

Lecture Notes in Mathematics

Edited by A. Dold, B. Eckmann and F. Takens

1457

*G. NOTAY, Solving
positive (semi)definite
linear systems by
preconditioned iterative
methods, pp. 105-125*

O. Axelsson L.Yu. Kolotilina (Eds.)

Preconditioned Conjugate Gradient Methods

Proceedings, Nijmegen 1989



Springer-Verlag

Table of Contents

Preface

Submitted papers

1 <i>Modified incomplete factorization strategies,</i> Robert Beauwens	1
2 <i>On some parallel preconditioned CG schemes,</i> R. Bramley, H.-C. Chen, U. Meier and A. Sameh	17
3 <i>Preconditioning indefinite systems arising from mixed finite element discretization of second-order elliptic problems,</i> R.E. Ewing, R.D. Lazarov, Peng Lu and PS. Vassilevski	28
4 <i>A class of preconditioned conjugate gradient methods applied to finite element equations – a survey on MIC methods,</i> Ivar Gustafsson	44
5 <i>Recent vectorization and parallelization of ITPACKV software package,</i> David R. Kincaid and Thomas C. Oppe	58
6 <i>On the sparsity patterns of hierarchical finite element matrices,</i> Jos Maubach	79
7 <i>Solving positive (semi) definite linear systems by preconditioned iterative methods,</i> Yvan Notay	105
8 <i>The convergence behaviour of preconditioned CG and CG-S,</i> H.A. van der Vorst	126
9 <i>Data reduction preconditioning for generalized conjugate gradient methods,</i> R. Weiss and A. Schonauer	137
10 <i>Analysis of a recursive 5-point / 9-point factorization method,</i> O. Axelsson and V. Eijkhout	154
11 <i>Iteration methods as discretization procedures,</i> O. Axelsson and W. Layton	174
List of speakers	195

SOLVING POSITIVE (SEMI)DEFINITE LINEAR SYSTEMS BY PRECONDITIONED ITERATIVE METHODS

Y. NOTAY *

Abstract. The use of preconditionings obtained by so called modified incomplete factorizations has become quite popular for the PCG solution of regular systems arising from the discretization of elliptic PDE's. Our purpose here is to review their recent extension to the singular case. Because such conditionings may themselves be singular, we first review the extension of the general theory of polynomial acceleration to the case of singular preconditionings. We emphasize that all results can be formulated in such a way that they cover both the regular and singular cases. Examples of application are given, displaying the superiority of the recently developed factorization strategies.

Key Words. Iterative methods for linear systems, acceleration of convergence, preconditioning.

1. Introduction. Let A be a symmetric positive (semi)definite $n \times n$ matrix and consider the consistent system

$$(1.1) \quad Ax = b$$

with $b \in \mathcal{R}(A)$ (see below for notation). In the regular case, the PCG method with preconditioning arising from modified incomplete factorizations has become a popular technique and our purpose here is to review its recent extension to the singular case. Such an extension was not straightforward because the associated preconditioning matrix may itself be singular and no theory of singularly preconditioned polynomially accelerated iterative schemes had been developed. It has therefore been performed in two stages, both of which will be reviewed here : (1) the extension of the general theory of preconditioned polynomial iterative methods to the case of possibly singular preconditionings, cf. [23], and (2) the more specific theory of modified incomplete factorizations of possibly singular Stieltjes matrices, cf. [24]. We shall additionally stress that all results can be formulated in such a way that they cover both the regular and singular cases with, in the latter event, both regular and singular preconditioners.

We are therefore first interested here in preconditioned (polynomial) iterative methods for solving (1.1). That is we first make some splitting of the system matrix

$$(1.2) \quad A = B - C$$

where the "preconditioner" or "preconditioning matrix" B is symmetric nonnegative definite with

$$(1.3) \quad \mathcal{N}(B) \subset \mathcal{N}(A)$$

Then, (1.1) is solved by an iterative scheme of the following form, starting with an arbitrary initial approximation x_0 :

* Université Libre de Bruxelles, Service de Métrologie Nucléaire, 50, av. F.D. Roosevelt, B-1050 Brussels, Belgium. Author's research is supported by the "Fonds National de la Recherche Scientifique" (Belgium) - Aspirant.

For $k = 0, 1, \dots$ until convergence do

$$(1.4) \quad \text{Solve} \quad Bg_k = b - Ax_k$$

$$(1.5) \quad \delta_k = g_k + \sum_{\substack{i=0 \\ k>i}}^{s-2} d_k^i \delta_{k-i-1}$$

$$(1.6) \quad x_{k+1} = x_k + a_k \delta_k$$

where $s > 0$ is the order of the method ($\delta_k = g_k$ for first order schemes) and where the parameters a_k and d_k^i are determined by the choice of an actual method. It should be noticed that the system (1.4) to be solved at each step is consistent if and only if $\mathcal{R}(A) \subset \mathcal{R}(B)$, i.e. because symmetry, if and only if $\mathcal{N}(B) \subset \mathcal{N}(A)$, giving rise to the condition (1.3). This remark also shows that, in the singular case, there is no reason to enforce $\mathcal{N}(B) = \{0\}$, i.e. to consider only regular preconditionings. We shall actually see in the following that all classical iterative schemes used in the positive definite case do extend to the semidefinite case, with both singular and nonsingular preconditionings.

On the other hand, regarding preconditioning methods by modified incomplete factorizations, the algebraic framework of the regular theory, cfr [3], [4], [6], [7], [8], [9], [10], [11], [12], [16], assumes that A is a nonsingular Stieltjes matrix (i.e. a positive definite matrix with nonpositive offdiagonal entries) and we shall show how this framework extends to the case of a singular Stieltjes matrix (i.e. a positive semidefinite matrix with nonpositive offdiagonal entries).

It will further follow from the general theory of polynomial acceleration that the spectral reduction technique (called spectral equivalence in the case of discrete PDE's) used in the regular case to reduce a general symmetric positive definite matrix A to Stieltjes one, also extends to the singular case.

General terminology and notation. All vectors belong to \mathcal{C}^n ; all matrices are $n \times n$ real matrices.

The symbol A^t , A^+ , $\mathcal{N}(A)$, $\mathcal{R}(A)$, $\sigma(A)$ and $\rho(A)$ denote, respectively, the transpose, the Moore-Penrose inverse, the null space, the range, the spectrum and the spectral radius of the matrix A .

By $P_{M,L}$ we denote the projector with null space L and range M (this notation implying that L and M are complementary subspaces).

If A is an $n \times n$ matrix and T a subspace of \mathcal{C}^n , we denote by $A|_T$ the linear operator in \mathcal{C}^n defined as the restriction of A to T .

The order relation between real matrices and vectors is the usual componentwise order: if $A = (a_{ij})$ and $B = (b_{ij})$ then $A \leq B$ ($A < B$) if $a_{ij} \leq b_{ij}$ ($a_{ij} < b_{ij}$) for all i, j ; A is called nonnegative (positive) if $A \geq 0$ ($A > 0$). If $A = (a_{ij})$, we denote by $diag(A)$ the (diagonal) matrix whose entries are $a_{ii}\delta_{ij}$ and we let $offdiag(A) = A - diag(A)$. By e we denote the vector with all components equal to unity; by a "(0,1) matrix", we understand a matrix whose nonzero entries are equal to unity.

We shall also need a few graph concepts; we refer to [15], [18] for general terminology about matrix graphs with the warning that, unless otherwise stated, all graphs considered here are ordered undirected graphs with node set $[1, n]$, i.e. the ordered set of the first n integers or (when subgraphs are considered) some subset of $[1, n]$. In addition, we introduce the following more specific graph notions.

DEFINITION 1.1. An increasing path in a graph is a path $i_0, i_1, i_2, \dots, i_t$ such that $i_0 < i_1 < i_2 < \dots < i_t$.

DEFINITION 1.2. A node k of a graph G is called a precursor (successor) of the node i of G if $\{i, k\}$ belongs to the edge set of G with $k < i$ ($k > i$). The set of precursors (successors) of i is denoted by $P(i)$ ($S(i)$).

DEFINITION 1.3. For any node i of a graph G , we define the ascent $As(i)$ of i as

$$As(i) = \{k \mid \text{There exists an increasing path from } k \text{ to } i\}.$$

It should be noticed that $i \in As(i)$ because a path of zero length is an increasing path.

DEFINITION 1.4. For any pair of nodes i and j of a graph G , we denote by

$$Pc(i, j) = P(i) \cap P(j)$$

their set of common precursors; we further define

$$Pc(G) = \bigcup_{\substack{i, j=1 \\ i \neq j}}^n Pc(i, j)$$

If G is the graph of an $n \times n$ matrix A , we also write $Pc(A)$ for $Pc(G)$.

2. Iterative schemes with possibly singular preconditionings.

2.1. Convergence analysis. For covering the singular case, we need to avoid assuming the matrices are invertible. The key of our ability to extend the regular theory lies thus in the use of generalized inverses. Some of their properties, needed for a good understanding of the paper, are recalled in the text. We refer to [13] for a more detailed exposition.

First, we assume in this paper that the operation (1.4) is achieved by the application of a linear operator, that is, there exists some matrix, say X , such that (1.4) reduces in practice to

$$(2.1) \quad g_k = X(b - Ax_k)$$

(as usual, it does not mean that X has to be computed explicitly). Now, $g = Xy$ is a solution of $Bg = y$ for all $y \in \mathcal{R}(B)$ if and only if $BXy = y$ for all $y \in \mathcal{R}(B)$, i.e. if and only if

$$(2.2) \quad B X B = B$$

that is, by definition, if and only if X is $\{1\}$ -inverse of B . It follows then [13] that there exists some subspace S complementary to $\mathcal{N}(B)$ such that

$$(2.3) \quad X B = P_{S, \mathcal{N}(B)}$$

which means that for any $y \in \mathcal{R}(B)$, $g = Xy$ is the unique solution to $Bg = y$ that belongs to S . Finally, it turns out from these remarks that, for all $y \in \mathcal{R}(B)$

$$(2.4) \quad Xy = P_{S, \mathcal{N}(B)} B^{(1)}y$$

where $B^{(1)}$ denotes any (other) $\{1\}$ -inverse of B .

The Moore-Penrose inverse B^+ of a matrix B is a particular $\{1\}$ -inverse which verifies some additional properties. When B is symmetric, it is symmetric and can be characterized by its eigenvectors and eigenvalues which are such that

$$(2.5) \quad B^+x = \lambda^+ x \Leftrightarrow Bx = \lambda x$$

with

$$(2.6) \quad \lambda^+ = \begin{cases} \lambda^{-1} & \text{if } \lambda \neq 0 \\ 0 & \text{otherwise} \end{cases}$$

Using (2.4) with $B^{(1)} = B^+$, we can rewrite the iterative scheme (1.4-6)

$$(2.7) \quad g_k = P_{S, \mathcal{N}(B)} B^+(b - Ax_k)$$

$$(2.8) \quad \delta_k = g_k + \sum_{\substack{i=0 \\ k>i}}^{k-2} d_k^i \delta_{k-i-1}$$

$$(2.9) \quad x_{k+1} = x_k + a_k \delta_k \quad k = 0, 1, 2, \dots$$

Further, it is of interest to also include in our analysis the schemes in which the solution is at each step projected in a prescribed subspace T complementary to $\mathcal{N}(A)$, i.e. the schemes where the following is used rather than (2.7) :

$$(2.10) \quad g'_k = P_{S, \mathcal{N}(B)} B^+(b - Ax_k) \quad ; \quad g_k = P_{T, \mathcal{N}(A)} g'_k$$

Finally, because (2.7), (2.10) are particular cases of

$$(2.11) \quad g_k = Q B^+(b - Ax_k)$$

where Q is some projector such that

$$(2.12) \quad \mathcal{N}(B) \subset \mathcal{N}(Q) \subset \mathcal{N}(A),$$

this formulation will be used from now on.

The scheme (2.11,8,9) is actually a polynomial acceleration of the basic unaccelerated iterative scheme

$$(2.13) \quad v_{k+1} = v_k + Q B^+(b - Av_k) \quad k = 0, 1, 2, \dots$$

and polynomially accelerated iterative schemes in the positive (semi)definite case with possibly singular preconditioning are studied in detail in [23]; we shall only briefly recall here the main issues of this work.

First, letting (x_k) be the sequence generated by (2.11,8,9) with given x_0 , defining

$$(2.14) \quad V = \mathcal{R}(B^+ A) \quad \text{and} \quad \tilde{Q} = P_{V, \mathcal{N}(A)}$$

and letting $\tilde{x}_k = \tilde{Q} x_k$ for all k , it is stated that, for any (semi) norm $|\cdot|$ whose kernel is $\mathcal{N}(A)$:

$$(2.15) \quad |x_k - x| = |\tilde{x}_k - \tilde{x}|$$

for all x, \tilde{x} such that $\tilde{x} = \tilde{Q} x$. Whence, the convergence analysis of the sequence (x_k) may equivalently be made on the sequence (\tilde{x}) . It is further shown that the latter is identical to the sequence obtained when using the same iterative method (i.e. the same s, d_k^i, a_k) with starting vector $\tilde{x}_0 = \tilde{Q} x_0$ in order to solve the regular system

$$(2.16) \quad A|_V x = b$$

with the regular preconditioning

$$(2.17) \quad (B^+ |_{\mathcal{R}(A)})^{-1} = B|_V$$

With the additional result

$$(2.18) \quad \sigma((B|_V)^{-1} A|_V) = \sigma(B^+ A) \setminus \{0\},$$

we reach the conclusion that all classical methods used in the regular case for the determination of the iteration parameters d_k^i, a_k apply to our general iterative scheme (2.11,8,9), provided that "the extremal values of $\sigma(B^{-1} A)$ " is understood as "the extremal values of $\sigma(B^+ A) \setminus \{0\}$ ". By way of conclusion, we find of some interest to mention here that this reduction to the regular case is convenient but however not necessary for proving our results. Indeed, using generalized inverses in the classical proofs for the regular case, we can cover directly both the regular and singular cases. The interested reader may refer to [22].

2.2. Practical schemes. We have already mentioned that all classical methods apply to our more general feature without any change except in the way of defining the extremal eigenvalues associated with the preconditioning. We shall therefore not review them, referring the reader to the litterature [1], [3], [17], [20]. A brief summary in which the singular case is included is also presented in [23].

We however reach some additional results for the following methods.

The extrapolation method. It is a first ($s = 1$) order scheme. For such a scheme, letting

$$P_k(\nu) = \prod_{i=0}^{k-1} (1 - a_i \nu) \quad \text{for } k \geq 1$$

$$M_k = \max_{\nu \in \sigma(B^+A) \setminus \{0\}} |p_k(\nu)|$$

the error evolution formula is given by

$$\|x - x_k\|_A \leq M_k \|x - x_0\|_A$$

where $\|\cdot\|_A = \sqrt{(\cdot, A)}$ denotes the A -(semi)norm and x any solution to (1.1). The extrapolation method is associated with $a_k = \tau$ for all $k \geq 0$, the basic unaccelerated iterative method being included with $\tau = 1$. Letting ν_{min} and ν_{max} denote respectively the smallest and the largest value of $\sigma(B^+A) \setminus \{0\}$, the method is convergent if $\tau < 2/\nu_{max}$ with then $M_k = [\max(|1 - \tau\nu_{min}|, |\tau\nu_{max} - 1|)]^k$. The optimum is thus reached for $\tau = 2/(\nu_{max} + \nu_{min})$, which leads to $M_k = [(\nu_{max} - \nu_{min})/(\nu_{max} + \nu_{min})]^k$.

When A is a regular Stieltjes matrix and $B = \text{diag}(A)$ (i.e. the point Jacobi preconditioner), the basic unaccelerated method ($\tau = 1$) is convergent (see [27]). Further, if the system to solve is the 5-point finite difference approximation of a second order elliptic PDE, $I - B^{-1}A$ is cyclic of index 2 [27] and therefore $\nu_{max} = 2 - \nu_{min}$, showing that the optimum is effectively reached for $\tau = 1$. On the other hand, when the PDE is a pure Neumann problem, A is (with the same 5-point finite difference discretization) a singular Stieltjes matrix (see definition below). With $B = \text{diag}(A)$, we have still that $I - B^{-1}A$ is cyclic of index 2, whence $\nu_{max} = 2$ and the unaccelerated method is no more convergent. If, however, one uses $a_0 = 1/2$ with $a_k = 1$ for $k \geq 1$, then $p_k(\nu) = (1 - \nu)^{k-1}(1 - \nu/2)$, showing that this method is convergent if and only if $\nu_{max} \leq 2$ with $M_k = [\max(|1 - \nu_{min}|, |\nu'_{max} - 1|)]^{k-1}$, where ν'_{max} denotes the maximal value of $\sigma(B^+A) \setminus \{2\}$. Moreover our assumptions imply $\nu'_{max} = 2 - \nu_{min}$, so that $M_k = [(\nu'_{max} - \nu_{min})/(\nu'_{max} + \nu_{min})]^{k-1}$, showing that the convergence rate is then about the same as in the regular case. The same remarks also hold for the block Jacobi preconditioner.

The steepest descent and the conjugate gradient method. These respectively first ($s = 1$) and second ($s = 2$) order methods present the particular feature that the coefficients a_k and d_k^c are computed during the iteration process according expressions which involve the vectors x_k and δ_k . Therefore, for that the convergence result given above applies, the coefficients are to be computed, in the singular case, not directly with the vectors x_k, δ_k , but with the vectors of the corresponding regular method which solve the regular system (2.16) with regular preconditioning (2.17). However, it is obtained in [23] that we may equivalently use the classical expressions with the vectors x_k, δ_k as they appear in the iterative scheme (2.11,8,9). Namely, for the conjugate gradient method, using

$$(2.19) \quad a_k = \frac{(b - Ax_k, \delta_k)}{(\delta_k, A\delta_k)}, \quad d_k^c = \frac{(b - Ax_k, g_k)}{(b - Ax_{k-1}, g_{k-1})}$$

we obtain the coefficients which minimize the A -(semi)norm $\|\cdot\|_A$ of the error among all the schemes (2.11,8,9). One can find in the literature ([5],[26]) some developments

about the convergence rate of this method in connection with the eigenvalue distribution. It turns out that all these developments extend to the singular case, provided that "the eigenvalue distribution in $\sigma(B^{-1}A)$ " is understood as "the eigenvalue distribution in $\sigma(B^+A) \setminus \{0\}$ ". Note that (2.19) allows the same implementation techniques as in the regular case. A proof of the optimality properties that covers directly both the regular and singular cases can be found in [22].

2.3. Numerical stability. For studying stability, we have to consider, rather than (2.11,8,9)

$$(2.20) \quad g_k = QB^+(b - Ax_k)$$

$$(2.21) \quad \delta_k = g_k + \sum_{\substack{i=0 \\ k>i}}^{s-2} d_k^i \delta_{k-i-1}$$

$$(2.22) \quad x_{k+1} = x_k + a_k \delta_k + \varepsilon_k \quad k = 0, 1, 2, \dots$$

where ε_k , assumed to be small, also synthesises the perturbations that may appear in (2.20) or (2.21). Now defining the sequence (\tilde{x}_k) by $\tilde{x}_k = \tilde{Q} x_k$ for all k (where \tilde{Q} is defined by (2.14)), it follows from the above mentioned results of [23] that (\tilde{x}_k) may be viewed as the sequence resulting from the perturbed iterative solution of the regular system $A|_V x = b$ with regular preconditioning $B|_V$. Thus, the stability analysis of an iterative method in the regular case applies to the sequence (\tilde{x}_k) . Further, one easily sees with Lemma 4.1 of [23] that for all $k \geq 1$

$$(2.23) \quad x_k = Q \tilde{x}_k + (I - Q \tilde{Q})x_o + \sum_{i=0}^{k-1} (I - Q \tilde{Q})\varepsilon_i$$

showing that (2.20-22) (with ε_k assumed small) will be stable if and only if the corresponding regular method is. The stability analyses made for the regular case extend thus to the singular case. (One easily verifies in addition that the particular implementation we suggest above for the extrapolation method in the singular case is stable because $|1 - a_k \nu| \leq 1$ for all $\nu \in \sigma(B^+A)$).

Besides these general aspects, Kaaschieter has pointed out in [19] that, in the singular case, some particular stability problems may appear when using the steepest descent or the conjugate gradient method. Indeed, we have to care that the condition $b \in \mathcal{R}(A)$ is not achieved in practice, due to roundoff errors. Generally, it is not cumbersome, because a small perturbation to b leads to a small term ε_k in (2.22), and if the iterative method used is stable, there will be no further problem. But, when using the steepest descent or the conjugate gradient method, we have to take into account that a_k is computed with the formula (2.19) in which b is involved. In order to see what may happen, let $b = b_o + \delta_b$, where $b_o \in \mathcal{R}(A)$ and $\delta_b \in \mathcal{N}(A)$, and let x be a solution to $Ax = b_o$. Then, for some given x_k and δ_k , we have

$$|x_k + a_k \delta_k - x|_A^2 = |x_k - x|_A^2 + a_k^2 |\delta_k|_A^2 + 2a_k(\delta_k, A(x_k - x))$$

so that we will have $|x_k + a_k \delta_k - x|_A \leq |x_k - x|_A$ if and only if

$$0 \leq a_k \leq \frac{2(\delta_k, A(x - x_k))}{(\delta_k, A\delta_k)} = \frac{2(\delta_k, b_o - Ax_k)}{(\delta_k, A\delta_k)}$$

or

$$0 \geq a_k \geq \frac{2(\delta_k, A(x - x_k))}{(\delta_k, A\delta_k)} = \frac{2(\delta_k, b_o - Ax_k)}{(\delta_k, A\delta_k)}$$

With $a_k = (\delta_k, b - Ax_k)/(\delta_k, A\delta_k)$ as in (2.19), we see that this condition will always be satisfied when $b = b_o$, while otherwise it is equivalent to

$$(2.24) \quad |(\delta_k, \delta_b)| \leq |(\delta_k, b_o - Ax_k)|$$

When $b_o - Ax_k$ becomes small, (2.24) may be no more satisfied, so that the iterative process starts diverging, as observed in [19]. However, if δ_b is only due to the roundoff errors (as it is guaranteed by projecting the right hand side on the range of A), the stopping criterion will usually be met far before $\|b_o - Ax_k\| \approx \|\delta_b\|$, so that (2.24) will always be satisfied. On the other hand, if one wants to compute the solution with great accuracy, one can prevent problems by using (2.10) with $T = \mathcal{R}(A)$, so that $\delta_k \in \mathcal{R}(A)$ for all $k \geq 0$. Then, since $\delta_b \in \mathcal{N}(A)$, we will have $|(\delta_k, \delta_b)| = 0$ within the roundoffs errors, so that (2.24) will hold even when $\|b_o - Ax_k\| \approx \|\delta_b\|$.

2.4. Conclusion. It follows from the results of this section that all iterative methods effective in the regular case can be used in the singular case, with both regular and singular preconditionings. Moreover, in the case of singular preconditioning, the convergence properties of the scheme are completely independent of the choice of the generalized inverse, i.e. on the way chosen for achieving (1.4). The convergence results can be expressed as in the regular case, provided that "the eigenvalue distribution in $\sigma(B^+A) \setminus \{0\}$ " is used as the right extension of "the eigenvalue distribution in $\sigma(B^{-1}A)$ ". This gives rise to the following generalized definition for the spectral condition number :

$$(2.25) \quad \kappa(B^+A) = \frac{\nu_{\max}(B^+A)}{\nu_{\min}(B^+A)}$$

with

$$(2.26) \quad \nu_{\max}(B^+A) = \max_{\nu \in \sigma(B^+A)} \nu, \quad \nu_{\min}(B^+A) = \max_{\substack{\nu \in \sigma(B^+A) \\ \nu \neq 0}} \nu$$

Finally, it is important to recall here the expressions obtained in [23] for ν_{\max} and ν_{\min} .

THEOREM 2.1. Let A, B be symmetric nonnegative definite with $\mathcal{N}(B) \subset \mathcal{N}(A)$, and let $\nu_{\max}(B^+A)$, $\nu_{\min}(B^+A)$ be given by (2.26). We have

$$(2.27) \quad \nu_{\max} = \max_{\substack{z \in \mathbb{C}^n \\ z \notin \mathcal{N}(B)}} \frac{(z, Az)}{(z, Bz)}$$

$$\begin{aligned}
 \nu_{\min}(B^+A) &= \min_{\substack{z \in \mathcal{R}(B^+A) \\ z \neq 0}} \frac{(z, Az)}{(z, Bz)} \\
 (2.28) \qquad &= \max_S \min_{\substack{z \in S \\ z \neq 0}} \frac{(z, Az)}{(z, Bz)} \\
 &\qquad S \oplus \mathcal{N}(A) = \mathcal{C}^n
 \end{aligned}$$

Most results about spectral bounds are obtained, in the regular case, by proving inequalities of the type $(z, Az) \leq c_1(z, Bz)$ or $(z, Az) \geq c_2(\lambda_1)(z, Bz)$ for all $z \in \mathcal{C}^n$, where λ_1 is the first eigenvalue of $D^{-1}A$, with $D = \text{diag}(A)$; see [3] for examples. It follows from Theorem 2.1 that the first type of inequality will give also an upper bound in the singular case, while, as will be seen below, the inequalities of the second type can generally be rewritten $(z, Az) \geq c_2(\lambda_{\min})(z, Bz)$, where λ_{\min} is the first nonzero eigenvalue of $D^{-1}A$, so that they give also a lower bound in the singular case.

Therefore, many regular conditioning analysis results can be rewritten with little handling so as to cover both the regular and singular cases. However, regarding the modified incomplete factorization methods, the generalization is not straightforward because the classical existence theorems do not work in the singular case. We summarize in the next section our new existence analysis that allows covering the singular case and, in Section 4, our results on conditioning analysis and factorization strategies.

3. Modified incomplete factorizations of Stieltjes matrices. In the regular case, the basic framework of the modified incomplete factorization methods assumes that the system matrix is a (regular) Stieltjes matrix. We first need to appropriately extend this notion to the singular case. To this aim, we use the following definition.

DEFINITION 3.1. *A real square matrix A is called an M -matrix if there exists a non negative number t such that*

$$tI - A \geq 0 \qquad \text{with} \qquad \rho(tI - A) \leq t;$$

a symmetric M -matrix is called a Stieltjes matrix.

It follows from Definition 3.1 that a Stieltjes matrix has nonpositive offdiagonal entries and that a symmetric matrix with non positive offdiagonal entries is a Stieltjes matrix if and only if it is nonnegative definite. The latter property has already been assumed in Section 1. The restriction concerning the offdiagonal entries is thus the only one we have to introduce from now on. However, it should be mentioned here that the modified incomplete factorization methods have actually a much wider scope of application. Indeed, the spectral reduction technique used in the regular case, called spectral equivalence in discrete PDE's applications ([3], [16]), readily extends to the semidefinite case. With this technique, an arbitrary symmetric positive (semi)definite matrix A may be reduced to the Stieltjes case provided that one can determine a Stieltjes matrix A_0 and positive numbers α, β with

$$(3.1) \qquad \alpha(z, A_0 z) \leq (z, Az) \leq \beta(z, A_0 z)$$

for all $z \in \mathbb{C}^n$; then, we have obviously $\mathcal{N}(A) = \mathcal{N}(A_0)$ and

$$\kappa(B^+ A) \leq \frac{\beta}{\alpha} \kappa(B^+ A_0)$$

readily follows from Theorem 2.1 for any preconditioner B such that $\mathcal{N}(B) \subset \mathcal{N}(A_0)$. Note that in [3], [16], inequalities of the type (3.1) are derived for finite element matrices by analysing element stiffness matrices. The latter are singular so that such analyses do apply to the singular case without any modification. The Stieltjes theory allows by this way to cover most discrete PDE's applications. Its extension to the singular case means, with this respect, that the pure Neumann problems are no more excluded.

The modified incomplete factorizations of a Stieltjes matrix A are based, in the regular case, upon the existence for such matrices of a positive vector x such that $Ax \geq 0$. The following theorem extends this result to the singular case (see [14] for a proof).

THEOREM 3.2. *Let $A = (a_{ij})$ be a Stieltjes matrix. Then, there exists a positive vector x such that $Ax \geq 0$. Further*

- (1) *If A is regular : $\exists x > 0 : \sum_{j=1}^i a_{ij} x_j > 0$ for all i*
- (2) *If A is irreducible and singular :*
 - a) $\exists x > 0 : \mathcal{N}(A) = \text{Span}\{x\}$
 - b) $\forall x : Ax \geq 0 \Rightarrow Ax = 0$

We recall now from [24] our definition of modified incomplete factorizations.

DEFINITION 3.3. *Let $A = (a_{ij})$ be a $n \times n$ Stieltjes matrix and let $x > 0$ be such that $Ax \geq 0$; let $\Lambda = (\lambda_i \delta_{ij})$ be a nonnegative diagonal matrix and $\beta = (\beta_{ij})$ a $(0,1)$ matrix; let $U = (u_{ij})$ be the $n \times n$ upper triangular matrix defined by the following algorithm : for $i=1, \dots, n$ set*

$$u_{ij} = a_{ij} - \beta_{ij} \sum_{k<i} u_{ki} u_{kk}^+ u_{kj} \quad \text{for } i < j \leq n$$

$$\begin{cases} (Ux)_i &= (Ax)_i + \lambda_i a_{ii} x_i - \sum_{k<i} u_{ki} u_{kk}^+ (Ux)_k \\ u_{ii} &= ((Ux)_i - \sum_{j>i} u_{ij} x_j) / x_i \end{cases}$$

where u_{kk}^+ is defined by

$$u_{kk}^+ = \begin{cases} u_{kk}^{-1} & \text{if } u_{kk} \neq 0 \\ 0 & \text{if } u_{kk} = 0 \end{cases}$$

We say that U is the upper triangular factor and $P = \text{diag}(U)$ the diagonal factor of the modified incomplete factorization

$$B = U^t P^+ U$$

of A associated with x, Λ and β .

Note that by (2.5), (2.6) we have $P^+ = (u_{ii}^+ \delta_{ij})$. Λ is often referred as the perturbation term, the factorizations with $\Lambda = 0$ being called unperturbed factorizations.

The following theorem contains our existence analysis.

THEOREM 3.4. *Let A be an irreducible $n \times n$ Stieltjes matrix and x a positive vector such that $Ax \geq 0$. Let U be the upper triangular factor of the modified incomplete factorization $B = U^t P^+ U$ (with $P = \text{diag}(U)$) of A associated with x, Λ and β ,*

where Λ is a nonnegative diagonal matrix and β a $(0,1)$ matrix, and let $D = \text{diag}(A)$; then :

- (1) U is an upper triangular M -matrix with $Ux \geq Ax \geq 0$;
- (2) B is symmetric and nonnegative definite with $Bx = Ax + \Lambda Dx$;
- (3) $\mathcal{N}(B) = \mathcal{N}(U)$;
- (4) $\mathcal{N}(B) \subset \mathcal{N}(A)$ if and only if, in the graph of U :

$$\left\{ \begin{array}{l} \forall i < n : S(i) \neq \phi \\ \text{or} \\ \forall i : S(i) = \phi \Rightarrow \exists j \in As(i) : (Ax + \Lambda Dx)_j > 0; \end{array} \right.$$

- (5) if $\mathcal{N}(B) \subset \mathcal{N}(A)$:

$$\mathcal{N}(B) = \begin{cases} \mathcal{N}(A) & \text{if } \Lambda = 0 \\ \{0\} & \text{otherwise.} \end{cases}$$

Proof. Statements (1), (2), (3), (5) are proven in [24]. The first sufficient condition (4) is also proven in [24] while the second readily implies by induction that $u_{ii} > 0$ for all i , hence $\mathcal{N}(B) = \{0\}$. For proving the necessity of condition (4), note first that if the second criterion is not met, there exists some i with $u_{ii} = 0$, hence B is singular by (3). But by (5), we have that B singular is compatible with $\mathcal{N}(B) \subset \mathcal{N}(A)$ if and only if A is singular and $\Lambda = 0$, and in the latter case, the necessity of the condition $i < n \Rightarrow S(i) \neq \phi$ is proven in [24]. ■

Note that the classical existence criterion used generally in the regular case to ensure $\mathcal{N}(B) = \{0\}$ (that is the semi strict diagonal dominance criterion [11]) assumes

$$\sum_{j=1}^i a_{ij} x_j > 0,$$

which implies

$$\forall i : S(i) = \phi \Rightarrow (Ax)_i > 0$$

and appears therefore clearly to be included in the second condition of (4) since $i \in As(i)$ by definition. The same remark also holds for Gustafsson existence analysis which only considers the cases where $\Lambda Dx > 0$. However, both do not cover the case of A singular with $\Lambda = 0$, giving rise to our first criterion which is also useful in the regular case because there are no more restrictions on the entries of A nor on the perturbation Λ , all being replaced by a graph condition to be satisfied by U . Since the graph of U includes that of A , it is fulfilled by U when it is by A ; otherwise, it requires either some fill-in of the matrix U or an appropriate reordering of the system matrix; note with this respect that the ordering procedure described in [10] leads anyway to a system matrix satisfying $S(i) \neq \phi$ for all $i < n$. Finally, an alternate technique, introduced in [24], may be used to prove that $\mathcal{N}(B) \subset \mathcal{N}(A)$ which simply consists in checking

$$(z, Az) \leq c(z, Bz)$$

for all $z \in \mathcal{C}^n$. The existence analysis may then be viewed as a part of the conditioning analysis, the existence being guaranteed for any factorization algorithm or "strategy" for which an upper eigenvalue bound is obtained.

Definition 3.3 leaves open the choice of x, Λ and β . The choice of x is limited in the regular case by practical considerations, while there is no choice at all in the singular case. About β , its optimization would require to allow the fill-in of U up to the point where the reduction of the number of iterations is compensated by the increase of the algorithm computational complexity. The choice of β may also be limited by the memory requirements and the data structure one gives to the computer program. In practice, one uses either the "incomplete factorization by position", for which β is *a priori* fixed, or the "incomplete factorization by value" in which β is dynamically determined. Such aspects lie outside the scope of the present work, and we shall concentrate in the next section on the choice of Λ , assuming that x and β are given. It should be noted that this assumption is actually not incompatible with a dynamic determination of β , because the latter deals with the offdiagonal entries of the approximate factor U , while Λ plays only a role in the determination of its diagonal entries.

4. Factorization strategies and conditioning analysis. We extend here to the singular case the factorization algorithms (or "strategies") developed in [10]. We shall state their conditioning analysis together with their other properties referring to [21], [24], [25] for a detailed exposition of eigenvalue bounds which includes the singular case. It should be noticed that, with the results of these works and those of the preceding sections, most remarks and comments made in [10] for the regular case actually apply to our more general framework. We shall therefore give a somewhat abrupt exposition, stressing only on the particularities involved by the singular case. On the other hand, it clearly follows from [10] that the strategies No.2, 3 and 4 proposed there are not essentially different. For brevity, we shall therefore only generalize here the strategy No.4 (our strategy No.2).

In each case, $A = (a_{ij})$ is a Stieltjes matrix, x a given positive vector such that $Ax \geq 0$ and $\beta = (\beta_{ij})$ a given $(0,1)$ matrix.

Strategy No.1. It consists in choosing $\Lambda = 0$, i.e. that B is the modified incomplete factorization of A associated with $x, \Lambda = 0$ and β . (See Definition 3.3 for the factorization algorithm). This strategy is also referred to as the unperturbed strategy. Its properties are displayed in the following theorem (see [24] for a proof).

THEOREM 4.1. *Let A be an irreducible Stieltjes matrix and U the upper triangular factor of the modified incomplete factorization $B = U^t P^+ U$ (with $P = \text{diag}(U)$) of A associated with x, Λ and β , where x is a positive vector such that $Ax \geq 0, \beta$ a $(0,1)$ matrix and $\Lambda = 0$. Let $0 \leq \tau \leq 1$ be given by*

$$\tau = \begin{cases} \max_{i \in P_c(U)} & \frac{((P-U)x)_i}{(Px)_i} & \text{if } P_c(U) \neq \phi \\ 0 & \text{otherwise} \end{cases}$$

We have :

$$(1) \quad \text{If } \tau < 1,$$

$$(4.1) \quad \mathcal{N}(B) \subset \mathcal{N}(A)$$

with

$$(4.2) \quad \nu_{\max}(B^+ A) \leq \frac{1}{1 - \tau}$$

(2) If $\mathcal{N}(B) \subset \mathcal{N}(A)$

$$(4.3) \quad \mathcal{N}(B) = \mathcal{N}(A)$$

with

$$(4.4) \quad \nu_{\min}(B^+ A) \geq 1$$

(3) In particular, if A is singular :

a) $Bx = Ux = Ax = 0$

b) $\mathcal{N}(B) \subset \mathcal{N}(A)$ if and only if $i < n \Rightarrow S(i) \neq \emptyset$

Depending on the case at hand, the bound (4.2) may be satisfying or not (see [10]). When A is singular, it however clearly follows from (3) that $\tau = 1$ except in the trivial case where $Pc(U) = \emptyset$ and therefore $B = A$. The bound (4.2) is thus useless in the singular case. Now, the conditioning analysis of this strategy has been improved in [9], [10], giving an upper bound compatible with $\tau = 1$, the diagonal dominance requirement being replaced by a graph condition to be satisfied by the graph of U . This theory has been further improved and extended to the singular case in [25]. It is not possible to summarize briefly these results because they require the introduction of a somewhat involved formalism; we therefore refer the reader to the above quoted works.

On the other hand, it follows from Theorem 4.1 that this strategy leads in the singular case to a singular preconditioning. As shown in Section 2, this is not anyway disturbing providing that we can achieve the step (1.4) : "solve $Bg = y$ " for any $y \in \mathcal{R}(B)$ by the use of a linear operator. Now, letting \tilde{U} be obtained from U by "shifting" (in the singular case) its last diagonal entry up to an arbitrary positive number, and setting $\tilde{P} = \text{diag}(\tilde{U})$, it can be shown that when $\mathcal{N}(U) = \text{Span}\{x\}$, \tilde{U} , \tilde{P} and $\tilde{B} = \tilde{U}^t \tilde{P}^{-1} \tilde{U}$ are regular matrices such that $g = \tilde{B}^{-1} y$ provides a solution to $Bg = y$ for all $y \in \mathcal{R}(B)$. In other words \tilde{B}^{-1} is a $\{1\}$ -inverse of B (cfr. [24] for a proof). From a practical point of view, it should be stressed that this "shift" is the only modification to implement in a computer program (that uses the PCG algorithm with the unperturbed modified incomplete factorization) to get it working for singular problems too.

Now, even with the above mentioned improved conditioning analysis, the upper bound obtained for this strategy may still be unsatisfying, giving rise to the following strategy, in which an upper bound is anyway guaranteed, at the expense of a decrease of the smallest eigenvalue $\nu_{\min}(B^+ A)$.

Strategy No.2. It consists in computing the upper triangular matrix $U = (u_{ij})$ according to the following algorithm, where $0 < \tau < 1$ is an *a priori* chosen parameter.

ALGORITHM 1.

For $i = 1, \dots, n$ set :

$$u_{ij} = a_{ij} - \beta_{ij} \sum_{k \in Pc(i,j)} \frac{u_{ki} u_{kj}}{u_{kk}}, \quad j = i+1, \dots, n$$

$$u_{ii} x_i = \begin{cases} \max(\tau^{-1} \sum_{j>i} u_{ij} x_j, (Ax)_i - \sum_{j>i} u_{ij} x_j - \sum_{k \in Pc(i)} u_{ki} u_{kk}^{-1} (Ux)_k) & \text{if } i \in Pc(U) \\ (Ax)_i - \sum_{j>i} u_{ij} x_j - \sum_{k \in Pc(i)} u_{ki} u_{kk}^{-1} (Ux)_k & \text{if } i \notin Pc(U) \end{cases}$$

It turns out that this algorithm can never break down since we will have $u_{ii} > 0$ for all i such that $u_{ij} \neq 0$ for some $j > i$.

The preconditioner is then given by $B = U^t P^+ U$, where $P = \text{diag}(U)$. Its properties are stated in the following theorem (see [24] for a proof).

THEOREM 4.2. *Let A be an irreducible Stieltjes matrix, x a positive vector such that $Ax \geq 0$ and β a $(0,1)$ matrix. Let U be the upper triangular matrix computed according Algorithm 1, where $0 < \tau < 1$ is some given parameter. Set $P = \text{diag}(U)$ and $B = U^t P^+ U$. We have :*

- (1) *There exists a nonnegative diagonal matrix $\Lambda = (\lambda_i \delta_{ij})$ such that U is the upper triangular factor of the modified incomplete factorization B of A , with respect to x, Λ and β . Further,*

$$\lambda_i a_{ii} x_i \begin{cases} = 0 & \text{if } i \notin Pc(U) \\ \leq \max(0, \tau^{-1}(1-\tau)^2((P-U)x)_i + (1-\tau)((U^t-U)x)_i - (Ax)_i) & \text{if } i \in Pc(U) \text{ with } P(i) \subset Pc(U) \\ \leq \max(0, \tau^{-1}((P-U)x)_i - (Ax)_i) & \text{otherwise} \end{cases}$$

- (2) *Except in the trivial case where $Pc(U) = \emptyset$ and therefore $B = A, U, P$ and B are regular with*

$$(4.5) \quad \nu_{\max}(B^+ A) \leq \frac{1}{1-\tau}$$

and

$$(4.6) \quad \nu_{\min}(B^+ A) \geq (1 + \min_{S \oplus \mathcal{N}(A) = \mathbb{C}^n} \max_{\substack{z \in S \\ z \neq 0}} \frac{(z, \Lambda Dz)}{(z, Az)})^{-1}$$

where $D = \text{diag}(A)$.

Note that the factorization algorithm proposed by Axelsson-Barker in [3] (eq. 7.18) is also based upon the choice of a parameter $\alpha > 0$ such that $1/\alpha$ is an upper bound on $\nu_{\max}(B^+ A)$. Clearly, the efficiency of such strategies depends on the smallest eigenvalue ($\nu_{\min}(B^+ A)$) behaviour. Now, one easily verifies with (1) of Theorem 4.2 that the perturbations λ_i involved by Algorithm 1 are anyway smaller¹ than those required for achieving Axelsson-Barker scheme when one uses $\alpha = 1 - \tau$ in the latter (which leads to the same value for the upper bound). Therefore, the smallest eigenvalue analysis developed in [3] also applies to Strategy No.2. It consists essentially in the proof of an inequality of the type $(z, \Lambda Dz) \leq c_1(z, Dz) + c_2(z, Az)$ which, extended to the singular case, leads together with (4.6) to $\nu_{\min}(B^+ A) \geq 1/(1 + c_2 + c_1 \lambda_{\min}^{-1})$ where λ_{\min} denotes the first nonzero eigenvalue of $D^{-1}A$.

This extension is included in an improved version of the Axelsson-Barker result, proven in [24]. Our improvements allow first to remove some hypotheses of the original version and secondly, by a more accurate analysis, to obtain bounds that are also good approximations of ν_{\min} (see [21]). The major conclusion we can deduce from these technical results is that, applied to discrete second order elliptic PDE's, for a family

¹ except possibly at nodes $i \in Pc(U)$ such that $Pc(U)$ does not contain $P(i)$; such nodes are however usually not met in practice

of problems $A_h x_h = b_h$ which differ only by the mesh size h , choosing $\tau_h = 1 - c_1 h$, they give $\nu_{\min}(B_h^+ A_h) \geq c_2$, where c_2 is independent of h ; further, with (4.5), this leads to $\kappa(B_h^+ A_h) \leq O(h^{-1})$.

To avoid the difficulties that would be involved by the use of the above mentioned results and other similar [8], [21] for a practical estimation of ν_{\min} , it is proposed in [10] an heuristic estimate. According to our general philosophy, we propose to generalize the latter by simply replacing $\lambda_1(D^{-1}A)$ by $\lambda_{\min}(D^{-1}A)$; i.e. we propose

$$(4.7) \quad \nu_{\min} \cong \frac{1}{1 + \frac{\langle \lambda \rangle}{\lambda_{\min}}}$$

where

$$(4.8) \quad \langle \lambda \rangle = \frac{(x, \Lambda D x)}{(x, D x)}$$

and

$$(4.9) \quad \lambda_{\min} = \min_{\substack{\lambda \in (D^{-1}A) \\ \lambda \neq 0}} \lambda$$

As in the regular case, (4.7) is justified from Theorem 4.2 by stating that, letting z_{\min} be such that $Az_{\min} = \lambda_{\min} z_{\min}$,

$$(4.10) \quad \min_{S \oplus \mathcal{N}(A) = \mathbb{C}^n} \max_{\substack{z \in S \\ z \neq 0}} \frac{(z, \Lambda D z)}{(z, A z)} \cong \frac{(z_{\min}, \Lambda D z_{\min})}{(z_{\min}, A z_{\min})}$$

and

$$(4.11) \quad \frac{(z_{\min}, \Lambda D z_{\min})}{(z_{\min}, D z_{\min})} \cong \frac{(x, \Lambda D x)}{(x, A x)}$$

generally hold for the matrices Λ as they appear in practical applications of approximate factorization algorithms like Algorithm 1.

For this strategy to be useful in practice, we have still to provide some rules for the determination of τ . To this aim, let us use the ratio of the bound (4.5) by the estimate (4.8) as estimate of $\kappa(B^+ A)$, and let further use

$$(4.12) \quad (x, \Lambda D x) \cong (x, \frac{(1-\tau)^2}{\tau}(P-U)x + (1-\tau)[\max(0, U^t - U)]x),$$

$$(4.13) \quad (x, (P-U)x) \cong \frac{1}{2}(x, D x);$$

we have then, with $\alpha = (1 - \tau)$

$$(4.14) \quad \kappa(B^+ A) \cong \frac{1}{\alpha} + \frac{\alpha}{1-\alpha} \frac{1}{2\lambda_{\min}} + \gamma$$

where γ is some constant independent of τ . The optimum is reached in (4.14) when

$$\alpha^2 = 2\lambda_{\min} \frac{1-2\alpha}{(1-\alpha)^2},$$

i.e. when $\lambda_{\min} \ll 1$:

$$(4.15) \quad 1 - \tau = \alpha \cong \sqrt{2\lambda_{\min}}$$

Then, (4.14) becomes

$$(4.16) \quad \kappa(B^+ A) \cong \sqrt{2\lambda_{\min}} + \gamma \cong \kappa(D^{-1} A) + \gamma$$

showing that Strategy No.2 allows generally when γ is not too large, to improve the original conditioning by an order of magnitude. Unfortunately, λ_{\min} is generally not known in advance so that (4.15) can generally not be achieved in practice. In discrete PDE's applications however, we have $\lambda_{\min} = O(h^{-2})$ so that we can deduce from (4.15) the following rule

$$(4.17) \quad 1 - \tau = \xi \frac{hS}{4V}$$

where h denotes the - or a typical - mesh size, V the area (volume) of the domain and S the length (area) of its boundary, and where ξ is a parameter. Our present experiments indicate that ξ has to be chosen not far from unity, $\kappa(B^+ A)$ being in that case generally relatively insensitive to its actual value. When using the preconditioned conjugate gradient process, we further observe that a sharp optimization of $\kappa(B^+ A)$ is not needed, due to the optimal convergence properties of the method. This is illustrated in the following section. Note that (4.17) is compatible with the above mentioned results $\nu_{\min} \geq O(1)$, $\nu_{\max} \leq O(h^{-1})$ and $\kappa(B^+ A) \leq O(h^{-1})$.

5. Example. We apply here the results of the preceding sections to the finite difference approximation of the Neumann problems

$$\begin{aligned} -\bar{\nabla} D(x, y) \bar{\nabla} u &= f \quad \text{on } \Omega =]0, 1[\times]0, 1[\\ D \frac{\partial u}{\partial n} &= 0 \quad \text{on } \partial\Omega \end{aligned}$$

where $D(x, y)$, $0 \leq x, y \leq 1$, is given by

Problem 1 :

$$D(x, y) = 1 \text{ for all } 0 \leq x, y \leq 1$$

Problem 2 :

$$D = \begin{cases} .010 & \text{if } 0 \leq x, y < \frac{1}{3} \\ 1000 & \text{if } \frac{2}{3} < x, y \leq 1 \\ 1 & \text{otherwise} \end{cases}$$

Problem 3 :

$$D = \begin{cases} 100 & \text{if } 0 \leq x, y < \frac{1}{3} \\ 1000 & \text{if } \frac{2}{3} < x, y \leq 1 \\ 1 & \text{otherwise} \end{cases}$$

In each case, we use an uniform mesh size $h = 1/N$, the order of the resulting linear system $Ax = b$ being $n = (N + 1)^2$. A is ordered according to the lexicographic ordering and we consider only modified incomplete factorizations with $\beta = 0$, while the positive vector to be used in e because $\mathcal{N}(A) = \text{Span}\{e\}$. For both strategies proposed in section 4 and for the Incomplete Cholesky preconditioner, we have computed ν_{max} and ν_{min} (given by (2.26)) for $N = 12, 24, 48, 96$. We also report the number k of preconditioned conjugate gradient (PCG) iterations necessary for meeting the relative residual error $\|r_k\|/\|r_0\| \leq \epsilon$, for $\epsilon = 10^{-3}, 10^{-5}, 10^{-8}$. In each case, $b = Au$, where u is the vector that samples the function $(1 + x)^2(1 + y)(2 - y)e^{xy}$. The initial approximation used is $x_0 = 0$.

Strategy No.1. $U = (u_{ij})$ is determined by

$$\text{offdiag}(U + U^t) = \text{offdiag}(A) \text{ and } Ue = 0$$

and we set $P = \text{diag}(U)$, $B = U^t P^+ U$. We have $\mathcal{N}(B) = \mathcal{N}(A)$, and, in order to perform iterations, we have still to choose a $\{1\}$ -inverse of B . For this purpose, we define $\tilde{U} = (\tilde{u}_{ij})$ by $\tilde{u}_{ij} = u_{ij}$ for all i, j , except $i = j = n$ while $\tilde{u}_{nn} = 1$ and we set $\tilde{P} = \text{diag}(\tilde{U})$ and $\tilde{B} = \tilde{U}^t \tilde{P}^{-1} \tilde{U}$. As mentioned in Section 4, \tilde{B}^{-1} is then a $\{1\}$ -inverse of B and we use it for the solution of our test problem by the PCG process. For completeness, we also mention the upper bound that would be obtained for this example by applying the results of [25].

Strategy No.2. U is here computed according to Algorithm 1, where τ is determined by (4.17) (i.e. $\tau = 1 - \xi/N$) with $\xi = .5, 1, 2..$ We also mention the estimate obtained for ν_{min} by the use of (4.7).

Incomplete Cholesky. Kaasschieter has proposed in [19] the use of the Incomplete Cholesky (IC) preconditioner for solving singular systems. For comparison purpose, we have included the latter in our experiments. Actually, one easily verifies (see [4]) that the Incomplete Cholesky preconditioner is identical to that obtained by a modified incomplete factorization with $x = e$, $\beta = 0$ and $\Lambda = (\lambda_i \delta_{ij})$ (dynamically) determined by

$$(5.1) \quad \lambda_i a_{ii} = \sum_{k \in P(i)} \sum_{\substack{j \in S(k) \\ j \neq i}} \frac{a_{ki} a_{kj}}{u_{kk}}$$

This allows us to use (4.7) to estimate ν_{min} .

TABLE I
Results for Problem 1

		ν_{min}		ν_{max}	up.b.	$\kappa(B^+A)$	Num. of iterations		
		est.					10^{-3}	10^{-5}	10^{-8}
N=12 n=169	S1	1.	1.	32.	48.	32.	12	17	25
	$\xi=.5$.83	.87	8.2	24.	9.9	10	15	21
	$\xi=1.$.65	.71	5.2	12.	8.0	10	15	21
	$\xi=2.$.38	.43	3.0	6.	7.8	10	15	21
	IC	.11	.10	1.2		11.5	11	16	22
N=24 n=625	S1	1.	1.	70.	96.	70.	17	27	39
	$\xi=.5$.83	.87	17.	48.	20.	15	22	32
	$\xi=1.$.65	.71	10.	24.	16.	14	20	29
	$\xi=2.$.40	.44	5.8	12.	15.	14	21	29
	IC	.28E-1	.28E-1	1.2		43.	19	30	38
N=48 n=2401	S1	1.	1.	150.	192.	150.	26	40	62
	$\xi=.5$.83	.87	34.	96.	41.	21	32	47
	$\xi=1.$.66	.71	21.	48.	32.	20	29	42
	$\xi=2.$.40	.44	11.	24.	29.	19	29	40
	IC	.73E-2	.73E-2	1.2		168.	36	55	70
N=96 n=9409	S1	1.	1.	242.	384.	242.	41	63	97
	$\xi=.5$.83	.87	70.	192.	84.	30	47	70
	$\xi=1.$.66	.71	42.	96.	64.	29	42	61
	$\xi=2.$.40	.44	23.	48.	57.	27	41	58
	IC	.18E-2	.18E-2	1.2		668.	71	95	136

The results are displayed in Tables 1,2,3 (S1 refers to Strategy No.1, $\xi=.5, 1., 2.$ to Strategy No.2 with the corresponding value of ξ ; est. refers to the estimate (4.7) of ν_{min} ; up.b. refers to the upper bound obtained in [25] when the Strategy No.1 is concerned, and to the upper bound (4.5) when it is Strategy No.2). The following remarks turn out :

- For the Strategy No.2, we observe that the value of ξ in (4.17) has little influence on $\kappa(B^+A)$ (except for Problem 3) and nearly not on the number of iterations.
- For Problems 1 and 2, both Strategies No.1 and 2 present a spectral condition number $O(h^{-1})$, entailing a number of iterations bounded by $O(h^{-1/2})$. The Strategy No.2 gives better results for Problem 1, while, for Problem 2, in spite of an higher spectral conditioning, the Strategy No.1 gives somewhat better results for realistic stopping criterions. For small ε , however, the Strategy No.2 is still better. Such behaviours can be explained by the superlinear convergence of the PCG process in connection with the eigenvalue distribution in $\sigma(B^+A)\setminus\{0\}$ (see [4], [5], [26]).

TABLE 2
Results for Problem 2

		ν_{min} est.		ν_{max} up.b.		$\kappa(B^+A)$	Num. of iterations		
		10^{-3}	10^{-5}	10^{-8}					
N=12 n=169	S1	1.	1.	14.	48.	14.	7	12	18
	$\xi=.5$.60	.55	5.7	24.	9.5	8	13	20
	$\xi=1.$.37	.34	4.1	12.	11.	9	13	20
	$\xi=2.$.17	.16	2.7	6.	15.7	9	14	21
	IC	.42E-1	.42E-1	1.2		29.5	11	15	21
N=24 n=625	S1	1.	1.	33.	96.	33.	11	18	29
	$\xi=.5$.60	.53	11.	48.	19.	12	20	28
	$\xi=1.$.38	.33	7.9	24.	21.	13	19	29
	$\xi=2.$.18	.16	5.0	12.	29.	12	20	30
	IC	.11E-1	.11E-1	1.2		114.	21	29	40
N=48 n=2401	S1	1.	1.	74.	192.	74.	17	27	46
	$\xi=.5$.60	.51	24.	96.	39.	18	29	41
	$\xi=1.$.38	.32	16.	48.	44.	18	27	40
	$\xi=2.$.18	.16	10.	24.	56.	18	28	42
	IC	.27E-2	.27E-2	1.2		455.	39	56	75
N=96 n=9409	S1	1.	1.	165.	384.	165.	25	42	69
	$\xi=.5$.60	.51	51.	192.	84.	27	40	61
	$\xi=1.$.38	.32	34.	96.	91.	27	38	59
	$\xi=2.$.18	.15	21.	48.	114.	25	35	59
	IC	.70E-3	.70E-3	1.2		1819.	76	113	148

- The Problem 3 presents some pathological features. First, the upper bound obtained in [25] for the unperturbed Strategy is no more $O(h^{-1})$, and the actual value of ν_{max} as well. Therefore, in spite of a nice value for κ obtained for moderate h , the method is less interesting with respect to very small h . Second, the first non zero eigenvalue of $D^{-1}A$ (with $D = \text{diag}(A)$) is about 100 times less than those of Problem 1 (it is a "quasi-singular" singular problem). This explains, by the relation (4.7), the very small values obtained for ν_{min} with the Strategy No.2. In spite of its behaviour which is still $O(h^{-1})$, $\kappa(B^+A)$ presents then a very high value. However, comparing the convergence behaviour with those observed for Problems 1 and 2, we see that only a very few extra iterations are needed. Again, this has to be explained by the super linear convergence in connection with the eigenvalue distribution. We further observe that this super linear convergence is particularly effective for small h and ϵ , i.e. when the Strategy No.1 fails to be completely satisfying.
- Regarding to the "model" Problem 1, we see that the number of iterations involved by the IC preconditioner is prohibitive for small and even moderate h . This is due to the spectral conditioning which is $O(h^{-2})$ for this method. This behaviour is readily explained by our estimate (4.7) (which is found accurate enough). Indeed, by (5.1) we have $\langle \lambda \rangle = O(1)$, so that (4.7) gives, with $\lambda_{min} \ll 1$:

TABLE 3
Results for Problem 3

		ν_{min} est.		ν_{max} up.b.		$\kappa(B^+A)$	Num. of iterations		
		10^{-3}	10^{-5}	10^{-8}					
N=12 n=169	S1	1.	1.	17.	100.	17.	8	13	19
	$\xi=.5$.38E-1	.38E-1	5.7	24.	152.	13	16	24
	$\xi=1.$.16E-1	.16E-1	4.1	12.	251.	12	16	22
	$\xi=2.$.63E-2	.63E-2	2.7	6.	428.	13	17	23
	IC	.14E-2	.14E-2	1.2		877.	15	18	23
N=24 n=625	S1	1.	1.	41.	311.	41.	13	20	33
	$\xi=.5$.35E-1	.35E-1	11.	48.	315.	18	24	33
	$\xi=1.$.15E-1	.15E-1	7.8	24.	508.	17	23	32
	$\xi=2.$.61E-2	.61E-2	5.0	12.	831.	18	24	32
	IC	.35E-3	.35E-3	1.2		3497.	28	33	42
N=48 n=2401	S1	1.	1.	110.	934.	110.	19	33	53
	$\xi=.5$.34E-1	.34E-1	23.	96.	692.	27	35	49
	$\xi=1.$.15E-1	.15E-1	16.	48.	1090.	24	33	45
	$\xi=2.$.59E-2	.59E-2	10.	24.	1701.	25	34	45
	IC	.87E-4	.87E-4	1.2		14030.	54	64	80
N=96 n=9409	S1	1.	1.	334.	2690.	334.	32	53	86
	$\xi=.5$.33E-1	.33E-1	50.	192.	1509.	38	50	71
	$\xi=1.$.15E-1	.15E-1	34.	96.	2311.	36	47	67
	$\xi=2.$.59E-2	.59E-2	21.	48.	3498.	37	49	63
	IC	.22E-4	.22E-4	1.2		56198.	105	127	156

$$\nu_{min}(B^+A) \cong \lambda_{min} / \langle \lambda \rangle = 0(\lambda_{min})$$

showing that the IC preconditioner fails to improve the original conditioning of an order of magnitude as both Strategies No.1 and 2 do. We further observe that, when passing from the "model" Problem 1 to the more realistic Problems 2 and 3, the convergence behaviour associated with the IC preconditioner deteriorates much more than those associated with the Strategies No.1 and 2. The IC method (often referred as the unmodified incomplete factorization method) should therefore, in our opinion, be considered as less robust than the modified incomplete factorization methods. For further comparison between modified and unmodified methods, see Axelsson [2].

6. Conclusion. The major conclusion we can draw from our results is that modified incomplete factorizations of semidefinite systems do not require a special theory or, otherwise stated, that their regular theory does cover the singular case provided that it is written in a formalism which does not exclude singularity by itself. From a practical point of view, it means that a computer program designed for the PCG solution of positive definite systems with preconditioning determined from a modified incomplete factorization can actually also work for semidefinite systems with nearly no changes.

Acknowledgements. I thank Professor R. BEAUWENS for useful comments and suggestions.

REFERENCES

- [1] O. AXELSSON, *Solution of linear equations : iterative methods*, in : *Sparse Matrix Techniques* (V.A. Barker, Editor), Lectures Notes in Mathematics No. 572, Springer-Verlag, 1977.
- [2] ———, *On the eigenvalue distribution of relaxed incomplete factorization methods and the rate of convergence of conjugate gradient methods*, Technical Report, Departement of Mathematics, Catholic University, Nijmegen, The Netherlands, 1989.
- [3] O. AXELSSON AND V. BARKER, *Finite Element Solution of Boundary Value Problems. Theory and Computation*, Academic Press, New York, 1984.
- [4] O. AXELSSON AND G. LINDSKOG, *On the eigenvalue distribution of a class of preconditioning methods*, *Numer. Math.*, 48 (1986), pp. 479–498.
- [5] ———, *On the rate of convergence of the preconditioned conjugate gradient method*, *Numer. Math.*, 48 (1986), pp. 499–523.
- [6] R. BEAUWENS, *Upper eigenvalue bounds for pencils of matrices*, *Lin. Alg. Appl.*, 62 (1984), pp. 87–104.
- [7] ———, *On Axelsson's perturbations*, *Lin. Alg. Appl.*, 68 (1985), pp. 221–242.
- [8] ———, *Lower eigenvalue bounds for pencils of matrices*, *Lin. Alg. Appl.*, 85 (1987), pp. 101–119.
- [9] ———, *Approximate factorizations with S/P consistently ordered M-factors*, *BIT*, 29 (1989), pp. 658–681.
- [10] ———, *Modified incomplete factorizations strategies*, submitted for publication in PCG Conference Proceedings, (1990).
- [11] R. BEAUWENS AND I. QUENON, *Existence criteria for partial matrix factorizations in iterative methods*, *Siam J. Numer. Anal.*, 13 (1976), pp. 615–643.
- [12] R. BEAUWENS AND R. WILMET, *Conditioning analysis of positive definite matrices by approximate factorizations*, *J. Comput. Appl. Math.*, 26 (1989), pp. 257–269.
- [13] A. BEN ISRAEL AND T. GREVILLE, *Generalized Inverses : theory and applications*, J. Wiley and Sons, New York, 1974.
- [14] A. BERMAN AND R. PLEMMONS, *Nonnegative Matrices in the Mathematical Sciences*, Academic Press, New York, 1979.
- [15] A. GEORGE AND J. LIU, *Computer Solution of Large Sparse Positive Definite Systems*, Prentice-Hall, Englewood Cliffs, 1981.
- [16] I. GUSTAFSSON, *Modified incomplete Cholesky (MIC) methods*, in : *Preconditioning Methods, Theory and Applications* (D.J. Evans, Editor), Gordon and Breach Science, 1983.
- [17] I. HAGEMAN AND D. YOUNG, *Applied Iterative Methods*, Academic Press, New York, 1981.
- [18] F. HARARY, *Graph Theory*, Addison-Wesley, Reading, 1969.
- [19] F. KAASSCHIETER, *Preconditioned conjugate gradients for solving singular systems*, *J. Comp. Appl. Math.*, 24 (1988), pp. 265–275.
- [20] V. LEBEDEV AND S. FINOGENOV, *Utilisation of ordered Chebyshev parameters in iterative methods*, *URSS Comput. Math. Math. Phys.*, 16 (1976), pp. 70–83.
- [21] M. MAGOLC, *Lower eigenvalue bounds for singular pencils of matrices*, in preparation.
- [22] Y. NOTAY, *Sur le conditionnement de matrices de Stieltjes par factorisations approchées*, Mémoire. Université Libre de Bruxelles, Brussels, Belgium, 1987.
- [23] ———, *Polynomial acceleration of iterative schemes associated with subproper splittings*, *J. Comp. Appl. Math.*, 24 (1988), pp. 153–167.
- [24] ———, *Incomplete factorizations of singular linear systems*, *BIT*, 29 (1989), pp. 682–702.
- [25] ———, *Conditioning of Stieltjes matrices by S/P consistently ordered approximate factorizations*, submitted for publication, (1990).
- [26] A. VAN DER SLUIS AND H. VAN DER VORST, *The rate of convergence of conjugate gradients*, *Numer. Math.*, 48 (1986), pp. 543–560.
- [27] R. VARGA, *Matrix iterative analysis*, Prentice Hall, Englewood Cliffs, 1962.