MATHEMATICS OF COMPUTATION, VOLUME 31, NUMBER 137
JANUARY 1977, PAGES 148–162

An Iterative Solution Method for Linear Systems of Which the Coefficient Matrix is a Symmetric M-Matrix

By J. A. Meijerink and H. A. van der Vorst

Abstract. A particular class of regular splittings of not necessarily symmetric M-matrices is proposed. If the matrix is symmetric, this splitting is combined with the conjugate-gradient method to provide a fast iterative solution algorithm. Comparisons have been made with other well-known methods. In all test problems the new combination was faster than the other methods.

1. Introduction. A time-consuming part of the numerical solution of partial differential equations using discretization methods is often the calculation of the solution of large sets of linear equations:

$$Ax = b, \tag{1.1}$$

where A is usually a sparse matrix.

In this paper, iterative solution methods will be presented which are restricted to equations where A is a symmetric M-matrix,* although symmetry is not required in most of the theorems. This type of matrix is often generated, e.g., by discretization of elliptic and parabolic differential equations. For an extensive study on this subject, see [7].

Most of the iterative methods are based on the following idea: If K is an arbitrary nonsingular matrix, then $A = K - R$ represents a splitting of the matrix A and associated with this splitting is an iterative method

$$Kx_{n+1} = (K - A)x_n + b = Rx_n + b \tag{1.2}$$

or

$$x_{n+1} = x_n + K^{-1}(b - Ax_n) = x_n + \Delta x_n. \tag{1.3}$$

The more K resembles A, the faster the method will converge. On the other hand, we have to solve the equation

$$K\Delta x_n = b - Ax_n, \tag{1.4}$$

during every iteration so K has to be such that only few calculations and not too much memory storage are required to achieve this. For instance, the choice of K to be the diagonal matrix equal to the diagonal of A leads to the Jacobi iterative method, while the Gauss-Seidel iterative method arises by choosing K to be the lower triangular part of A. For both these choices the solution of (1.4) is straightforward.

Received May 5, 1975; revised January 16, 1976.
AMS (MOS) subject classifications (1970). Primary 65F10, 65N20.

* A matrix $A = (a_{ij})$ is an M-matrix if $a_{ij} \leqslant 0$ for $i \neq j$, A is nonsingular and $A^{-1} \geqslant 0$.

For other choices of K, the direct solution of (1.4) is equivalent to the LU-decomposition of K and the solution of the equations

$$Ly_n = b - Ax_n \tag{1.5}$$

and

$$U\Delta x_n = y_n. \tag{1.6}$$

The choice of K most ideal for the iteration process is A, since only one iteration is needed, but the LU-decomposition of A requires a large number of calculations and much memory storage, since L and U are usually considerably less sparse than A. This suggests we look for matrices $K = LU$ which resemble A, with L and U almost as sparse as A.

In [6], Stone presents a method that is based on this idea.

In Section 2 we shall introduce another class of such matrices K. We shall call this class "Incomplete LU-decompositions of A". It will be proven that this class is not empty and that the splitting $A = K - R$ is a regular splitting** which implies that the iterative method (1.2) will converge.

In Section 3 we shall discuss the stability of incomplete LU-decompositions.

In Section 4 a successful combination with the conjugate-gradient method will be described for symmetric matrices.

In Section 5 two special types of incomplete decompositions are proposed, while in Section 6 results are presented, discussed and compared with results of other familiar iterative methods.

2. Incomplete LU-Decompositions.

Notation. A lower triangular $n \times n$ matrix is denoted by $L = (l_{ij})$, so $l_{ij} = 0$ if $i < j$, and an upper triangular $n \times n$ matrix by $U = (u_{ij})$.

As mentioned in the introduction, a matrix K approximating A has to be constructed such that the L and U belonging to K are sparse. This can be realized by making an LU-decomposition of A, during which elements are neglected in the L and U matrices in appropriate places. That is the reason that we shall call $K = LU$ an "incomplete LU-decomposition of A".

Theorem 2.3 guarantees the existence of incomplete LU-decompositions. In these L and U, zeros may occur in arbitrary off-diagonal places, which can be chosen in advance. These places (i, j) will be given by the set

$$P \subset P_n \equiv \{(i, j) \mid i \neq j, 1 \leqslant i \leqslant n, 1 \leqslant j \leqslant n\}.$$

Note that P_n contains all pairs of indices of off-diagonal matrix entries. The various algorithms arise by choosing these places. Some choices for special matrices will be described in more detail in Section 5.

In the proof of Theorem 2.3 the incomplete LU-decomposition is obtained via Gauss elimination. The proof requires two theorems about operations on M-matrices.

**For $n \times n$ real matrices A, K and R, $A = K - R$ is a regular splitting of the matrix A if K is nonsingular, $K^{-1} \geqslant 0$ and $R \geqslant 0$.

The first theorem shows that the matrix that arises from an M-matrix after one elimination step is again an M-matrix.

THEOREM 2.1 (KY FAN [2, p. 44]). *If $A = (a_{ij})$ is an M-matrix, then $A^1 = (a_{ij}^1)$ is so, where A^1 is the matrix that arises by eliminating the first column of A using the first row.*

The second theorem will be used to be able to omit appropriate nondiagonal elements during the construction of the incomplete LU-decomposition of A.

THEOREM 2.2. *Let $A = (a_{ij})$ be an $n \times n$ M-matrix and let the elements of $B = (b_{ij})$ satisfy the relations*

$$a_{ij} \leqslant b_{ij} \leqslant 0 \quad \text{for } i \neq j$$

and $0 < a_{ii} \leqslant b_{ii}$. Then B is also an M-matrix.

Proof. The proof is essentially the same as a proof given by Varga [7, Proof of Theorem 3.12].

Let D_A be a diagonal matrix whose diagonal entries are given by $d_{ii} = 1/a_{ii}$, and let D_B be defined in the same way. Let Q_A and Q_B be defined by

$$Q_A = I - D_A A \quad \text{and} \quad Q_B = I - D_B B.$$

Since A is an M-matrix, the spectral radius $\rho(Q_A)$ of Q_A satisfies

$$\rho(Q_A) < 1 \qquad \text{(see [7, Theorem 3.10])},$$

and as from the assumptions it follows that $0 \leqslant Q_B \leqslant Q_A$, we have

$$\rho(Q_B) \leqslant \rho(Q_A) < 1 \qquad [7, \text{Theorem } 2.8].$$

From [7, Theorem 3.10] it follows that B is an M-matrix. □

THEOREM 2.3. *If $A = (a_{ij})$ is an $n \times n$ M-matrix, then there exists for every $P \subset P_n$ a lower triangular matrix $L = (l_{ij})$, with unit diagonal ($l_{ii} = 1$), an upper triangular matrix $U = (u_{ij})$ and a matrix $R = (r_{ij})$ with*

$$\begin{aligned} l_{ij} &= 0 \quad \text{if } (i, j) \in P, \\ u_{ij} &= 0 \quad \text{if } (i, j) \in P, \\ r_{ij} &= 0 \quad \text{if } (i, j) \notin P, \end{aligned}$$

such that the splitting $A = LU - R$ is regular. The factors L and U are unique.

Proof. The proof of this theorem also gives a way to construct L and U. The construction process consists of $n - 1$ stages. The kth stage consists of subtracting from the current coefficient matrix the elements with indices (k, j) and $(i, k) \in P$ and then reducing the matrix in the usual way. So let us define the matrices

$$A^k = (a_{ij}^k), \qquad \tilde{A}^k = (\tilde{a}_{ij}^k), \qquad L^k = (l_{ij}^k) \text{ and } R^k = (r_{ij}^k)$$

by the relations:

$$\left.\begin{aligned} A^0 &= A \\ \tilde{A}^k &= A^{k-1} + R^k \\ A^k &= L^k \tilde{A}^k \end{aligned}\right\} \quad \text{for } k = 1, 2, \ldots, n-1.$$

Here the matrix R^k is defined by

$$\begin{aligned} r_{kj}^k &= -a_{kj}^{k-1}, \quad \text{if } (k, j) \in P, \\ r_{ik}^k &= -a_{ik}^{k-1}, \quad \text{if } (i, k) \in P \text{ and all other } r_{ij}^k \text{ are equal to zero.} \end{aligned}$$

L^k is equal to the unit matrix, except for the kth column, which written row-wise, is as follows

$$\left[0, 0, 0, \ldots, 1, -\frac{\tilde{a}^k_{k+1\,k}}{\tilde{a}^k_{kk}}, -\frac{\tilde{a}^k_{k+2\,k}}{\tilde{a}^k_{kk}}, \ldots, -\frac{\tilde{a}^k_{nk}}{\tilde{a}^k_{kk}}\right].$$

From this it can easily be seen that A^k is the matrix that arises from $\tilde{A}^k$ by eliminating the lowermost $n-k$ elements in the kth column using the kth row.

$A^0 = A$ is an M-matrix, so $R^1 \geqslant 0$. From Theorem 2.2 it follows that $\tilde{A}^1$ is an M-matrix. Therefore $L^1 \geqslant 0$ and applying Theorem 2.1 we see that A^1 is an M-matrix. Continuing in this manner, we can prove that

$$\left.\begin{array}{l} A^k \text{ is an } M\text{-matrix} \\ \tilde{A}^k \text{ is an } M\text{-matrix} \\ L^k \geqslant 0 \\ R^k \geqslant 0 \end{array}\right\} \quad \text{for } k = 1, \ldots, n-1.$$

From the definitions it follows immediately that

$$\begin{aligned} L^k R^m &= R^m \quad \text{if } k < m, \\ A^{n-1} &= L^{n-1}\tilde{A}^{n-1} = L^{n-1}A^{n-2} + L^{n-1}R^{n-1} \\ &= L^{n-1}L^{n-2}\tilde{A}^{n-2} + L^{n-1}R^{n-1} = \cdots = L^{n-1}L^{n-2}\cdots L^1A^0 \\ &\quad + L^{n-1}L^{n-2}\cdots L^1R^1 + L^{n-1}L^{n-2}\cdots L^2R^2 + \cdots + L^{n-1}R^{n-1}. \end{aligned}$$

By combining these equations, we find

$$A^{n-1} = L^{n-1}L^{n-2}\cdots L^1(A + R^1 + R^2 + \cdots + R^{n-1}).$$

Let us now define $U = A^{n-1}$, $L = (L^{n-1}L^{n-2}\cdots L^1)^{-1}$ and $R = R^1 + R^2 + \cdots + R^{n-1}$ then $LU = A + R$, $(LU)^{-1} \geqslant 0$ and $R \geqslant 0$, so the splitting $A = LU - R$ is regular. The uniqueness of the factors L and U follows from equating the elements of A and LU for $(i, j) \notin P$, and from the fact that L has a unit diagonal. □

For the case where A is in addition symmetric and thus positive definite Theorem 2.4 gives a symmetric variant of the preceding theorem. This states that a symmetric incomplete LU-decomposition can be achieved which contains zeros in a symmetric pattern of places indicated in advance.

THEOREM 2.4. *If A is a symmetric M-matrix, there exists for each $P \subset P_n$ having the property that $(i, j) \in P$ implies $(j, i) \in P$, a uniquely defined lower triangular matrix L and a symmetric nonnegative matrix R, with $l_{ij} = 0$ if $(i, j) \in P$ and $r_{ij} = 0$ if $(i, j) \notin P$, such that the splitting $A = LL^T - R$ is a regular splitting.*

Proof. This theorem follows directly from the fact that Choleski decomposition is equivalent to gaussian elimination except for a diagonal matrix. This extra diagonal matrix does not affect places which contain zeros. □

From the previous theorems the convergence of the method defined in (1.2) – (1.4) follows immediately, this is formulated in Theorem 2.5.

THEOREM 2.5. *If A, L, U and R are defined as in Theorem* 2.3, *the iterative method*

$$LUx_{i+1} = Rx_i + b, \qquad i \geqslant 0,$$

will converge to the solution of $Ax = b$ for every choice of x_0.

Proof. This is an immediate consequence of Theorem 3.13 of Varga [7]. □

By properly choosing $P \subset P_n$, we obtain a number of well-known methods: $P = P_n$ results in the point Jacobi method, and $P = \{(i, j) \mid i < j\}$ results in the point Gauss-Seidel method. Also, line and block variants of these two methods can be obtained by a proper choice of P. So Jacobi and Gauss-Seidel methods are a subclass of methods based on incomplete LU-decompositions, which are themselves a subclass of methods based on regular splittings.

3. Numerical Stability. The question which now arises is whether the construction of an incomplete LU-decomposition is stable. In order to answer this question, we need Theorem 3.1. This theorem indicates the effect on the decomposition process of replacing off-diagonal elements in the matrix by nonpositive elements that are smaller in absolute value, as well as the effect of replacing diagonal elements by larger ones.

THEOREM 3.1. *Let $A = (a_{ij})$ and $B = (b_{ij})$ be defined as in Theorem 2.2. Let A^1 and B^1 be the matrices that arise from A and B by eliminating the first column using the first row.*

Then,

$$a_{ij}^1 \leqslant b_{ij}^1 \leqslant 0, \qquad 0 < a_{ii}^1 \leqslant b_{ii}^1$$

and B^1 is an M-matrix.

Proof.

$$a_{ij}^1 = a_{ij} - \frac{a_{i1}}{a_{11}} a_{1j} \leqslant b_{ij} - \frac{b_{i1}}{b_{11}} b_{1j} = b_{ij}^1 \quad \text{for } i \neq 1, j \neq 1.$$

From these relations it follows that $b_{ij}^1 \leqslant 0$, for $i \neq j$. Now A^1 is an M-matrix (Theorem 2.1), and $a_{ii}^1 > 0$ is a property of M-matrices [7, Theorem 3.10], so from Theorem 2.2 B^1 is also an M-matrix. □

The following theorem states that the incomplete LU-decomposition process is more stable than the complete LU-decomposition process (without partial pivoting).

THEOREM 3.2. *If A is an M-matrix, then the construction of an incomplete LU-decomposition is at least as stable as the construction of a complete decomposition $A = LU$ without any pivoting.*

Proof. Let $\tilde{A}$ be the matrix that is obtained by setting some off-diagonal elements of A to zero in the first column and in the first row (compare Theorem 2.3). Let $\tilde{L}_1$ be the gaussian elimination matrix for the first elimination step on $\tilde{A}$, and L_1 be the same for A.

Then it is obvious that the elements of $\tilde{L}_1$ are not larger in absolute value than the elements of L_1. From Theorem 2.1 and Theorem 2.2 it follows that $\tilde{A}_1$ and A_1 are M-matrices, while Theorem 3.1 states that $\tilde{A}^1 \geqslant A^1$. From repeated application of Theorem 3.1 it follows that the elements of the gaussian elimination matrices $\tilde{L}_k$, in each stage of the incomplete LU-decomposition process, are not larger in absolute value than the elements of the gaussian elimination matrices L_k that arise in the complete decomposition process. This gives the desired result (see [8], [9]). □

COROLLARY 1. *If A is a symmetric M-matrix, then the construction of an incomplete LL^T-decomposition is at least as stable as Choleski's process.*

Note. It is well known that in general LU-decomposition without pivoting is not a very satisfactory process. Therefore, we consider the practical situation, where A is a diagonally dominant M-matrix. It is easy to see that gaussian elimination preserves the diagonal dominance of the matrix. Therefore gaussian elimination, in this case, is identical with Crout LU-decomposition with partial pivoting, the latter being fairly stable.

4. An Accelerated Method for Symmetric Systems of Equations. If the $n \times n$ matrix A of the linear system of equations $Ax = b$ is an M-matrix, and LU is an incomplete decomposition of A, the iterative process, defined in Theorem 2.5, generates a sequence $\{x_n\}_{n>0}$ that converges to x. From simple analysis it follows that

$$x_i = x_0 - \binom{i}{1}[(LU)^{-1}A]\,(x_0 - x) + \binom{i}{2}[(LU)^{-1}A]^2(x_0 - x) - \cdots + (-1)^i\binom{i}{i}[(LU)^{-1}A]^i(x_0 - x). \tag{4.1}$$

If A is a symmetric M-matrix (hence, positive definite), an upperbound for the error $\|x_i - x\|_A^2 \equiv (A(x_i - x), x_i - x)$ is given by

$$\|x_i - x\|_A^2 \leqslant [\max\{|1 - \lambda_{\min}|, |1 - \lambda_{\max}|\}]^{2i}\|x_0 - x\|_A^2, \tag{4.2}$$

with

$$(x, y) \equiv \sum_{j=1}^{n} x_j y_j,$$

$\lambda_{\min}$ is the smallest eigenvalue of $(LU)^{-1}A$, and

$\lambda_{\max}$ the largest eigenvalue.

For this special case of A a symmetric M-matrix, symmetric incomplete LL^T-decomposition can be combined with the method of conjugate gradients. This leads to a similar scheme as (4.1), which is known to be faster. For the discussion of this combined method and its main properties, results from [1] and [3] are used. Let M be a square nonsingular $n \times n$ matrix and let H and K be positive definite symmetric $n \times n$ matrices, $N \equiv M^*HM$ and $T \equiv KN$, then a conjugate-gradient method to solve the equation $Mx = b$ is defined by

x_0 an arbitrary initial approximation to x,

$$r_0 = b - Mx_0, \quad g_0 = M^*Hr_0, \quad p_0 = Kg_0,$$

$$\left.\begin{aligned} \alpha_i &= (g_i, p_i)/(p_i, Np_i) = (g_i, Kg_i)/(p_i, Np_i) \\ x_{i+1} &= x_i + \alpha_i p_i \\ r_{i+1} &= K - Mx_{i+1} = r_i - \alpha_i Mp_i \\ g_{i+1} &= M^*Hr_{i+1} = g_i - \alpha_i Np_i \\ \beta_i &= -(Np_i, Kg_{i+1})/(p_i, Np_i) = (g_{i+1}, Kg_{i+1})/(g_i, Kg_i) \\ p_{i+1} &= Kg_{i+1} + \beta_i p_i \end{aligned}\right\}, \quad i = 1, 2, \ldots. \tag{4.3}$$

This method has the following theoretical properties:

(1) the sequence $\{x_i\}_{i\geq 0}$ converges to the solution x within n iterations.

(2) the conjugate-gradient method minimizes $\|x_i - x\|_N$ for all i, among all algorithms of the form

$$x_i = x_0 + P_{i-1}(T)T(x - x_0), \tag{4.4}$$

where P_{i-1} is a polynomial of degree $i-1$.

(3)

$$\|x_i - x\|_N^2 \leq \left(\frac{\sqrt{c}-1}{\sqrt{c}+1}\right)^{2i} \|x_0 - x\|_N^2, \tag{4.5}$$

where $c = \lambda_{\max}(T)/\lambda_{\min}(T)$.

From the choice $M = A$, $H = A^{-1}$ and $K = (LL^T)^{-1}$ which results in $N = A$ and $T = (LL^T)^{-1}A$, it follows that the iterative method defined in Theorem 2.5 is of the form (4.4) and hence from property (2) it follows that the combined method will converge at least as fast. Also the two upper bounds for the errors show a substantial difference. For this choice the iteration scheme can be written as:

x_0 is an arbitrary initial approximation to x,

$r_0 = b - Ax_0, \quad p_0 = (LL^T)^{-1}r_0,$

$$\left.\begin{aligned} \alpha_i &= \frac{(r_i, [LL^T]^{-1}r_i)}{(p_i, Ap_i)} \\ x_{i+1} &= x_i + \alpha_i p_i \\ r_{i+1} &= r_i - \alpha_i A p_i \\ \beta_i &= \frac{(r_{i+1}, [LL^T]^{-1}r_{i+1})}{(r_i, [LL^T]^{-1}r_i)} \\ p_{i+1} &= [LL^T]^{-1}r_{i+1} + \beta_i p_i \end{aligned}\right\}, \quad i = 0, 1, 2, \ldots . \tag{4.6}$$

Remark. The inequality (4.5) does not take advantage of the distribution of the eigenvalues of T, while the conjugate-gradient method does so. Therefore the upper-bound (4.5) might be pessimistic. This happens especially when most of the eigenvalues of T are clustered in small intervals compared to the interval $[\lambda_{\min}(T), \lambda_{\max}(T)]$.

5. Two Applications of Incomplete Decomposition. For a special type of matrix, two different incomplete decompositions will be introduced in this section. The matrix-equation arises from five-point discrete approximations to the second-order selfadjoint elliptic partial differential equation:

$$-\frac{\partial}{\partial x}A(x, y)\frac{\partial}{\partial x}u(x, y) - \frac{\partial}{\partial y}B(x, y)\frac{\partial}{\partial y}u(x, y) + C(x, y)u(x, y) = D(x, y) \tag{5.1}$$

with $A(x, y), B(x, y) > 0$, $C(x, y) \geq 0$, and $(x, y) \in R$, where R is a square region, and with suitable boundary conditions on R. The resulting symmetric positive definite diagonally dominant nth order matrix $A = (a_{ij})$ is schematically shown in Figure 1. Places of zero entries are given by

$$P^* = \{(i, j) \mid |i - j| \neq 0, 1, m\}, \tag{5.2}$$

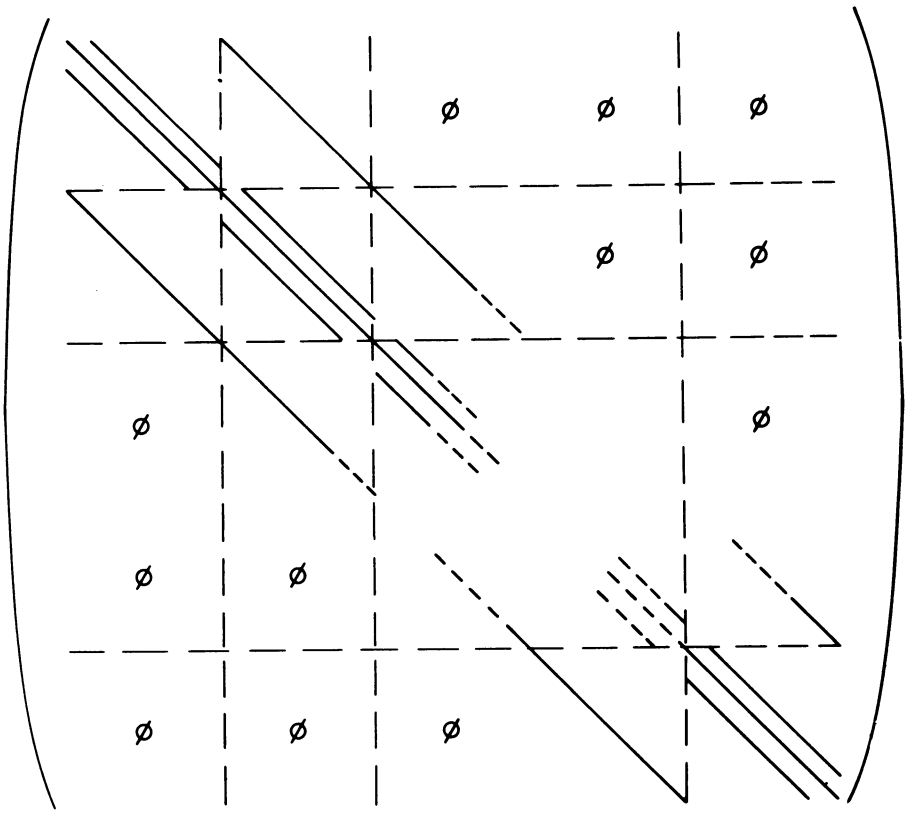


FIGURE 1

where m is the half bandwidth of the matrix. For the derivation of such linear systems see references [6] and [7].

The elements of the diagonal of A are denoted by a_i, the upper-diagonal elements are denoted by b_i and the elements of the mth upper diagonal are denoted by c_i, where i is the index of the row of A in which the respective elements occur. Theorem 2.4 guarantees the existence of incomplete symmetric decompositions for A. Our first application considers the incomplete decomposition that arises in the decomposition process when all elements are ignored in those places where A has zero entries. This variant is characterized by P^*.

In the following it will be convenient to write the incomplete decomposition in the form: LDL^T, where D is a diagonal matrix. If the elements of D are denoted by $\tilde{d}_i$ and the elements of L^T are denoted analogous to A by $\tilde{a}_i$, $\tilde{b}_i$ and $\tilde{c}_i$, then the following recurrent relations hold for these elements:

$$\begin{aligned} &\tilde{b}_i = b_i, \quad \tilde{c}_i = c_i, \\ &\tilde{a}_i = \tilde{d}_i^{-1} = a_i - \tilde{b}_{i-1}^2 \tilde{d}_{i-1} - \tilde{c}_{i-m}^2 \tilde{d}_{i-m}, \end{aligned} \qquad i = 1, 2, \ldots, n, \tag{5.3}$$

where elements that are not defined should be replaced by zeros.

Note that, once we have stored A, we need only compute and store the diagonal matrix D. Note also that when this process is programmed for a two-level store computer, for very large matrices, the diagonal D can be constructed by taking successive parts of A into fast core. Equation (1.4) reduces to

$$LDL^T\Delta x_i = b - Ax_i,$$

and it is well known that Δx_i can be determined by solving successively

$$Ly_i = b - Ax_i, \quad Dz_i = y_i$$

and

$$L^T\Delta x_i = z_i.$$

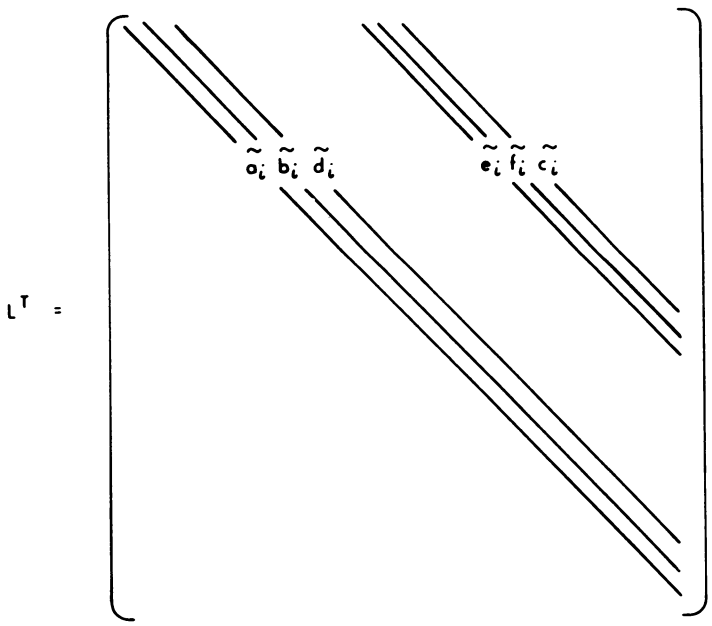
This first variant of incomplete decompositions for the type of matrices described here, in combination with the conjugate gradient method, is referred to by ICCG(0) further on (i.e., Incomplete Choleski & Conjugate Gradients, with 0 extra diagonals).

The second variant is characterized by

$$P^3 \equiv \{(i, j) \mid |i - j| \neq 0, 1, 2, m - 2, m - 1, m\}.$$

In this case, the symmetric decomposition is written as LL^T.

The elements of L^T are denoted as follows, where i is counted row-wise again:



With the above notations, the elements of L^T can be recursively computed as follows:

$$\begin{aligned}
&\tilde{a}_i = \sqrt{a_i - \tilde{b}_{i-1}^2 - \tilde{d}_{i-2}^2 - \tilde{e}_{i-m+2}^2 - \tilde{f}_{i-m+1}^2 - \tilde{c}_{i-m}^2},\\
&\tilde{b}_i = (b_i - \tilde{d}_{i-1}\tilde{b}_{i-1} - \tilde{c}_{i-m+1}\tilde{f}_{i-m+1} - \tilde{f}_{i-m+2}\tilde{e}_{i-m+2})/\tilde{a}_i,\\
&\tilde{d}_i = -\tilde{c}_{i-m+2}\tilde{e}_{i-m+2}/\tilde{a}_i, \quad \tilde{e}_i = -(\tilde{c}_{i-2}\tilde{d}_{i-2} + \tilde{f}_{i-1}\tilde{b}_{i-1})/\tilde{a}_i,\\
&\tilde{f}_i = -\tilde{c}_{i-1}\tilde{b}_{i-1}/\tilde{a}_i, \quad \tilde{c}_i = c_i/\tilde{a}_i \quad \text{for } i = 1, 2, \ldots, n.
\end{aligned} \tag{5.4}$$

Elements not defined should be replaced by zeros. It should be remarked that it is also possible to avoid the square root computations by a slight modification of the Eqs. (5.4).

This second variant, in combination with the conjugate-gradient method, will be referred to as ICCG(3), as it has three more diagonals at each side than the original matrix A.

6. Numerical Results and Comparison with Other Methods. In this section, results are presented and compared with results of other iterative methods, for two special cases of Eq. (5.1). However, it should be mentioned that similar results have been obtained in other cases.

We first discuss briefly the different iterative methods. N will denote the order of the matrix A.

ICCG(0). This variant of incomplete decomposition is discussed in Section 5. Each iteration of ICCG(0) needs $\simeq$ 16N multiplications.

ICCG(3). For a discussion see Section 5. Each ICCG(3)-iteration needs $\simeq$ 22N multiplications.

SLOR. Successive Line Over-Relaxation needs $\simeq$ 6N multiplications each iteration if intermediate results are stored.

Conjugate Gradients. See Section 4, where for this case LL^T should be replaced by the identity matrix. Each iteration needs $\simeq$ 10N multiplications. If the matrix has 'property A', this can be reduced by a factor 2 [5].

SIP. The Strongly Implicit Procedure has been described in detail by Stone [6]. Each iteration needs $\simeq$ 22N multiplications.

In interpreting the results of the various methods, it should be noted that any initial work, such as the work necessary for the estimation of iteration-parameters or the computational work for the decompositions of the ICCG methods, was neglected. This did not affect the conclusions seriously, because this initial work will in general be negligible compared to the computational work needed for even a small number of iterations. The methods are compared on the basis of computational work, which was measured, rather arbitrarily, using the total number of multiplications.

The number of multiplications needed for each iteration is mentioned above. In the figures, the number of multiplications required for one single iteration of ICCG(3), i.e. 22N multiplications, was chosen as the unit for the computational work.

Example 1. Equation (5.1) is considered over the square region $0 \leqslant x \leqslant 1$, $0 \leqslant y \leqslant 1$, with $A(x, y) = B(x, y) = 1$, $C(x, y) = D(x, y) = 0$ and the boundary conditions $\partial u/\partial x = 0$ for $x = 0$ and $x = 1$, $\partial u/\partial y = 0$ for $y = 1$ and $u = 1$ for $y = 0$. A uniform rectangular mesh was chosen, with $\Delta x = 1/31$ and $\Delta y = 1/31$, which resulted in a linear system of 992 equations. The solution of (5.1) is known to be $u(x, y) = 1$, and as initial starting vector for the iterative scheme described in Section 4, a vector was chosen with all entries random between 0 and 2. This was done to prevent fast convergence by coincident. The results are plotted in Figure 2.

Example 2. In order to illustrate the power of the ICCG methods, also for more

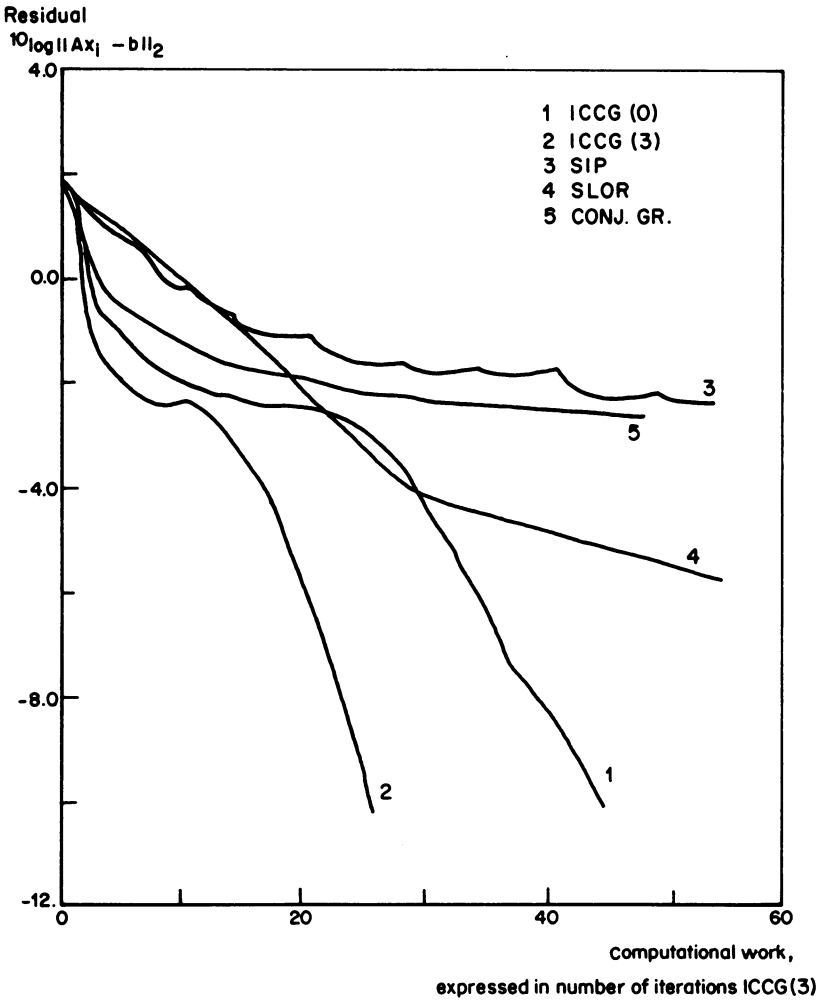
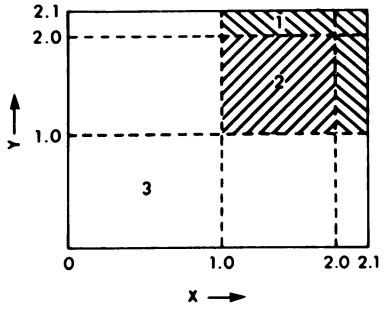


FIGURE 2. *Results for Example* 1

practical nonuniform situations, a problem suggested by Varga [7, Appendix B] was considered. Equation (5.1) holds on R, where R is the square region $0 \leqslant x, y \leqslant 2.1$, as shown below.



On the boundary of R the boundary conditions are $\partial u/\partial n = 0$. Further $D(x, y) = 0$ over R and the functions A, B and C are defined by

Region	$A(x, y)$	$B(x, y)$	$C(x, y)$
1	1.0	1.0	0.02
2	2.0	2.0	0.03
3	3.0	3.0	0.05

A uniform rectangular mesh was chosen with mesh spacing 0.05, so a system of 1849 equations in 1849 unknowns resulted. The solution of this problem is known to be $u(x, y) = 0$; as starting vector for all iterative methods, a vector was chosen similar to the one in Example 1. The iteration results are plotted in Figure 3.

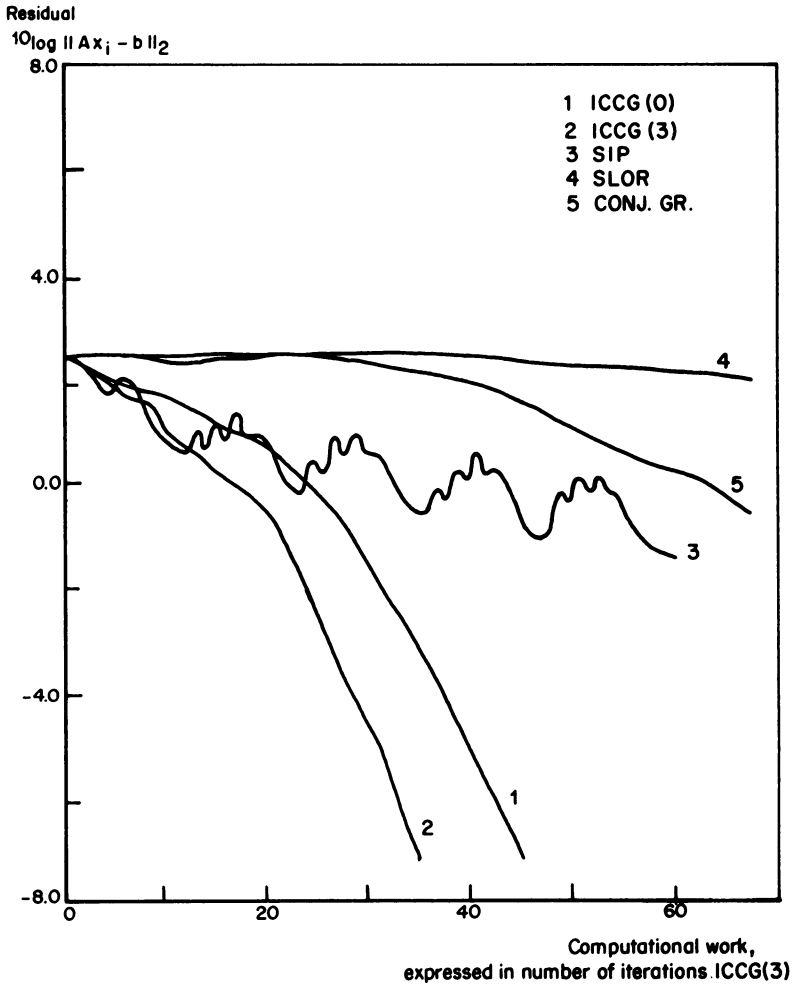
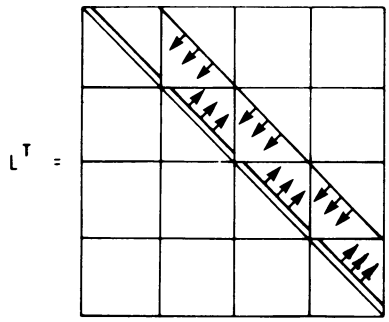


FIGURE 3. *Results for Example* 2

These few examples give some impression of the kind of convergence that is typical for the ICCG methods. In order to explain this phenomenon, a complete Choleski-decomposition of the type of matrix, introduced in Section 5, is considered. It is then observed that the nonzero entries in the full decomposition decrease rapidly in magnitude in the directions pointed out below.



As is known, Choleski-decomposition is a stable process, therefore it might be expected that setting some of the smaller elements to zero, results in an incomplete Choleski-decomposition, which will be like the full decomposition. Thus, the matrix $(LL^T)^{-1}A$, where LL^T is an incomplete decomposition, should resemble the identity matrix in some way, or more precisely, $(LL^T)^{-1}A$ will have all eigenvalues close to 1.0. The fact that conjugate gradients gives fast convergence for matrices with the latter property explains to some extent the fast convergence of the ICCG methods.

In order to give an impression of the eigenvalues of $(LL^T)^{-1}A$ for both the ICCG(0) and ICCG(3) methods, a smaller version of the matrix arising in Example 1 has been chosen. In fact, the choice $\Delta x = 1/5$ and $\Delta y = 1/6$, resulted in a matrix of order 36. In Figure 4 all the eigenvalues of A, $(L_0L_0^T)^{-1}A$ and $(L_3L_3^T)^{-1}A$ are plotted. The lower index indicates which ICCG method is considered.

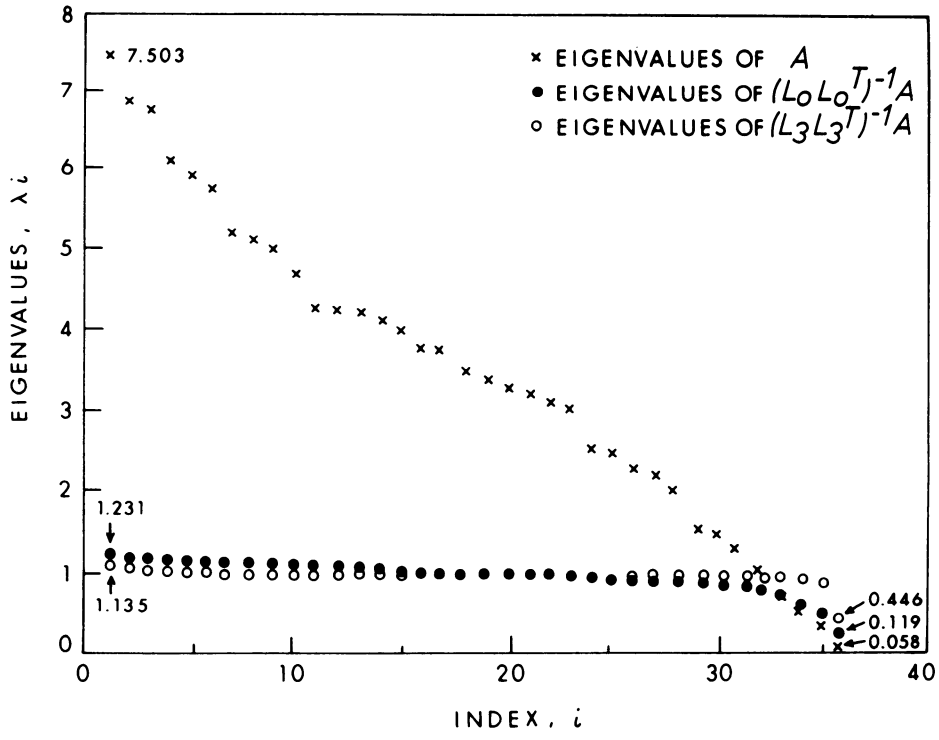


FIGURE 4. *Eigenvalues of* A, $(L_1L_1^T)^{-1}A$ *and* $(L_3L_3^T)^{-1}A$

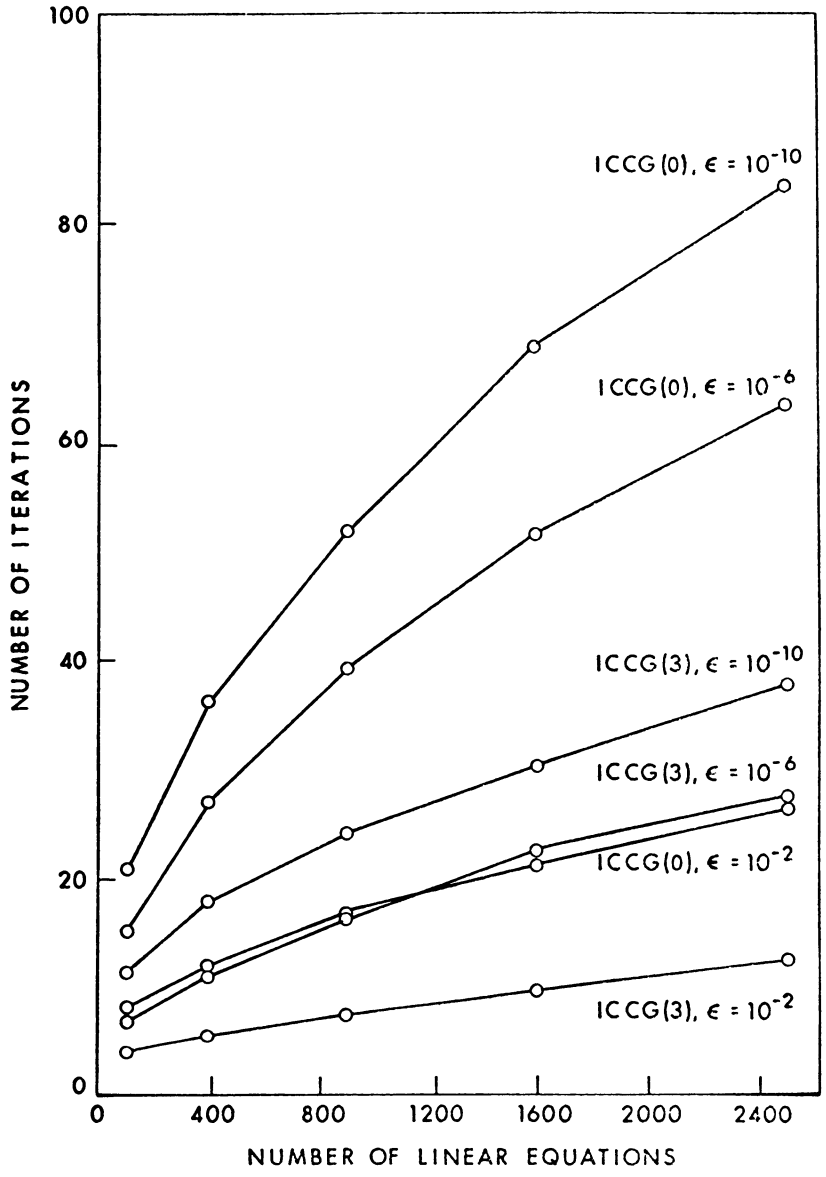


FIGURE 5. *Effect of number of equations on the rate of convergence*

It follows from formula (4.5) that the error $\|x_i - x\|_A$ is multiplied at each step by at most $r = (\sqrt{c} - 1)/(\sqrt{c} + 1)$. This helps explain the fast convergence; for A, $(L_0 L_0^T)^{-1}A$, and $(L_3 L_3^T)^{-1}A$, respectively, we find $r = .84$, $r_0 = .53$ and $r_3 = .23$.

Finally, for the linear equations arising in Example 1, the influence of the order of the matrix on the number of iterations required to reach a certain precision was checked for both ICCG(0) and ICCG(3).

Therefore several uniform rectangular meshes have been chosen, with mesh spacings varying from $\sim 1/10$ up to $\sim 1/50$. This resulted in linear systems with matrices of order 100 up to about 2500. In each case it was determined how many iterations were necessary, in order that the magnitude of each entry of the residual vector was below some fixed small number ϵ, when starting with $x_0 = 0$.

In Figure 5 the number of iterations are plotted against the order of the matrices for $\epsilon = 10^{-2}$, $\epsilon = 10^{-6}$ and $\epsilon = 10^{-10}$. It can be seen that the number of iterations, necessary to get the residual vector sufficiently small, increases only slowly for increasing order of the matrix.

7. **Conclusions.** In the examples, both ICCG methods appeared to be far superior to all the other iterative methods mentioned, except possibly CG when the matrix has 'property A' [5].

If the solution of the linear system is calculated by complete Choleski, the total number of multiplications is given approximately by $n(m + 1)(m + 2)/2 + 2n(m + 1)$ [8], where n is the order of the matrix and $2m$ the bandwidth. For $n = 900$ this amount of work is equivalent to about 25 ICCG(3) iterations (at this time storage aspects are not considered).

This implies that both ICCG methods can compete with direct solution with regard to computational work, if we are satisfied with not too high an accuracy. From Figure 5 it can be seen that for larger matrices the ICCG methods are to be preferred even more.

These statements also hold if the direct method takes advantage of the very sparse structure of the matrices. In this case Price and Coats [4] showed that the total number of multiplications for the direct method can be reduced by a factor 6, compared to the number mentioned above.

Finally, we would like to observe that the ICCG methods have also been applied very successfully in practice, in solving both two- and three-dimensional problems.

Koninklijke/Shell
Exploratie & Produktie Laboratorium
Rijswijk, The Netherlands

Academic Computer Centre Utrecht
Utrecht, The Netherlands

1. J. W. DANIEL, "The conjugate gradient method for linear and nonlinear operator equations," *SIAM J. Numer. Anal.*, v. 4, 1967, pp. 10–26. MR 36 #1076.

2. KY FAN, "Note on *M*-matrices," *Quart. J. Math. Oxford Ser.* (2), v. 11, 1960, pp. 43–49. MR 22 #8024.

3. M. R. HESTENES, *The Conjugate-Gradient Method for Solving Linear Systems,* Proc. Sympos. Appl. Math., vol. VI, Numerical Analysis, McGraw-Hill, New York, 1956, pp. 83–102. MR 18, 824.

4. H. S. PRICE & K. H. COATS, "Direct methods in reservoir simulation," *Soc. Petroleum Engrs. J.*, v. 14, 1974, pp. 295–308.

5. J. K. REID, "The use of conjugate gradients for systems of linear equations possessing 'Property A'," *SIAM J. Numer. Anal.*, v. 9, 1972, pp. 325–332. MR 46 #4697.

6. H. L. STONE, "Iterative solution of implicit approximations of multidimensional partial differential equations," *SIAM J. Numer. Anal.*, v. 5, 1968, pp. 530–558. MR 38 #6780.

7. R. S. VARGA, *Matrix Iterative Analysis,* Prentice-Hall, Englewood Cliffs, N.J., 1962. MR 28 #1725.

8. J. H. WILKINSON & C. REINSCH, *Linear Algebra,* Springer-Verlag, Berlin and New York, 1971.

9. J. H. WILKINSON, *The Algebraic Eigenvalue Problem,* Clarendon Press, Oxford, 1965. MR 32 #1894.