A note on iterative refinement for seminormal equations

Miroslav Rozložník[a,*], Alicja Smoktunowicz[b,*], Jiří Kopal[c]

[a] *Institute of Computer Science, Academy of Sciences of the Czech Republic, Pod vodárenskou věží 2, CZ-182 07 Prague 8, Czech Republic*
[b] *Faculty of Mathematics and Information Science, Warsaw University of Technology, Pl.Politechniki 1, Warsaw, 00-661, Poland*
[c] *Technical University of Liberec, Institute of Novel Technologies and Applied Informatics, Hálkova 6, CZ-461 17 Liberec, Czech Republic*

Abstract

We present a roundoff error analysis of the method for solving linear least squares problem $\min_x \|b - Ax\|$ with full column rank matrix A, using only factors Σ and V from the SVD decomposition of $A = U\Sigma V^T$. This method (called SNE_{SVD} here) is an analogue of the method of seminormal equations (SNE_{QR}), where the solution is computed from $R^T Rx = A^T b$, using the factor R from the QR factorization of A. Such methods have practical applications when A is large and sparse and if one needs to solve least squares problems with the same matrix A and multiple right-hand sides. However, they are not numerically stable for all A. We analyze one step of fixed precision iterative refinement to improve the accuracy of the SNE_{SVD} method. We show that under mild conditions this method (called $CSNE_{SVD}$) pro-

*Corresponding author

**The work of M. Rozložník was supported by Grant Agency of the Academy of Sciences of the Czech Republic under the project IAA100300802. The work of J. Kopal was supported by the Ministry of Education of the Czech Republic under the project no. 7822/115.

Email addresses: miro@cs.cas.cz (Miroslav Rozložník), A.Smoktunowicz@mini.pw.edu.pl (Alicja Smoktunowicz), jiri.kopal@tul.cz (Jiří Kopal)

duces a forward stable solution. We illustrate our analysis by numerical experiments.

Keywords: Condition number; numerical stability; normal equations.

2000 MSC: 65F10, 65G50, 15A12

Dedicated to Paul Van Dooren on the occasion of his 60th Birthday

1. Introduction

We study the numerical properties of some correction methods for seminormal equations for solving linear least squares problem

$$\min_x \|b - Ax\|, \tag{1}$$

where $A \in \mathbb{R}^{m\times n}$ has full column rank, $n = rank(A) \leq m$ and $b \in \mathbb{R}^m$. There exists only one solution $x_* \in \mathbb{R}^n$ to (1) and x_* satisfies the normal equations

$$A^T A x_* = A^T b. \tag{2}$$

Therefore, $x_* = A^\dagger b$, where $A^\dagger = (A^T A)^{-1} A^T$ denotes the pseudoinverse of A. There are many algorithms for solving (1) using factorization of the matrix A. For example, if we apply Householder QR decomposition: $A = QR$ where $Q \in \mathbb{R}^{m\times n}$ is left orthogonal (i.e. $Q^T Q = I_n$) and $R \in \mathbb{R}^{n\times n}$ is nonsingular upper-triangular, then the normal equations can be written as $R^T R x_* = R^T Q^T b$, so we can simply solve the system $R x_* = Q^T b$. However, if $m \gg n$ we can use only the R-factor, we do not store Q and solve the *seminormal equations* (SNE) (cf.[1])

$$R^T R x_* = A^T b. \tag{3}$$

However, numerical properties of these two methods are very different (cf.[2]–[1]). Similar approach can be applied to other factorization of A. For example, if $A = UCV^T$ where $U \in \mathbb{R}^{m\times n}$ is left orthogonal, $V \in \mathbb{R}^{n\times n}$ is orthogonal and C has a simple structure (triangular, bidiagonal, diagonal), then we can consider the class of equations

$$C^T C(V^T x_*) = V^T(A^T b). \tag{4}$$

One of the goals of SNE in (3) is that we do not need to have Q, while in (4) we do not store the matrix U but we need the matrix V. The dimension n is often much smaller than the other dimension m, so the solution of (4) still can be very efficient.

We study the numerical properties of the four algorithms algorithms summarized in Table 1. Our analysis motivated mostly by the Å. Björck's paper [1] and the paper of Å. Björck and C.C. Paige [2] (see also [4]–[6]). Through the paper we assume that the computed matrix $\tilde{C}$ in floating point arithmetic is obtained by a backward stable algorithm. It means that there exists *exactly orthogonal* matrices $\hat{U}$ and $\hat{V}$ such that

$$\hat{A} = A + \Delta A = \hat{U}\tilde{C}\hat{V}^T, \quad \|\Delta A\| \le \mathcal{O}(u)\,\|A\| \tag{5}$$

and the computed matrix $\tilde{V}$ is close to $\hat{V}$, i.e.

$$\tilde{V} = \hat{V} + \Delta V, \quad \|\Delta V\| \le \mathcal{O}(u). \tag{6}$$

It is interesting to notice that $\hat{C}^T\hat{C} = \hat{V}^T(\hat{A}^T\hat{A})\hat{V}$. Hence our computed solution $\tilde{x}$ of (4) will be often related to the solution of the perturbed system of normal equations $\hat{A}^T\hat{A}\hat{x} = \hat{A}^T b$.

Algorithm I ($\mathbf{SNE_{QR}}$) Compute the upper-triangular factor $R \in \mathbb{R}^{n\times n}$ of Householder QR decomposition of A: $A = QR$, where $Q \in \mathbb{R}^{m\times n}$ is left orthogonal. Don't store Q. Solve the seminormal equations $R^T Rx = A^T b$.
Algorithm II ($\mathbf{SNE_{SVD}}$) Find $V \in \mathbb{R}^{n\times n}$ and $\Sigma \in \mathbb{R}^{n\times n}$ of the SVD decomposition of A: $A = U\Sigma V^T$, where $U \in \mathbb{R}^{m\times n}$ is left orthogonal. Don't store U! Solve the seminormal equations $\Sigma^2(V^T x) = V^T(A^T b)$.
Algorithm III ($\mathbf{CSNE_{QR}}$) Solve $R^T Rx = A^T b$ for x by Algorithm I. Compute $r = b - Ax$. Solve $R^T R\Delta x = A^T r$ for Δx (using Algorithm I). Update $x_{new} = x + \Delta x$.
Algorithm IV ($\mathbf{CSNE_{SVD}}$) Solve $\Sigma^2(V^T x) = V^T(A^T b)$ for x by Algorithm II. Compute $r = b - Ax$. Solve $\Sigma^2(V^T \Delta x) = V^T(A^T r)$ for Δx (using Algorithm II). Update $x_{new} = x + \Delta x$.

Table 1: Algorithms description

The paper is organized as follows. Section 2 is devoted to the sensitivity of the least squares problem and Section 3 to the numerical stability of algorithms for computing the least squares solution. In Section 4 we study the numerical properties of the SNE_{SVD} method based on the SVD of A. In this paper we consider only diagonal C, where the system (4) can be solved very accurately, for the case when C is bidiagonal we refer to [7]. In Section 5 we give a roundoff error analysis of the corrected seminormal equations $CSNE_{SVD}$. Numerical experiments are given in Section 6. Through the paper we use only the 2-norm and assume the floating point arithmetic with machine precision u.

2. Perturbation analysis

We recall some important facts on the sensitivity of the solution $x_* = A^\dagger b$ to small perturbations in A and b. Let us consider first *a perturbation in the vector b*. Let $\hat{x} = A^\dagger(b + \Delta b)$ be the least-squares solution to the perturbed system $\min_x \|(b+\Delta b) - Ax\|$ and let us assume that $\|\Delta b\| \leq \epsilon \|b\|$. Thus, the difference between the solution $\hat{x}$ and the solution of the original problem $x_* = A^\dagger b$ is equal to $\hat{x} - x_* = A^\dagger \Delta b$, so

$$\frac{\|\hat{x} - x_*\|}{\|x_*\|} \leq \epsilon \kappa_b(A, b), \quad \kappa_b(A, b) = \frac{\|A^\dagger\| \|b\|}{\|x_*\|}, \tag{7}$$

where $\kappa_b(A, b)$ is *the condition number the least squares problem with respect to small perturbation in b*. Since A has full column rank we have $\|b\| \leq \|A\| \|x_*\|$, so $\kappa_b(A, b) \leq \kappa(A)$, where $\kappa(A) = \|A\| \, \|A^\dagger\|$ is the standard condition number of the matrix A. How does $\kappa_b(A, b)$ compare with $\kappa(A)$? In fact, it can be arbitrarily smaller than $\kappa(A)$ for particular b. For more details we refer to the recent paper of Dopico and Molera [8].

Now we consider the second case: *a perturbation in the matrix A*. We recall Wedin's results (cf. [11, pp. 49–51], [3, pp. 26–31], [9]) on the sensitivity of the least squares solution to small perturbations in A. In this case the *augmented system approach* is very helpful. Notice that the normal equations (2) may be written as $A^T r_* = 0$ with $r_* = b - Ax_*$. Thus we get the augmented system for r_* and x_* in the form

$$\begin{pmatrix} I & A \\ A^T & 0 \end{pmatrix} \begin{pmatrix} r_* \\ x_* \end{pmatrix} = \begin{pmatrix} b \\ 0 \end{pmatrix}. \tag{8}$$

The exact formula for the inverse of M is known and thus we have

$$\begin{pmatrix} r_* \\ x_* \end{pmatrix} = \begin{pmatrix} I - AA^\dagger & (A^\dagger)^T \\ A^\dagger & -(A^T A)^{-1} \end{pmatrix} \begin{pmatrix} b \\ 0 \end{pmatrix}. \tag{9}$$

Theorem 2.1. *Assume that $A \in \mathbb{R}^{m\times n}$ has a full column rank with $m \geq n$ and $b \in \mathbb{R}^m$. Let $\hat{x} \neq 0$ be the exact solution to (1) for slightly perturbed problem $\min_x \|b - \hat{A}x\|$, where*

$$\hat{A} = A + \Delta A, \quad \|\Delta A\| \leq \epsilon \|A\|.$$

Let $\hat{r} = b - \hat{A}\hat{x}$ and $r_* = b - Ax_*$ be the residuals for the original and the perturbed least squares problem. Assuming $\epsilon\kappa(A) < 1$ it follows that $rank(\hat{A}) = rank(A) = n$

$$\|\hat{x} - x_*\| \leq \frac{\epsilon\,\kappa(A)}{1 - \epsilon\kappa(A)} \left(\|x_*\| + \|A^\dagger\|\|r_*\|\right) + \epsilon\kappa(A)\|x_*\|, \tag{10}$$

$$\|\hat{r} - r_*\| \leq \epsilon\|A\| \left(\|x_*\| + \|A^\dagger\|\|r_*\|\right). \tag{11}$$

Remark 2.1. We can rewrite the bound (10) as follows

$$\frac{\|\hat{x} - x_*\|}{\|x_*\|} \leq 2\epsilon\,\kappa_{LS}(A, b) + \mathcal{O}((\epsilon\,\kappa(A))^2), \tag{12}$$

where

$$\kappa_{LS}(A,b) = \kappa(A)\left(1 + \kappa(A)\frac{\|r_*\|}{\|A\|\,\|x_*\|}\right) \tag{13}$$

is called the *condition number of the least squares problem* (cf.[3, p. 31]). Notice that $\kappa_{LS}(A,b) \geq \kappa(A) \geq \kappa_b(A,b)$. The ratio $\kappa(A)\frac{\|r_*\|}{\|A\|\,\|x_*\|}$ (*"incompatibility factor"*) plays an important role here. If $r_* = b - Ax_* = 0$ (i.e. the system is compatible) then $\kappa_{LS}(A,b) = \kappa(A)$. The situation is different in case when this factor is large and the term proportional to $\kappa^2(A)$ dominates in $\kappa_{LS}(A,b)$. We introduce also the *total condition number of the least squares problem* as a measure of the sensitivity of x_* to small perturbations in both A and b:

$$cond(A,b) = \kappa_{LS}(A,b) + \kappa_b(A,b). \tag{14}$$

3. Numerical stability

In this paper we study the forward stability of algorithms for computing the least squares solution. More precisely, if the approximate solution $\tilde{x}$ satisfies the bound

$$\|\tilde{x} - x_*\| \leq \ \mathcal{O}(u)\,\kappa(A)\left(\|x_*\| + \|A^\dagger\|\|r_*\|\right) \tag{15}$$

then we call $\tilde{x}$ a *forward stable solution* to the least squares problem (1). We see that we can rewrite the inequality (15) as

$$\frac{\|\tilde{x} - x_*\|}{\|x_*\|} \leq \mathcal{O}(u)\,\kappa_{LS}(A,b). \tag{16}$$

Å. Björck's proved that the seminormal equations method based on the backward stable QR factorization of A, i.e. $R^T R x_* = A^T b$, is not *forward stable* (cf. [1]). More general case was considered by Björck and C.C.Paige

(cf. [2]) for augmented systems (8). However, in [2] they indicate that if $\mathcal{O}(u)\kappa^2(A) < 1$ then one one step of iterative refinement called as the *corrected seminormal equations* method ($CSNE_{QR}$)) produces forward stable solution to (1) in sense of (15). In addition, it was noted in conclusions that the $CSNE_{QR}$ method is not forward stable for all matrices A such that $\mathcal{O}(u)\kappa(A) < 1$. Our main results on SNE_{SVD} and $CSNE_{SVD}$ in Sections 4 and 5 will be of similar type but they assume the use of the SVD decomposition instead of the QR factorization of A.

In our error analysis we obtain error bounds similar to (15) but in terms of the perturbed matrix $\hat{A} = A + \Delta A$ and the vectors $\hat{x}$ and $\hat{r}$ associated with the perturbed least squares problem with $\hat{A}$ and b. Here $\|\Delta A\| \leq \mathcal{O}(u)\|A\|$, where $\mathcal{O}(u) = c(m,n)u$ ($c(m,n)$ always means a modest constant depending on m and n here). Assume the hypothesis of Theorem 2.1 and let the computed vector $\tilde{x}$ satisfies the bound

$$\|\tilde{x} - \hat{x}\| \leq \mathcal{O}(u)\kappa(\hat{A}) \left(\|\hat{x}\| + \|\hat{A}^\dagger\| \|\hat{r}\| \right).$$

Notice that $\|\tilde{x} - x_*\| \leq \|\tilde{x} - \hat{x}\| + \|\hat{x} - x_*\|$, $\|\hat{x}\| \leq \|x_*\| + \|\hat{x} - x_*\|$ and $\|\hat{r}\| \leq \|r_*\| + \|\hat{r} - r_*\|$. Now it is natural to use the result on the sensitivity of the singular values. It is clear that under assumption $\mathcal{O}(u)\kappa(A) < 1$ we can give bounds $\|\hat{A}\| \leq \|A\| + \|\Delta A\|$ and $\|\hat{A}^\dagger\| \leq \frac{\|A^\dagger\|}{1 - \|\Delta A\| \|A^\dagger\|}$. The condition numbers are given as $\kappa(A) = \|A\|\|A^\dagger\|$ and $\kappa(\hat{A}) = \|\hat{A}\|\|\hat{A}^\dagger\|$. Then we have

$$\kappa(\hat{A}) \leq \kappa(A) \frac{1 + \mathcal{O}(u)\kappa(A)}{1 - \mathcal{O}(u)\kappa(A)}.$$

Considering somewhat stronger assumption $\mathcal{O}(u)\kappa(A) \ll 1$ and using Theorem 2.1 (with $\epsilon = \mathcal{O}(u)$) we obtain the bound in terms of the matrix A and

the vectors x_* and r_* associated with the original problem (1)

$$\|\tilde{x} - x_*\| \leq \mathcal{O}(u)\kappa(A)\left(\|x_*\| + \|A^\dagger\|\|r_*\|\right). \tag{17}$$

We can rewrite the inequality (17) into the formula for relative error

$$\frac{\|\tilde{x} - x_*\|}{\|x_*\|} \leq \mathcal{O}(u)\kappa_{LS}(A, b)$$

indicating that $\tilde{x}$ is a forward stable solution to the original problem (1).

Now we explain how $\kappa_b(\hat{A}, b)$ is related to $\kappa_b(A, b)$. We assume that $\mathcal{O}(u)\kappa_{LS}(A, b) < 1$. Since $\hat{x} = x_* + (\hat{x} - x_*)$ we can write

$$\|\hat{x}\| \geq \|x_*\| - \|\hat{x} - x_*\| \geq \|x_*\|\left(1 - \mathcal{O}(u)\kappa_{LS}(A, b)\right)$$

leading together with $\|\hat{A}\| \geq \|A\| - \|\Delta A\|$ to the final bound for $\kappa_b(\hat{A}, b)$

$$\kappa_b(\hat{A}, b) \leq \frac{\kappa_b(A, b)}{1 - \mathcal{O}(u)\kappa_{LS}(A, b)}.$$

4. Error analysis of Algorithm II

Lemma 4.1. *Let $A \in \mathbb{R}^{m\times n}$ has full column rank, $n = rank(A) \leq m$ and $b \in \mathbb{R}^m$. Assume that there exist orthogonal matrices $\hat{U}$ and $\hat{V}$ such that the computed matrices $\tilde{\Sigma}$ and $\tilde{V}$ satisfy*

$$\hat{A} = A + \Delta A = \hat{U}\tilde{\Sigma}\hat{V}^T, \quad \|\Delta A\| \leq O(u)\|A\| \tag{18}$$

and the computed matrix $\tilde{V}$ is close to $\hat{V}$, i.e.

$$\tilde{V} = \hat{V} + \Delta V, \quad \|\Delta V\| \leq \mathcal{O}(u). \tag{19}$$

Assume that $rank(\hat{A}) = n$ and $u\kappa(\hat{A}) < 1$. Let $\hat{x}$ be the exact solution to the system $\hat{A}^T\hat{A}\,\hat{x} = \hat{A}^T b$. Then the computed solution $\tilde{x}$ in floating point arithmetic by Algorithm II satisfies

$$\hat{A}^T\hat{A}\,(I + F)\tilde{x} = \hat{A}^T b + e, \tag{20}$$

where

$$\|F\| \leq \mathcal{O}(u), \quad \|e\| \leq \mathcal{O}(u)\|\hat{A}\|\|b\|. \tag{21}$$

Proof. It is easy to see that $fl(A^T b) = (A + E_1)^T b, \quad \|E_1\| \leq \mathcal{O}(u)\|A\|$ and $fl(\tilde{V}^T fl(A^T b)) = \hat{V}^T(I+F_1)(A+E_1)^T b = \hat{V}^T(A+E_2)^T b$, where $\|F_1\| \leq \mathcal{O}(u)$ and $\|E_2\| \leq \mathcal{O}(u)\|A\|$. Then there exist F_2 and F_3 such that $\tilde{\Sigma}^2\tilde{y} = (I + F_2)\hat{V}^T(A + E_2)^T b = \hat{V}^T(I + F_3)(A + E_2)^T b$, so

$$\tilde{\Sigma}^2\tilde{y} = \hat{V}^T(A + E_3)^T b, \tag{22}$$

where $\|F_i\| \leq \mathcal{O}(u)$ for $i = 2, 3$ and $\|E_3\| \leq \mathcal{O}(u)\|A\|$. We see that $\tilde{x} = fl(\tilde{V}\tilde{y}) = \hat{V}(I + F_4)\tilde{y}, \quad \|F_4\| \leq \mathcal{O}(u)$. ¿From this we obtain

$$\tilde{y} = \hat{V}^T(I + F_4)^{-1}\tilde{x} = \hat{V}^T(I + F)\tilde{x}, \quad \|F\| \leq \mathcal{O}(u). \tag{23}$$

This together with (22) gives $\tilde{\Sigma}^2\hat{V}^T(I + F)\tilde{x} = \hat{V}^T(A + E_3)^T b$, which we rewrite as $\hat{V}\tilde{\Sigma}^2\hat{V}^T(I + F)\tilde{x} = (A + E_3)^T b$. Now we use the equality $\hat{A}^T\hat{A} = \hat{V}\tilde{\Sigma}^2\hat{V}^T$ and we see that slightly perturbed $\tilde{x}$ satisfies $\hat{A}^T\hat{A}\,(I + F)\tilde{x} = (A+E_3)^T b$, which can be written in the form of (20) with $e = (E_3-\Delta A)^T b$. Clearly, $\|e\| \leq (\|E_3\| + \|\Delta A\|)\|b\| \leq \mathcal{O}(u)\|A\|\|b\|$. Since $\|A\| \leq [1 + \mathcal{O}(u)]\|\hat{A}\|$, the bounds (21) follow immediately. □

Theorem 4.1. *Assuming the hypothesis of Lemma 4.1 we have*

$$\tilde{x} - \hat{x} = (\hat{A}^T\hat{A})^{-1}e - F\tilde{x}, \quad \hat{x} = \hat{A}^\dagger b, \tag{24}$$

with the following two bounds

$$\|\tilde{x} - \hat{x}\| \leq \mathcal{O}(u)\kappa(\hat{A})\,\|\hat{A}^\dagger\|\|b\| \tag{25}$$

and

$$\|A(\tilde{x} - \hat{x})\| \leq \mathcal{O}(u)\kappa(\hat{A})\,\|b\| + \mathcal{O}(u)\|\hat{A}\|\|\tilde{x}\|. \tag{26}$$

Proof. It is clear from (20) that $\tilde{x} - \hat{x}$ satisfies (24) with the bound $\|\tilde{x} - \hat{x}\| \leq \|(\hat{A}^T\hat{A})^{-1}\|\|e\| + \|F\|\|\tilde{x}\|$. From (23) we have $\|F\| \leq \mathcal{O}(u)$. Using $\tilde{x} = (\tilde{x} - \hat{x}) + \hat{x}$, we have

$$\|\tilde{x} - \hat{x}\| \leq \|(\hat{A}^T\hat{A})^{-1}\|\|e\| + \mathcal{O}(u)\|\hat{x}\|. \tag{27}$$

Since $\hat{x} = \hat{A}^\dagger b$, we get $\|\hat{x}\| \leq \|\hat{A}^\dagger\|\|b\|$. It is clear that $\|(\hat{A}^T\hat{A})^{-1}\| = \|\hat{A}^\dagger\|^2$. ¿From (21), (27) and the inequality $\kappa(\hat{A}) \geq 1$ we obtain (25). To prove (26) we multiply the equation (24) by $\hat{A}$ and get $A(\tilde{x} - \hat{x}) = (\hat{A}^\dagger)^T e - \hat{A}F\tilde{x}$. From this it follows that

$$\|A(\tilde{x} - \hat{x})\| \leq \|\hat{A}^\dagger\|\|e\| + \mathcal{O}(u)\|\hat{A}\|\|\tilde{x}\|.$$

This together with (21) gives (25). □

Remark 4.1. Note that $\kappa_b(\hat{A}, b) = \frac{\|\hat{A}^\dagger\|\|b\|}{\|\hat{x}\|}$ is the condition number of $\hat{x} = \hat{A}^\dagger b$ to small perturbation in b. If $\kappa_b(\hat{A}, b)$ is of order unity, then $\frac{\|\tilde{x}-\hat{x}\|}{\|\hat{x}\|} \leq \mathcal{O}(u)\,\kappa(\hat{A})$ and in this case Algorithm II produces a forward stable solution to the problem (1).

5. Error analysis of Algorithm IV

Theorem 5.1. *Let $\tilde{x}$, $\tilde{r}$, $\Delta\tilde{x}$ and $\tilde{x}_{new}$ be computed by Algorithm IV in floating point arithmetic. Under the hypothesis of Lemma 4.1 and assuming that $\mathcal{O}(u)\kappa(\hat{A})\kappa_b(\hat{A}, b) < 1$ we have*

$$\frac{\|\tilde{x}_{new} - \hat{x}\|}{\|\hat{x}\|} \leq \mathcal{O}(u)\left(\kappa(\hat{A}) + \kappa_b(\hat{A}, b) + \kappa^2(\hat{A})\frac{\|b - \hat{A}\hat{x}\|}{\|\hat{A}\|\|\hat{x}\|}\right). \tag{28}$$

Proof. We have $\tilde{x}_{new} = (I+D)(\tilde{x}+\Delta\tilde{x})$ with $\|D\| \leq u$. This gives $\tilde{x}_{new} - \hat{x} = (I+D)((\tilde{x}+\Delta\tilde{x}) - \hat{x}) + D\hat{x}$. Hence we get the following bound

$$\|\tilde{x}_{new} - \hat{x}\| \leq (1+u)\|(\tilde{x}+\Delta\tilde{x}) - \hat{x}\| + u\|\hat{x}\|. \tag{29}$$

All we need is to bound the term $\|(\tilde{x}+\Delta\tilde{x}) - \hat{x}\|$. By the same approach as in Lemma 4.1 and Theorem 4.1 we obtain

$$\Delta\tilde{x} = \hat{A}^{\dagger}\tilde{r} + (\hat{A}^T\hat{A})^{-1}g - G\Delta\tilde{x}, \tag{30}$$

$$\|G\| \leq \mathcal{O}(u), \quad \|g\| \leq \mathcal{O}(u)\|\hat{A}\|\|\tilde{r}\|. \tag{31}$$

The computed residual $\tilde{r} = b - \hat{A}\tilde{x} + f$ satisfies the identity

$$\tilde{r} = \hat{r} - \hat{A}(\tilde{x} - \hat{x}) + f, \quad \|f\| \leq \mathcal{O}(u)(\|b\| + \|\hat{A}\|\|\tilde{x}\|), \tag{32}$$

where $\hat{r} = b - \hat{A}\hat{x}$. Due to $\hat{A}^{\dagger}\hat{A} = I$ and $\hat{A}^{\dagger}\hat{r} = 0$ it follows from (32) that $\hat{A}^{\dagger}\tilde{r} = -(\tilde{x} - \hat{x}) + \hat{A}^{\dagger}f$ and so (30) can be rewritten as $\Delta\tilde{x} = -(\tilde{x} - \hat{x}) + \hat{A}^{\dagger}f + (\hat{A}^T\hat{A})^{-1}g - G\Delta\tilde{x}$. Thus, $(\tilde{x}+\Delta\tilde{x}) - \hat{x} = \hat{A}^{\dagger}f + (\hat{A}^T\hat{A})^{-1}g - G\Delta\tilde{x}$. This together with the assumption $\|G\| \leq \mathcal{O}(u)$ leads to

$$\|(\tilde{x}+\Delta\tilde{x}) - \hat{x}\| \leq \|\hat{A}^{\dagger}f\| + \|(\hat{A}^T\hat{A})^{-1}g\| + \mathcal{O}(u)\|\Delta\tilde{x}\|. \tag{33}$$

In order to bound $\|\Delta\tilde{x}\|$, we write $\|\Delta\tilde{x}\| \leq \|\tilde{x}+\Delta\tilde{x}\| + \|\tilde{x}\|$. On the other hand, $\tilde{x}+\Delta\tilde{x} = (I+D)^{-1}\tilde{x}_{new}$, so $\|\tilde{x}+\Delta\tilde{x}\| \leq \frac{1}{1-u}\|\tilde{x}_{new}\|$, hence $u\|\Delta\tilde{x}\| \leq \mathcal{O}(u)\|\tilde{x}_{new}\| + u\|\tilde{x}\|$. Then (33) can be further bounded as

$$\|(\tilde{x}+\Delta\tilde{x}) - \hat{x}\| \leq \|\hat{A}^{\dagger}f\| + \|(\hat{A}^T\hat{A})^{-1}g\| + \mathcal{O}(u)\|\tilde{x}_{new}\| + u\|\tilde{x}\|. \tag{34}$$

First we consider $\|\hat{A}^{\dagger}f\|$. It follows from (32) that

$$\|\hat{A}^{\dagger}f\| \leq \|\hat{A}^{\dagger}\|\|f\| \leq \mathcal{O}(u)\|\hat{A}^{\dagger}\|\|b\| + \mathcal{O}(u)\kappa(\hat{A})\|\tilde{x}\|. \tag{35}$$

Now we want to bound $\|(\hat{A}^T\hat{A})^{-1}g\|$. Notice that from (31) we have $\|(\hat{A}^T\hat{A})^{-1}g\| \leq \|\hat{A}^\dagger\|^2\|g\| \leq \mathcal{O}(u)\kappa(\hat{A})\|\hat{A}^\dagger\|\|\tilde{r}\|$. Due to (32) we get $\|\tilde{r}\| \leq \|\hat{r}\| + \|\hat{A}(\tilde{x} - \hat{x})\| + \|f\|$. Taking into account (26), (32) and $\kappa(\hat{A}) \geq 1$ we get $\|\tilde{r}\| \leq \|\hat{r}\| + \mathcal{O}(u)\kappa(\hat{A})\|b\| + \mathcal{O}(u)\|\hat{A}\|\|\tilde{x}\|$. We conclude that

$$\|(\hat{A}^T\hat{A})^{-1}g\| \leq \mathcal{O}(u)\kappa(\hat{A})\|\hat{A}^\dagger\|\|\hat{r}\| + \mathcal{O}(u)\kappa(\hat{A})\|\hat{A}^\dagger\|[\|\hat{A}(\tilde{x} - \hat{x})\| + \|f\|]. \quad (36)$$

Since we have $\|\tilde{x}\| \leq \|\hat{x}\| + \|\tilde{x} - \hat{x}\| \leq \|\hat{x}\| + \mathcal{O}(u)\kappa(\hat{A})\|\hat{A}^\dagger\|\|b\|$ and $\|\tilde{x}_{new}\| \leq \|\hat{x}\| + \|\tilde{x}_{new} - \hat{x}\|$, the bounds (29) together with (34), (35) and (36) give then the statement (28).

Remark 5.1. Notice that $\kappa_b(\hat{A}, b) \leq \kappa(\hat{A})$, so (28) can be read as follows

$$\frac{\|\tilde{x}_{new} - \hat{x}\|}{\|\hat{x}\|} \leq \mathcal{O}(u)\left(\kappa(\hat{A}) + \kappa^2(\hat{A})\frac{\|b - \hat{A}\hat{x}\|}{\|\hat{A}\|\|\hat{x}\|}\right). \quad (37)$$

Thus, $\tilde{x}_{new}$ is a *forward stable solution* to the least squares problem $\min_x \|b - \hat{A}x\|$, and under reasonable conditions also to the original least squares problem (1), see the discussion in Section 3.

□

6. Numerical experiments

In this section we illustrate our theoretical results. All experiments are performed using MATLAB with unit roundoff $u = 1.1 \cdot 10^{16}$. We assume two extreme cases, where the least squares solution x_* is equal to right singular vector corresponding to the smallest or to the largest singular value of the matrix A. As it will be clear from results, the correction step is important especially for problems with solution close (equal) to the right singular vectors corresponding to the largest singular values. The first problem with

dimensions $m = 20$ and $n = 7$ is defined by the singular value decomposition of the matrix $A = U\Sigma V^T$, where U and V are orthogonal matrices with corresponding dimensions generated by the *orthog* subroutine (we consider only the first n columns for matrix U). The matrix Σ is a diagonal matrix with singular values given as $\sigma_i(A) = 10^{6-1.5i}$ for $i = 1, \ldots, n$. It is clear then that $\kappa(A) = 10^9$. Thus we consider two problems with the solutions x_1 and x_2 given by two columns in the matrix V corresponding to two extremal singular values, respectively. This leads to the right hand side vectors $b_1 = Ax_1$ and $b_2 = Ax_2$. Finally we take vector h as the scaled $n + 1$-th column of the orthogonal matrix that generates the matrix U with $\|h\| = \sigma_n(A)$. It is clear then that $A^T h = 0$ and $\kappa_{LS}(A, b) \approx \kappa(A)$. The second set of problems is defined in a similar way. The only difference is in the $m = 10000$, $n = 500$ and in the definition of the matrix Σ, where singular values are given as $\sigma_i(A) = 10^{4.5-9(i-1)/499}$ for $i = 1, \ldots, 500$. Tables 2–5 summarize our numerical results. The first rows correspond to the case of the over-determined systems $Ax_1 = b_1$ and $Ax_2 = b_2$ with the classical solutions x_1 and x_2. In the subsequent rows we have increased the residual norm of the last squares problem via the appropriate scaling of the vector h. The norms of relative errors are scaled by $cond(A, b)$ and thus correspond to the theoretical bound (28). We see that SNE_{SVD} is quite satisfactory enough if the solution x_1 is equal to the right singular vector corresponding to the smallest singular value. This is no longer true for the solution x_2 equal to the right singular vector corresponding to the norm of A. In this case the refinement step in the CSNE_{SVD} method can significantly improve the accuracy of the computed solution. In addition our problems still meet the assumptions of Theorem

b	$cond(A,b)$	$\kappa_b(A,b)$	$\frac{\|\tilde{x}-x_1\|}{\|x_1\|*cond(A,b)}$	$\frac{\|\tilde{x}_{new}-x_1\|}{\|x_1\|*cond(A,b)}$
b_1	2.0000e+09	1.0000e+09	5.9605e-08	3.3112e-15
$b_1 + h$	3.0000e+09	1.0000e+09	3.9737e-08	1.1983e-15
$b_1 + 10^1 \cdot h$	1.2000e+10	1.0000e+09	9.9341e-09	4.6631e-16
$b_1 + 10^2 \cdot h$	1.0200e+11	1.0000e+09	1.1687e-09	9.5519e-17
$b_1 + 10^3 \cdot h$	1.0020e+12	1.0000e+09	1.1897e-10	3.4409e-17
$b_1 + 10^4 \cdot h$	1.0002e+13	1.0000e+09	1.7878e-11	2.5554e-17
$b_1 + 10^5 \cdot h$	1.0000e+14	1.0000e+09	1.1921e-12	4.5836e-17
$b_1 + 10^6 \cdot h$	1.0000e+15	1.0000e+09	1.1921e-13	3.7842e-17
$b_1 + 10^7 \cdot h$	1.0000e+16	1.0000e+09	1.1921e-14	5.1597e-17

Table 2: Results for the least squares problem with $m = 20$ and $n = 7$.

5.1 and so the method computes forward stable approximate solutions.

7. Conclusions

In this paper we have considered two methods for solution the least squares problems which are based only on the factors Σ and V from the SVD decomposition of the matrix A. We have shown that while SNE_{SVD} method based on the solution of the system (4) is not forward stable, one step of fixed precision iterative refinement in $CSNE_{SVD}$ improves the accuracy of the computed approximate solution. We have shown that under the condition $\mathcal{O}(u)\kappa(A)\kappa_b(A,b) < 1$ this method produces a forward stable solution, while it is not difficult to construct a numerically nonsingular example (satisfying $\mathcal{O}(u)\kappa(A) < 1$) for which the $CSNE_{SVD}$ method fails in computing the forward stable solution.

b	$cond(A,b)$	$\kappa_b(A,b)$	$\frac{\|\tilde{x}-x_2\|}{\|x_2\|*cond(A,b)}$	$\frac{\|\tilde{x}_{new}-x_2\|}{\|x_2\|*cond(A,b)}$
b_2	1.0000e+09	1.0000e+00	4.6511e-17	1.5901e-18
$b_2 + h$	2.0000e+09	1.4142e+00	5.0153e-17	1.5554e-17
$b_2 + 10^1 \cdot h$	1.1000e+10	1.0050e+01	1.9985e-17	3.3175e-17
$b_2 + 10^2 \cdot h$	1.0100e+11	1.0000e+02	1.4897e-17	3.4555e-17
$b_2 + 10^3 \cdot h$	1.0010e+12	1.0000e+03	2.1718e-17	9.4127e-18
$b_2 + 10^4 \cdot h$	1.0001e+13	1.0000e+04	4.6702e-17	2.5224e-17
$b_2 + 10^5 \cdot h$	1.0000e+14	1.0000e+05	1.1828e-17	2.4759e-17
$b_2 + 10^6 \cdot h$	1.0000e+15	1.0000e+06	3.2526e-17	7.1239e-18
$b_2 + 10^7 \cdot h$	1.0000e+16	1.0000e+07	4.2960e-17	4.2055e-17

Table 3: Results for the least squares problem with $m = 20$ and $n = 7$.

b	$cond(A,b)$	$\kappa_b(A,b)$	$\frac{\|\tilde{x}-x_1\|}{\|x_1\|*cond(A,b)}$	$\frac{\|\tilde{x}-x_1\|}{\|x_1\|*cond(A,b)}$
b_1	2.0000e+09	1.0000e+09	1.3146e-07	4.2228e-15
$b_1 + h$	3.0000e+09	1.0000e+09	7.4301e-08	3.2382e-15
$b_1 + 10^1 \cdot h$	1.2000e+10	1.0000e+09	1.4348e-08	6.9681e-16
$b_1 + 10^2 \cdot h$	1.0200e+11	1.0000e+09	2.5637e-09	9.4715e-17
$b_1 + 10^3 \cdot h$	1.0020e+12	1.0000e+09	1.9640e-10	9.2667e-18
$b_1 + 10^4 \cdot h$	1.0002e+13	1.0000e+09	3.0798e-11	6.2056e-18
$b_1 + 10^5 \cdot h$	1.0000e+14	1.0000e+09	2.0705e-12	2.2348e-17
$b_1 + 10^6 \cdot h$	1.0000e+15	1.0000e+09	2.0001e-13	1.0037e-17
$b_1 + 10^7 \cdot h$	1.0000e+16	1.0000e+09	2.2046e-14	1.6284e-18

Table 4: Results for the least squares problem with $m = 10000$ and $n = 500$.

b	$cond(A,b)$	$\kappa_b(A,b)$	$\frac{\|\tilde{x}-x_2\|}{\|x_2\|*cond(A,b)}$	$\frac{\|\tilde{x}_{new}-x_2\|}{\|x_2\|*cond(A,b)}$
b_2	1.0000e+09	1.0000e+00	2.1639e-17	4.4147e-19
$b_2 + h$	2.0000e+09	1.4142e+00	1.1273e-17	3.0412e-18
$b_2 + 10^1 \cdot h$	1.1000e+10	1.0050e+01	4.1139e-18	8.0375e-18
$b_2 + 10^2 \cdot h$	1.0100e+11	1.0001e+02	4.9674e-18	5.2661e-18
$b_2 + 10^3 \cdot h$	1.0010e+12	1.0000e+03	1.3189e-17	8.7706e-18
$b_2 + 10^4 \cdot h$	1.0001e+13	1.0000e+04	1.2039e-17	1.3397e-17
$b_2 + 10^5 \cdot h$	1.0000e+14	1.0000e+05	4.7011e-18	4.4827e-18
$b_2 + 10^6 \cdot h$	1.0000e+15	1.0000e+06	3.4720e-18	7.7549e-18
$b_2 + 10^7 \cdot h$	1.0000e+16	1.0000e+07	7.5257e-18	6.8175e-18

Table 5: Results for the least squares problem with $m = 10000$ and $n = 500$.

References

[1] Å. Björck, *Stability analysis of the method of seminormal equations for linear least squares problems*, Linear Algebra Appl. 88/89 (1987) 31–48.

[2] Å. Björck and C.C. Paige, *Solution of augmented linear systems using orthogonal factorizations*, BIT 34 (1994) 1–24.

[3] Å. Björck, *Numerical Methods for Least Squares Problems*, SIAM, Philadelphia, PA, USA (1996).

[4] Å. Björck, Iterative refinement of linear least squares solutions I, BIT 7 (1967) 257–278.

[5] Å. Björck, Iterative refinement of linear least squares solutions II, BIT 8 (1968) 8–30.

[6] Å. Björck, Iterative refinement and reliable computing. In Reliable Nu-

merical Computation, M. G. Cox and S. J. Hammarling, editors, Oxford University Press (1990) 249–266.

[7] R. Byers and H. Xu, *A new scaling for Newton's iteration for the polar decomposition and its backward stability*, SIAM Journal on Matrix Analysis and Applications 30 (2008), 822–843.

[8] F. M. Dopico and J. M. Molera, *Accurate solution on structured linear systems via rank-revealing decompositions*, published online in IMA Journal of Numerical Analysis (2011) (doi: 10.1093/imanum/drr023).

[9] N. J. Higham, *Accuracy and Stability of Numerical Algorithms*, Second Edition, SIAM, Philadelphia 2002.

[10] S. J. Leon, Å. Björck, and W. Gander, *Gram-Schmidt orthogonalization: 100 years and more*, a manuscript (2011).

[11] C. L. Lawson and R. J. Hanson, *Solving least squares problems*, Prentice-Hall Inc., Englewood Cliffs, N.J. (1974).

[12] J. H. Wilkinson, *The algebraic eigenvalue problem*, Oxford University Press (1965).