COMPUTATIONAL & APPLIED MATHEMATICS

Volume 26, N. 1, pp. 95–128, 2007

ISSN 0101-8205
www.scielo.br/cam

Matrix differential equations and inverse preconditioners

JEAN-PAUL CHEHAB

Laboratoire de Mathématiques Paul Painlevé, UMR 8524
Université de Lille 1, Bât. M2, 59655 Villeneuve d'Ascq cedex, France; and
Laboratoire de Mathématiques, UMR 8628
Equipe Analyse Numérique et EDP, Bât. 425, Université Paris XI
91405 Orsay cedex, France

E-mail: chehab@math.univ-lille1.fr

Abstract. In this article, we propose to model the inverse of a given matrix as the state of a proper first order matrix differential equation. The inverse can correspond to a finite value of the independent variable or can be reached as a steady state. In both cases we derive corresponding dynamical systems and establish stability and convergence results. The application of a numerical time marching scheme is then proposed to compute an approximation of the inverse. The study of the underlying schemes can be done by using tools of numerical analysis instead of linear algebra techniques only. With our approach, we recover some known schemes but also introduce new ones. We derive in addition a masked dynamical system for computing sparse inverse approximations. Finally we give numerical results that illustrate the validity of our approach.

Mathematical subject classification: 65F10, 65F35, 65L05, 65L12, 65L20, 65N06.

Key words: matrix differential equation, numerical schemes, numerical linear algebra, preconditioning.

1 Introduction

The modern methods we have at our disposal for solving linear systems of equations such as the preconditioned versions of GMRES [19] or BI-CGSTAB [21],

#657/06. Received: 14/III/06. Accepted: 20/X/06.

are robust and apply to many situations; they are intensively used for the numerical solution of large sparse linear systems coming out from PDE discretisation. For this reason the preconditioning is still now a central topic in numerical linear algebra since it is the common universal approach to accelerate the solution of a linear system by an iterative method. Preconditioning a matrix practically leads to improve its spectral properties, by, e.g., concentrating the spectrum of the preconditioned matrix. There is not an efficient general method for building a preconditioner for a given nonsingular matrix : a large number of approaches have been developed depending on the properties of the considered matrix, surveys are presented, e.g., in [4, 19].

Let $\mathbf{P}$ be a $n \times n$ regular matrix and $b \in \mathbb{R}^n$. The preconditioning of the numerical solution of the linear system

$$\mathbf{P}u = b, \tag{1.1}$$

(with a descent method) consists in solving at each step of the iterative process, an additional linear system

$$\mathbf{K}v = c, \tag{1.2}$$

where $\mathbf{K}$ is the preconditioning matrix. Of course, the additional computation carried by the solution of (1.2) is convenient when this system is easy to solve and when $\mathbf{K}$ (respectively $\mathbf{K}^{-1}$) resembles $\mathbf{P}$ (resp. $\mathbf{P}^{-1}$).

When the preconditioning of (1.1) is obtained by approaching $\mathbf{P}$, system (1.2) must be easy to solve. In the other case, the approximation of $\mathbf{P}^{-1}$, which defines the so-called inverse preconditioner, leads to a trivial solution of (1.2).

Inverse preconditioners can be built in many ways: by minimizing an objective functional (the Frobenius norm of the residual, [10]), by incomplete sparse factorization [9], or also by building proper convergent sequences, see [5, 10] in which the authors have presented sequences generated by descent methods such as MINRES or Newton-like schemes. The polynomial preconditioning, which consists in approaching the inverse of a matrix by a proper polynomial, has been developed and implemented for parallel computers, see [19].

In this paper we propose to model the inverse of a regular matrix as a state of a first order Matrix Differential Equation (MDE). This state can correspond to the solution of the MDE for a finite value of the independent variable, but

also to an equilibrium point, depending on the equation. In such a way, the implementation of any numerical integration can produce an approximation of $\mathbf{P}^{-1}$, say an inverse preconditioner of $\mathbf{P}$. The underlying schemes involve at least one multiplication of two matrices at each iterations so the terms of the sequence of inverse preconditioners become denser even if the matrix $\mathbf{P}$ is sparse. However, it is possible to derive a masked dynamical system which preserve a given density pattern making our method suitable for computing sparse inverse preconditioners. From a technical point of view, the advantage of the dynamical system approach is to use the classical tools of numerical analysis of differential equations for studying the processes.

This approach is very flexible since the construction of the numerical scheme is subjected to the choice of the modelling ODE and of a time marching scheme; it allows also to study the schemes by using classical mathematical tools of ODE analysis and of numerical analysis of ODEs [15, 16, 20]. We mention that the use of differential equation modeling for solving systems of equations, including linear algebra problems, was considered in other situations: in [14, 17] for generating flows of matrices that preserve eigenvalues, singular values; in [8] for generating fixed point methods, the solution being defined as a stable steady state; in [13] for computing the square root of a matrix, by integrating a Riccati matrix differential equation (see also R. Bellman's book [3], chap 10).

The article is organized as follows: in Section 2 we study a Riccati differential equation whose solution is the inverse at finite time of one of the data; the derivation, the stability analysis and the study of approximation scheme is given. In Section 3 we consider Matrix differential equations for which one of the steady states is the inverse of a datum of the equation. In Section 4 we concentrate on the construction of sparse inverse preconditioners by considering the numerical integration of a so-called masked differential equation, when a sparsity pattern is fixed; error estimates are derived. Finally in Section 5 we give numerical illustration on the solution of linear systems when using the approximate inverses built in the previous sections.

2 Inverse at finite time

2.1 Derivation of the equation

Let $P(t)$ and $Q(t)$ be two square matrices, depending on the scalar variable t which belongs to an interval I. We assume that the coefficients of both P and Q are differentiable functions of t. We have

$$\frac{dP(t)Q(t)}{dt} = \frac{dP(t)}{dt}Q(t) + P(t)\frac{dQ(t)}{dt}, \quad \forall t \in I.$$

Assume that $P(t)$ is regular, i.e. invertible, for all t in I and consider the particular situation $Q(t) = P^{-1}(t)$, $\forall t \in I$. Then we have

$$\frac{dP(t)Q(t)}{dt} = 0.$$

So,

$$\begin{aligned} \frac{dQ(t)}{dt} &= -Q(t)\frac{dP(t)}{dt}Q(t) \text{ or, equivalently,} \\ \frac{dP(t)}{dt} &= -P(t)\frac{dQ(t)}{dt}P(t), \end{aligned} \tag{2.1}$$

for all $t \in I$. If $P(t)$ is supposed to be known, then $Q(t)$ can be computed by integrating the differential matrix equation:

$$\begin{cases} \frac{dQ(t)}{dt} = -Q(x)\frac{dP(t)}{dt}Q(t), \quad t \in I, \\ Q(0) = P^{-1}(0). \end{cases} \tag{2.2}$$

Q is hence the solution of a matrix Riccati differential equation.

Let now $\mathbf{P}$ be a regular $n \times n$ matrix and Id, the $n \times n$ identity matrix. Now, the basic idea consists in defining $P(t)$ as a simple path function of regular matrices between $P(0)$ easy to invert ($Q(0) = P^{-1}(0)$) and $P(1) = \mathbf{P}$. We consider the function

$$P(t) = (1-t)Id + t\mathbf{P}, \quad t \in [0,1]. \tag{2.3}$$

Assume that $P(t)$ is invertible for all t in $[0,1]$. The Matrix $Q(t) = P^{-1}(t)$ satisfies the Cauchy problem

$$\begin{cases} \frac{dQ(t)}{dt} = -Q(t)(\mathbf{P} - Id)Q(t) \quad t \in I, \\ Q(0) = Id, \end{cases} \tag{2.4}$$

and $\mathbf{P}^{-1} = Q(1)$; we assume that $[0, 1] \subset \bar{I}$.

We have the following result:

Lemma 1. *$P(t)$ is regular for all t in $[0, 1]$ iff $SP(\mathbf{P}) \subset \mathbb{R}^2 \setminus \{(t, 0), t \leq 0\}$ where $SP(\mathbf{P})$ denotes the spectrum of $\mathbf{P}$.*

Proof. The eigenvalues of $P(t)$ are the numbers

$$S(t) = (1 - t) + t\lambda \neq 0, \ \lambda \in SP(\mathbf{P}).$$

Taking the real and the imaginary parts of this expression, we have

$$\phi_1(t) = (1 - t) + t\Re(\lambda), \phi_2(t) = t\Im(\lambda).$$

Let us look to necessary and sufficient conditions for having $\phi_1(t) = \phi_2(t) = 0$ for same t. By continuity, it is easy to see that $\phi_1(t) = 0$ if and only if $\Re(\lambda) \leq 0$. $\phi_2(t)$ vanishes for $t = 0$ (but $P(0) = Id$) or for $\Im(\lambda) = 0$.

In conclusion $S(t)$ vanishes if and only if there exists $\lambda \in SP(\mathbf{P})$ such that $\Re(\lambda) \leq 0$ and $\Im(\lambda) = 0$. □

Particularly, Lemma 1 applies when $\mathbf{P}$ is positive definite, such as, e.g., discretization matrices of elliptic operators.

Remark 1. We can consider $P(t) = (1 - t)\mathbf{P}_0 + t\mathbf{P}$ with $\mathbf{P}_0$ a preconditioner of $\mathbf{P}$. Of course, in this case, Lemma 1 applies replacing $\mathbf{P}$ by $\mathbf{P}_0^{-1}\mathbf{P}$. Same considerations can be made with a more general path

$$P(t) = (1 - \phi(t))\mathbf{P}_0 + \phi(t)\mathbf{P},$$

with

$$\phi(t) : [0, 1] \to [0, 1], \quad \phi \in C^1([0, 1]), \quad \phi'(t) > 0, t \in]0, 1[.$$

2.2 Stability results

Let us now give some notations and technical results which will be used along the article.

2.2.1 *Matrix norms*

Let M be a $n \times n$ matrix. We denote by $\|M\|$ any matrix norm of M and particularly, $\|M\|_2$ and $\|M\|_F$ the 2-norm and the Frobenius norm of M, respectively. We shall use also the notation $\|v\|_2$ for the 2 norm of a vector of $\mathbb{R}^n$, there will be no ambiguity in practice.

2.2.2 *Hadamard Matrix Product*

We denote by $R * M$ the Hadamard product of R and M:

$$(M * R)_{i,j} = R_{i,j} M_{i,j}.$$

2.2.3 *Matrix scalar product*

We will use the following scalar product:

$$\ll R, M \gg = \sum_{i,j=1}^{n} R_{i,j} M_{i,j},$$

which coincide with the sum of the coefficient of the Hadamard product of R and M. We also use the euclidean scalar product of vector of $\mathbb{R}^n$ that we note by $< \cdot, \cdot >$.

We begin with the following very simple but useful technical result:

Lemma 2. *Let R and S be two $n \times n$ matrices. We have the inequalities*

(i) $\displaystyle\sum_{i,j=1}^{n} |(R^2 * R^2)_{i,j}| \leq \|R\|_F^4,$

(ii) $\displaystyle\sum_{i,j=1}^{n} |(R^2 * S)_{i,j}| \leq \|R\|_F^2 \|S\|_F,$

(iii) $\|S\|_2 \leq \|S\|_F.$

Proof. Assertion (i) follows from a simple application of Cauchy-Schwarz inequality.

Let us prove (ii). We have

$$\begin{aligned}|(R^2 * S)_{i,j}| &= |\left(\sum_{k=1}^{n} R_{i,k} R_{k,j}\right) S_{i,j}|,\\ &\qquad \text{(using Cauchy Schwarz inequality in } \mathbb{R}^n),\\ &\leq \left(\sum_{k=1}^{n} R_{i,k}^2\right)^{1/2} \left(\sum_{k=1}^{n} R_{k,j}^2\right)^{1/2} |S_{i,j}|.\end{aligned}$$

We now take the sum of these terms for $i, j = 1, \cdots n$. We obtain

$$\begin{aligned}\sum_{i,j=1}^{n} |(R^2 * S)_{i,j}| &\leq \sum_{i,j=1}^{n} |S_{i,j}| \left(\sum_{k=1}^{n} R_{i,k}^2\right)^{1/2} \left(\sum_{k=1}^{n} R_{k,j}^2\right)^{1/2},\\ &\qquad \text{(using Cauchy Schwarz inequality in } \mathbb{R}^{n^2}),\\ &\leq \left(\sum_{i,j=1}^{n} S_{i,j}^2\right)^{1/2} \left(\sum_{i,j=1}^{n} (\sum_{k=1}^{n} R_{i,k}^2 \sum_{k=1}^{n} R_{k,j}^2\right)^{1/2},\\ &\leq \|R\|_F^2 \|S\|_F.\end{aligned}$$

Assertion (iii) is classical and obtained by applying Cauchy-Schwarz inequality to $\|Sv\|_2^2 = \sum_{i=1}^{n} \left(\sum_{j=1}^{n} S_{i,j} v_j\right)$, for $v \in \mathbb{R}^n$, $\|v\|_2 = 1$. □

At this point, we can establish a stability result:

Proposition 1. *Assume that* $Id - \mathbf{P}\mathbf{P}_0^{-1}$ *satisfy the assumptions of Lemma* 1. *We set* $S(t) = (\mathbf{P} - \mathbf{P}_0)Q(t)$, *where* $Q(t)$ *solves the equation*

$$\begin{cases} \dfrac{dQ(t)}{dt} = -Q(t)(\mathbf{P} - \mathbf{P}_0)Q(t) \quad t \in I, \\ Q(0) = \mathbf{P}_0^{-1}. \end{cases} \tag{2.5}$$

Assume that $\|S(0)\|_F < 1$. *Then* $S(t)$ *exists for all* t *in* $[0, 1]$ *and*

$$\|S(t)\|_F \leq \frac{1}{\left(1 - \dfrac{1}{\|Id - \mathbf{P}\mathbf{P}_0^{-1}\|_F}\right)^2}.$$

Proof. Multiplying on the left each term of (2.7) by $\mathbf{P} - \mathbf{P}_0$, we obtain

$$\frac{dS(t)}{dt} = -S(t)^2.$$

We now take the Hadamard product of each term with $S(t)$, and consider the sum of all indices $i, j = 1, \cdots, n$. We find

$$\begin{aligned} \frac{1}{2}\frac{d\|S(t)\|_F^2}{dt} &= -\sum_{i,j=1}^{n} (S(t)^2)_{i,j} S(t)_{i,j}, \\ &\qquad \text{((ii) of Lemma 2)}, \\ &\leq \|S(t)\|_F^3. \end{aligned}$$

Hence, $\|S(t)\|_F^2 \leq y(t)$, where $y(t)$ is the solution of the differential equation

$$\begin{cases} \dfrac{dy(t)}{dt} = 2y(t)^{3/2} \\ y(0) = \|S(0)\|_F^2 \end{cases} \tag{2.6}$$

We find

$$y(t) = \frac{1}{\left(\frac{1}{\sqrt{y(0)}} - t\right)^2} = \frac{1}{\left(\frac{1}{\|S(0)\|_F} - t\right)^2}.$$

Since $\|S(0)\|_F < 1$, $y(t)$ remains bounded and $y(t) \leq y(1)$. □

Another stability result can be derived when assuming both $\mathbf{P}$ and $\mathbf{P}_0$ to be symmetric, positive definite (SPD). More precisely we have the next result:

Lemma 3. *Assume that both* $\mathbf{P}$ *and* $\mathbf{P}_0$ *are SPD. Then* $Q(t) = P(t)^{-1}$ *is SPD for all* $t \in [0, 1]$.

Proof. It suffices to prove that $P(t)$ is SPD for all $t \in [0, 1]$. The proof is straightforward starting from the definition of $P(t)$:

$$P(t) = (1 - t)\mathbf{P}_0 + t\mathbf{P}.$$ □

2.3 *Construction of an inverse preconditioner by numerical integration*

Let us subdivide $I = [0, 1]$ into N subintervals of the same length $\delta = 1/N$, the step-length. The application of any (stable) time marching scheme to equation (2.4) generates a sequence $Q_k, k = 1, \cdots, N$, Q_k being an approximation of $Q(k/N)$. In particular, since $Q(1) = \mathbf{P}^{-1}$, Q_N will be an inverse preconditioner for the matrix $\mathbf{P}$.

For each time integration scheme, a method for computing a preconditioner is derived. We consider the following cases.

Forward Euler Scheme
We consider the sequence

$$\begin{cases} Q_0 = Id \\ \texttt{For k=0,...,N-1} \\ Q_{k+1} = Q_k - \frac{1}{N} Q_k(\mathbf{P} - Id)Q_k \end{cases} \tag{2.7}$$

We have $Q_N \simeq \mathbf{P}^{-1}$.

Second order Adams Bashforth (AB2)
We consider the sequence

$$\begin{cases} Q_0 = Id \\ \texttt{Computation of } Q_1 \texttt{ by RK2} \\ K_0 = Q - \frac{1}{2N} Q_0(\mathbf{P} - Id)Q_0 \\ Q_1 = Q_0 - \frac{1}{N} K_0(\mathbf{P} - Id)K_0 \\ \texttt{For k=1,...,N-1} \\ Q_{k+1} = Q_k - \frac{1}{2N} \left(3Q_k(\mathbf{P} - Id)Q_k - Q_{k-1}(\mathbf{P} - Id)Q_{k-1}\right) \end{cases} \tag{2.8}$$

We have $Q_N \simeq \mathbf{P}^{-1}$.

Of course, further methods can be derived by considering, e.g., Runge-Kutta or more general Adams-Bashforth schemes, but, in practice, it is important to find a compromise between the accuracy and the cost of the computation, since each iteration requires (at least) the multiplication of three matrices.

It is easy to see that the above schemes consist in approaching $\mathbf{P}^{-1}$ with a polynomial of $\mathbf{P}$, $\mathcal{P}_N(\mathbf{P})$, whose coefficients are matrix independent. The degree of $\mathcal{P}_N(\mathbf{P})$ grows exponentially with N. For instance, we have the following expressions of $\mathcal{P}_N(\mathbf{P})$ when it is seen as a one variable function:

Euler's

$$N=1 \qquad \mathcal{P}_N(t) = 2-t$$

$$N=2 \qquad \mathcal{P}_N(t) = -\frac{1}{8}(t-3)(t^2-4t+7)$$

$$N=3 \qquad \mathcal{P}_N(t) = -\frac{1}{2187}(t-4)(t^2-5t+13)(t^4-10t^3+42t^2-85t+133)$$

AB2's

$$N=1 \qquad \mathcal{P}_N(t) = \frac{13}{4}-\frac{15}{4}t+\frac{7}{4}t^2-\frac{1}{4}t^3$$

$$\begin{aligned} N=2 \qquad \mathcal{P}_N(t) = {} & \frac{16019}{4096}-\frac{25173}{4096}t+\frac{20815}{4096}t^2-\frac{10217}{4096}t^3+\frac{3225}{4096}t^4 \\ & -\frac{639}{4096}t^5+\frac{69}{4096}t^6-\frac{3}{4096}t^7 \end{aligned}$$

$$\begin{aligned} N=3 \qquad \mathcal{P}_N(t) = {} & \frac{515661916}{1088391168}-\frac{3581936773}{362797056}t+\frac{1484035553}{120932352}t^2 \\ & -\frac{11394203831}{1088391168}t^3+\frac{2404253335}{362797056}t^4-\frac{391144243}{120932352}t^5 \\ & +\frac{1352868157}{1088391168}t^6-\frac{46022507}{120932352}t^7+\frac{1251827}{13436928}t^8 \\ & -\frac{19757677}{1088391168}t^9+\frac{1008073}{362797056}t^{10}-\frac{39389}{120932352}t^{11} \\ & +\frac{30455}{1088391168}t^{12}-\frac{595}{362797056}t^{13}+\frac{7}{120932352}t^{14} \\ & -\frac{1}{1088391168}t^{15} \end{aligned}$$

These polynomials are approximations of the function $t \mapsto \frac{1}{t}$, as illustrated in Figure 1.

Remark 2. Both of the numerical integration schemes given above lead to compute the approximate inverse with a polynomial of $\mathbf{P}$. This is hence a polynomial preconditioning. Several approaches of polynomial preconditioning have been proposed: they are based on truncated Neumann series [11] or based on

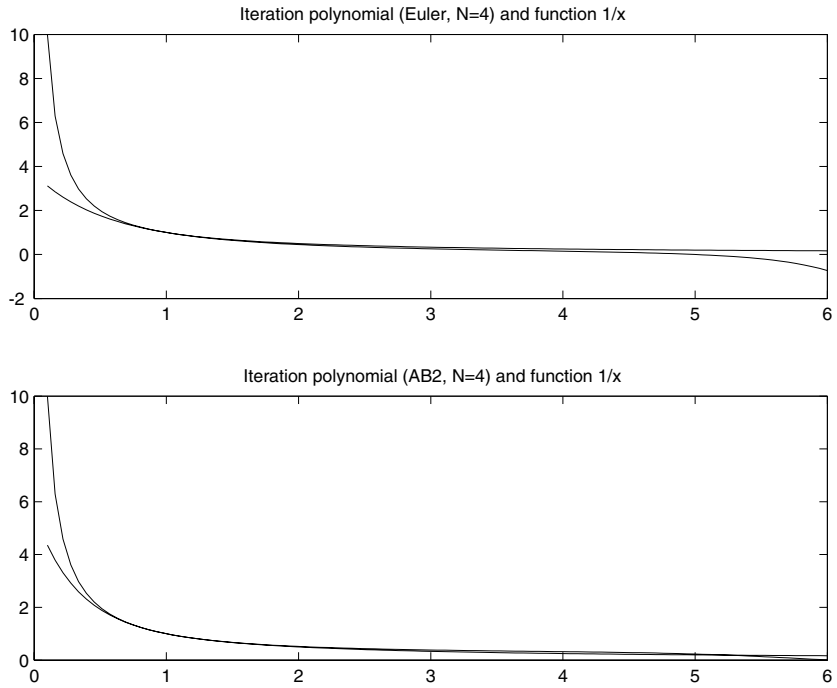

Figure 1 – The function $\frac{1}{x}$ and the iteration polynomials a) Euler, $N = 4$, b) AB2, $N = 4$.

orthogonal polynomial [1], see [2, 4] for a review. However, the point of view here is different and the underlying polynomial are also different.

At this point we give a convergence result for the Euler method (scheme (2.9)). More precisely, we give a consistency error bound. We have the

Theorem 1. *Assume that* $\mathbf{P} - \mathbf{P}_0$ *is regular and satisfies the hypothesis of Lemma* 1. *Let* Q_N, *be the approximation of* $Q(1)$ *obtained by replacing*

$$\frac{dQ(t)}{dt}\left(\frac{k}{N}\right) \quad by \quad \frac{Q\left(\frac{k}{N}\right) - Q\left(\frac{k-1}{N}\right)}{\frac{1}{N}}.$$

Assume that the solution of (2.7) *is* C^2. *Then*

$$\|Q(1) - Q_N\|_2 \leq \frac{1}{2N}\|(\mathbf{P} - \mathbf{P}_0)^{-1}\|_2 \frac{1}{\left(1 - \frac{1}{|Id - \mathbf{P}\mathbf{P}_0^{-1}|_F}\right)^2}.$$

Proof. We have

$$\begin{aligned} Q(1) - Q(0) &= \int_0^1 \frac{dQ(t)}{dt} dt \\ &= \sum_{k=1}^{N} \int_{\frac{k-1}{N}}^{\frac{k}{N}} \frac{dQ(t)}{dt} dt \end{aligned}$$

Let k be fixed and let $t \in]\frac{k-1}{N}, \frac{k}{N}[$. There exists $t_0 \in]\frac{k-1}{N}, t[$ such that

$$\begin{aligned} \frac{dQ(t)}{dt} &= \frac{dQ}{dt}\left(\frac{k-1}{N}\right) + \left(t - \frac{k-1}{N}\right)\frac{d^2Q}{dt^2}(t_0), \\ &\quad (Q(t) \text{ being solution of (2.7)}) \\ &= -Q\left(\frac{k-1}{N}\right)(\mathbf{P}-\mathbf{P}_0)Q\left(\frac{k-1}{N}\right) \\ &\quad + 2\left(t - \frac{k-1}{N}\right) Q(t_0)(\mathbf{P}-\mathbf{P}_0)Q(t_0)(\mathbf{P}-\mathbf{P}_0)Q(t_0), \\ &= -Q\left(\frac{k-1}{N}\right)(\mathbf{P}-\mathbf{P}_0)Q\left(\frac{k-1}{N}\right) \\ &\quad + 2\left(t - \frac{k-1}{N}\right)(\mathbf{P}-\mathbf{P}_0)^{-1}\left((\mathbf{P}-\mathbf{P}_0)Q(t_0)\right)^3. \end{aligned}$$

Hence,

$$\begin{aligned} &\| \int_{\frac{k-1}{N}}^{\frac{k}{N}} \frac{dQ(t)}{dt} dt - \frac{1}{N}\left(-Q\left(\frac{k-1}{N}\right)(\mathbf{P}-\mathbf{P}_0)Q\left(\frac{k-1}{N}\right)\right) \|_2 \\ &\qquad \leq \frac{1}{N^2}\|(\mathbf{P}-\mathbf{P}_0)^{-1}\|_2 \sup_{t\in[0,1]} \|(\mathbf{P}-\mathbf{P}_0)Q(t)\|_2^3. \end{aligned}$$

Therefore, we have the estimate:

$$\begin{aligned} \|Q(1) - Q_N\|_2 &\leq \sum_{k=1}^{N} \frac{1}{N^2}\|(\mathbf{P}-\mathbf{P}_0)^{-1}\|_2^2 \sup_{t\in[0,1]} \|(\mathbf{P}-\mathbf{P}_0)Q(t)\|^3, \\ &\leq \frac{1}{N}\|(\mathbf{P}-\mathbf{P}_0)^{-1}\|_2 \sup_{t\in[0,1]} (\|\mathbf{P}-\mathbf{P}_0)Q(t)\|_F^3), \\ &\leq \frac{1}{N}\|(\mathbf{P}-\mathbf{P}_0)^{-1}\|_2 y(1). \end{aligned}$$

where $y(t)$ is solution of (2.8), and

$$y(1) = \frac{1}{\left(1 - \frac{1}{\|Id - \mathbf{P}\mathbf{P}_0^{-1}\|_F}\right)^2}.$$

□

The Euler scheme preserves the symmetry. In particular we can prove that for $\mathbf{P}$, $\mathbf{P}_0$, $\mathbf{P}-\mathbf{P}_0$ SPD, and for N large enough, the matrices Q_k generated by (2.9) are SPD for $k = 0, \cdots, N$.

3 Inverse matrix as steady state

3.1 The equations

Another way to reach $\mathbf{P}^{-1}$ is to consider differential equations for which one of the steady states is $Q = \mathbf{P}^{-1}$. We consider the two following equations:

$$\begin{cases} \dfrac{dQ(t)}{dt} = Q(t)\left(Id - \mathbf{P}Q(t)\right), \\ Q(0) = Q_0, \end{cases} \tag{3.11}$$

which is a Riccati matrix differential equation and its linearized version

$$\begin{cases} \dfrac{dQ(t)}{dt} = Id - \mathbf{P}Q(t), t \geq 0 \\ Q(0) = Q_0. \end{cases} \tag{3.12}$$

In both equations $\mathbf{P}^{-1}$ is a steady state.

Remark 3. We can also proceed as in Section 2: we consider equation (2.4) with the path function $P(t)$:

$$P(t) = (1 - e^{-t})\mathbf{P} + e^{-t}\mathbf{P}_0.$$

It is easy to see that $P(t)$ is invertible for all $t \geq 0$ iff $\mathbf{P}\mathbf{P}_0^{-1}$ satisfies the assumptions of Lemma 1, see also Remark 1. The differential equation satisfied by $Q(t)$ is then

$$\begin{cases} \dfrac{dQ(t)}{dt} = e^{-t}Q(t)(\mathbf{P} - \mathbf{P}_0)Q(t), t \geq 0, \\ Q(0) = Q_0. \end{cases}$$

We now give sufficient conditions for obtaining the convergence $\lim_{t\to+\infty} Q(t) = \mathbf{P}^{-1}$. We propose two approaches. The first one consists in deriving bounds of the Frobenius norm of the solution, assuming that the initial data is close enough to the steady state. The second one concentrates on the symmetric definite positive case.

We begin with the following result:

Proposition 2. *Let $Q(t)$ be the solution of the matrix differential equation* (3.11). *Assume that $\|Id - \mathbf{P}Q_0\|_F < 1$. Then $\lim_{t\to\infty} Q(t) = \mathbf{P}^{-1}$.*

Proof. The matrix $R(t) = Id - \mathbf{P}Q(t)$ satisfies the equation

$$\frac{dR(t)}{dt} = -R(t) + R^2(t).$$

Then, taking the Hadamard product of each terms with $R(t)$ and taking the sum of all the coefficients, we obtain

$$\frac{1}{2}\frac{d\|R(t)\|_F^2}{dt} + \|R(t)\|_F^2 = -\sum_{i,j}(R^2 * R)_{i,j}.$$

By the first assertion of Lemma 2 (with $S = R$), we have

$$\frac{1}{2}\frac{d\|R(t)\|_F^2}{dt} + \|R(t)\|_F^2 \leq \|R(t)\|_F^3.$$

From the previous inequality, we infer that $\|R(t)\|_F^2$ is bounded from below by the solution of the scalar differential equation

$$\begin{cases} \dfrac{dy(t)}{dt} = -2y(t)(1 - \sqrt{y(t)}) \\ y(0) = \|Id - \mathbf{P}Q_0\|_F^2 \end{cases}$$

We have

$$y(t) = \frac{1}{\left(1 + \left(\dfrac{1}{\sqrt{y(0)}} - 1\right)e^t\right)^2},$$

hence the result. □

This last results insures the existence of solution and the convergence to $\mathbf{P}^{-1}$ for initial conditions closed enough to the steady state; however no other properties of $\mathbf{P}$ or of $\mathbf{Q}_0$ are required. We now give an existence and a convergence result in the symmetric positive definite case. We have the

Proposition 3. *Assume that $\mathbf{P}$ and $Q(0)$ are SPD matrices. Then $Q(t)$ is SPD for all $t \geq 0$ and $\lim_{t\to\infty} Q(t) = \mathbf{P}^{-1}$.*

Proof. If $Q(t)$ is regular for all t, then $U(t) = Q(t)^{-1}$ satisfies the differential equation

$$\begin{cases} \dfrac{dU(t)}{dt} = -U(t) + \mathbf{P}, \ x \geq 0, \\ U(0) = Q^{-1}(0). \end{cases}$$

We prove the proposition by studying $U(t)$. We have

$$U(t) = (1 - e^{-t})\mathbf{P} + e^{-t}U_0,$$

from which we infer that $U(t)$ is SPD for all $t \geq 0$. Indeed, $U(t)$ is symmetric as sum of symmetric matrices, and for every $w \in \mathbb{R}^n$, we have

$$\langle U(t)w, w\rangle = (1 - e^{-t})\langle \mathbf{P}w, w\rangle + e^{-t}\langle \mathbf{P}_0 w, w\rangle > 0,$$

since both $\mathbf{P}$ and $\mathbf{P}_0$ are assumed to be positive definite. Furthermore, we have immediately $\lim_{t\to\infty} U(t) = \mathbf{P}$. Therefore $U(t)$ is SPD for all $t \geq 0$. In conclusion $Q(t)$ exists and is SPD for all $t \geq 0$ and $\lim_{x\to\infty} Q(t) = \mathbf{P}^{-1}$. □

Proposition 4. *Let $Q(t)$ be the solution of* (3.12). *Assume $\mathbf{P}$ is positive definite. Then $\lim_{x\to\infty} Q(t) = \mathbf{P}^{-1}$. Moreover if $\mathbf{P}$ and Q_0 are SPD and commute with $\mathbf{P}$, then $\mathbf{Q(t)}$ is also SPD for all $t \geq 0$.*

Proof. As usual, we introduce the residual matrix $R(t) = Id - \mathbf{P}Q(t)$ which here satisfies the equation

$$\frac{dR(t)}{dt} = -\mathbf{P}R(t),$$

whose solution is

$$R(t) = e^{-t\mathbf{P}}R(0).$$

Hence, if $\mathbf{P}$ is positive definite, $\lim_{x\to\infty} R(t) = 0$.

From the expression of $R(t)$ we infer

$$Q(t) = (Id - e^{-t\mathbf{P}})\mathbf{P}^{-1} + e^{-t\mathbf{P}}Q_0.$$

$Q(t)$ is thus SPD when Q_0 and $\mathbf{P}$ are SPD and Q_0 and $\mathbf{P}$ commute.

Let us now establish the convergence in the Frobenius norm. If $\mathbf{P}$ is positive definite, there exists a strictly positive real number α such that

$$\alpha \sum_{i=1}^{n} u_i^2 \leq \langle \mathbf{P}u, u \rangle = \sum_{i=1}^{n} \left(\sum_{k=1}^{n} \mathbf{P}_{i,k} u_k \right) u_k, \forall u \in \mathbb{R}^n,\ u = (u_1, \cdots, u_n)^t,$$

the number α possibly depending on n.

Taking the Hadamard product of each term of the differential equation and summing on all indices i, j, we get

$$\frac{1}{2} \frac{d\|R\|_F^2}{dt} + \sum_{i,j=1}^{n} \left(\sum_{k=1}^{n} \mathbf{P}_{i,k} R_{k,j} \right) R_{i,j} = 0.$$

Therefore,

$$\frac{1}{2} \frac{d\|R\|_F^2}{dt} + \alpha \|R\|_F^2 \leq 0.$$

By integration of each side of the last inequality, we obtain

$$\|R(t)\|_F \leq e^{-\alpha t} \|R(0)\|_F. \qquad \square$$

3.2 Construction of preconditioners by numerical integration

We introduce the discrete residual $R_k = Id - \mathbf{P}Q_k$. The numerical integration of equation (3.11) by forward Euler's method generates the sequence R_k which satisfies the recurrence relation:

$$R_{k+1} = (1 - \Delta t_k) R_k + \Delta t_k R_k^2.$$

We remark that for $\Delta t_k = 1$, the convergence is quadratic whenever $\|R_0\| < 1$, where $\|.\|$ is any matrix norm. We recover in this case the Newton method derived from the equation in one variable $\frac{1}{t} - r = 0$, see [10].

Let us study the general case.

Theorem 2. *We have the following results:*

(i) $\Delta_k = \Delta t$. *Assume that*

$$\rho(R_0) < 1 \quad \textit{and} \quad \Delta t < \frac{2}{1 - \rho(R_0)}.$$

Then $\lim_{k \to \infty} Q_k = \mathbf{P}^{-1}$.

(ii) *Assume that* $\|R_0\|_F < 1$ *and that*

$$0 < \Delta t_k < \frac{2}{1 + \|R_k\|_F} \forall k.$$

Then, $\lim_{k\to\infty} \|R_k\|_F = 0$. *Moreover the convergence is quadratic for* $\Delta t_k = 1$.

(iii) *Assume that* $\mathbf{P}$ *and* Q_0 *are symmetric, then* Q_k *is also symmetric for all* $k \geq 0$.

Proof. From the relation

$$R_{k+1} = R_k \left((1 - \Delta t) Id + \Delta t R_k\right),$$

we deduce that the convergence is guaranteed if $\rho((1 - \Delta t) Id + \Delta t R_k) < 1$, say if

$$\rho(R_0) < 1 \text{ and } 0 < \Delta t < \frac{2}{1 - \rho(R_0)}.$$

The first condition is verified, e.g., when $Q_0 = \gamma \mathbf{P}^T$ with $0 < \gamma < \frac{2}{\rho(\mathbf{P}\mathbf{P}^T)}$. Notice that if $\mathbf{P}$ is positive definite, we can take $Q_0 = \gamma Id$ with $0 < \gamma < \frac{2}{\rho(\mathbf{P})}$, Hence the assertion (i).

Let k be fixed. We have

$$R_{k+1} * R_{k+1} = (1 - \Delta t_k)^2 R_k * R_k + 2\Delta t_k (1 - \Delta t_k) R_k * R_k^2. + (\Delta t_k)^2 R_k^2 * R_k^2.$$

Taking the sum of all the indices, we obtain

$$\begin{aligned}
\|R_{k+1}\|_F^2 &= (1 - \Delta t_k)^2 \|R_k\|_F^2 \\
&\quad + 2\Delta t_k (1 - \Delta t_k) \sum_{i,j=1}^{n} \left(\sum_{m=1}^{n} (R_k)_{i,m} (R_k)_{m,j} \right) (R_k)_{i,j} \\
&\quad + (\Delta t_k)^2 \sum_{i,j=1}^{n} \left(\sum_{m=1}^{n} (R_k)_{i,m} (R_k)_{m,j} \right)^2, \\
&\quad \text{(applying Lemma 2)} \\
&\leq (1 - \Delta t_k)^2 \|R_k\|_F^2 + 2\Delta t_k |1 - \Delta t_k| \|R_k\|_F^3 + (\Delta t_k)^2 \|R_k\|_F^4, \\
&\leq \|R_k\|_F^2 \left(|1 - \Delta t_k| + \Delta t_k \|R_k\|_F\right)^2.
\end{aligned}$$

Therefore, if $M_k = |1 - \Delta t_k| + \Delta t_k \|R_k\|_F < 1$, say if $\|R_k\|_F < 1$ and $0 < \Delta t_k < \frac{2}{1+\|R_k\|_F}$, then $\|R_{k+1}\|_F < M_k \|R_k\|_F$ with $M_k < 1$. The contraction holds in particular when $0 < \Delta t_k \leq 1$ and it is easy to prove by induction that if $\|R_0\|_F < 1$ then $\|R_{k+1}\|_F < M\|R_k\|_F$, with $M < 1$. The convergence follows.

The particular case $\Delta t_k = 1$ gives directly the estimate

$$\|R_k\|_F \leq \|R_0\|_F^{2^k}.$$

The convergence in Frobenius norm is then quadratic in this case if $\|R_0\|_F < 1$. The point (ii) is proved.

Finally, if Q_0 and $\mathbf{P}$ are SPD, then using the relation $Q_{k+1} = (1 + \Delta t_k)Q_k - \Delta t_k Q_k \mathbf{P} Q_k$, we show easily by induction that Q_k is symmetric for all $k \geq 0$. This completes the proof. □

Let us now consider the implementation of the Euler scheme to (3.12). The following sequence of matrices is generated:

$$\begin{cases} Q_0 \text{ given} \\ \texttt{For k=0,...} \\ Q_{k+1} = Q_k + \Delta t_k (Id - \mathbf{P} Q_k) \end{cases} \tag{3.13}$$

We have the

Theorem 3. *Assume* $\mathbf{P}$ *is positive definite and, for simplicity, that* $\Delta_k t = \Delta t$. *Then*

(i) *If* $0 < \Delta t < \frac{2}{\rho(\mathbf{P})}, \forall k \geq 0$. *Then,* Q_k, *the sequence generated by the scheme* (3.13) *converges to* $\mathbf{P}^{-1}$.

(ii) *Assume in addition that* $\mathbf{P}$ *is symmetric and* Q_0 *is SPD. Assume that* Q_0 *and* $\mathbf{P}$ *commute. Then* Q_k *is symmetric for all* $k \geq 0$. *Moreover if* $\alpha_0 - \Delta t \frac{M\|Id - \Delta t P Q_0\|_2}{1-M} > 0$ *then* Q_k *is SPD for all* $k \geq 0$, *where we have set* $M = \|Id - \Delta t P\|_2$,

$$\alpha_0 = \min_{x \in \mathbb{R}^n, \|x\|_2 = 1} \langle Q_0 x, x \rangle.$$

Proof. Assume first that $\Delta t_k = \Delta t$. Using the same notations, we have,

$$R_{k+1} = (Id - \Delta t \mathbf{P}) R_k.$$

Thus, $R_k \to 0$ if and only if $0 < \Delta t < \dfrac{2}{\rho(\mathbf{P})}$.

Let us now study the convergence in the Frobenius norm. Since $\mathbf{P}$ is positive definite, we can define

$$0 < \alpha = \min_{u \in \mathbb{R}^n, \|u\|_2 = 1} \frac{\langle Pu, u \rangle}{\langle u, u \rangle}.$$

We have

$$R_{k+1} * R_{k+1} = (Id - \Delta t \mathbf{P}) R_k * (Id - \Delta t \mathbf{P}) R_k.$$

Hence, taking the sum of all indices, we obtain after simplifications

$$\|R_{k+1}\|_F^2 + 2\Delta t \sum_{i,j} \sum_{m=1}^{n} \mathbf{P}_{i,m} R_{m,j} R_{i,j} = \|R_k\|_F^2 + (\Delta t)^2 \sum_{i,j} \left(\sum_{m=1}^{n} \mathbf{P}_{i,m} R_{m,j} \right)^2.$$

Therefore

$$\|R_{k+1}\|_F^2 + 2\alpha \Delta t \|R_k\|_F^2 \leq \|R_k\|_F^2 + (\Delta t)^2 \|P\|_F^2 \|R_k\|_F^2.$$

Finally

$$\|R_{k+1}\|_F^2 \leq \left(1 - 2\alpha \Delta t + (\Delta t)^2 \|P\|_F^2\right) \|R_k\|_F^2,$$

which gives the (sufficient) stability condition

$$0 < \Delta t < \frac{2\alpha}{\|P\|_F^2}.$$

Now, one can show by induction that if Q_0 and $\mathbf{P}$ commute, then Q_k and $\mathbf{P}$ commute also for all $k \geq 0$. Then, proceeding also by induction, it can be shown that Q_k is symmetric for all $k \geq 0$. Notice that the condition $\mathbf{P} Q_0 = Q_0 \mathbf{P}$ is simply verified, e.g., with the choice $Q_0 = Id$.

Now we set

$$\alpha_k = \min_{x \in \mathbb{R}^n, \|x\|_2 = 1} \frac{\langle Q_k x, x \rangle}{\langle x, x \rangle}, \forall k \geq 0.$$

Let $x \in \mathbb{R}^n$, $\|x\|_2 = 1$. We have

$$\begin{aligned} \langle Q_{k+1} x, x \rangle &= \langle Q_k x, x \rangle - \Delta t \langle R_k x, x \rangle, \\ &\geq \alpha_k - \Delta t \|R_k\|_2. \end{aligned}$$

But $\|R_k\| \leq \|Id - \Delta t\mathbf{P}\|_2^k \|R_0\|_2 = M^k \|R_0\|_2$. Thus

$$\alpha_{k+1} \geq \alpha_k - \Delta t M^k \|R_0\|_2,$$

and therefore

$$\alpha_k \geq \alpha_0 - \Delta t \frac{M(1-M^k)\|R_0\|_2}{1-M} \geq \alpha_0 - \Delta t \frac{M\|R_0\|_2}{1-M}.$$

This completes the proof. □

By analogy between Euler's method and Richardson's iterations, it is natural to compute Δt_k such as minimizing $\|R_{k+1}\|_F$. We have

$$R_{k+1} * R_{k+1} = (Id - \Delta t_k \mathbf{P})R_k * (Id - \Delta t_k \mathbf{P})R_k.$$

Taking the sum on all indices, we obtain, after the usual simplifications

$$\|R_{k+1}\|_F^2 = \|R_k\|_F^2 + (\Delta t_k)^2 \sum_{i,j=1}^{n} ((\mathbf{P}R_k) * (\mathbf{P}R_k))_{i,j} - 2\Delta t_k \sum_{i,j=1}^{n} ((\mathbf{P}R_k) * R_k)_{i,j}.$$

It follows that $\|R_{k+1}\|_F$ is minimized for

$$\Delta x_k = \frac{\sum_{i,j=1}^{n} ((\mathbf{P}R_k) * R_k)_{i,j}}{\|\mathbf{P}R_k\|_F^2} = \frac{\ll \mathbf{P}R_k, R_k \gg}{\ll \mathbf{P}R_k, \mathbf{P}R_k \gg},$$

and we recover the iterations proposed in [10], see also Section 4.

3.3 Steepest descent-like Schemes

The computation of a steady state by an explicit scheme can be speeded up by enhancing the stability domain of the scheme since it allows to use larger time steps; in this situation the accuracy of a time marching scheme is not fundamental. We can derive more stable methods by using parametrized one step schemes and to fit the parameters, not for increasing the accuracy such as in the classical schemes (Heun's, Runge Kutta's), but for improving the stability.

For example, in [8] it was defined a method for computing iteratively fixed points with larger descent parameter starting from a specific numerical time scheme. It consists in integrating the differential equation

$$\begin{cases} \dfrac{dU}{dt} = F(U), \\ U(0) = U_0, \end{cases} \tag{3.14}$$

by the two steps scheme

$$\begin{cases} K_1 = F(U^k), \\ K_2 = F(U^k + \Delta t K_1), \\ U^{k+1} = U^k + \Delta t\,(\alpha K_1 + (1-\alpha) K_2)\,. \end{cases} \tag{3.15}$$

Here α is a parameter to be fixed. This scheme allows a larger stability as compared to the Forward Euler scheme. More precisely, when $F(U) = b - \mathbf{P}U$.

Lemma 4. *Assume that* $\mathbf{P}$ *is positive definite, then the scheme is convergent iff*

$$\alpha < \frac{7}{8} \quad \text{and} \quad \Delta t < \frac{1}{(1-\alpha)\rho(\mathbf{P})}.$$

Of course, one can define iteratively α and Δt such as minimizing the euclidean norm of the residual, exactly as in the steepest descent method. The residual equation is

$$r^{k+1} = \left(I - \Delta t_k\, \mathbf{P} + (1-\alpha_k)(\Delta t_k)^2\, \mathbf{P}^2\right) r^k. \tag{3.16}$$

Hence

$$\begin{aligned} \|r^{k+1}\|^2 \;=\; & \|r^k\|^2 - 2\Delta t_k \langle \mathbf{P}r^k, r^k\rangle + (\Delta t_k)^2 \|\mathbf{P}r^k\|^2 \\ & + 2(1-\alpha_k)(\Delta t_k)^2 \langle \mathbf{P}^2 r^k, r^k\rangle - 2(1-\alpha_k)(\Delta t_k)^3 \langle \mathbf{P}^2 r^k, \mathbf{P}r^k\rangle \\ & + (1-\alpha_k)^2 (\Delta t_k)^4 \langle \mathbf{P}^2 r^k, \mathbf{P}^2 r^k\rangle. \end{aligned}$$

We set for convenience

$$\begin{aligned} & a = \|r^k\|^2, \qquad b = \langle \mathbf{P}r^k, r^k\rangle, \qquad c = \|\mathbf{P}r^k\|^2, \\ & d = \langle \mathbf{P}^2 r^k, r^k\rangle, \quad e = \langle \mathbf{P}^2 r^k, \mathbf{P}r^k\rangle, \quad f = \langle \mathbf{P}^2 r^k, \mathbf{P}^2 r^k\rangle. \end{aligned}$$

$\|r^{k+1}\|$ is minimized for the following definition of the parameters:

$$\Delta t_k = \frac{fb - ed}{fc - e^2}, \;\; \alpha_k = (fc - e^2)\, \frac{eb - cd}{(fb - ed)^2}.$$

This gives rise to the steepest descent method derived from (3.15).

4 Sparse inverse preconditioners

The iterative processes generated by numerical integration of the differential equations require at least a product of two matrices at each iteration. Hence, at each iteration, the inverse preconditioner matrix becomes denser, even if the initial data and the matrix to invert are sparse.

We propose here a simple way to derive a dropping strategy from the numerical integration of a matrix differential equation. The notations are the same as in the previous sections. Consider the equation

$$\begin{cases} \frac{dQ}{dt} = Id - \mathbf{P}Q, \\ Q(0) = Q_0. \end{cases} \tag{4.17}$$

Here $\mathbf{P}$ is a positive definite matrix so $\lim_{x\to\infty} Q(x) = \mathbf{P}^{-1}$, a shown in section 2.

4.1 Derivation of the equations

Now, let $\mathcal{F}$ be a $n \times n$ matrix with coefficients 0 or 1. The Hadamard product $\mathcal{F} * \mathbf{P}$ returns a matrix whose coefficients are those of $\mathbf{P}$ which have the same indices as the non null coefficients of $\mathcal{F}$, so $\mathcal{F}$ is a filter matrix which selects a sparsity pattern. More precisely, we have

$$(\mathcal{F} * \mathbf{P})_{i,j} = \begin{cases} \mathbf{P}_{i,j} & \text{if } \mathcal{F}_{i,j} = 1, \\ 0 & \text{else.} \end{cases}$$

We assume that $\mathcal{F}_{i,i} = 1, i = 1, \cdots n$, so $\mathcal{F} * Id = Id$, where Id is the $n \times n$ identity matrix.

At this point, we consider the Hadamard product of each term of (4.17) with $\mathcal{F}$. We obtain the system

$$\begin{cases} \frac{d\mathcal{F} * Q}{dt} = Id - \mathcal{F} * (\mathbf{P}Q), \\ \mathcal{F} * Q(0) = \mathcal{F} * Q_0. \end{cases} \tag{4.18}$$

For deriving an autonomous equation with a sparse matrix S as unknown, we approach $\mathcal{F} * (\mathbf{P}Q)$ by $\mathcal{F} * (\mathbf{P}\mathcal{F}Q)$ and we obtain the new system

$$\begin{cases} \frac{dS}{dt} = Id - \mathcal{F} * (\mathbf{P}S), \\ \mathcal{F} * S(0) = \mathcal{F} * Q_0. \end{cases} \tag{4.19}$$

The matrix $S(t)$ is sparse for all t. Indeed, we have the

Lemma 5. *The matrix equation* (4.19) *has a unique solution* $S(t) \in C^1(]0, +\infty[$ *and*

$$\mathcal{F} * S(t) = S(t), \forall t \geq 0.$$

Proof. The existence and the uniqueness of $S(t)$ is established by using standard arguments.

We have

$$S(t) = S(0) + \int_0^t (Id - \mathcal{F} * (\mathbf{P}S))ds$$

Hence, since $\mathcal{F} * S(0) = S(0)$, we can write

$$\begin{aligned} \mathcal{F} * S(t) &= S(0) + \int_0^t \mathcal{F} * (Id - \mathcal{F} * (\mathbf{P}S)ds \\ &= S(0) + \int_0^t (Id - \mathcal{F} * (\mathbf{P}S)ds. \end{aligned}$$

□

We now will show that $S(t)$ is an approximation of $Q(t)$.

4.2 A priori estimates

We have the following result:

Theorem 4.

$$\begin{aligned} \|S - Q\|_F^2 \leq\ & 2 \times \Big\{ \|(\mathbf{1} - \mathcal{F}) * \mathbf{P}^{-1}\|_F^2 \|e^{-t\mathbf{P}} - Id\|_F^2 \|Id - \mathbf{P}Q_0\|_F^2 \\ & + \frac{1}{2\alpha} \|(\mathbf{1} - \mathcal{F}) * \mathbf{P}^{-1}\|_F^2 \|Id - \mathbf{P}Q_0\|_F^2 \int_0^t e^{-\alpha(t-s)} \|e^{-s\mathbf{P}} - Id\|_F^2 ds \Big\}, \end{aligned}$$

where $\mathbf{1}$ *is the neutral element of the Hadamard product,* ($\mathbf{1}_{i,j} = 1, i, j = 1, \cdots n$).

Proof. Taking the difference of the equations (4.19) and (4.18), we get

$$\frac{dS - \mathcal{F} * Q}{dt} = -\mathcal{F} * (\mathbf{P}(S - \mathcal{F} * Q)) + \mathcal{F} * (\mathbf{P}(\mathcal{F} * Q - Q)). \quad (4.20)$$

The difference between (4.18) and (4.17) gives

$$\frac{d\mathcal{F}*Q-Q}{dt}=(\mathbf{1}-\mathcal{F})*\mathbf{P}Q. \tag{4.21}$$

From (4.20), we infer

$$\begin{aligned}\frac{1}{2}\,\frac{d\|S-\mathcal{F}*Q\|_F^2}{dt} &= -\ll \mathcal{F}*(\mathbf{P}(S-\mathcal{F}*Q)),\, S-\mathcal{F}*Q\gg \\ &\quad +\ \ll \mathcal{F}*(\mathbf{P}(\mathcal{F}*Q-Q)),\, S-\mathcal{F}*Q\gg .\end{aligned} \tag{4.22}$$

Now, since

$$\ll \mathcal{F}*(\mathbf{P}(S-\mathcal{F}*Q)),\, S-\mathcal{F}*Q\gg = \ll \mathbf{P}(S-\mathcal{F}*Q)),\, S-\mathcal{F}*Q\gg ,$$

we can write

$$\begin{aligned}\frac{1}{2}\,\frac{d\|S-\mathcal{F}*Q\|_F^2}{dt}+\ll \mathbf{P}(S-\mathcal{F}*Q)),\, S-\mathcal{F}*Q\gg = \\ +\ll \mathbf{P}(\mathcal{F}*Q-Q),\, S-\mathcal{F}*Q\gg .\end{aligned} \tag{4.23}$$

We let $\alpha=\min\limits_{Q\in\mathcal{M}_n(\mathbb{R})}\dfrac{\ll \mathbf{P}Q,\,Q\gg}{\ll Q,\,Q\gg}$ and we deduce from the previous equation

$$\begin{aligned}&\frac{1}{2}\,\frac{d\|S-\mathcal{F}*Q\|_F^2}{dt}+\alpha\|S-\mathcal{F}*Q\|_F^2 \\ &\qquad\le \ll \mathbf{P}(\mathcal{F}*Q-Q),\,S-\mathcal{F}*Q\gg, \\ &\qquad\quad \text{(applying Young's inequality)}, \\ &\qquad\le\ \frac{\eta}{2}\|S-\mathcal{F}*Q\|_F^2+\frac{1}{2\eta}\|\mathcal{F}*Q-Q\|_F^2.\end{aligned} \tag{4.24}$$

Here η is a strictly positive real number which will be chosen later on. We now must derive estimates for $\|\mathcal{F}*Q-Q\|_F$. From the direct integration of (4.17), we get

$$\mathbf{P}Q=Id-e^{-t\mathbf{P}}\,(Id-\mathbf{P}Q_0)\,.$$

Therefore,

$$\frac{d\mathcal{F}*Q-Q}{dt}=(\mathbf{1}-\mathcal{F})*\left(Id-e^{-t\mathbf{P}}\,(Id-\mathbf{P}Q_0)\right),$$

so

$$
\begin{aligned}
(\mathcal{F} * Q - Q)(t) &= (\mathcal{F} * Q - Q)(0) \\
&\quad + \int_0^t (\mathbf{1} - \mathcal{F}) * \left(Id - e^{-t\mathbf{P}} (Id - \mathbf{P}Q_0) \right) ds, \qquad (4.25) \\
&= \int_0^t (\mathbf{1} - \mathcal{F}) * e^{-s\mathbf{P}} (Id - \mathbf{P}Q_0) ds, \qquad (4.26) \\
&= (\mathbf{1} - \mathcal{F}) * \int_0^t e^{-s\mathbf{P}} (Id - \mathbf{P}Q_0)\, ds, \qquad (4.27) \\
&= (\mathbf{1} - \mathcal{F}) * \mathbf{P}^{-1} \left(e^{-t\mathbf{P}} - Id \right) (Id - \mathbf{P}Q_0). \qquad (4.28)
\end{aligned}
$$

We then can write

$$
\|(\mathcal{F} * Q - Q)(t)\|_F \leq \|(\mathbf{1} - \mathcal{F}) * \mathbf{P}^{-1}\|_F \|e^{-t\mathbf{P}} - Id\|_F \|Id - \mathbf{P}Q_0\|_F .
$$

Substituting this last inequality in (4.24), we get

$$
\begin{aligned}
\frac{1}{2} \frac{d\|S - \mathcal{F} * Q\|_F^2}{dt} &+ \alpha \|S - \mathcal{F} * Q\|_F^2 \leq \frac{\eta}{2} \|S - \mathcal{F} * Q\|_F^2 \\
&+ \frac{1}{2\eta} \|(\mathbf{1} - \mathcal{F}) * \mathbf{P}^{-1}\|_F^2 \|e^{-t\mathbf{P}} - Id\|_F^2 \|Id - \mathbf{P}Q_0\|_F^2
\end{aligned}
$$

Now we choose $\eta = \alpha$ and we integrate this inequality:

$$
\begin{aligned}
\|(S - \mathcal{F} * Q)(t)\|_F^2 &\leq \|(S - \mathcal{F} * Q)(0)\|_F^2 e^{-\alpha t} \\
+ \frac{1}{2\alpha} \|(\mathbf{1} - \mathcal{F}) * \mathbf{P}^{-1}\|_F^2 \|Id - \mathbf{P}Q_0\|_F^2 &\int_0^t e^{-\alpha(t-s)} \|e^{-s\mathbf{P}} - Id\|_F^2 ds .
\end{aligned}
$$

Finally, summing this last estimate with $\|(\mathcal{F} * Q - Q)(t)\|_F^2$ we obtain

$$
\begin{aligned}
\|S - Q\|_F^2 &\leq 2(\|S - F * Q\|_F^2 + \|F * Q - Q\|_F^2) \\
&\leq 2 \times \Big\{ \|(\mathbf{1} - \mathcal{F}) * \mathbf{P}^{-1}\|_F^2 \|e^{-t\mathbf{P}} - Id\|_F^2 \|Id - \mathbf{P}Q_0\|_F^2 \\
&\quad + \frac{1}{2\alpha} \|(\mathbf{1} - \mathcal{F}) * \mathbf{P}^{-1}\|_F^2 \|Id - \mathbf{P}Q_0\|_F^2 \int_0^t e^{-\alpha(t-s)} \|e^{-s\mathbf{P}} - Id\|_F^2 ds \Big\}
\end{aligned}
$$

□

We conclude this section with an important remark. The error bounds that we derive do not insure that the sparse preconditioner $S(t)$ is invertible for all t and at least for $t \geq t_0$. However, in practice, the numerical implementations of time marching schemes for computing $S(t)$, t large, produce invertible matrices.

5 Numerical illustrations

5.1 Inverse matrix approximation

Consider the problem

$$-\Delta u + a(x, y)\frac{\partial u}{\partial x} + b(x, y)\frac{\partial u}{\partial y} = f \quad \text{in } \Omega =]0, 1[^2 \tag{5.29}$$

$$u = 0 \quad \text{on } \partial\Omega \tag{5.30}$$

We discretize this problem by second order finite differences on a $n \times n$ grid and we define $\mathbf{P}$ as the underlying matrix. The numerical results we present were obtained by using Matlab 6 on a cluster of Bi-processor 800 (Pentium III) at Université Paris XI, Orsay, France.

5.1.1 Integration of finite time inverse matrix differential equation

We first consider the problem with

$$a(x, y) = 30e^{y^2-x^2},\ b(x, y) = 50\sin(72x(1-x)y) * \sin(3\pi y),\ n = 30$$

(the matrix is of size 900×900) and a Chebyshev Mesh in both directions. In Figure 2 we have compared the preconditioners obtained with 2 iterations of Euler (Euler(2)), of Adams-Bashforth (AB(2)), of Fourth order Runge Kutta (RK4(2)). We observe that the more accurate is the integration method, the more concentrated is the spectrum of the preconditioned matrix.

5.1.2 Sparse inverse preconditioner case

We consider here the sparse approximation of the inverse of the finite differences discretization matrix of the operator

$$-\Delta + 500\partial_x + 20\partial_y$$

on the domain $]0, 1[^2$ with homogeneous Dirichlet boundary conditions, on a regular grid. Here the sparsity pattern is defined by the $n^2 \times n^2$ symmetric mask-matrix $\mathcal{F}$ as follows

$$\mathcal{F}_{i,j} = 1 \text{ if } |i-j| \leq 2 \text{ or if } |i-j \pm n| \leq 1 \mathcal{F}_{i,j} = 0 \text{ in the other cases.}$$

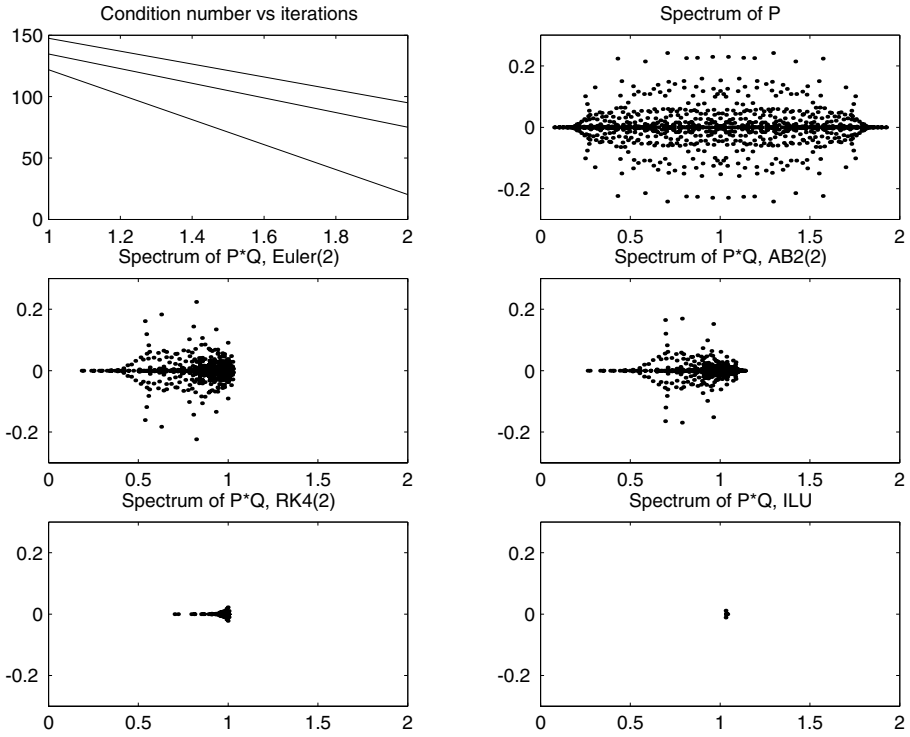

Figure 2 – Spectrum of $\mathbf{P}$, $Euler(2)\mathbf{P}$, $AB2(2)\mathbf{P}$, $RK4(2)\mathbf{P}$ and $ILU\mathbf{P}$.

In Figure 3 we have represented approximations of the inverse matrix that are obtained by a tresholding of the coefficient at the level ϵ, for different values of ϵ. This shows that a sparse approximation can be considered in this case.

As we can see in Figure 4, very few iterations are needed to obtain the convergence. Of course the residual do not converge to 0 because the approximation of the inverse is sparse. This is agree with error estimates of the continuous equations: a saturation is expected. In Figure 5, we have plotted the spectrum of the preconditioned matrix. We observe that the inverse preconditioner provides a concentration of the spectrum.

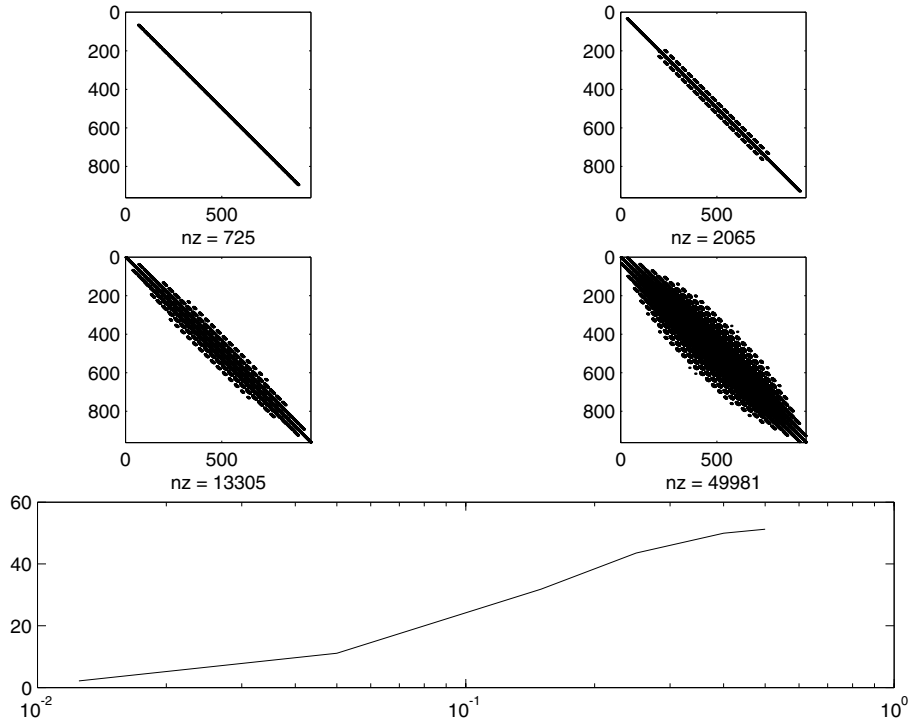

Figure 3 – Coefficients of $\mathbf{P}^{-1}$ greater (in modulus) than $\epsilon = 0.5$ (fig. (a)), $\epsilon = 0.4$ (fig. (b)), $\epsilon = 0.25$ (fig. (c)), $\epsilon = 0.15$ (fig. (d)), norm of the error for the filtered inverse matrix vs ϵ, (e).

5.2 Preconditioned descent methods

The reduction of the condition number as well as the concentration of the spectrum of the preconditioned matrix allows faster convergence of descent methods.

As an illustration, we apply the explicit preconditioner computed above to the numerical solution of the convection diffusion problem. To this end, we use the preconditioned BiCgstab method [21].

The discretization matrix is the same as above. The discrete problem to solve reads

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

We prepare the system by diagonal preconditioning, and we consider the equiv-

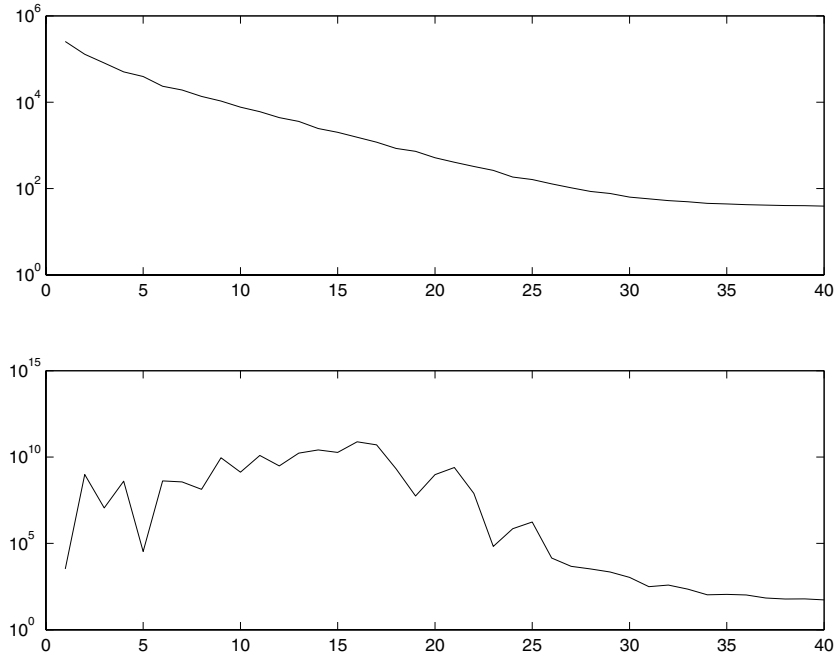

Figure 4 – Computation of sparse inverse preconditioner: Residual vs iterations (above) and condition number vs iterations (below).

alent problem

$$\text{diag}(\mathbf{P})^{-1}\mathbf{P}x = \text{diag}(\mathbf{P})^{-1}b.$$

The exact solution x_e is a random vector and $b = \mathbf{P}x_e$.

In Figure 6 we have represented the residual (respectively the error) versus the iteration when using Bicgstab and various preconditioned versions; the explicit preconditioners Q were here generated by, in the one hand, with two iterations of Euler, of AB2 and of RK4, and, in the other hand, with an ILU factorization with $\epsilon = 10^{-2}$ as tolerance. The Euler and the AB2 preconditioners improve the convergence of the unpreconditioned method, with respective rates 2 and 3. The RK4 preconditioner is comparable to the ILU one.

In Figure 6, we have illustrated the improvement of the convergence carried by the sparse inverse preconditioner computed on the previous subsection.

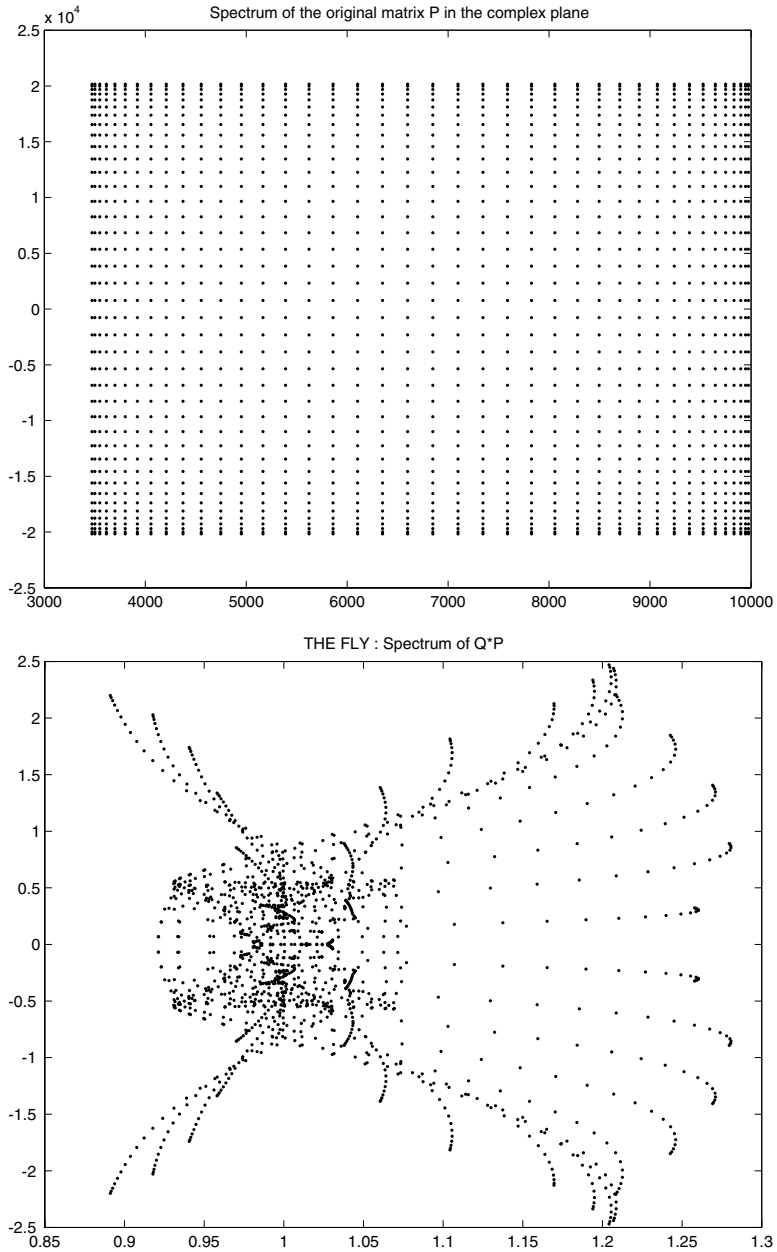

Figure 5 – Spectrum of the original matrix (top), (the fly) Spectrum the preconditioned matrix (bottom), in the complex plane.

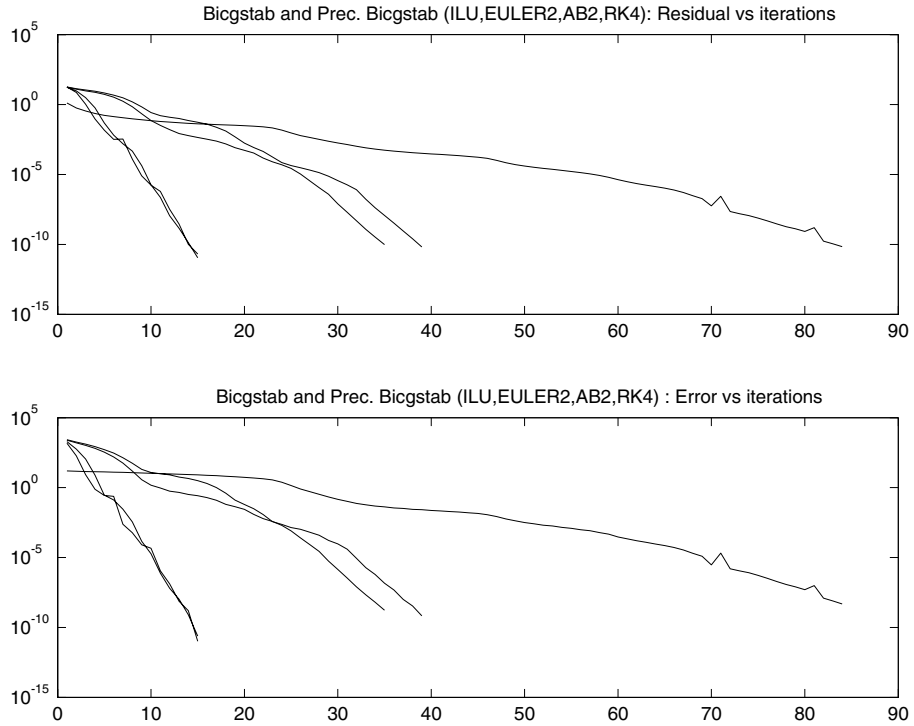

Figure 6 – Comparison of the preconditioners : Euler(2), AB(2), RK4(2) and ILU, Size of the system : 961×961, a) Residual vs iterations, b) error vs iterations.

6 Concluding remarks

The approach we have developed here is simple, rather general and seems to apply to a large class of matrices. The advantage of this technique is to study the underlying approximations with simple analysis tools; we recover in addition particular sequences of inverse preconditioners ([4, 10, 9]) and introduce new ones. The iterative schemes we introduced in this article are all based on approximation of the inverse by a proper polynomial: they can be considered as polynomial preconditioners in spite they are not automatically related to the ones proposed (e.g.) by [1], the point of view being here different. This suggests as a feature to analyze them by using an approach coming from the approximation theory.

The masked matrix differential equation approach allows to build simply effi-

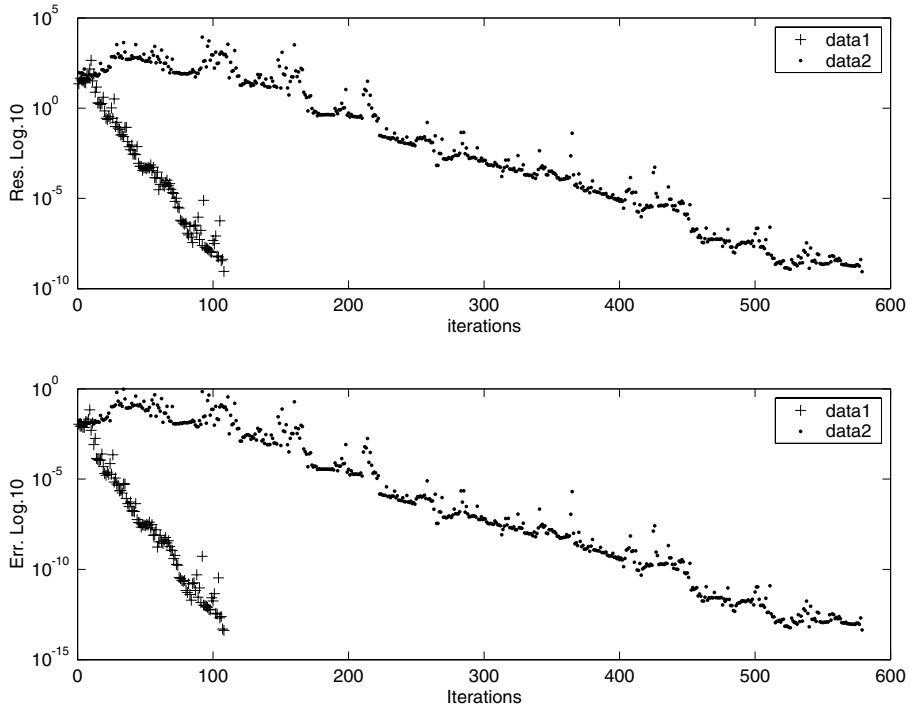

Figure 7 – Sparse inverse preconditionning for convection-diffusion problem. a) Residual vs iterations; b) Error vs iterations.

cients sparse inverse preconditioners for a fixed sparsity pattern. A natural next feature would be to develop dropping strategies for improving the method.

We have applied here a dynamical modeling approach to the construction of inverse preconditioners. A similar approach can be developed for the solution of linear as well as non linear systems of equations, deriving numerical schemes from special dynamical systems.

The examples we give are coming out from PDE's discretization and are rather academic, but is it a first step to be considered before developing and applying the schemes to large scales problems.

Acknowledgments. The author thanks Y. Chitour and J. Laminie, from Université Paris-Sud, for fruitful remarks.

REFERENCES

[1] Ashby, Manteuffel and Otto, A comparison of adaptive Chebyshev and least squares polynomial preconditioning for Hermitian positive definite linear systems, SIAM J. Sci. Stat. Comput., **13**(1) (1992), 1–29.

[2] O. Axelsson, *Iterative solution methods*. Cambridge University Press, Cambridge, 1994. xiv+654 pp.

[3] R.E. Bellman, *Introduction to Matrix Analysis*, Mcgraw-Hill (New York), 1970 – 2nd ed.

[4] M. Benzi, Preconditioning techniques for large linear systems: a survey. J. Comput. Phys., **182**(2) (2002), 418–477.

[5] C. Brezinski, *Projection Methods for Systems of Equations*, North-Holland, 1997.

[6] C. Brezinski, Dynamical systems and sequence transformation, J. Phys. A: Math. Gen., **34** (2001), 10659–10669.

[7] C. Brezinski, Difference and differential equations and convergence acceleration algorithms. SIDE III-symmetries and integrability of difference equations (Sabaudia, 1998), 53–63, CRM Proc. Lecture Notes, 25, Amer. Math. Soc., Providence, RI, 2000.

[8] C. Brezinski and J.-P. Chehab, Nonlinear hybrid procedures and fixed point iterations, Numer. Func. Anal. Opt., **19**(5-6) (1998), 415–487.

[9] G. Castro, J. Laminie, M. Sarrazin and A. Seghier, Factorized Sparse Inverse Preconditioner using Generalized Reflection Coefficients, Prépublication d'Orsay numéro 28 (10/7/1997).

[10] E. Chow and Y. Saad, Approximate Inverse Preconditioning for Sparse-Sparse Iterations, SIAM Journal of Scientific Computing, **19** (1998), 995–1023.

[11] Eisenstat, Ortega and Vaughan, Efficient polynomial preconditioning for the conjugate gardient method, SIAM J. Sci Stat. Comp., **11**(5) (1990), 859–872.

[12] A. Cuyt and L. Wuytack, *Nonlinear Methods in Numerical Analysis*, North-Holland, Amsterdam, 1987.

[13] F. Dubois and A. Saidi, Unconditionally stable scheme for Riccati equation, ESAIM Proc, **8** (2000), 39–52.

[14] U. Helmke and J.B. Moore, *Optimization and Dynamical Systems*, Comm. Control Eng. Series, Springer, London, 1994.

[15] M.W. Hirsch and S. Smale, *Differential Equations, Dynamical Systems and Linear Algebra*, Academic Press, London, 1974.

[16] J.H. Hubbard and B.H. West, *Differential equations. A dynamical Systems Approach. Part I: Ordinary Differential Equations*, Springer Verlag, New-York, 1991.

[17] J.B. Moore, R.E. Mahony and U. Helmke, Numerical Gradient Algorithms for eigenvalue and singular value calculations, SIMAX, **15**(3) (1994), 881–902.

[18] Y. Saad, Practical implementation of polynomial preconditioning for Conjugate Gradient, SIAM J. Sci. Stat. Comp., **6** (1985), 865–881.

[19] Y. Saad, *Iterative Methods for Sparse Linear Systems*, SIAM, 1996.

[20] A.M. Stuart, Numerical analysis of dynamical systems, in *Acta Numerica, 1994*, Cambridge University Press, Cambridge, 1994, pp. 467–572.

[21] H.A. Van der Vorst, Bi-CGSTAB: a fast and smoothly converging variant of Bi-CG for the solution of nonsymmetric linear systems, SIAM J. Sci. Stat. Comput., **13** (1992), 631–644.