MATHEMATICS OF COMPUTATION, VOLUME 34, NUMBER 150
APRIL 1980, PAGES 473–497

An Incomplete Factorization Technique for Positive Definite Linear Systems

By T. A. Manteuffel

Abstract. This paper describes a technique for solving the large sparse symmetric linear systems that arise from the application of finite element methods. The technique combines an incomplete factorization method called the shifted incomplete Cholesky factorization with the method of generalized conjugate gradients. The shifted incomplete Cholesky factorization produces a splitting of the matrix A that is dependent upon a parameter α. It is shown that if A is positive definite, then there is some α for which this splitting is possible and that this splitting is at least as good as the Jacobi splitting. The method is shown to be more efficient on a set of test problems than either direct methods or explicit iteration schemes.

1. Introduction. High speed computers have made possible truly three-dimensional models. The increased complexity of such models is transmitted to the linear systems that often form an integral component of the model. The linear systems that result from an application of finite element methods can be especially complex. This work was motivated by a model of structural deformation in three dimensions that produced linear systems with 18,000 unknowns, a half-bandwidth of 1,300, and an average of 115 nonzeros per equation. The direct solution of such a system would require approximately 15,000,000,000 arithmetic operations and 23,000,000 words of storage. For that reason iterative methods were pursued; see Section 6.

Characteristically, finite element methods yield positive definite linear systems that are sparse, but much more dense than their finite difference counterpart. The zero structure of finite element matrices reflects the complexity of the domain of the model and irregularity of the mesh. There is little pattern. These matrices do not possess Young's property-A, (Young [27, p. 41]), nor are they M-matrices,* nor are they diagonally dominant. We can, however, extend some of the results for these special cases.

The iterative technique to be described is a variant of the incomplete Cholesky factorization–generalized conjugate gradient method (ICCG) described by Meijerink and van der Vorst [17]. In their paper it is shown that if A is an M-matrix it can always be split into

$$A = M - N, \tag{1.1}$$

Received July 14, 1978; revised March 26, 1979.
AMS (MOS) subject classifications (1970). Primary 65F10.
*A matrix $A = (a_{ij})$ is an M-matrix if $a_{ij} \leqslant 0$ for $i \neq j$, A is nonsingular, and $A^{-1} \geqslant 0$.

0025-5718/80/0000-0057/$07.25

where

$$M = L\Sigma U \tag{1.2}$$

and L and U are sparse unit lower and upper triangular matrices and Σ a positive diagonal matrix. The splitting (1.1) is a regular splitting. If A is also symmetric, then M will be symmetric and this splitting can be used in conjunction with the generalized conjugate gradient algorithm. If A is not an M-matrix, this splitting may not be a regular splitting, and in fact, the factorization of M may not always be positive; see Section 3. This paper will describe the implementation of a similar procedure called the shifted incomplete Cholesky factorization (SIC) on general symmetric positive definite matrices and discuss its efficiency.

Unlike the SIP splitting (Stone [24]) which is only applicable to 5-point difference matrices, the SIC splitting can be applied to any positive definite matrix regardless of structure. The SIC splitting has an advantage over the SSOR splitting (Axelsson [1]) in that a more accurate approximate inverse may be brought into play. In fact, one can move from the extreme of no splitting to the extreme of complete factorization with iterative refinement by adjusting the amount of fill-in allowed in the factorization (1.2).

Section 2 will establish notation with a brief description of the generalized conjugate gradient iteration. In Section 3 the result of Meijerink and van der Vorst will be extended to H-matrices.** It will be shown that for M-matrices allowing more fill-in improves the condition of the splitting. In Section 4 it will be shown that the procedure is invariant to diagonal scaling, and a measure of the positivity will be proposed. The shifted incomplete Cholesky factorization will be discussed in Section 5. It will be shown that a factorization is possible for any positive definite matrix and that the resulting splitting is at least as good as the Jacobi splitting. It will be shown that for a Stieltjes matrix,*** the best shift is nonpositive. Numerical results on a test problem will be discussed in Section 6.

This paper is mainly concerned with symmetric positive definite matrices. However, the results of Sections 3 and 4 will be given in terms of possibly nonsymmetric matrices. Many of the details have been omitted. They may be found, along with a description of the implementation of the procedure, in the report [16].

2. Generalized Conjugate Gradient Iteration. The conjugate gradient method was first described by Hestenes and Stiefel [9] in 1952. Because numerical properties of the conjugate gradient algorithm differed from the theoretical properties, the method saw little use as a method for solving linear systems until 1970 when it was shown by Reid [22] to be highly effective on some large sparse systems. The generalized conjugate gradient method, described by Hestenes [10] and later by Faddeev and Faddeeva [5] and Daniel [3], was shown to be effective for some large sparse systems by Concus, Golub, and O'Leary [2].

**The matrix $A = (a_{ij})$ is an H-matrix if the matrix $B = (b_{ij})$ with $b_{ii} = a_{ii}$, $b_{ij} = -|a_{ij}|$, $i \neq j$ is an M-matrix.

***A Stieltjes matrix is a symmetric M-matrix.

Given the linear system

$$Ax = b, \tag{2.1}$$

we split A into

$$A = M - N, \tag{2.2}$$

where M is easily invertible and the behavior of M in some sense approximates the behavior of A. The preconditioned system

$$M^{-1}Ax = M^{-1}b \tag{2.3}$$

may have much better condition than the original system (2.1) and, therefore, be easier to solve by iterative techniques.

If A and M are both positive definite, then $M^{1/2}$ exists and the system

$$\tilde{A}y = \tilde{b}, \tag{2.4}$$

where

$$\tilde{A} = M^{-1/2}AM^{-1/2}, \quad \tilde{b} = M^{-1/2}b, \quad y = M^{+1/2}x,$$

is positive definite and equivalent to (2.1). The generalized conjugate gradient algorithm is equivalent to applying conjugate gradients to the system (2.4).

The algorithm is as follows: given x_0, let

$$r_0 = b - Ax_0, \quad \text{(residual)},$$

$$h_0 = M^{-1}r_0, \quad \text{(generalized residual)},$$

$$p_0 = h_0.$$

The general step is

$$\begin{aligned} x_{i+1} &= x_i + \alpha_i p_i, \quad \alpha_i = \frac{\langle r_i, p_i\rangle}{\langle Ap_i, p_i\rangle} = \frac{\langle r_i, h_i\rangle}{\langle Ap_i, p_i\rangle}; \\ r_{i+1} &= r_i - \alpha_i Ap_i; \quad h_{i+1} = M^{-1}r_{i+1}; \\ p_{i+1} &= h_{i+1} + \beta_i p_i, \quad \beta_i = \frac{-\langle Ap_i, h_{i+1}\rangle}{\langle Ap_i, p_i\rangle} = \frac{\langle r_{i+1}, h_{i+1}\rangle}{\langle r_i, h_i\rangle}. \end{aligned} \tag{2.5}$$

With $M = I$, the identity, this is the same as the conjugate gradient method.

It has been shown that the convergence properties of the conjugate gradient algorithm of system (2.4) depend upon the entire spectrum of $\tilde{A}$ (Greenbaum [8]). However, upper bounds for the $\tilde{A}$-norm of the error can be found in terms of the condition of $\tilde{A}$,

$$K(\tilde{A}) = \|\tilde{A}\| \cdot \|\tilde{A}^{-1}\|,$$

where $\|\tilde{A}\|$ is the spectral norm of the operator. These bounds are useful for determining stopping criteria for the algorithm. If we let e_i be the error of system (2.1) and $\tilde{e}_i$ be the error of system (2.4) after i steps of iteration (2.5), then the relative

error of system (2.4) is bounded by

$$\frac{\|\tilde{\mathbf{e}}_i\|}{\|\mathbf{y}\|} \leqslant K(\tilde{A}) \frac{\|\tilde{\mathbf{r}}_i\|}{\|\tilde{\mathbf{b}}\|} = K(\tilde{A}) \left(\frac{\langle \mathbf{r}_i, \mathbf{h}_i \rangle}{\langle M^{-1}\mathbf{b}, \mathbf{b} \rangle} \right)^{1/2}; \tag{2.6}$$

and the relative error of system (2.1) is bounded by

$$\frac{\|\mathbf{e}_i\|}{\|\mathbf{x}\|} \leqslant (K(M))^{1/2} \frac{\|\tilde{\mathbf{e}}_i\|}{\|\mathbf{y}\|} \leqslant (K(M))^{1/2} K(\tilde{A}) \left(\frac{\langle \mathbf{r}_i, \mathbf{h}_i \rangle}{\langle M^{-1}\mathbf{b}, \mathbf{b} \rangle} \right)^{1/2}. \tag{2.7}$$

Since M is constructed it will be possible to estimate $K(M)$; see Section 4. An estimate of $K(\tilde{A})$ may also be computed by exploiting the relationship between the Lanczos algorithm and the conjugate gradient algorithm; cf. Concus, Golub, and O'Leary [2], Manteuffel [16]. Numerical experience has shown that good estimates of the largest eigenvalue are achieved after a very few steps, but estimates of the smallest eigenvalue have poor relative accuracy until the algorithm converges. This is due to the fact that the algorithm minimizes the error in the A-norm ($\tilde{A}$-norm) which neglects the smaller eigenvalues. The A-norm can be considered to be a weighted norm, where error in the direction of an eigenvector of A is weighted by the corresponding eigenvalue. Because of the larger weights, error in the direction of the eigenvectors associated with large eigenvalues is suppressed first. This is borne out in experimental results by comparing the l_2-norm of the error with the A-norm of the error (see Section 6).

3. Incomplete Factorization. If A is positive definite, a triangular decomposition of A yields

$$A = L\Lambda U, \qquad \Lambda = \operatorname{diag}(\rho_1, \ldots, \rho_N), \qquad \rho_i > 0, \tag{3.1}$$

where L and U are unit lower and unit upper triangular. Incomplete factorization is a method by which the positive definite matrix A is split into $A = M - R$, where

$$M = \hat{L}\Sigma\hat{U}, \qquad \Sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_n) \tag{3.2}$$

and $\hat{L}$ and $\hat{U}$ are unit lower and unit upper triangular. If A is symmetric, then $C = \hat{L}\,\Sigma^{1/2}$ yields $M = CC^T$, which is known as incomplete Cholesky factorization.

The incomplete Cholesky factorization was first described by Varga [25] as a method of constructing a regular splitting of certain finite difference operators. Meijerink and van der Vorst [17] showed how it could be applied to M-matrices with arbitrary zero-structure, and that $\hat{L}$ and $\hat{U}$ could be constructed to have a predetermined zero-structure. Given G, a set of ordered pairs of integers (i, j), $1 \leqslant i, j \leqslant N$, we construct $\hat{L}$ and $\hat{U}$ so that $\hat{l}_{ij} \neq 0$ $(i > j)$, $\hat{u}_{ij} \neq 0$ $(i < j)$ only if $(i, j) \in G$. We will refer to G as the nonzero set of the factorization.

Once G has been chosen the factorization (3.2) is defined recursively by

$$\sigma_i = a_{ii} - \sum_{k=1}^{i-1} \hat{l}_{ik}\hat{u}_{ki}\sigma_k, \tag{3.3}$$

and for $j = i + 1, \ldots, N$,

$$\sigma_i \hat{l}_{ji} = \begin{cases} a_{ji} - \sum_{k=1}^{i-1} \hat{l}_{jk}\hat{u}_{ki}\sigma_k, & (j, i) \in G, \\ 0, & (j, i) \notin G, \end{cases} \tag{3.4}$$

$$\sigma_i \hat{u}_{ij} = \begin{cases} a_{ij} - \sum_{k=1}^{i-1} \hat{l}_{ik}\hat{u}_{ki}\sigma_k, & (i, j) \in G, \\ 0, & (i, j) \notin G. \end{cases} \tag{3.5}$$

If we write $M = \hat{L}\Sigma\hat{U} = A + R$, we see we have the exact factorization of the matrix $A + R$. A quick calculation shows that A and M match on the diagonal and the nonzero set G.

Suppose that the nonzero set G includes the nonzero set of A. Then M and A match at each nonzero of A. The graph of M looks like the graph of A with a few more edges. For each edge common to both the graph of A and the graph of M the corresponding matrix elements match. If the matrix A is the result of the discretization of a physical problem, then the graph of A often represents the physical domain (see Figures 6.1 and 6.2). It is instructive to view M as an approximation of A with more connectivity. It is the magnitude as well as the location of these extra edges that determine the usefulness of M as an approximation.

When implementing this procedure the elements of $\hat{L}$ and $\hat{U}$ are computed only if they are not to be set to zero later. Thus, the entire computation of $\hat{L}$, $\hat{U}$, and Σ can be carried out in the storage space that $\hat{L}$ and $\hat{U}$ will eventually occupy. The diagonal matrix Σ can be stored over the unit diagonal of L. If A is symmetric, then $\hat{L} = \hat{U}^T$ and only $\hat{L}$ need be stored.

The decomposition defined in (3.3)–(3.5) will be stable as long as $\sigma_i \neq 0$ at each step. In addition, we would like $\sigma_i > 0, \forall i$. Otherwise, $M^{-1}A$ would be indefinite. In general, indefinite systems are much harder to solve using iterative techniques than definite systems. If A is symmetric and the GCG iteration is to be used to accelerate the splitting, then the iteration as defined in Section 2 may break down unless M is positive definite. We make the following definition.

Definition. The incomplete factorization of the matrix A using nonzero set G is said to be *positive* if

$$\sigma_i > 0, \quad i = 1, \ldots, N. \tag{3.6}$$

Meijerink and van der Vorst [17] (see also [6], [26]) showed that if A is an M-matrix, then the incomplete factorization using any nonzero set will be at least as positive as complete factorization. Their method of proof involves showing that the class of M-matrices is invariant to setting off-diagonal terms to zero. The result that the class is also invariant under one step of triangular decomposition completes the proof. We can expand their result to H-matrices.

LEMMA 3.1. *Let $A = (a_{ij})$ be an M-matrix. If $A_k = (a_{ij}^{(k)})$ is the submatrix remaining after $k - 1$ steps of triangular factorization and $\hat{A}_k = (\hat{a}_{ij}^{(k)})$ is the submatrix*

remaining after $k-1$ steps of incomplete factorization with any nonzero set G, then

$$
\begin{aligned}
&0 < a_{ii}^{(k)} \leqslant \hat{a}_{ii}^{(k)}, && i = k, \ldots, N, \\
&a_{ij}^{(k)} \leqslant \hat{a}_{ij}^{(k)} \leqslant 0, && j = k, \ldots, N,\ j \neq i.
\end{aligned} \tag{3.7}
$$

Proof. The proof follows from the proofs leading up to [17, Theorem 3.2]. □

COROLLARY 3.2. *If the complete factorization of the M-matrix A yields*

$$A = L\Lambda U, \qquad \Lambda = \operatorname{diag}(\rho_1, \ldots, \rho_N),$$

and the incomplete factorization with some nonzero set G yields

$$A = \hat{L}\Sigma\hat{U} - R, \qquad \Sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_N),$$

then

$$
\begin{aligned}
&0 < \rho_i \leqslant \sigma_i, && i = 1, \ldots, N, \\
&l_{ji} \leqslant \hat{l}_{ji} \leqslant 0, \\
&u_{ij} \leqslant \hat{u}_{ij} \leqslant 0, && j = i+1, \ldots, N.
\end{aligned} \tag{3.8}
$$

Proof. The proof follows from the lemma above since

$$\rho_i = a_{ii}^{(i)}, \qquad \sigma_i = \hat{a}_{ii}^{(i)},$$

and

$$
\begin{aligned}
l_{ji} &= \frac{1}{a_{ii}^{(i)}} a_{ji}^{(i)}, & u_{ij} &= \frac{1}{a_{ii}^{(i)}} a_{ij}^{(i)}, \\
\hat{l}_{ji} &= \frac{1}{\hat{a}_{ii}^{(i)}} \hat{a}_{ji}^{(i)}, & \hat{u}_{ij} &= \frac{1}{\hat{a}_{ii}^{(i)}} \hat{a}_{ij}^{(i)}. \quad \square
\end{aligned}
$$

If A is an M-matrix, then not only do we have positivity for any nonzero set G, but the pivots of the incomplete factorization are larger than those of the complete factorization. We can use this result to show that the pivots will be positive for an H-matrix as well.

COROLLARY 3.3. *If $A = (a_{ij})$ is an H-matrix with positive diagonal elements and $\hat{A} = (\hat{a}_{ij})$, $\hat{a}_{ii} = a_{ii}$, $\hat{a}_{ij} = -|a_{ij}|$ is the associated M-matrix, then incomplete factorization with any nonzero set G will yield*

$$
\begin{aligned}
&A = L\Sigma U - R, && \Sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_N), \\
&\hat{A} = \hat{L}\hat{\Sigma}\hat{U} - \hat{R}, && \hat{\Sigma} = \operatorname{diag}(\hat{\sigma}_1, \ldots, \hat{\sigma}_N),
\end{aligned} \tag{3.9}
$$

with $0 < \hat{\sigma}_i \leqslant \sigma_i$, $i = 1, \ldots, N$,

$$
\begin{aligned}
&\hat{l}_{ji} \leqslant -|l_{ji}| \leqslant 0, \\
&\hat{u}_{ij} \leqslant -|u_{ij}| \leqslant 0, && j = i+1, \ldots, N.
\end{aligned} \tag{3.10}
$$

Proof. The proof follows from induction on the equations (3.3)–(3.5). □

Thus, the incomplete factorization of a positive definite H-matrix with any nonzero set will be positive. Note that a diagonally dominant matrix is an H-matrix.

The pivots produced by the incomplete factorization of an H-matrix A are larger than the pivots produced by the incomplete factorization of the associated M-matrix $\hat{A}$ using the same nonzero set. However, they are not necessarily larger than the pivots produced by complete factorization of A. Consider the H-matrix

$$\begin{pmatrix} 4 & 1 & 0 & -1 \\ 1 & 4 & 1 & 0 \\ 0 & 1 & 4 & 1 \\ -1 & 0 & 1 & 4 \end{pmatrix}.$$

Complete decomposition will yield

$$\rho_4 = \frac{196}{56}.$$

Incomplete decomposition with nonzero set

$$G = \begin{Bmatrix} (1,2), (1,4), (2,3), (3,4) \\ (2,1), (4,1), (3,2), (4,3) \end{Bmatrix},$$

will yield

$$\sigma_4 = \frac{195}{56}.$$

If the nonzero set is large enough, then incomplete factorization becomes complete factorization and the splitting has perfect condition. For M-matrices it can be shown that the improvement in condition is essentially monotone with respect to the nonzero set.

COROLLARY 3.4. *If $G_1 \subseteq G_2$ are two nonzero sets which yield incomplete factorizations of the M-matrix A,*

$$A = \hat{L}\hat{\Sigma}\hat{U} - \hat{R}, \qquad \hat{\Sigma} = \operatorname{diag}(\hat{\sigma}_1, \ldots, \hat{\sigma}_N),$$

$$A = L\Sigma U - R, \qquad \Sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_N),$$

respectively, then

$$\begin{aligned} &0 < \sigma_i \leqslant \hat{\sigma}_i, && i = 1, \ldots, N, \\ &l_{ji} \leqslant \hat{l}_{ji} \leqslant 0, \\ &u_{ij} \leqslant \hat{u}_{ij} \leqslant 0, && j = i+1, \ldots, N. \end{aligned} \tag{3.11}$$

Proof. The proof is analogous to the proof Lemma 3.1 and Corollary 3.2 where incomplete factorization with nonzero set G_2 is substituted for complete factorization. □

Note that we may write the unit triangular matrices as

$$L = I - T, \qquad U = I - V,$$

where T and V are strictly triangular. We have

$$L^{-1} = (I + T + T^2 + \cdots + T^{N-1}), \qquad U^{-1} = (I + V + V^2 + \cdots + V^{N-1}). \tag{3.12}$$

Consider the complete factorization of an M-matrix A and the incomplete factorizations with nonzero sets $G_1 \subseteq G_2$ denoted as

$$\begin{aligned} A &= (I - T)\Lambda(I - V), \\ A &= M_2 - R_2 = (I - T_2)\Sigma_2(I - V_2) - R_2, \\ A &= M_1 - R_1 = (I - T_1)\Sigma_1(I - V_1) - R_1. \end{aligned} \tag{3.13}$$

If A is an M-matrix, Eqs. (3.8) and (3.11) yield

$$0 \leqslant \Lambda \leqslant \Sigma_2 \leqslant \Sigma_1, \quad 0 \leqslant T_1 \leqslant T_2 \leqslant T, \quad 0 \leqslant V_1 \leqslant V_2 \leqslant V. \tag{3.14}$$

In light of (3.12) we see that

$$A^{-1} \geqslant M_2^{-1} \geqslant M_1^{-1} \geqslant 0. \tag{3.15}$$

As we increase the nonzero set, M^{-1} moves term by term closer to A^{-1}. This would lead one to suspect that the condition of the splitting with M_2 is better than the condition of the splitting with M_1. This can be shown for M-matrices up to a factor of 2. First, we will need the following result due to Perron and Frobenius; see Varga [26, pp. 30, 47].

THEOREM 3.5. *Let $A \geqslant 0$ term by term. Then, there is a positive (real) eigenvalue of A equal to the spectral radius of A. Further, if $0 \leqslant |B| \leqslant A$, then the spectral radius of A is greater than or equal to the spectral radius of B.*

We have the following result:

THEOREM 3.6. *Let $G_1 \subseteq G_2$ be two nonzero sets that yield incomplete factorizations on the M-matrix A as in* (3.13). *Let μ_i be the eigenvalue of smallest modulus of $\tilde{A}_i = M_i^{-1}A$ for $i = 1, 2$. Then*

$$0 < \mu_1 \leqslant \mu_2; \tag{3.16}$$

and if A is symmetric,

$$\frac{1}{\mu_i} \leqslant K(\tilde{A}_i) \leqslant \frac{2 - \mu_i}{\mu_i}, \quad i = 1, 2. \tag{3.17}$$

Proof. The second result follows from the fact that the incomplete factorization of an M-matrix with any nonzero set yields a regular splitting; see Meijerink and van der Vorst [17, p. 151] and Varga [26, p. 87]. To see the first result, let

$$W_i = \Sigma_i^{1/2}(I - V_i)^T A^{-1}(I - T_i)\Sigma_i^{1/2}.$$

We have that $\tilde{A}_i^{-1} = A^{-1}M_i$ is similar to W_i. Now

$$\begin{aligned} W_i &= \Sigma_i^{1/2}(I - V_i)^T(I - V)^{-T}\Lambda^{-1}(I - T)^{-1}(I - T_i)\Sigma_i^{1/2} \\ &= \Sigma_i^{1/2}((I - V) + (V - V_i))^T(I - V)^{-T}\Lambda^{-1}(I - T)^{-1}((I - T) + (T - T_i))\Sigma_i^{1/2} \\ &= \Sigma_i^{1/2}(I + (I - V)^{-1}(V - V_i))^T\Lambda^{-1}(I + (I - T)^{-1}(T - T_i))\Sigma_i^{1/2}. \end{aligned}$$

From (3.14) each term in the above representation of W_i is positive and yields

$$W_1 \geqslant W_2 \geqslant 0.$$

Since incomplete factorization yields a regular splitting, the spectral radius of $I - \tilde{A}_i$ is less than 1 for $i = 1, 2$. The eigenvalues of $\tilde{A}_i$ and thus W_i have positive real part for $i = 1, 2$. By the Perron-Frobenius theorem, W_i has a positive eigenvalue, say ω_i, equal to its spectral radius and further $\omega_2 \leqslant \omega_1$. Thus,

$$0 < \mu_1 = \frac{1}{\omega_1} \leqslant \mu_2 = \frac{1}{\omega_2},$$

which completes the proof. □

4. Measure of Positivity. Although many applications yield matrices that are M-matrices or H-matrices, the application of finite element techniques seldom does. The above sufficient conditions are very strict, and certainly not necessary. In fact, it is not even necessary for A to be positive definite for the incomplete factorization with *some* nonzero set to yield positive pivots, even through complete factorization of a symmetric indefinite matrix must yield nonpositive pivots. In general, positivity depends upon the nonzero set. We know that if A is positive definite there is some nonzero set for which the incomplete factorization is positive. In this section we will discuss a way to measure how positive a factorization is.

If one of the pivots becomes negative then it is clear that the factorization is not positive. However, even if all the pivots stay positive it is important that they not become to small. A small pivot would yield a nearly singular M and thus $M^{-1}A$ would most likely have poor condition. We may ask: How small is too small? We need a way to compare the size of the pivot to the size of the elements of A.

Consider the positive definite matrix $A = (a_{ij})$. Since $a_{ii} \neq 0$ we can rewrite the system

$$A\mathbf{x} = \mathbf{b}$$

as

$$\tilde{A}\mathbf{y} = \tilde{\mathbf{b}}, \quad \tilde{A} = DAD, \quad D\mathbf{y} = \mathbf{x}, \quad \tilde{\mathbf{b}} = D\mathbf{b}, \tag{4.1}$$

where $D = \operatorname{diag}(\cdots a_{ii}^{-1/2} \cdots)$ is a diagonal scaling matrix. Here $\tilde{A}$ has unit diagonal.† If we perform an incomplete factorization on $\tilde{A}$ to get pivots $\Sigma = \operatorname{diag}(\sigma_1, \ldots, \sigma_N)$, then it is easy to see that $\sigma_1 = 1$, $\sigma_i \leqslant 1$, $i = 2, \ldots, N$. Hence,

$$S = 1/\min_i (\sigma_i) \tag{4.2}$$

yields a good measure of the positivity. We next show that the ICCG procedure is invariant to diagonal scaling. We use the following easy result.

---

†It is shown in Forsythe and Straus [7] that this is the best scaling in terms of $K(A)$ if A has 'property-A'.

LEMMA 4.1. *Let $\tilde{A} = PAQ$ where P and Q are nonsingular diagonal matrices. If an incomplete factorization on A with nonzero set G is positive and yields*

$$A = M - R = L\Sigma U - R,$$

then an incomplete factorization on $\tilde{A}$ with nonzero set G will yield

$$\tilde{A} = \tilde{M} - \tilde{R} = \tilde{L}\tilde{\Sigma}\tilde{U} - \tilde{R},$$

where

$$\tilde{L} = PLP^{-1}, \quad \tilde{U} = Q^{-1}L^TQ, \quad \tilde{\Sigma} = P\Sigma Q, \quad \tilde{R} = PRQ.$$

Thus, $M^{-1}A$ is similar to $\tilde{M}^{-1}\tilde{A}$ since

$$\tilde{M}^{-1}\tilde{A} = (PMQ)^{-1}(PAQ) = Q^{-1}M^{-1}AQ.$$

Since the convergence of the GCG iteration is dependent upon the entire spectrum of $M^{-1}A$, the ICCG procedure is invariant to diagonal scaling. Thus, we may scale the system before factoring and use S as a measure of positivity. Notice also that

$$S = 1/\min(\sigma_i) \leqslant K(M). \tag{4.3}$$

The measure of positivity can be used to approximate $K(M)$ in the bound (2.7).

Let us assume that A has been diagonally scaled as in (4.1). If we knew the spectrum of $M^{-1}A$, we could predict the convergence of the GCG iteration. Unfortunately, this information is computationally unavailable at the time of the splitting. In fact, even $K(A)$ may not be known accurately. We do know that for the optimal splitting, the complete factorization, we have $M^{-1} = A^{-1}$, $K(M) = K(A)$. It is also clear that the condition of M should not be allowed to be significantly larger than the condition of A. If an estimate of $K(A)$ is available, then a reasonable computational strategy is to accept the factorization only if

$$S \leqslant K(A). \tag{4.4}$$

This bound has worked well in practice. In fact, in the examples to be described later the factorizations which performed best corresponded to values of S many orders of magnitude smaller than $K(A)$.

5. Shifted Incomplete Cholesky Factorization. In the remainder of this paper we will be concerned with symmetric positive definite systems. We will assume that the nonzero set is symmetric also; that is, $(i, j) \in G$ implies $(j, i) \in G$. The most convenient nonzero set in terms of storage and handling is the nonzero set of A. There are many examples of symmetric positive definite matrices for which incomplete factorization with this nonzero set is not positive. We can still apply this method by performing an incomplete factorization on a matrix that is close to A. One way to do this is to make A more nearly diagonally dominant. Suppose we write

$$A = D - B,$$

where D is the diagonal of A and B is the off-diagonal. Clearly, there is some value of α such that the matrix

$$(1 + \alpha)D - B \tag{5.1}$$

is diagonally dominant. Incomplete factorization of this shifted matrix will be positive. Although diagonal dominance is sufficient, it is not in general necessary and an incomplete factorization may be positive for much smaller values of α. This motivates the following splitting of A.

Consider the pencil of matrices

$$A(\alpha) = D - \frac{1}{1 + \alpha} B. \tag{5.2}$$

Suppose that incomplete factorization of $A(\alpha)$ with nonzero set G yields

$$A(\alpha) = L\Sigma L^T - R(\alpha) = M(\alpha) - R(\alpha); \tag{5.3}$$

(here L, Σ depend upon α also) then, we may write

$$\begin{aligned} A = D - B &= D - \frac{1}{1 + \alpha} B - \frac{\alpha}{1 + \alpha} B \\ &= M(\alpha) - \left(R(\alpha) + \frac{\alpha}{1 + \alpha} B\right) = M(\alpha) - N(\alpha). \end{aligned} \tag{5.4}$$

For each α for which the incomplete factorization is positive, we have a splitting of A. With $\alpha = 0$ the splitting is the incomplete factorization with nonzero set G.

Consider the difference between A and $M(\alpha)$. For $\alpha = 0$, $M(\alpha)$ matches A on the diagonal and on the nonzero set G; that is, if $(i, j) \in G$, then $a_{ij} = m_{ij}$. For $\alpha > 0$, some of the error is shifted onto the off-diagonal terms of M.

In incomplete Cholesky factorization, just as in complete Cholesky factorization, the factors L and Σ are rational functions of the elements of $A(\alpha)$. Since for all α sufficiently large $A(\alpha)$ is diagonally dominant and thus the incomplete factorization is positive, we have

$$\lim_{\alpha \to \infty} M(\alpha) = D, \qquad \lim_{\alpha \to \infty} N(\alpha) = B, \qquad \lim_{\alpha \to \infty} R(\alpha) = 0. \tag{5.5}$$

If the nonzero set G includes the nonzero set of A, then the limit in (5.5) is quickly achieved. A straightforward calculation (which we omit) yields

$$\sigma_i = a_{ii} - \frac{1}{(1 + \alpha)^2} \sum_{k<i} \frac{a_{ik}^2}{a_{kk}} + O\left(\frac{1}{(1 + \alpha)^3}\right), \tag{5.6}$$

$$l_{ji} = \frac{1}{(1 + \alpha)} \frac{a_{ji}}{a_{ii}} - \frac{1}{(1 + \alpha)^2} \frac{1}{a_{ii}} \sum_{k<i} \frac{a_{ik}a_{jk}}{a_{kk}} + O\left(\frac{1}{(1 + \alpha)^3}\right), \tag{5.7}$$

$$r_{ji} = \frac{1}{(1 + \alpha)^2} \sum_{k<i} \frac{a_{ik}a_{jk}}{a_{kk}} + O\left(\frac{1}{(1 + \alpha)^3}\right). \tag{5.8}$$

Thus, as α increases the terms in $N(\alpha)$ quickly approach B. From (5.4) and (5.8) we can write

$$N(\alpha) = \frac{\alpha}{1+\alpha} B + O\left(\frac{1}{(1+\alpha)^2}\right),$$

so that

$$D - B = M(\alpha) - \frac{\alpha}{1+\alpha} B + O\left(\frac{1}{(1+\alpha)^2}\right),$$

or

$$M(\alpha) = D - \frac{1}{1+\alpha} B + O\left(\frac{1}{(1+\alpha)^2}\right). \tag{5.9}$$

The generalized conjugate gradient algorithm is sensitive to the spectrum of $M^{-1}(\alpha)A$. Suppose we write

$$M^{-1}(\alpha) = (D^{-1} + E).$$

A simple calculation shows that

$$E = \frac{1}{1+\alpha} D^{-1}BD^{-1} + O\left(\frac{1}{(1+\alpha)^2}\right),$$

so that

$$\begin{aligned} M^{-1}(\alpha)A &= \left(D^{-1} + \frac{1}{1+\alpha} D^{-1}BD^{-1} + O\left(\frac{1}{(1+\alpha)^2}\right)\right)(D - B) \\ &= I - D^{-1}B + \frac{1}{1+\alpha}(D^{-1}B - (D^{-1}B)^2) + O\left(\frac{1}{(1+\alpha)^2}\right). \end{aligned} \tag{5.10}$$

In the limit this is equivalent to a Jacobi splitting. From Section 2 we see that the generalized conjugate gradient algorithm on the Jacobi splitting is equivalent to the conjugate gradient algorithm on the diagonally scaled system (4.1). Let the eigenvalues of $-D^{-1}B$ be given by $\eta_1 \leqslant \eta_2 \leqslant \cdots \leqslant \eta_N$ and the eigenvalues of $M^{-1}(\alpha)A$ be given by $0 < \mu_1 \leqslant \mu_2 \leqslant \cdots \leqslant \mu_N$. Then for large α we have

$$\mu_i \sim 1 + \eta_i - \frac{1}{1+\alpha}(\eta_i + \eta_i^2). \tag{5.11}$$

If we differentiate the right-hand side with respect to α we have

$$\frac{\partial \mu_i}{\partial \alpha} \sim \frac{1}{(1+\alpha)^2}(\eta_i + \eta_i^2). \tag{5.12}$$

If $\eta_i > 0$, then in the limit μ_i is increasing to $1 + \eta_i$ or $\mu_i < 1 + \eta_i$ for sufficiently large α. If $\eta_i < 0$, then, because $I - D^{-1}B$ is positive definite, we must have $-1 < \eta_i < 0$. Thus, μ_i is decreasing to $1 + \eta_i$, or $\mu_i > 1 + \eta_i$ for sufficiently large α. We formalize this discussion with the following theorem. First, let

$$\tilde{A}(\alpha) = M^{-1/2}(\alpha)AM^{-1/2}(\alpha) \tag{5.13}$$

and

$$\tilde{A} = \lim_{\alpha \to \infty} \tilde{A}(\alpha) = D^{-1/2} A D^{-1/2}. \tag{5.14}$$

THEOREM 5.1. *Suppose the nonzero set G includes the nonzero set of A. If $D^{-1}B$ is indefinite, then there exists $\alpha_u > 0$ such that for $\alpha > \alpha_u$, $K(\tilde{A}(\alpha))$ is strictly increasing and*

$$K(\tilde{A}(\alpha)) < K(\tilde{A}). \tag{5.15}$$

Moreover, if $\mu_i(\alpha)$ is an eigenvalue of $\tilde{A}(\alpha)$ and μ_i is the corresponding eigenvalue of $\tilde{A}$, then for $\mu_i > 1$ and α sufficiently large, $\mu_i(\alpha)$ is strictly increasing and

$$1 \leqslant \mu_i(\alpha) < \mu_i, \tag{5.16}$$

and for $\mu_i < 1$ and α sufficiently large, $\mu_i(\alpha)$ is strictly decreasing and

$$\mu_i < \mu_i(\alpha) \leqslant 1. \tag{5.17}$$

Proof. The result is clear from the discussion above. A more rigorous argument can be made from the fact that (5.10) is actually a matrix each of whose elements is a rational function in $1/(1 + \alpha)$. Thus, there is an α_u large enough so that for $\alpha > \alpha_u$ both sides of (5.12) have the same sign. If $\alpha_u < \alpha_1 < \alpha_2$, then the splitting $\tilde{A}(\alpha_1)$ will perform better for any initial error than the splitting $\tilde{A}(\alpha_2)$ when accelerated by conjugate gradients because of the eigenvalue structure defined by (5.11), (5.16), and (5.17) (see Greenbaum [8]). □

For large α, $K(\tilde{A}(\alpha))$ is majorized by $K(\tilde{A})$. It is also clear that if α were allowed to decrease without bound there would be some α for which $M(\alpha)$ has a zero pivot. Let α_l be the first such α as α decreases from α_u. We have

$$\lim_{\alpha \to \alpha_l +} K(\tilde{A}(\alpha)) = +\infty,$$

because $M^{-1}(\alpha)$ becomes unbounded. We would like to find α that satisfies

$$\min_{\alpha_l \leqslant \alpha \leqslant \alpha_u} K(\tilde{A}(\alpha)), \tag{5.18}$$

or more specifically, find the α for which the convergence properties of $\tilde{A}(\alpha)$ are best. The relationship between α and $K(\tilde{A}(\alpha))$ is not clear, but experimental results seem to indicate that it is roughly a unimodal function on the interval (α_l, α_u). It is not difficult to approximate α_l by trial and error in that the factorization will not be positive for $\alpha < \alpha_l$ and the measure of positivity proposed in (4.3) will be large for α near α_l. In the test problems the best α in terms of convergence was very close to α_l. The minimum of $K(\tilde{A}(\alpha))$ occurred at a larger value of α and was fairly insensitive to overestimation. In short, a good guess was good enough. More will be said about the choice of α in the next section.

We can combine the results of Section 3 with the above discussion to show that, up to a factor of 2, the best value of α for a Stieltjes matrix is $\alpha \leqslant 0$. Observe that $\alpha_l < 0$ for a Stieltjes matrix because the factorization is positive for $\alpha = 0$. If A is not Stieltjes but is diagonally dominant, the best value of α may be nonzero.

THEOREM 5.2. *Let $A = (I - T)\Lambda(I - T)^T$ be a Stieltjes matrix. Suppose a shifted incomplete Cholesky factorization with nonzero set G yields*

$$A = (I - T(\alpha))\Sigma(\alpha)(I - T(\alpha))^T - \left(R(\alpha) + \frac{\alpha}{1+\alpha} B\right).$$

If $\alpha_l < \alpha_1 < \alpha_2 < \infty$, then

$$0 \leqslant R(\alpha_2) \leqslant R(\alpha_1). \tag{5.19}$$

If in addition $0 \leqslant \alpha_1 < \alpha_2$, then

$$0 \leqslant \Lambda \leqslant \Sigma(\alpha_1) \leqslant \Sigma(\alpha_2), \tag{5.20}$$

$$0 \leqslant T(\alpha_2) \leqslant T(\alpha_1) \leqslant T, \tag{5.21}$$

$$A^{-1} \geqslant M^{-1}(\alpha_1) \geqslant M^{-1}(\alpha_2) \geqslant 0. \tag{5.22}$$

Moreover, if $\mu_1(\alpha)$ is the smallest eigenvalue of $\tilde{A}(\alpha)$, then $\mu_1(\alpha)$ is maximized over the interval $[0, \infty)$ at $\alpha = 0$ and

$$\frac{1}{\mu_1(\alpha)} \leqslant K(\tilde{A}(\alpha)) \leqslant \frac{2 - \mu_1(\alpha)}{\mu_1(\alpha)}.$$

Proof. Let $A^{(k)} = (a_{ij}^{(k)})$, $A^{(k)}(\alpha) = (a_{ij}^{(k)}(\alpha))$ be the submatrix remaining after $k - 1$ steps of complete factorization on A and shifted incomplete factorization with nonzero set G on $A(\alpha) = D - (1/(1+\alpha))B$, respectively. If $\alpha_1 < \alpha_2$, then as in (3.7) we have

$$\begin{aligned} &0 < a_{ii}^{(1)}(\alpha_1) \leqslant a_{ii}^{(1)}(\alpha_2), && i = 1, \ldots, N, \\ &a_{ij}^{(1)}(\alpha_1) \leqslant a_{ij}^{(1)}(\alpha_2) \leqslant 0, && j = 1, \ldots, N,\ j \neq i. \end{aligned} \tag{5.23}$$

If in addition $0 \leqslant \alpha_1 < \alpha_2$, then we have

$$\begin{aligned} &0 < a_{ii}^{(1)} \leqslant a_{ii}^{(1)}(\alpha_1) \leqslant a_{ii}^{(1)}(\alpha_2), && i = 1, \ldots, N, \\ &a_{ij}^{(1)} \leqslant a_{ij}^{(1)}(\alpha_1) \leqslant a_{ij}^{(1)}(\alpha_2) \leqslant 0, && j = 1, \ldots, N,\ j \neq i. \end{aligned} \tag{5.24}$$

The induction proof of Lemma 3.1 follows through intact. We may replace the superscript (1) by (k) in (5.23) and (5.24). Now

$$r_{ij}(\alpha) = \begin{cases} -a_{ij}^{(j)}(\alpha), & (i, j) \notin G, \\ 0, & (i, j) \in G, \end{cases}$$

which yields (5.19). For $0 \leqslant \alpha_1 < \alpha_2$, the proof of (5.20), (5.21), and (5.22) is equivalent to the chain of proofs leading up to (3.14).

To see that $\mu_1(\alpha)$ has a maximum at $\alpha = 0$, consider $\alpha > 0$. Then, as in Theorem 3.6, we have that $\tilde{A}^{-1}(\alpha)$ is similar to

$$\Sigma^{1/2}(\alpha)(I + (I - T)^{-1}(T - T(\alpha)))^T \Lambda^{-1}(I + (I - T)^{-1}(T - T(\alpha)))\Sigma^{1/2}(\alpha).$$

By the Perron-Frobenius Theorem 3.5, the spectral radius of $\tilde{A}^{-1}(\alpha)$ is monotonically increasing in α.

The last statement of the theorem follows from the fact that the shifted incomplete factorization of a Stieltjes matrix yields a regular splitting for any $\alpha \geqslant 0$. □

6. Numerical Results.

(a) *Motivation.* As was mentioned in the introduction, this work was motivated by a three-dimensional model of the structural deformation of a cylinder with varying thickness and holes (see Figure 6.1). Assuming linear deformation and using isoparametric 20-node brick finite elements, the model required the solution of a positive definite linear system with approximately 18,000 unknowns, a half-bandwidth of 1,300 and 1,040,000 nonzero elements in the upper triangular part. The number of unknowns varied slightly according to the boundary conditions imposed. The condition of the matrix was estimated to be on the order of 10^8.

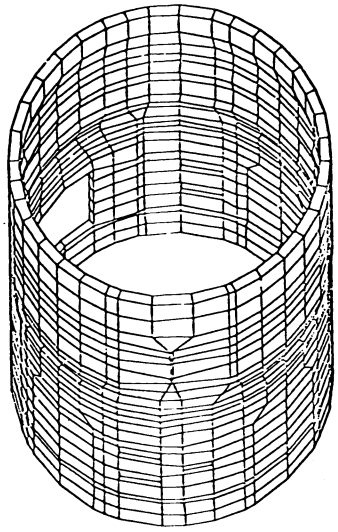

FIGURE 6.1
Motivating problem: Cylinder

A direct solution to this system was sought but never achieved, due to the large amount of storage required for the upper triangular band (23,000,000) and problems with the computer system. It was estimated from partial runs to require 9,000 CP seconds on the CDC-7600 at Sandia Laboratories in Albuquerque, or 50,000–70,000 CP seconds on the CDC-6600 at Sandia Laboratories in Livermore.

Conjugate gradients with diagonal scaling, that is, acceleration of the Jacobi splitting, was only moderately successful. This implementation required 2,000,000 words of storage for the matrix. A series of runs were made in which the algorithm was restarted at each run using the solution from the previous run as an initial guess. (It is well known that restarting, rather than resuming the iteration will slow convergence.) Some of the runs were made on the CDC-7600 and some on the CDC-6600. The l_2-norm of the residual was reduced by a factor of 10^{-6} after a total of about 4,000 iterations which is estimated to cost 7,000 CP seconds on the CDC-7600 or 40,000–45,000 CP seconds on the CDC-6600 (see Table 6.1).

The shifted incomplete Cholesky factorization was more successful. It required 3,000,000 words of storage. Using nonzero set G equal to the nonzero set of A, the factorization required 700 CP seconds on the CDC-6600. The problem was solved for several sets of boundary conditions on the CDC-6600. Using ad hoc values of α ranging from .05 to .10, an acceptable solution was achieved after approximately 200 iterations or 6,000 CP seconds. A solution of much higher resolution was achieved in approximately 700 iterations or 20,000 CP seconds (see Table 6.1).

TABLE 6.1

	Time (6600)	Error Reduction	Storage
DIRECT	50,000–70,000	------	23,000,000
CG	40,000–45,000	10^{-6}	2,000,000
SICCG	6,000–10,000	10^{-5}	3,000,000
	20,000–25,000	10^{-10}	

It is clear that the savings in both time and storage were significant. Also notice that the time required to perform the factorization was small compared to the overall effort. In such a problem it is feasible to spend time searching for a good value of the parameter α.

(b) *Test Problem.* The SICCG procedure was tested extensively on much smaller problems—where the advantage over direct methods is no longer clear. The algorithm behaved similarly in each case. The remainder of this section will be devoted to exploring in depth the results from one test problem.

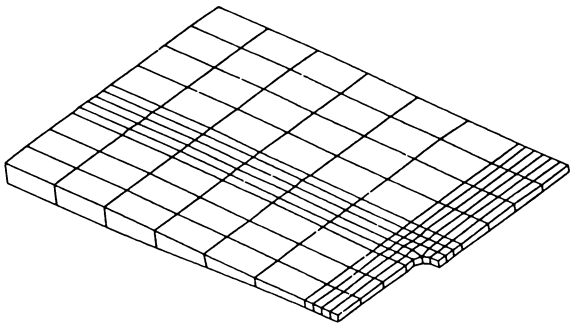

FIGURE 6.2

Test problem: Tapered slab

The test problem was a three-dimensional model of structural deformation in a tapered slab (see Figure 6.2). Again linear deformation was assumed and isoparametric 20-node brick finite elements were used. The boundary conditions corresponded to pressing the thin edge against a wall like a wedge. This gave a system with 3090 unknowns, a half-bandwidth of 286 and 170,000 nonzeros in the upper triangular part.

The matrix was scaled symmetrically to have unit diagonal as in (4.1). The condition of the system was estimated to be on the order of 10^{+8}.

With known solution, the following norms can be easily constructed as the iteration proceeds.

1. $\langle \mathbf{e}, \mathbf{e} \rangle = \|\mathbf{e}\|^2$ l_2-norm of error,
2. $\langle M\mathbf{e}, \mathbf{e} \rangle = \|\mathbf{e}\|_M^2$ M-norm of error,
   $= \|\tilde{\mathbf{e}}\|^2$ l_2-norm of error of system (2.4),
3. $\langle A\mathbf{e}, \mathbf{e} \rangle = \|\mathbf{e}\|_A^2$ A-norm of error,
4. $\langle \mathbf{r}, \mathbf{r} \rangle = \|\mathbf{r}\|^2$ l_2-norm of residual,
   $= \|\mathbf{e}\|_{A^2}^2$ A^2-norm of error,
5. $\langle \mathbf{r}, \mathbf{h} \rangle = \langle \tilde{\mathbf{r}}, \tilde{\mathbf{r}} \rangle$ l_2-norm of residual of system (2.4),
   $= \|\tilde{\mathbf{e}}\|_{\tilde{A}^2}^2$ $\tilde{A}^2$-norm of error of system (2.4).

Recall from Section 2 that $\|\mathbf{e}\|_A = \|\tilde{\mathbf{e}}\|_{\tilde{A}}$.

Figure 6.3 demonstrates the behavior of the log of the relative error in each of the above norms for the test problem with $\alpha = .005$ using graph-set G equal to the graph-set of A. Time is measured in CP seconds on the CDC-6600. The initial plateau represents the time required for the factorization.

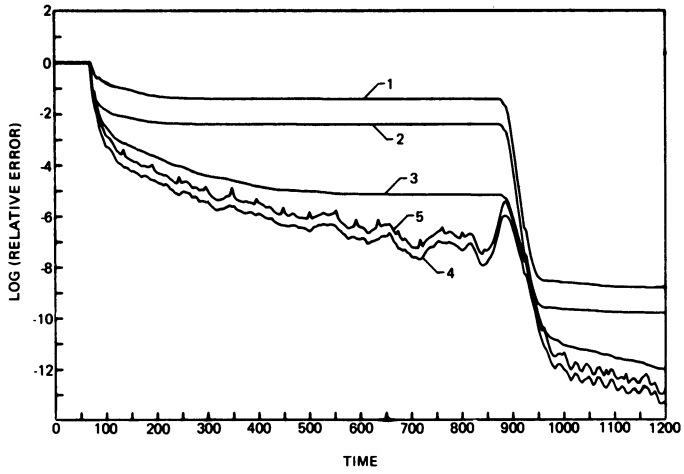

FIGURE 6.3

Relative error in 5 norms

Figure 6.3 demonstrates several points of interest. The error appears to fall off of a cliff when the algorithm finally converges. This behavior is more pronounced when α is near its optimal value. This may be due to bunching of the eigenvalues of $M^{-1}A$ so that the minimal polynomial of $M^{-1}A$ can be approximated at its roots by a polynomial of much smaller degree. Such behavior was demonstrated by D. Kershaw [13] for certain M-matrices.

Notice that the two norms that can be computed without knowledge of the solution, namely the residuals, oscillate whereas the other three decrease monotonically. The A-norm will decrease monotonically because the algorithm is optimal with respect to this norm, as will the M-norm (Hestenes and Stiefel [9, p. 416]). Also, notice

that the residual errors actually increase just before the cliff. A user unaware of this characteristic might tend to abandon the iteration on the brink of convergence.

Notice that the l_2-norms $\|e\|$ and $\|\tilde{e}\|$ lag behind the other norms. This is again due to the fact that the method minimizes the A-norm of the error. The A-norm can be considered to be a weighted norm, where the error in the direction of an eigenvector is weighted by the corresponding eigenvalue. Because of the larger weights, the error in the direction of eigenvectors associated with large eigenvalues is suppressed first.

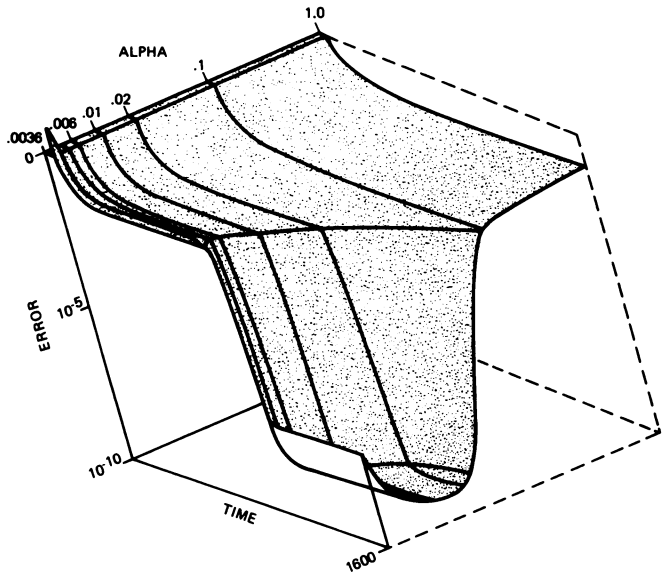

FIGURE 6.4

Convergence surface: Error versus time versus α

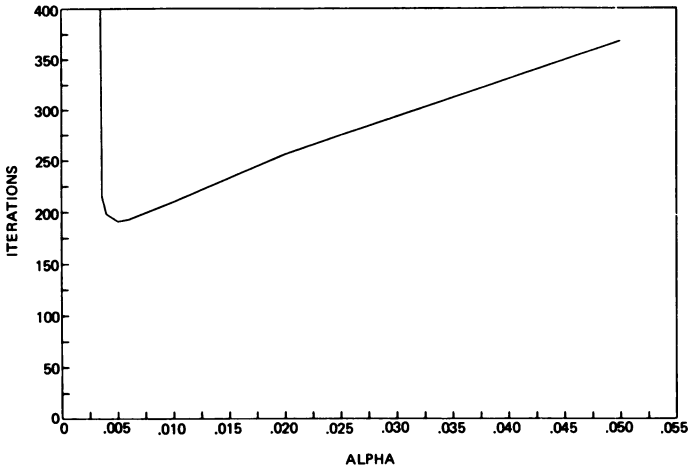

FIGURE 6.5

Iterations to convergence

The l_2-norm of the error is frequently the norm of greatest interest among those measured. It was decided to use this norm to measure error. The cliff gives an easy criteria for convergence. As α was varied, the time required to reach the cliff varied. Figure 6.4 shows a three-dimensional surface representing the log of the relative error in the l_2-norm versus CP time on the 6600 for different values of α using nonzero set G equal to the nonzero set of A. Figure 6.5 gives a "top view" of this surface by comparing the number of iterations required to reach the cliff for various values of α. It was found experimentally that the factorization was not positive for $\alpha < .0036$ and that convergence was fastest for $\alpha_b = .0055$. Notice that for $\alpha = .05$, an order of magnitude larger, convergence still occurred within a reasonable number of iterations.

Near α_b the factorization was very positive. Table 6.2 shows the measure of positivity, S, for various values of α. Even $\alpha = .0036$ gave a surprisingly small S.

TABLE 6.2

	α	λ_{max}	S
	.0036	36.0	146.7
	.004	14.0	37.2
	.0045	8.95	26.2
	.005	6.62	22.0
$\alpha_b =$	.0055	5.29	20.6
	.006	4.42	19.5
	.007	3.37	17.8
	.010	2.21	15.4
	.011	1.92	14.9
	.012	1.77	14.4
$\alpha_c =$	.015	1.51	13.1
	.020	1.40	11.5
	.050	1.34	7.24
	.10	1.30	4.64
	1.6	2.03	1.21
	9.0	3.90	1.02
	100	5.50	1.00019
	∞	5.78	1.0

(c) *Choosing* α. Figure 6.6 shows a superposition of the graph of the iterations versus α of Figure 6.5 with a graph of $K(\tilde{A}(\alpha))$ versus α. Notice that the minimum of $K(\tilde{A}(\alpha))$ does not correspond exactly with the value of α that produced convergence in the fewest number of iterations. This pattern was repeated in other test problems. This may be due to a bunching of eigenvalues of $K(\tilde{A}(\alpha))$ that occurs near the optimal value of α.

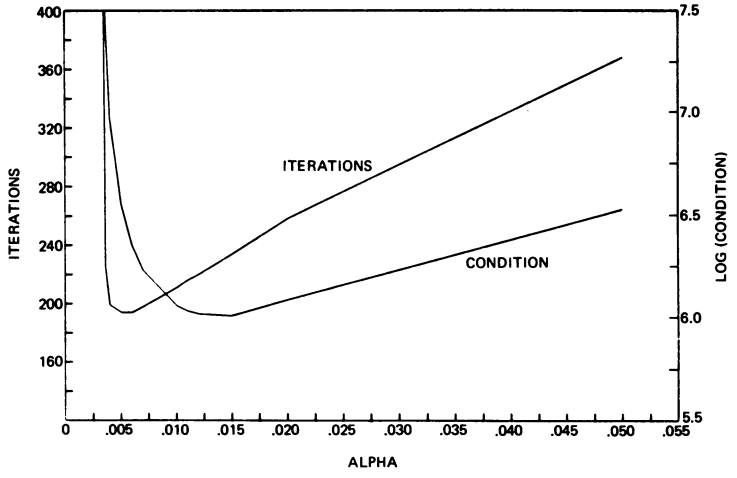

FIGURE 6.6

Iterations to convergence and condition

Figure 6.6 shows that in the absence of prior knowledge of α_b, the value α_c that minimizes $K(\tilde{A}(\alpha))$ will still yield acceptable results. Unfortunately, good estimates of $K(\tilde{A}(\alpha))$ are only available after convergence occurs. Figure 6.7 shows a superposition of the graph of the log of the relative l_2-norm of the error and the log of the estimated condition versus iterations for $\alpha = .005$. The condition is estimated by the procedure described in Section 2. Notice that the estimated condition is increasing until after convergence.

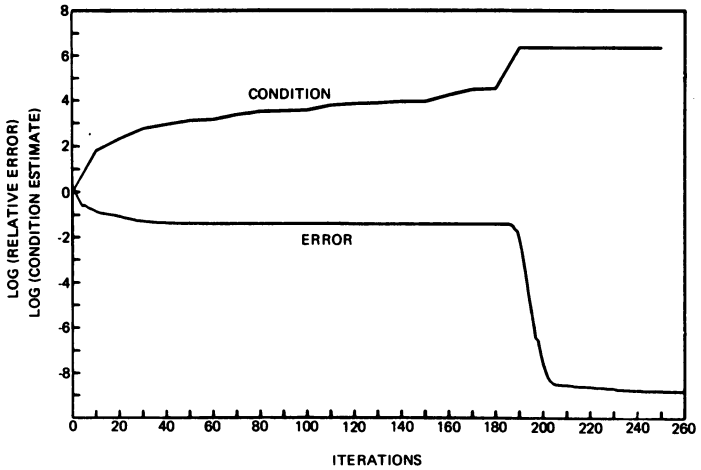

FIGURE 6.7

Condition estimate and relative error

However, a good estimate of the largest eigenvalue of $\tilde{A}(\alpha)$ is available after only a few steps. In each test the estimate of the largest eigenvalue was accurate to 2 places after only 10 steps of the iteration. Figure 6.8 shows a graph of the largest eigenvalue of $\tilde{A}(\alpha)$ versus α. The same data is presented in Table 6.2. As α goes to infinity, the largest eigenvalue approaches the largest eigenvalue of the Jacobi splitting of A.

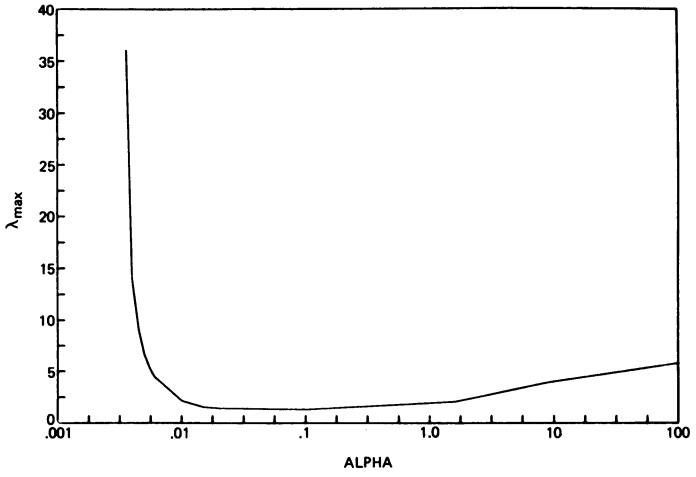

FIGURE 6.8

Largest eigenvalue of splitting

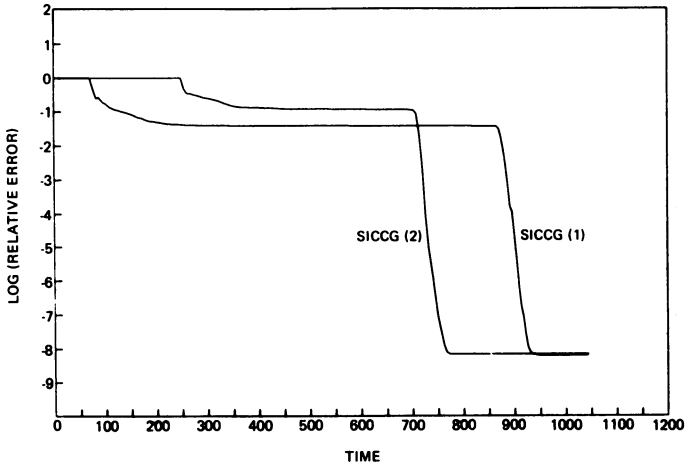

FIGURE 6.9

Comparison of splittings with G(1) and G(2)

Notice that α_b yields a largest eigenvalue, $\lambda_{max} = 5.29$, which is on the same order of magnitude as the largest eigenvalue of the Jacobi splitting, $\lambda_{max} = 5.78$. On the other hand, α_c yields a largest eigenvalue $\lambda_{max} = 1.51$. In the absence of mathematical analysis to better explain the choice of α, the parameter was manipulated so as to yield a largest eigenvalue of $\tilde{A}(\alpha)$ between 2 and the

largest eigenvalue of the associated Jacobi splitting. The largest eigenvalue was computed by performing the factorization, then iterating for ten steps and finding an estimate of the largest eigenvalue as described in Section 2. This procedure gave good results on a variety of test problems.

(d) *Extended Nonzero Set.* Using nonzero set G equal to the nonzero set of A may be thought of as allowing fill-in in location (i, j) of L whenever unknown i and unknown j are neighbors in the graph of A. Suppose we extend this association and allow fill-in in location (i, j) of L whenever unknown i and unknown j have a common neighbor in the graph of A. Not all such locations will actually fill in. We need only have $(i, j) \in G$ whenever there exists $k \leqslant j < i$ such that (i, k) and (j, k) are in the nonzero set of A.

TABLE 6.3

ALPHA	STEPS	CONDITION	λmax
.0005	82	.361 + 7	51.9
.00051	80	.184 + 7	26.3
α_b = .0006	78	.407 + 6	5.64
.0007	79	.259 + 6	3.46
.0010	82	.167 + 6	2.04
α_c = .0015	91	.148 + 6	1.55
.0020	98	.152 + 6	1.39
.0030	111	.175 + 6	1.28

We will refer to the nonzero set of A as $G(1)$ and this first level of extension as $G(2)$.†† Clearly, the nonzero set may be extended any number of levels until all possible fill-in is accounted for. With increasing levels it is assumed that $M^{-1}(\alpha)$ more closely approximates A^{-1} and thus fewer iterations are required. However, more work and storage is required for a factorization with a higher level of extension and each iteration requires more work. The optimal level of extension will depend upon the problem as well as the computer to be used. In the work of Meijerink and van der Vorst [17] the first level of extension was an improvement, while in the work of Kershaw [13] it was not an improvement in terms of total work. For the test problem above the first level extension was an improvement. Figure 6.9 shows a comparison of the log of the l_2-norm of the relative error using nonzero sets $G(1)$, $\alpha = .005$, and $G(1)$, $\alpha = .0015$ versus CP time. Notice that the plateau representing the factorization is much larger for $G(2)$ but the total time is significantly less. Figure 6.10

††This corresponds to ICCG(3) in Meijerink and van der Vorst [17] and Kershaw [13].

shows the iterations to convergence and the condition of $\tilde{A}(\alpha)$ for various values of α. The largest eigenvalue is given in Table 6.3. In this example, the best value of α gave a largest eigenvalue close to the largest eigenvalue of the Jacobi splitting.

In general, one can expect the smallest α for which factorization is positive to be smaller for a higher level of extension. This is not strictly true. In one test problem the factorization with $G(1)$ was positive for $\alpha = 0$ while the factorization with $G(2)$ was not.

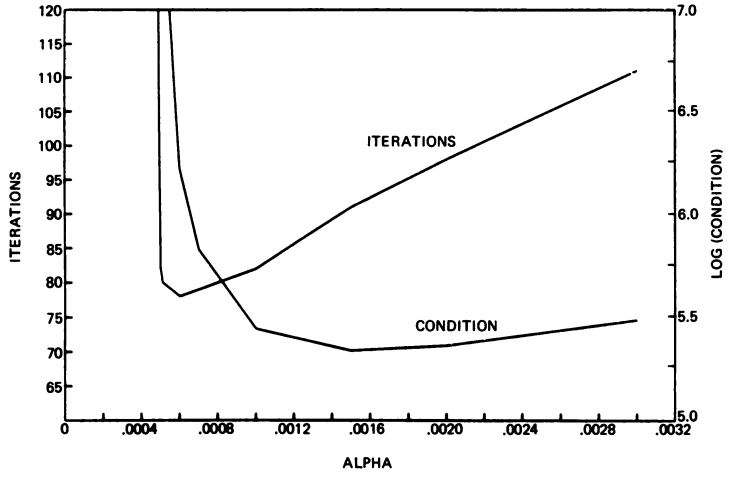

FIGURE 6.10

Iterations to convergence and condition for extended nonzero set $G(2)$

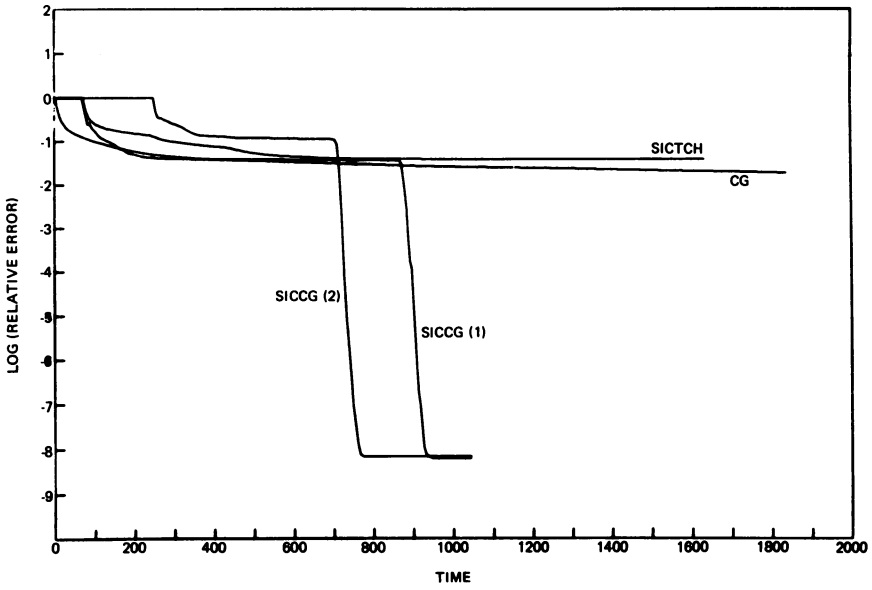

FIGURE 6.11

Method comparison

(e) *Comparison to Other Methods.* In order to compare this method with other iterative methods, the test problem was run using straight conjugate gradients of the scaled matrix (CG) and a Tchebychev acceleration of the shifted incomplete Cholesky factorization (SICTCH) with dynamic estimation of parameters (Manteuffel [15]). Figure 6.11 shows the results. This figure takes into account the differences in time required to perform the various iterations.

7. **Conclusion.** The shifted incomplete factorization provides a method for splitting any symmetric positive definite matrix and accelerating by a conjugate gradient iteration. We have seen in Section 5 that this splitting is at least as good as the Jacobi splitting.

The SIC splitting has an advantage over the SIP splitting (Stone [24]) in that the matrix need not be a 5-point difference matrix. It has an advantage over the SSOR splitting (Axelsson [1]) in that one can extend the graph-set to higher levels and bring a more accurate approximate inverse into play. In fact, one can readily move from the extreme of no splitting to the extreme of complete factorization with iterative refinement by adjusting the graph-set.

The results on the motivating problem showed that significant savings in both work and storage can be realized over direct methods on large problems. This savings should increase for even larger problems. The results also showed that with nonzero set $G(1)$ or $G(2)$ the factorization was inexpensive compared to the total solution time. Thus, a trial and error approach for finding an acceptable value of α as described in Section 6 is feasible. Hopefully, further analysis will reveal an a priori estimate of α_b for certain classes of matrices.

Applied Mathematics Division
Sandia Laboratories
Livermore, California 94550

1. O. AXELSSON, *On Preconditioning and Convergence Acceleration in Sparse Matrix Problems*, Report CERN 74-10 of the CERN European Organization for Nuclear Research, Data Handling Division, Laboratory I, 8 May, 1974.

2. P. CONCUS, G. H. GOLUB & D. P. O'LEARY, *A Generalized Conjugate Gradient Method for the Numerical Solution of Elliptic Partial Differential Equations*, Lawrence Berkeley Laboratory Publ. LBL-4604, Berkeley, Calif., 1975.

3. J. W. DANIEL, "The conjugate-gradient method for linear and non-linear operator equations," *SIAM J. Numer. Anal.*, v. 4, 1967, pp. 10–26.

4. M. ENGELI, T. GINSBURG, H. RUTISHAUSER & E. STIEFEL, *Refined Iterative Methods for Computation of the Solution and the Eigenvalues of Self-Adjoint Boundary Value Problems*, Birkhäuser Verlag, Basel/Stuttgart, 1959.

5. D. K. FADDEEV & V. N. FADDEEVA, *Computational Methods of Linear Algebra*, Freeman, San Francisco, Calif., and London, 1963.

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

7. G. FORSYTHE & E. G. STRAUS, "On best conditioned matrices," *Proc. Amer. Math. Soc.*, v. 6, 1955, pp. 340–345.

8. A. GREENBAUM, *Comparison of Splittings Used With the Conjugate Gradient Method*, Lawrence Livermore Laboratories Report UCRL-80800, Livermore, Calif., March 1978.

9. M. HESTENES & E. STIEFEL, "Methods of conjugate gradients for solving linear systems," *J. Res. Nat. Bur. Standards*, v. 49, 1952, pp. 409–436.

10. M. R. HESTENES, "The conjugate-gradient method for solving linear systems," *Proc. Sympos. Appl. Math.*, vol. 6, *Numerical Analysis*, McGraw-Hill, New York, 1956, pp. 83–102.

11. A. S. HOUSEHOLDER, *The Theory of Matrices in Numerical Analysis*, Blaisdell, New York, 1964.

12. S. KANIEL, "Estimates for some computational techniques in linear algebra," *Math. Comp.*, v. 20, 1966, pp. 369–378.

13. D. S. KERSHAW, *The Incomplete Cholesky-Conjugate Gradient Method for the Iterative Solution of Systems of Linear Equations*, Lawrence Livermore Laboratory Report UCRL-78333, Livermore, Calif., 1976.

14. C. LANCZOS, "An iteration method for the solution of the eigenvalue problem of linear differential and integral operators," *J. Res. Nat. Bur. Standards*, v. 45, 1950, pp. 255–282.

15. T. A. MANTEUFFEL, "The Tchebychev iteration for nonsymmetric linear systems," *Numer. Math.*, v. 28, 1977, pp. 307–327.

16. T. A. MANTEUFFEL, *The Shifted Incomplete Cholesky Factorization*, Sandia Laboratories Report SAND 78-8226, May 1978.

17. J. A. MEIJERINK & H. A. VAN DER VORST, "An iterative solution method for linear systems of which the coefficient matrix is a symmetric M-matrix," *Math. Comp.*, v. 31, 1977, pp. 148–162.

18. D. O'LEARY, *Hybrid Conjugate Gradient Algorithms*, Ph.D. Thesis, Computer Science Dept., Stanford Univ., 1975.

19. C. C. PAIGE, *The Computation of Eigenvalues and Eigenvectors of Very Large Sparse Matrices*, Ph.D. Thesis, London Univ., Institute of Computer Science, 1971.

20. C. C. PAIGE, "Computational variants of the Lanczos method for the eigenproblem," *J. Inst. Math. Appl.*, v. 10, 1972, pp. 373–381.

21. C. C. PAIGE & M. A. SAUNDERS, *Solution of Sparse Indefinite Systems of Equations and Least Squares Problems*, Standard Reports, STAN-CS-73-399, Nov. 1973.

22. J. K. REID, "On the method of conjugate gradients for the solution of large sparse systems of linear equations," *Proc. Conference on "Large Sparse Sets of Linear Equations"*, Academic Press, New York, 1971.

23. G. W. STEWART, *Introduction to Matrix Computation*, Academic Press, New York, 1973.

24. H. L. STONE, "Iterative solutions of implicit approximations of multidimensional partial differential equations," *SIAM J. Numer. Anal.*, v. 5, 1968, p. 530.

25. R. S. VARGA, "Factorization and normalized iterative methods," *Boundary Problems in Differential Equations* (R. E. Langer, Ed.), Univ. of Wisconsin Press, Madison, 1960.

26. R. S. VARGA, *Matrix Iterative Analysis*, Prentice-Hall, Englewood Cliffs, N. J., 1962.

27. D. YOUNG, *Iterative Solution of Large Linear Systems*, Academic Press, New York, 1972.