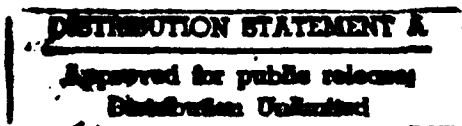DTIC
SELECTE
SEP.2 5 1992
S D

Perturbation Theory for the Solution
of Systems of Linear Equations

Shivkumar Chandrasekaran, Ilse Ipsen

Research Report YALEU/DCS/RR-866
October 1991

YALE UNIVERSITY
DEPARTMENT OF COMPUTER SCIENCE

Abstract.

We present expressions for absolute and relative errors in individual components of the solution to systems of linear equations. We consider three kinds of linear systems: non-singular, underdetermined of full row rank, and least squares of full column rank. No assumptions regarding the structure or distribution of the perturbations are required.

Our expressions for component-wise *relative* errors allow the following conclusions: For *any* linear system there is at least one solution component whose sensitivity to perturbations is proportional to the condition number of the matrix; but – depending on the relation between right-hand side and matrix – there may exist components that are much better conditioned. For a least squares problem, the sensitivity of the components also depends on the right-hand side and may be as high as the square of the condition number. Least squares problems are therefore always more receptive to ill-conditioning than linear systems.

In addition, we show that the component-wise relative errors for linear systems are reduced by column scaling only if column scaling manages to reduce the perturbations. Regarding underdetermined linear systems of full column rank, the problem of finding the minimal-norm solution can be formulated so that the same analysis as for least squares problems is applicable here as well.

Finally, we define component-wise condition numbers that measure the sensitivity of the solution components to perturbations. They have simple geometric interpretations and can be computed and estimated as efficiently as the conventional condition numbers.

Perturbation Theory for the Solution
of Systems of Linear Equations

Shivkumar Chandrasekaran, Ilse Ipsen

92   9 24 021

92-25797

1 Introduction

Most people would probably believe that there is nothing left to be done when it comes to error analysis for the solution of linear systems of equations and linear least squares problems, especially where perturbation analysis without regard to a particular algorithm is concerned. So, why yet another paper on the subject?

We want to demonstrate that a careful perturbation analysis is capable of providing a realistic assessment of the error and reliable measures of the sensitivity of the solution to perturbations in the data.

In particular, we derive expressions for the errors in *individual* components of the solution vector. These expressions give rise to realistic and efficiently computable error bounds. The derivations of the error expressions require no restrictions on the structure or distribution of the perturbations. Without any knowledge of the underlying algorithm, we can therefore obtain a great deal of information about the sensitivity of individual solution components to perturbations in the data – much more, in fact, than what is provided by conventional perturbation results.

1.1 Motivation

Consider the solution of a system of linear equations $Ax = b$ with non-singular coefficient matrix A. The *computed solution* $\bar{x}$, which is usually different from the true solution x, can be viewed as the true solution to a *perturbed* system $(A + F)\bar{x} = b + f$. Let's assume we do not know which algorithm was used for the computation of $\bar{x}$, so we have no knowledge about the structure of the perturbations F and f.

Only very infrequently, e.g. [4, 15], does one try to assess the error in *individual* solution components. The conventional way of assessing the error in $\bar{x}$, as compared to the true solution x, is to estimate an upper bound on the *norm*-based[1] relative error $\|\bar{x} - x\|/\|x\|$. The most commonly used first-order bound is

$$\frac{\|\bar{x} - x\|}{\|x\|} \lesssim \kappa(A)(\rho_A + \rho_b),$$

where the condition number $\kappa(A) = \|A\| \|A^{-1}\| \geq 1$ acts as an amplifier for the relative perturbations in the data $\rho_A = \|F\|/\|A\|$ and $\rho_b = \|f\|/\|b\|$. This norm-based bound has led to a rule of thumb: If, for instance, $\kappa(A)$ is about 10^3, and the size of the relative perturbations is about 10^{-7}, then the computed solution $\bar{x}$ can be expected to be accurate to $7 - 3 = 4$ significant digits.

In many situations this type of error assessment is just fine – unless, however, the individual components of the solution have physical significance as, for example, in statistical applications [21]. Consider the linear system $Ax = b$, where

$$A = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \qquad b = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \qquad x = \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

Suppose the computed solution is $\bar{x} = \begin{pmatrix} 1 \\ \epsilon \end{pmatrix}$, where ϵ is a very small positive number. Then $\bar{x}$ can be viewed as the true solution to the perturbed system

$$A + F = A = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \qquad b + f = \begin{pmatrix} 1 \\ \epsilon \end{pmatrix}.$$

---

[1] The following inequalities hold for any vector p-norm and induced matrix norm; see Section 2 in [12], for instance. In this paper we use the two-norm.

DTIC QUALITY INSPECTED 3

1

Because A is the identity matrix, $\kappa(A) = 1$, and the above error bound tells us that $\|\bar{x} - x\|/\|x\| \lesssim \epsilon$. So the error in the solution seems to be no more than the error in the data, which is all we are entitled to. However, the second component of the computed solution has component-wise relative error[2]

$$\frac{\bar{x}_2 - x_2}{\bar{x}_2} = \frac{\epsilon - 0}{\epsilon} = 1,$$

and is thus totally wrong. Therefore, a small bound on the norm-based error does not guarantee accuracy in *individual* components of the computed solution.

Of course, you could argue now that this should have been anticipated. Since x_2 is zero, hence small in magnitude, one should not expect to compute it correctly in the first place. Accordingly, we could account for it by estimating the error in each component $\bar{x}_i$ of the computed solution via

$$\frac{|\bar{x}_i - x_i|}{|x_i|} \leq \frac{\|\bar{x} - x\|}{|x_i|} \lesssim \kappa(A) \frac{\|x\|}{|x_i|} (\rho_A + \rho_b),$$

provided $x_i \neq 0$. The amplifiers for the relative perturbations are now the condition number, as well as the size of an individual component relative to the whole solution. This modification yields a correct assessment for the errors in individual solution components of the above example.

Unfortunately, we have not really fixed the problem. The condition number $\kappa(A)$ can still severely over estimate the error in some solution components, as the following 4 × 4 linear system demonstrates.

$$A = \begin{pmatrix} 0.4919 & 0.1112 & -0.6234 & -0.6228 \\ -0.5050 & -0.6239 & 0.0589 & 0.0595 \\ 0.5728 & -0.0843 & 0.7480 & 0.7483 \\ -0.4181 & 0.7689 & 0.2200 & 0.2204 \end{pmatrix}, \quad b = \begin{pmatrix} 0.4351 \\ -0.1929 \\ 0.6165 \\ -0.8022 \end{pmatrix}.$$

The first three columns of A are nearly orthogonal while the last two columns are almost identical. Both the two-norm condition number $\kappa_2(A)$ and Skeel's condition number [19] are larger than 10^3. Note that the matrix is not ill-scaled.

But the 'component-wise condition numbers' that we will introduce in this paper turn out to be

$$< 1.1, \quad < 1.1, \quad > 10^3, \quad > 10^3.$$

This means that the first two components of x are well-conditioned and the remaining two are ill-conditioned, regardless of the perturbations. To illustrate this, compare the 'exact' solution computed with 16-digit arithmetic

$$x^T = (\ 1.000075414240576 \quad -.5000879795933286 \quad -.0242511388797165 \quad .02624513955005858\ ),$$

with the solution computed with 4-digit arithmetic, which can be viewed as the solution to a perturbed problem,

$$\bar{x}^T = (\ 1.000 \quad -.5003 \quad -.0589 \quad .06090\ ).$$

As predicted by our component-wise condition numbers, the first two components are accurate to almost four digits, whereas the last two have no accuracy whatsoever. As far as we know no other existing condition numbers can predict the well-conditioning of the first two components of this system.

Therefore, the conventional norm-based bounds are apparently not able to estimate the accuracy of individual components correctly. We hope to have now provided enough motivation for the need to study *component-wise* relative errors and the sensitivity to perturbations of *individual* solution components.

---

[2] Whenever $x_i = 0$ while $\bar{x}_i \neq 0$, the component-wise relative error has $\bar{x}_i$ instead of x_i in the denominator.

1.2 Overview

Given a linear system $Ax = b$ of full column rank and a perturbed system $(A + F)\bar{x} = b + f$, we derive expressions for the error in individual components of the computed solution $\bar{x}$. Our work is more general than that of Skeel [19] on component-wise perturbations and that of Stewart [22] on stochastic perturbations because we make no assumptions about the perturbations F and f, either their size, structure or distribution.

In particular, we show that there is always one component of the solution vector whose sensitivity to relative perturbations is proportional to the condition number of the matrix; but – depending on the right-hand side – there may exist components that are much better conditioned. Therefore the conventional upper bounds on norm-wise relative errors are as tight as possible, and if they are pessimistic it is because they represent an inadequate means of measuring the error.

We *derive* condition numbers for individual components for the solution of a linear system, which we call 'component-wise condition numbers'. We thus associate with a linear system $Ax = b$ not a single condition number but a *set* of condition numbers. Our work, although developed independently, can therefore be considered a continuation of Stewart's work on collinearity in regression problems [21]. The singular value decomposition, often used to determine the conventional condition number of a matrix, provides a basis for the column space but does not relate this basis to the columns of the matrix. In contrast, Stewart's condition numbers are designed to expose the most linearly dependent columns of a matrix. They are embedded in our component-wise condition numbers, whose purpose is not only to recognise linearly dependent columns but also to reflect the relationship between matrix and right-hand side. We provide a geometric interpretation for Stewart's condition numbers and demonstrate that they are 'inherent' in the inverse of the matrix.

All of our results also hold for the solution of linear least squares problems $\min_y \|Ay - b\|$ of full column rank. The set of component-wise condition numbers for a least squares problem contains those for a linear system as a subset, hence the sensitivity of some solution components may be much lower than the condition number. In particular, we show that there is a component of the solution vector whose sensitivity to relative perturbations equals at least the product of condition number and $\tan \theta$, where θ is the angle between the right-hand side and the column space of the matrix; the sensitivity can be as high as the product of $\tan \theta$ and the square of the condition number. Least squares problems are therefore always more receptive to ill-conditioning than linear systems.

In addition, we show that the component-wise relative errors for linear systems are reduced by column scaling only if column scaling manages to reduce the perturbations. Regarding underdetermined linear systems of full column rank, the problem of finding the minimal-norm solution can be formulated so that the same analysis as for least squares problems is applicable.

The expressions for the errors in the solution of least squares problems and underdetermined linear systems can be used, for instance, to obtain perturbation results for the computation of left and right inverses of matrix.

In Section 2 we present the basic ideas contained in this paper. We derive them from first principles, keeping technical details to a minimum. Section 3 and Appendix 2 contain a detailed perturbation theory for the solution of linear systems of full column rank, and Section 4 extends it to the solution of least squares problems of full column rank. The treatment of full rank least squares problems is extended to the solution of underdetermined linear systems of full row rank in Section 5. In Section 6 we discuss the efficient computation and estimation of component-wise condition numbers. In particular, we show how to compute them via updating QR decompositions, and how to estimate them by means of conventional condition numbers estimators. A short summary

of the paper is followed by Appendix 1, where expressions for the left-inverse of a matrix are derived in order to justify our choice of condition numbers as a natural measure of sensitivity.

Although we concentrate on component-wise *relative* errors, expressions for component-wise *absolute* errors are also included; the corresponding condition numbers can be computed as easily as those for conventional norm-based errors.

1.3 Summary of Notation

We give a brief summary of frequently used notation for easy reference. This notation is also introduced in the body of the paper whenever it appears for the first time.

The norm $\| \cdot \|$ represents the two-norm, and e_i stands for the ith column of the identity matrix I, whose order will be clear from the context. The column space of a matrix A, $\{c : Ax = c\}$, is represented by $\mathcal{R}(A)$ and its nullspace, $\{x : Ax = 0\}$ by $\mathrm{Ker}(A)$. The subspace in real n-space $\Re^n$ that is orthogonal to the space span$\{v_1, \ldots, v_k\}$ spanned by $n \times 1$ vectors $v_1, \ldots, v_k$ is denoted by $\mathrm{span}^\perp\{v_1, \ldots, v_k\}$.

The columns of a $n \times m$ matrix A are denoted by a_i, and if A is of rank m the rows of its left-inverse $A^\dagger$ are denoted by r_i^T,

$$A = (\, a_1 \ \ \ldots \ \ a_m \,), \qquad A^\dagger = \begin{pmatrix} r_1^T \\ \vdots \\ r_m^T \end{pmatrix}.$$

The singular value decomposition (SVD) of a $n \times m$ matrix A, $n \geq m$, is represented as $A = U\Sigma V^T$, where U is a $n \times n$ orthogonal matrix, V is a $m \times m$ orthogonal matrix, and the $m \times n$ diagonal matrix Σ has as its diagonal elements the singular values of A in descending order $\sigma_1 \geq \ldots \geq \sigma_m \geq 0$. The two-norm condition number of a full-rank matrix A is denoted by $\kappa(A) = \|A\|\,\|A^\dagger\|$.

If x solves the least squares problem $\min_y \|Ay - b\|$ then the residual is denoted by $r = b - Ax$.

2 The Basic Ideas

We start out by illustrating the ideas that led us to pursue a component-wise perturbation analysis; this is done by studying perturbations in the right-hand side only. We also restrict ourselves to the solution of full-rank least squares problems until Section 5 where the results are extended to the solution of underdetermined linear systems of full row rank.

As for notation, $\| \cdot \|$ represents the two-norm, and e_i stands for the ith column of the identity matrix I.

2.1 Motivation

The first theorem gives a simple geometric interpretation of the components of the solution x to a full-rank least squares problem $\min_y \|Ay - b\|$. An individual solution component can be expressed

as a product of three factors: the length of a row in the left-inverse $A^\dagger$, the length of the right-hand side and the angle between the two.

Theorem 1 *Given a $n \times m$ matrix A of rank m, denote by r_i^T the rows of its left-inverse $A^\dagger$,*

$$A^\dagger = (A^T A)^{-1} A^T = \begin{pmatrix} r_1^T \\ \vdots \\ r_m^T \end{pmatrix}.$$

Then the components x_i of the solution x to the least squares problem $\min_y \|Ay - b\|$ are given by

$$x_i = r_i^T b = \|r_i\| \, \|b\| \cos \beta_i, \qquad 1 \le i \le m,$$

where β_i is the angle between r_i and b.

Proof: The vector x solves $\min_y \|Ay - b\|$ if and if only it solves the normal equations $A^T A x = A^T b$. So $x = A^\dagger b$, which implies $x_i = r_i^T b = \|r_i\| \, \|b\| \cos \beta_i$, where β_i is the angle between r_i and b. ∎

Already in [20] Stewart recognised the importance of the $\|r_i\|$ for the purpose of detecting almost linearly dependent columns in A. In fact, it turns out that length and angles associated with the r_i indicate the sensitivity of *individual* components of the solution x to perturbations in the right-hand side.

Theorem 2 *Given a matrix A of full column rank, let $x \ne 0$ solve $\min_y \|Ay - b\|$ and let $\bar{x}$ solve $\min_y \|Ay - (b + f)\|$.*

Denote by ψ_i the angle between r_i and f. Then

$$\bar{x}_i = x_i + r_i^T f = x_i + \|r_i\| \, \|f\| \cos \psi_i.$$

If $x_i \ne 0$ and $\epsilon_b = \|f\|/\|b\|$ then

$$\begin{aligned}
\frac{\bar{x}_i - x_i}{x_i} &= \frac{1}{\cos \beta_i} \epsilon_b \cos \psi_i \\
&= \frac{\|b\|}{\|A\| \, \|x\|} \frac{\|x\|}{x_i} \|A\| \, \|r_i\| \, \epsilon_b \cos \psi_i.
\end{aligned}$$

Proof: According to Theorem 1,

$$\bar{x}_i = r_i^T (b + f) = r_i^T b + r_i^T f = x_i + r_i^T f = x_i + \|r_i\| \, \|f\| \cos \psi_i,$$

where ψ_i is the angle between r_i and f. Since $0 \ne x_i = r_i^T b = \|r_i\| \, \|b\| \cos \beta_i$ we have

$$\frac{\bar{x}_i - x_i}{x_i} = \frac{r_i^T f}{r_i^T b} = \frac{1}{\cos \beta_i} \frac{\|f\|}{\|b\|} \cos \psi_i.$$

∎

The theorem states that the absolute perturbation $\|f\| \cos \psi_i$ in $\bar{x}_i - x_i$ is amplified by $\|r_i\|$. In the first expression for the relative error, the perturbation $\epsilon_b \cos \psi_i$ is amplied by $1/\cos \beta_i$. That is, the 'more orthogonal' b is to r_i, the smaller is $\cos \beta_i$, and the larger is the amplification of the

relative perturbation. Therefore, the component-wise relative error is likely to increase, the more orthogonal r_i is to the right-hand side.

Comparing the two amplifiers we see that the amplifier $\|r_i\|$ in the absolute error only refers to the matrix and ignores b, while the amplifier $1/\cos\beta_i$ in the first expression for the relative error describes a relationship between the matrix and the right-hand side.

The second expression for the relative error in Theorem 2 is more conventional and perhaps easier to interpret. It consists of the relative perturbation $\epsilon_b \cos\psi_i$, amplified by three factors: the magnitude of x_i relative to $\|x\|$; the term $\|A\|\,\|r_i\|$, which describes the condition of the matrix and will be studied more closely in Section 2.2; and the term $\frac{\|b\|}{\|A\|\|x\|}$, which is common to all components and describes the relation between matrix and right-hand side. If we denote by $\kappa(A) = \|A\|\,\|A^\dagger\|$ the condition number of the matrix A then $\|A\|\,\|r_i\|$ can be bounded by

$$1 = \|e_i^T\| = \|e_i^T A^\dagger A\| \le \|e_i^T A^\dagger\|\,\|A\| = \|A\|\,\|r_i\| \le \kappa(A),$$

A lower bound for $\frac{\|b\|}{\|A\|\|x\|}$, provided $x \ne 0$, is

$$\frac{\|b\|}{\|A\|\,\|x\|} = \frac{1}{\|A\|}\frac{1}{\frac{\|A^\dagger b\|}{\|b\|}} \ge \frac{1}{\kappa(A)}.$$

In case of a linear system $Ax = b$,

$$\frac{\|b\|}{\|A\|\,\|x\|} = \frac{\|Ax\|}{\|A\|\,\|x\|} \le 1,$$

otherwise it can be unbounded since b may be almost orthogonal to all rows of $A^\dagger$.

Therefore, the component-wise relative error tends to be large for those components x_i whose size is small in comparison to $\|x\|$, or whose matrix condition number $\|A\|\,\|r_i\|$ is large, or whose right-hand side is nearly orthogonal to all rows of $A^\dagger$. The three amplification factors in the second expression for the relative error in Theorem 2 provide a clear separation of the factors responsible for the loss of accuracy in the computed solution: relative magnitude of the solution components, matrix condition, and relationship between matrix and right-hand side.

In Sections 3 and 4 we show that the same quantities that determine the sensitivity to right-hand side perturbations also determine the sensitivity to perturbations in the matrix. First, though, we relate them to more established ways of measuring sensitivity.

2.2   Relation to Singular Values

The goal of this section is to compare the amplification factors for the usual norm-based errors with those for our new component-wise errors.

Because the two-norm condition number $\kappa(A) = \|A\|\,\|A^\dagger\|$ equals the ratio of the extreme singular values of A, we can relate the $\|r_i\|$ to the singular values of A and obtain the following well-known inequalities.

Theorem 3 *Let A be a $n \times m$ matrix of rank m with singular values $\sigma_1 \ge \ldots \ge \sigma_m > 0$, and denote by r_i^T the rows of $A^\dagger$. Then*

$$\sigma_m \le \frac{1}{\|r_i\|} \le \sigma_1, \qquad \sigma_m \le \min_k \frac{1}{\|r_k\|} \le \sqrt{m}\,\sigma_m.$$

6

If $\|r_{max}\| = \max_k \|r_k\|$ *then*

$$\|A\| \, \|r_{max}\| \leq \|A\| \, \|A^\dagger\| \leq \sqrt{m}\|A\| \, \|r_{max}\|.$$

Proof: The singular values of the left inverse $A^\dagger$ are $1/\sigma_i$, Section 5.5.4 in [12], hence $1/\sigma_1 \leq \|r_i\| \leq 1/\sigma_m$, giving the first set of inequalities.

Let $A = U\Sigma V^T$ be the singular value decomposition of A. The last row $e_m^T V^T$ of V^T is a vector with unit two-norm in $\Re^m$, so at least one of its components, say the jth, must be of magnitude $1/\sqrt{m}$. Hence the jth row r_j of $A^\dagger$ satisfies

$$\|r_j\| = \|U\Sigma(\Sigma^T\Sigma)^{-1}V^T e_j\| = \|\Sigma(\Sigma^T\Sigma)^{-1}V^T e_j\| \geq |e_m^T\Sigma(\Sigma^T\Sigma)^{-1}V^T e_j| \geq \frac{1}{\sqrt{m}}\frac{1}{\sigma_m},$$

yielding the second set of inequalities.

The last set of inequalities comes from $\|r_{max}\| \leq \|A^\dagger\| = 1/\sigma_m$. ∎

Applying Theorem 3 to the second expression for the component-wise error in Theorem 2 shows that there must exist a component $\bar{x}_k$ for which

$$\frac{|\bar{x}_k - x_k|}{|x_k|} \geq \frac{1}{\sqrt{m}}\frac{\|b\|}{\|A\|\,\|x\|}\,\kappa(A)\,\frac{\|x\|}{|x_k|}\,\epsilon_b|\cos\psi_k|.$$

Therefore, the sensitivity of x_k to right-hand side perturbations is proportional to the condition number of A whenever the right-hand side has an appropriate direction, that is, whenever $\frac{\|b\|}{\|A\|\,\|x\|}$ is not too small

We briefly take a closer look at this last condition. When $Ax = b$ and b is a singular vector associated with the smallest singular value σ_m of A, $\|A^\dagger b\|/\|b\| = 1/\sigma_m = \|A^\dagger\|$, then

$$\frac{\|b\|}{\|A\|\,\|x\|} = \frac{1}{\kappa(A)}, \qquad \text{and} \qquad \frac{\|b\|}{\|A\|\,\|x\|}\,\|A\|\,\|r_i\| = \frac{\|r_i\|}{\|A^\dagger\|} \leq 1.$$

According to the expressions for the errors in Theorem 2, the sensitvity of all solution components to right-hand side perturbations is then solely determined by their relative magnitude.

The existence of a row of $A^\dagger$ whose norm approximates $1/\sigma_m$ well, as evidenced by Theorem 3, underlies the rank-revealing QR factorisations, which first appeared in [11, 13], and are further analysed and refined in [20, 10, 6, 21]. In the simplest case, the goal of a rank-revealing QR factorisation is to determine the most linearly dependent column of a matrix A. To this end one performs the QR factorisation $AP = QR$, where Q has orthonormal columns, R is upper triangular and the permutation matrix P is chosen so as to minimise the trailing diagonal element $(R)_{mm}$ of R. Then the inverse of this element, $1/|(R)_{mm}| = \|e_m^T R^{-1}\| = \|r_m^T\|$, is as large as possible, and thus close to $1/\sigma_m$.

While Theorem 3 states that at least one $\|r_j\|$ approximates the smallest singular value well, the following corollary indicates that each $\|r_i\|$ cannot stray too far away from some singular value.

Theorem 4 *Let A be a $n \times m$ matrix of rank m with singular values $\sigma_1 \geq \ldots \geq \sigma_m > 0$, and let $\|A\|_F = \sqrt{\sum_{i,j} a_{ij}^2}$ denote the Frobenius norm of A.*

7

If the $\|r_i\|$ are ordered by increasing norm, $\|r_{j_1}\| \leq \ldots \leq \|r_{j_m}\|$, then

$$\sum_{i=1}^{m} \|r_i\|^2 = \|A^\dagger\|_F^2 = \frac{1}{\sigma_1^2} + \ldots + \frac{1}{\sigma_m^2}, \qquad \sum_{i=1}^{k} \frac{1}{\sigma_i^2} \leq \sum_{i=1}^{k} \|r_{j_i}\|^2, \qquad 1 \leq k \leq m-1.$$

Proof: The equality results from the invariance of the Frobenius norm under orthogonal transformations, Section 2.5.3 in [12].

The inequalities are obtained by applying the proof of Theorem 4.3.26 in [17] to the singular values of $A^\dagger$. ■

Remark 1 *It is important to realise that the looseness of the inequalities in Theorem 9 depends on how close the right singular vector matrix of A is to a permutation matrix: if $A = U\Sigma V^T$ is the SVD of A then*

$$\|r_i\| = \|U\Sigma(\Sigma^T\Sigma)^{-1}V^T e_i\| = \|\Sigma(\Sigma^T\Sigma)^{-1}V^T e_i\|.$$

Thus, if V is a permutation matrix (this includes diagonal matrices) then we can find indices that achieve the bounds in Theorem 9 since $\|r_i\| = 1/\sigma_k$ for some k.

2.3 Conventional Error Bounds

In this section we present a rather unconventional way of deriving bounds on the *norm-based* relative error, by making use of the theorems from the previous sections.

An expression for the absolute norm-based error in the infinity-norm is available from Theorem 2,

$$\|\bar{x} - x\|_\infty = \max_i \{\|r_i\| \|f\| |\cos\psi_i|\}.$$

Dividing this by $\|x\|$ results in a mixed-norm relative error

$$\frac{\|\bar{x} - x\|_\infty}{\|x\|} = \max_i \{\|A\| \|r_i\| \frac{\|b\|}{\|A\| \|x\|} \epsilon_b |\cos\psi_i|\},$$

where $\epsilon_b = \|f\|/\|b\|$. Denoting by $\kappa(A) = \|A\| \|A^\dagger\|$ the condition number of A, we obtain an upper bound for the norm-based relative error from Theorem 3,

$$\frac{\|\bar{x} - x\|}{\|x\|} \leq \sqrt{m} \frac{\|\bar{x} - x\|_\infty}{\|x\|} \leq \sqrt{m}\kappa(A) \frac{\|b\|}{\|A\| \|x\|} \epsilon_b.$$

In case of a linear system $Ax = b$, $\|b\| \leq \|A\| \|x\|$ and the bound simplifies to

$$\frac{\|\bar{x} - x\|}{\|x\|} \leq \sqrt{m}\kappa(A)\epsilon_b.$$

In this last form, the upper bound agrees with the conventional bounds. Its amplification factor for the perturbations consists of the condition number $\kappa(A)$ of the matrix but ignores the relationship between matrix and right-hand side.

Theorem 3 also comes in handy for the derivation of the lower bound

$$\frac{\|\bar{x} - x\|}{\|x\|} \geq \frac{\|\bar{x} - x\|_\infty}{\|x\|} = \max_i \{\|A\| \|r_i\| \frac{\|b\|}{\|A\| \|x\|} \epsilon_b |\cos\psi_i|\} \geq \frac{1}{\sqrt{m}}\kappa(A) \frac{\|b\|}{\|A\| \|x\|} \epsilon_b \mu,$$
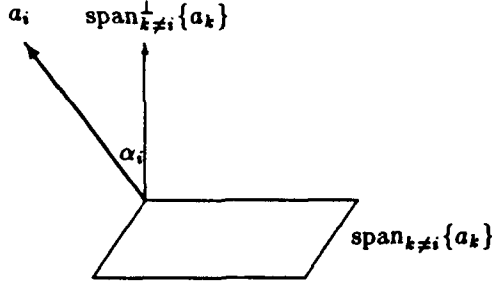
Figure 1: Angles Associated with Columns.

where $\mu = \max_i \{\|r_i\| \, |\cos\psi_i|\} / \max_k \|r_k\|$.

To summarise, we have derived lower and upper bounds on the norm-wise relative error for perturbations restricted to the right-hand side,

$$\frac{1}{\sqrt{m}} \kappa(A) \frac{\|b\|}{\|A\| \|x\|} \epsilon_b \mu \leq \frac{\|\bar{x} - x\|}{\|x\|} \leq \sqrt{m} \kappa(A) \frac{\|b\|}{\|A\| \|x\|} \epsilon_b.$$

In the absence of knowledge about the value of $\cos\psi_i$ we have to assume the worst case $\mu = 1$, which implies that the norm-based error bound is tight. Therefore the conventional upper bounds are as good as possible – given that one has chosen to measure a *norm*-based error. We have therefore shown that, if the norm-wise bounds give unsatisfying information, it is not because the bounds are loose but rather because an unsatisfying way of measuring the error was adopted in the first place.

When $Ax = b$ and b is a singular vector associated with the smallest singular value σ_m of A, $\|A^\dagger b\|/\|b\| = 1/\sigma_m = \|A^\dagger\|$, then $\|A\| \|x\|/\|b\| = \kappa(A)$ and

$$\frac{1}{\sqrt{m}} \epsilon_b \mu \leq \frac{\|\bar{x} - x\|}{\|x\|} \leq \sqrt{m} \epsilon_b.$$

In this special case the norm-wise relative error is of about the same magnitude as the perturbation in the right-hand side and does not depend on the condition number of A, an observation already made by Chan and Foulser [7].

2.4 Geometric Interpretation

We have seen so far that individual components of the solution x to a full-rank least squares problem $\min_y \|Ay - b\|$ can be expressed as $x_i = \|r_i\| \|b\| \cos\beta_i$, where r_i^T is the ith row of $A^\dagger$ and β_i is the angle between r_i and b; that $\|r_i\|$ and $1/\cos\beta_i$ determine the sensitivity of x_i to perturbations in b; and that at least one $1/\|r_j\|$ approximates the smallest singular value of A well.

Now we want to give a geometric interpretation of the $\|r_i\|$ in terms of the columns in the original matrix A. This will allow us to determine how exactly the linear independence of the columns of A and their relationship to b affects the sensitivity of individual solution components to perturbations.

As for notation, the column space of a matrix A is represented by $\mathcal{R}(A)$ and its nullspace by $\text{Ker}(A)$. The subspace in real n-space $\Re^n$ that is orthogonal to the space $\text{span}\{v_1, \ldots, v_k\}$ spanned by $n \times 1$ vectors $v_1, \ldots, v_k$ is denoted by $\text{span}^\perp\{v_1, \ldots, v_k\}$.

We first show that the size of the $\|r_i\|$ reflects the linear dependence of the ith column of A on all others.

Theorem 5 *Given a $n \times m$ matrix A of rank m, denote by a_i its columns, and by r_i^T the rows of its left-inverse $A^\dagger$,*

$$A = (\, a_1 \quad \ldots \quad a_m \,), \qquad A^\dagger = (A^T A)^{-1} A^T = \begin{pmatrix} r_1^T \\ \vdots \\ r_m^T \end{pmatrix}.$$

Then $\mathcal{R}((A^\dagger)^T) = \mathcal{R}(A)$ and

$$\|r_i\| = \frac{1}{\|a_i\| \cos \alpha_i}, \qquad 1 \leq i \leq m,$$

where $-\frac{1}{2}\pi < \alpha_i < \frac{1}{2}\pi$ is the angle between r_i and a_i.

Proof: Because A has full column rank, $A^T A$ is non-singular, and $Ax = A(A^T A)^{-1} z = (A^\dagger)^T z$, where $z = A^T A x$, which implies that $\mathcal{R}((A^\dagger)^T) = \mathcal{R}(A)$.

The ith diagonal element of $I = A^\dagger A$ satisfies $1 = r_i^T a_i = \|r_i\| \, \|a_i\| \cos \alpha_i$, where α_i is the angle between r_i and a_i. Hence $\cos \alpha_i > 0$, so $-\frac{1}{2}\pi < \alpha_i < \frac{1}{2}\pi$, and $\|r_i\| = \frac{1}{\|a_i\| \cos \alpha_i}$. ∎

Because $e_i^T = r_i^T A$, r_i is orthogonal to all columns of A except for a_i, that is $r_i \in \text{span}^{\perp}_{k \neq i}\{a_k\}$, see Figure 1. Theorem 11 and Corollary 5 of Appendix 1 show that the ith row r_i^T of $A^\dagger$ has the same direction as the residual in the least squares approximation of column a_i by the remaining columns: if A_i contains all columns of A except for a_i then

$$r_i^T = e_i^T A^\dagger = \frac{1}{\|a_i\| \cos \alpha_i} \, \frac{1}{\|\hat{a}_i\|} \hat{a}_i^T,$$

where $-\hat{a}_i = A_i z - a_i$ is the residual for the solution z to the least squares problem $\min_y \|A_i y - a_i\|$. In other words, $\hat{a}_i$ is the projection of a_i onto the orthogonal complement of $\mathcal{R}(A_i)$, and r_i has the same direction as $\hat{a}_i$.

With regard to the length of r_i, it follows that

$$\|r_i\| = \frac{1}{\|\hat{a}_i\|} = \frac{1}{\|a_i\| \cos \alpha_i}.$$

This means, the better the remaining columns A_i approximate a_i the smaller is the residual $\|\hat{a}_i\|$ and the larger is $\|r_i\|$. That is, the more linearly dependent a_i is on the other columns, the larger is $\|r_i\|$.

The relationship between the length of r_i and the norm of the residual is already known. In [21] Stewart uses a different argument to show that

$$\|\hat{a}_i\| = \min_y \|A_i y - a_i\| = \frac{1}{\|r_i\|}.$$

Our contribution here is to provide more justification for the choice of r_i as an indicator of sensitivity. Because r_i is a multiple of the residual $\hat{a}_i$, the residual is inherent in A and thus represents a most natural choice for sensitivity measure.

Our geometric interpretation of the rows of the left-inverse justifies the use of rank-revealing QR factorisations to determine the most linearly dependent column of a matrix. If the permutation matrix P for the QR factorisation $AP = QR$ is chosen so that the trailing diagonal element $|(R)_{mm}|$

of R is minimal, then the residual $1/\|r_m^T\| = 1/\|e_m^T R^{-1}\| = |(R)_{mm}|$ is minimal. This implies that the last column of AP is the column that can be best approximated by all other columns and so is the most linearly dependent among all columns.

The individual components of the solution x to a least squares problem $\min_y \|Ay - b\|$ can be expressed as

$$x_i = \|r_i\| \|b\| \cos \beta_i = \frac{\|b\| \cos \beta_i}{\|a_i\| \cos \alpha_i}.$$

The denominator of x_i indicates the linear dependence of column a_i on all others, while the numerator indicates the contribution of the right-hand side b in $\text{span}_{k \neq i}^\perp \{a_k\}$. In detail, for fixed b, the smaller the contribution of a_i outside the space spanned by the other columns, the larger is x_i. Or, the smaller the contribution of a_i outside the space spanned by the other columns, the more x_i has to make up for the weakness of a_i in the direction $\text{span}_{k \neq i}^\perp \{a_k\}$. Moreover, the shorter a_i is, the larger x_i has to be because it has to make up for the shortness of a_i.

We can also apply the geometric interpretations to the errors resulting from perturbations f in the right-hand side. The expression for the absolute error from Theorem 2,

$$\bar{x}_i - x_i = \frac{\|f\| \cos \psi_i}{\|a_i\| \cos \alpha_i},$$

contains a large amplification factor $\frac{1}{\|a_i\| \cos \alpha_i}$ if column a_i is short or lies almost in the space spanned by the other columns. The relative error

$$\frac{\bar{x}_i - x_i}{x_i} = \frac{1}{\cos \beta_i} \epsilon_b \cos \psi_i$$

contains a large amplification factor $1/\cos \beta_i$ if b lies almost in the space spanned by the other columns or in $\text{Ker}(A^T) = \text{Ker}(A^\dagger)$ (in the latter case the right-hand side of the normal equations is zero). Note that the amplification factor for the absolute error only reflects the linear independence of the matrix columns, yet ignores their relation to the right-hand side.

2.5 Implications for Column Scaling

A diagonal column scaling D of the least squares problem $\min_y \|Ay - b\|$ to $\min_z \|(AD)z - b\|$, where $D = (d_{ij})$ is a non-singular diagonal matrix, changes only the lengths of the columns but not the angles, so

$$z_i = \frac{\|b\|}{\|d_{ii} a_i\|} \frac{\cos \beta_i}{\cos \alpha_i}.$$

In case of a column equilibrated matrix AD, Section 3.5.2 in [12], and [24, 25], where the diagonal matrix D is chosen so that all columns of AD have identical length, the condition number of AD comes from the largest angle α_{max} of A, as

$$\frac{1}{\cos \alpha_{max}} \leq \|AD\| \|(AD)^\dagger\| \leq \frac{\sqrt{m}}{\cos \alpha_{max}},$$

according to Theorem 3. This bound already appeared in a different form in [21].

Van der Sluis has shown that a column equilibrated matrix A has the lowest condition number among all matrices of the form AD [24]. This would suggest that one solve only linear systems and

11

least squares problems with column equilibrated matrices so as to minimise the condition number in

$$\frac{\|\bar{x} - x\|}{\|x\|} \leq \sqrt{m} \kappa(A) \frac{\|b\|}{\|A\| \|x\|} \epsilon_b.$$

However, the condition number occurs in an *upper bound*!

An examination of the first expression for the component-wise relative error in Theorem 2 shows that none of the angles change when the columns of A are multiplied by non-zero scalars. In particular, if we consider instead the system $(AD)z = b$, where $z = D^{-1}x$, then the computed solution $\bar{z}$ satisfies a perturbed system $AD\bar{z} = b + g$. Postmultiplication of A by D corresponds to premultiplication of $A^\dagger$ by D^{-1}, which changes only the lengths of the rows r_i^T in $A^\dagger$ but preserves the angles β_i between b and r_i. Hence the amplification factor $1/\cos \beta_i$ remains invariant under column scaling.

Therefore, if perturbations are restricted to the right-hand side, then column scaling is only beneficial if it manages to decrease the relative perturbations $\epsilon_b \cos \psi_i$ in the component-wise relative error (this could occur, for instance, if column scaling brings about a different choice of pivots in Gaussian elimination).

2.6 Summary

The main result of Section 2 is the pair of expressions for the component-wise relative errors in a full-rank least squares problem when perturbations are restricted to the right-hand side (Theorem 2).

Suppose $x \neq 0$ solves the least squares problem $\min_y \|Ay - b\|$, and $\bar{x}$ solves the corresponding problem $\min_y \|Ay - (b + f)\|$ with a perturbed right-hand side. The relative error in an individual component of $\bar{x}$ can be expressed as

$$\frac{\bar{x}_i - x_i}{x_i} = \frac{1}{\cos \beta_i} \epsilon_b \cos \psi_i,$$

where β_i is the angle between b and the ith row of $A^\dagger$, ψ_i is angle between f and the ith row of $A^\dagger$, and $\epsilon_b = \|f\|/\|b\|$. Thus, the component-wise relative error consists of a relative perturbation $\epsilon_b \cos \psi_i$, amplified by $1/\cos \beta_i$. This amplification factor is large if b is almost orthogonal to the ith row of $A^\dagger$; that is, if b lies almost in the space spanned by the other columns or in $\text{Ker}(A^T) = \text{Ker}(A^\dagger)$.

Because β_i depends only on the direction but not the length of the ith row of $A^\dagger$, column scaling of A is only beneficial if it manages to decrease the relative perturbations $\epsilon_b \cos \psi_i$.

We also gave a second expression for the relative error

$$\frac{\bar{x}_i - x_i}{x_i} = \frac{\|b\|}{\|A\| \|x\|} \frac{\|x\|}{x_i} \|A\| \|r_i\| \epsilon_b \cos \psi_i,$$

which provides a clear separation of the factors responsible for the loss of accuracy in the computed solution: relative magnitude of the solution components $\|x\|/x_i$; matrix condition $\|A\| \|r_i\|$; and relationship between matrix and right-hand side $\frac{\|b\|}{\|A\| \|x\|} \geq \frac{1}{\kappa(A)}$, where $\kappa(A) = \|A\| \|A^\dagger\|$ is the matrix condition number. In case of a linear system $Ax = b$,

$$\frac{\|b\|}{\|A\| \|x\|} = \frac{\|Ax\|}{\|A\| \|x\|} \leq 1,$$

otherwise there is no bound as b may be almost orthogonal to all rows of $A^\dagger$.

The component-wise relative error tends to be large for those components x_i whose size is small in comparison to $\|x\|$, or whose matrix condition number $\|A\| \|r_i\|$ is large, or whose right-hand side is nearly orthogonal to all rows of $A^\dagger$. Moreover, Theorem 3 shows that there must be at least one component x_k for which $\|A\| \|r_k\| \geq \kappa(A)/\sqrt{m}$. In the special case when $Ax = b$ and b is a singular vector associated with the smallest singular value of A,

$$\frac{\|b\|}{\|A\| \|x\|} = \frac{1}{\kappa(A)}, \qquad \text{and} \qquad \frac{\|b\|}{\|A\| \|x\|} \|A\| \|r_i\| \leq 1,$$

and the sensitvity of all solution components to right-hand side perturbations is solely determined by their relative magnitude.

In the next section we derive expressions for component-wise relative errors when perturbations in the matrix are also allowed. For simplicity we start with linear systems, and consider least squares problems separately in the subsequent section.

3   Perturbation Results for Linear Systems

We derive expressions for component-wise errors in a linear system of full column rank when both matrix and right-hand side are perturbed. From these expressions we derive *component-wise condition numbers* for the individual components of the solution. The expressions for the component-wise errors are used in turn to derive upper bounds for the norm-based errors that are essentially equal to the conventional upper bounds. We conclude that the norm-based bounds are as tight as possible. If they turn out to be pessimistic then this is because one has chosen to measure the norm of the error instead of the error in individual components.

3.1   Component-Wise Errors

A computed solution $\bar{x}$ to a linear system $Ax = b$ can be viewed as the exact solution to a perturbed system $(A + F)\bar{x} = b + f$. We will determine how the error in the components of $\bar{x}$ is affected by the perturbations F and f.

The first theorem investigates the effect of perturbations in the matrix.

Theorem 6 *Given a matrix A of full column rank and $b \neq 0$ such that $Ax = b$, let the computed solution $\bar{x} \neq 0$ satisfy $(A + F)\bar{x} = b$.*

Denote by ψ_i the angle between r_i and $F\bar{x}$. Then

$$\bar{x}_i = x_i - \frac{\|F\bar{x}\| \cos \psi_i}{\|a_i\| \cos \alpha_i}.$$

If $x_i \neq 0$ and $\epsilon_A = \frac{\|F\bar{x}\|}{\|A\| \|\bar{x}\|}$ then

$$\begin{aligned}
\frac{\bar{x}_i - x_i}{x_i} &= -\frac{1}{\cos \beta_i} \frac{\|F\bar{x}\|}{\|b\|} \cos \psi_i \\
&= -\frac{\|\bar{x}\|}{x_i} \|A\| \|r_i\| \, \epsilon_A \cos \psi_i.
\end{aligned}$$

13

Proof: In Theorem 2 we set $f = -F\bar{x}$ to get

$$\bar{x}_i = x_i - r_i^T F\bar{x} = x_i \left(1 - \frac{1}{\cos \beta_i} \frac{\|F\bar{x}\|}{\|b\|} \cos \psi_i\right)$$

$$= x_i - \frac{\|F\bar{x}\| \cos \psi_i}{\|a_i\| \cos \alpha_i}.$$

Dividing the whole equation by x_i gives the expressions for the component-wise relative error. ∎

The first expression for the component-wise error says that the more b lies in $\mathrm{span}^{\perp}_{k \neq i}\{a_k\}$, the more sensitive is x_i to relative perturbations. However a large $1/\cos \beta_i$ does not necessarily imply that b has little contribution in a_i. In fact, if $b = a_i$ and $1/\cos \alpha_i$ is large then $1/\cos \beta_i$ will also be large – in this case $\cos \beta_i$ reflects the linear dependence of the columns of A.

We interpret the second expression for the component-wise relative error in Theorem 6 as follows: the first term, $\|x\|/x_i$, represents the relative magnitude of x_i; the second term, $\|A\| \|r_i\| = \frac{\|A\|}{\|a_i\| \cos \alpha_i}$, represents the linear dependence of the ith column of A on all other columns; and the last term $\epsilon_A \cos \psi_i$ represents a relative perturbation for the matrix in the context of the given linear system. The component-wise relative error tends to be large for those components x_i whose size is small in comparison to $\|\bar{x}\|$, or whose associated column is short in length or nearly linearly dependent on the other columns. The two amplification factors in the second expression for the relative error in Theorem 6 provide a clear separation of the factors responsible for loss of accuracy in the computed solution: relative magnitude of solution components and linear dependence of matrix columns.

In comparison to the error from right-hand side perturbations in Theorem 2, the error from matrix perturbations in Theorem 6 does not contain the term $\frac{\|b\|}{\|A\| \|x\|}$, which describes the relationship between matrix and right-hand side. According to Theorem 3 we conclude that there *always* exists a component x_k whose sensitivity to relative perturbations in the matrix is on the order of $\kappa(A)$. This is in contrast to right-hand side perturbations, where b has to lie in a certain direction for the sensitivity to be proportional to the condition number.

Before resolving this apparent contradiction (in particular, when the perturbations are due to backward errors from algorithms, which can be shuffled back and forth between matrix and right-hand side), we first give an expression for the component-wise relative error for a linear system when matrix and right-hand side are perturbed simultaneously.

Corollary 1 *Given a matrix A of full column rank, and $b \neq 0$ such that $Ax = b$. Let $\bar{x} \neq 0$ satisfy $(A + F)\bar{x} = b + f$.*

Denote by $\psi_{F,i}$ the angle between r_i and $F\bar{x}$, and by $\psi_{f,i}$ the angle between r_i and f. Then

$$\bar{x}_i = x_i + \frac{\|f\| \cos \psi_{f,i} - \|F\bar{x}\| \cos \psi_{F,i}}{\|a_i\| \cos \alpha_i}.$$

If $x_i \neq 0$ and

$$\epsilon_b = \frac{\|f\|}{\|b\|}, \qquad \epsilon_A = \frac{\|F\bar{x}\|}{\|A\| \|\bar{x}\|}$$

then

$$\frac{\bar{x}_i - x_i}{x_i} = -\frac{1}{\|b\| \cos \beta_i} [\|F\bar{x}\| \cos \psi_{F,i} - \|f\| \cos \psi_{f,i}]$$

14

$$= -\frac{\|\bar{x}\|}{x_i} \|A\| \|r_i\| \left[\epsilon_A \cos \psi_{F,i} - \frac{\|b\|}{\|A\| \|\bar{x}\|} \epsilon_b \cos \psi_{f,i} \right]$$

The second expression for the relative error allows us to state that, in general, for *every* linear system there exists a solution component whose sensitivity is proportional to the condition number, because the term that could avoid this, $\frac{\|b\|}{\|A\|\|\bar{x}\|}$, multiplies only the right-hand side perturbations but not the matrix perturbations. In addition, the following theorem shows that for any $\bar{x} \neq 0$ the perturbations can always be allocated to the matrix.

The following theorem helps to resolve the discrepancies in the rôles of right-hand side and matrix perturbations. It also justifies the representation of the matrix perturbation in the form $\epsilon_A = \frac{\|F\bar{x}\|}{\|A\|\|\bar{x}\|}$. The bounds on the norm-based relative error [12, 23], usually contain the term $\rho_A = \|F\|/\|A\|$ as the representative for the matrix perturbation. But $\epsilon_A \leq \rho_A$ and, as it turns out, ϵ_A constitutes the smallest possible matrix perturbation.

Theorem 7 *Given a matrix A of full column rank and $b \neq 0$ such that $Ax = b$, and a computed solution $\bar{x} \neq 0$, let F_{min} be the perturbation of smallest Frobenius norm among all perturbations F that satisfy $(A + F)\bar{x} = b$ (F_{min} also has smallest two-norm among all such perturbations).*

If $x_i \neq 0$ and $\epsilon_{min} = \|F_{min}\|/\|A\|$ then

$$\frac{\bar{x}_i - x_i}{x_i} = \frac{\|A\| \|\bar{x}\|}{\|b\|} \frac{1}{\cos \beta_i} \epsilon_{min} \cos \psi_i,$$

where ψ_i is the angle between $F_{min}\bar{x}$ and r_i.

If $\epsilon_{res} = \|b - A\bar{x}\|/\|b\|$ is the relative residual then

$$\epsilon_{min} = \frac{\|b\|}{\|A\| \|\bar{x}\|} \epsilon_{res}, \qquad \frac{1}{\kappa(A)} \frac{\|x\|}{\|\bar{x}\|} \epsilon_{res} \leq \epsilon_{min} \leq \frac{\|x\|}{\|\bar{x}\|} \epsilon_{res}.$$

Proof: If $f = b - A\bar{x}$ is the residual then F_{min} is given by, [18] and Theorem III.2.16 in [23],

$$F_{min} = -\frac{f\bar{x}^T}{\bar{x}^T \bar{x}}$$

and satisfies

$$f = -F_{min}\bar{x} \qquad \text{and} \qquad \|F_{min}\| = \frac{\|f\|}{\|\bar{x}\|},$$

where the second equality comes about because F_{min} has rank one. Substituting $\|f\| = \|F_{min}\| \|\bar{x}\|$ into the first expression for the component-wise relative error from Theorem 2 yields the expression for the error.

The relation between ϵ_{min} and ϵ_{res} comes about as $\epsilon_{res} = \|f\|/\|b\|$ and $\|F_{min}\| = \|f\|/\|\bar{x}\|$. ∎

The proof of Theorem 7 makes clear that, given $Ax = b$ and $\bar{x}$, the smallest matrix perturbation satisfies

$$(A + F_{min})\bar{x} = b, \qquad \epsilon_{min} = \frac{\|F_{min}\|}{\|A\|} = \frac{\|F\bar{x}\|}{\|A\| \|\bar{x}\|},$$

which is exactly the matrix perturbation ϵ_A in Theorem 6.

Moreover, for a given computed solution $\bar{x}$ one can define two perturbations: the minimal matrix perturbation ϵ_{min}; and the relative residual ϵ_{res}, which reflects the relationship between matrix and right-hand side. If the magnitude of the computed solution is not totally off, i.e. if $\|x\| \approx \|\bar{x}\|$, then ϵ_{min} is of the same order of magnitude or smaller than ϵ_{res}. According to Sections 2.1 and 2.2, ϵ_{res} is much larger than ϵ_{min} whenever b lies nearly in the direction of a singular vector associated with the smallest singular value of A (provided the directions of x and $\bar{x}$ are not too different).

Regarding the interpretation of error bounds and the determination of amplification factors, one must therefore be careful about deciding whether to allocate the perturbations to the matrix or to the right-hand side. We continue this discussion in the context of norm-wise error bounds in Section 3.4.

In Theorem 7 the relative errors in the different components differ only in $\cos \psi_i / \cos \beta_i$, while the term $\|A\| \|\bar{x}\| / \|b\|$ is common to all components. Because $1/\cos \beta_i \geq 1$,

$$\frac{|\bar{x}_i - x_i|}{|x_i|} \geq \frac{\|A\| \|\bar{x}\|}{\|b\|} \epsilon_{min} |\cos \psi_i|,$$

so *all* components of x are sensitive to *matrix* perturbations if $\|A\| \|\bar{x}\| / \|b\|$ is large. In particular, if b lies along the direction of a singular vector associated with the smallest singular value of A then $\|A\| \|\bar{x}\| / \|b\| \approx \kappa(A)$. Together with the results from Section 2.3 this implies that the solution components are extremely sensitive to matrix perturbations exactly when they are insensitive to right-hand side perturbations.

The expressions for the component-wise errors in this section contain not only the data A and b, but also the result $\bar{x}$. In Appendix 2 we show how to express the relative errors entirely in terms of the data; although the perturbations take a slightly different form, the magnification factors for the perturbations continue to be $1/\cos \alpha_i$ and $1/\cos \beta_i$.

3.2   Examples

Now we give two examples to illustrate the previous results. The first example demonstrates that a matrix with perfectly conditioned columns may give rise to a linear system with extremely sensitive solution components.

Example 1 *If A is an orthogonal matrix then $\frac{\|A\|}{\|a_i\| \cos \alpha_i} = 1$ and, according to Theorem 6, $\|\bar{x}\|/x_i$ is the sole term responsible for error magnification. Thus, as we already know from the norm-wise bounds, a solution vector with small as well as large components suffers from large error amplification in the small components.*

This also comes out if we consider instead the angles

$$\frac{1}{\cos \alpha_i} = \frac{1}{\|a_i\| \cos \alpha_i} = 1, \qquad \frac{1}{\cos \beta_i} = \frac{\|x\|}{x_i}, \qquad \frac{1}{\|b\| \cos \beta_i} = \frac{1}{x_i},$$

where the next to last equality comes about because $\|b\| = \|x\|$.

In contrast to the first example, the second one shows that even a very ill-conditioned matrix may have robust solution components. It is a generalisation of the example presented in Section 1.1.

16

Example 2 *Consider a 4×4 orthogonal matrix $A = (\begin{array}{cccc} a_1 & a_2 & a_3 & a_4 \end{array})$ and define a one-parameter family of matrices by*

$$A(\lambda) = \left(\begin{array}{cccc} a_1 & a_2 & a_3 & \frac{1}{\sqrt{1+\lambda^2}}(\lambda a_3 + a_4) \end{array} \right).$$

We see that $A(0) = A$, a well-conditioned matrix, and that $A(\infty)$ is a singular matrix. For all λ, $\|A(\lambda)\| \leq 2$. When $\lambda < \infty$, the inverse is given by

$$[A(\lambda)]^{-1} = \left(\begin{array}{c} a_1^T \\ a_2^T \\ a_3^T - \lambda a_4^T \\ \sqrt{1+\lambda^2} a_4^T \end{array} \right),$$

from which we can compute

$$\cos \alpha_1 = \|a_1\| \cos \alpha_1 = \cos \alpha_2 = \|a_2\| \cos \alpha_2 = 1$$
$$\cos \alpha_3 = \|a_3\| \cos(\alpha_3) = \cos \alpha_4 = \|a_4\| \cos(\alpha_4) = \frac{1}{\sqrt{1+\lambda^2}}.$$

Thus as $\lambda \to \infty$ the matrix $A(\lambda)$ becomes increasingly singular. Its condition number behaves like $O(\lambda)$. Note that the matrix $A(\lambda)$ is column equilibrated, so the ill-conditioning is a result of small angles rather than short columns.

Consider a linear system $A(\lambda)x(\lambda) = b$, where the right-hand side is independent of λ and can be represented as $b = \tau_1 a_1 + \tau_2 a_2 + \tau_3 a_3 + \tau_4 a_4$. Then

$$\cos \beta_1 = \frac{\tau_1}{\|b\|}, \qquad \cos \beta_2 = \frac{\tau_2}{\|b\|}, \qquad \cos \beta_3 = \frac{\tau_3 - \lambda \tau_4}{\|b\|\sqrt{1+\lambda^2}}, \qquad \cos \beta_4 = \frac{\tau_4}{\|b\|}.$$

The solution vector is given by

$$x(\lambda) = \left(\begin{array}{cccc} \tau_1 & \tau_2 & \tau_3 - \lambda \tau_4 & \sqrt{1+\lambda^2} \tau_4 \end{array} \right)^T.$$

The values of x_1 and x_2 are independent of λ, and so are $\|a_j\| \cos \alpha_j$ and $\cos \beta_j$ for $j = 1, 2$. So the sensitivity of the components x_1 and x_2 depends solely on their size relative to x. If, for instance, $|x_1| \gg |x_i|$ for $i \neq 1$ then Corollary 1 says that the error in x_1 is not amplified – independent of the values of λ and the condition number of $A(\lambda)$.

3.3 Condition Numbers and Column Scaling

For a linear system $Ax = b$ with full-rank coefficient matrix A and non-zero right-hand side b, Corollary 1 presents two different expressions for the component-wise relative error in the computed solution $\bar{x}$: suppose $\bar{x} \neq 0$ satisfies $(A + F)\bar{x} = b + f$, and

$$\epsilon_b = \frac{\|f\|}{\|b\|}, \qquad \epsilon_A = \frac{\|F\bar{x}\|}{\|A\|\|\bar{x}\|}$$

then

$$\begin{aligned} \frac{\bar{x}_i - x_i}{x_i} &= -\frac{1}{\|b\| \cos \beta_i} \left[\|F\bar{x}\| \cos \psi_{F,i} - \|f\| \cos \psi_{f,i} \right] \\ &= -\frac{\|\bar{x}\|}{x_i} \|A\| \|r_i\| \left[\epsilon_A \cos \psi_{F,i} - \frac{\|b\|}{\|A\|\|\bar{x}\|} \epsilon_b \cos \psi_{f,i} \right] \end{aligned}$$

17

Sections 2.2 and 3.1 explain that under certain circumstances the factor $\frac{\|b\|}{\|A\|\|x\|}$ causes the sensitivity of large solution components to *right-hand side* perturbations to be independent of matrix conditioning. We now ignore $\frac{\|b\|}{\|A\|\|x\|}$ because it does not affect the sensitivity of the solution to *perturbations in the matrix*.

The term $\|A\|\,\|r_i\| = \|A\|\,\|e_i^T A^\dagger\| \leq \|A\|\,\|A^\dagger\|$ represents a condition number 'restricted to' x_i. Already in 1970 van der Sluis [25, 26] realised the need to distinguish the conditioning of individual components of x and the fact that the conditioning depends on the relative size of a component. He introduces the notion of 'ith column condition number of A', $\|A^{-1}\|\,\|a_i\|$, and derives the similar looking norm-wise relative error bound (here $f = 0$)

$$\frac{\|\bar{x} - x\|}{\|x\|} \leq \frac{\|F\|}{\|A\|} \sum_i \|A^{-1}\|\,\|a_i\| \frac{|x_i|}{\|x\|}.$$

He also acknowledges the importance of angles on the conditioning of the matrix: if each column is well separated from the space spanned by the other columns then the solution components are likely to be insensitive to perturbations, page 251 in [26].

According to van der Sluis's bounds one naturally concludes that column equilibrated matrices (all of whose columns have identical norm) should give rise to solution components with identical sensitivity to perturbations. Yet, the amplification factor $\frac{1}{\|b\|\cos\beta_i}$ in the first expression for the component-wise relative error is independent of column scaling. So essentially the conclusions of Section 2.5 remain valid when, in addition to the right-hand side, the matrix is also perturbed: the component-wise relative error decreases under column scaling only if column scaling manages to reduce the perturbation $\|F\bar{x}\|\cos\psi_{F,i} - \|f\|\cos\psi_{f,i}$. Note that we could have also expressed the error as

$$\frac{\bar{x}_i - x_i}{x_i} = -\frac{1}{\cos\beta_i}\left[\frac{\|F\bar{x}\|}{\|b\|}\cos\psi_{F,i} - \epsilon_b \cos\psi_{f,i}\right],$$

in which case the amplification factors for the relative perturbations $\|F\bar{x}\|/\|b\|$ and ϵ_b remain invariant under column scaling. However, when $f = 0$ we know from Theorem 7 that

$$\frac{\|F\bar{x}\|}{\|b\|} = \frac{\|b - A\bar{x}\|}{\|b\|} = \frac{\|A\|\,\|\bar{x}\|}{\|b\|}\epsilon_A,$$

where $\|A\|\,\|\bar{x}\|/\|b\|$ can be as large as $\kappa(A)$. This means that the perturbation $\|F\bar{x}\|/\|b\|$ may be proportional to the condition number of the matrix. Finally, Lemma 1 of Appendix 2 states that the amplification factors for the error $(\bar{x}_i - x_i)/x_i$ remain invariant under column scaling when perturbations are restricted to column i of the matrix.

Although the amplification factors in the *second* expression for the error above do change under column scaling, they have the advantage of representing easily computable a posteriori error estimates: we show in Section 6 how to estimate $\|A\|\,\|r_i\|$ efficiently with available condition number estimators.

Due to the deliberations in this and the previous sections we feel justified in introducing a new set of condition numbers.

Definition 1 *Let $Ax = b$ be a linear system with $n \times m$ matrix A of rank m and $b \neq 0$, and let $\bar{x} \neq 0$ be the computed solution. Denote by $r_i^T = e_i^T A^\dagger$ the ith row of the left-inverse of A and by β_i the angle between b and r_i, $1 \leq i \leq m$.*

The quantities

$$\frac{\|\bar{x}\|}{|x_i|}, \qquad \|A\|\,\|r_i\|, \qquad 1 \leq i \leq m,$$

18

are called 'component-wise condition numbers for the linear system $Ax = b$'. We also refer to them as 'condition numbers for x_i'.

The validity of this definition is also justified by earlier work of Stewart [21], page 72, who introduces the 'collinearity indices' $\kappa_i = \|a_i\| \, \|r_i\|$. The squares of the collinearity indices are known as 'variance inflation factors' in statistics [21]. From Theorem 5 we see that $\kappa_i = 1/\cos\alpha_i$. They represent the scaling-invariant version of $\|A\| \, \|r_i\|$ and appear as amplification factors in the error expressions of Appendix 2. The main difference between our and Stewart's condition numbers is that the collinearity indices are designed to reflect the linear dependence of the matrix columns, while our component-wise condition numbers measure the conditioning of the linear system: matrix plus right-hand side.

3.4 Conventional Error Bounds

As we did already in Section 2.3, we now relate our component-wise results to the conventional norm-based upper bounds on relative errors.

Corollary 2 *Given a $n \times m$ matrix A of rank m and $b \neq 0$ such that $Ax = b$, as well as $\bar{x} \neq 0$ with $(A + F)\bar{x} = b + f$, let*

$$\epsilon_b = \frac{\|f\|}{\|b\|}, \qquad \epsilon_A = \frac{\|F\bar{x}\|}{\|A\| \, \|\bar{x}\|}.$$

Then

$$\frac{1}{\sqrt{m}}\kappa(A) \max_i \left| \frac{\|b\|}{\|A\| \, \|x\|}\epsilon_b \mu_{f,i} - \frac{\|\bar{x}\|}{\|x\|}\epsilon_A \mu_{F,i} \right| \leq \frac{\|\bar{x} - x\|}{\|x\|} \leq \sqrt{m}\kappa(A) \left[\frac{\|b\|}{\|A\| \, \|x\|}\epsilon_b + \frac{\|\bar{x}\|}{\|x\|}\epsilon_A \right],$$

where $\mu_{f,i} = \|r_i\| \cos\psi_{f,i} / \max_k \|r_k\|$, $\mu_{F,i} = \|r_i\| \cos\psi_{F,i} / \max_k \|r_k\|$, and $\psi_{f,i}$ and $\psi_{F,i}$ are the respective angles between the ith row r_i of $A^\dagger$ with f and $F\bar{x}$.

Proof: Multiplying the last equation in Corollary 1 by x_i and applying the infinity-norm gives

$$\|\bar{x} - x\|_\infty = \|\bar{x}\| \max_i \left\{ \|A\| \, \|r_i\| \left| \frac{\|b\|}{\|A\| \, \|\bar{x}\|}\epsilon_b \cos\psi_{f,i} - \epsilon_A \cos\psi_{F,i} \right| \right\},$$

yielding the upper bound

$$\frac{\|\bar{x} - x\|_\infty}{\|x\|} \leq \max_i \left\{ \|A\| \, \|r_i\| \right\} \left[\frac{\|b\|}{\|A\| \, \|x\|}\epsilon_b + \frac{\|\bar{x}\|}{\|x\|}\epsilon_A \right] \leq \kappa(A) \left[\frac{\|b\|}{\|A\| \, \|x\|}\epsilon_b + \frac{\|\bar{x}\|}{\|x\|}\epsilon_A \right].$$

The lower bound is obtained from Theorem 3

$$\frac{\|\bar{x} - x\|_\infty}{\|x\|} \geq \frac{1}{\sqrt{m}}\kappa(A) \max_i \left| \frac{\|b\|}{\|A\| \, \|x\|}\epsilon_b \mu_{f,i} - \frac{\|\bar{x}\|}{\|x\|}\epsilon_A \mu_{F,i} \right|.$$

The mixed-norm error is replaced by the two-norm error by means of the inequalities $\|y\|_\infty \leq \|y\| \leq \sqrt{m}\|y\|_\infty$ for any m-vector y, Section 2.2.2 in [12]. ∎

We arrive at the same conclusion as in Section 2.3, where only perturbations in the right-hand side were allowed. Without knowledge about $\cos\psi_{F,i}$ and $\cos\psi_{f,i}$ the norm-based bounds are as

tight as possible. In fact, under weaker conditions we have derived essentially the same upper bound as the one commonly found in the literature for non-singular linear systems. In Section III.2.3 in [23], for instance, one finds that, subject to the condition $\|A^{-1}F\| < 1$,

$$\frac{\|\bar{x} - x\|}{\|x\|} \leq \frac{\kappa(A)}{1 - \kappa(A)\rho_A}(\rho_A + \epsilon_b), \qquad \rho_A = \frac{\|F\|}{\|A\|},$$

while Corollary 2 gives

$$\frac{\|\bar{x} - x\|}{\|x\|} \leq \sqrt{m}\kappa(A)\frac{\|\bar{x}\|}{\|x\|}(\rho_A + \epsilon_b).$$

In this last, commonly used form the norm-based bound ignores any relationship between matrix and right-hand side.

Chan and Foulser [7] intended to remedy this ignorance of the right-hand side by modifying the bound as follows. Let

$$A = U\Sigma V^T, \qquad \text{where} \quad U = (u_1 \quad \dots \quad u_n), \qquad \sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_n > 0,$$

be the SVD of a non-singular matrix A with singular values σ_i and right singular vectors u_i. According to Theorem 1 in [7], if $A\bar{x} = b + f$ and P_k is the orthogonal projection onto span $\{u_{n-k+1}, \dots, u_n\}$

$$\frac{\|\bar{x} - x\|}{\|x\|} \leq \frac{\sigma_{n-k+1}}{\sigma_n}\left(\frac{\|P_k b\|}{\|b\|}\right)^{-1}\epsilon_b.$$

Chan and Foulser [7] conclude that if, for some k, a large fraction of b lies in span $\{u_{n-k+1}, \dots, u_n\}$ and if $\sigma_{n-k+1} \approx \sigma_n$ then x 'is relatively insensitive to perturbations in b'. For instance, if $b = u_n$ then $P_1 b = b$,

$$\frac{\|\bar{x} - x\|}{\|x\|} \leq \epsilon_b,$$

and we conclude that x is insensitive to perturbations in b.

The interpretation of Theorem 1 given in [7] is valid if f represents the *input* error in the data b. However we do not agree with the application of Theorem 1 in the case when f represents a backward error chosen to satisfy $A\bar{x} = b + f$. As we discussed in Section 3.1, f depends on the size of $\bar{x}$. From Theorem 7 we know that $F_{min} = -\frac{f\bar{x}^T}{\bar{x}^T\bar{x}}$ is the perturbation of smallest two-norm and Frobenius norm satisfying $(A + F_{min})\bar{x} = b$, and that

$$\frac{|\bar{x}_i - x_i|}{|x_i|} = \frac{\|A\|\|\bar{x}\|}{\|b\|}\frac{1}{|\cos\beta_i|}\epsilon_{min}|\cos\psi_i| \geq \frac{\|A\|\|\bar{x}\|}{\|b\|}\epsilon_{min}|\cos\psi_i|.$$

When $b = u_n$ the common term $\|A\|\|\bar{x}\|/\|b\|$ is approximately σ_1/σ_n, and the sensitivity of *all* solution components is proportional to the condition number. A slightly different argument based on the second statement in Theorem 7,

$$\epsilon_{min} = \frac{\|F_{min}\|}{\|A\|} = \frac{\|b\|}{\|A\|\|\bar{x}\|}\epsilon_b,$$

implies that for $b = u_n$ we have $\epsilon_b \approx \kappa(A)\epsilon_{min}$, and the ill-conditioning is merely hidden in the perturbation ϵ_b. Consequently, all components of x are extremely sensitive to perturbations if A is ill-conditioned, which disagrees with the interpretation by Chan and Foulser.

3.5 Summary

The main results in this section are expressions for the component-wise errors in a linear system of full column rank when perturbations are allowed in both the matrix and the right-hand side (Corollary 1).

Suppose A is a matrix of full column rank and $b \neq 0$ such that $Ax = b$. Let $\bar{x} \neq 0$ satisfy $(A + F)\bar{x} = b + f$. If

$$
\epsilon_b = \frac{\|f\|}{\|b\|}, \qquad \epsilon_A = \frac{\|F\bar{x}\|}{\|A\|\,\|\bar{x}\|}
$$

then the relative error in an individual component of $\bar{x}$ can be expressed in two ways,

$$
\begin{aligned}
\frac{\bar{x}_i - x_i}{x_i} &= -\frac{1}{\|b\|\cos\beta_i}\left[\|F\bar{x}\|\cos\psi_{F,i} - \|f\|\cos\psi_{f,i}\right] \\
&= -\frac{\|\bar{x}\|}{x_i}\|A\|\,\|r_i\|\left[\epsilon_A\cos\psi_{F,i} - \frac{\|b\|}{\|A\|\,\|\bar{x}\|}\epsilon_b\cos\psi_{f,i}\right],
\end{aligned}
$$

where r_i^T is the ith row of the pseudo-inverse $A^\dagger$, β_i is the angle between b and r_i, and $\psi_{F,i}$ and $\psi_{f,i}$ are error angles.

In the first expression, the amplification factors $\frac{1}{\|b\|\cos\beta_i}$ are invariant under column scaling. Hence the component-wise relative error decreases under column scaling only if column scaling manages to reduce the size of the perturbations.

In the second expression, the perturbations are amplified by two terms: $\|\bar{x}\|/x_i$ represents the relative magnitude of x_i, and $\|A\|\,\|r_i\|$ represents the dependence of the ith column of A on all other columns. Hence the component-wise relative error tends to be large for those components x_i whose size is small in comparison to $\|\bar{x}\|$, or whose associated column is short in length or nearly linearly dependent on the other columns.

Theorem 7 demonstrates that any $\bar{x} \neq 0$ can be viewed as the solution to a linear system whose perturbations affect only the matrix and leave the right-hand side clean. Hence the amplification factor for each x_i is at least $\|A\|\,\|r_i\|$. According to Theorem 3 there always exists a component x_k for whom $\|A\|\,\|r_k\|$ is on the order of $\kappa(A)$. Thus any linear system contains a solution component whose sensitivity to relative perturbations is proportional to the condition number of the matrix.

The quantities

$$
\frac{\|\bar{x}\|}{|x_i|}, \qquad \|A\|\,\|e_i^T A^\dagger\|, \qquad 1 \leq i \leq m,
$$

are called *component-wise condition numbers for the linear system $Ax = b$*.

4   Perturbation Results for Least Squares Problems

We saw in Section 2.1 that, provided the matrix has full column rank, perturbations in the right-hand side have the same effect for both linear systems and least squares problems. However this is not true for perturbations in the matrix. A perturbation F in a linear system $(A + F)\bar{x} = b$ represents a 'linear disturbance', and it does not interact with the right-hand side. In contrast, a perturbation F in a least squares problem $\min_y \|(A + F)y - b\|$ represents a 'quadratic disturbance', because the

perturbed problem is equivalent to the linear system $(A+F)^T(A+F)\bar{x} = (A+F)^T b$. In addition, the perturbation F causes a second quadratic disturbance if the right-hand side is perturbed separately.

As in Section 3 we derive expressions for component-wise errors in a least squares problem when both matrix and right-hand side are perturbed. There exists a component of the solution vector whose sensitvity equals at least the product of condition number and $\tan\theta$, where θ is the angle between the right-hand side and the column space of the matrix; the sensitivity can be as high as the product of $\tan\theta$ and the square of the condition number. Least squares problems are therefore always more sensitive to ill-conditioning than linear systems.

Finally, from the expressions for the component-wise errors we derive an upper bound on the norm-based error that is essentially equal to the first-order term in the conventional bound.

The perturbation results in this section can be applied to the computation of the left-inverse by expressing it as the solution $X = A^{\dagger}$ to the least squares problem $\min_Y \|AY - I\|$, and computing one column of X at a time. Norm-based perturbation results for pseudo-inverses can be found, for instance, in Section III of [23].

4.1   Component-Wise Errors

Now we consider perturbations in the coefficient matrix of a full-rank least squares problem $\min_y \|Ay - b\|$. The following theorem shows that the component-wise errors in a least squares problem consist of the errors in a linear system, plus a further term.

Theorem 8 *Given a matrix A of full column rank, let $x \neq 0$ solve $\min_y \|Ay - b\|$ and let $\bar{x} \neq 0$ solve $\min_y \|(A + F)y - (b + f)\|$.*

Let
$$r_i^T = e_i^T A^{\dagger}, \qquad q_i^T = e_i^T (A^T A)^{-1}, \qquad \bar{r} = b + f - (A + F)\bar{x} \neq 0,$$
then
$$\bar{x}_i = x_i - \frac{\|F\bar{x}\| \cos\psi_{F,i} - \|f\| \cos\psi_{f,i}}{\|a_i\| \cos\alpha_i} + \frac{\|F^T \bar{r}\| \cos\omega_{q,i}}{\|a_i\|^2 \cos^2\alpha_i},$$
where $\psi_{F,i}$ is the angle between $F\bar{x}$ and r_i, $\psi_{f,i}$ is the angle between f and r_i, ω_i is the angle between r_i and $\bar{r}$, and $\omega_{q,i}$ is the angle between $F^T \bar{r}$ and q_i.

If $x_i \neq 0$ and $\epsilon_{A,r} = \frac{\|F^T \bar{r}\|}{\|A^T\| \|\bar{r}\|}$ then
$$\begin{aligned} \frac{\bar{x}_i - x_i}{x_i} &= L + \frac{\|\bar{r}\|}{\|b\|} \frac{1}{\cos\beta_i} \cos\omega_i \\ &= L + \frac{\|\bar{r}\|}{\|A\| \|\bar{x}\|} \frac{\|\bar{x}\|}{x_i} \|q_i\| \|A\|^2 \epsilon_{A,r} \cos\omega_{q,i}, \end{aligned}$$
where
$$L = -\frac{1}{\|b\| \cos\beta_i} [\|F\bar{x}\| \cos\psi_{F,i} - \|f\| \cos\psi_{f,i}] = -\frac{\|\bar{x}\|}{x_i} \|A\| \|r_i\| \left[\epsilon_A \cos\psi_{F,i} - \frac{\|b\|}{\|A\| \|\bar{x}\|} \epsilon_b \cos\psi_{f,i}\right]$$
is the component-wise relative error in a linear system solution from Corollary 1.

22

Proof: The vector x solves $\min_y \|Ay - b\|$ if and only if

$$\begin{pmatrix} I & A \\ A^T & 0 \end{pmatrix} \begin{pmatrix} r \\ x \end{pmatrix} = \begin{pmatrix} b \\ 0 \end{pmatrix},$$

where $r = b - Ax$, and the inverse of the coefficient matrix equals

$$\begin{pmatrix} I & A \\ A^T & 0 \end{pmatrix}^{-1} = \begin{pmatrix} I - AA^\dagger & (A^\dagger)^T \\ A^\dagger & -(A^T A)^{-1} \end{pmatrix}.$$

Moreover the vector $\bar{x}$ solves

$$\begin{pmatrix} I & A + F \\ (A + F)^T & 0 \end{pmatrix} \begin{pmatrix} \bar{r} \\ \bar{x} \end{pmatrix} = \begin{pmatrix} b + f \\ 0 \end{pmatrix}, \quad \text{or} \quad \begin{pmatrix} I & A \\ A^T & 0 \end{pmatrix} \begin{pmatrix} \bar{r} \\ \bar{x} \end{pmatrix} = \begin{pmatrix} b \\ 0 \end{pmatrix} - \begin{pmatrix} F\bar{x} - f \\ F^T \bar{r} \end{pmatrix}.$$

We can now apply Theorem 2 to the above system, whose right-hand side perturbation equals $-\begin{pmatrix} F\bar{x} - f \\ F^T \bar{r} \end{pmatrix}$. If $q_i^T = e_i^T (A^T A)^{-1}$ then

$$\bar{x}_i = x_i - r_i^T (F\bar{x} - f) - r_i^T \bar{r} = x_i - r_i^T (F\bar{x} - f) + q_i^T F^T \bar{r}$$

because $F^T \bar{r} = -A^T \bar{r}$. Expressing the inner products in terms of cosines and norms gives

$$\begin{aligned}
\bar{x}_i &= x_i - \frac{\|F\bar{x}\| \cos \psi_{F,i} - \|f\| \cos \psi_{f,i}}{\|a_i\| \cos \alpha_i} - \frac{\|\bar{r}\| \cos \omega_i}{\|a_i\| \cos \alpha_i} \\
&= x_i - \frac{\|F\bar{x}\| \cos \psi_{F,i} - \|f\| \cos \psi_{f,i}}{\|a_i\| \cos \alpha_i} + \|q_i\| \|F^T \bar{r}\| \cos \omega_{q,i},
\end{aligned}$$

where $\psi_{F,i}$ is the angle between r_i and $F\bar{x}$, $\psi_{f,i}$ is the angle between r_i and f, ω_i is the angle between r_i and $\bar{r}$, and $\omega_{q,i}$ is the angle between q_i and $F^T \bar{r}$.

The second expression for the relative error follows from

$$\frac{\|q_i\| \|F^T \bar{r}\|}{x_i} = \frac{\|\bar{x}\|}{x_i} \|q_i\| \|A\|^2 \frac{\|\bar{r}\|}{\|A\| \|\bar{x}\|} \frac{\|F^T \bar{r}\|}{\|A^T\| \|\bar{r}\|}.$$

∎

The above theorem contains as special cases the perturbation results derived before. If $\bar{x}$ happens to solve the linear system $(A + F)\bar{x} = b + f$ then $\bar{r} = 0$, and, as the first expression for the relative error shows, the errors reduce to those in a linear system from Corollary 1. If F happens to be zero then $A^T \bar{r} = 0$, so $\cos \omega_i = 0$ and the errors reduce to those due to pure right-hand side perturbations from Theorem 2.

Theorem 8 provides two expressions for the component-wise relative error, they differ in the form of the additional term due to the least squares nature of the problem. We will now examine them in turn.

The perturbation in the first expression for the relative error in Theorem 8 is $\cos \omega_i$. As argued above, $\cos \omega_i$ is zero whenever F is zero. The perturbation is amplified by $1/\cos \beta_i$, which indicates how linearly dependent b is on the space $\text{span}_{k \neq i}\{a_k\}$. The term $\|\bar{r}\|/\|b\|$ is independent of i, hence present in the relative error of *each* solution component. If θ is the angle between b and $\mathcal{R}(A)$ then we show in Section 4.3 that $\frac{\|r\|}{\|b\|} = \sin \theta \leq 1$, where $r = b - Ax$ is the exact residual. Thus, if $\|\bar{r}\| \approx \|r\|$ then $\|\bar{r}\|/\|b\|$ controls the influence on the relative error from the additional term due to the least squares problem. This term has a greater influence on the error when the distance of b to $\mathcal{R}(A)$ is

23

large: $\mathcal{R}(A) = \mathcal{R}((A^\dagger)^T)$ by Theorem 5, so if b is almost orthogonal to $r_i^T = e_i^T A^\dagger$ then $\cos\beta_i \approx 0$ and $1/\cos\beta_i$ is very large. The advantage of the first expression for the component-wise relative error in Theorem 8 is the invariance of its amplification factor under column scaling. However it seems to be difficult to get a handle on $\cos\omega_i$.

That is the reason why Theorem 8 contains an alternative expression for the component-wise relative error. Although individual factors in the alternative expression change under column scaling, we find them easier to interpret. The relative perturbation $\epsilon_{A,r}\cos\omega_{q,i}$ in the least squares term is amplified by three factors. The first factor represents, as in the error for linear system solution, the size of the component x_i relative to $\|\bar{x}\|$. The second factor $\|q_i\|\,\|A\|^2$ has the bounds

$$(\|r_i\|\,\|A\|)^2 \le \|q_i\|\,\|A\|^2 = \|e_i^T(A^T A)^{-1}\|\,\|A\|^2 \le \|(A^T A)^{-1}\|\,\|A\|^2 = \kappa^2(A),$$

where we have made use of the inequality $\|q_i\| \ge \|r_i\|^2$ from Corollary 6 in the Appendix and the fact that $\|A^T A\| = \|A\|^2$. From Theorem 3 we know that there exists a row r_k of $A^\dagger$ whose norm approximates $A^\dagger$ to a factor of $\sqrt{m}$. Hence there exists at least one component x_k for which

$$\|q_k\|\,\|A\|^2 \ge \frac{1}{m}\kappa^2(A).$$

The third factor multiplying the relative perturbation, $\frac{\|\bar{r}\|}{\|A\|\,\|\bar{x}\|}$, describes the relationship between matrix and right-hand side. It will be examined by considering instead the exact quantity $\frac{\|r\|}{\|A\|\,\|x\|}$. A few paragraphs ago we introduced the angle θ between b and $\mathcal{R}(A)$, so

$$\frac{\|r\|}{\|b\|} = \sin\theta, \qquad \frac{\|Ax\|}{\|b\|} = \frac{\|b-r\|}{\|b\|} = \cos\theta.$$

This implies the equality

$$\frac{\|r\|}{\|A\|\,\|x\|} = \frac{\|Ax\|}{\|A\|\,\|x\|}\tan\theta.$$

Since $\|x\| = \|A^\dagger A x\| \le \|A^\dagger\|\,\|Ax\|$ we get the bounds

$$\frac{1}{\kappa(A)}\tan\theta \le \frac{\|r\|}{\|A\|\,\|x\|} \le \tan\theta.$$

Combining all the bounds for the exact quantities shows that for some x_k

$$\frac{1}{m}\kappa(A)\tan\theta \le \frac{\|r\|}{\|A\|\,\|x\|}\|q_k\|\,\|A\|^2 \le \kappa^2(A)\tan\theta.$$

Consequently, whenever $\|\bar{r}\| \approx \|r\|$ and $\|\bar{x}\| \approx \|x\|$, there exists a solution component whose sensitivity equals at least the product of condition number and $\tan\theta$, and it may be as high as the product of $\tan\theta$ and squared condition number.

Given a computed solution $\bar{x}$ to a linear system, Theorem 7 shows how to construct minimal-size perturbations F, and gives expressions for the corresponding component-wise relative errors. In the case of least squares problems, unfortunately, we do not know how to construct minimal-size perturbations for a given computed solution $\bar{x}$. Therefore the analogue of Theorem 7 for least squares problems below is not nearly as strong. When the exact residual $r = b - Ax = 0$ the expression below equals the one in Theorem 7.

Theorem 9 *Given a matrix A of full column rank, let $x \neq 0$ solve $\min_y \|Ay - b\|$. Denote the computed solution by $\bar{x} \neq 0$, the exact residual by $r = b - Ax$, and the 'computable residual' by $r_c = b - A\bar{x}$.*

24

If $x_i \neq 0$ then

$$\frac{\bar{x}_i - x_i}{x_i} = -\frac{1}{\cos \beta_i} \frac{\sqrt{\|r_c\|^2 - \|r\|^2}}{\|b\|} \cos \psi_i,$$

where ψ_i is the angle between $r_c - r$ and the ith row r_i of $A^\dagger$.

Proof: According to Theorem III.5.5 in [23], the computed solution $\bar{x}$ solves the least squares problem $\min_y \|(A + F_0)y - b\|$, where

$$F_0 = \frac{(r_c - r)\bar{x}^T}{\bar{x}^T \bar{x}}.$$

When $r = 0$ then F_0 equals F_{min} from Theorem 7.

We want to substitute F_0 into the second expression for the component-wise relative error in Theorem 8. To this end, note that by construction of F_0, $F_0\bar{x} = r_c - r$. Moreover, $F_0^T \bar{r} = 0$ because

$$\bar{r} = b - (A + F_0)\bar{x} = r_c - F_0\bar{x} = r,$$

and

$$0 = (A + F_0)^T \bar{r} = A^T r + F_0^T \bar{r} = F_0^T \bar{r}$$

due to $A^T r = 0$ and the first part of the proof of Theorem 8.

Theorem 8 gives therefore

$$\frac{\bar{x}_i - x_i}{x_i} = -\frac{1}{\cos \beta_i} \frac{\|r_c - r\|}{\|b\|} \cos \psi_i,$$

where ψ_i is the angle between r_i. At last, $\|r_c - r\|^2 = \|r_c\|^2 - \|r\|^2$ as $r^T r_c = r^T (A(x - \bar{x}) + r) = \|r\|^2$. ∎

We can now define the set of condition numbers for least squares problems, it contains the set of condition numbers for linear systems.

Definition 2 *Let $x \neq 0$ solve the least squares problem $\min_y \|Ay - b\|$ with $n \times m$ matrix A of rank m. Let $\bar{x} \neq 0$ be the computed solution with residual $\bar{r} \neq 0$. If $q_i = e_i^T (A^T A)^{-1}$ and $r_i^T = e_i^T A^\dagger$ then the quantities*

$$\frac{\|\bar{x}\|}{|x_i|}, \qquad \|A\| \|r_i\|, \qquad \frac{\|\bar{r}\|}{\|A\| \|\bar{x}\|} \|A\|^2 \|q_i\|, \qquad 1 \leq i \leq m,$$

are called component-wise condition numbers for the least squares problem $\min_y \|Ay - b\|$. *We also refer to them as* condition numbers for x_i.

4.2   Example and Conventional Error Bounds

We now modify Example 2 for linear systems to illustrate that a least squares problems with a very ill-conditioned matrix may have extremely insensitive solution components.

Example 3 *Consider a 4×4 orthogonal matrix $A = (\begin{array}{cccc} a_1 & a_2 & a_3 & a_4 \end{array})$ and define a one-parameter family of rectangular matrices by*

$$A(\lambda) = \left(\begin{array}{ccc} a_2 & a_3 & \frac{1}{\sqrt{1+\lambda^2}}(\lambda a_3 + a_4) \end{array}\right).$$

We see that $A(0)$ is a perfectly conditioned matrix, and that $A(\infty)$ has two linearly dependent columns. When $\lambda < \infty$, the left-inverse is given by

$$[A(\lambda)]^\dagger = \begin{pmatrix} r_1^T \\ r_2^T \\ r_3^T \end{pmatrix} = \begin{pmatrix} a_2^T \\ a_3^T - \lambda a_4^T \\ \sqrt{1+\lambda^2} a_4^T \end{pmatrix},$$

from which we can compute

$$\|r_1\| = 1, \qquad \|r_2\| = \|r_3\| = \sqrt{1+\lambda^2}$$

and

$$[A(\lambda)^T A(\lambda)]^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & \frac{\lambda}{\sqrt{1+\lambda^2}} \\ 0 & \frac{\lambda}{\sqrt{1+\lambda^2}} & 1 \end{pmatrix}^{-1} = \begin{pmatrix} q_1^T \\ q_2^T \\ q_3^T \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1+\lambda^2 & -\lambda\sqrt{1+\lambda^2} \\ 0 & -\lambda\sqrt{1+\lambda^2} & 1+\lambda^2 \end{pmatrix}$$

with

$$\|q_1\| = 1, \qquad \|q_2\| = \|q_3\| = \sqrt{(1+\lambda^2)(1+2\lambda^2)}.$$

Thus as $\lambda \to \infty$ the matrix $A^T(\lambda)A(\lambda)$ becomes increasingly singular, and its condition number behaves like $O(\lambda^2)$.

Consider a least squares problem $\min_y \|A(\lambda)x(\lambda) - b\|$, where the right-hand side is independent of λ and can be represented as $b = \tau_1 a_1 + \tau_2 a_2 + \tau_3 a_3 + \tau_4 a_4$. Solution vector and residual are given by

$$x(\lambda) = \begin{pmatrix} \tau_2 & \tau_3 - \lambda\tau_4 & \sqrt{1+\lambda^2}\tau_4 \end{pmatrix}^T, \qquad r = \tau_1 a_1.$$

The value of x_1 is independent of λ, and so are $\|r_1\|$, $\|q_1\|$ and $\cos\beta_1 = \tau_2/\|b\|$. Hence the sensitivity of x_1 depends solely on its size relative to x, and the distance $\|r\|/\|b\| = |\tau_1|/\|b\|$ of b to the column space of A. If, for instance, $|x_1| \gg |x_i|$ and $|\tau_1| \ll |\tau_i|$ for $i \neq 1$ then Theorem 8 says that the error in x_1 is not amplified – independent of the values of λ and the condition number of $A(\lambda)^T A(\lambda)$.

At last we derive a norm-wise upper bound from the second expression for the component-wise error in Theorem 8, which turns out to be almost identical to the well-known first-order bound.

Corollary 3 *Given a matrix A of full column rank, let $x \neq 0$ solve $\min_y \|Ay - b\|$ and let $\bar{x} \neq 0$ solve $\min_y \|(A+F)y - (b+f)\|$. Suppose $r = b - Ax \neq 0$ and $\bar{r} = b + f - (A+F)\bar{x} \neq 0$.*

If $\max\left\{ \frac{\|F\|}{\|A\|}, \frac{\|f\|}{\|b\|} \right\} \leq \epsilon$ then

$$\frac{\|\bar{x} - x\|_\infty}{\|x\|} \leq \epsilon \left[\kappa(A) \left(\frac{\|b\|}{\|A\|\,\|x\|} + \frac{\|\bar{x}\|}{\|x\|} \right) + \frac{\|\bar{r}\|}{\|r\|} \kappa^2(A) \frac{\|r\|}{\|A\|\,\|x\|} \right],$$

where

$$\frac{1}{\kappa(A)} \tan\theta \leq \frac{\|r\|}{\|A\|\,\|x\|} \leq \tan\theta$$

and θ is the angle between b and $\mathcal{R}(A)$.

Proof: Multiplying the second expression for the relative error in Theorem 8 by x_i gives

$$\|\bar{x} - x\|_\infty = \|\bar{x}\| \max_i \left\{ \|A\|\,\|r_i\| \left[\epsilon_A \cos\psi_{F,i} - \frac{\|b\|}{\|A\|\,\|\bar{x}\|} \epsilon_b \cos\psi_{f,i} \right] + \frac{\|\bar{r}\|}{\|A\|\,\|\bar{x}\|} \|q_i\|\,\|A\|^2 \epsilon_{A,r} \cos\omega_{q,i} \right\},$$

26

where

$$\epsilon_b = \frac{\|f\|}{\|b\|}, \qquad \epsilon_A = \frac{\|F\bar{x}\|}{\|A\|\,\|\bar{x}\|}, \qquad \epsilon_{A,r} = \frac{\|F^T\bar{r}\|}{\|A^T\|\,\|\bar{r}\|}$$

are all bounded above by ϵ. The proof follows from these inequalities and dividing by $\|x\|$. The upper and lower bounds on $\frac{\|r\|}{\|A\|\,\|x\|}$ were derived in Section 4.1. ∎

The upper bound on the norm-wise error from Corollary 3 resembles very much the first-order bound on page 229 in [12], where, subject to the condition that $\epsilon < \kappa(A)$,

$$\frac{\|\bar{x} - x\|}{\|x\|} \leq \epsilon \left[\kappa(A) \left(\frac{\|b\|}{\|A\|\,\|x\|} + 1 \right) + \kappa^2(A)\tan\theta \right] + O(\epsilon^2).$$

If one refrains from replacing $\frac{\|r\|}{\|A\|\,\|x\|}$ by its upper bound $\tan\theta$ then the norm-based conventional bounds are essentially tight. Unfortunately the justification is not as strong as that for linear systems because we do not have clean lower bounds on the norm-based error. In related work, Van der Sluis [27] has shown that for any least squares problem there exist perturbations that render the upper bound on the norm-based error proportional to the square of the condition number.

4.3   The Residual

In this section we briefly examine the error in individual components of the residual $r = b - Ax$. To this end denote by $P = AA^\dagger = A(A^TA)^{-1}A^T$ the orthogonal projector on the column space $\mathcal{R}(A)$ of A, and by $P^\perp = I - P$ the orthogonal projector on the orthogonal complement of $\mathcal{R}(A)$. Hence, if $r = b - Ax$ is the residual of the least squares problem $\min_y \|Ay - b\|$ with solution x, we can write $r = P^\perp b$.

Remember that the residual of the computed solution $\bar{x}$ is represented by $\bar{r} = (A + F)\bar{x} - (b + f)$ and that it is different from the 'computable residual' $b - A\bar{x}$.

Theorem 10 *Given a matrix A of full column rank, let $x \neq 0$ solve $\min_y \|Ay - b\|$ so that $r = b - Ax \neq 0$, and let $\bar{x}$ solve $\min_y \|(A + F)y - (b + f)\|$ so that $\bar{r} = b + f - (A + F)\bar{x} \neq 0$.*

Let

$$w_i = A^\dagger e_i, \qquad p_i = Pe_i, \qquad p_i^\perp = P^\perp e_i, \qquad r = \begin{pmatrix} z_1 \\ \vdots \\ z_n \end{pmatrix}, \qquad \bar{r} = \begin{pmatrix} \bar{z}_1 \\ \vdots \\ \bar{z}_n \end{pmatrix},$$

then

$$\begin{aligned}
\bar{z}_i &= z_i - \|p_i^\perp\| \left(\|F\bar{x}\|\cos\phi_{F,i} - \|f\|\cos\phi_{f,i} \right) + \|p_i\|\,\|\bar{r}\|\cos\nu_i \\
&= z_i - \|p_i^\perp\| \left(\|F\bar{x}\|\cos\phi_{F,i} - \|f\|\cos\phi_{f,i} \right) - \|w_i\|\,\|F^T\bar{r}\|\cos\nu_{w,i},
\end{aligned}$$

where $\phi_{F,i}$ is the angle between $F\bar{x}$ and $p_i^\perp$, $\phi_{f,i}$ is the angle between f and $p_i^\perp$, ν_i is the angle between p_i and $\bar{r}$, and $\nu_{r,i}$ is the angle between $F^T\bar{r}$ and w_i.

Let

$$\epsilon_b = \frac{\|f\|}{\|b\|}, \qquad \epsilon_A = \frac{\|F\bar{x}\|}{\|A\|\,\|\bar{x}\|}, \qquad \epsilon_{A,r} = \frac{\|F^T\bar{r}\|}{\|A^T\|\,\|\bar{r}\|}.$$

27

If $z_i \neq 0$ and γ_i is the angle between $p_i^\perp$ and b then

$$\frac{\bar{z}_i - z_i}{z_i} = \frac{1}{\|b\| \cos \gamma_i} \left[-\|F\bar{x}\| \cos \phi_{F,i} + \|f\| \cos \phi_{f,i} + \frac{\|p_i\|}{\sqrt{1 - \|p_i\|^2}} \|\bar{r}\| \cos \nu_i \right]$$

$$= \frac{\|\bar{r}\|}{z_i} \left[-\|p_i^\perp\| \frac{\|A\| \|\bar{x}\|}{\|\bar{r}\|} (\epsilon_A \cos \phi_{F,i} - \frac{\|b\|}{\|A\| \|\bar{x}\|} \epsilon_b \cos \phi_{f,i}) - \|w_i\| \|A\| \epsilon_{A,r} \cos \nu_{w,i} \right].$$

Proof: The proof proceeds in a manner similar to that of Theorem 8. From the proof of Theorem 8 we know

$$\begin{pmatrix} I & A \\ A^T & 0 \end{pmatrix} \begin{pmatrix} r \\ x \end{pmatrix} = \begin{pmatrix} b \\ 0 \end{pmatrix}, \qquad \begin{pmatrix} I & A+F \\ (A+F)^T & 0 \end{pmatrix} \begin{pmatrix} \bar{r} \\ \bar{x} \end{pmatrix} = \begin{pmatrix} b+f \\ 0 \end{pmatrix},$$

which implies

$$\begin{pmatrix} I & A \\ A^T & 0 \end{pmatrix} \begin{pmatrix} \bar{r} \\ \bar{x} \end{pmatrix} = \begin{pmatrix} b \\ 0 \end{pmatrix} - \begin{pmatrix} F\bar{x} - f \\ F^T\bar{r} \end{pmatrix} \quad \text{or} \quad \begin{pmatrix} \bar{r} \\ \bar{x} \end{pmatrix} = \begin{pmatrix} r \\ x \end{pmatrix} - \begin{pmatrix} I - AA^\dagger & (A^\dagger)^T \\ A^\dagger & -(A^TA)^{-1} \end{pmatrix} \begin{pmatrix} F\bar{x} - f \\ F^T\bar{r} \end{pmatrix}.$$

So,

$$\bar{r} = r - P^\perp(F\bar{x} - f) - (A^\dagger)^T F^T \bar{r}.$$

Note also that $\|p_i^\perp\|^2 = 1 - \|p_i\|^2$, due to the symmetry and idempotence of the orthogonal projectors P and $P^\perp = I - P$, Section 2.6.1 in [12]. ∎

We start by examining the first expression for the relative error in Theorem 10. The first amplification factor contains the angle γ_i between b and $\mathcal{R}^\perp(A)$. If b is close to $\mathcal{R}(A)$ then all γ_i are large and $\cos \gamma_i$ are small. Hence, a large $1/\cos \gamma_i$ signals a small component z_i of the residual r.

As already mentioned in Section 4.1, the factor $\|\bar{r}\|/\|b\|$ approximates $\|r\|/\|b\| = \sin \theta$, where θ is the angle between b and $\mathcal{R}(A)$. For the sake of completeness we briefly derive this well-known relation. On one hand, the properties of orthogonal projectors imply that

$$\|r\|^2 = \|b - Ax\|^2 = \|(I - P)b\|^2 = \|b\|^2 - \|Pb\|^2,$$

and thus

$$\frac{\|r\|}{\|b\|} = \sqrt{1 - \left(\frac{\|Pb\|}{\|b\|}\right)^2}.$$

On the other hand, the definition of an angle between two subspaces, Section 12.4.3 in [12], implies that the angle θ between b and $\mathcal{R}(A)$ satisfies

$$\cos \theta = \max_{z \neq 0} \frac{b^T}{\|b\|} \frac{Pz}{\|Pz\|} = \max_{z \neq 0} \frac{(Pb)^T Pz}{\|b\| \|Pz\|}.$$

Substituing b for z and using the Cauchy-Schwartz inequality leads to the bounds

$$\frac{\|Pb\|}{\|b\|} \leq \frac{\|Pb\|^2}{\|b\| \|Pb\|} \leq \max_{z \neq 0} \frac{(Pb)^T Pz}{\|b\| \|Pz\|} \leq \frac{\|Pb\| \|Pz\|}{\|b\| \|Pz\|} = \frac{\|Pb\|}{\|b\|},$$

and $\cos \theta = \|Pb\|/\|b\|$. Finally,

$$\frac{\|r\|}{\|b\|} = \sqrt{1 - \cos^2 \theta} = \sin \theta.$$

Therefore, $\|\bar{r}\|/\|b\|$ can be expected to be large when b is far away from the column space of A.

28

As for the third amplification factor, an analogous derivation shows that $\|p_i\| = \cos \tau_i$, where τ_i is the angle between e_i and $\mathcal{R}(A)$, so

$$\frac{\|p_i\|}{\sqrt{1 - \|p_i\|^2}} = \frac{\cos \tau_i}{\sin \tau_i} = \cot \tau_i.$$

Hence, when e_i is close to $\mathcal{R}(A)$ then the angle τ_i is small, and $\cot \tau_i$ is large. Like the first amplification factor this one also signals a small z_i.

Now we consider the second expression for the relative error in Theorem 10. The first amplification factor $\|\bar{r}\|/z_i$ represents the magnitude of z_i with regard to the whole vector r, which means that small components of r tend to be more sensitive to perturbations than large components.

Furthermore, we know that there always exists a row w_k^T of $A^\dagger$ for which $\|w_k\| \geq \frac{1}{\sqrt{n}}\|A^\dagger\|$ and $\|w_k\| \|A\| \geq \frac{1}{\sqrt{n}}\kappa(A)$. Hence there always exists a component of the residual whose sensitivity to perturbations is proportional to the condition number of A. This also shows up in the remaining amplification factor for particular right-hand side vectors. There always exists a column $p_j^\perp$ for which

$$\|p_j^\perp\| \geq \frac{1}{\sqrt{n}}\|P^\perp\| = \frac{1}{\sqrt{n}}\frac{\|P^\perp\| \|b\|}{\|b\|} \geq \frac{1}{\sqrt{n}}\frac{\|P^\perp b\|}{\|b\|} = \frac{1}{\sqrt{n}}\frac{\|r\|}{\|b\|}.$$

Therefore,

$$\|p_j^\perp\|\frac{\|A\| \|x\|}{\|r\|} \geq \frac{1}{\sqrt{n}}\frac{\|r\|}{\|b\|}\frac{\|A\| \|x\|}{\|r\|} = \frac{1}{\sqrt{n}}\frac{\|A\| \|x\|}{\|b\|}.$$

As demonstrated in Section 3.1, if b is close to a singular direction of A associated with a small vector then the factor $\|A\| \|x\|/\|b\|$ is close to the condition number $\kappa(A)$.

We conclude that at least one component of the residual in a full-rank least squares has a sensitivity proportional to the condition number of the matrix. Note that it is now the *columns* of $A^\dagger$ that determine the sensitivity of the error rather than the rows.

From the expressions in Theorem 10 one can derive an upper bound on the norm-wise relative error in the residual.

Corollary 4 *Given a matrix A of full column rank, let $x \neq 0$ solve $\min_y \|Ay - b\|$ such that $r = b - Ax \neq 0$, and let $\bar{x}$ solve $\min_y \|(A + F)y - (b + f)\|$ such that $\bar{r} = b + f - (A + F)\bar{x} \neq 0$.*

If $\max\left\{\frac{\|F\|}{\|A\|}, \frac{\|f\|}{\|b\|}\right\} \leq \epsilon$ *then*

$$\frac{\|\bar{r} - r\|_\infty}{\|b\|} \leq \epsilon\left[\|P^\perp\|\frac{\|A\| \|\bar{x}\|}{\|b\|}\left(1 + \frac{\|b\|}{\|A\| \|\bar{x}\|}\right) + \kappa(A)\frac{\|\bar{r}\|}{\|b\|}\right] \leq \epsilon\left[\kappa(A)\left(\frac{\|\bar{x}\|}{\|x\|} + \frac{\|\bar{r}\|}{\|r\|}\right) + 1\right].$$

The second upper bound is almost identical to the first-order bound (5.3.9) in [12]

$$\frac{\|\bar{r} - r\|}{\|b\|} \leq \epsilon(2\kappa(A) + 1)\min\{1, m - n\} + O(\epsilon^2).$$

The term $\min\{m - n, 1\}$ accounts for the possibility $m = n$ where, since A has full rank, $r = 0$ and $P^\perp = 0$.

4.4 Summary

In this section we have shown that the errors in individual components of a full-rank least squares problem $\min_y \|Ay - b\|$, when perturbations are allowed in the matrix and right-hand side, consist of the linear system errors plus a term whose influence depends on the relation between the right-hand side vector and the column space of the matrix (Theorem 8).

Given a matrix A of full column rank, let $x \neq 0$ solve $\min_y \|Ay - b\|$ and let $\bar{x} \neq 0$ solve $\min_y \|(A + F)y - (b + f)\|$. Furthermore, let $\bar{r} = b + f - (A + F)\bar{x} \neq 0$ and q_i be the ith row of $(A^T A)^{-1}$. If $x_i \neq 0$ the component-wise relative error can be expressed as

$$\frac{\bar{x}_i - x_i}{x_i} = L + \frac{\|\bar{x}\|}{x_i} \frac{\|\bar{r}\|}{\|A\|\,\|\bar{x}\|} \|q_i\|\,\|A\|^2 \, \epsilon_{A,r} \cos \omega_{q,i},$$

where L is the component-wise relative error for linear system solution, $\epsilon_{A,r} = \frac{\|F^T \bar{r}\|}{\|A^T\|\,\|\bar{r}\|}$, and $\omega_{q,i}$ represents an error angle.

There exists at least one component x_k of the solution vector for whom

$$\frac{1}{m} \kappa(A) \tan \theta \leq \frac{\|r\|}{\|A\|\,\|x\|} \|q_k\|\,\|A\|^2 \leq \kappa^2(A) \tan \theta,$$

where θ is the angle between b and $\mathcal{R}(A)$. If $\|\bar{r}\| \approx \|r\|$ and $\|\bar{x}\| \approx \|x\|$ then these bounds are also valid for the term $\frac{\|\bar{r}\|}{\|A\|\,\|\bar{x}\|} \|q_i\|\,\|A\|^2$.

The quantities

$$\frac{\|\bar{x}\|}{|x_i|}, \qquad \|A\|\,\|e_i^T A^\dagger\|, \qquad \frac{\|\bar{r}\|}{\|A\|\,\|\bar{x}\|} \|A\|^2\|e_i^T (A^T A)^{-1}\|, \qquad 1 \leq i \leq m,$$

are called component-wise condition numbers for the least squares problem $\min_y \|Ay - b\|$.

The expressions for the component-wise error in the residual demonstrate that there exists at least one component of the residual whose sensitivity is proportional to the condition number of the matrix (Theorem 10).

As in the case of linear system solution, we use the expressions for the component-wise errors to derive the conventional norm-based error bounds for the solution and the residual. We conclude that the conventional bounds are essentially as tight as possible but our justification is not as strong as for linear systems.

5 Underdetermined Linear Systems

In this section we discuss the solution of linear systems $Ax = b$ when A is a $n \times m$ matrix, $n \leq m$, whose rank is m. Since this system may have infinitely many solutions, we want to compute the solution of minimal norm.

5.1 Minimal-Norm Solution

Any solution x of $Ax = b$ can be uniquely represented as the sum of its constituents in the nullspace and row space of A, $x = x^K + x^R$ with $x^R \in \mathcal{R}(A^T)$ and $x^K \in \text{Ker}(A)$. If $A^\dagger = A^T(AA^T)^{-1}$ is the

right-inverse of A then $A^\dagger A$ is the orthogonal projector onto the row space of A, Section 5.5.4 in [12], and $x^R = A^\dagger A x$.

Note that the component x^R in the rowspace is the same for all x. For suppose that were not the case, then there would exist x and y with

$$Ax = b, \quad Ay = b, \quad x = x^R + x^K, \quad y = y^R + y^K, \quad y^R \neq x^R,$$

and $x^R, y^R \in \mathcal{R}(A^T)$. These conditions imply that

$$0 = A(x - y) = A(x^R - y^R) + A(x^K - y^K) = A(x^R - y^R).$$

Hence $x^R - y^R$ is in both $\mathcal{R}(A^T)$ and $\text{Ker}(A)$. But this is only possible if $x^R - y^R = 0$ due to the orthogonality of the spaces $\mathcal{R}(A^T)$ and $\text{Ker}(A)$, Section 2.6.2 in [12]. So x^R must be unique.

Since x^R is already in the row space of A, an orthogonal projection onto the row space leaves x^R unchanged, $x^R = A^\dagger A x = A^\dagger b$, and $x = x^K + A^\dagger b$. According to the orthogonality of $\text{Ker}(A)$ and $\mathcal{R}(A^T)$, $\|x\|^2 = \|x^K\|^2 + \|x^R\|^2$, which is minimised for $x^K = 0$. Therefore $x^R = A^\dagger b$ is the minimal-norm solution to the linear system $Ax = b$.

5.2 'Duality'

In the previous section we established that the minimal-norm solution x^R to a linear system $Ax = b$ of full row rank must lie in the rowspace of A, so there is a vector y such that $x^R = -A^T y$. Hence x^R satisfies

$$x^R + A^T y = 0, \qquad A x^R = b.$$

In other words, x^R is part of the solution to the non-singular linear system

$$\begin{pmatrix} I & A^T \\ A & 0 \end{pmatrix} \begin{pmatrix} x^R \\ y \end{pmatrix} = \begin{pmatrix} 0 \\ b \end{pmatrix}.$$

Now remember that the solution x^C to the least squares problem $\min_y \|Ay - b\|$ of full column rank satisfies

$$r + A x^C = b, \qquad A^T r = 0.$$

In other words, x^C is part of the solution to the non-singular linear system

$$\begin{pmatrix} I & A \\ A^T & 0 \end{pmatrix} \begin{pmatrix} r \\ x^C \end{pmatrix} = \begin{pmatrix} b \\ 0 \end{pmatrix}.$$

Therefore the solution of the full row-rank linear system and the solution of the full column-rank least squares problems constitute 'dual' problems: the norms of r and x^R are minimised, so r corresponds to x^R while x^C corresponds to y.

Thus a sensitivity analysis of the component-wise errors for the underdetermined system yields results similar to those of Chapter 3 for the least squares problem, and we will only give a brief sketch here. The exact solution x^R and the computed solution $\bar{x}^R$, satisfy the linear systems

$$\begin{pmatrix} I & A^T \\ A & 0 \end{pmatrix} \begin{pmatrix} x^R \\ y \end{pmatrix} = \begin{pmatrix} 0 \\ b \end{pmatrix}, \quad \begin{pmatrix} I & (A+F)^T \\ A+F & 0 \end{pmatrix} \begin{pmatrix} \bar{x}^R \\ \bar{y} \end{pmatrix} = \begin{pmatrix} 0 \\ b+f \end{pmatrix}.$$

Hence

$$\begin{pmatrix} \bar{x}^R \\ \bar{y} \end{pmatrix} = \begin{pmatrix} x^R \\ y \end{pmatrix} - \begin{pmatrix} I - A^\dagger A & A^\dagger \\ (A^\dagger)^T & -(AA^T)^{-1} \end{pmatrix} \begin{pmatrix} F^T \bar{y} \\ F\bar{x}^R - f \end{pmatrix}$$

and

$$\bar{x}^R = x^R - P^\perp F^T \bar{y} - A^\dagger (F\bar{x}^R - f).$$

Like the residual r in the least squares problem, the sensitivity of the components of x^R depends on the rows of the projector $P^\perp$; and like the solution x^C in the least squares problem, the sensitivity of x^R is also governed by the rows of $A^\dagger$.

The perturbation results can be applied to the computation of the right-inverse by expressing it as the minimal-norm solution to the linear system $AX = I$, where I is the $n \times n$ identity matrix.

6 Computation and Estimation of Component-Wise Condition Numbers

In this section we discuss how to compute and to estimate the component-wise condition numbers for linear systems and least squares problems, defined, respectively, in Sections 2.5 and 4.1.

Let $x \neq 0$ be the solution to the least squares problem $\min_y \|Ay - b\|$ with $n \times m$ matrix A of rank m. If $\bar{x} \neq 0$ is the computed solution with residual $\bar{r} \neq 0$ then component-wise condition numbers are

$$\frac{\|\bar{x}\|}{|x_i|}, \quad \|A\| \, \|r_i\|, \qquad \frac{\|\bar{r}\|}{\|A\| \, \|\bar{x}\|} \|A\|^2 \|q_i\|, \qquad 1 \leq i \leq m,$$

where

$$r_i^T = e_i^T A^\dagger, \quad q_i^T = e_i^T (A^T A)^{-1}.$$

The condition numbers for a linear system form a subset of those for a least squares problem.

Numerical issues in the computation of the $\|r_i\|$, due to the potential ill-conditioning of A, are addressed in [20], and in the context of statistical errors in [21]. Now let us consider the computational requirements.

The matrix two-norm $\|A\|$ can be bounded by the one-norm $\|A\|_1 = \max_{1 \leq i \leq m} \|a_i\|$ via, Section 2.3.2 in [12],

$$\frac{1}{\sqrt{n}} \|A\|_1 \leq \|A\| \leq \sqrt{m} \|A\|_1.$$

The computation of the m vector norms $\|a_i\|$ for $\|A\|_1$ requires a total of $O(mn)$ operations. The relative sizes $\|\bar{x}\|/|x_i|$ can be estimated a posteriori from the computed solution $\bar{x}$ in $O(m)$ operations. The term $\frac{\|\bar{r}\|}{\|A\| \, \|\bar{x}\|}$ can be estimated a posteriori from the computed residual $r_c = b - A\bar{x}$ in $O(mn)$ operations. This leaves the computation of the $\|r_i\|$ and $\|q_i\|$.

If a factorisation of A is available then upper bounds on the $\|r_i\|$ can be determined in $O(n^2)$ operations, as shown in Section 6, and an estimate of the $\|q_i\|$ can be obtained by making use of the inequality $\|q_i\| \geq \|r_i\|^2$ from Corollary 6 in Appendix 1.

In the following sections we discuss how to compute the $\|r_i\|$ from the QR decomposition of A, how to compute them in the special case where A is bi- or tridiagonal, and how to estimate them from a decomposition of A.

To sum up, if A^{-1} or a factorisation of A is available then the component-wise condition numbers can be computed or estimated with a total of $O(n^2)$ operations.

6.1 Computation of Condition Numbers from the QR Decomposition

Let

$$A = Q \begin{pmatrix} R \\ 0 \end{pmatrix}$$

be the QR decomposition of A, where Q is a $n \times n$ orthogonal matrix, and R is a $m \times m$ non-singular upper triangular matrix. To compute $\|r_i\|$ and $\|q_i\|$ it is sufficient to work with R instead of A, as we will now show.

We have

$$q_i^T = e_i^T (A^T A)^{-1} = e_i^T R^{-1} R^{-T} = v_i^T R^{-T},$$

where $v_i = R^{-T} e_i$. This means, once v_i has been computed, q_i can be determined from v_i by solving the triangular system $R q_i = v_i$ in $O(m^2)$ operations. As $A^\dagger = (A^T A)^{-1} A^T$ we get

$$r_i^T = q_i^T A^T = v_i^T R^{-T} A^T = (v_i^T \quad 0) Q^T,$$

so $\|r_i\|$ can be determined directly from v_i via $\|r_i\| = \|v_i\|$. Hence, $\|q_i\|$ and $\|r_i\|$ can be obtained from the ith row v_i^T of R^{-1}.

In order to accomplish this efficiently we first consider the case $i = m$. The upper triangular structure of R implies that $v_m = R^{-T} e_m = \frac{1}{\rho} e_m$, where ρ is the element of R in position (m, m). So $\|r_m\|$ can be determined from the bottom element of R via $\|r_m\| = 1/|\rho|$ – without inverting R. Substituting $v_m = \frac{1}{\rho} e_m$ in q_m yields $R q_m = \frac{1}{\rho} e_m$. Therefore, if a QR decomposition of A is available, $\|r_m\|$ is available right away and the computation of q_m requires merely the solution of a $m \times m$ triangular system.

This process can be carried out for all i, and is described in [20] for the computation of $\|r_i\|$. After choosing a permutation matrix P that moves column i of A to the last position, and performing a QR factorisation of the permuted matrix AP, proceed as for $i = m$ in order to obtain $\|r_i\|$ and $\|q_i\|$. The proof of Corollary 6 in the Appendix establishes the correctness of this procedure. One does not have to perform the QR factorisation from scratch for each permutation P. Gragg and Stewart [13] show how to efficiently 'update' the QR factorisation from one permutation to the next in $O(m^2)$ operations, see also Section 12.6 in [12].

The next section discusses the efficient computation of the $\|r_i\|$ for non-singular bidiagonal and tridiagonal matrices.

6.2 Computation of Condition Numbers for Bi- and Tridiagonal Matrices

In [14] Higham gives algorithms for computing $\|A^{-1}\|_\infty$ when A is bi- or tridiagonal. We modify these algorithms to compute $\|r_i\|$, where $r_i^T = e_i^T A^{-1}$.

When

$$A = \begin{pmatrix} a_{11} & a_{12} & & & & \\ & a_{22} & a_{23} & & & \\ & & a_{33} & \ddots & & \\ & & & \ddots & a_{m-1,m} & \\ & & & & a_{mm} \end{pmatrix}$$

is a $m \times m$ non-singular bidiagonal matrix, the elements α_{ij} of its inverse are given by [14]

$$\alpha_{ij} = \begin{cases} 0 & \text{if } i > j \\ 1/a_{ii} & \text{if } i = j \\ -a_{i,i+1}\alpha_{i+1,j}/a_{i,i} & \text{if } i < j \end{cases} .$$

These expressions and $\|r_i\|^2 = \sum_{j=i}^{m} \alpha_{ij}^2$ lead to the following modification of Algorithm 2.1 in [14] for the computation of all $\|r_i\|^2$,

$$\|r_m\|^2 = \frac{1}{a_{mm}^2}, \qquad \|r_i\|^2 = \frac{1 + a_{i,i+1}^2 \|r_{i+1}\|^2}{a_{ii}^2}, \qquad m-1 \geq i \geq 1.$$

This algorithm requires a total of $5m$ operations. It incurs no round-off error from cancellation because all quantities involved are non-negative.

When A is a $m \times m$ tridiagonal matrix we modify Algorithm 4.2 in [14] to compute the two-norm of the columns of A^{-T}. Assume that

$$A = \begin{pmatrix} a_{11} & a_{12} & & & & \\ a_{21} & a_{22} & a_{23} & & & \\ & a_{32} & a_{33} & \ddots & & \\ & & \ddots & \ddots & a_{m-1,m} & \\ & & & a_{m,m-1} & a_{mm} \end{pmatrix}$$

be irreducible, that is, $a_{i,i+1}$ and $a_{i+1,i}$ are non-zero. Otherwise one can either introduce small perturbations to make all $a_{i,i+1}$ and $a_{i+1,i}$ non-zero, or one can treat A as a block tridiagonal matrix with diagonal blocks that are irreducible tridiagonal [14].

The inverse of a tridiagonal irreducible matrix A^T can be represented by means of two vectors y and z as [5, 14, 29]

$$(A^{-T})_{ij} = \begin{cases} y_i z_j p_j & \text{if } i \leq j \\ z_i y_j p_j & \text{if } i \geq j \end{cases},$$

where

$$p_1 = 1, \qquad p_{i+1} = \prod_{j=1}^{i} \left(\frac{a_{j+1,j}}{a_{j,j+1}} \right) = p_i \frac{a_{i+1,i}}{a_{i,i+1}}, \qquad 1 \leq i \leq m-1.$$

The products p_j can be computed recursively in $2m$ operations. To illustrate the computation of y and z we write out in full the representation of A^{-T},

$$A^{-T} = \begin{pmatrix} y_1 z_1 & y_1 z_2 & y_1 z_3 & \cdots & y_1 z_m \\ y_1 z_2 & y_2 z_2 & y_2 z_3 & \cdots & y_2 z_m \\ y_1 z_3 & y_2 z_3 & y_3 z_3 & \cdots & y_3 z_m \\ \vdots & \vdots & \vdots & & \vdots \\ y_1 z_m & y_2 z_m & y_3 z_m & \cdots & y_m z_m \end{pmatrix} \begin{pmatrix} 1 & & & & \\ & p_2 & & & \\ & & p_3 & & \\ & & & \ddots & \\ & & & & p_m \end{pmatrix}.$$

This implies $A^{-T}e_1 = y_1 z$ and $A^{-T}e_m = z_m p_m y$. Set $y_1 = 1$. Since A^T is tridiagonal, one can solve for $y_2, \ldots, y_m$ from equations $1, \ldots, m-1$ of $A^T y = (z_m p_m)^{-1} e_m$; and use equation m to solve for z_m. Then use $A^T z = e_1$ to solve for $z_{m-1}, \ldots, z_1$.

34

Since column i of A^{-T} equals

$$r_i^T = p_i (\, y_1 z_i \quad y_2 z_i \quad \cdots \quad y_{i-1} z_i \quad y_i z_i \quad y_i z_{i+1} \quad \cdots \quad y_i z_m \,),$$

we have $\|r_i\|^2 = p_i^2(z_i^2 \sum_{j=1}^i y_j^2 + y_i^2 \sum_{j=i+1}^m z_j^2)$. All $\|r_i\|^2$ can now be computed in a total of $O(m)$ operations by accumulating the two sets of partial sums

$$s_1 = y_1^2, \qquad s_i = s_{i-1} + y_i^2, \qquad 2 \le i \le m,$$

and

$$t_m = z_m^2, \qquad t_i = t_{i+1} + z_i^2, \qquad m - 1 \ge i \ge 1,$$

and then combining them into

$$\|r_i\|^2 = p_i^2(z_i^2 s_i + y_i^2 t_{i+1}), \qquad 1 \le i \le m.$$

When A is a Hessenberg matrix one can probably use the expressions for A^{-1} in [5] to compute $\|r_i\|^2$ in a fashion similar to that for tridiagonal matrices.

6.3   Estimation of Component-Wise Condition Numbers

When A is a $m \times m$ triangular matrix upper bounds for the $\|r_i\|$ can be computed in $O(m^2)$ operations by making use of ideas from condition number estimators for triangular matrices [16], as we will now show. An estimate of the $\|q_i\|$ can be obtained from the inequality $\|q_i\| \ge \|r_i\|^2$ from Corollary 6 in Appendix 1.

Since A is triangular, $(A^{-1})_{ii} = 1/a_{ii}$ and $1/|a_{ii}| \le \|r_i\| \le \|r_i\|_1$. Instead of A, we will work with its *comparison matrix* $C(A) = (c_{ij})$ of A [3], which is defined as

$$c_{ij} = \begin{cases} |a_{ii}| & \text{if } i = j \\ -|a_{ij}| & \text{if } i \ne j \end{cases}$$

and satisfies the component-wise inequalities

$$C(A)^{-1} \ge 0, \qquad |A^{-1}| \le C(A)^{-1}$$

because it is an M-matrix [28]. The first inequality implies that the ith element of $C(A)^{-T}e$ equals $\|C(A)^{-T}e_i\|_1$, where e is the vector of all ones, while the second one implies $\|r_i\| \le \|r_i\|_1 \le \|C(A)^{-T}e_i\|_1$. Hence all $\|C(A)^{-T}e_i\|_1$ can be computed with a total of $O(m^2)$ operations by solving the system $C(A)^T y = e$.

When A is a general, non-singular matrix, the $\|r_i\|$ may be estimated by applying the above estimator to the LU factors of A. Let $A = LU$, where L is a lower triangular and U is an upper triangular matrix. If $C(L)$ and $C(U)$ are the respective comparison matrices of L and U then

$$|L^{-1}| \le C(L)^{-1}, \qquad |U^{-1}| \le C(U)^{-1}, \qquad |A^{-1}| \le |U^{-1}||L^{-1}| \le C(U)^{-1}C(L)^{-1}.$$

Thus, $\|r_i\| \le \|r_i\|_1 \le \|C(L)^{-T}C(U)^{-T}e_i\|_1$, where $\|C(L)^{-T}C(U)^{-T}e_i\|_1$ is the ith element of $C(L)^{-T}C(U)^{-T}e$ and e is the vector of all ones.

Therefore, if the LU decomposition of the $m \times m$ matrix A is available, an upper bound on the $\|r_i\|$ can be obtained by solving the two triangular systems $C(L)^T y = e$ and $C(U)^T z = y$ in $O(m^2)$ operations. Of course, the same is true for the Cholesky decomposition $A = L^T L$ of a symmetric positive-definite matrix A.

7 Summary

Traditonally, the error in the computed solution $\bar{x}$ of a system of linear equations $Ax = b$ has been estimated from the *norm* of the relative error $\|\bar{x} - x\|/\|x\|$.

Amomg the advantages of norm-based errors are straight-forward perturbation analyses as well as clear, simple error bounds. For example,

$$\frac{\|\bar{x} - x\|}{\|x\|} \stackrel{.}{\leq} \kappa(A, x)\epsilon$$

represents an approximate upper bound on the norm of the error in $\bar{x}$, where ϵ is the size of the relative perturbation in the data A and b, and the condition number $\kappa(A, x)$ determines the sensitivity of the solution x to perturbations in the data.

In the simplest case, when $\kappa(A, x)$ equals the two-norm condition number of the matrix A, the error bound frequently turns out to be rather pessimistic. A more realistic condition number, such as Skeel's [19], is obtained by exploiting the structure in perturbations from error analyses of algorithms for linear system solution, e.g. [2]. Still, as we illustrated in Section 1.1, the condition numbers are prone to overestimating the error in *individual* components of $\bar{x}$.

As a consequence, we decided to pursue a perturbation analysis for *individual* components of the solution x without making any assumptions about the perturbations. The resulting expressions for the component-wise relative errors $|\bar{x}_i - x_i| / |x_i|$ are simple and easy to interpret.

The terms that multiply, and possibly amplify, the perturbations in the component-wise errors are called component-wise condition numbers. Besides being amenable to nice geometric interpretations, they are able to reveal the existence of solution components that are much better conditioned than existing condition numbers would lead us to believe. Moreover, barring any restrictions on the perturbations, we showed that there is always one component of x whose condition number is proportional to $\kappa_2(A)$.

We conclude that *no* norm-based relative error bound can ever predict the presence of well-conditioned components in x, assuming no restrictions on the perturbations. Therefore, our component-wise condition numbers are *essential*.

It would be interesting to examine how restrictions on the structure or distribution of perturbations affect the component-wise condition numbers.

Acknowledgements

8 Appendix 1: Expressions for Left-Inverses

We give 'block' expressions for the left-inverse of a matrix with full column rank, which involve angles associated with columns of the matrix, and one can clearly see that each part of the left-inverse has a geometrical interpretation. We used these expressions in Section 2.4 to justify our

choice of measures of sensitivity in component-wise relative errors for the solution of linear systems and least squares problems.

In order to explain the emergence of angles, we start off with some geometric interpretations. Given a $n \times m$ matrix $A = (a_1 \ \ldots \ a_m)$, $n \geq m$, with columns a_i the matrix $A^T A$ contains information about the angles π_{ij} between individual columns a_i and a_j:

$$(A^T A)_{ij} = a_i^T a_j = \|a_i\| \|a_j\| \cos \pi_{ij},$$

see also [1], page 65.

In contrast, the inverse matrix $(A^T A)^{-1}$ provides information about the angles associated with an individual column and the subspace spanned by all other columns, as the next theorem shows. Thus, while $A^T A$ contains 'local' information about columns, the inverse provides a more 'global' view.

Theorem 11 *Let $A = (a_1 \ \ A_1)$ be a $n \times m$ matrix, $n \geq m$, of full column rank, where a_1 represents the first column of A and A_1 the remaining columns. Let c be the solution of the least squares approximation of a_1 by the columns of A_1, so*

$$\|\hat{a}_1\| = \min_y \|A_1 y - a_1\|, \qquad -\hat{a}_1 = A_1 c - a_1,$$

where $\hat{a}_1$ is the residual. Similarly, let d^T be the least squares approximation of the columns of A_1 by a_1,

$$\|\check{A}_1\| = \min_y \|a_1 y^T - A_1\| \qquad -\check{A}_1 = a_1 d^T - A_1.$$

Then

$$(A^T A)^{-1} = \begin{pmatrix} (\hat{a}_1^T a_1)^{-1} & \\ & (\check{A}_1^T A_1)^{-1} \end{pmatrix} \begin{pmatrix} 1 & -c^T \\ -d & I \end{pmatrix}.$$

Proof: Verify that multiplication of the above expressions by $A^T A$ gives the identity.

These particular expressions for the inverse are based on the derivation of a formula for partial correlation coefficients in [9]. If

$$M = \begin{pmatrix} X & Y^T \\ Y & W \end{pmatrix}$$

is a symmetric positive-definite matrix then its inverse can be written as [8]

$$M^{-1} = \begin{pmatrix} (X - Y^T W^{-1} Y)^{-1} & -X^{-1} Y (W - Y X^{-1} Y^T)^{-1} \\ -W^{-1} Y (X - Y^T W^{-1} Y)^{-1} & (W - Y X^{-1} Y^T)^{-1} \end{pmatrix}.$$

Apply this formula to

$$A^T A = \begin{pmatrix} a_1^T a_1 & a_1^T A_1 \\ A_1^T a_1 & A_1^T A_1 \end{pmatrix},$$

and use the fact [9] that the inverse of the $(1, 1)$ element of $(A^T A)^{-1}$ equals

$$a_1^T a_1 - a_1^T A_1 (A_1^T A_1)^{-1} A_1^T a_1 = a_1^T \left(I - A_1 (A_1^T A_1)^{-1} A_1^T \right) a_1 = \hat{a}_1^T a_1,$$

where $I - A_1 (A_1^T A_1)^{-1} A_1^T$ is the orthogonal projector [12], page 7ɔ, onto $\mathcal{R}^{\perp}(A_1)$.

Thus $\hat{a}_1$ is the projection of a_1 onto the orthogonal complement $\mathcal{R}^{\perp}(A_1)$ of the column space of A_1 in $\mathfrak{R}^n$, while $\check{A}_1$ is the projection of A_1 onto the orthogonal complement $\mathcal{R}^{\perp}(a_1)$ of a_1 in $\mathfrak{R}^n$,

$$\hat{a}_1 = \left(I - A_1(A_1^T A_1)^{-1}A_1^T\right)a_1, \qquad \check{A}_1 = \left(I - a_1(a_1^T a_1)^{-1}a_1^T\right)A_1.$$

∎

It is easy to show that by applying a permutation to the columns of A, the above theorem can be made to distinguish column a_i of A instead of column a_1.

Theorem 12 *With the notation of Theorem 11, the left-inverse $(A^T A)^{-1}A^T$ of A can be written as*

$$A^{\dagger} = (A^T A)^{-1}A^T = \begin{pmatrix} (\hat{a}_1^T a_1)^{-1}\hat{a}_1^T \\ (\check{A}_1^T A_1)^{-1}\check{A}_1^T \end{pmatrix}.$$

Proof: Multiply the expression in Theorem 11 by A^T. ∎

One can clearly see that the first row of $(A^T A)^{-1}A^T$ is orthogonal to $\mathcal{R}(A_1)$, while the remaining rows are orthogonal to a_1. Now we can express the elements of the left-inverse of A in terms of angles associated with columns of A.

Corollary 5 *The ith row r_i^T of the left-inverse $(A^T A)^{-1}A^T$ of A has the same direction as the residual in the least squares approximation of a_i by the remaining columns,*

$$r_i^T = e_i^T (A^T A)^{-1}A^T = \frac{1}{\hat{a}_i^T a_i}\hat{a}_i^T = \frac{1}{\|a_i\| \cos \alpha_i}\frac{1}{\|\hat{a}_i\|}\hat{a}_i^T,$$

while its length is inversely proportional to that of the residual,

$$\|\hat{a}_i\| = \|a_i\| \cos \alpha_i = \frac{1}{\|r_i\|}.$$

Because the ith row r_i of the inverse is a multiple of the residual $\hat{a}_i$, its norm is a most natural criterion for measuring how well the ith column of a matrix can be approximated by the other columns. As mentioned in Section 2.4, Stewart already proved the relationship

$$\|\hat{a}_i\| = \min_y \|A_i y - a_i\| = \frac{1}{\|r_i\|}.$$

The following corollary is needed in the derivation of the component-wise error for least squares problems in Section 4.1. It states that the length of a row in $(A^T A)^{-1}$ is greater than the square of the length of the corresponding row in $A^{\dagger}$ (they are equal when A is non-singular).

Corollary 6 *If A is a $n \times m$ matrix of rank m, $r_i^T = e_i^T A^{\dagger}$, and $q_i^T = e_i^T (A^T A)^{-1}$ then $\|q_i\| \geq \|r_i\|^2$.*

Proof: Let

$$AP = Q \begin{pmatrix} R \\ 0 \end{pmatrix}$$

38

be the QR decomposition of A with column permutations, where P is a $m \times m$ permutation matrix, Q is a $n \times n$ orthogonal matrix, and R is a $m \times m$ non-singular upper triangular matrix. If $e_i = Pe_j$ then

$$q_i^T = e_i^T(A^TA)^{-1} = e_j^T P^T(A^TA)^{-1}PP^T = e_j^T R^{-1}R^{-T}P^T = v_j^T R^{-T}P^T,$$

where $v_j = R^{-T}e_j$. As $A^\dagger = (A^TA)^{-1}A^T$ we get

$$r_i^T = q_i^T A^T = v_j^T R^{-T}P^T A^T = (v_j^T \quad 0)Q^T.$$

Hence $\|r_i\| = \|v_j\|$ and $\|q_i\| = \|R^{-1}v_j\|$.

For a fixed P consider the index i for which $j = m$. The upper triangular structure of R implies that $v_m = R^{-T}e_m = \frac{1}{\rho}e_m$, where ρ is the element of R in position (m, m). So $\|r_i\| = 1/|\rho|$. Substituting $v_m = \frac{1}{\rho}e_m$ in q_i yields $q_i = PR^{-1}v_m = \frac{1}{\rho}PR^{-1}e_m$. Hence $\|q_i\| \geq 1/\rho^2$ and $\|q_i\| \geq \|r_i\|^2$.

In order to prove the corollary for all rows, choose a sequence of permutation matrices such that $e_i = Pe_m$ for $1 \leq i \leq m$. ■

Remark 2 *The angles α_i associated with the columns of A are equal to the corresponding angles associated with the rows of $(A^TA)^{-1}A^T$.*
To see why, recall that α_i is the angle between a_i and r_i, where r_i constitute the columns of $B = \left[(A^TA)^{-1}A^T\right]^T$, and A^T is the right-inverse of B with rows a_i^T. Therefore the relevant angles are those between r_i and a_i, which are just α_i.

Appendix 2: More Perturbation Results

Here we derive perturbation results for linear systems that, unlike those in Section 3, do not contain computed quantities in the error expressions. This is possible because the perturbations are expressed differently.

Let $\bar{x}$ be the computed solution of the linear system $Ax = b$ and the exact solution of $(A+F)\bar{x} = b$. In order to assess in more detail the effect of perturbations in the matrix on the component-wise relative error, we distinguish two cases regarding perturbations in the matrix A: perturbations confined to the column corresponding to x_i, and perturbations of all columns except for the one corresponding to x_i.

Denote by a_i and f_i the respective columns of A and $A + F$,

$$A = (a_1 \quad \ldots \quad a_{i-1} \quad a_i \quad a_{i+1} \quad \ldots a_m), \qquad F = (f_1 \quad \ldots \quad f_{i-1} \quad f_i \quad f_{i+1} \quad \ldots f_m).$$

We distinguish the ith columns a_i and f_i from the remaining columns by introducing

$$A_i = (a_1 \quad \ldots \quad a_{i-1} \quad a_{i+1} \quad \ldots \quad a_m), \qquad F_i = (f_1 \quad \ldots \quad f_{i-1} \quad f_{i+1} \quad \ldots \quad f_m).$$

As before, α_i denotes the angle between a_i and and its projection on $\text{span}^\perp_{k \neq i}\{a_k\}$, while β_i denotes the angle between b and the projection of a_i onto $\text{span}^\perp_{k \neq i}\{a_k\}$. Due to the particular expressions for the perturbations, the following results are formulated in terms of the projections $\hat{a}_i$ of a_i onto $\text{span}^\perp_{k \neq i}\{a_k\}$, rather than in terms of the rows r_i^T of $A^\dagger$. This represents only a small change, as according to Corollary 5 in Appendix 1 $\|\hat{a}_i\| = \|a_i\|\cos\alpha_i = 1/\|r_i\|$.

First consider the case when only column a_i is perturbed, that is, $F_i = 0$. In Section 6 of [21] Stewart discusses this situation in the context of errors in regression variables. Due to his assumptions with regard to the statistical nature of the errors, it is difficult to compare his and our results.

The following lemma, whose proof appears at the end of this section, applies to both linear systems and least squares problems. It shows that in case of linear systems the *effect* on x_i of the perturbation f_i is making itself felt in terms of its size $\|f_i\|$ and in terms of its distance to the space of the other columns. In some sense, the effect of f_i is confined to column a_i – although the direction of f_i itself is arbitrary. The reason is that the space spanned by the remaining columns has dimension $n - 1$. So its orthogonal complement $\text{span}^{\perp}_{k \neq i}\{a_k\}$ has dimension one. But both, a_i and f_i, are projected into this one-dimensional space in order to make up $\bar{x}_i$. Therefore, the effect of any perturbation in the ith column is confined to the one-dimensional space $\text{span}^{\perp}_{k \neq i}\{a_k\}$. Consequently, unless $a_i + f_i$ is zero, there is no question about the right-hand side b remaining in the column space of A, so $\cos \beta_i$ does not enter the relative error for $\bar{x}_i$.

In the case of least squares problems, the effect of f_i still remains confined to column a_i but now the relation between perturbation and right-hand side matters as well, and $\cos \beta_i$ enters the picture. The smaller the contribution of a_i to b outside the space of the other columns, the more the angle between f_i and b matters. Due to the quadratic nature of the least squares problems we get squared condition numbers in front of second-order error terms.

Lemma 1 *Given matrices A and $A + F$ of full column rank with $F_i = 0$, let x solve $\min_y \|Ay - b\|$ and $\bar{x}$ solve $\min_y \|(A + F)y - b\|$. Let ϕ_i be the angle between f_i and the projection $\hat{a}_i$ of a_i onto $\text{span}^{\perp}_{k \neq i}\{a_k\}$, and $\rho_i = \|f_i\|/\|a_i\|$.*

If $x_i = 0$ then $\bar{x}_i = 0$.

If $x_i \neq 0$ then

$$\bar{x}_i = x_i \frac{1 + \rho_i \frac{\cos \phi_{f,i} \cos \phi_{b,i}}{\cos \alpha_i \cos \beta_i}}{1 + 2\rho_i \frac{\cos \phi_i}{\cos \alpha_i} + \rho_i^2 \frac{\cos^2 \phi_{f,i}}{\cos^2 \alpha_i}}, \qquad \frac{\bar{x}_i - x_i}{x_i} = -\frac{\rho_i}{\cos \alpha_i} \frac{2 \cos \phi_i + \rho_i \frac{\cos^2 \phi_{f,i}}{\cos \alpha_i} - \cos \phi_{f,i} \frac{\cos \phi_{b,i}}{\cos \beta_i}}{1 + 2\rho_i \frac{\cos \phi_i}{\cos \alpha_i} + \rho_i^2 \frac{\cos^2 \phi_{f,i}}{\cos^2 \alpha_i}},$$

where $\phi_{b,i}$ is the angle between b and the projection $\hat{f}_i$ of f_i onto $\text{span}^{\perp}_{k \neq i}\{a_k\}$, and $\phi_{f,i}$ is the angle between f_i and $\hat{f}_i$.

If $x_i \neq 0$ and A is non-singular then

$$\bar{x}_i = x_i \frac{1}{1 + \frac{\cos \phi_i}{\cos \alpha_i} \rho_i}, \qquad \frac{\bar{x}_i - x_i}{x_i} = -\frac{\rho_i}{\cos \alpha_i} \frac{\cos \phi_i}{1 + \frac{\cos \phi_i}{\cos \alpha_i} \rho_i}.$$

The relative perturbation $\rho_i \cos \phi_i$ associated with a linear system is amplified by the factor $1/\cos \alpha_i$, which is independent of the righthand side in contrast to Corollary 1. Therefore, the more the corresponding column a_i lies in the space $\text{span}_{k \neq i}\{a_k\}$ spanned by the remaining columns, the larger the relative error in the ith component of $\bar{x}$.

With the abbreviations $\kappa_i = 1/\cos \alpha_i$ and $\rho_i^c = -\rho_i \cos \phi_i$, the above component-wise relative error for non-singular linear systems can be written as

$$\frac{\bar{x}_i - x_i}{x_i} = \frac{\kappa_i \rho_i^c}{1 - \kappa_i \rho_i^c},$$

40

which closely resembles the norm-based relative error for the perturbed system $(A + F)\bar{x} = b$

$$\frac{\|\bar{x} - x\|}{\|x\|} \leq \frac{\kappa(A)\rho(A)}{1 - \kappa(A)\rho(A)}, \qquad \rho(A) = \frac{\|F\|}{\|A\|}.$$

Now consider the case when all columns, except for the one associated with x_i, are perturbed, that is $f_i = 0$. Given the original linear system $Ax = b$, where A is a $n \times n$ non-singular matrix, let the perturbed system $(A + F)\bar{x} = b + f$ satisfy $\|A^{-1}\|\,\|F\| < 1$. This means that $A + F$ is also non-singular because $A + F = A(I + A^{-1}F)$, and $I + A^{-1}F$ is non-singular if $\|A^{-1}F\| \leq \|A^{-1}\|\,\|F\| < 1$ [12], Lemma 2.3.3. The proof of the following lemma appears at the end of this section.

Lemma 2 *Let $Ax = b$ and $(A + F)\bar{x} = b$ with $f_i = 0$, and $\|A^{-1}\|\,\|F\| < 1$.*

Define the matrix
$$\Delta_i = F_i \left(I + (A_i^T A_i)^{-1} A_i^T F_i\right)^{-1} (A_i^T A_i)^{-1} A_i^T,$$

where
$$\|\Delta_i\| \leq \frac{\kappa(A)\rho_A}{1 - \kappa(A)\rho_A}, \qquad \rho_A = \frac{\|F\|}{\|A\|},$$

and the related relative errors
$$\rho_{a,i} = \frac{\|\Delta_i a_i\|}{\|a_i\|}, \qquad \rho_{b,i} = \frac{\|\Delta_i b\|}{\|b\|}.$$

Let $\phi_{a,i}$ be the angle between $\Delta_i a_i$ and the projection $\hat{a}_i$ of a_i onto $\mathrm{span}^{\perp}_{k \neq i}\{a_k\}$, and $\phi_{b,i}$ the angle between $\Delta_i b$ and $\hat{a}_i$.

If $x_i \neq 0$ then
$$\bar{x}_i = \frac{\hat{a}_i^T (I - \Delta_i) b}{\hat{a}_i^T (I - \Delta_i) a_i} = x_i \frac{1 - \frac{\cos \phi_{b,i}}{\cos \beta_i} \rho_{b,i}}{1 - \frac{\cos \phi_{a,i}}{\cos \alpha_i} \rho_{a,i}}, \qquad \frac{\bar{x}_i - x_i}{x_i} = \frac{\frac{\cos \phi_{a,i}}{\cos \alpha_i} \rho_{a,i} - \frac{\cos \phi_{b,i}}{\cos \beta_i} \rho_{b,i}}{1 - \frac{\cos \phi_{a,i}}{\cos \alpha_i} \rho_{a,i}}.$$

If $x_i = 0$ then
$$\bar{x}_i = -\frac{\frac{\|\Delta_i b\| \cos \phi_{b,i}}{\|a_i\| \cos \alpha_i}}{1 - \frac{\|\Delta_i a_i\| \cos \phi_{a,i}}{\|a_i\| \cos \alpha_i}}.$$

The vectors $\Delta_i a_i$ and $\Delta_i b$ are in the column space of F_i. If the column space of F_i is a subset of the column space of A, then F_i is orthogonal to $\mathrm{span}^{\perp}_{k \neq i}\{a_k\}$, and $\cos \phi_{a,i}$ and $\cos \phi_{b,i}$ are zero, which implies $\bar{x}_i = x_i$. The angles $\phi_{a,i}$ and $\phi_{b,i}$ are bounded above by the angle between $\mathrm{span}^{\perp}_{k \neq i}\{a_k\}$ and the column space of F_i. As before, the relative error in the ith component of $\bar{x}$ is the larger the more the right-hand side b and the corresponding column a_i lie in the space spanned by the remaining columns.

Proof of Lemma 1

From Theorem 1 and Corollary 5 we have
$$\bar{x}_i = \frac{(\hat{a}_i + \hat{f}_i)^T b}{(\hat{a}_i + \hat{f}_i)^T (a_i + f_i)},$$

where $\hat{a}_i$ and $\hat{f}_i$ are the respective projections of a_i and f_i onto $\text{span}^{\perp}_{k \neq i}\{a_k\}$.

We first prove the easier case when A is non-singular and $\text{span}^{\perp}_{k \neq i}\{a_k\}$ has dimension one. So $\hat{f}_i$ must be a multiple of $\hat{a}_i$, i.e. $\hat{f}_i = \lambda \hat{a}_i$ for some real number λ. Therefore

$$\bar{x}_i = \frac{(1 + \lambda)\|\hat{a}_i\|\, \|b\| \cos \beta_i}{(1 + \lambda)\hat{a}_i^T(a_i + f_i)} = \frac{\|\hat{a}_i\|\, \|b\| \cos \beta_i}{\hat{a}_i^T(a_i + f_i)},$$

where β_i is the angle between $\hat{a}_i$ and b.

If ϕ_i is the angle between f_i and $\hat{a}_i$, then the denominator of $\bar{x}_i$ equals

$$\hat{a}_i^T(a_i + f_i) = \|\hat{a}_i\|\, \|a_i\| \cos \alpha_i + \|\hat{a}_i\|\, \|f_i\| \cos \phi_i.$$

Therefore,

$$\bar{x}_i = \frac{\|b\| \cos \beta_i}{\|a_i\| \cos \alpha_i + \|f_i\| \cos \phi_i} = x_1 \frac{1}{1 + \frac{\cos \phi_i}{\cos \alpha_i}\frac{\|f_i\|}{\|a_i\|}},$$

since $x_i = \|b\| \cos \beta_i / (\|a_i\| \cos \alpha_i)$ by Theorem 1 and Corollary 5.

In the general case when $\text{span}^{\perp}_{k \neq i}\{a_k\}$ may have dimension greater than one, the projection $\hat{f}_i$ is not a multiple of $\hat{a}_i$. So,

$$\bar{x}_i = x_i \frac{1 + \frac{\hat{f}_i^T b}{\hat{a}_i^T b}}{1 + 2\frac{\hat{a}_i^T f}{\hat{a}_i^T a_i} + \frac{\hat{f}_i^T f_i}{\hat{a}_i^T a_i}}.$$

Denote by $\phi_{b,i}$ the angle between $\hat{f}_i$ and b, and by $\phi_{f,i}$ the angle between $\hat{f}_i$ and f_i. So,

$$\hat{f}_i^T b = \|\hat{f}_i\|\, \|b\| \cos \phi_{b,i}, \qquad \hat{f}_i^T f_i = \|\hat{f}_i^T\|\, \|f_i\| \cos \phi_{f,i}.$$

From Corollary 5 we know that $\|\hat{a}_i\| = \|a_i\| \cos \alpha_i$. In a similar fashion we can show that $\|\hat{f}_i\| = \|f_i\| \cos \phi_{f,i}$. This gives the expression for the relative error in the general case.

Clearly $\bar{x}_i = 0$ whenever $x_i = 0$.

Proof of Lemma 2

Similar to the proofs in Appendix 1, it suffices to show the statement for $\bar{x}_1$.

From Theorem 1 and Corollary 5 we know $\bar{x}_1 = a_1^T \overline{P} b / a_1^T \overline{P} a_1$, where

$$\overline{P} = I - (A_1 + F_1)\left((A_1 + F_1)^T(A_1 + F_1)\right)^{-1}(A_1 + F_1)^T.$$

We consider numerator and denominator of $\bar{x}_1$ separately.

In order to get rid of the inverse inside the projector $\overline{P}$ we would like to apply the Sherman-Morrison-Woodbury formula [12], page 51, in such a way that the inverse of the matrix sum is a scalar. This is possible if the sum consists of rank-one matrices.

To this end decompose F_1 into its components in $\text{span}_{k \neq 1}\{a_k\}$ and $\text{span}^{\perp}_{k \neq 1}\{a_k\}$,

$$F_1 = A_1(A_1^T A_1)^{-1}A_1^T F_1 + \hat{F}_1, \qquad \text{where} \quad \hat{F}_1 = (I - A_1(A_1^T A_1)^{-1}A_1^T)F_1.$$

Hence

$$A_1 + F_1 = A_1 Z + \hat{F}_1, \qquad \text{where} \quad Z = I + (A_1^T A_1)^{-1} A_1^T F_1,$$

and

$$(A_1 + F_1)^T (A_1 + F_1) = (A_1 Z + \hat{F}_1)^T (A_1 Z + \hat{F}_1) = Z^T A_1^T A_1 Z + \hat{F}_1^T \hat{F}_1$$

since $A_1^T \hat{F}_1 = 0$. By means of the singular value decomposition of A_1 and the interlacing of singular values of A_1 and A, [12], Section 8.3.1, we can show that $\|(A_1^T A_1)^{-1} A_1^T F_1\| \leq \|A^{-1}\| \|F\| < 1$, so by Lemma 2.3.3 in [12] Z is non-singular. Thus,

$$\overline{P} = I - (A_1 + \hat{F}_1 Z^{-1})(A_1^T A_1 + Z^{-T} \hat{F}_1^T \hat{F}_1 Z^{-1})^{-1}(A_1 + \hat{F}_1 Z^{-1})^T.$$

As each column of $\hat{F}_1$ is in $\text{span}_{k \neq 1}^{\perp}\{a_k\}$, and $\text{span}_{k \neq 1}^{\perp}\{a_k\}$ has dimension one, $\hat{F}_1$ has rank one and we can write $\hat{F}_1 = \hat{a}_1 f^T$ for some $(n-1) \times 1$ vector f (the fact that, in case of least squares problems, the rank of F exceeds one has sofar prevented us from proving this theorem in the more general case). Now we can apply the Sherman-Morrison-Woodbury formula to

$$(A_1^T A_1 + Z^{-T} \hat{F}_1^T \hat{F}_1 Z^{-1})^{-1} = (A_1^T A_1 + \hat{a}_1^T \hat{a}_1 Z^{-T} f f^T Z^{-1})^{-1}$$
$$= (A_1^T A_1)^{-1} - \hat{a}_1^T \hat{a}_1 (A_1^T A_1)^{-1} Z^{-T} f [1 + \hat{a}_1^T \hat{a}_1 f^T Z^{-1} (A_1^T A_1)^{-1} Z^{-T} f]^{-1} f^T Z^{-1} (A_1^T A_1)^{-1}$$

Distinguishing the scalars

$$\sigma = \hat{a}_1^T \hat{a}_1 = \hat{a}_1^T a_1, \qquad \zeta = f^T Z^{-1} (A_1^T A_1)^{-1} Z^{-T} f$$

gives

$$\overline{P} = I - (A_1 + \hat{F}_1 Z^{-1}) \left((A_1^T A_1)^{-1} - \frac{\sigma}{1 + \sigma \zeta} (A_1^T A_1)^{-1} Z^{-T} f f^T Z^{-1} (A_1^T A_1)^{-1} \right) (A_1 + \hat{F}_1 Z^{-1})^T.$$

Multiply the terms in the product and use $y = A_1 (A_1^T A_1)^{-1} Z^{-T} f$ to get

$$\overline{P} = I - \left(A_1 (A_1^T A_1)^{-1} A_1^T + \frac{1}{1 + \sigma \zeta} (\hat{a}_1 y^T - \sigma y y^T + y \hat{a}_1^T + \zeta \hat{a}_1 \hat{a}_1^T) \right).$$

The numerator of $\hat{x}_1$ equals

$$a_1^T \overline{P} b = \frac{1 - a_1^T y}{1 + \sigma \zeta} (\hat{a}_1^T b - \sigma y^T b) = \frac{1 - a_1^T y}{1 + \sigma \zeta} \hat{a}_1^T (I - \hat{F}_1 Z^{-1}(A_1^T A_1)^{-1} A_1^T) b = \hat{a}_1^T (I - \Delta_1) b,$$

where $\Delta_1 = F_1 Z^{-1}(A_1^T A_1)^{-1} A_1^T$. Similarly, the denominator equals

$$a_1^T \overline{P} a_1 = \frac{1 - a_1^T y}{1 + \sigma \zeta} (\hat{a}_1^T a_1 - \sigma y^T a_1) = \frac{1 - a_1^T y}{1 + \sigma \zeta} \hat{a}_1^T (I - \hat{F}_1 Z^{-1}(A_1^T A_1)^{-1} A_1^T) a_1 = \hat{a}_1^T (I - \Delta_1) a_1.$$

Therefore

$$\bar{x}_1 = \frac{\hat{a}_1^T (I - \Delta_1) b}{\hat{a}_1^T (I - \Delta_1) a_1},$$

which represents the first equality for x_1 in the statement of the theorem.

Let β_1 be the angle between $\hat{a}_1$ and b, and $\phi_{b,1}$ be the angle between $\hat{a}_1$ and $\Delta_1 b$. The numerator of x_1 satisfies

$$\hat{a}_1^T b - \hat{a}_1^T \Delta_1 b = \|\hat{a}_1\| \|b\| \cos \beta_1 - \|\hat{a}_1\| \|\Delta_1 b\| \cos \phi_{b,1} = \|\hat{a}_1\|(\|b\| \cos \beta_1 - \|\Delta_1 b\| \cos \phi_{b,1}).$$

Similarly, with α_1 being the angle between $\hat{a}_1$ and a_1 and with $\phi_{a,1}$ being the angle between $\hat{a}_1$ and $\Delta_1 a_1$, the denominator satisfies

$$\hat{a}_1^T a_1 - \hat{a}_1^T \Delta_1 a_1 = \|\hat{a}_1\| \|a_1\| \cos \alpha_1 - \|\hat{a}_1\| \|\Delta_1 a_1\| \cos \phi_{a,1} = \|\hat{a}_1\|(\|a_1\| \cos \alpha_1 - \|\Delta_1 a_1\| \cos \phi_{a,1}).$$

Hence,

$$\bar{x}_1 = \frac{\|b\| \cos \beta_1 - \|\Delta_1 b\| \cos \phi_{b,1}}{\|a_1\| \cos \alpha_1 - \|\Delta_1 a_1\| \cos \phi_{a,1}}.$$

If $x_1 = 0$ then

$$\bar{x}_1 = -\frac{\|\Delta_1 b\| \cos \phi_{b,1}}{\|a_1\| \cos \alpha_1 - \|\Delta_1 a_1\| \cos \phi_{a,1}},$$

and otherwise

$$\bar{x}_1 = \frac{\|b\| \cos \beta_1}{\|a_1\| \cos \alpha_1} \frac{1 - \frac{\|\Delta_1 b\|}{\|b\|} \frac{\cos \phi_{b,1}}{\cos \beta_1}}{1 - \frac{\|\Delta_1 a_1\|}{\|a_1\|} \frac{\cos \phi_{a,1}}{\cos \alpha_1}} = x_1 \frac{1 - \frac{\|\Delta_1 b\|}{\|b\|} \frac{\cos \phi_{b,1}}{\cos \beta_1}}{1 - \frac{\|\Delta_1 a_1\|}{\|a_1\|} \frac{\cos \phi_{a,1}}{\cos \alpha_1}}$$

by Theorem 1 and Corollary 5. This establishes the second equality for x_1 in the statement of the theorem and the expression for the relative error in x_1.

References

[1] T. ANDERSON, *An Introduction to Multivariate Statistical Analysis, Second Edition*, John Wiley and Sons, 1984.

[2] M. ARIOLI, J. DEMMEL, AND I. DUFF, *Solving sparse linear systems with sparse backward error*, SIAM J. Matrix Anal. Appl., 10 (1989), pp. 165–90.

[3] A. BERMAN AND R. PLEMMONS, *Nonnegative Matrices in the Mathematical Sciences*, Academic Press, 1979.

[4] A. BJÖRCK, *Component-wise perturbation analysis and error bounds for linear least squares solutions*, BIT, 31 (1991), pp. 238–44.

[5] W. CAO AND W. STEWART, *A note on inverses of Hessenberg-like matrices*, Linear Algebra and its Applications, 76 (1986), pp. 233–40.

[6] T. CHAN, *Rank revealing QR factorizations*, Linear Algebra and its Applications, 88/89 (1987), pp. 67–82.

[7] T. CHAN AND D. FOULSER, *Effectively well-conditioned linear systems*, SIAM J. Sci. Stat. Comput., 9 (1988), pp. 963–69.

[8] R. COTTLE, *Manifestations of the Schur complement*, Linear Algebra and its Applications, 8 (1974), pp. 189–211.

[9] J. DELOSME AND I. IPSEN, *From Bareiss' algorithm to the stable computation of partial correlations*, Journal of Computational and Applied Mathematics, 27 (1989), pp. 53–91. Also in: *Parallel Algorithms for Numerical Linear Algebra (Advances in Parallel Computing, 1)*, H. van der Vorst and P. van Dooren, eds., North Holland, 1990.

[10] L. FOSTER, *Rank and null space calculations using matrix decomposition without column interchanges*, Linear Algebra and its Applications, 74 (1986), pp. 47–71.

[11] G. GOLUB, V. KLEMA, AND G. STEWART, *Rank degeneracy and least squares problems*, Tech. Report STAN-CS-76-559, Computer Science Department, Stanford University, 1976.

[12] G. GOLUB AND C. VAN LOAN, *Matrix Computations*, The Johns Hopkins Press, 1989.

[13] W. GRAGG AND G. STEWART, *A stable variant of the secant method for solving nonlinear equations*, SIAM J. Numer. Anal., 13 (1976), pp. 889–903.

[14] N. HIGHAM, *Efficient algorithms for computing the condition number of a tridiagonal matrix*, SIAM J. Sci. Stat. Comput., 7 (1986), pp. 150–65.

[15] ———, *Error analysis of the Björck-Pereyra algorithms for solving Vandermonde systems*, Numer. Math., 50 (1987), pp. 613–32.

[16] ———, *A survey of condition number estimation for triangular matrices*, SIAM Review, 29 (1987), pp. 575–96.

[17] R. HORN AND C. JOHNSON, *Matrix Analysis*, Cambridge University Press, 1985.

[18] J. RIGAL AND J. GACHES, *On the compatibility of a given solution with the data of a linear system*, JACM, 14 (1967), pp. 543–8.

[19] R. SKEEL, *Scaling for numerical stability in gaussian elimination*, JACM, 26 (1979), pp. 494–526.

[20] G. STEWART, *Rank degeneracy*, SIAM J. Sci. Stat. Comput., 5 (1984), pp. 403–13.

[21] ———, *Collinearity and least squares regression*, Statistical Science, 2 (1987), pp. 68–100.

[22] ———, *Stochastic perturbation theory*, Siam Review, 32 (1990), pp. 579–610.

[23] G. STEWART AND J. SUN, *Matrix Perturbation Theory*, Academic Press, 1990.

[24] A. VAN DER SLUIS, *Condition numbers and equilibration of matrices*, Numer. Math., 14 (1969), pp. 14–23.

[25] ———, *Condition, equilibration and pivoting in linear algebraic systems*, Numer. Math., 14 (1970), pp. 74–86.

[26] ———, *Stability of solutions of linear algebraic systems*, Numer. Math., 14 (1970), pp. 246–51.

[27] ———, *Stability of the solutions of linear least squares problems*, Numer. Math., 23 (1974), pp. 241–54.

[28] R. VARGA, *Matrix Iterative Analysis*, Prentice-Hall, 1962.

[29] T. YAMAMOTO AND Y. IKEBE, *Inversion of band matrices*, Linear Algebra and its Applications, 24 (1979), pp. 105–11.