

M-Matrix Theory and Recent Results
in Numerical Linear Algebra

Richard S. Varga*

Department of Mathematics
Kent State University
Kent, OH 44242

*Dedicated to Garrett Birkhoff on the occasion of his 65th
birthday, January 10, 1976.

§1. Introduction.

That the theory of M-matrices, as introduced by Ostrowski [19] in 1937, provides fundamental tools in the analysis of problems in numerical linear algebra, in particular in the iterative solution of large sparse systems of linear equations, is of course well known to all. What is perhaps surprising is that the theory of M-matrices continues to be the underlying theme, whether directly or indirectly, for many recent contributions in numerical linear algebra. The aim of this short contribution is sketch how the theory of M-matrices serves as a basis for some recent contributions.

We remark that the material presented here is a condensation of the material to appear in [2], [26], and [27].

* Research supported in part by the Air Force Office of Scientific Research under Grant AFOSR-74-2729, and by the Energy Research and Development Administration (ERDA) under Grant E(11-1)-2075.

§2. Notation and Terminology.

For a positive integer n , let $\langle n \rangle := \{1, 2, \dots, n\}$, let $\mathbb{C}^{n,n}$ denote the collection of all $n \times n$ complex matrices $A = [a_{i,j}]$, and let $\mathbb{C}_{\pi}^{n,n}$ denote the subset of matrices of $\mathbb{C}^{n,n}$ with all diagonal elements nonzero.

Similarly, let \mathbb{R}^n denote the complex n -dimensional vector space of all column vectors $v = [v_1, v_2, \dots, v_n]^T$, where $v_i \in \mathbb{C}$ for all $i \in \langle n \rangle$. The restriction to real entries or components similarly defines $\mathbb{R}_{\pi}^{n,n}$, $\mathbb{R}_{\pi\pi}^{n,n}$, and $\mathbb{R}_{\pi\pi\pi}^n$.

For any $A \in \mathbb{C}^{n,n}$, set $\text{spec}[A] := \{\lambda : \det(A - \lambda I) = 0\}$, and call $\rho(A) := \max\{|\lambda| : \lambda \in \text{spec}[A]\}$ the spectral radius of A . Next, for $v \in \mathbb{R}^n$, we write $v > 0$ or $v \geq 0$, respectively, if $v_i > 0$ or if $v_i \geq 0$ for all $i \in \langle n \rangle$. Similarly, for $A = [a_{i,j}] \in \mathbb{R}^{n,n}$, we write $A > \theta'$ or $A \geq \theta'$, respectively, if $a_{i,j} > 0$ or if $a_{i,j} \geq 0$ for all $i, j \in \langle n \rangle$. Also, given $A = [a_{i,j}] \in \mathbb{C}_{\pi}^{n,n}$, then $|A| \in \mathbb{R}^{n,n}$ is defined by $|A| := [|a_{i,j}|]$. Next, given $A = [a_{i,j}] \in \mathbb{C}_{\pi\pi}^{n,n}$, the set of matrices defined by

$$(2.1) \quad \Omega(A) := \{B = [b_{i,j}] \in \mathbb{C}^{n,n} : |b_{i,j}| = |a_{i,j}| ; i, j \in \langle n \rangle\}$$

is called the equimodular set of matrices associated with the given matrix A .

For any $A = [a_{i,j}] \in \mathbb{C}_{\pi\pi\pi}^{n,n}$, we can decompose each $B = [b_{i,j}]$ in $\Omega(A)$ into the sum

$$(2.2) \quad B = D(B) - L(B) - U(B),$$

where $D(B) = \text{diag}[b_{1,1}, b_{2,2}, \dots, b_{n,n}]$ is nonsingular, and where $L(B)$ and $U(B)$ are respectively strictly lower and strictly upper triangular matrices. From the decomposition

of (2.2), three familiar iteration matrices can be defined.

For any $\omega > 0$,

$$(2.3) \quad J_{\omega}(B) := \omega(D(B))^{-1} \{L(B)+U(B)\} + (1-\omega)I$$

is the (point) Jacobi overrelaxation iteration matrix for B,

$$(2.4) \quad \mathcal{L}_{\omega}(B) := [D(B)-\omega L(B)]^{-1} \{ (1-\omega)D(B)+\omega U(B) \}$$

is the (point) successive overrelaxation iteration matrix (SOR matrix) for B, and

$$(2.5) \quad S_{\omega}(B) := [D(B)-\omega U(B)]^{-1} \{ (1-\omega)D(B)+\omega L(B) \} [D(B)-\omega L(B)]^{-1} \cdot \\ \{ (1-\omega)D(B)+\omega U(B) \}$$

is the (point) symmetric successive overrelaxation iteration matrix (SSOR matrix) for B.

Next, to define M-matrices, consider any $B = [b_{i,j}] \in \mathbb{R}^{n,n}$ with $b_{i,j} \leq 0$ for all $i \neq j$ with $i, j \in \langle n \rangle$. Then, B can be expressed as the difference

$$(2.6) \quad B = \tau_B \cdot I - C(B),$$

where $\tau_B := \max\{b_{i,i} : i \in \langle n \rangle\}$, and where $C(B) = [c_{i,j}] \in \mathbb{R}^{n,n}$ satisfying $C(B) \geq \theta$, has its entries given by

$$(2.7) \quad c_{i,i} = \tau_B - b_{i,i} \geq 0; \quad c_{i,j} = -b_{i,j}, \quad i \neq j; \quad i, j \in \langle n \rangle.$$

Following Ostrowski [19], we say that such a matrix B is a nonsingular M-matrix iff $\tau_B > \rho(C(B))$. Next, given any $A = [a_{i,j}] \in \mathbb{C}^{n,n}$, we define, as in [19], its comparison matrix $\mathfrak{M}(A) = [\alpha_{i,j}] \in \mathbb{R}^{n,n}$ by

$$(2.8) \quad \alpha_{i,i} = |a_{i,i}|; \quad \alpha_{i,j} = -|a_{i,j}|, \quad i \neq j; \quad i, j \in \langle n \rangle,$$

and A is defined, again as originally in Ostrowski [19], to be a nonsingular H-matrix iff $\mathfrak{M}(A)$ is a nonsingular M-matrix.

How does one recognize if a given matrix $A \in \mathbb{C}^{n,n}$ is a nonsingular H-matrix, or equivalently, if $\mathfrak{M}(A)$ is a non-singular M-matrix? Perhaps best known to most readers (cf. [25, p. 83]) is the following: $\mathfrak{M}(A)$ is a nonsingular M-matrix iff $(\mathfrak{M}(A))^{-1} \geq 0$. But this is just one of the many known equivalent conditions for a nonsingular M-matrix which has been studied in the classic and famous papers of Taussky [23], Fan [10], and Fiedler and Pták [11]. This last paper lists thirteen equivalent conditions for a nonsingular M-matrix!

§3. Main Result and Commentary.

In analogy with the previous mentioned papers by Taussky [23], Fan [10], and Fiedler and Pták [11], we collect now some known, as well as some new, equivalent conditions for a given $A \in \mathbb{C}_{\pi}^{n,n}$ to be a nonsingular H-matrix, and then we comment briefly on how these equivalent conditions relate to recent results (back to 1968) in the literature.

Theorem 1. For any $A = [a_{i,j}] \in \mathbb{C}_{\pi}^{n,n}$, $n \geq 2$, the following are equivalent:

- i) A is a nonsingular H-matrix, i.e., $\mathfrak{M}(A)$ is a nonsingular M-matrix;
- ii) there exists a $\underline{u} \in \mathbb{R}^n$ with $\underline{u} > 0$ such that $\mathfrak{M}(A) \cdot \underline{u} > 0$;
- iii) $\mathfrak{M}(A)$ is of generalized positive type, i.e., there exists a $\underline{u} \in \mathbb{R}^n$ such that
 - a) $\underline{u} > 0$, $\mathfrak{M}(A) \cdot \underline{u} \geq 0$, and $\{i \in \langle n \rangle : \mathfrak{M}(A) \cdot \underline{u}_i > 0\}$ is nonempty;
 - b) for each $i_0 \in \langle n \rangle$ with $\mathfrak{M}(A) \cdot \underline{u}_{i_0} = 0$,

there exist indices i_1, i_2, \dots, i_r in $\{n\}$ with $a_{i_k, i_{k+1}} \neq 0$, $0 \leq k \leq r-1$,

such that $\Omega(A) \cdot \underline{u} > 0$;

- iv) there exists a $\underline{u} \in \mathbb{R}^n$ with $\underline{u} > 0$ such that $\Omega(A) \cdot \underline{u} \geq 0$, and such that $\sum_{j \leq i} a_{i,j} u_j > 0$ for all $i \in \{n\}$, where $\Omega(A) := [a_{i,j}]$ (cf. (2.8));
- v) there exist respectively lower triangular and upper triangular nonsingular M-matrices L and U such that $\Omega(A) = L \cdot U$;

- vi) for any $B \in \Omega(A)$, $\rho(J_1(B)) \leq \rho(|J_1(B)|) = \rho(J_1(\Omega(A))) < 1$;
- vii) for any $B \in \Omega(A)$ and any $0 < w < 2/[1 + \rho(J_1(B))]$, $\rho(J_w(B)) \leq w \cdot \rho(J_1(B)) + |1-w| < 1$;
- viii) for any $B \in \Omega(A)$ and any $0 < w < 2/[1 + \rho(|J_1(B)|)]$, $\rho(\mathcal{L}_w(B)) \leq w \cdot \rho(|J_1(B)|) + |1-w| < 1$;
- ix) for any $B \in \Omega(A)$ and any $0 < w < 2/[1 + \rho(|J_1(B)|)]$, $\rho(S_w(B)) < 1$.

That ii) is equivalent to i) is known, and is due to Fan [10]; that v) is equivalent to i) is known, and is due to Fiedler and Pták [11]. The definition of generalized positive type in iii) is a slight extension of a definition due to Bramble and Hubbard [7]; that iii) is equivalent to i) can be found in Varga [26], Rheinboldt [20], and More [17]. That iv) is equivalent to i) is new, and is due to Beauwens [4]. That vi), vii), and viii) are equivalent to i) is essentially new, and can be found in Varga [26]; similarly, that ix) is equivalent to i) is new, and can be found in Alefeld and Varga [2].

We now briefly comment on how certain recent results in the literature (back to 1968) are equivalent to, or weaker than, the listed conditions of Theorem 1. To begin, James and Riha [14] have defined $A = [a_{i,j}] \in \mathbb{C}^{n,n}$ to have generalized column diagonal dominance if there exists a $\underline{u} = [u_1, u_2, \dots, u_n]^T \in \mathbb{R}^n$ with $\underline{u} > \underline{0}$ such that

$$(3.1) \quad |a_{i,i}|u_i > \sum_{\substack{j \in \langle n \rangle \\ j \neq i}} |a_{i,j}|u_j, \quad \text{for all } i \in \langle n \rangle.$$

This definition, however, is precisely equivalent to ii) of Theorem 1, i.e., there exists a $\underline{u} \in \mathbb{R}^n$ with $\underline{u} > \underline{0}$ such that $\mathfrak{M}(A) \cdot \underline{u} > \underline{0}$. These authors in essence show [14, Theorem 4], under the added (unnecessary) assumption that A is irreducible, that ii) of Theorem 1 is equivalent to the combined hypotheses that $P(J_1 \mathfrak{M}(A)) < 1$ and $P(\mathcal{L}_w \mathfrak{M}(A)) < 1$ for $0 < w \leq 1$, which is weaker than the separate equivalences of ii), vi), and viii) of Theorem 1. Similarly, James [13, Theorem 1] shows that strict irreducible diagonal dominance for A (cf. [25, p. 23]), a stronger assumption than ii) of Theorem 1, implies that $P(\mathcal{L}_1(A)) < 1$, which is weaker than viii) of Theorem 1.

Next, given $A = [a_{i,j}] \in \mathbb{C}^{n,n}$, suppose that A is diagonally dominant, i.e., $\mathfrak{M}(A) \cdot \underline{\zeta} \geq \underline{0}$, where $\underline{\zeta} := [1, 1, \dots, 1]^T$. If A is singular, so that A is not a nonsingular H-matrix, then from i) and iii) of Theorem 1, $\mathfrak{M}(A)$ cannot be of generalized positive type with respect to the vector $\underline{\zeta}$.

Without going into detail, we simply remark that negating the property of generalized positive type (with respect to $\underline{\zeta}$) duplicates the main result of Erdelsky [9, Theorem 1]. In defining his Zeilensummenbedingung, Bohl [5, 6]

weakens the assumption for generalized positive type matrices in iii) of Theorem 1 by allowing $\underline{u} \in \mathbb{R}^n$ in iii a) to satisfy

$\underline{u} \geq 0$, but then immediately shows [6, Satz 2.1] that if $\underline{u} \in \mathbb{R}^n$ with $\underline{u} \geq 0$ satisfies the remaining conditions of iii), then in fact $\underline{u} > 0$. Consequently, Bohl's Zeilensummenbedingung and the hypothesis of generalized positive type, iii) of Theorem 1, are equivalent. For $A \in \mathbb{R}^{n,n}$, with $A \geq 0$, Bohl [6, Satz 2.2] shows that i), ii), and iii) of Theorem 1 are equivalent when applied to $I-A$.

Defining for any $A = [a_{i,j}] \in \mathbb{C}^{n,n}$ its induced operator norm $\|A\|_\infty$ by

$$(3.1) \quad \|A\|_\infty := \max_{i \in \langle n \rangle} \left\{ \sum_{j \in \langle n \rangle} |a_{i,j}| \right\},$$

Schäfke [21, Satz 1], improving on a paper by Walter [28], gives six equivalent conditions for an $A = [a_{i,j}] \in \mathbb{R}^{n,n}$ satisfying $A \geq 0$ and $\|A\|_\infty \leq 1$, to have $\rho(A) < 1$. This can be viewed as finding equivalent conditions on A for $I-A$ to be a nonsingular M-matrix. Condition 4 of [21, Satz 1], in fact, reduces to ii) of Theorem 1 in this case.

Next, Kulisch [15, Theorem 1], as a special case, establishes that $\rho(\mathcal{L}_w(B)) < 1$ for any $0 < w \leq 2/[1+\rho(|J_1(B)|)]$ and for any B with $\rho(|J_1(B)|) < 1$, and deduces [15, Cor. 1.3] that B , being either strictly or irreducibly diagonally dominant, is sufficient for $\rho(|J_1(B)|) < 1$. This last deduction follows more generally from the equivalence of iii) and vi) of Theorem 1. See also Apostolatos and Kulisch [3].

Continuing, Elsner [8] gives the definition of a verallgemeinerter Zeilensummenkriterium, which turns out to be precisely condition ii) of Theorem 1, applied to the matrix $I-A$. As consequences of his definition, Elsner in essence shows that ii) implies iii) of Theorem 1, and that ii) implies the convergence of the Gauss-Seidel iterative method, a special case $w=1$ of viii) of Theorem 1.

Next, in the iterative solution of nonlinear systems of equations, a number of authors have contributed results which, in the linear case, relate to various parts of Theorem 1. For example, Müller [18] formulates the concept [18, Def. 5] of a chained weakly contractive system which, when applied to the linear matrix equation $(I-A)\underline{x} = \underline{b}$, reduces precisely to iii) of Theorem 1, i.e., $\mathfrak{M}(I-A)$ is of generalized positive type. In the spirit of Theorem 1 and the work of Schäfke [21], Müller [18, Sätze 4, 5, 5a] develops ten consequences and equivalences (when $A \geq 0$) of his chained weakly contractive system in the linear case, as, for example, the equivalence of iii), iii), and vi) of Theorem 1. In a similar vein, Rheinboldt [20, Thm. 4.4] deduces, as a special case of such nonlinear investigations, the equivalence of i), ii), and iii) of Theorem 1, while More [17] gives the definition in the linear case of Ω -diagonally dominant matrices, which turns out to be equivalent to the assumption that $\mathfrak{M}(A)$ be of generalized positive type (cf. iii) of Theorem 1) with respect to the vector $\underline{\zeta} = [1, 1, \dots, 1]^T$. More then shows [17, Thm. 4.7] the equivalence of iii) and i) of Theorem 1.

Continuing our discussion, in Young [29, p. 43], one can deduce the equivalence of i) and vi) of Theorem 1, and it is shown [29, p. 107] that if A is irreducibly diagonally dominant, a stronger hypothesis than iii) of Theorem 1, then $\rho(J_{\omega}(A)) < 1$ and $\rho(J_{\omega}(A)) < 1$ for any $0 < \omega \leq 1$, which are respectively weaker than vii) and viii) of Theorem 1. It is also shown [29, p. 126] that i) of Theorem 1 implies $\rho(J_{\omega}(\mathfrak{M}(A))) < 1$ for any $0 < \omega < 2/[1+\rho(J_1(\mathfrak{M}(A)))]$, which is a special case of viii).

Next, it is interesting to note that Beauwens [4], who introduces the condition iv) of Theorem 1, calls this

property when $\underline{u} = \underline{\zeta}$, lower semi-strictly diagonal dominance, and shows [4] that this property, coupled with irreducibility (cf. [25, p. 19]), is equivalent with irreducible diagonal dominance. Next, the result of Jacobsen [12] and Meijerink and Van der Vorst [16] is that a nonsingular symmetric M-matrix (which is necessarily positive definite) can be factored as $G \cdot G^T$, where G is a nonsingular triangular M-matrix, which in essence is weaker than the equivalence of i and v .

Finally, in Shrivakumar and Chew [22], one finds as the main result that the special case $\underline{u} = \underline{\zeta}$ of Theorem 1, implies that A is nonsingular, which is weaker than the equivalence of i) and ii) of Theorem 1.

§4. On Bounding $\|A^{-1}\|_\infty$.

In a recent paper, Varah [24] established

Theorem A. Assume that $A = [a_{i,j}] \in \mathbb{C}^{n,n}$, $n \geq 2$, is strictly diagonally dominant, i.e.,

$$(4.1) \quad \{|a_{i,i}| - \sum_{\substack{j \in \langle n \\ j \neq i}} |a_{i,j}| > 0, \quad \text{for all } i \in \langle n \rangle,$$

and set

$$(4.2) \quad \alpha := \min_{i \in \langle n \rangle} \{|a_{i,i}| - \sum_{\substack{j \in \langle n \rangle \\ j \neq i}} |a_{i,j}| \}.$$

Then (cf. (3.1)),

$$(4.3) \quad \|A^{-1}\|_\infty \leq 1/\alpha,$$

and

Theorem B. Assume that $A = [a_{i,j}] \in \mathbb{C}^{n,n}$ and A^T are both strictly diagonally dominant, and set
 $\beta := \min_{i \in \langle n \rangle} \{|a_{i,i}| - \sum_{\substack{j \in \langle n \rangle \\ j \neq i}} |a_{j,i}| \}$. Then, the smallest

singular value $\sigma_n(A)$ of A can be bounded below by

$$(4.4) \quad \sigma_n(A) := (\|A^{-1}\|_2)^{-1} \geq \sqrt{\alpha \beta}.$$

We first remark that Theorem A is known in the literature: see Ahlberg and Nilsson [1, p. 96]. Now, using the theory of M-matrices, we show how Theorems A and B can be generalized. First, the assumption in Theorem A that A is strictly diagonally dominant is equivalent to assuming that $\mathfrak{M}(A) \cdot \underline{\zeta} > \underline{0}$, whence ii) of Theorem 1 is satisfied with the particular vector $\underline{u} = \underline{\zeta} = [1, 1, \dots, 1]^T$. Thus, from the equivalence of i) and ii) of Theorem 1, assuming that A is a nonsingular H-matrix implies (after a simple normalization) that the set

$$(4.5) \quad U_A := \{\underline{u} \in \mathbb{R}^n : \underline{u} > \underline{0}, \mathfrak{M}(A) \cdot \underline{u} > \underline{0}, \text{ and } \|\underline{u}\|_\infty = 1\}$$

is nonempty. Then, define

$$(4.6) \quad f_A(\underline{u}) := \min_{i \in \{1, \dots, n\}} \{\mathfrak{M}(A) \cdot \underline{u}\}_i, \quad \text{for any } \underline{u} \in U_A.$$

The generalization (cf. [27]) of Theorem A is

Theorem 2. If $A \in \mathbb{C}^{n,n}$, $n \geq 2$, is a nonsingular H-matrix, then

$$(4.7) \quad \sup_{B \in \Omega_A} \|B^{-1}\|_\infty = \|\mathfrak{M}(A)^{-1}\|_\infty = \frac{1}{\max\{f_A(\underline{u}) : \underline{u} \in U_A\}}.$$

Note that if A , as in Theorem A, is strictly diagonally dominant, then $\underline{\zeta} \in U_A$, and $f_A(\underline{\zeta}) = \alpha$, where α is defined in (4.2). Recalling that A is an element of Ω_A from (2.1), we see that (4.7) of Theorem 2 implies (4.3) of Theorem A. Similarly, the following (cf. [27]) then generalizes Theorem B.

Theorem 3. If $A \in \mathbb{C}^{n,n}$, $n \geq 2$, is a nonsingular H-matrix, then (cf. (4.4))

$$(4.8) \quad \sigma_n(A) \geq \{f_A(\underline{u}) \cdot f_{A^T}(\underline{v})\}^{1/2}, \quad \text{for any } \underline{u} \in U_A, \underline{v} \in U_{A^T}.$$

References

1. J. H. Ahlberg and E. N. Nilson, "Convergence properties of the spline fit", J. SIAM 11(1963), 95-104.
2. G. Alefeld and R. S. Varga, "Zur Konvergenz des symmetrischen Relaxationsverfahrens", Numerische Mathematik (to appear).
3. N. Apostolatos and U. Kulisch, "Über die Konvergenz des Relaxationsverfahrens bei nicht-negativen und diagonaldominanten Matrizen", Computing 2(1967), 17-24.
4. Robert Beauwens, "Semi-strict diagonal dominance", SIAM J. Numer. Anal. (to appear).
5. Erich Bohl, Monotone: Lösbarkeit und Numerik bei Operatorgleichungen, Springer Tracts in Natural Philosophy, 25(1974).
6. Erich Bohl, "Über eine Zeillensummenbedingung bei L-Matrizen", Lecture Notes in Mathematics 395 (Sammelband der Tagung über Numerische Lösung nichtlinearer partieller Differential- und Integrodifferential-Gleichungen, in Oberwolfach), Springer, 1974, 247-263.
7. J. H. Bramble and B. E. Hubbard, "On a finite difference analogue of an elliptic boundary value problem which is neither diagonally dominant nor of non-negative type", J. Math. and Phys. 43(1964), 117-132.
8. L. Elsner, "Bemerkungen zum Zeillensummenkriterium", Zeit. Angew. Math. Mech. 49(1966), 211-214.
9. P. J. Erdelsky, "A general theorem on dominant-diagonal matrices", Linear Algebra Appl. 1(1968), 203-209.
10. K. Fan, "Topological proof for certain theorems on matrices with non-negative elements", Monatsh. Math. 62(1958), 219-237.

11. Miroslav Fiedler and Vlastimil Pták, "On matrices with non-positive off-diagonal elements and positive principal minors", Czech. Math. J. 12(87)(1962), 382-400.
12. D. H. Jacobsen, "Factorization of symmetric M-matrices", Linear Algebra and Appl. 9(1974), 275-278.
13. K. R. James, "Convergence of matrix iterations subject to diagonal dominance", SIAM J. Numer. Anal. 10(1973), 478-484.
14. K. R. James and W. Riha, "Convergence criteria for successive overrelaxation", SIAM J. Numer. Anal. 12(1974), 137-143.
15. U. Küllisch, "Über reguläre Zerlegungen von Matrizen und einige Anwendungen", Numer. Math. 11(1968), 444-449.
16. J. A. Meijerink and N. A. Van der Vorst, "Iterative solution of linear systems arising from discrete approximation to partial differential equations", J. Computational Physics (to appear).
17. Jorge J. More', "Nonlinear generalizations of matrix diagonal dominance with application to Gauss-Seidel iterations", SIAM J. Numer. Anal. 9(1972), 357-378.
18. Karl Hans Müller, "Zum schwachen Zeilensummenkriterium bei nichtlinearen Gleichungssystemen", Computing 7(1971), 153-171.
19. A. M. Ostrowski, "Über die Determinanten mit überwiegender Hauptdiagonale", Comment. Math. Helv. 10(1937), 69-96.
20. Werner C. Rheinboldt, "On M-functions and their application to nonlinear Gauss-Seidel iterations and to network flows", J. Math. Anal. Appl. 32(1970), 274-307.
21. F. W. Schäfke, "Zum Zeilensummenkriterium", Numer. Math. 12(1968), 448-453.