b} \hat{x}^H + \Delta \hat{A} (I - \hat{A}^+ \hat{A}). \quad (5.7)$$

*Proof.* From (5.5) we have

$$\Delta \hat{A} \hat{A}^+ \hat{A} = \Delta \hat{A} \hat{A}^H \hat{A}^{+H} = -\Delta \hat{b} \hat{b}^H \hat{A}^{+H} = -\Delta \hat{b} \hat{x}^H, \quad (5.8)$$

which is essentially the latter formula in (5.7). The former formula now follows from

$$\hat{r} = b - A\hat{x} = \Delta \hat{b} - \Delta \hat{A} \hat{x} = \Delta \hat{b} - \Delta \hat{A} \hat{A}^+ \hat{A} \hat{x} = \Delta \hat{b} + \Delta \hat{b} \hat{x}^H \hat{x}$$

[in the second step consistency and in the last step (5.8) has been used]. ■

**REMARK 5.11.** If  $k = n$ , then  $\hat{A}^+ \hat{A} = I$ ; hence in this case  $(\Delta \hat{A}, \Delta \hat{b})$  can be expressed entirely in terms of  $\hat{x}$  and  $\hat{r}$ , viz

$$\begin{pmatrix} \Delta \hat{A} \\ \Delta \hat{b} \end{pmatrix} = \frac{\hat{r} \begin{pmatrix} -\hat{x}^H \\ 1 \end{pmatrix}}{1 + \|\hat{x}\|^2} \quad (5.9)$$

This result is in accordance with Theorem 3.1(c) (with  $p = n + 1$ ,  $q = n$ ), since it is easily shown that  $\|(\Delta \hat{A} \begin{smallmatrix} : \\ \Delta \hat{b} \end{smallmatrix})\|_F = \|\hat{r}\| / (1 + \|\hat{x}\|^2)^{1/2}$  is a singular value of  $(A \begin{smallmatrix} : \\ b \end{smallmatrix})$  as well as of  $(\Delta \hat{A} \begin{smallmatrix} : \\ \Delta \hat{b} \end{smallmatrix})$  with  $(-\hat{x}^H \begin{smallmatrix} : \\ 1 \end{smallmatrix})^H$  and  $\hat{r}$  as (unnormalized) right and left singular vectors. Compare also Golub [4] for another approach for the case  $k = n$ .

**COROLLARY 5.12.** *If the assumptions of Theorem 5.9 are satisfied, then  $\hat{x} = \tilde{x} + O(\epsilon^2)$ ,  $\hat{r} = \tilde{r} + O(\epsilon^2)$  with  $\tilde{r} := b - A\tilde{x}$ . Hence  $\Delta \hat{b} = (1 + \|\tilde{x}\|^2)^{-1} \tilde{r} + O(\epsilon^2)$  and, if  $k = n$ ,  $\Delta \hat{A} = -(1 + \|\tilde{x}\|^2)^{-1} \tilde{r} \tilde{x}^H + O(\epsilon^2)$ .*

**REMARK 5.13.** Until now we have only considered nearest pairs in the sense of the Frobenius distance, i.e., in terms of Sec. 1.3, the case  $T = \alpha I$ ,  $\alpha$  a scalar.

Since orthogonal projection commutes with column scaling, the formula (5.4) and Lemma 5.8 remain valid if  $(\hat{A}, \hat{b})$  is nearest to  $(A, b)$  in the sense

that  $\|(\Delta\hat{A} \begin{smallmatrix} \vdots \\ \hat{A} \end{smallmatrix})T^{-1}\|_F$  is minimal, where  $T$  is a matrix as in Sec. 1.3. Hence the observation following Lemma 5.8 applies also to nearest pairs in this sense (nevertheless, Theorem 5.9 needs adjustment if  $T \neq \alpha I$ , since then  $\|\Delta\hat{A}\|_F \leq \epsilon$  is not implied).

If  $T \neq \alpha I$  we get instead of (5.5)

$$\Delta\hat{A}\tilde{T}^{-2}\hat{A}^H + \tau^{-2}\Delta\hat{b}\hat{b}^H = 0 \quad (5.10)$$

with  $\tilde{T} = \text{diag}(t_{11}, \dots, t_{nn})$ ,  $\tau = t_{n+1, n+1}$ .

This means that then the proof of Theorem 5.10 is no longer valid. Now one might observe that for

- (1)  $k = n$ ,  $T$  arbitrary nonsingular,
- (2)  $k < n$ ,  $\tilde{T} = \alpha I$ ,

we can easily adapt theorem and proof. However, these are just the uninteresting cases, since it is then allowed to scale matrix and right hand side beforehand, i.e. to consider the pair  $(A\tilde{T}^{-1}, \tau^{-1}b)$  and apply to this pair the procedures and theory developed for  $T = I$  (the actual problem in the case  $k < n$ ,  $\tilde{T} \neq \alpha I$  is that for a rank  $k$  matrix  $P$  the generalized inverse of  $P\tilde{T}$  is, in general, unequal to  $\tilde{T}^{-1}P^+$ ).

### 5.3. Perturbation formulae

If, as before,  $(A_0, b_0)$  is a consistent rank  $k$  pair and  $(A, b) = (A_0, b_0) + (\delta A, \delta b)$ , where  $(\delta A, \delta b)$  is of order  $\epsilon$ , say, we might wish to approximate the best rank  $k$  approximation  $(\hat{A}, \hat{b})$  to  $(A, b)$  in terms of  $A_0, b_0, \delta A, \delta b$ . Since  $(\Delta\hat{A} \begin{smallmatrix} \vdots \\ \hat{A} \end{smallmatrix})$  can be found from the last  $n - k + 1$  singular values and vectors of  $(A \begin{smallmatrix} \vdots \\ b \end{smallmatrix})$ , and  $(A \begin{smallmatrix} \vdots \\ b \end{smallmatrix})$  is close to  $(A_0 \begin{smallmatrix} \vdots \\ b_0 \end{smallmatrix})$ , whose  $n - k + 1$  last singular values are zero, we may use the perturbation theory for the singular value decomposition (Stewart [9]). However, first order approximations for  $(\hat{A}, \hat{b})$  can be readily derived from the results obtained so far.

**THEOREM 5.14.** *Let  $(A_0, b_0)$  be a consistent rank  $k$  pair,  $(A, b) = (A_0, b_0) + (\delta A, \delta b)$  with  $\|(\delta A \begin{smallmatrix} \vdots \\ \delta b \end{smallmatrix})\|_F \leq \epsilon$ , and  $(\hat{A}, \hat{b})$  the rank  $k$  pair nearest to  $(A, b)$  (which is unique and consistent if  $\epsilon < \frac{1}{2}\|A_0^+\|^{-1}$ ; cf. Theorem 5.4). Then, if  $x_0 = A_0 + b_0$ ,*

$$\hat{b} = b_0 + \delta b - \frac{(I - A_0 A_0^+)(\delta b - \delta A x_0)}{1 + \|x_0\|^2} + O(\epsilon^2), \quad (5.11)$$

$$\begin{aligned} \hat{A} = A_0 + \delta A + & \frac{(I - A_0 A_0^+)(\delta b - \delta A x_0)x_0^H}{1 + \|x_0\|^2} \\ & - (I - A_0 A_0^+)\delta A(I - A_0^+ A_0) + O(\epsilon^2). \end{aligned} \quad (5.12)$$

*Proof.* Since  $\hat{b} = b_0 + \delta b - \Delta \hat{b}$ , (5.11) follows from Theorem 5.10 together with (4.10) and (4.7). Similarly we have  $\hat{A} = A_0 + \delta A - \Delta \hat{A}$ , in which the component  $\Delta \hat{A} \hat{A}^+ \hat{A}$  follows again from Theorem 5.10. In order to approximate  $\Delta \hat{A} (I - \hat{A}^+ \hat{A})$  we observe that  $(I - \hat{A} \hat{A}^+) (\Delta \hat{A} - \delta A) (I - A_0^+ A_0) = (I - \hat{A} \hat{A}^+) (A_0 - \hat{A}) (I - A_0^+ A_0) = 0$ . Hence, since  $\hat{A}^+ \Delta \hat{A} = 0$ ,  $\Delta \hat{A} (I - \hat{A}^+ \hat{A}) = (I - A_0 A_0^+) \delta A (I - A_0^+ A_0) + O(\epsilon^2)$ . Now (5.12) follows directly. ■

REMARK 5.15. The formulae (5.11) and (5.12) may be combined to

$$\begin{aligned} (\hat{A} \mid \hat{b}) &= (A_0 \mid b_0) + (\delta A \mid \delta b) - (I - A_0 A_0^+) (\delta A \mid \delta b) \\ &\times \left\{ \left( \begin{array}{c|c} I - A_0 A_0^+ & 0 \\ \hline - & 0 \end{array} \right) + (1 + \|x_0\|^2)^{-1} \begin{pmatrix} -x_0 \\ \hline 1 \end{pmatrix} \begin{pmatrix} -x_0^H & 1 \end{pmatrix} \right\} + O(\epsilon^2). \end{aligned} \quad (5.13)$$

We note that  $I - A_0 A_0^+ = I - (A_0 \mid b_0) (A_0 \mid b_0)^+$  is the orthogonal projector on  $\ker (A_0^H) = \ker ((A_0 \mid b_0)^H)$ . The factor between curly braces can be shown to be equal to  $I - (A_0 \mid b_0)^+ (A_0 \mid b_0)$ , i.e. the orthogonal projector on  $\ker ((A_0 \mid b_0))$ . This result can also be found by elaboration of Stewart's results on perturbations of the singular value decomposition [9, §6].

REMARK 5.16. First order perturbation formulae for  $\Delta \hat{x}$  and  $\hat{r} := b - A\hat{x}$  are implied in (4.20) and (4.21).

## REFERENCES

- 1 S. N. Afriat, Orthogonal and oblique projectors and the characteristics of pairs of vector spaces, *Proc. Cambridge Philos. Soc.* 53: 800–816 (1956).
- 2 A. Ben-Israel and T. N. E. Greville, *Generalized Inverses*, Wiley, New York, 1974.
- 3 C. Eckart and G. Young, The approximation of one matrix by another of lower rank. *Psychometrika* 1:211–218 (1936).
- 4 G. Golub, Some modified matrix eigenvalue problems, *SIAM Rev.* 15: 318–334 (1973).
- 5 G. H. Golub and P. A. Businger, Linear least squares solutions by Householder transformations, *Numer. Math.* 7: 269–276 (1965).
- 6 M. G. Kendall and A. Stuart, *The Advanced Theory of Statistics*, Vol. 2, 3rd ed., Griffin, London, 1973.
- 7 C. L. Lawson and R. J. Hanson, *Solving Least Squares Problems*, Prentice-Hall, Englewood Cliffs, N.J., 1974.

- 8 G. W. Stewart, *Introduction to Matrix Computations*, Academic Press, New York, 1973.
- 9 G. W. Stewart, Error and perturbation bounds for subspaces associated with certain eigenvalue problems, *SIAM Rev.* 15: 727–764 (1973).
- 10 G. W. Stewart, On the perturbation of pseudo-inverses, projections and linear least squares problems, *SIAM Rev.* 19: 634–662 (1977).
- 11 P. A. Wedin, Perturbation theory for pseudo-inverses, *Nordisk Tidskr. Informationsbehandling (BIT)* 13: 217–232 (1973).

*Received 4 January 1979; revised 16 February 1979*