Numerical Linear Algebra with Applications, Vol. 1(6), 511–532 (1994)

DRIC: a Dynamic Version of the RIC Method

Yvan Notay

Service de Métrologie Nucléaire, Université Libre de Bruxelles (C.P. 165), 50, Av. F.D. Roosevelt, B-1050 Brussels, Belgium

A new incomplete factorization method is proposed, differing from previous ones by the way in which the diagonal entries of the triangular factors are defined. A comparison is given with the dynamic modified incomplete factorization methods of Axelsson–Barker and Beauwens, and with the relaxed incomplete Cholesky method of Axelsson and Lindskog. Theoretical arguments show that the new method is at least as robust as both previous ones, while numerical experiments made in the discrete PDE context show an effective improvement in many practical circumstances, particularly for anisotropic problems.

KEY WORDS Iterative methods for linear systems Acceleration of convergence Preconditioning

1. Introduction

The preconditioned conjugate gradient method with preconditioning obtained by incomplete factorization of the system matrix is now a popular technique for the iterative solution of large sparse linear systems. For systems arising from the discretization of second order elliptic PDEs, it is becoming generally admitted that so-called 'modified' Incomplete Cholesky (IC) factorizations are more efficient than the standard (or 'unmodified') one. In the former methods, the diagonal part of the upper triangular factor is computed so that the preconditioner satisfies a (generalized) row-sum condition. Because the 'unperturbed' version, for which the preconditioner and the system matrix have same row-sum, may suffer from an unpredictable behaviour of its associated largest eigenvalue(s) (see e.g. [3,9,18]), appropriate perturbations are generally introduced so as to satisfy a prescribed upper eigenvalue bound.

Several recent works have focused therefore on the design of strategies according to which perturbations should be introduced. From a practical point of view, a user friendly method should not depend on the analysis of the associated PDE but rather be usable in a 'black-box' fashion, and when a parameter-dependency cannot be avoided, a rule should be provided for its determination.

CCC 1070–5325/94/060511–22

Received August 1992
Revised June 1994

A first such method was that initiated by Axelsson [1], generalized by Gustafsson [13], whose *dynamic* version, first proposed by Axelsson and Barker [3], has been reformulated in a more convenient way by Beauwens [9]. It will be referred to as DMIC (Dynamic Modified Incomplete Cholesky). It has been studied over many years and found very robust for discrete second-order elliptic PDE problems with *isotropic* coefficients. A theoretical analysis of this robustness has further been provided in [17], which, as will be seen, leads one on the other hand to expect a weak behaviour of the method in presence of strong anisotropies.

Another worthwhile candidate for preconditioning is the Relaxed Incomplete Cholesky (RIC) method by Axelsson and Lindskog [4]. It was introduced on a more empirical basis, but, following a point of view already used in [2] within the framework of a model problem analysis, we shall prove here a general upper spectral bound for the associated preconditioning matrix, and develop accordingly a specific conditioning analysis. A comparison with DMIC on this basis will lead to a less favourable asymptotic behaviour for RIC, but at the same time display its insensitivity to anisotropies.

The main result of this paper is a new upper eigenvalue bound, on the basis of which we develop a new perturbation technique, which uses perturbations similar to those of RIC, but dynamically modulated in much the same way as in DMIC. It will be seen that the new method, which we call DRIC (Dynamic Relaxed Incomplete Cholesky), combines the favourable asymptotic behaviour of DMIC with the weak sensitivity of RIC to anisotropies, and is therefore at least as robust as both previous techniques.

To conform with a widespread usage, we shall refer in this paper to the modified incomplete factorization method *without perturbations* (row-sum criterion satisfied exactly) as MIC. However, we warn the reader against confusion with the original MIC method by Gustafsson [13], which involves perturbations and turns out to be more or less equivalent to DMIC, see the comments in section 3 or [19] for a direct comparison.

The remainder of this paper is organized as follows. General terminology and notation are summarized below; sections 2, 3, 4 and 5 contain, respectively, some general considerations about incomplete factorizations, a summary of the analysis of DMIC, our analysis of RIC, and the developments which lead to the definition of DRIC; the algorithm and basic properties of each method are summarized in section 6, and they are numerically compared for a few typical examples in section 7.

General terminology and notation. All vectors belong to $\mathcal{C}^n$; all matrices are $n \times n$ real matrices. The symbols $A^t, \mathcal{N}(A)$ and $\sigma(A)$ denote, respectively, the transpose, the null space and the spectrum of the matrix A. 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 $offdiag(A) = A - diag(A)$. The Hadamard product $A{*}B$ of the matrices $A = (a_{ij})$ and $B = (b_{ij})$ of the same dimensions is defined by element by element multiplication: $(A * B)_{ij} = a_{ij}b_{ij}$. The unit matrix with respect to the Hadamard multiplication, denoted ϵ, is the matrix whose entries are all equal to unity, while $e = (1 \dots 1)^t$.

As usual when dealing with approximate factorization methods, we shall assume that the system matrix is a Stieltjes matrix, more general cases arising from the discrete PDE context being included through the reduction techniques presented in [3,11,15]. To cover the singular case at once, as done in our preceding works (see e.g. [14]), we use an extended notion for Stieltjes matrices and define a Stieltjes matrix as any symmetric nonnegative definite matrix with nonpositive offdiagonal entries.

2. Perturbed and unperturbed methods

In *modified* incomplete factorization methods, the diagonal entries of the upper triangular factor $U = P - F$, with $P = diag(U)$, are determined from the relation

$$Bx = Ax + \Lambda Dx \tag{2.1}$$

where $B = U^t P^{-1} U$ is the approximate factorization, $D = diag(A)$, $\Lambda = (\lambda_i \delta_{ij})$ a diagonal matrix of prescribed nonnegative perturbation parameters, and x a positive vector such that $Ax \geq 0$. The existence of such a positive vector is known for Stieltjes matrices; when A is diagonally dominant, $x = e$ is the standard choice.

Setting $\beta_{ij} = 1$ if fill-in is allowed in the position (i, j) and $\beta_{ij} = 0$ otherwise, the factorization algorithm is then as follows:

$$For \quad i = 1, \dots, n \; :$$

$$u_{ij} = a_{ij} - \beta_{ij} \sum_{k<i} \frac{u_{ki} u_{kj}}{p_{kk}} , \qquad j = i + 1, \dots, n \tag{2.2}$$

$$p_{ii} = \lambda_i a_{ii} + x_i^{-1} \left((Ax)_i + (Fx)_i - \sum_{k<i} \frac{u_{ki}((P - F)x)_k}{p_{kk}} \right) \tag{2.3}$$

We give this algorithm for theoretical reference only since (see section 6 below) practical implementations generally avoid the explicit computation of the quantities referred to in (2.3).

In general, the positive definiteness of P is not guaranteed and has to be ensured by imposing some criterion, see e.g. [3,14]. However, with the DMIC, RIC (with $-1 \leq \omega < 1$) and DRIC methods, which are our main concern here, P is always positive definite when A is a Stieltjes matrix. We refer to the above quoted works for a detailed analysis.

The method corresponding to $\Lambda = 0$ is the 'unperturbed' method, which we also refer to as MIC. 'Perturbed' methods are those where appropriately chosen perturbations are introduced, generally so as to be consistent with some conditioning analysis.

In the earliest paper which demonstrates the efficiency of modified incomplete factorization methods [1], Axelsson used such a perturbation technique, although it was not presented as such due to the lack of any reference to an unperturbed case.

This raised later the question of the necessity of these pertubations, which was subsequently investigated by Beauwens *et al.* [6–9,16]. It is unfortunately not possible to summarize briefly the main issues of these works, except (very roughly) by saying the following: while it is true that there are problems for which the unperturbed method is as fast as the perturbed ones, and even some quasi-singular problems (i.e. with a nearly zero eigenvalue) for which it is faster than expected for any perturbed method, on the other hand, there has up to now been no general means to obtain a robust preconditioner on the basis of unpertubed factorizations. That is why we shall concentrate here on the discussion and comparison of perturbed methods.

All perturbed methods are based on the fact that the main drawback of unperturbed MIC, which is the sometimes unpredictable behaviour of its associate largest eigenvalue(s), can be removed by using perturbations λ_i so as to satisfy a prescribed upper eigenvalue bound.

A standard but sharp lowest eigenvalue estimate is then

$$\nu_{\min}(B^{-1}A) \geq \left(1 + \nu_{\max}(A^{-1}\Lambda D)\right)^{-1} \tag{2.4}$$

(see e.g. [9] for a proof and [14] for the extension to the singular case). Further, (2.4) extends as follows [17] :

$$\nu_{\min}^{(p)}(B^{-1}A) \geq \left(1 + \nu_{\max}^{(p)}(A^{-1}\Lambda D)\right)^{-1} \tag{2.5}$$

where $\nu_{\min}^{(p)}$ and $\nu_{\max}^{(p)}$ denote to the pth eigenvalue by respectively increasing and decreasing order.

The lowest eigenvalue's behaviour thus depends essentially on the size of the perturbations, so we may compare the different techniques by evaluating the perturbations needed to achieve the desired upper bound. This will be the basis of our analysis in the following sections.

3. The DMIC method

The upper bound on which the DMIC method is based first appeared in Dupont, Kendall and Rachford [12]. This result has further been generalized by Beauwens and Quenon in [10], where they prove that a modified incomplete factorization $B = U^t P^{-1} U$ of a Stieltjes matrix A (as defined in section 1) satisfies

$$\nu_{\max}(B^{-1}A) \leq \frac{1}{1-\tau} \tag{3.1}$$

provided that the upper triangular factor satisfies

$$\tau Px \geq Fx \tag{3.2}$$

Look now at the recurrence relation (2.3) which determines P. Without strict diagonal dominance of A ($(Ax)_i = 0$), and neglecting the term $\sum_{k<i} p_{kk}^{-1}(-u_{ki})((P-F)x)_k$, we are left with $p_{ii} \geq x_i^{-1}(Fx)_i + \lambda_i a_{ii}$, so that to satisfy $\tau(Px)_i \geq (Fx)_i$ for some $\tau < 1$ seems to require $\lambda_i = \mathcal{O}(1-\tau)$, and it is clear that, with such perturbations, one will hardly improve the original conditioning by combining the upper bound (3.1) with the lower bound (2.4).

A decisive step was carried out by Axelsson [1], who developed a careful analysis of the term $\sum_{k<i} p_{kk}^{-1}(-u_{ki})((P-F)x)_k$, responsible for the 'propagation' of the diagonal dominance from a node k to its neighbours i with $i > k$. He showed that, for a certain class of second-order discrete elliptic PDEs, and assuming a natural ordoring of the unknowns, perturbations $\lambda_i = \mathcal{O}(h^2)$ were sufficient to ensure $1 - \tau = \mathcal{O}(h)$, so that the combination of (3.1) and (2.4) gives now $\kappa(B^{-1}A) \leq \mathcal{O}(h^{-1})$, which is an order of magnitude better than the original conditioning.

Axelsson's result was limited to some model problems with Dirichlet boundary conditions and no strong discontinuities. The next step was taken by Gustafsson [13], who showed that these limitations might be overcome by putting $\mathcal{O}(h)$ perturbations at nodes corresponding to Neumann boundary conditions and to lines/planes of discontinuity. This was sufficient to ensure $1 - \tau = \mathcal{O}(h)$ while a careful analysis of (2.4) showed that $\nu_{\min} = \mathcal{O}(1)$.

This method works quite well, with the drawback that a judicious choice of the perturbations requires in general an analysis of the considered PDE. This became unnecessary with the algebraic reformulation by Axelsson and Barker [3], which showed that (3.2) with $\tau = 1 - \alpha$ was ensured by using perturbations computed according to

$$p_{ii}^{(0)} = x_i^{-1}\left((Ax)_i + (Fx)_i - \sum_{k<i} \frac{u_{ki}((P-F)x)_k}{p_{kk}}\right) \tag{3.3}$$

$$\alpha_i^{(0)} = 1 - \frac{(Fx)_i}{p_{ii}^{(0)} x_i} \tag{3.4}$$

$$\lambda_i = \begin{cases} 0 & \text{if } \alpha_i^{(0)} \geq \alpha \\ \dfrac{\alpha^2}{1-\alpha}\,\dfrac{p_{ii}^{(0)}}{a_{ii}} + \dfrac{\alpha}{1-\alpha}\,\dfrac{\max\left[((F-F^t)x)_i\,,\,0\right]}{a_{ii}x_i} & \text{otherwise} \end{cases} \tag{3.5}$$

With $\alpha = \mathcal{O}(h)$, these perturbations are not much different from Gustafsson's, a standard analysis giving, for a natural ordering, $((F - F^t)x)_i \leq \mathcal{O}(h)$ for all interior nodes except along the lines/planes of discontinuity. The improvement is that their computation is now part of the factorization algorithm and requires only the choice of the parameter α. Additionally, Axelsson and Barker produced a more general proof that perturbations $\mathcal{O}(h)$ at some nodes are consistent with $\nu_{\min} = \mathcal{O}(1)$ provided that the number of such nodes is no larger than $\mathcal{O}(h^{1-d})$, where d is the dimension of the space in which the problem is solved.

Finally, an ultimate development is due to Beauwens [9], who proposed to replace (3.5) by

$$\lambda_i = \begin{cases} 0 & \text{if } \alpha_i^{(0)} \geq \alpha \\ \dfrac{\alpha - \alpha_i^{(0)}}{1-\alpha}\,\dfrac{p_{ii}^{(0)}}{a_{ii}} & \text{otherwise} \end{cases} \tag{3.6}$$

that is, to compute dynamically at each node the perturbation strictly necessary to satisfy (3.2) with $\tau = 1 - \alpha$. The method is further refined by showing that it is unnecessary to perturb nodes that are not at the origin of fill-in elements (i.e. nodes k such that there are not at least two nodes, $i,\ j,\ j > i > k$, such that $u_{ki}u_{kj} \neq 0$).

$\nu_{\max}(B^{-1}A) \leq \alpha^{-1}$ is here automatically satisfied, so that we are left with analysing the involved perturbations. This was carried out in [14], where we showed that

$$\lambda_i \leq \max\left[0\,,\ \frac{\alpha^2}{1-\alpha}\,\frac{(Fx)_i}{a_{ii}x_i} + \alpha\,\frac{((F - F^t - A)x)_i}{a_{ii}x_i}\right] \tag{3.7}$$

Hence, see (3.5), Axelsson–Barker's lowest eigenvalue analysis *a fortiori* applies to this version of the method, which we properly call DMIC. It presents only small differences with the previous versions from the conditioning point of view, so that its main advantage is that it is more elegant and easier to program. This motivates our extensive reference to DMIC in the following, but one has to remember that, when such practical details are not of concern, the same comments apply to all of the earlier methods quoted above.

Another merit of the algebraic results (3.5), (3.7) which we would like to point out is that, together with the restriction '$\lambda_i = \mathcal{O}(h)$ at no more than $\mathcal{O}(h^{1-d})$ nodes', they provide a very simple means to distinguish the orderings that are convenient for (perturbed) modified incomplete factorization.

Now, all these results do not disclose why the DMIC method has some serious drawbacks, and are actually mainly concerned with the asymptotic behaviour of the method for $h \to 0$, while, from a practical point of view, one is rather interested with the actual behaviour for realistic problem sizes.

These considerations gave rise to the more accurate analysis developed in [17]. There, assuming as usual that the discretization is sufficiently regular so that it makes sense to use the average mesh size h as a parameter, we first prevent any trouble from its dimensional dependency by defining the reduced average mesh size :

$$h_0 = \frac{hS}{4V} \tag{3.8}$$

where V denotes the area/volume of the domain Ω in which the PDE is solved and S the length/area of its boundary.

Then, summarizing very roughly, our main result is that, assuming a natural ordering of the unknowns, one has, for the DMIC method with $\alpha \approx h_0$,

$$\nu_{\min}^{(p)}(B^{-1}A) \approx \left(1 + \frac{h_0^2}{\lambda_{\min}^{(p)}(D^{-1}A)}\right)^{-1} \tag{3.9}$$

where $\nu_{\min}^{(p)}(B^{-1}A)$ and $\lambda_{\min}^{(p)}(D^{-1}A)$ denote the pth (nonzero) eigenvalue of the preconditioned and unpreconditioned system respectively.

Hence, the conditioning is close to h_0^{-1} if and only if $h_0^{-2}\lambda_{\min}^{(1)}(D^{-1}A)$ is close to 1. A weaker robustness condition is obtained by taking into account the optimal convergence properties of the conjugate gradient method. Namely, a particular case of the bound (2.20) in Axelsson and Lindskog [5] gives

$$k_\varepsilon \le \operatorname{int}\left[\frac{1}{2}\sqrt{\frac{\nu_{\max}}{\nu_{\min}^{(p)}}}\left(\ln\frac{2}{\varepsilon} + \sum_{i=1}^{p-1}\ln\frac{\nu_{\min}^{(p)}}{\nu_{\min}^{(i)}}\right)\right] + (p-1)\operatorname{int}\left[\sqrt{\frac{\nu_{\max}}{\nu_{\min}^{(p)}}} + 1\right] + 1 \tag{3.10}$$

for the number of iterations k_ε to reduce the relative error in the A-(semi)norm by a factor ε. (See [20] for details, where this bound is further shown to be valid also in the presence of rounding errors.) From this one readily deduces that $k_\varepsilon \le c h_0^{-\frac{1}{2}} \ln \varepsilon^{-1}$ with c not too far from unity, provided that

$$h_0^{-2}\lambda_{\min}^{(p)}(D^{-1}A) \tag{3.11}$$

is close to 1 for some small integer p. This is much weaker than when restricted to $p = 1$ and explains why the DMIC method often exhibits faster convergence rates than unperturbed MIC, even in examples for which the opposite result would be expected on the basis of the spectral condition number only.

This robustness condition seems generally satisfied for isotropic problems. Difficulties arise, however, in the case of anisotropic problems. An example is the 5-point finite difference approximation with uniform mesh size $h = 1/N$ of

$$-\ \partial_{xx}^2 u \ -\ \zeta\ \partial_{yy}^2 u \ =\ f \tag{3.12}$$

on the unit square, with Dirichlet boundary conditions on the bottom ($y = 0$) and top ($y = 1$) boundaries, and Neumann boundary conditions elsewhere. The eigenvalues of the unpreconditioned system are given by

$$\lambda_{k\ell}(D^{-1}A) = \frac{1}{1+\zeta}\left(\sin^2\frac{k\pi}{2N} + \zeta\sin^2\frac{\ell\pi}{2N}\right) \tag{3.13}$$

for $k = 0, \ldots, N$ and $\ell = 1, \ldots, N-1$. For small values of ζ, the lowest eigenvalues are

$$\lambda_{\min}^{(\ell)}(D^{-1}A) = \frac{\zeta}{1+\zeta}\sin^2\frac{\ell\pi}{2N}, \qquad \ell = 1, 2, 3, \ldots \tag{3.14}$$

and condition on (3.11) is not satisfied, (3.9) giving

$$\nu_{\min}^{(\ell)}(B^{-1}A) \approx \left(1 + \frac{4(1+\zeta)}{\zeta\pi^2\ell^2}\right)^{-1}, \qquad \ell = 1, 2, 3, \ldots \tag{3.15}$$

Note, on the other hand, that we still have $N^2\lambda_{\min}^{(1)}(D^{-1}A) \geq 1$ for all $\zeta \geq 1$, $N \geq 2$, so that no trouble occurs for large ζ. This means that *not all* anisotropic problems yield difficulties for the DMIC method.

4. The RIC method

In the RIC method, the perturbations are such that

$$Bx = Ax + (1-\omega)\, offdiag(B-A)x \tag{4.1}$$

for some $-1 \leq \omega < 1$. Since, letting $E = F^t$,

$$offdiag(B-A) = (\epsilon - \beta) * EP^{-1}F \tag{4.2}$$

where ϵ is the matrix with all entries equal to unity and β is be the symmetric matrix with diagonal unity such that $(\beta)_{ij} = \beta_{ij}$ for $1 \leq i < j \leq n$, this amounts to using (2.2), (2.3) with

$$\lambda_i = \frac{1}{a_{ii}x_i}(1-\omega)\sum_{j\neq i}(1-\beta_{ij})\sum_{k<i,j}\frac{u_{ki}u_{kj}x_j}{p_{kk}} \tag{4.3}$$

In [2], $2/(1-\omega)$ is shown to be an upper spectral bound for $B^{-1}A$ for a model problem. We give in the next section a general proof (not restricted to model problems) of this result. Since the recommended choice for ω is

$$\omega = 1 - \delta h_0 \tag{4.4}$$

with δ close to 1 (see [4]), the method may then be seen as an alternate perturbation technique to DMIC, similarly formulated to satisfy a 'target' upper bound.

The drawback of RIC is that the perturbations are $\mathcal{O}(h)$ at nearly all nodes, so that (2.5) gives

$$\nu_{\min}^{(p)}(B^{-1}A) \approx \frac{1}{1 + \mathcal{O}(h^{-1})} \tag{4.5}$$

and the method does not have the same favourable asymptotic behaviour as DMIC.

Consider now, for illustration, the discrete PDE problem (3.12) referred to at the end of the preceding section. Using $x = e$, the lexicographic ordering and a factorization scheme without fill-in allowed ($\beta_{ij} = 0 \; \forall \, i \neq j$), one can easily confirm that

$$\lambda_i \;\approx\; \frac{\zeta(1-\omega)}{(1+\zeta)^2} \tag{4.6}$$

at nearly all nodes. Hence, for $1 - \omega = 2h_0 = 2N^{-1}$ and small ζ, one obtains from (2.5)

$$\nu_{\min}^{(\ell)}(B^{-1}A) \;\approx\; \left(1 + \frac{8\,N}{(1+\zeta)\pi^2\ell^2}\right)^{-1} , \quad \ell = 1, 2, 3, \dots \tag{4.7}$$

which is effectively $\mathcal{O}(h)$, but actually better than (3.15) as long as $N \leq (1+\zeta)^2/2\zeta$.

Note that, on the other hand, we have anyway $\nu_{\min}(B^{-1}A) \geq 1/2$ when $\zeta \geq N$, so that, here again, the situation is much more favourable for large ζ due to the insensitivity of the eigenvalues of $D^{-1}A$ to the anisotropy ratio ζ. This may even be exploited as done in [2] to prove a better conditioning by using a smaller value for ω; we refer the reader without further comments to the latter work for more details, since our interest here is the discussion of more difficult problems.

5. The DRIC method

The weakness of DMIC is due to the fact that it uses perturbations proportional to $(Fx)_i$, and thus does not take advantage of the very small value of the neglected fill-in to reduce the amount of perturbation in the presence of strong anisotropies. On the other hand, RIC has a less favourable asymptotic behaviour because it uses $\mathcal{O}(h)$ perturbations everywhere while DMIC exploits the propagation of the diagonal dominance to put $\mathcal{O}(h)$ perturbations only at some selected nodes.

It is therefore tempting to propose a new technique which, like RIC, uses perturbations proportional to the neglected fill-in, but, like DMIC, takes into account the diagonal dominance of U. Of course, we first need to prove that such perturbations are sufficient to achieve a similar upper spectral bound. This is the main purpose of Theorem 5.1 below. We first prove the following lemma :

Lemma 5.1. *Let $F \geq 0$ be a strictly upper triangular matrix and P, Q nonnegative diagonal matrices. Let B be a matrix such that*

$$\begin{aligned} Bx &\geq 0 \\ offdiag(B) &= offdiag((P-E)Q(P-F)) \end{aligned}$$

where x is some positive vector and $E = F^t$. If $Px \geq Fx$, then B is nonnegative definite.

Proof Let P_0 be the diagonal matrix such that $P_0 x = Fx$, and set $B_1 = (P_0 - E)Q(P_0 - F)$. B_1 is nonnegative definite by construction while $B_2 = B - B_1$ satisfies

$$\begin{aligned} B_2 x &\geq 0 \\ offdiag(B_2) &= -(P-P_0)QF - EQ(P-P_0) \end{aligned}$$

showing together with $Px \geq Fx = P_0 x \Rightarrow P \geq P_0$ that B_2 is a Stieltjes matrix, whence the conclusion. ■

Theorem 5.1. *Let A be a symmetric nonnegative definite matrix, x a positive vector such that $Ax \geq 0$, $U = P - F$ with $P = diag(U)$ an upper triangular M matrix such that $Ux \geq 0$, and β, $0 \leq \beta \leq \epsilon$, a symmetric matrix such that,*

$$F + E \ \geq \ -offdiag(A) \ + \ \beta * offdiag(EP^+F) \tag{5.1}$$

where $E = F^t$ and P^+ is the diagonal matrix defined from P by

$$(P^+)_{ii} = \begin{cases} p_{ii}^{-1} & if \ \ p_{ii} \neq 0 \\ 0 & if \ \ p_{ii} = 0 \end{cases} \tag{5.2}$$

If

$$B = (P - E)P^+(P - F) \tag{5.3}$$

satisfies

$$Bx \ \geq \ \mu_0^{-1} Ax \ + \ (\epsilon - \beta) * offdiag(E\ \Gamma P^+ F)x \tag{5.4}$$

for some number $\mu_0 \geq 1$ and some nonnegative diagonal matrix $\Gamma = (\gamma_i \delta_{ij})$, and if, letting τ_i, $0 \leq \tau_i \leq 1$, $i = 1, \ldots, n$, be such that

$$\tau_i (Px)_i \ \geq \ (Fx)_i \tag{5.5}$$

one has

$$\tau_i (2 - \gamma_i) \ < \ 2 \qquad for\ all\ i \tag{5.6}$$

then

$$(z, Az) \ \leq \ \mu(z, Bz) \quad for\ all\ z \in \mathcal{C}^n \tag{5.7}$$

with

$$\mu \ = \ \max\left[\mu_0\,,\ \max_{1 \leq i \leq n} \frac{2}{2 - (2 - \gamma_i)\tau_i}\right] \tag{5.8}$$

Proof Let $\Theta = (\theta_i \delta_{ij})$ be such that

$$\theta_i = \begin{cases} 0 & \text{if } \ \tau_i \leq 1 - \frac{1}{\mu} \\ 1 - \tau_i^{-1}(1 - 1/\mu) & \text{otherwise} \end{cases}$$

We have $0 \leq \Theta \leq \mu^{-1} I$ and therefore the matrix

$$B_1 \ = \ \Delta_1 \ - \ \mu^{-1}(F + E + offdiag(A)) \ + \ \beta * offdiag(E\ \Theta P^+ F)$$

with Δ_1 diagonal and such that $B_1 x \ = \ 0$, is a Stieltjes matrix and hence nonnegative definite. Let further $T = (t_i \delta_{ij})$ be such that

$$t_i = \begin{cases} 1 - \frac{1}{\mu} & \text{if } \ \tau_i \leq 1 - \frac{1}{\mu} \\ \tau_i & \text{otherwise} \end{cases}$$

Since $t_i \geq \tau_i$ for all i, one has with (5.5) that $TPx \geq Fx$, and therefore the matrix

$$B_2 = (TP - E)(I - \Theta)P^+(TP - F) + \Delta_2$$

with Δ_2 diagonal and such that $B_2 x = 0$, is nonnegative definite by application of Lemma 5.1.

On the other hand, $P\ P^+F = F$ (because, with $Ux \geq 0$, $p_{ii} = 0$ for some i implies $u_{ij} = 0$ for all $j > i$) while $T(I - \Theta) = (1 - 1/\mu)I$. An elementary calculation shows then that

$$B - \frac{1}{\mu}A - B_1 - B_2 = \Delta_3 + (\epsilon - \beta) * offdiag(E\ \Theta P^+ F) \qquad (5.9)$$

where Δ_3 is diagonal. From (5.4) one has then

$$\Delta_3 x \geq (\epsilon - \beta) * offdiag(E(\Gamma - \Theta)P^+ F)x$$

while, by (5.8), $(1 - \frac{\gamma_i}{2})\tau_i \leq (1 - 1/\mu)$ for all i, so that either $\tau_i \leq (1 - 1/\mu)$ or $\gamma_i \geq 2(1 - \tau_i^{-1}(1 - 1/\mu))$. This implies $\Gamma \geq 2\Theta$, showing that the right-hand side of (5.9) is nonnegative definite, whence the conclusion. ■

It is readily seen that, within the framework of modified incomplete factorizations of a Stieltjes matrix A, all assumptions but (5.6) are automatically satisfied with $\mu_0 = 1$. On the other hand, we do not assume the regularity of P since, when A is regular, the positive definiteness of B directly follows from (5.7). Hence, for any factorization algorithm such that all assumptions are satisfied with μ no greater than some prescribed $\overline{\mu}$, Theorem 5.1 proves that B is a positive definite preconditioner, and that $\upsilon_{\max}(B^{-1}A) \leq \overline{\mu}$. This nice property extends to the singular case, since, as is easily confirmed from the results in [14], for a modified incomplete factorization B of an irreducible singular Stieltjes matrix A, (5.7) is not compatible with B singular except when $offdiag(B) = offdiag(A)$, in which case B is an exact factorization of A.

Let us now show that the upper bounds for DMIC and RIC are particular cases of Theorem 5.1. This is obvious for DMIC since $Bx \geq Ax$ implies (5.4) with $\Gamma = 0$, and (5.5), (5.6), (5.8) reduce then to $\mu = 1/(1 - \tau)$ with $\tau < 1$ such that $\tau Px \geq Fx$. On the other hand, the claimed upper bound for RIC follows from the fact that (4.1) and (4.2) imply (5.4) with $\Gamma = (1 - \omega)I$, whence $\mu \leq 2/(1 - \omega)$ since $\tau_i \leq 1$ for all i.

Now, an alternative possibility of guaranteeing $\mu \leq \alpha^{-1}$ for some prescribed α consists in computing at the ith step of the factorization algorithm $p_{ii}^{(0)}$ and $\alpha_i^{(0)}$ by (3.3), (3.4) as in DMIC, and, whenever $\alpha_i^{(0)} \leq \alpha$, adding to all nodes $j > i$ such that $u_{ij} \neq 0$ the perturbation

$$\lambda_j^{(i)} = \frac{2}{a_{jj}x_j} \sum_{j' \neq j} (1 - \beta_{jj'}) \frac{\alpha - \alpha_i^{(0)}}{1 - \alpha_i^{(0)}} \frac{u_{ij}u_{ij'}x_{j'}}{p_{ii}} \qquad (5.10)$$

Thus, $\lambda_j = \sum_{i<j} \lambda_j^{(i)}$ for all j. One then can easily check that Theorem 5.1 applies with $\tau_i = 1 - \alpha_i^{(0)}$ for all i and

$$\gamma_i = \begin{cases} 0 & \text{if } \alpha_i^{(0)} \geq \alpha \\ 2(\alpha - \alpha_i^{(0)})/(1 - \alpha_i^{(0)}) & \text{otherwise} \end{cases} \qquad (5.11)$$

so that this perturbation technique, which is just the DRIC method, guarantees as claimed $\mu \le \alpha^{-1}$. We shall further see in the next section that it is quite easy to implement, needing only a slight modification of any program which already works for RIC or DMIC.

We are unfortunately not able to derive an upper bound on the perturbations involved by this method. However, we can make an asymptotic analysis for the model example (3.12) of section 3 with $x = e$, the lexicographic ordering and a factorization scheme without fill-in allowed. Indeed, consider the limit case where N goes to infinity, so that it makes sense to define $(\lambda,\ \alpha^{(0)},\ p,\ p^{(0)}) = \lim_{i\to\infty}(\lambda_i,\ \alpha_i^{(0)},\ p_{ii},\ p_{ii}^{(0)})$ and set $p = (1+\zeta)(1+\delta)$, $p^{(0)} = (1+\zeta)(1+\delta_0)$. Since $\lambda = 2\lim_{i\to\infty}\lambda_{i+1}^{(i)}$, (3.3), (3.4), (5.10) and $p_{ii} = \lambda_i a_{ii} + p_{ii}^{(0)}$ give respectively, for $i \to \infty$,

$$\delta_0 = \frac{\delta}{1+\delta} \tag{5.12}$$

$$\alpha^{(0)} = \frac{\delta_0}{1+\delta_0} \tag{5.13}$$

$$\lambda = 2(\alpha - \alpha^{(0)})\frac{\zeta}{(1+\zeta)^2} \tag{5.14}$$

$$\delta = 2\lambda + \delta_0 \tag{5.15}$$

Therefore, combining (5.15) and (5.12),

$$\lambda = \frac{1}{2}(\delta - \delta_0) = \frac{\delta^2}{2(1+\delta)} \tag{5.16}$$

and, substituting the latter value in (5.14), one gets from (5.13)

$$\frac{\delta^2}{1+\delta} = \frac{4\zeta}{(1+\zeta)^2}\left(\alpha - \frac{\delta_0}{1+\delta_0}\right) \tag{5.17}$$

or, using (5.12) again

$$\frac{\delta}{1+2\delta} + \frac{\delta^2}{1+\delta}\,\frac{(1+\zeta)^2}{4\zeta} = \alpha \tag{5.18}$$

Hence, neglecting either the first or the second term of the LHS,

$$\lambda = \frac{\delta^2}{2(1+\delta)} \le \min\left[\frac{\alpha^2}{2(1-2\alpha)},\ \frac{2\alpha\zeta}{(1+\zeta)^2}\right] \tag{5.19}$$

The first of these numbers is very similar to our estimate (3.7) of the DMIC perturbations (where in the present case $(Fe)_i/a_{ii} = 1/2$ and $((F - F^t - A)e)_i = 0$ for all interior nodes), while the second is, for $1-\omega = 2\alpha$, i.e. for the same value of the target upper bound, identical to our estimate (4.6) of the RIC perturbations. This means that the method uses perturbations very close to to the DMIC ones when $\alpha \ll \zeta \ll \alpha^{-1}$, i.e. for isotropic to moderately anisotropic problems, and very close to to the RIC ones when $\alpha^{-1} \ll \max[\zeta, \zeta^{-1}]$, i.e. in strongly anisotropic cases. In other words, in this example, DRIC automatically shifts to the strongest method, DMIC or RIC, depending on the case at hand.

More generally, one may deduce from this illustrative example that, in isotropic regions, the DRIC perturbations are similar to the DMIC ones, the propagation of the diagonal dominance acting in much the same way to ensure $\alpha - \alpha_i^{(0)} \approx \alpha^2$ at most nodes, while, in anisotropic regions, the DRIC perturbations are much smaller because the latter method benefits from the smaller value of the neglected fill-in. On the other hand, it is obvious from (4.3) and (5.10) that the DRIC perturbations cannot exceed the RIC ones. Therefore, for the same value of the target upper bound, the behaviour of the lowest eigenvalues of DRIC will at least be as good as the best behaviour from DMIC and RIC, showing that the new method is at least as robust as the two previous ones.

Moreover, considering now problems presenting isotropic and anisotropic regions as well, one expects an effective improvement over both previous techniques, since DRIC requires much smaller perturbations than DMIC in anisotropic regions and much smaller than RIC in isotropic ones.

6. Summary of algorithms and basic properties

We summarize below the algorithms associated with each method. Since they only differ by the way in which some parameters are chosen, we can present them as a single algorithm with different options. To clarify the presentation, we use a pseudo-programming language and restrict ourselves to diagonally dominant matrices. (The general Stieltjes case requires only minor modifications.)

Algorithm 6.1. *Initialize : $u_{ij} := a_{ij}$ for all $1 \leq i \leq j \leq n$, and choose the parameters ω and α associated with the RIC and DMIC or DRIC methods, respectively, such that*

$$\begin{cases} -1 \leq \omega < 1 & \textit{for RIC} \\ 0 < \alpha < 1 & \textit{for DMIC} \\ 0 < \alpha \leq 1 & \textit{for DRIC} \end{cases}$$

Execute then, for $k = 1, \dots, n-1$:

```
$\alpha_k := 1 - \frac{-\sum_{i>k} u_{ki}}{u_{kk}}$
    if  IC : $\omega_k := 0$
    if  RIC : $\omega_k := \omega$
    if  MIC : $\omega_k := 1$
    if  DMIC: $\begin{cases} \omega_k := 1 \\ \textit{if } \alpha_k < \alpha \ : \ u_{kk} := -(1-\alpha)^{-1} \sum_{i>k} u_{ki} \end{cases}$
    if  DRIC : $\omega_k := \min\left(\frac{2(1-\alpha)}{(1-\alpha_k)} - 1 \, , \, 1\right)$
for all  $i > k$ such that $u_{ki} \neq 0$ :
        $u_{ii} := u_{ii} - \frac{u_{ki}^2}{u_{kk}}$
        for all $j > i$ such that $u_{kj} \neq 0$ :
            if fill-in is allowed in position $(i, j)$: $u_{ij} := u_{ij} - \frac{u_{ki} u_{kj}}{u_{kk}}$
```

$$\textit{otherwise:}\ \begin{cases} u_{ii} & := & u_{ii} - \omega_k \frac{u_{ki}u_{kj}}{u_{kk}} \\ u_{jj} & := & u_{jj} - \omega_k \frac{u_{ki}u_{kj}}{u_{kk}} \end{cases}$$

Note that the IC (MIC) method is the same as the RIC one with $\omega = 0$ ($\omega = 1$) while RIC with $\omega = -1$ and DRIC with $\alpha = 1$ are identical.

$A = (a_{ij})$ being an irreducible (nonstrictly) diagonally dominant Stieltjes matrix, the algorithm cannot fail and it results an upper triangular matrix $U = (u_{ij})$ which is regular except possibly for MIC factorizations (see [3,14] for existence criteria). Letting

$$B = U^t P^{-1} U \tag{6.1}$$

where $P = diag(U)$ be the associated preconditioner, it follows further from Theorem 5.1 that

$$\nu_{\max}(B^{-1}A) \le \begin{cases} 2 & \text{for IC} \\ \frac{2}{1-\omega} & \text{for RIC} \\ \alpha^{-1} & \text{for DMIC \& DRIC} \end{cases}$$

In the discrete PDE context, a practical rule for the determination of the parameter α is provided by

$$\alpha = \xi h_0 \tag{6.2}$$

with ξ close to 1 and h_0 given by (3.8). When h_0 is not available, one may use

$$\alpha \approx n^{-\frac{1}{d}} \tag{6.3}$$

where d is the dimension of the space in which the PDE is solved.

7. Numerical results

We have tested the different methods on the linear systems resulting from the 5-point finite difference approximation with uniform mesh size h of :

$$-\ \partial_x a_x \partial_x\, u \ -\ \partial_y a_y \partial_y\, u \ =\ f$$

on the unit square, with :

$$\begin{cases} u = 0 & \text{for } 0 \le x \le 1,\ y = 0 \\ \partial_n u = 0 & \text{on the remaining part of the boundary} \end{cases}$$

and

$$\text{Problem 1:}\quad a_x = a_y = \begin{cases} 100 & \text{in } (\frac{1}{4}, \frac{3}{4}) \times (\frac{1}{4}, \frac{3}{4}) \\ 1 & \text{elsewhere} \end{cases}$$

$$\text{Problem 2:}\quad \begin{cases} a_x & = & \begin{cases} 100 & \text{in } (\frac{1}{4}, \frac{3}{4}) \times (\frac{1}{4}, \frac{3}{4}) \\ 1 & \text{elsewhere} \end{cases} \\ a_y & = & a_x/100 \end{cases}$$

$$\text{Problem 3:}\quad \begin{cases} a_x &= \begin{cases} 100 & \text{in } (\frac{1}{4},\frac{3}{4}) \times (\frac{1}{4},\frac{3}{4}) \\ 1 & \text{elsewhere} \end{cases} \\ a_y &= a_x/10^4 \end{cases}$$

$$\text{Problem 4:}\quad \begin{cases} a_x &= 1 \\ a_y &= \begin{cases} 100 & \text{in } (\frac{1}{4},\frac{3}{4}) \times (\frac{1}{4},\frac{3}{4}) \\ 1 & \text{elsewhere} \end{cases} \end{cases}$$

$$\text{Problem 5:}\quad \begin{cases} a_x &= 1 \\ a_y &= \begin{cases} 10^4 & \text{in } (\frac{1}{4},\frac{3}{4}) \times (\frac{1}{4},\frac{3}{4}) \\ 1 & \text{elsewhere} \end{cases} \end{cases}$$

In each case, two right-hand sides are considered. First,

$$f = f_1 = \begin{cases} 100 & \text{in } (\frac{1}{4},\frac{3}{4}) \times (\frac{1}{4},\frac{3}{4}) \\ 0 & \text{elsewhere} \end{cases}$$

and next

$f = f_2$ such that the vector which samples the function $u = (1+x)^2$ $(1+y)(2-y)e^{xy}$ is the solution of the discrete system.

Note that the three categories of isotropic, purely anisotropic and mixed problems are represented in this set. We have deliberately disregarded model examples with constant coefficients, and further anisotropic problems for which the eigenvalues of the unpreconditioned system do not strongly depend on the anisotropy ratio since all the methods work quite well in such cases.

The results are reported in Tables 1–5 for the preconditioners associated with the lexicographic ordering (starting at the bottom left corner) and a factorization scheme without fill-in allowed ($\beta_{ij} = 0 \; \forall \; i \neq j$); δ appears in rule (4.4) for RIC and ξ in rule (6.2) for DMIC & DRIC. For $\varepsilon = 10^{-4}$ and $\varepsilon = 10^{-8}$, we display the number of iterations to reduce the relative residual error by a factor ε, It1 referring to the experiment with f_1 and It2 to that with f_2. From the computed eigenvalues, $\nu_{\min}^{(p)}$, $p = 1, \ldots, 7$ and $\nu_{\max}$, we have also determined the integer p_m for which the upper bound (3.10) $k_\varepsilon^{(p)}$ on the number of iterations is minimal, and we display the corresponding values.

For DMIC, RIC and DRIC it is fair to say that the convergence rate depends mainly on the highest eigenvalues and a few lowest eigenvalues, and, more particularly, on the reduced spectral condition number

$$\kappa^{(p_m)} = \frac{\nu_{\max}}{\nu_{\min}^{(p_m)}} \tag{7.1}$$

Note that the bound is less accurate when the method is less interesting or presents results relatively unstable with respect to the right-hand side, which comes from the fact that, when the preconditioning fails to produce a good clustering of all eigenvalues except a few small ones, the optimal convergence properties of the conjugate gradient method act more efficiently than what the bound (3.10) indicates, and thus partially compensates for the relatively unsatisfying behaviour of the method concerning the eigenvalue distribution.

Further comments :

- It is confirmed that the DMIC method has $\mathcal{O}(1)$ lowest eigenvalues with an $\mathcal{O}(h_0^{-1})$

Table 1. Numerical results for Problem 1

			$h_0^{-1} = 32$			
	$\varepsilon = 10^{-4}$			$\varepsilon = 10^{-8}$		
method	It1	It2	$k_\varepsilon^{(p_m)}$	It1	It2	$k_\varepsilon^{(p_m)}$
IC	35	40	$48^{(2)}$	51	57	$72^{(4)}$
MIC	33	24	$73^{(1)}$	51	43	$141^{(1)}$
DMIC, $\xi = 1.$	25	24	$29^{(2)}$	36	36	$46^{(2)}$
DMIC, $\xi = 2.$	24	24	$28^{(2)}$	36	35	$41^{(3)}$
RIC, $\delta = 1.$	20	24	$27^{(2)}$	31	33	$40^{(3)}$
RIC, $\delta = 2.$	22	25	$28^{(2)}$	32	35	$42^{(3)}$
DRIC, $\xi = 1.$	24	24	$29^{(2)}$	36	36	$45^{(2)}$
DRIC, $\xi = 2.$	23	23	$27^{(2)}$	34	34	$41^{(3)}$
DRIC, $\xi = 4.$	23	24	$27^{(2)}$	34	35	$41^{(2)}$

			$h_0^{-1} = 32$		
method	$\nu_{\min}^{(2)}$	$\nu_{\min}^{(3)}$	$\nu_{\max}$	$\kappa^{(2)}$	$\kappa^{(3)}$
DMIC, $\xi = 1.$	.838	.943	11.2	13.3	11.9
DMIC, $\xi = 2.$	.659	.816	6.94	10.5	8.51
RIC, $\delta = 1.$	.506	.609	4.80	9.50	7.89
RIC, $\delta = 2.$	.336	.420	3.52	10.5	8.39
DRIC, $\xi = 1.$	.868	.956	11.3	13.0	11.8
DRIC, $\xi = 2.$	.715	.878	7.16	10.0	8.16

			$h_0^{-1} = 128$			
	$\varepsilon = 10^{-4}$			$\varepsilon = 10^{-8}$		
method	It1	It2	$k_\varepsilon^{(p_m)}$	It1	It2	$k_\varepsilon^{(p_m)}$
IC	129	154	$163^{(2)}$	197	217	$275^{(4)}$
MIC	88	63	$267^{(1)}$	144	118	$515^{(1)}$
DMIC, $\xi = 1.$	52	49	$59^{(2)}$	78	76	$94^{(2)}$
DMIC, $\xi = 2.$	48	47	$54^{(2)}$	72	71	$81^{(3)}$
RIC, $\delta = 1.$	50	56	$62^{(2)}$	74	78	$93^{(4)}$
RIC, $\delta = 2.$	57	62	$71^{(2)}$	82	86	$106^{(4)}$
DRIC, $\xi = 1.$	52	49	$59^{(2)}$	72	75	$93^{(2)}$
DRIC, $\xi = 2.$	48	47	$54^{(2)}$	72	70	$81^{(3)}$
DRIC, $\xi = 4.$	50	49	$53^{(2)}$	73	73	$81^{(2)}$

			$h_0^{-1} = 128$		
method	$\nu_{\min}^{(2)}$	$\nu_{\min}^{(3)}$	$\nu_{\max}$	$\kappa^{(2)}$	$\kappa^{(3)}$
DMIC, $\xi = 1.$	.838	.942	47.1	56.2	50.0
DMIC, $\xi = 2.$	.661	.849	28.1	42.5	33.1
RIC, $\delta = 1.$	.177	.206	8.93	50.4	43.4
RIC, $\delta = 2.$	.097	.108	6.47	66.5	59.9
DRIC, $\xi = 1.$	.847	.947	47.3	55.8	49.9
DRIC, $\xi = 2.$	.677	.858	28.4	41.9	33.1

Table 2. Numerical results for Problem 2

	$h_0^{-1} = 32$					
	$\varepsilon = 10^{-4}$			$\varepsilon = 10^{-8}$		
method	It1	It2	$k_\varepsilon^{(p_m)}$	It1	It2	$k_\varepsilon^{(p_m)}$
IC	36	38	$57^{(6)}$	45	45	$72^{(6)}$
MIC	77	41	$281^{(1)}$	125	86	$543^{(1)}$
DMIC, $\xi = 1.$	44	42	$84^{(4)}$	57	54	$110^{(6)}$
DMIC, $\xi = 2.$	45	44	$96^{(6)}$	55	55	$123^{(6)}$
RIC, $\delta = 1.$	38	36	$63^{(2)}$	50	48	$88^{(4)}$
RIC, $\delta = 2.$	38	38	$57^{(4)}$	47	47	$77^{(4)}$
DRIC, $\xi = 1.$	38	37	$56^{(4)}$	48	47	$77^{(4)}$
DRIC, $\xi = 2.$	38	36	$55^{(4)}$	48	47	$72^{(5)}$
DRIC, $\xi = 4.$	37	37	$57^{(4)}$	45	45	$69^{(6)}$

	$h_0^{-1} = 32$				
method	$\nu_{\min}^{(4)}$	$\nu_{\min}^{(6)}$	$\nu_{\max}$	$\kappa^{(4)}$	$\kappa^{(6)}$
DMIC, $\xi = 1.$	.445	.722	20.0	45.0	39.4
DMIC, $\xi = 2.$	.176	.359	12.2	69.4	56.4
RIC, $\delta = 1.$	.936	1.00	26.8	28.6	26.8
RIC, $\delta = 2.$	.768	.975	15.7	20.5	16.2
DRIC, $\xi = 1.$	.913	1.00	19.4	21.3	19.4
DRIC, $\xi = 2.$	.692	8.89	12.0	17.3	13.5

	$h_0^{-1} = 128$					
	$\varepsilon = 10^{-4}$			$\varepsilon = 10^{-8}$		
method	It1	It2	$k_\varepsilon^{(p_m)}$	It1	It2	$k_\varepsilon^{(p_m)}$
IC	134	142	$210^{(6)}$	166	172	$266^{(6)}$
MIC	436	169	$567^{(1)}$	724	460	>999
DMIC, $\xi = 1.$	107	98	$148^{(4)}$	141	134	$190^{(6)}$
DMIC, $\xi = 2.$	117	112	$172^{(6)}$	149	145	$217^{(7)}$
RIC, $\delta = 1.$	99	98	$143^{(4)}$	134	131	$182^{(6)}$
RIC, $\delta = 2.$	99	95	$132^{(6)}$	130	129	$169^{(6)}$
DRIC, $\xi = 1.$	96	87	$129^{(4)}$	132	125	$175^{(4)}$
DRIC, $\xi = 2.$	96	92	$125^{(6)}$	128	124	$159^{(6)}$
DRIC, $\xi = 4.$	104	103	$141^{(6)}$	133	132	$178^{(7)}$

	$h_0^{-1} = 128$				
method	$\nu_{\min}^{(4)}$	$\nu_{\min}^{(6)}$	$\nu_{\max}$	$\kappa^{(4)}$	$\kappa^{(6)}$
DMIC, $\xi = 1.$	.461	.760	64.1	139.	84.4
DMIC, $\xi = 2.$	.185	.383	40.7	220.	106.
RIC, $\delta = 1.$	.538	.826	66.6	124.	80.7
RIC, $\delta = 2.$	.284	.637	39.8	140.	123.
DRIC, $\xi = 1.$	.705	.923	71.0	101.	77.1
DRIC, $\xi = 2.$	.343	.745	40.7	119.	54.6

Table 3. Numerical results for Problem 3

			$h_0^{-1} = 32$			
	$\varepsilon = 10^{-4}$			$\varepsilon = 10^{-8}$		
method	It1	It2	$k_\varepsilon^{(p_m)}$	It1	It2	$k_\varepsilon^{(p_m)}$
IC	35	4	$61^{(6)}$	36	36	$84^{(7)}$
MIC	14	5	$282^{(1)}$	29	17	$544^{(1)}$
DMIC, $\xi = 1.$	76	8	$638^{(7)}$	82	83	$794^{(7)}$
DMIC, $\xi = 2.$	106	11	$853^{(7)}$	112	112	>999
RIC, $\delta = 1.$	37	7	$91^{(4)}$	44	41	$128^{(4)}$
RIC, $\delta = 2.$	37	8	$78^{(5)}$	43	40	$104^{(5)}$
DRIC, $\xi = 1.$	37	8	$78^{(5)}$	43	41	$104^{(5)}$
DRIC, $\xi = 2.$	36	7	$65^{(6)}$	39	38	$83^{(6)}$
DRIC, $\alpha = 2.$	35	35	$61^{(6)}$	36	36	$88^{(7)}$

			$h_0^{-1} = 32$		
method	$\nu_{\min}^{(4)}$	$\nu_{\min}^{(6)}$	$\nu_{\max}$	$\kappa^{(4)}$	$\kappa^{(6)}$
DMIC, $\xi = 1.$	.0051	.011	21.5	4173.	1876.
DMIC, $\xi = 2.$	.0019	.0042	13.0	6882.	3084.
RIC, $\delta = 1.$	.967	1.00	61.2	63.4	61.2
RIC, $\delta = 2.$	.780	1.00	30.9	39.7	25.4
DRIC, $\xi = 1.$	.785	1.00	31.7	40.3	31.7
DRIC, $\xi = 2.$	.511	.999	15.9	31.1	15.9

			$h_0^{-1} = 128$			
	$\varepsilon = 10^{-4}$			$\varepsilon = 10^{-8}$		
method	It1	It2	$k_\varepsilon^{(p_m)}$	It1	It2	$k_\varepsilon^{(p_m)}$
IC	131	22	$232^{(6)}$	135	136	$295^{(6)}$
MIC	37	6	$564^{(1)}$	67	33	>999
DMIC, $\xi = 1.$	181	10	>999	194	193	>999
DMIC, $\xi = 2.$	192	19	>999	204	193	>999
RIC, $\delta = 1.$	145	26	$252^{(6)}$	185	156	$323^{(6)}$
RIC, $\delta = 2.$	141	19	$229^{(6)}$	169	154	$292^{(6)}$
DRIC, $\xi = 1.$	140	19	$229^{(6)}$	166	152	$292^{(6)}$
DRIC, $\xi = 2.$	138	14	$230^{(6)}$	155	152	$292^{(6)}$
DRIC, $\alpha = 2.$	131	24	$232^{(6)}$	135	135	$295^{(6)}$

			$h_0^{-1} = 128$		
method	$\nu_{\min}^{(4)}$	$\nu_{\min}^{(6)}$	$\nu_{\max}$	$\kappa^{(4)}$	$\kappa^{(6)}$
DMIC, $\xi = 1.$	.0053	.012	85.3	16.E3	7177.
DMIC, $\xi = 2.$	.0020	.0044	51.2	26.E3	12.E3
RIC, $\delta = 1.$	.552	1.00	241.	436.	241.
RIC, $\delta = 2.$	.286	.663	122.	426.	183.
DRIC, $\xi = 1.$	.293	.680	125.	426.	183.
DRIC, $\xi = 2.$	.148	.341	62.8	425.	184.

Table 4. Numerical results for Problem 4

	$h_0^{-1} = 32$					
	$\varepsilon = 10^{-4}$			$\varepsilon = 10^{-8}$		
method	It1	It2	$k_\varepsilon^{(p_m)}$	It1	It2	$k_\varepsilon^{(p_m)}$
IC	39	37	$56^{(3)}$	54	54	$81^{(4)}$
MIC	27	20	$37^{(1)}$	46	39	$71^{(1)}$
DMIC, $\xi = 1.$	28	28	$50^{(3)}$	42	42	$71^{(4)}$
DMIC, $\xi = 2.$	32	33	$51^{(4)}$	45	45	$67^{(5)}$
RIC, $\delta = 1.$	30	26	$38^{(3)}$	41	39	$57^{(4)}$
RIC, $\delta = 2.$	30	29	$41^{(3)}$	42	38	$62^{(3)}$
DRIC, $\xi = 1.$	24	23	$38^{(2)}$	38	38	$56^{(3)}$
DRIC, $\xi = 2.$	24	23	$33^{(3)}$	37	36	$49^{(3)}$
DRIC, $\xi = 4.$	26	25	$35^{(3)}$	38	37	$51^{(4)}$

	$h_0^{-1} = 32$				
method	$\nu_{\min}^{(3)}$	$\nu_{\min}^{(4)}$	$\nu_{\max}$	$\kappa^{(3)}$	$\kappa^{(4)}$
DMIC, $\xi = 1.$	.558	.789	17.5	31.3	22.1
DMIC, $\xi = 2.$	.236	.422	8.88	37.5	21.0
RIC, $\delta = 1.$	.560	.763	12.0	20.1	15.8
RIC, $\delta = 2.$	.406	.521	9.08	22.4	17.4
DRIC, $\xi = 1.$	.971	1.00	17.0	17.5	17.0
DRIC, $\xi = 2.$	.732	.941	9.73	13.3	9.73

	$h_0^{-1} = 128$					
	$\varepsilon = 10^{-4}$			$\varepsilon = 10^{-8}$		
method	It1	It2	$k_\varepsilon^{(p_m)}$	It1	It2	$k_\varepsilon^{(p_m)}$
IC	149	150	$215^{(3)}$	230	231	$315^{(4)}$
MIC	68	43	$108^{(1)}$	114	89	$209^{(1)}$
DMIC, $\xi = 1.$	64	59	$99^{(3)}$	95	91	$141^{(4)}$
DMIC, $\xi = 2.$	67	65	$96^{(4)}$	96	93	$130^{(5)}$
RIC, $\delta = 1.$	83	72	$102^{(3)}$	118	110	$154^{(4)}$
RIC, $\delta = 2.$	101	85	$125^{(3)}$	143	130	$187^{(4)}$
DRIC, $\xi = 1.$	57	52	$84^{(3)}$	88	84	$126^{(3)}$
DRIC, $\xi = 2.$	57	54	$74^{(4)}$	83	80	$106^{(4)}$
DRIC, $\xi = 4.$	63	60	$81^{(4)}$	89	86	$111^{(5)}$

	$h_0^{-1} = 128$				
method	$\nu_{\min}^{(3)}$	$\nu_{\min}^{(4)}$	$\nu_{\max}$	$\kappa^{(3)}$	$\kappa^{(4)}$
DMIC, $\xi = 1.$	.571	.802	70.8	124.	88.3
DMIC, $\xi = 2.$	.246	.446	34.9	142.	78.4
RIC, $\delta = 1.$	.224	.281	32.1	143.	114.
RIC, $\delta = 2.$	.122	.153	25.3	208.	165.
DRIC, $\xi = 1.$	.805	.960	71.0	88.1	73.9
DRIC, $\xi = 2.$	.427	.745	35.7	125.	106

Table 5. Numerical results for Problem 5

			$h_0^{-1} = 32$			
		$\varepsilon = 10^{-4}$			$\varepsilon = 10^{-8}$	
method	It1	It2	$k_\varepsilon^{(p_m)}$	It1	It2	$k_\varepsilon^{(p_m)}$
IC	37	33	$58^{(3)}$	57	50	$84^{(4)}$
MIC	27	10	$42^{(1)}$	45	31	$80^{(1)}$
DMIC, $\xi = 1.$	109	108	$331^{(7)}$	126	124	$421^{(7)}$
DMIC, $\xi = 2.$	125	126	$381^{(7)}$	137	136	$485^{(7)}$
RIC, $\delta = 1.$	31	18	$44^{(2)}$	43	34	$66^{(3)}$
RIC, $\delta = 2.$	31	21	$45^{(3)}$	47	35	$68^{(4)}$
DRIC, $\xi = 1.$	27	17	$44^{(2)}$	39	33	$58^{(3)}$
DRIC, $\xi = 2.$	26	23	$38^{(3)}$	39	32	$53^{(4)}$
DRIC, $\xi = 4.$	27	22	$40^{(3)}$	39	33	$54^{(4)}$

			$h_0^{-1} = 32$		
method	$\nu_{\min}^{(3)}$	$\nu_{\min}^{(5)}$	$\nu_{\max}$	$\kappa^{(3)}$	$\kappa^{(5)}$
DMIC, $\xi = 1.$	.0073	.024	18.0	2464.	759.
DMIC, $\xi = 2.$	.0027	.0087	8.86	3295.	1013.
RIC, $\delta = 1.$	.607	.833	15.2	25.1	18.3
RIC, $\delta = 2.$	.411	.669	11.1	26.9	16.5
DRIC, $\xi = 1.$	.983	1.00	17.6	17.9	17.6
DRIC, $\xi = 2.$	.615	.945	10.3	16.7	10.9

			$h_0^{-1} = 128$			
		$\varepsilon = 10^{-4}$			$\varepsilon = 10^{-8}$	
method	It1	It2	$k_\varepsilon^{(p_m)}$	It1	It2	$k_\varepsilon^{(p_m)}$
IC	156	122	$225^{(4)}$	236	207	$332^{(4)}$
MIC	69	23	$147^{(1)}$	115	72	$284^{(1)}$
DMIC, $\xi = 1.$	301	294	$678^{(7)}$	341	336	$862^{(7)}$
DMIC, $\xi = 2.$	391	389	$738^{(7)}$	434	436	$936^{(1)}$
RIC, $\delta = 1.$	100	66	$134^{(3)}$	145	114	$206^{(4)}$
RIC, $\delta = 2.$	116	82	$158^{(3)}$	171	132	$238^{(4)}$
DRIC, $\xi = 1.$	71	55	$114^{(4)}$	101	89	$152^{(5)}$
DRIC, $\xi = 2.$	73	59	$103^{(5)}$	102	86	$134^{(6)}$
DRIC, $\xi = 4.$	80	66	$110^{(5)}$	104	96	$140^{(7)}$

			$h_0^{-1} = 128$		
method	$\nu_{\min}^{(3)}$	$\nu_{\min}^{(5)}$	$\nu_{\max}$	$\kappa^{(3)}$	$\kappa^{(5)}$
DMIC, $\xi = 1.$	.0074	.025	82.8	11.E3	3339.
DMIC, $\xi = 2.$	.0028	.0093	36.2	13.E5	3901.
RIC, $\delta = 1.$	.227	.296	56.2	248.	190.
RIC, $\delta = 2.$	.123	.161	41.2	334.	256.
DRIC, $\xi = 1.$	.395	1.00	80.0	202.	80.0
DRIC, $\xi = 2.$	.204	.648	41.3	203.	63.8

highest eigenvalue. It behaves nicely on isotropic problems but cannot be recommended for strongly anisotropic problems where it presents, as expected, too many degenerate lowest eigenvalues ($\nu_{\min}^{(i)} \ll 1$) to produce a favourable convergence rate.

- The lowest eigenvalues of the RIC method act roughly like

$$\nu_{\min}^{(p)} \approx \frac{1}{1+c_p h_0^{-1}} \tag{7.2}$$

and even slightly worse. The asymptotic behaviour of RIC is thus, as expected, inferior to that of DMIC. However, even for the isotropic Problem 1 and the slightly anisotropic Problem 4, the method provides nearly as good results for values of h_0^{-1} up to 128. This is explained by an actual highest eigenvalue behaviour better than $\mathcal{O}(h_0^{-1})$ together with the fact that (7.2) is not so bad when h_0^{-1} is only moderately large and c_p is kept sufficiently small for reasonable p by an appropriate choice of δ. This reasoning also explains the undesirable sensitivity of the results with respect to the parameter.

- In strongly anisotropic problems, the highest eigenvalue associated with RIC behaves effectively like $\mathcal{O}(h_0^{-1})$, but the method is better than DMIC because the lowest eigenvalues are insensitive to the anisotropy ratio, as can be seen by comparing Table 2 with Table 3 and Table 4 with Table 5.
- Regarding these numerical results, MIC is not a bad method but has a rather unpredictable behaviour. For instance, the conditionings for Problems 2 and 3 are equally bad, but the method converges very quickly on Problem 3 because all eigenvalues are clustered around a few values.
- IC behaves relatively well on Problem 3 for which all methods present $\mathcal{O}(h_0^{-1})$ numbers of iterations. This may be explained by the eigenvalue estimates of RIC :

$$\begin{aligned}\nu_{\min}^{(p)} &\approx \frac{1}{1+c_p'(1-\omega)h_0^{-2}}\\ \nu_{\max} &\approx \frac{c'}{1-\omega}\end{aligned}$$

Hence, the estimated conditioning is minimal for $\omega = -1$ and nearly optimal for $\omega = 0$ where the method reduces to IC. One may thus recommend one of these methods or DRIC with $\alpha = 2$ in strongly and purely anisotropic problems. One should, however, be cautious with this conclusion because it does not extend to moderately anisotropic problems, while DRIC with the usual rule for the choice of α provides nearly as good results.

- As expected, the DRIC method acts like DMIC on isotropic problems and like RIC on purely anisotropic ones (with in both cases even a slight improvment). It further represents a noteworthy improvement on problems presenting isotropic and anisotropic regions as well. The highest eigenvalue is always $\mathcal{O}(h_0^{-1})$ while the lowest eigenvalues are $\mathcal{O}(1)$ in isotropic cases and otherwise about as given by (7.2). Hence the number of iterations is $\mathcal{O}(h_0^{-\frac{1}{2}})$ only for Problem 1. But, except for Problem 3, it remains however much better than $\mathcal{O}(h_0^{-1})$ because, on the one hand, as already discussed, (7.2) is not so bad for realistic h_0, while, on the other hand, it is seen that on Problems 2, 4 and 5 one has about :

$$\kappa^{(p_m)}\,|_{h_0^{-1}=128} \approx 4\kappa^{(p_m)}\,|_{h_0^{-1}=32}$$

Hence, the 'effective' spectral condition number is $\mathcal{O}(h_0^{-1})$, with the restriction that, unlike in isotropic cases, the value p_m increases with h_0^{-1}, and thus that the number of iterations increases slightly faster than $\mathcal{O}(h_0^{-\frac{1}{2}})$ because of the ε-independent term in (3.10). This is confirmed by the experiment: considering the actual number of iterations, one sees that the additional cost to reduce the error from 10^{-4} to 10^{-8} is effectively $\mathcal{O}(h_0^{-\frac{1}{2}})$.

- It follows from all these considerations, together with the theoretical developments of sections 2–5, that DRIC, in our opinion, best combines robustness and efficiency. Considering the experimental results obtained here on a set of 'difficult' problems, the method with $\xi = 2$ provides nearly always the best eigenvalue ratio and smallest actual number of iterations. Further it is not very sensitive to the choice of the parameter made according to the prescribed rule (6.2).

Acknowledgements

This work presents research results of the Belgian Incentive Program 'Information Technology'—Computer Science of the future, initiated by the Belgian State—Prime Minister's Service—Science Policy Office (Contract No. IT/IF/14). The scientific responsibility is assumed by its author.

Author's research is supported by the Belgian 'Fonds National de la Recherche Scientifique'.

We thank Professor R. Beauwens for his useful comments and suggestions and one of the referees for his numerous helpful remarks and corrections.

REFERENCES

1. O. Axelsson. A generalized SSOR method. *BIT*, 13, 443–467, 1972.
2. O. Axelsson. Condition number estimates for elliptic difference problems with anisotropy. Technical report, Department 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, 479–498, 1986.
5. O. Axelsson and G. Lindskog. On the rate of convergence of the preconditioned conjugate gradient method. *Numer. Math.*, 48, 499–523, 1986.
6. R. Beauwens. Upper eigenvalue bounds for pencils of matrices. *Lin. Alg. Appl.*, 62, 87–104, 1984.
7. R. Beauwens. On Axelsson's perturbations. *Lin. Alg. Appl.*, 68, 221–242, 1985.
8. R. Beauwens. Approximate factorizations with S/P consistently ordered M-factors. *BIT*, 29, 658–681, 1989.
9. R. Beauwens. Modified incomplete factorization strategies. In O. Axelsson and L. Kolotilina, editors. *Preconditioned Conjugate Gradient Methods*, pages 1–16. Lectures Notes in Mathematics No. 1457, Springer Verlag, New York, 1990.
10. R. Beauwens and L. Quenon. Existence criteria for partial matrix factorizations in iterative methods. *Siam J. Numer. Anal.*, 13, 615–643, 1976.
11. R. Beauwens and R. Wilmet. Conditioning analysis of positive definite matrices by approximate factorizations. *J. Comput. Appl. Math.*, 26, 257–269, 1989.
12. T. Dupont, R. Kendall, and H. Rachford. An approximate factorization procedure for solving self-adjoint elliptic difference equations. *SIAM J. Numer. Anal.*, 5, 559–573, 1968.
13. I. Gustafsson. Stability and rate of convergence of modified Cholesky factorization methods. Research Report 79.02R, Dept. of Computer Sciences, Chalmers University. of Technology and University of Goteborg, Goteborg, Sweden, 1979.

14. Y. Notay. Solving positive (semi)definite linear systems by preconditioned iterative methods. In O. Axelsson and L. Kolotilina, editors. *Preconditioned Conjugate Gradient Methods*, pages 105–125. Lectures Notes in Mathematics No. 1457, Springer Verlag, New York, 1990.
15. Y. Notay. *Résolution iterative de systèmes linéaires par factorisations approchées.* Ph.D. thesis, Service de Métrologie Nucléaire, Université Libre de Bruxelles, Brussels, Belgium, 1991.
16. Y. Notay. Conditioning of Stieltjes matrices by S/P consistently ordered approximate factorizations. *Appl. Numer. Math.*, 10, 381–396, 1992.
17. Y. Notay. On the robustness of modified incomplete factorization methods. *Inter. J. Computer Math.*, 40, 121–141, 1992.
18. Y. Notay. Upper eigenvalue bounds and related modified incomplete factorization strategies. In R. Beauwens and P. de Groen, editors. *Iterative Methods in Linear Algebra*, pages 551–562. North-Holland, Amsterdam, 1992.
19. Y. Notay. A new incomplete factorization method. In W. Hackbusch and G. Wittum, editors. *Incomplete Decomposition (ILU) - Algorithms, Theory and Applications*, pages 551–562. NNFM, V 41, Vieweg, Braunschweig, 1993.
20. Y. Notay. On the convergence rate of the conjugate gradients in presence of rounding errors. *Numer. Math.*, 65, 301–317, 1993.