Consiglio Nazionale delle Ricerche

ISTITUTO DI ELABORAZIONE
DELLA INFORMAZIONE

PISA

ERROR ESTIMATES IN SOLVING LINEAR SYSTEMS

M. Arioli, F. Romani

Nota interna B85-06
Agosto 1985

ERROR ESTIMATES IN SOLVING LINEAR SYSTEMS

M. ARIOLI and F. ROMANI
Istituto di Elaborazione dell'Informazione - CNR,
Via S.Maria 46, PISA, ITALY.

ABSTRACT

Given a linear system $A \underline{x} = \underline{b}$, with a real square nonsingular coefficient matrix, the error on the solution $\underline{x}$ is studied with respect to data perturbations and rounding errors of the computation.

Assuming local errors to be independent random variables, the expected value of the total error is computed as a function of $\underline{x}$, say $e(\underline{x})$. The mean of $e(\underline{x})$ in the unitary ball is then computed, obtaining statistical estimates to the errors. Moreover, the influence of diagonal scaling on the stability of the computation is studied.

These results are applied to the solution of triangular systems, to Gaussian elimination and orthogonalization techniques.

Subject Classification: AMS(MOS) 65G05, 65F05; CR: G.1.3.

1. INTRODUCTION AND PRELIMINARIES

Let us consider a linear system $A \underline{x} = \underline{b}$ where $A = (a_{ij})$ is a real nonsingular nxn matrix. In this work we study the behaviour of the error on the solution $\underline{x}$ taking into account the perturbations of input data and the rounding errors of the computation.

In Section 2 statistical estimates of the mean square error, in presence of data perturbations, are derived. In Section 3 these results are used to estimate the influence of local errors in the solution of a linear system. Section 4 deals with the scaling of the problem data and the effects of equilibration on the total error. In Section 5 the algorithmic errors in the solution of triangular systems are analyzed and in Sections 6 and 7 Gaussian elimination and orthogonalization techniques are studied.

The notation $A^{-1} = Z = (z_{ij})$ is used. The symbol $\| \cdot \|_p$ denotes the Holder p-norm (p real positive number or infinite) and $\| \cdot \|_F$ denotes the Frobenius norm of matrices as well as of three-way arrays (i.e. the square root of the sum of the squares of all the entries). Moreover $|Y|$ denotes the array of the absolute values of the entries of Y, and the symbol * denotes the Hadamard product (i.e. componentwise multiplication) between two arrays of the same size; $\rho(M)$ denotes the spectral radius of a square matrix M. All summation indices are intended to range from 1 to n when not otherwise indicated.

E(y) denotes the expected values of the random variable y; when $\underline{y}$ is an array of random variables, E($\underline{y}$) is the array of the expected values of the single entries.

The classical condition number of a nonsingular matrix is

$$k_p(A) = ||Z||_p \; ||A||_p .$$

We will also use the Skeel condition number which is defined as

$$C_p(A) = || \; |Z| \; |A| \; ||_p \qquad [12].$$

Let us now define the domain B_n and its measure as

$$B_n = \{ \underline{x} \mid ||\underline{x}||_2 = 1 \}, \qquad \Gamma = \int_{B_n} d\underline{x}.$$

Then the mean of a vectorial function f($\underline{x}$), assuming $\underline{x}$ to be uniformely distributed in the unitary ball, can be written

$$\mathop{\mathrm{Mean}}_{||\underline{x}||_2=1} f(\underline{x}) = \Gamma^{-1} \int_{B_n} f(\underline{x}) \; d\underline{x}.$$

The theoretical results derived in this paper have been extensively tested with numerical experiments. Four classes of matrices have been chosen, i.e. Vandermonde, Rice [10], Toeplitz and random matrices. For each class a parameter generates different matrices. Additional classes can be derived by diagonal scaling of the original matrices. Four types of scaling have been used:

1) column perturbation,

$$a'_{ij} = a_{ij} \, c^{j},$$

where c > 1 is a constant empirically chosen.

2) row perturbation

$$a'_{ij} = c^{i} \, a_{ij},$$

where c > 1 is a constant empirically chosen.

3) column scaling,

$a'_{ij} = a_{ij} \, d_j$, with d_j chosen to minimize $C_\infty(A')$.

4) row scaling,

$a'_{ij} = d_i \, a_{ij}$, with d_i chosen to minimize $k_\infty(A')$.

In the following graphs the abscissas represent the base 10 logarithm of the Skeel condition number (using the maximum norm) of the various matrices, and the ordinates, also in logarithmic scale, may represent other condition numbers or the errors resulting from the application of an algorithm. Data points are connected by straight lines to evidentiate the behaviour of the matrices in the same class.

In order to evaluate the mean algorithmic error, m linear systems are solved with the solution vectors randomly chosen in the unitary ball with uniform distribution. The following quantity is then computed and plotted

$$\mathrm{eps}^{-1}\left(m^{-1}\sum_{j=1}^{m}||\eta^{(j)}||_2^2/||x^{(j)}||_2^2\right)^{1/2} =$$

$$= \mathrm{eps}^{-1}\left(m^{-1}\sum_{j=1}^{m}\sum_{i=1}^{n}\eta_i^{(j)\,2}\right)^{1/2},$$

where $\eta^{(j)} = (\eta_i^{(j)})$ is the vector of errors of the solution of the j-th system, and eps is the machine precision related to the word length used to represent the matrices. For t-digit β-base floating point arithmetic with rounding one has

$\mathrm{eps} = \beta^{1-t}/2$ see [15, p. 6]).

In our experiments m=100.

When local errors are produced by the representation of real numbers in the computer or by single arithmetic operations, the quantity eps is related with to the mean m and the variance s² of the resulting errors. More in detail, using a floating point arithmetic with rounding it is common to assume that local representation and roundoff relative errors are independent random variables, uniformely distributed between -eps/2 and eps/2 [7,8,9]. Therefore we can write

$$(1.1)\qquad \begin{cases} m = 0; \\ s^2 = \mathrm{eps}^2/12. \end{cases}$$

As noted by Oppenheim [9], empirical studies have shown that the distribution is not quite uniform, so that s² is proportional to eps² with a proportionality constant slightly

less than 1/12.

In performing matrix operations the accuracy of the result is limited by the finite precision of the arithmetic. The choice of word length influences both the amount of space required to store the matrices and the time spent in computations. The trivial choice is to use the same word length both to store the matrices and to perform the operations. In this case, usually many digits of the intermediate matrices involved in the computation and of the result are less of significance. A classical alternative consists in using multiple precision arithmetic to perform the most critical operations (e.g. the accumulation of scalar products) [17,2]. Recently some authors proposed a technique which allows computing arithmetic expressions to least significant bit accuracy at the expense of a little computational overhead [6,11].

Both these techniques allow a proper use of the computer storage and a better control of the errors, moreover on the modern computers, the resulting computational overhead is not too high.

On the basis of these considerations we will assume in the following that elementary operations on matrices are performed in multiple precision or with maximal accuracy arithmetic so that all the digits in the intermediate matrices representation are accurate. This implies that relative errors on these matrices can be considered independent, uniformely distributed random variables with mean 0 and variance eps²/12.

2. PROPAGATION ERROR IN SOLVING LINEAR SYSTEMS

Let us recall here some known results about the error produced in the evaluation of rational functions.

Let $f(v_1, v_2, \ldots, v_n)$ be a rational function of n variables, computed with the following straight line algorithm

$$\left.\begin{aligned} r_1 &= v_1;\\ r_2 &= v_2;\\ &\ \vdots\\ r_n &= v_n; \end{aligned}\right\} \qquad \text{input data}$$

$$\begin{aligned} r_{n+1} &= r_{j'_1} \,(\mathrm{op})\, r_{j''_1};\\ r_{n+2} &= r_{j'_2} \,(\mathrm{op})\, r_{j''_2}; & \qquad 1 \le j'_i, j''_i < n+i,\\ &\ \vdots & (\mathrm{op}) \in \{+,-,x,:\},\\ f = r_{n+p} &= r_{j'_p} \,(\mathrm{op})\, r_{j''_p}; \end{aligned}$$

Taking into account the errors on the input data and the errors in the arithmetic operations, the actual computation can be described as follows.

$$\begin{aligned} r'_1 &= v_1 (1+e_1),\\ r'_2 &= v_2 (1+e_2),\\ &\ \vdots\\ r'_n &= v_n (1+e_n), \end{aligned}$$

$$\begin{aligned} r'_{n+1} &= (\, r'_{j'_1} \,(\mathrm{op})\, r'_{j''_1} \,)\,(1+e_{n+1}),\\ r'_{n+2} &= (\, r'_{j'_2} \,(\mathrm{op})\, r'_{j''_2} \,)\,(1+e_{n+2}),\\ &\ \vdots\\ f' = r'_{n+p} &= (\, r'_{j'_p} \,(\mathrm{op})\, r'_{j''_p} \,)\,(1+e_{n+p}), \end{aligned}$$

where $e_1, e_2, \ldots, e_n$ are the relative errors on input data, and e_{n+i}, are the local relative errors on r_{n+i}, $p=1,2,\ldots,n$.

Expanding the expression of f' and assuming that nonlinear terms can be neglected we find the following expression for the total linearized error $\Delta f \simeq f-f'$

$$\text{(2.1)} \qquad Df = \sum_{i=1}^{n+p} \frac{\partial f}{\partial r_i}\, r_i\, e_i$$

The first n terms of the sum give the so called inherent error, (i.e. the propagation of the data errors on the result); the remaining terms give the algorithmic error.

If the local errors e_i are assumed to be random variables, it is possible to derive the mean quadratic deviation

$$E(\Delta f^2) = \sum_{i=1}^{n+k} \sum_{j=1}^{n+k} \frac{\partial f}{\partial r_i} \frac{\partial f}{\partial r_j}\, r_i\, r_j\, E(e_i e_j).$$

Moreover, if the random variables e_i are independent then

$$E(\Delta f^2) = \sum_{i=1}^{n+k} \left(\frac{\partial f}{\partial r_i}\right)^2 r_i^2 \; E(e_i^2).$$

Let us consider a perturbation of the system $A\underline{x} = \underline{b}$. Let A' and $\underline{b}'$ be the perturbed values of A and $\underline{b}$, respectively. Matrix A' can be expressed as $A + A*E'$ where $E' = (e'_{ij})$ is the matrix of the relative error terms of the entries of A. This representation is not unique, and we assume that $e'_{ij} = 0$ if $a_{ij} = 0$. Analogously we write $\underline{b}' = \underline{b} + \underline{b}*\underline{e}''$ with $e''_i = 0$ if $b_i = 0$.

The perturbed system will be $A'\underline{y} = \underline{b}'$, or, equivalently,

$$A\,(I + Z\,A*E')\,\underline{y} = \underline{b} + \underline{b}*\underline{e}''.$$

In the following we assume $\|Z\,A*E'\|_2 < 1$. Under this hypothesis matrix A' is nonsingular and the perturbed system has a unique solution x'.

From (2.1) we can derive the linearized propagation error $\Delta\underline{x} \simeq \underline{x}-\underline{x}'$

$$(2.2)\quad \Delta x_r = \sum_i \left(\sum_j \frac{\partial x_r}{\partial a_{ij}} \, a_{ij} \, e'_{ij} + \frac{\partial x_r}{\partial b_i} \, b_i \, e''_i \right).$$

It is easy to see that $\dfrac{\partial x_r}{\partial b_i} = z_{ri}$. Moreover from the relation $\dfrac{\partial Z}{\partial a_{rs}} = - Z \dfrac{\partial A}{\partial a_{rs}} Z$, see [5, p. 68], using $\underline{x} = Z\,\underline{b}$, it follows $\dfrac{\partial x_r}{\partial a_{ij}} = - z_{ri} \, x_j$. Hence equation (2.2) becomes

$$\Delta x_r = - \sum_i z_{ri} \sum_j a_{ij} \, x_j \, e'_{ij} - b_i \, e''_i =$$

$$= - \sum_{i\;j} z_{ri} \, a_{ij} \, x_j \, (e'_{ij} - e''_i).$$

In matrix notation

$$(2.3)\qquad \Delta\underline{x} = - Z\,(A * E)\,\underline{x},$$

where $E = (e_{ij})$, $e_{ij} = e'_{ij} - e''_i$.

From the vectorial expression of the error it is common to derive some scalar quantities which measure the numerical difficulties to solve the problem.

The first approach is to bound the absolute value of the error, using a vectorial norm. Obviously

$$\max_{\|\underline{x}\|_p = 1} \|\Delta\underline{x}\|_p = \|\,Z\,(A*E)\,\|_p$$

This measure can be related to the Skeel condition number. The following proposition is easy to be proved.

PROPOSITION 2.1

Let $|e_{ij}| < \varepsilon$, $1 \le i,j \le n$. Then

$$(2.4) \qquad \max_{\|\underline{x}\|_p = 1} \|\Delta \underline{x}\|_p \le \varepsilon \, C_p(A).$$

Moreover, using the maximum norm, there exists a matrix E for which (2.4) holds with the equality sign. ∎

We assume errors to be arrays of random variables whose instances are independent from the values of A and $\underline{x}$. In this case the error $\Delta\underline{x}$ is a random vector whose distribution depends on A, $\underline{x}$ and the distribution of E. The expected value of $\Delta\underline{x}^*\Delta\underline{x}$ has the following expression

$$E(\Delta x_r^2) = E\left(\left[\sum_{i\,j} z_{ri}\, a_{ij}\, x_j\, e_{ij}\right]^2\right) =$$

$$= \sum_{i\,j\,p\,s} z_{ri}\, a_{ij}\, z_{rp}\, a_{ps}\, x_j\, x_s\, E(e_{ij}\, e_{ps}).$$

Let $e(\underline{x}) = E(\|\Delta\underline{x}\|_2^2 / \|\underline{x}\|_2^2)$, then

$$e(\underline{x}) = \sum_{r\,i\,j\,p\,s} z_{ri}\, a_{ij}\, z_{rp}\, a_{ps}\, E(e_{ij}\, e_{ps})\, x_j\, x_s / \|\underline{x}\|_2^2 .$$

Using the results presented in Appendix A, we can prove the following proposition.

PROPOSITION 2.2

$$\underset{\|\underline{x}\|_2 = 1}{\mathrm{Mean}} \; e(\underline{x}) = \sum_{r\,i\,j\,p} z_{ri}\, a_{ij}\, z_{rp}\, a_{pj}\, E(e_{ij}\, e_{pj})/n,$$

∎

Assuming data errors to be independent random perturbations of A and $\underline{b}$, with the same mean m and variance s², from Lemma B.1 of Appendix B we get

$E(e_{ij}\, e_{ps}) = s^2 \delta_{ip} (1+d_{js})$, where δ_{pq} is the Kronecker symbol.

It is useful to introduce now the three way array

$$\mathbb{A} = (a_{ijk}), \qquad a_{ijk} = z_{ij}\, a_{jk}$$

which is called the tensor associated to the matrix A. With this notation it readily follows that

$$(2.5) \quad \underset{\|\underline{x}\|_2 = 1}{\mathrm{Mean}} \; e(\underline{x}) = (2\, s^2/n) \sum_{r\,i\,j} z_{ri}^2\, a_{ij}^2 = 2\, s^2 \, \|\mathbb{A}\|_F^2 /n.$$

It is also easy to prove the relation

$$\max_{\|\underline{x}\|_2 = 1} e(\underline{x}) \le 2\, s^2 \, \|\mathbb{A}\|_F^2 .$$

The quantity $\|\mathbb{A}\|_F$ is called the tensorial condition of A.

It is worth noting that the quantities $C_\infty(A)$ and $\|\mathbb{A}\|_F$ are the maximum and Frobenius norms of the same rectangular matrix.

Let $B = (b_{rk})$, $b_{rk} = |z_{ri}|\,|a_{ij}|$, $1 \le i,j,r \le n$, $k = i + n\,(j-1) \le n$, we have $C_\infty(A) = \|B\|_\infty$ and $\|\,|A|\,\|_F = \|B\|_F$, thus the tensorial and Skeel condition have a similar behaviour.

For many classes of matrices the classical condition number too is close to the above quantities and can be used to bound the propagation error. On the other hand there exist classes of matrices for which the ratio between classical and Skeel condition number is not bounded (e.g. diagonal matrices).

In the following, in order to evidentiate the dependence of the algorithmic errors on the various condition numbers, we will use test matrices for which classical and Skeel condition numbers have a different behaviour.

The numerical experiments suggested to use

1) Vandermonde column perturbed, i.e.

$$a_{ij} = (a^{i-1}\,c)^{j-1}, \quad \begin{cases} a & \text{parameter,} \\ c & \text{perturbation constant,} \end{cases}$$

2) Vandermonde column scaled, i.e.

$$a_{ij} = (a^{i-1})^{j-1}\,d_j, \quad \text{with } d_j \text{ chosen to minimize } C\,(A).$$

In Fig. 1 and Fig. 2 you can see the classical condition $k_\infty(A)$ and the tensorial condition plotted versus the Skeel condition $C_\infty(A)$ in a logarithmic scale, for these two types of matrices (with n=5).

FIG. 1.

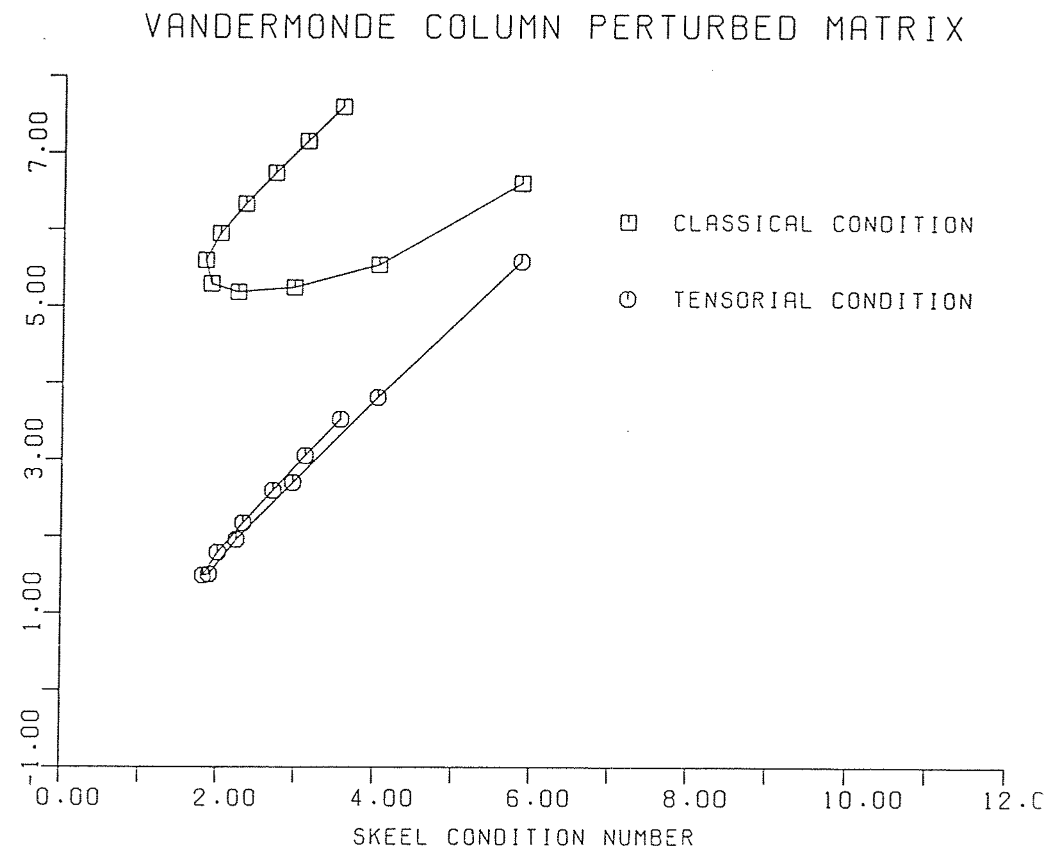

Fig. 1. Classical and tensorial condition numbers plotted versus the Skeel condition number for a 5x5 Vandermonde column perturbed matrix.

FIG.2.

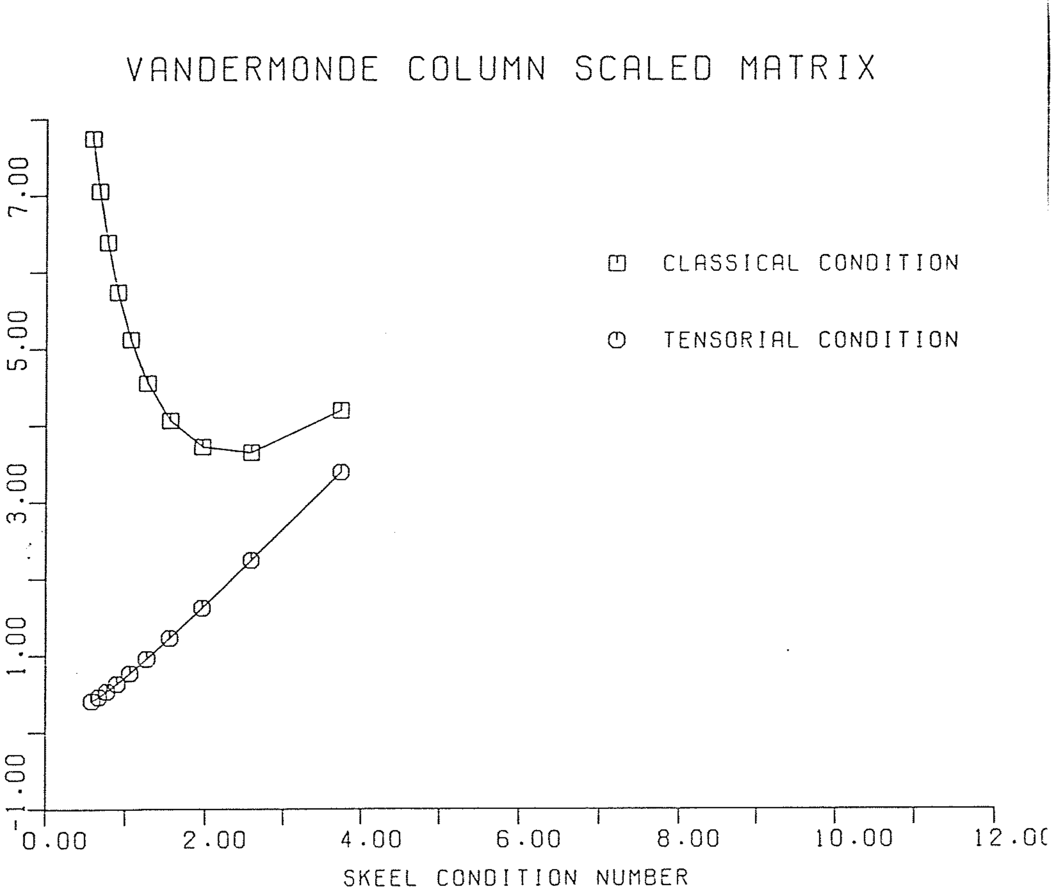

Fig. 2. Classical and tensorial condition numbers plotted versus the Skeel condition number for a 5x5 Vandermonde column scaled matrix.

3. TOTAL ERROR IN SOLVING LINEAR SYSTEMS

A direct method for solving a linear system $A\,\underline{x} = \underline{b}$ can be viewed as a sequence of rational transformations

$$(A^{(0)} \mid \underline{b}^{(0)}) \rightarrow (A^{(1)} \mid \underline{b}^{(1)}) \rightarrow \dots \rightarrow (A^{(t)} \mid \underline{b}^{(t)}),$$

which from the nx(n+1) matrix $(A \mid \underline{b}) = (A^{(0)} \mid \underline{b}^{(0)})$ lead to the matrix $(A^{(t)} \mid \underline{b}^{(t)}) = (I \mid \underline{x})$, with the equivalence conditions

$$A^{(i)}\,\underline{x} = \underline{b}^{(i)}, \quad i=0,1,\dots,t.$$

Applying the considerations of section 2, it is readily seen that each transformation contributes to the total linearized error with a term

$$(3.1)\quad \Delta x_r^{(k)} = \sum_i \left(\sum_j \frac{\partial x_r}{\partial a_{ij}^{(k)}} a_{ij}^{(k)} e'^{(k)}_{ij} + \frac{\partial x_r}{\partial b_i^{(k)}} b_i e''^{(k)}_i \right),$$

where $$e'^{(k)}_{ij} = \begin{cases} 0, \text{ if } a_{ij}^{(k)} \text{ is identically } 0 \text{ or } a_{ij}^{(k)} = a_{ij}^{(k-1)}, \\ \text{the relative error on } a_{ij}^{(k)}, \text{ otherwise.} \end{cases}$$

Analogously $e''^{(k)}_i$ is the relative error of $b_i^{(k)}$. Moreover $e'^{(0)}_{ij}$ and $e''^{(0)}_i$ denote the errors on input data $A^{(0)}$ and $\underline{b}^{(0)}$.

In matrix notation, using (2.3) we get

$$\Delta\underline{x}^{(k)} = - Z^{(k)} (A^{(k)} * E^{(k)}) \underline{x},$$

where $E^{(k)} = (e_{ij}^{(k)})$, $e_{ij}^{(k)} = e'^{(k)}_{ij} - e''^{(k)}_{i}$

Therefore the inherent, algorithmic and total errors become

$$\Delta\underline{x}^{(0)}, \quad \sum_{k=1}^{t} \Delta\underline{x}^{(k)} \quad \text{and} \quad \sum_{k=0}^{t} \Delta\underline{x}^{(k)}, \quad \text{respectively.}$$

Let $|e_{ij}^{(k)}| < \varepsilon_k$, $1 \le i,j \le n$, $0 \le k \le t$, then the total error $\Delta\underline{x}$ can be related to the Skeel condition numbers of the intermediate matrices of the solving process, i.e.

$$\max_{\|\underline{x}\|_p=1} \|\Delta\underline{x}\|_p \le \sum_{k=0}^{t} \varepsilon_k C_p(A^{(k)}).$$

Now let us assume errors to be matrices of random variables whose instances are independent from the values of A and $\underline{x}$. In this case the total error is a random vector whose distribution depends on A, $\underline{x}$ and the distribution of local errors.

Let $e_k(\underline{x}) = E(\|\Delta\underline{x}^{(k)}\|_2^2 / \|\underline{x}\|_2^2)$, then, assuming that the errors of different steps of the solving process are mutually independent, we obtain

$$e(\underline{x}) = \sum_{k=1}^{t} e_k(\underline{x}).$$

and analogously

$$\operatorname*{Mean}_{\|\underline{x}\|_2=1} e(\underline{x}) = \sum_{k=0}^{t} \operatorname*{Mean}_{\|\underline{x}\|_2=1} e_k(\underline{x})$$

4. INFLUENCE OF DIAGONAL SCALING ON THE ERROR

Scaling is one of the most commonly used preconditioning techniques. It consists in multiplying rows and columns of the matrix A by suitable factors before solving the system.

Let U, V be diagonal positive nxn matrices. The system A $\underline{x}$ = $\underline{b}$ can be written U A $\underline{x}$ = U $\underline{b}$ and solved in two steps

Row scaling

(4.a') compute $\underline{z} = U\,\underline{b}$ and $F = U\,A$;

(4.b') solve $F\,\underline{x} = \underline{z}$.

Analgously we can write A V $\underline{y}$ = $\underline{b}$ and solve with the following algorithm.

Column scaling

(4.a") compute $F = A\,V$;

(4.b") solve $F\,\underline{y} = \underline{b}$;

(4.c") compute $\underline{x} = V\,\underline{y}$.

Finally the two forms of scaling can be combined

Complete scaling

(4.a) compute $\underline{z} = U\,\underline{b}$ and $F = U\,A\,V$;

(4.b) solve $F\,\underline{y} = \underline{z}$;

(4.c) compute $\underline{x} = V\,\underline{y}$.

These processes obviously do not affect the inherent error. Some questions naturally arise about the numerical behaviour of the scaling.

i) How condition numbers of F and A differ ?

ii) How the algorithm used to solve the system changes due to the scaling? (e.g. scaling affects the choice of pivots in Gaussian elimination).

iii) When the algorithm does not change, how much is the error on the solution of the problem sensitive to the scaling itself?

For which concerns question (i), it is remarkable that the Skeel condition number is invariant under row scaling. Moreover, when the maximum norm is used, the problem is completely solved, namely

$$\min_{U \in D} k_\infty(UA) = C_\infty(A),$$

where D denotes the class of positive diagonal matrices. If A is irreducible it is easy to prove that

$$\min_{U,V \in D} k_\infty(UAV) = \min_{V \in D} C_\infty(AV) = \rho(|A^{-1}|\,|A|), \ [1,14].$$

Answering to question (ii) need the knowledge of the

properties of the algorithm used to solve (4.b). The discussion will be made in the following, according to the particular situations arising.

We want now to answer question (iii). First we consider the error introduced by the scaling process. When the diagonal entries of U and V are integer powers of b (the base of the arithmetic) no error is introduced by the scaling process. Otherwise, the relative error induced by steps (4.c) and (4,c") is bounded by the machine precision eps; the mean of the error is 0 and the variance is $eps^2/12$, moreover steps (4.a') and (4.a") introduce a perturbation on the matrix which also can be bounded by eps. The propagation of the errors in the scaling process can be estimated with the techniques of Section 2.

Let us consider an algorithmic process which solves linear systems with a sequence of tranformations as in Section 3. When no error is introduced in the scaling process or the error itself is disregarded, the following sufficient conditions for the invariance of the error under diagonal scaling can be stated. (A similar theorem has been proved by Bauer [1] for the Gaussian elimination algorithm).

PROPOSITION 4.1

Given a diagonal scaling A -> U A V, if the following conditions are satisfied

a) the intermediate matrices $(A^{(k)} \mid b^{(k)})$ are transformed into $(U A^{(k)} V \mid U b^{(k)})$;

b) the statistical distribution of $e'^{(k)}_{ij}$, $e''^{(k)}_{i}$ does not change,

then the error $D\underline{x}$ remains unchanged.

Proof

Under the above hypoteses, we can write

$$\Delta y^{(k)}_r = \sum_i -z^{(k)}_{ri} v^{-1}_{rr} u^{-1}_{ii} \sum_j u_{ii} a^{(k)}_{ij} v_{jj} y_j e'^{(k)}_{ij} - u_{ii} b^{(k)}_i e''^{(k)}_i =$$

$$= - v^{-1}_{rr} \sum_i z^{(k)}_{ri} \sum_j a^{(k)}_{ij} v_{jj} y_j e'^{(k)}_{ij} - b^{(k)}_i e''^{(k)}_i .$$

Since $x_i = v_{ii} y_i$ and $\Delta x_r = v_{rr} \Delta y_r$, the thesis follows.

∎

Note that the cases of row and column scaling can be treated by restricting the hypotheses of proposition 4.1 with the insertion of the conditions V=I or U=I, respectively.

From these facts we can draw the following conclusions about the opportunity of diagonal scaling.

a) Scaling is useless when the algorithmic error grows as the Skeel or tensorial condition unless integer powers of β are used. In fact the perturbation of the matrix, produced by the scaling, induces a propagation error of order not lower than the algorithmic error which the scaling would reduce.

b) Complete, row or column scaling are not useful when proposition 4.1 can be applied.

c) When the scaling is used to modify the algorithm (like in Gaussian elimination), the values of the entries of the matrix should be preserved e.g. by using integer powers of β or a weighted pivoting strategy (see section 6).

5. ALGORITHMIC ERROR IN SOLVING TRIANGULAR SYSTEMS

Let A be lower triangular. The classical algorithm to solve the system is structured as follows

$$x_1 := b_1 / a_{11} ;$$

$$x_i := (b_i - \sum_{j=1}^{i-1} l_{ij} x_j) / l_{ii} , \quad i=2,\ldots,n.$$

The computed solution $\underline{x}'$ can be interpreted as the exact solution of perturbed system $(A + A*G)\ \underline{x} = \underline{b}$ [3], with

$$G = (g_{ij}), \quad g_{ij} = \begin{cases} 0, & \text{if } i<j, \\ e_{i1}, & \text{if } j=1, \\ \prod_{k=i}^{2i-1} (1+e_{ik}) - 1, & \text{if } i=j, \\ (1+e_{ij}) \prod_{k=i}^{i+j-2} (1+e_{ik}) - 1, & \text{otherwise.} \end{cases}$$

The e_{ij} are the relative errors of the single arithmetic operations and are assumed to be independent random variables of mean 0 and variance s^2.

To compute the mean error using proposition 3.2 we have to evaluate $E(g_{ij}\, g_{pj})$. We get

$$\begin{cases} E(g_{ij}\, g_{pj})=0, & \text{if } i\neq p,\\ E(g_{ij}\, g_{ij}) = (1+s^2)^j - 1 \simeq j\, s^2 . \end{cases}$$

Finally the mean algorithmic error in solving triangular systems can be expressed as follows

$$\underset{\|\underline{x}\|_2=1}{\text{Mean}}\ e(\underline{x}) = (s^2/n) \sum_i \sum_r z_{ri}^2 \sum_j j\, a_{ij}^2 \leq$$

$$\leq s^2 \sum_i \sum_r z_{ri}^2 \sum_j a_{ij}^2 = s^2 \, \|A\|_F^2 ,$$

It is readily seen that if A is upper triangular, the error analysis leads to the same result.

If a multiple precision arithmetic with machine precision eps' is used, the algorithmic error is bounded by

$$eps'^2 \, \|A\|_F^2 /12.$$

On the other hand, the error due to the representation of A in the machine word is given by

$$eps^2 \, \|A\|_F^2 \; / \; 6\; n.$$

Therefore, if n eps' << 2 eps then the algorithmic error can be ignored.

6. ALGORITHMIC ERROR IN GAUSSIAN ELIMINATION

Consider a linear system $A\,\underline{x} = \underline{b}$ and assume the pivoting strategy to have been already applied in the form of a suitable permutation. The Gaussian elimination algorithm can be interpreted as a sequence of n-1 rational transformations

$$(A^{(0)}\,|\underline{b}^{(0)}) \to (A^{(1)}\,|\underline{b}^{(1)}) \to \dots \to (A^{(n-1)}\,|\underline{b}^{(n-1}),$$

which lead the nx(n+1) matrix $(A \mid \underline{b}) = (A^{(0)}\,|\underline{b}^{(0)})$ to the matrix $(A^{(n-1)}\,|\underline{b}^{(n-1)})$, with $A^{(n-1)}$ upper triangular.

Each transformation has the form

$$a_{ij}^{(k)} = \begin{cases} a_{ij}^{(k-1)}, & \text{if } i=k,\\ a_{ij}^{(k-1)} - a_{ik}^{(k-1)}\, a_{kj}^{(k-1)} / a_{kk}^{(k-1)}, & \text{if } k < i,j < n,\\ 0, & \text{otherwise,} \end{cases}$$

$$b_i^{(k)} = \begin{cases} b_i^{(k-1)}, & \text{if } i=k,\\ b_i^{(k-1)} - a_{ik}^{(k-1)}\, b_k^{(k-1)} / a_{kk}^{(k-1)}, & \text{if } k < i < n, \end{cases}$$

The total error of the algorithmic process becomes

$$(6.1)\quad \Delta x_r = \sum_{k=1}^{n-1} \sum_{i=k+1}^{n} \sum_{j=k+1}^{n} -z_{ri}^{(k)} a_{ij}^{(k)} x_j (e'^{(k)}_{ij} - e''^{(k)}_{i}).$$

Note that $Z^{(k)}$ differs from $Z^{(k-1)}$ in the k-th column only. This implies that in (6.1) the elements of Z can be used, i.e.

$$(6.2)\quad \Delta x_r = \sum_{k=1}^{n-1} \sum_{i=k+1}^{n} \sum_{j=k+1}^{n} -z_{ri} a_{ij}^{(k)} x_j (e'^{(k)}_{ij} - e''^{(k)}_{i}).$$

The error analysis can be considerably simplified by using multiple precision or maximal accuracy techniques for the arithmetic operations.

We assume that the errors due to computations are negligible when compared to the errors due to the representation of the intermediate matrices. Thus we can consider the entries

$$e'^{(k)}_{ij}, \quad e''^{(k)}_{i}, \quad 1 \le k \le t, \quad k+1 \le i,j \le n$$

to be independent random variables with mean 0 and variance s^2. (As noted in the introduction $s^2 = eps^2/12$, where eps is the related to the length of the memory word). With this assumption also the error due to the solution of the triangular system can be ignored. Using lemma B.1 we obtain the following result.

$$(6.3)\quad \underset{\|\underline{x}\|_2=1}{\text{Mean}}\ e(\underline{x}) = (2\,s^2/n) \sum_{k=1}^{n-1} \sum_{r=1}^{n} \sum_{i=k+1}^{n} \sum_{j=k+1}^{n} z_{ri}^2\, a_{ij}^{(k)\,2}.$$

Upper bounds to (6.2) and (6.3) can be derived by using the bound to the grow of $|a_{ij}^{(k)}|$, which in turn derive from the chosen pivoting strategy. Namely, let $\alpha = \underset{i\,j}{\text{Max}} |a_{ij}|$ and let $|a_{ij}^{(k)}| \le \alpha\, g(n)$ [16], and $|e'^{(k)}_{ij}|, |e''^{(k)}_{ij}| \le \varepsilon$. Then

$$\underset{\|\underline{x}\|=1}{\text{Max}} \|D\underline{x}\| \le \varepsilon\, a\, g(n)\, n(n-1)/2\ \|Z\|,$$

and

$$\underset{\|\underline{x}\|_2=1}{\text{Mean}}\ e(\underline{x}) = (2\,s^2 \alpha^2 g(n)^2/n) \sum_{k=1}^{n-1} (n-k) \sum_{i=k+1}^{n} \sum_{r=1}^{n} z_{ri}^2 <$$

$$< s^2\, \alpha^2\, g(n)^2\, (n-1)\ \|Z\|_F.$$

It is readily seen that Proposition 4.1 hold for Gaussian elimination when the scaling does not affect the pivoting strategy, therefore, as proved also in [1,3] for Gaussian elimination, scaling is not useful. The influence of scaling on the pivoting strategy and the overall error has been studied extensively in [12,13].

We have tested three pivoting strategies, namely

Gauss total pivot;

Gauss column pivot;

Gauss column weighted.

The last one is a column pivoting technique where in the k-th step the pivot a_{ik} is chosen according to the relation

$$a_{ik}^{2} / \left(\sum_{p=k}^{n} a_{ip}^{2} \right) = \underset{k\leq j\leq n}{\mathrm{Max}} \; a_{jk}^{2} / \left(\sum_{q=k}^{n} a_{jq}^{2} \right), \quad \text{see [15].}$$

The results of this test are presented in Figures 3 and 4.

7. ALGORITHMIC ERROR IN ORTHOGONALIZATION METHODS

The solution of linear systems using orthogonalization techniques consists in reducing the system to triangular form by multiplying the matrix of coefficients by appropriate orthogonal matrices.

We have

$$A^{(0)} = A, \quad A^{(i)} = P^{(i)} A^{(i-1)} \quad 1\leq i<n, \quad A^{(n-1)} = R \text{ upper triangular,}$$

$$P^{(i)^{T}} P^{(i)} = I \,, \quad Q^{T} = P^{(n-1)} \ldots P^{(2)} P^{(1)} \,, \quad Q^{T} Q = I,$$

$$\underline{b}^{(0)} = \underline{b}, \quad \underline{b}^{(i)} = P^{(i)} \underline{b}^{(i-1)} \quad 1\leq i<n, \quad \underline{b}^{(n-1)} = Q^{T} \underline{b},$$

and finally $Q^{T} A = R$, i.e. $A = Q R$.

The matrices $P^{(i)}$ can be elementary Householder matrices or a product of n+1-i plane rotations in the Givens method [4].

The algorithm, denoted as QR algorithm, is structured as follows

FIG.3.

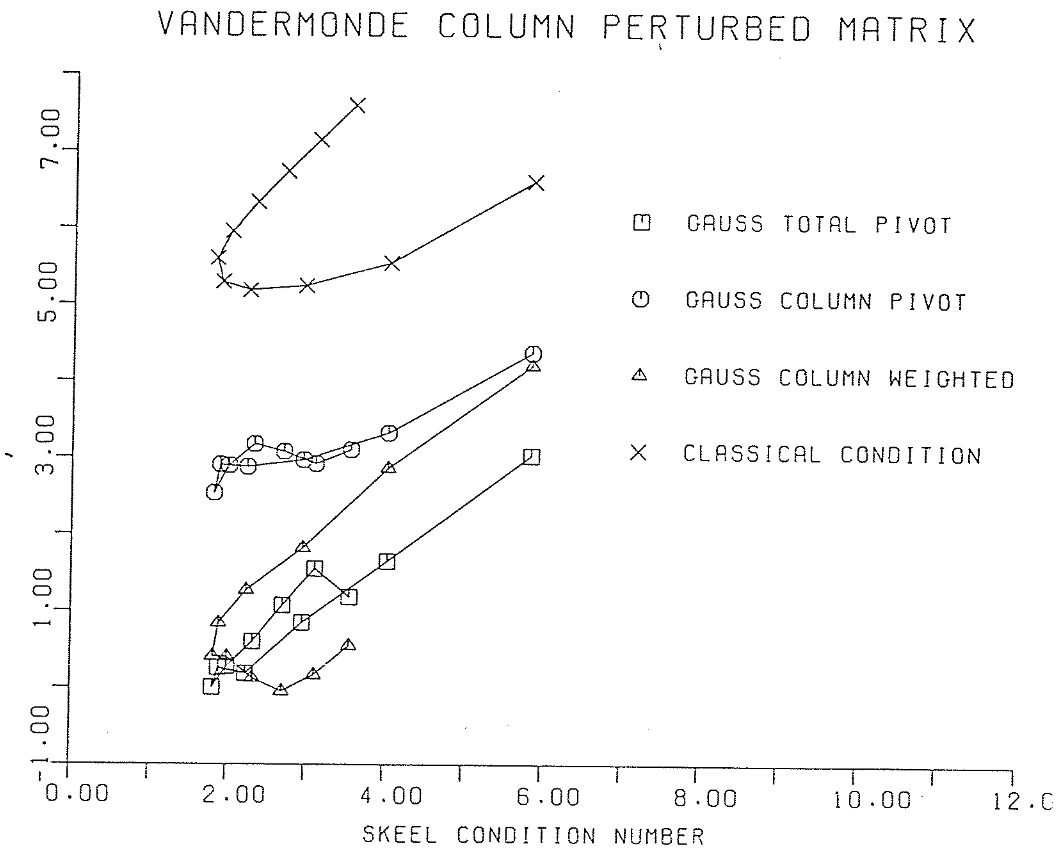

Fig. 3. Classical condition number and mean algorithmic errors of different Gaussian elimination techniques plotted versus the Skeel condition number for a 5x5 Vandermonde column-perturbed matrix.

FIG.4.

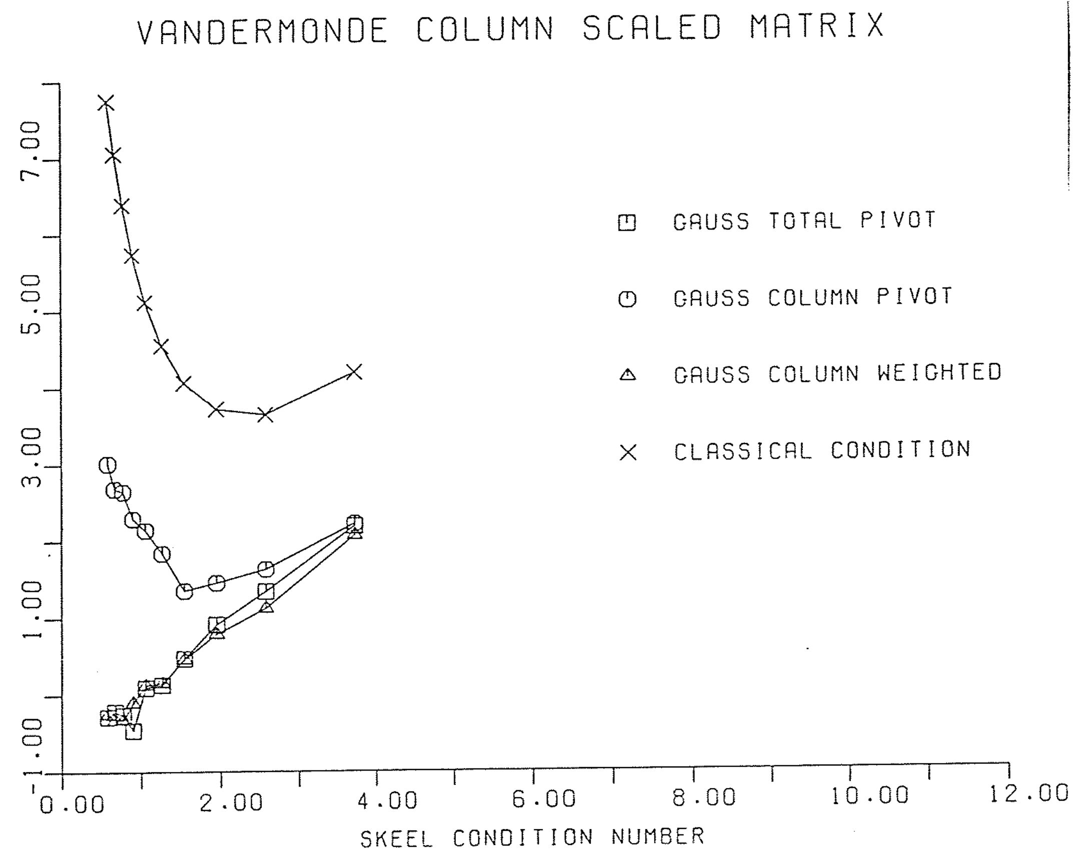

Fig. 4. Classical condition number and mean algorithmic errors of different Gaussian elimination techniques plotted versus the Skeel condition number for a 5x5 Vandermonde column scaled matrix.

Steps (1...n-1): Reduce A to upper triangular form;

Step (n): Solve $R\,\underline{x} = Q^T\,\underline{b}$.

Both in Householder and Givens method the matrix $P^{(i)}$ will modify only the last n+1-i rows and columns of $A^{(i)}$. Thus the total error has the expression

$$(7.1)\quad \Delta x_r = \sum_{k=1}^{n-1}\sum_{i=k}^{n}\sum_{j=k}^{n} -z_{ri}^{(k)}\, a_{ij}^{(k)}\, x_j\, e_{ij}^{(k)}.$$

With the same assumptions made in the previous section the value of the mean error is

$$\underset{\|\underline{x}\|_2=1}{\text{Mean}}\; e(\underline{x}) = (2\,s^2/n)\sum_{k=1}^{n-1}\sum_{r=1}^{n}\sum_{i=k}^{n}\sum_{j=k}^{n} z_{ri}^{(k)\,2}\; a_{ij}^{(k)\,2} \le$$

$$\le (2\,s^2/n)\sum_{k=1}^{n} \|Z^{(k)}\|_F^2\; \|A^{(k)}\|_F^2 .$$

Whence

$$(7.2)\quad \underset{\|\underline{x}\|_2=1}{\text{Mean}}\; e(\underline{x}) \le 2\,s^2\,\|Z\|_F^2\,\|A\|_F^2$$

Proposition 4.1 can be applied to QR algorithm if V=I. This means that only row scaling can be useful for QR algorithm.

It is worth noting that applying the QR algorithm to the transpose of A we can derive a similar decomposition which will reduce A to lower triangular form. Namely we have

$A^{(0)} = A, \quad A^{(i)} = A^{(i-1)} P^{(i)} \quad 1 \le i < n, \quad A^{(n-1)} = L$ lower triangular;

$P^{(i)^T} P^{(i)} = I \;, \quad Q^T = P^{(1)} P^{(2)} \ldots P^{(n-1)} \;, \quad Q^T Q = I;$

and $A Q^T = L \;, \quad A = L Q$ and $A Q^T Q x = \underline{b}.$

Thus, this algorithm, denoted as LQ algorithm, is structured as follows

Steps (1...n-1): Reduce A to lower triangular form;

Step (n): Solve $L \underline{y} = \underline{b}$;

Step (n+1): Solve $Q \underline{x} = \underline{y}$, i.e. compute $\underline{x} = Q^T \underline{y}$.

The error analysis of the LQ algorithm is simple and leads to better results. Since the transformation matrices do not change the length of both the rows of A and the columns of Z, we can write

$$(7.3) \quad \underset{||\underline{x}||_2 = 1}{\text{Mean}} \; e(\underline{x}) = (2 s^2/n) \sum_{k=1}^{n-1} || /A^{(k)} ||_F^2 \le 2 s^2 || /A ||_F^2 .$$

The error due to step (n+1) has to be considered. It should be noted that Q is orthogonal and its tensorial condition is not greater than n. Thus the square of the mean error due to the representation of Q in the machine word is not greater than eps²/6 and, by using multiple precision arithmetic the algorithmic error too can be arbitrarily small. Therefore the error of step (n+1) does not depend on the condition of A and is of the order of the machine precision eps.

Proposition 4.1 can be applied to LQ algorithm if U=I. Then only column scaling would be useful for LQ algorithm, but, because the error has the same behaviour of the tensorial condition, LQ is not improved by any scaling.

Comparing (7.2) with (7.3) we note that the bound for LQ algorithm is better because the classical condition can be arbitrarily larger than the tensorial one. The question arises whether the bound for the QR algorithm is tigth. Figures 5,6 and 7 show that, for some Rice (n=10) and Vandermonde (n=5) matrices, the QR algorithm presents larger errors than LQ and the behavior of QR errors is similar to the classical condition as suggested by (7.2) (both algorithms were implemented using Householder transformations).

It is worth noting that, if the row scaling which is optimal with respect to maximum norm, is performed for QR algorithm, the classical condition of the scaled matrix becomes equal to the Skeel condition and the resulting algorithm has the same error behaviour than LQ.

8. CONCLUSION

The statistical error analysis of the solution of a linear system has led to introduce the three way array /A. The tensorial and Skeel condition numbers are norms of a matrix with the same entries of A and measure the condition of the problem.

The stability of the problem under diagonal scaling has been

FIG.6.

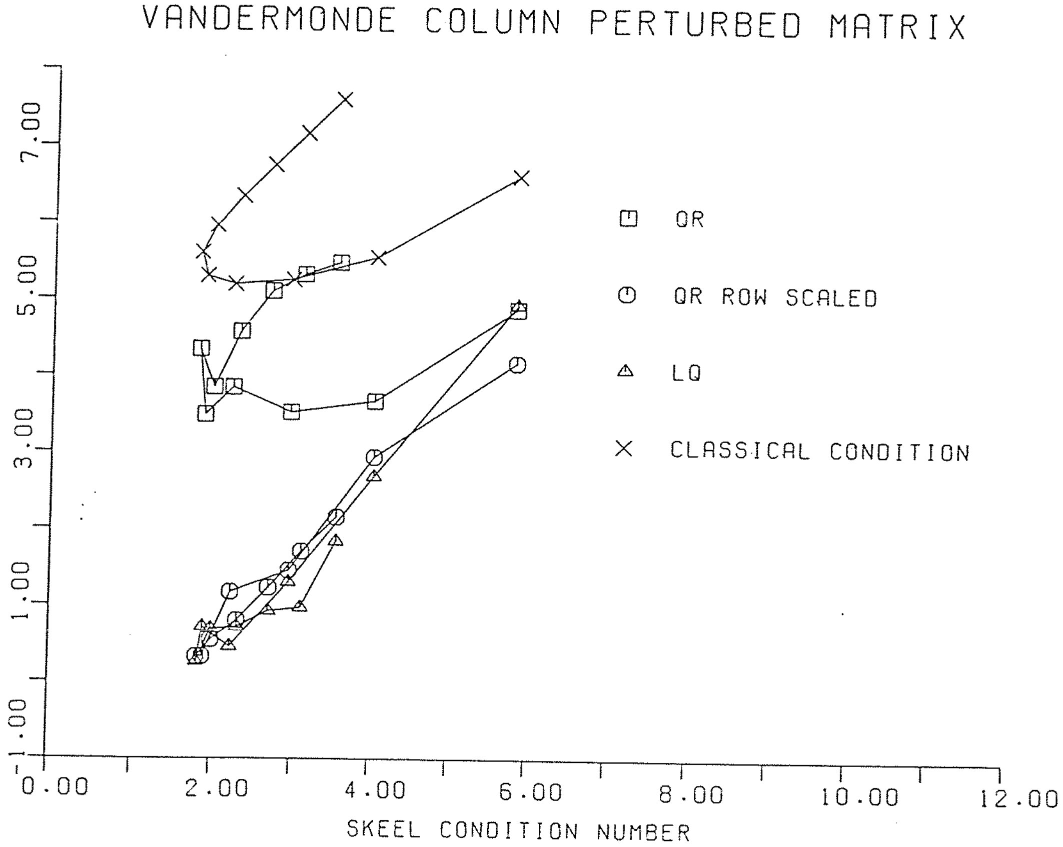

Fig. 6. Classical condition number and mean algorithmic errors of different orthogonalization techniques plotted versus the Skeel condition nnmber for a 5x5 Vandermonde column perturbed matrix.

FIG.5.

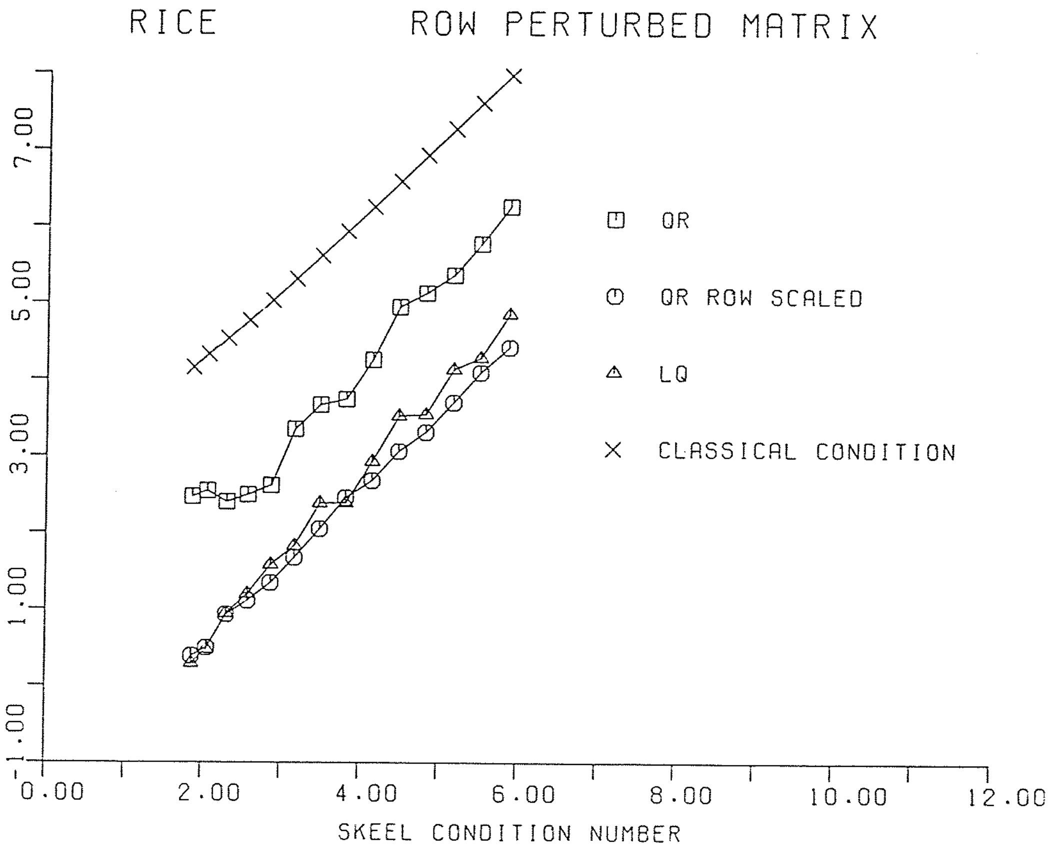

Fig. 5. Classical condition number and mean algorithmic errors of different orthogonalization techniques plotted versus the Skeel condition number for a 10x10 Rice row perturbed matrix.

FIG.7.

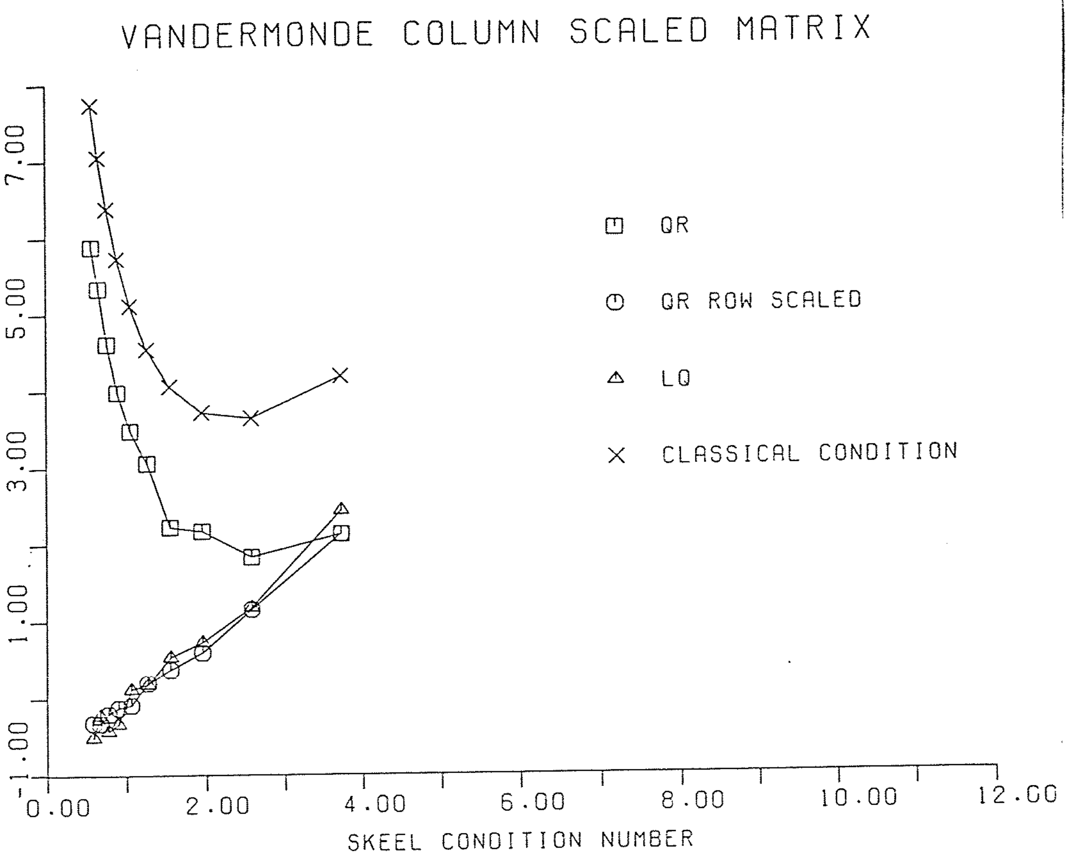

Fig. 7. Classical condition number and mean algorithmic errors of different orthogonalization techniques plotted versus the Skeel condition number for a 5x5 Vandermonde column scaled matrix.

studied and some criteria have been found to determine whether the scaling is useful.

For Gaussian elimination bounds similar to the classical ones have been found. Moreover the scaling resulted has no influence on the final error (as pointed out also by Bauer) nevertheless weigthed pivoting tecniques are useful to improve the stability of algorithms.

For orthogonalization algorithms an essential asymmetry between QR and LQ error behaviour has been found. Errors of QR seem to be related to the classical condition whereas errors of LQ can be bounded using the tensorial condition. Finally row scaling is useful when applied to QR algorithm.

APPENDIX A

Let we prove some elementary lemmas.

LEMMA A.1

$$\Gamma^{-1}\int_{B_n} x_h \, d\underline{x} = 0, \quad h=1,\dots,n.$$

The proof follows from the symmetry of the integration field. ■

LEMMA A.2

$$\Gamma^{-1}\int_{B_n} x_h^2 \, d\underline{x} = 1/n, \quad h=1,\dots,n.$$

The proof follows from the relation

$$\Gamma = {}_{B_n} ||\underline{x}||_2^2 \, d\underline{x} = \sum_i \int_{B_n} x_i^2 \, d\underline{x} = n \int_{B_n} x_h^2 \, dx.$$

■

LEMMA A.3

$$\Gamma^{-1}\int_{B_n} x_h x_k \, d\underline{x} = \delta_{hk}/n.$$

The proof follows from lemma A.1 and A.2. ■

APPENDIX B

Let $E = (e_{ij})$, $e_{ij} = e'_{ij} - e''_i$, $i=1,\dots,n$, $j=1,\dots,n$ be a matrix of random variables, where e'_{ij} and e''_i are independent random variables with mean m and variance s². The following lemma holds.

LEMMA B.1

$$E(e_{hk}\, e_{pq}) = s^2 (\delta_{hp}\delta_{kq} + \delta_{hp}).$$

Proof

We have

$$E(e_{hk}\, e_{pq}) = E(e'_{hk}\, e'_{pq}) + E(e''_h\, e''_p) - E(e'_{hk}\, e''_p) - E(e'_{pq}\, e''_h).$$

Hence

$$E(e'_{hk}\, e'_{pq}) = \begin{cases} E(e'^{2}_{hk}) = m^2 + s^2 & \text{if } h=p \text{ and } k=q; \\ E(e'_{hk})E(e'_{pq}) = m^2 & \text{otherwise.} \end{cases}$$

$$E(e''_h\, e''_p) = \begin{cases} E(e''^{2}_{h}) = m^2 + s^2 & \text{if } h=p; \\ E(e''_h)\ E(e''_p) = m^2 & \text{otherwise.} \end{cases}$$

And the thesis easily follows. ■

REFERENCES

1. Bauer, F.L.: Optimally Scaled Matrices. Numer. Math. 5 73-87 (1963).

2. Brent R.P.: Fast Multiple-Precision Evaluation of Elementary Functions. J. Assoc. Comput. Mach. 23, 242-251 (1976).

3. Forsythe, G., Moler, C.B.: Computer Solution of Large Linear Algebraic Systems. Englewood Cliffs: Prentice Hall 1967.

4. Golub, G.H., Van Loan, C.F.: Matrix Computations. North Oxford Academic: Oxford 1983.

5. Graham, A.: Kronecker Product and Matrix Calculus with Applications. Chichester: Ellis Horwood 1981.

6. Kulisch, U.W., Miranker W.L.: Computer Arithmetic in Theory and Practice. New york: Academic Press 1981.

7. Liu, B., Kaneko, T.: Error Analysis of Digital Filters Realized with Floating Point Arithmetic. Proc. IEEE 57, 1735-1747 (1969).

8. Liu, B., Kaneko, T.: Accumulation of Roundoff Errors in Fast Fourier Transforms. J. Assoc. Comput. Mach. 17, 637-654 (1970).

9. Oppenheim, A.V., Weinstein, C.J.: Effects of Finite Register Length in Digital Filtering and the Fast Fourier Transform. Proc. IEEE 60, 957-976 (1972).

10. Rice, J.R.: Matrix Computation and Mathematical Software. New York: McGraw Hill 1981.

11. Rump, S.M., Böhm, H.: Least Significant Bit Evaluation of Arithmetic Expressions in Single Precision. Computing 30, 189-199 (1983).

12. Skeel, R.D.: Scaling for Numerical Stability in Gaussian Elimination. J. Assoc. Comput. Mach. 26 494-526 (1979).

13. Skeel, R.D.: Effects of Equilibration on Residual Size for Partial Pivoting. SIAM J. Numer. Anal. 18, 449-454 (1981).

14. Sluis, van der, A.: Condition numbers and equilibration of matrices. Numer. Math. 14, 14-23 (1969).

15. Stoer, J., Burlisch, R.: Introduction to Numerical Analysis. New York: Springer 1980.

16. Wilkinson, J.H.: Error analysis of direct methods of matrix inversion. J. Assoc. Comput. Mach. 8, 281-330 (1961).

17. Wilkinson, J.H.: Rounding Errors in Algebraic Processes. Englewood Cliffs: Prentice Hall 1963.