Numer. Math. 14, 14–23 (1969)

Condition Numbers and Equilibration of Matrices

A. VAN DER SLUIS

Received March 6, 1969

Introduction

In numerical linear algebra one meets condition numbers $\|A\|\,\|A^{-1}\|$ and similar quantities such as $(\max|a_{ij}|)\,\|A^{-1}\|$ and $\|A_j\|\,\|A^{-1}\|$, where $A=(a_{ij})$ and A_j is the j-th column of A. The norms are very diverse.

The problem then is to determine a row- and/or column-scaling of A which minimizes the quantity under consideration.

It is the purpose of this paper to specify a class of such quantities for which those scalings can be given explicitly. The results will be extensions of some results in [2]. They will also hold for non-square matrices. All proofs will be completely elementary.

Also, in some cases where the minimizing scaling cannot be given explicitly, it can be said how far at most for a certain scaling the quantity under consideration may be away from its minimum.

Conventions

$\mathfrak{M}_{mn}$ will denote the set of real or complex $m\times n$ matrices, $m\geqq n$, and A will always be an element of $\mathfrak{M}_{mn}$. A^H will denote the transposed of $\bar{A}$. $\mathfrak{D}_m$ and $\mathfrak{D}_n$ will denote the class of non-singular real or complex $m\times m$ or $n\times n$ diagonal matrices.

X and Y will always denote real or complex cartesian spaces of dimension n and m respectively, and with norms $\|.\|_\beta$ and $\|.\|_\alpha$ respectively.

All of $\mathfrak{M}_{mn}$, $\mathfrak{D}_m$, $\mathfrak{D}_n$, X and Y will be real or all of them will be complex.

This induces the quantities $\mathrm{lub}_{\alpha\beta}(A)=\max\limits_{x\neq 0}\|Ax\|_\alpha/\|x\|_\beta$ and $\mathrm{glb}_{\alpha\beta}(A)=\min\limits_{x\neq 0}\|Ax\|_\alpha/\|x\|_\beta$ for any $A\in\mathfrak{M}_{mn}$.

lub_{pq} and glb_{pq}, p and q real numbers or ∞, will denote the case that $\|.\|_\alpha$ and $\|.\|_\beta$ are the Hölder p- and q-norm respectively, defined as $\|x\|_p=(\Sigma|x_j|^p)^{1/p}$, and similarly for q.

$\|A\|_p=\mathrm{lub}_{pp}(A)$.

$|x|$ will denote the vector whose coordinates are the moduli of the corresponding coordinates of x; similarly for a matrix $|A|$.

A vector $x\in X$, $x\neq 0$, will be called a *maximizing* vector for A with respect to the given norms if $\|Ax\|_\alpha/\|x\|_\beta=\mathrm{lub}_{\alpha\beta}(A)$. Also, x will be called a *minimizing* vector if $\|Ax\|_\alpha/\|x\|_\beta=\mathrm{glb}_{\alpha\beta}(A)$.

Def. will indicate a definition, **Th.** will indicate a small or intermediate result.

1. Monotonic Matrix Functions

Def. 1.1. A vectornorm is called *absolute* if $\|x\| = \||x|\|$, it is called *monotic* if $|x| \leqq |y| \Rightarrow \|x\| \leqq \|y\|$ (cf. [1]).

We recall that the properties of absoluteness and monotonicity are equivalent (cf. [1], Theorem 2, where this property is proved for complex spaces, but the proof carries over to real spaces).

Def. 1.2. A vectornorm will be called *strongly monotonic* if it is monotonic and moreover $|x| \leqq |y| \wedge |x| \neq |y| \Rightarrow \|x\| < \|y\|$.

Th. 1.3. Any Hölder norm of index $<\infty$ is strongly monotonic.

Def. 1.4. A non negative function ϕ on $\mathfrak{M} \subset \mathfrak{M}_{mn}$ will be called *left-*, *right-*, or *two-sided monotonic* if for all $A \in \mathfrak{M}$ either

$$\mathfrak{D}_m \mathfrak{M} = \mathfrak{M} \quad \text{and} \quad \phi(DA) \leqq \phi(A) \max |d_{ii}| \quad \text{for all} \quad D \in \mathfrak{D}_m \tag{1.5}$$

or

$$\mathfrak{M} \mathfrak{D}_n = \mathfrak{M} \quad \text{and} \quad \phi(AD) \leqq \phi(A) \max |d_{ii}| \quad \text{for all} \quad D \in \mathfrak{D}_n \tag{1.6}$$

or both are satisfied.

Def. 1.7. Moreover we shall say that ϕ is *strongly left-*, *right-* or *two-sided monotonic at* A if in (1.5) or (1.6) or both the $<$ sign holds as soon as not all $|d_{ii}|$ are equal.

We list a few obvious properties.

Th. 1.8. If ϕ is left-monotonic then $\phi(A) \min |d_{ii}| \leqq \phi(DA) \leqq \phi(A) \max |d_{ii}|$.

Th. 1.9. If ϕ is right-monotonic then $\phi(A) \min |d_{ii}| \leqq \phi(AD) \leqq \phi(A) \max |d_{ii}|$.

Th. 1.10. If $|D| = I$ and ϕ is left- or right monotonic then $\phi(DA) = \phi(A)$ or $\phi(AD) = \phi(A)$ respectively.

Th. 1.11. If all matrices of $\mathfrak{M}$ have inverses and ϕ is left- or right-monotonic on $\mathfrak{M}$ then the function ψ defined by $\psi(A) = 1/\phi(A^{-1})$ is right- or left-monotonic respectively.

Theorem 1.12. The functions $\mathrm{lub}_{\alpha\beta}$ and $\mathrm{glb}_{\alpha\beta}$ are left-monotonic on $\mathfrak{M}_{mn}$ if and only if $\|.\|_\alpha$ is an absolute vectornorm.

Proof. If $\|.\|_\alpha$ is an absolute vectornorm then for $D \in \mathfrak{D}_m$, $\mathrm{lub}_{\alpha\alpha}(D) = \max |d_{ii}|$ (cf. [1], Theorem 3). Hence

$$\frac{\|DAx\|_\alpha}{\|x\|_\beta} \leqq \mathrm{lub}_{\alpha\alpha}(D) \frac{\|Ax\|_\alpha}{\|x\|_\beta} = \max |d_{ii}| \frac{\|Ax\|_\alpha}{\|x\|_\beta} \quad \text{for all} \quad x \in X, \quad x \neq 0.$$

This proves the if-part.

To prove the only-if part, we note that absolute norms are characterized by the fact that $\|Dy\|_\alpha = \|y\|_\alpha$ for any $D \in \mathfrak{D}_m$ with $|D| = \mathrm{I}$ and any $y \in Y$. Hence, if $\|.\|_\alpha$ is not absolute there exists a $D \in \mathfrak{D}_m$ with $|D| = I$ such that $\|Dy\|_\alpha > \|y\|_\alpha$

for a certain $y \in Y$. There exists a matrix $A \in \mathfrak{M}_{mn}$ and a vector $x \in X$ such that $y = Ax$ and $\|y\|_\alpha = \mathrm{lub}_{\alpha\beta}(A)\|x\|_\beta$ (cf. 6.1). Then $\mathrm{lub}_{\alpha\beta}(DA) > \mathrm{lub}_{\alpha\beta}(A)$ and hence $\mathrm{lub}_{\alpha\beta}$ cannot be left-monotonic (cf. 1.10).

There also exists a $D \in \mathfrak{D}_m$ with $|D| = I$ such that $\|Dy\|_\alpha < \|y\|_\alpha$ for a certain $y \in Y$. Put $\|Dy\|_\alpha = \|y\|_\alpha/(1+\varepsilon)$. There exists a matrix $A \in \mathfrak{M}_{mn}$ and a vector $x \in X$ such that $y = Ax$ and $\|y\|_\alpha < \mathrm{glb}_{\alpha\beta}(A)\|x\|_\beta(1+\varepsilon)$ (cf. 6.3). Hence $\mathrm{glb}_{\alpha\beta}(DA) < \mathrm{glb}_{\alpha\beta}(A)$ and hence $\mathrm{glb}_{\alpha\beta}$ cannot be left-monotonic. ‖

Theorem 1.13. The functions $\mathrm{lub}_{\alpha\beta}$ and $\mathrm{glb}_{\alpha\beta}$ are right-monotonic on $\mathfrak{M}_{mn}$ if and only if $\|.\|_\beta$ is an absolute vectornorm.

Proof. If $\|.\|_\beta$ is an absolute vectornorm then

$$\frac{\|A\,Dx\|_\alpha}{\|x\|_\beta} = \frac{\|Ay\|_\alpha}{\|D^{-1}y\|_\beta} \leqq \frac{\|Ay\|_\alpha}{\|y\|_\beta}\,\mathrm{lub}_{\beta\beta}(D) = \frac{\|Ay\|_\alpha}{\|y\|_\beta}\max|d_{ii}|$$

for all $x \in X$, $x \neq 0$, and $y = Dx$.

This proves the if-part.

If $\|.\|_\beta$ is not absolute, there exists a $D \in \mathfrak{D}_n$ with $|D| = I$ such that $\|Dx\|_\beta > \|x\|_\beta$ for a certain $x \in X$. Put $Dx = x'$. There exists a matrix $A \in \mathfrak{M}_{mn}$ such that $\|Ax'\|_\alpha = \mathrm{lub}_{\alpha\beta}(A)\|x'\|_\beta$ (cf. 6.1). Then

$$\frac{\|A\,Dx\|_\alpha}{\|x\|_\beta} > \frac{\|Ax'\|_\alpha}{\|x'\|_\beta} = \mathrm{lub}_{\alpha\beta}(A)$$

and hence $\mathrm{lub}_{\alpha\beta}(AD) > \mathrm{lub}_{\alpha\beta}(A)$; thus $\mathrm{lub}_{\alpha\beta}$ cannot be right-monotonic (cf. 1.10).

There also exists a $D \in \mathfrak{D}_n$ with $|D| = I$ such that $\|Dx\|_\beta < \|x\|_\beta$ for a certain $x \in X$. Put $\|Dx\|_\beta = \|x\|_\beta/(1+\varepsilon)$ and put $Dx = x'$. There exists a matrix $A \in \mathfrak{M}_{mn}$ such that $\|Ax'\|_\alpha < \mathrm{glb}_{\alpha\beta}(A)\|x'\|_\beta(1+\varepsilon)$ (cf. 6.3). Hence $\mathrm{glb}_{\alpha\beta}(AD) < \mathrm{glb}_{\alpha\beta}(A)$. ‖

As a corollary we have

Th. 1.14. If $\|.\|_\alpha$ and $\|.\|_\beta$ are Hölder norms of any index, the functions $\mathrm{lub}_{\alpha\beta}$ and $\mathrm{glb}_{\alpha\beta}$ are two-sided monotonic.

A different class of two-sided monotonic matrixfunctions is given by

Th. 1.15. Any matrixnorm which is obtained by considering an $m \times n$ matrix as element of the $m \times n$ dimensional Cartesian space with an absolute norm is two-sided monotonic.

Consequently:

Th. 1.16. $(\Sigma\,|a_{ij}|^p)^{1/p}$, $1 \leqq p \leqq \infty$, and in particular the Frobenius norm $\|A\|_F = \sqrt{\Sigma\,|a_{ij}|^2}$ and the function $\max|a_{ij}|$, is two-sided monotonic on $\mathfrak{M}_{mn}$.

More generally, but still a special case of 1.15:

Th. 1.17. Ostrowski's composite norm H, defined by $H(A) = h[h_1(A_1), \ldots, h_n(A_n)]$ (cf. [5], 3.1)), where h is monotonic in the positive orthant (in the sense of [1], (3.1)) and each h_i is absolute, is two-sided monotonic on $\mathfrak{M}_{mn}$. The same holds for H' defined as $H'(A) = H(A^H)$.

2. Equilibration Theorems

Def. 2.1. In the following B will be a matrix with *m rows* if B is left-multiplied by $D \in \mathfrak{D}_m$, B will have *n columns* if B is right-multiplied by $D \in \mathfrak{D}_n$. Also, $\|.\|_\gamma$ will denote any vectornorm.

Def. 2.2. For any two matrices A and B and any two matrix functions ϕ and ψ we define

$$\varkappa(B, A) = \psi(B)/\phi(A).$$

if the right hand side exists.

We then have the following two theorems

Theorem 2.3. Column Equilibration Theorem. If $\psi(B) = \max_j \|B_j\|_\gamma$ (where B_j denotes the j-th column of B) and ϕ is right-monotonic on $A\,\mathfrak{D}_n$ and $\widetilde{D} \in \mathfrak{D}_n$ is such that $B\widetilde{D}$ is column-equilibrated in the sense of $\|.\|_\gamma$ (i.e. all columns of $B\widetilde{D}$ have equal γ-norm). Then

(a) $$\varkappa(B\widetilde{D}, A\widetilde{D}) = \min_{D \in \mathfrak{D}_n} \varkappa(BD, AD)$$

(b) Any matrix D for which the minimum in (a) is attained may be obtained by multiplying $\widetilde{D}$ by a diagonalmatrix whose diagonal elements have equal modulus if and only if ϕ is strongly right-monotonic at $A\widetilde{D}$ (hence in this case column-equilibration of the argument of ψ is also a necessary condition for the minimum to be attained).

Proof. (a) It is no restriction to assume that $\widetilde{D} = I$, i.e. that B already is column-equilibrated. Then $\max_j \|BD_j\|_\gamma = \max_j |d_{jj}|\cdot \max_j \|B_j\|_\gamma$ and $\phi(AD) \leqq (\max_j |d_{jj}|)\,\phi(A)$.

(b) Trivial. ‖

Theorem 2.4. Row Equilibration Theorem. If $\psi(B) = \max_j \|(B^H)_j\|_\gamma$ (where $(B^H)_j$ denotes the j-th column of B^H, i.e. actually the j-th row of $\overline{B}$) and ϕ is left-monotonic on $\mathfrak{D}_m A$ and $\widetilde{D} \in \mathfrak{D}_m$ is such that $\widetilde{D}B$ is row-equilibrated in the sense of $\|.\|_\gamma$ (i.e. all columns of $(\widetilde{D}B)^H$ have equal γ-norm). Then

(a) $$\varkappa(\widetilde{D}B, \widetilde{D}A) = \min_{D \in \mathfrak{D}_m} \varkappa(DB, DA)$$

(b) Any matrix D for which the minimum in (a) is attained may be obtained by multiplying $\widetilde{D}$ by a diagonal matrix whose diagonal elements have equal modulus if and only if ϕ is strongly left-monotonic at $\widetilde{D}A$.

These equilibration theorems just serve to set a pattern. They are actually equivalent. They are so trivial that the applications can just as well be proved directly. The surprising thing, though, is that the minimizing $\widetilde{D}$ are determined by B only and do not at all depend on A, whereas uniqueness of $\widetilde{D}$ (apart from multiplication by diagonal matrices whose diagonal elements have equal modulus) is determined by ϕ and $A\widetilde{D}$ or $\widetilde{D}A$ only.

In 2.3, $\psi(B)$ obviously is a composite norm of B in the sense of [5], (3.1); a slight generalization can be obtained by allowing the norm $\|.\|_\gamma$ to be a different one for each column of B, i.e. $\psi(B) = \max_j \|B_j\|_{\gamma_j}$. Similarly for 2.4.

A few obvious applications of the equilibration theorems are contained in

Theorem 2.5. Let A be square or non-square. In the case that A is square the norm $\|.\|^*$ may be any Hölder norm or the Frobenius norm.

(a) $\varkappa(A)=\|A\|_\infty\|A^{-1}\|^*$ or $\|A\|_\infty/\mathrm{glb}_{pq}(A)$. Then $\varkappa(\tilde{D}A)$ is minimal if in $\tilde{D}A$ all rows have equal 1-norm.

(b) $\varkappa(A)=(\max|a_{ij}|)\|A^{-1}\|^*$ or $(\max|a_{ij}|)/\mathrm{glb}_{pq}(A)$. Then $\varkappa(\tilde{D}A)$ is minimal if in $\tilde{D}A$ all rows have equal ∞-norm.

(c) $\varkappa(A)$ as in (b). Then $\varkappa(A\tilde{D})$ is minimal if in $A\tilde{D}$ all columns have equal ∞-norm.

(d) $\varkappa(A)=\|A\|_1\|A^{-1}\|^*$ or $\|A\|_1/\mathrm{glb}_{pq}(A)$. Then $\varkappa(A\tilde{D})$ is minimal if in $A\tilde{D}$ all columns have equal 1-norm.

(e) $\varkappa_i(a)=\|A_i\|_\infty\|A^{-1}\|^*$ or $\|A_i\|_\infty/\mathrm{glb}_{pq}(A)$. Then $\varkappa_i(\tilde{D}A)$ is minimal if all coordinates of $\tilde{D}A_i$ have equal modulus.

(f) $\varkappa(A)=\|A\|_\infty/\|A\|^*$. Then $\varkappa(\tilde{D}A)$ is minimal if in $\tilde{D}A$ all rows have equal 1-norm.

(g) $\varkappa(A)=\|A\|_1/\|A\|^*$. Then $\varkappa(A\tilde{D})$ is minimal if in $A\tilde{D}$ all columns have equal 1-norm.

Proof. (a) $B=A$, $\|.\|_\gamma=\|.\|_1$; (b) $B=A$, $\|.\|_\gamma=\|.\|_\infty$; (c) $B=A$, $\|.\|_\gamma=\|.\|_\infty$; (d) $B=A$, $\|.\|_\gamma=\|.\|_1$; (e) $B=A_i$, $\|.\|_\gamma=\|.\|_p$; (f) $B=A$, $\|.\|_\gamma=\|.\|_1$; (g) $B=A$, $\|.\|_\gamma=\|.\|_1$. ‖

Obviously, [2], Theorem IIa, corollary, is a consequence of 2.5 (a). However, a generalization of [2], Theorem IIa itself, too, can be easily proved directly:

Th. 2.6. If $\varkappa(B,A)=\|B\|_\infty\|A^{-1}\|_\infty$ and B has no row consisting entirely out of zeros, then $\min\limits_{D\in\mathfrak{D}_n}\varkappa(DB,DA)=\||A^{-1}|\,|B|\|_\infty$.

Proof. $\||A^{-1}|\,|B|\|_\infty$ does not change when A and B are left-multiplied by $D\in\mathfrak{D}_n$ (note that $m=n$). We may therefore assume that in each row of B the sum of the moduli is 1. Now, if $\boldsymbol{e}_n$ denotes the n-dimensional vector with coordinates 1 only, then $\boldsymbol{e}_n$ is a maximizing vector with respect to the ∞-norm of any non-negative matrix with n columns. Similarly $\boldsymbol{e}_p$. Therefore, since $|B|\,\boldsymbol{e}_p=\boldsymbol{e}_n$ if B has p columns,

$$\||A^{-1}|\,|B|\|_\infty=\||A^{-1}|\,|B|\,\boldsymbol{e}_p\|_\infty=\||A^{-1}|\,\boldsymbol{e}_n\|_\infty=\||A^{-1}|\|_\infty=\|A^{-1}\|_\infty=\varkappa(B,A).\quad\|$$

3. Approximate Minimization

In view of the special form of ψ in the equilibration theorems, these theorems give conditions for the minimizing $\tilde{D}$ only for a limited class of condition numbers. From [2], Lemma I and the subsequent remarks it is seen that in other cases the Perron-eigenvectors of certain matrices play a role, and hence it cannot be expected that the minimizing $\tilde{D}$ can, in general, be easily determined.

Therefore we ask in those cases how well equilibration performs. In this respect we have the following obvious Theorems 3.1 and 3.3:

Theorem 3.1. Let B be any matrix and let ψ satisfy

$$p \max_j \|BD_j\|_\gamma \leqq \psi(BD) \leqq q \max_j \|BD_j\|_\gamma \tag{3.2}$$

where p and q are independent of $D \in \mathfrak{D}_n$. If ϕ is right-monotonic on $A\mathfrak{D}_n$ and $\tilde{D} \in \mathfrak{D}_n$ is such that $B\tilde{D}$ is column equilibrated in the sense of $\|.\|_\gamma$, then

$$\varkappa(B\tilde{D}, A\tilde{D}) \leqq \frac{q}{p} \inf_{D \in \mathfrak{D}_n} \varkappa(BD, AD).$$

Theorem 3.3. Let B be any matrix and let ψ satisfy

$$p \max_j \|((DB)^H)_j\|_\gamma \leqq \psi(DB) \leqq q \max_j \|((DB)^H)_j\|_\gamma \tag{3.4}$$

where p and q are independent of $D \in \mathfrak{D}_m$. If ϕ is left-monotonic on $\mathfrak{D}_m A$ and $\tilde{D} \in \mathfrak{D}_m$ is such that $\tilde{D}B$ is row-equilibrated in the sense of $\|.\|_\gamma$, then

$$\varkappa(\tilde{D}B, \tilde{D}A) \leqq \frac{q}{p} \inf_{D \in \mathfrak{D}_m} \varkappa(DB, DA).$$

Application:

Theorem 3.5. Assumptions as in 2.5. Also, let $\varkappa(A) = \|A\|_2 \|A^{-1}\|^*$ or $\|A\|_F \|A^{-1}\|^*$ or $\|A\|_2/\mathrm{glb}_{pq}(A)$ or $\|A\|_F/\mathrm{glb}_{pq}(A)$. Then

(a) $\varkappa(\tilde{D}A)$ is no more than a factor $\sqrt{m}$ away from its minimum if in $\tilde{D}A$ all rows have equal 2-norm.

(b) $\varkappa(A\tilde{D})$ is no more than a factor $\sqrt{n}$ away from its minimum if in $A\tilde{D}$ all columns have equal 2-norm.

Proof. (a) $B = A$, $\|.\|_\gamma = \|.\|_2$; $\max \|(B^H)_j\|_2 \leqq \|B\|_2 \leqq \|B\|_F \leqq (\sqrt{m}) \max \|(B^H)_j\|_2$.
(b) $B = A$, $\|.\|_\gamma = \|.\|_2$; $\max \|B_j\|_2 \leqq \|B\|_2 \leqq \|B\|_F \leqq (\sqrt{n}) \max \|B_j\|_2$. ‖

4. Applications to Symmetric Scaling

As a corollary of the results in the previous section we have

Theorem 4.1. Let $\varkappa(P) = \|P\|_2 \|P^{-1}\|_2$ for any non-singular $n \times n$ matrix P. Then, if P is positive definite and hermitean,

$$\varkappa(P) \leqq n \min_{D \in \mathfrak{D}_n} \varkappa(D^H P D)$$

if in P all diagonal elements are equal.

Proof. P can be written as $A^H A$, $A \in \mathfrak{M}_{nn}$, and obviously in A all columns have the same 2-norm. Also $\varkappa(P) = (\varkappa(A))^2$. Now apply 3.5 (b). ‖

This theorem has obvious applications to the matrices arising from least squares problems and from discretized elliptic differential equations. In the former case it is remarkable that only the *smaller* dimension of the $m \times n$ matrix of the least squares system enters 4.1, and since in this case n usually is not very large, symmetric scaling for equal diagonal elements can be considered as reasonably optimal.

The theorem reminds the reader to a result in [3], where it is proved that if in P all diagonal elements are equal and P has Young's "property A", then

$\varkappa(P) = \min_{D \in \mathfrak{D}_n} \varkappa(D^H P D)$. Thus particularly in the case of discretized differential equations our theorem may fall far behind the result in [3]. However, just in the case of lacunary systems we can improve 4.1 considerably.

We first have the following lemma, which may also have a wider interest.

Th. 4.2. Let A be an $m \times n$ matrix with at most q non-zero elements in any column or at most q non-zero elements in any row. Then $\|A\|_2 \leqq (\sqrt{q}) \max \|(A^H)_i\|_2$ or $\|A\|_2 \leqq (\sqrt{q}) \max \|A_j\|_2$ respectively.

Proof. It is sufficient to consider the case of at most q non-zero elements in a column since $\|A^H\|_2 = \|A\|_2$. Then

$$\|A x\|_2^2 = \sum_i \Big| \sum_j a_{ij} x_j \Big|^2 \leqq \sum_i \left(\|(A^H)_i\|_2^2 \sum\nolimits^{(i)} |x_j|^2 \right) \leqq \left(\max_i \|(A^H)_i\|_2^2 \right) \sum_i \sum\nolimits^{(i)} |x_j|^2,$$

where $\sum^{(i)}$ indicates that the summation index j assumes only those values for which $a_{ij} \neq 0$. Hence $\sum_i \sum^{(i)}_j |x_j|^2 \leqq q \|x\|_2^2$. ‖

In some interesting cases this theorem performs remarkably well. E.g. for the Dirichlet problem on a square for $\Delta u = 0$, where Δu is approximated by $\{u(x+h, y) + u(x-h, y) + u(x, y+h) + u(x, y-h) - 4u(x, y)\}/h^2$, a positive definite symmetric matrix arises with diagonal elements 4 and at most 4 elements -1 and further zeros on each row. This matrix is known to have a largest eigenvalue ~ 8 (cf. [4], p. 230), and hence its 2-norm is approximately equal to 8. Our 4.2 gives $(\sqrt{5})\,(\sqrt{20}) = 10$. Admittedly, this result is not as good as the upper bound 8 which is derived from Gershgorin's circle theorem. This is not surprising, however, since for hermitean matrices our 4.2 actually is a consequence of this theorem, as is easily verified. But 4.2 can be used for non-symmetric matrices also.

We now have the following refinement of 4.1.

Theorem 4.3. Assumptions as in 4.1. If, moreover, P has at most q non-zero elements in any row then

$$\varkappa(P) \leqq q \min_{D \in \mathfrak{D}_n} \varkappa(D^H P D)$$

if in P all diagonal elements are equal.

Proof. Let A be any $n \times n$ matrix such that $Q = A^H A$ has at most q non-zero elements in any row and column. Then $\|A\|_2^2 = \|Q\|_2 \leqq (\sqrt{q}) \max \|Q_j\|_2$ (cf. 4.2). Also $q_{ij} = (A_i)^H A_j$, hence $|q_{ij}| \leqq \|A_i\|_2 \|A_j\|_2$. Since Q_j has at most q non-zero elements, $\|Q_j\|_2^2 \leqq q \max \|A_i\|_2^4$. Therefore, $\|A\|_2 \leqq (\sqrt{q}) \max \|A_i\|_2$.

Now writing $P = A^H A$, we have $\|A D\|_2 \leqq (\sqrt{q}) \max \|A D_i\|_2$. Since P has equal diagonal elements, all columns of A have equal 2-norm. Therefore $\varkappa(A) \leqq (\sqrt{q}) \min \varkappa(A D)$ (cf. 3.1). ‖

5. Uniqueness of Minimizing Scaling

Assertions 2.3 (b) and 2.4 (b) say that under circumstances the matrices D for which $\varkappa(BD, AD)$ or $\varkappa(DB, DA)$ is minimized, are essentially unique.

Unfortunately, these circumstances — strong monotonicity of ϕ — usually cannot easily be verified.

There are norms for which strong monotonicity of glb (and for the usual condition numbers this is the important case) is rare:

Th. 5.1. If $\phi(A) = 1/\|A^{-1}\|_\infty$ then ϕ is not strongly right-monotonic at A unless all rows of A^{-1} have equal 1-norm. If $\phi(A) = 1/\|A^{-1}\|_1$ then ϕ is not strongly left-monotonic at A unless all columns of A^{-1} have equal 1-norm.

However, for many other practically important norms glb is strongly monotonic for an overwhelming majority of matrices A (even if $A = B$), e.g.

Th. 5.2. (a) If $\phi(A) = 1/\|A^{-1}\|_1$ and A^{-1} has a maximal column (i.e. a column whose 1-norm equals $\|A^{-1}\|_1$) without an element 0 then ϕ is strongly right-monotonic at A.

(b) If $\phi(A) = 1/\|A^{-1}\|_\infty$ and A^{-1} has a maximal row (i.e. a row whose 1-norm equals $\|A^{-1}\|_\infty$) without an element 0 then ϕ is strongly left-monotonic at A.

Also:

Th. 5.3. If $\phi(A) = 1/\|A^{-1}\|^*$, $\|.\|^*$ a norm as in 1.15, then ϕ is strongly right- and left-monotonic at A if the norm on the $m \times n$ dimensional space (cf. 1.15) is strongly monotonic (m and n are now equal).

For other norms the following result has some relevance:

Theorem 5.4. (a) If $\phi(A) = \mathrm{glb}_{\alpha\beta}(A)$, $\|.\|_\beta$ strongly monotonic, and A has at least one minimizing vector x_0 without a coordinate 0 then ϕ is strongly right-monotonic at A.

(b) If $\phi(A) = \mathrm{glb}_{\alpha\beta}(A)$, $\|.\|_\alpha$ strongly monotonic, and A has at least one minimizing vector x_0 such that $A x_0$ has no coordinate 0, then ϕ is strongly left-monotonic at A.

Proof.

$$\text{(a)}\quad \mathrm{glb}_{\alpha\beta}(AD) = \min_{x \neq 0} \frac{\|ADx\|_\alpha}{\|x\|_\beta} = \min_{x \neq 0} \frac{\|Ax\|_\alpha}{\|D^{-1}x\|_\beta} \leqq \frac{\|Ax_0\|_\alpha}{\|D^{-1}x_0\|_\beta} < \frac{\|Ax_0\|_\alpha}{\|x_0\|_\beta} \max |d_{ii}|$$

if not all $|d_{ii}|$ are equal.

$$\text{(b)}\quad \mathrm{glb}_{\alpha\beta}(DA) = \min_{x \neq 0} \frac{\|DAx\|_\alpha}{\|x\|_\beta} \leqq \frac{\|DAx_0\|_\alpha}{\|x_0\|_\beta} < \frac{\|Ax_0\|_\alpha}{\|x_0\|_\beta} \max |d_{ii}|$$

if not all $|d_{ii}|$ are equal. ‖

The difficulty, of course, is to know whether the condition on x_0 is satisfied. Although the author cannot prove this, he believes that at least in the case that $\phi(A) = 1/\|A^{-1}\|_p$, $1 < p < \infty$, the set of matrices that do not satisfy the condition, has measure 0 in the set of all matrices. He can prove, however, that if $\phi(A) = 1/\|A^{-1}\|_p$, $1 < p < \infty$, and A^{-1} is a positive matrix, then A satisfies the conditions of 5.4. This shows at least that the set of matrices satisfying 5.4 has positive measure, so that one certainly cannot *trust* that there is any real freedom left in choosing the scaling matrix if one wants to minimize $\varkappa$.

Another question, of course, is what freedom one is likely to gain in this respect if it is only required to get $\varkappa$ within a given factor (e.g. 2 or n) from its minimum.

6. Appendix

In this appendix we mention and prove a few results on the existence of matrices with given maximizing or minimizing vectors or given images of such vectors.

Theorem 6.1. For any $x \in X$, $x \neq 0$, and any $y \in Y$ and any pair of norms there exists an $A \in \mathfrak{M}_{mn}$ such that $y = Ax$ and x is a maximizing vector for A.

This theorem has no immediate parallel for minimizing vectors:

Th. 6.2. If $n \geqq 2$ it is not true that for any pair of norms and any $x \in X$, $x \neq 0$, there exist an $A \in \mathfrak{M}_{mn}$ and a $y \in Y$, $y \neq 0$, such that $y = Ax$ and x is a minimizing vector for A. Neither is it true that for any $y \in Y$, $y \neq 0$, and any pair of norms there exist an $A \in \mathfrak{M}_{mn}$ and an $x \in X$, $x \neq 0$, such that $y = Ax$ and x is a minimizing vector for A.

This implies that the mapping A in 6.1 cannot be required to be *injective* (i.e. to have an inverse). However:

Theorem 6.3. For any $x \in X$, $x \neq 0$, and any $y \in Y$, $y \neq 0$, and any pair of norms and any $\varepsilon > 0$ there exists an $A \in \mathfrak{M}_{mn}$ such that $y = Ax$ and $\|Ax\|_\alpha / \|x\|_\beta < (1+\varepsilon)\,\mathrm{glb}_{\alpha\beta}(A)$.

Proof of 6.1. Let L be any linear functional on the variety of scalar multiples of x. Now extend L to the whole of X without increasing the norm of L (Hahn-Banach theorem). Then x remains a maximizing vector of L. Finally, define A by $At = \frac{Lt}{Lx} y$ for all $t \in X$. ‖

Proof of 6.2, first part. Provide X and Y with the Hölder 1-norm and 2-norm respectively. Let $A \in \mathfrak{M}_{mn}$ have rank n (since otherwise $\mathrm{glb}(A) = 0$). Take $x = (1, 0, 0, \ldots)^H$ and $x' = (0, 1, 0, \ldots)^H$. Then $\|x + \lambda x'\|_1 = 1 + |\lambda|$. If $Ax' \perp Ax$ then $\|Ax + \lambda Ax'\|_2$ behaves as $p + q\lambda^2$, p and q positive constants (since Ax' and Ax cannot vanish), for small values of λ and hence x is no minimizing vector. If $\neg(Ax' \perp Ax)$ then $\|Ax + \lambda Ax'\|_2 \leqq \|Ax\|_2$ for some values of $\lambda \neq 0$, and again x is no minimizing vector. ‖

Proof of 6.2, second part. Provide X and Y with the Hölder 2-norm and ∞-norm respectively. Let $A \in \mathfrak{M}_{mn}$ have rank n (since otherwise $\mathrm{glb}(A) = 0$). Take $y = (1, 0, 0, \ldots)^H$, and suppose that $y = Ax$. Since A has rank n, there certainly is an $x' \in X$ such that $y' = Ax' \neq 0$, but has first coordinate 0. Then $\|y + \lambda y'\|_\infty$ is constant for small values of λ. However, $\|x + \lambda x'\|_2 > \|x\|_2$ for some small values of λ. Hence, x is no minimizing vector. ‖

Proof of 6.3. Let Y' be any n-dimensional subspace of Y which contains y. Then there exists a mapping $B: Y' \to X$ such that $x = By$ and y is a maximizing vector for B (cf. 6.1), but this mapping may be singular. In any neighbourhood of B there exist non-singular mappings $B': Y' \to X$. Thus for any $\eta > 0$ there exists a non-singular B' such that $\|B'y - By\|_\beta < \eta \|By\|_\beta$ and $\mathrm{lub}_{\beta\alpha}(B') < (1+\eta)\,\mathrm{lub}_{\beta\alpha}(B)$.

Now take $\eta < 1$. Then for any p and $q \in X$ with $\|q - p\|_\beta < \eta \|p\|_\beta$ there exists a non-singular mapping $C: X \to X$ such that $\mathrm{lub}_{\beta\beta}(C - I) < \eta$ and $Cp = q$. Indeed, we can take $C - I$ such that $(C - I)\,p = q - p$ and p is a maximizing vector for $C - I$ (cf. 6.1); the non-singularity follows from $\mathrm{lub}_{\beta\beta}(C - I) < \eta < 1$. Hence C^{-1} exists and $\mathrm{lub}_{\beta\beta}(C^{-1}) < 1/(1-\eta)$.

Applying this with $p = By = x$ and $q = B'y$ we get

(i) $x = C^{-1}B'y$;

(ii) $\mathrm{lub}_{\beta\alpha}(C^{-1}B') < \mathrm{lub}_{\beta\alpha}(B)\,(1+\eta)/(1-\eta)$.

Hence

$$\frac{\|y\|_\alpha}{\|x\|_\beta} = \frac{1}{\mathrm{lub}_{\beta\alpha}(B)} < \frac{1}{\mathrm{lub}_{\beta\alpha}(C^{-1}B')}\,\frac{1+\eta}{1-\eta} = \mathrm{glb}_{\alpha\beta}\left((C^{-1}B')^{-1}\right)\frac{1+\eta}{1-\eta}.$$

Thus, if η is small enough, $A = (C^{-1}B')^{-1}$ satisfies all requirements. ‖

Acknowledgement. The author recalls with pleasure instructive talks on this subject with Prof. F. L. Bauer (Munich, Germany).

References

1. Bauer, F. L., Stoer, J., Witzgall, C.: Absolute and monotonic norms. Num. Math. **3**, 257–264 (1961).
2. — Optimally scaled matrices. Num. Math. **5**, 73–87 (1963).
3. Forsythe, G. E., Straus, E. G.: On best conditioned matrices. Proc. Amer. Math. Soc. **6**, 340–345 (1955).
4. — Wasow, W. R.: Finite difference methods for partial differential equations. New York: John Wiley 1960.
5. Ostrowski, A.: Über Normen von Matrizen. Math. Zeitschr. **63**, 2–18 (1955).

A. van der Sluis
Rekencentrum Rijksuniversiteit
Budapestlaan 6
Utrecht, Netherlands