Serdica
Bulgariacae mathematicae publicationes

Сердика
Българско математическо списание

AN EASILY COMPUTED UPPER BOUND FOR THE CONDITION NUMBER OF A MATRIX

WLADIMIR POPOV

We introduce an easily computed characteristic of non-singular matrices: $w(A)=(n^{-1}\Sigma a_{ij}^2)^{1/2}/|\det(A)|^{1/n}$ and show that it can be used to obtain: 1) an upper bound for $k_2(A)=\|A\|_2\|A^{-1}\|_2$; 2) additional information concerning the singular spectrum of A (when some other estimate of $k_2(A)$ is known).

The condition number of a non-singular n by n matrix is defined as $k(A)=\|A\|\cdot\|A^{-1}\|$ where $\|\cdot\|$ is some matrix norm. (We shall denote by $k_1(A)$, $k_2(A)$, and $k_\infty(A)$ the condition numbers corresponding to $\|\cdot\|_1$, $\|\cdot\|_2$, and $\|\cdot\|_\infty$ respectively.)

Following inequalities are well known (cf. [3]):

$$\frac{\|x^*-x\|}{\|x\|}\leq\frac{\|b^*-b\|}{\|b\|}\cdot k(A) \quad \text{for } Ax=b,\ Ax^*=b^*,\ x\neq 0, \tag{1}$$

$$\frac{\|x^*-x\|}{\|x\|}\leq\frac{\|A^*-A\|}{\|A\|}\cdot k(A) \quad \text{for } Ax=b,\ A^*x^*=b,\ x\neq 0. \tag{2}$$

Therefore the knowledge of $k(A)$ provides valuable information about the reliability of the solution of the linear system $Ax=b$ when A or b are perturbed. However, the direct computation of $k(A)$ is time consuming ($O(n^3)$ for k_1, k_2 and k_∞) and is usually not adjustable in computational practice.

In some cases estimates of $k(A)$ can be obtained (at the rate of $O(n^2)$ extra operations) as a by-product while solving $Ax=b$. A good algorithm for estimating $k_1(A)$ is discussed in [1] and [2], and implemented in the LINPACK package. The estimate obtained by this algorithm is a lower bound of $k_1(A)$ and gives a reliable indication of ill-condition. It must be pointed out, however, that the assertion of its being close to $k_1(A)$ is probabilistic and a matrix can not be proved to be well conditioned using this estimate.

In this paper we consider the characteristic

$$w(A)=(n^{-1}\sum_1^n\sigma_i^2)^{1/2}/(\prod_1^n\sigma_i)^{1/n}, \tag{3}$$

where $\sigma_1,\ldots,\sigma_n$ are the singular values of A.

When $Ax=b$ is solved by direct methods using triangular factorization (Gauss, Cholesky, QR etc.) $w(A)$ is computed at the rate of $O(n^2)$ extra operations and can be used to obtain:

1) an upper bound for $k_2(A)$;

2) additional information concerning the distribution of the singular spectrum of the matrix (when another estimate of $k(A)$ is known).

Let us recollect some classical results which will be used later:

SERDICA, Bulgaricae mathematicae publicationes, Vol. 15, 1989, p. 192—196

For any real n by n matrix A there exist two orthogonal matrices U and V and a diagonal matrix $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_n)$ such that

$$A = U\Sigma V^T, \quad \sigma_1 \geqq \sigma_2 \geqq \ldots \geqq \sigma_n \geqq 0. \tag{4}$$

This singular value decomposition is unique and $\sigma_1^2, \ldots, \sigma_n^2$, are the eigenvalues of AA^T,

If A is non-singular, then

$$\sigma_1 = \| A \|_2, \quad \sigma_n = \| A^{-1} \|_2^{-1}, \quad k_2(A) = \sigma_1/\sigma_n. \tag{5}$$

Now let A be a non-singular matrix and let $w(A)$ be defined by (3).

Theorem 1.

(i) $w(A) \geqq 1$,

(ii) $w(A) = (n^{-1} \Sigma a_{ij}^2)^{1/2} / |\det(A)|^{1/n}$,

Proof: (i) follows from the quadratic-geometric mean inequality.

(ii) Since $\lambda_1 = \sigma_1^2, \ldots, \lambda_n = \sigma_n^2$ are the eigenvalues of AA^T:

$$\Sigma\sigma_i^2 = \Sigma\lambda_i = \mathrm{tr}(AA^T) = \Sigma a_{ij}^2, \quad (\Pi\sigma_i)^2 = \Pi\lambda_i = \det(AA^T) = \det^2(A).$$

The computation of Σa_{ij}^2 takes n^2 additions and n^2 multiplications and $\det(A)$ is obtained by $O(n)$ multiplications from the triangular factors in which A is, may be implicitly, decomposed when $Ax = b$ is solved by direct methods. Thus the calculation of $w(A)$ instead of executing the algorithm of [1] (requiring about $5n^2$ operations) can provide considerable saving of time, especially for small values of n. It is particularly advantageous on specialized computers (array or matrix processors, multiprocessor systems); in this case the computation of $w(A)$ amounts to several vector operations and can be performed simultaneously with the factorization of the matrix. Now we are going to give an upper bound for $k_2(A)$ in terms of $w(A)$.

For $x = (x_1, \ldots, x_n)$, $x_i > 0$ let us denote: $m_2(x) = (n^{-1}\Sigma x_i^2)^{1/2}$, $m_0(x) = (\Pi x_i)^{1/n}$, $w(x) = m_2(x)/m_0(x)$.

Lemma 1. *Let* $x = (x_1, \ldots, x_{n+m})$, $x_i > 0$, n, $m \geqq 1$

$x' = (x_{i_1}, \ldots, x_{i_n})$, $x'' = (x_{j_1}, \ldots, x_{j_m})$, $\{i_1, \ldots, i_n\} \cup \{j_1, \ldots, j_m\} = \{1, 2, \ldots, n+m\}$

Then $w(x) \geqq w(x')^{n/(n+m)} w(x'')^{m/(m+m)}$ *the equality being reached iff* $m_2(x') = m_2(x'')$

Proof: (i) $m_0(x) = (\prod_{k=1}^{n+m} x_k)^{1/(n+m)} = m_0(x')^{n/(n+m)} m_0(x'')^{m/(n+m)}$

(ii) $m_2(x) = (\frac{n}{n+m}(n^{-1}\sum_1^n x_{i_k}^2) + \frac{m}{n+m}(m^{-1}\sum_1^m x_{j_k}^2))^{1/2}$

$\geqq ((n^{-1}\sum_1^n x_{i_k}^2)^{n/(n+m)} (m^{-1}\sum_1^m x_{j_k}^2)^{m/(n+m)})^{1/2} = m_0(x')^{n/(n+m)} m_0(x'')^{m/(n+m)}$.

The inequality follows from the Bernoulli inequality: $\mu a + (1-\mu)b \geqq a^\mu b^{(1-\mu)}$ for a, $b > 0$, $0 \leqq \mu \leqq 1$ where the equality is reached iff $a = b$.

Theorem 2. *Let* A *be a non-singular* n *by* n *matrix,* $w = w(A)$, $k = k_2(A)$. *Then* $k \leqq w^n + (w^{2n} - 1)^{1/2}$.

Proof: (i) $n = 2$; in this case k is uniquely determined by w:

$$w^2 = \frac{1}{2}(\sigma_1^2 + \sigma_2^2)/\sigma_1\sigma_2 = \frac{1}{2}(\sigma_1/\sigma_2 + \sigma_2/\sigma_1) = \frac{1}{2}(k + 1/k)$$

and k, $1/k$ are the roots of the quadratic equation $x^2-2w^2x+1=0$, i, e. $k=w^2+(w^4-1)^{1/2}$ (since $k\geq 1$).

(ii) $n>2$; as in the case $n=2$ it can be seen that

$$k=w(\sigma_1,\ \sigma_n)^2+(w(\sigma_1,\ \sigma_n)^4-1)^{1/2} \tag{6}$$

and has its maximal value when $w(\sigma_1,\ \sigma_n)$ is maximal. By Lemma 1 and Theorem 1 (i) $w(\sigma_1,\ \sigma_n)^{2/n}\cdot w(\sigma_2,\ \ldots,\ \sigma_{n-1})^{((n-2)/n}\leq w,\ w(\sigma_2,\ \ldots,\ \sigma_{n-1})\geq 1$ consequently

$$w(\sigma_1,\ \sigma_n)\leq w^{n/2}. \tag{7}$$

Now the assertion of the theorem follows from (6) and (7).

Note that, although it is reached for singular spectra of the form $\sigma_2^2=\ldots\sigma_{n-1}^2=c$, $\sigma_1^2=c+c(1-1/w^{2n})^{1/2}$, $\sigma_n^2=c-c(1-1/w^{2n})^{1/2}$, the upper bound given by Theorem 2 will be a severe overestimate in most cases when n is large. Therefore the calculation of a large $w(A)$ does not mean that A is ill-conditioned. A small $w(A)$, however, provides the full guaranty that A is "good". In a series of numerical experiments $w(A)$ was calculated for several hundreds of matrices, generated at random from different distributions. The most probable values of $w(A)$ were in the range 1.6 to 1.8. This means that, for most matrices A, the estimate $k_2(A)\leq 2w^n(A)$ proves A to be:
for $n\leq 50$ — non-singular with respect to the standard double precision (56 bits);
for $n\leq 30$ — non-singular in single precision (24 bits) and well conditioned in double precision;
for $n\leq 20$ — well conditioned in single precision.
(Remember that the time-saving from calculating $w(A)$ instead of using the LINPACK algorithm for estimating $k(A)$ is significant just for small values of n).

For $n>60$ the estimate of Theorem 2 is usually impractical and some other estimate for $k(A)$ must be used. Nevertheless, the computation of $w(A)$ in this case may also be helpful, providing additional information about the structure of the singular spectrum of A. Suppose we have two 10×10 matrices A_1 and A_2 with the singular spectra $\sigma(A_1)=(1,\ldots,\ 1,\ 10^{-6})$ and $\sigma(A_2)=(1,\ 10^{-6},\ldots,\ 10^{-6})$. Since $k_2(A_1)=k_2(A_2)=10^6$ they are both singular in single precision. On the other hand, they are quite different: A_1 is approximately equal to a singular matrix of rank n-1, while A_2 is "nearly singular" of rank 1; i. e. the linear system $A_2x=b$ is much stronger overdetermined than $A_1x=b$. Some information about the "rank" of a nearly singular matrix can be used to choose a specialized algorithm for solving ill-conditioned systems, or to provide some probabilistic estimates for the propagation of errors in A and b. Such information can be obtained (when $k_2(A)=\sigma_1/\sigma_n$ or its estimate is known) from the characteristic $w(A)$ in which all singular values of A take part. Thus, in our example, $w(A_1)=3.78$, $w(A_2)=79432.87$. Next theorem gives upper and lower bounds of $w(A)$ for fixed $k_2(A)$ and an estimate of the number of "small" singular values in $\sigma(A)$.

Lemma 2. *Let* $a_1,\ a_2,\ \ldots,\ a_{n-1}$ *be positive and let* $f(x)=w(a_1,\ \ldots,\ a_{n-1},\ x)=(n^{-1}(x^2+\Sigma a_i^2))^{1/2}(x\Pi a_i)^{-1/n}$ *for* $x>0$ *then*

i) *f has its minimum in* $x_0=m_2(a_1,\ \ldots,\ a_{n-1})=((n-1)^{-1}\Sigma a_i^2)^{1/2}$,

ii) *f is decreasing in* $(0,\ x_0)$ *and increasing in* $(x_0,\ +\infty)$.

Proof: Let us denote $a=\Sigma_{i=1}^{n-1}a_i^2$, $b=\Pi_{i=1}^{n-1}a_i^2$, $\xi=x^2$ and consider the function

$$g(\xi)=n^nf^{2n}(x)=(a+\xi)^n/b\xi.$$

Obviously f is increasing (decreasing, has a local extremum) in x if and only if g is thus in $\xi=x^2$.

$$g'(\xi)=\frac{(a+\xi)^{n-1}}{b\xi^2}((n-1)\xi-a)$$

has (for $\xi>0$) a unique zero in $\xi_0=(n-1)^{-1}a=m_2^2(a_1, \ldots, a_{n-1})$.

For $\xi<\xi_0$ $g'(\xi)$ is negative, and for $\xi>\xi_0$ — positive, q. e. d.

Theorem 3. *If A is a non-singular $n\times n$ matrix with $k_2(A)=c>1$,* $((1+c^2)/2c)^{1/n} \leqq w(A)\leqq(1+(c^2-1)\varkappa)^{1/2}c^{\varkappa}$, *where* $\varkappa=\max(n^{-1}, (21n(c)^{-1}-(c^2-1)^{-1})$.

Proof: We can suppose that the singular spectrum of A is of the form $c=\sigma_1\geqq\sigma_2 \geqq \ldots \geqq\sigma_n=1$. (If not so, we can multiply $A=U^T \operatorname{diag}(\sigma_1, \ldots, \sigma_n)V$ by $D=\operatorname{diag}(\sigma_n^{-1}, \ldots, \sigma_n^{-1})$, obtaining $B=DA$ with $k_2(A)=k_2(B)$, $w(A)=w(B)$, $\sigma(B)=(c, \sigma_2/\sigma_n, \ldots, \sigma_{n-1}/\sigma_n, 1)$).

By Lemma 1

$$w(A)\geqq w(\sigma_2, \ldots, \sigma_{n-1})^{(n-2)/n} w(\sigma_1, \sigma_n)^{2/n}\geqq w_{2,0}(c, 1)^{2/n}=((1+c^2)/2c)^{1/n},$$

which proves the first inequality. (Lemma 1 also implies that this bound is reached for $\sigma_2=\ldots=\sigma_n=(2^{-1}(1+c^2))^{1/2}$).

To prove the upper bound we first observe that, as a trivial consequence of Lemma 2, the maximum of $w(A)$ (for fixed c) is reached for some singular spectrum of the form

$$\sigma^{(k)}=(\underbrace{c, \ldots, c}_{k \text{ times}}, \underbrace{1, \ldots, 1}_{(n-k) \text{ times}}). \tag{8}$$

In this case $w(A)=w(\sigma^{(k)})=(n^{-1}(kc^2+n-k))^{1/2}c^{-k/n}$.

Now consider the function

$$f(\varkappa)=(1+(c^2-1)\varkappa)\,.\,c^{-2\varkappa}.$$

Obviously $f(k/n)=w^2(\sigma^{(k)})$ and $w(\sigma^{(k_1)})\geqq w(\sigma^{(k_2)})$ if and only if $f(k_1/n)\geqq f(k_2/n)$.

$$f'(\varkappa)=[(c^2-1)-(1+(c^2-1)\varkappa)\,.\,21n(c)]\,.\,c^{-2\varkappa}$$

has a unique zero $\varkappa_0=1/21n(c)-1/(c^2-1)$ in the interval $(0, 1)$. f' is positive in $(0, \varkappa^0)$ and negative for $\varkappa>\varkappa_0$, thus $f(\varkappa_0)$ is a local maximum. Consequently $w(\sigma^{(k)})$ has its maximum in $[n\varkappa_0]$ or $[n\varkappa_0]+1$. For $\varkappa_0\leqq 1/n$ this can be only $[n\varkappa_0]+1=1$, q. e. d.

For $n>1$, $c<1$ let us consider the function

$$w_{n,c}(x)=[n^{-1}x(c^2-1)+1]^{1/2}\,.\,c^{-x/n}. \tag{9}$$

The proof of Theorem 3 implies that $W_{n,c}(x)$ is a convex function in $[1, n]$ which has its maximum in $x_0=n/21\,n(c)-n/(c^2-1)$.

If A is a $n\times n$ matrix with $k_2(A)=c$,

$$W_{n,c}(x_0)\geqq w(A)\geqq 1=W_{n,c}(n)$$

and

$$w(A)=W_{n,c}(\xi) \text{ for some } \xi\in[x_0, n].$$

Let us call this ξ the pseudorank of A and denote it by $\operatorname{pr}(A)$. The calculation of $\operatorname{pr}(A)$ for a "nearly singular" matrix A gives some information about the dimension of the subspace of R^n over which A is "well conditioned", i. e. about the number of those singular values of A which are "large". In the particular case when $\sigma(A)$ is of the form (8), where c is large enough to provide $n/21n(c)-n/(c^2-1)\leqq k$, $\operatorname{pr}(A)=k$.

On the other hand, if $w=w(A)$ and some estimate p for $\operatorname{pr}(A)$ are known, the equation

$$w=W_{n,c}(p) \tag{10}$$

can be used to obtain a more realistic estimate of $c=k_2(A)$.

Suppose, for example, that $w(A)=2.0$ has been computed for a 20×20 matrix A. Theorem 2 gives the upper bound $2.2^{20}\cong 2.10^6$ for $k_2(A)$. If, however, it is known that $\mathrm{pr}(A)\leqq 18$ (if, e. g., three rows of A are nearly parallel), (10) can be solved for c with $w=2$, $p=18$ to obtain the much stricter bound $k_2(A)\leqq 1735$. Next lemma provides the basis for such estimates:

Lemma 3. Let A be a $n\times n$ matrix, $w=w(A)$, $p\geqq \mathrm{pr}(A)$, and let c be a solution of (10). Then $k_2(A)\leqq c$.

Proof. Let $c_1=k_2(A)$ and $p_1=\mathrm{pr}(A)$. Then $w=W_{n,c_1}(p_1)$ and, since $p\geqq p_1\geqq(n/21n(c_1)-n/(c_1^2-1)$, we have $W_{n,c_1}(p_1)\geqq W_{n,c_1}(p)$.

Now let us suppose $c_1>c$. The definition (9) of $W_{n,c}$ implies that, for fixed n and $x>0$, $W_{n,c}(x)$ is increasing by c in the interval $(1, +\infty)$ and thus

$$w=W_{n,c_1}(p_1)\geqq W_{n,c_1}(p)>W_{n,c}(p)=w$$

what is absurd. Consequently $c_1\leqq c$.

Lemma 3 can also be used to obtain "conditional" esimates of $k(A)$ of the form

$$\text{"}\mathrm{pr}(A)\geqq n-1 \text{ and } k(A)<c_1\text{"} \text{ or } \text{"}\mathrm{pr}(A)\geqq n-2 \text{ and } k(A)<c_2\text{"}\ldots,$$

where $c_1, c_2, \ldots$ is a (rapidly) decreasing sequence of numbers.

REFERENCES

1. A. C. Cline et al. An Estimate for the Condition Number of a Matrix. *SIAM J. Numer. Anal.*, **16**, 1979, 368-375.
2. J. J. Dongarra et al. LINPACK Users' Guide. Philadelphia, 1979.
3. G. Forsythe, C. Moler. Computer Solution of Linear Algebraic Systems. Englewood Cliffs. N. J., 1967.

Received 25. 05. 1984
Revised 20. 10. 1988