NUMERICAL LINEAR ALGEBRA WITH APPLICATIONS
Numer. Linear Algebra Appl. 2001; **8**:1–16 *Prepared using* nlaauth.cls *[Version: 2000/03/22 v1.0]*

On the solution of large Sylvester-observer equations

D. Calvetti[1], B. Lewis[2], and L. Reichel[2]

[1]*Department of Mathematics, Case Western Reserve University, Cleveland, OH 44106.*
[2]*Department of Mathematics and Computer Science, Kent State University, Kent, OH 44242.*

SUMMARY

The design of a Luenberger observer for large control systems is an important problem in Control Theory. Recently, several computational methods have been proposed by Datta and collaborators. The present paper discusses numerical aspects of one of these methods, described by Datta and Saad (1991). Copyright © 2001 John Wiley & Sons, Ltd.

KEY WORDS: Arnoldi process, Sylvester equation, Luenberger observer, partial fraction representation, control system.

1. INTRODUCTION

Consider the control system

$$\begin{array}{rcl} \dot{\hat{x}}(t) & = & M\hat{x}(t) + B\hat{u}(t), \qquad \hat{x}(0) = \hat{x}_0, \qquad t \geq 0, \\ \hat{y}(t) & = & C\hat{x}(t), \end{array} \tag{1}$$

where $M \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times k}$ and $C \in \mathbb{R}^{m\times n}$, and the functions $\hat{x}(t) \in \mathbb{R}^n$ and $\hat{u}(t) \in \mathbb{R}^k$ are defined for $t \geq 0$. Throughout this paper we will assume that n is large and $m \ll n$.

In many situations of practical interest neither the initial state $\hat{x}_0$ nor the states $\hat{x}(t)$ for $t > 0$ are explicitly known. A popular approach to gain information about $\hat{x}(t)$ is to design a control systems related to (1), whose states $x(t)$ approximate $\hat{x}(t)$; see, e.g., Datta [9, Chapter 12] or Kailath [15, Chapter 4] for discussions.

Luenberger [16] proposed the construction of an approximation $x(t)$ of $\hat{x}(t)$ as follows. Introduce the control system

$$\dot{x}(t) = H^T x(t) + G^T \hat{y}(t) + D\hat{u}(t), \qquad x(0) = x_0, \qquad t \geq 0, \tag{2}$$

email: D. Calvetti (dxc57@po.cwru.edu), B. Lewis (blewis@mcs.kent.edu), L. Reichel (reichel@mcs.kent.edu)

Contract/grant sponsor: [1]Research supported in part by NSF grant DMS-9806702. [2]Research supported in part by NSF grant DMS-9806413.

Copyright © 2001 John Wiley & Sons, Ltd.

Received 21 January 2000
Revised 5 February 2001
Accepted 6 April 2001

where $H, G \in \mathbb{R}^{m\times m}$, $D \in \mathbb{R}^{m\times k}$ are matrices to be determined. The system (2) is commonly referred to as a Luenberger observer. Assume that the spectra $\lambda(H)$ of H and $\lambda(M)$ of M satisfy

$$\lambda(H) \cap \lambda(M) = \emptyset. \tag{3}$$

The property (3) secures that the Sylvester equation

$$X^T M - H^T X^T = G^T C \tag{4}$$

has a unique solution $X^T \in \mathbb{R}^{m\times n}$; see, e.g., [13, Chapter 4] for a discussion. Let

$$A = M^T. \tag{5}$$

For future reference, we express equation (4) in the form

$$AX - XH = C^T G. \tag{6}$$

We will now show that the difference between $x(t)$ and $X^T\hat{x}(t)$, where $\hat{x}(t)$ solves (1), converges to zero as t increases, provided that the matrices H and D in (2) are chosen in a suitable manner. Differentiating the difference

$$e(t) = x(t) - X^T\hat{x}(t) \tag{7}$$

and using (1) and (2) yields

$$\begin{aligned} \dot{e}(t) &= H^T x(t) + G^T\hat{y}(t) + D\hat{u}(t) - X^T(M\hat{x}(t) + B\hat{u}(t)), \\ &= H^T e(t) + (H^T X^T + G^T C - X^T M)\hat{x}(t) + (D - X^T B)\hat{u}(t). \end{aligned} \tag{8}$$

Letting $D = X^T B$ and substituting the relation (4) into (8) shows that

$$e(t) = \exp(Ht)(x_0 - X^T\hat{x}_0), \qquad t \geq 0. \tag{9}$$

Assume that the matrix H is asymptotically stable, i.e., that every eigenvalue of H has negative real part. Then it follows from (9) that the difference (7) converges to zero as t increases.

The above construction of an approximation $x(t)$ of $\hat{x}(t)$ showed necessary conditions on the matrix H. However, no requirements were imposed on the matrices X and G. In order to reduce the sensitivity of $x(t)$ to perturbations, the following additional conditions are often imposed:

1. All eigenvalues of H have smaller real part than any eigenvalue of A,
2. X is well conditioned, and
3. G is such that the matrix pair $\{H, G\}$ is controllable, i.e.,

$$\operatorname{rank}([G, H - zI_m]) = m, \qquad \forall z \in \mathbb{C}, \tag{10}$$

where I_m denotes the $m \times m$ identity matrix. Moreover, it is desirable that the distance to uncontrollability

$$d_{uc}(H, G) = \min_{z\in\mathbb{C}} \sigma_m([G, H - zI_m])$$

is not "tiny," where $\sigma_m([G, H - zI_m])$ denotes the mth singular value of the matrix $[G, H - zI_m]$ and the singular values are enumerated in decreasing order.

Copyright © 2001 John Wiley & Sons, Ltd.

Prepared using nlaauth.cls

We focus on control systems (1) with large, possibly sparse, matrices A, and therefore ignore solution methods that require the transformation of the matrix A to a condensed form or demand knowledge of all eigenvalues of A. For references to methods that are well suited for small to medium sized systems (6); see Datta and Saad [11] as well as the survey papers by Datta [7, 8].

One approach to solve (6) is to first choose the matrices G and H and then solve (6) for X by a solution method for Sylvester equations, such as the iterative methods discussed in [4, 14]. However, these methods do not guarantee that the solution matrix X is well conditioned.

Datta and collaborators have developed several elegant methods for the solution of large-scale Sylvester-observer equations (6); see [1, 8, 10, 11, 12]. These methods are based on the Arnoldi process and use the matrix A, defined by (5), only to evaluate matrix-vector products. Sparsity or other structure of A, such as Toeplitz structure, may enable rapid evaluation of the matrix-vector products. Moreover, these methods yield well conditioned matrices X. An insightful and thorough discussion on Sylvester-observer equations is provided in the forthcoming book by Datta [9].

The present paper discusses numerical aspects of the Datta-Saad method [11]. This method allows an arbitrary choice of distinct eigenvalues of the matrix H. We investigate how the location of the eigenvalues of H affects the performance of the algorithm and propose a strategy for choosing these eigenvalues. In order to keep our presentation simple, we only discuss the method of Datta and Saad [11], however, our analysis and strategy for choosing eigenvalues of H also applies, mutatis mutandis, to the related methods developed by Datta and collaborators. A preliminary discussion of the topic of the present paper can be found in [5].

2. THE DATTA-SAAD METHOD

The solution method for the Sylvester-observer equation (6) by Datta and Saad [11] is based on the Arnoldi process. The matrix A is used only in matrix-vector product evaluations and this makes the method well-suited to the solution of large-scale Sylvester-observer equations. The method is designed for the special case when the right-hand side matrix $C^T G$ in (6) is of rank one. Modifications that allow a right-hand side matrix of higher rank are presented by Bischof et al. [1] and Datta and Hetti [10].

Let the matrices A and C be given by (1). Following Datta and Saad [11], we choose G to be the identity matrix. Then, clearly, equation (10) is satisfied and $d_{uc}(H, G) \geq 1$. We may assume that the rank-one matrix C^T is of the form ce_m^T, where $c \in \mathbb{R}^n$ and e_m denotes the mth axis vector. This particular form of C can be obtained by an initial orthogonal transformation. Thus, we may write equation (6) as

$$AX - XH = ce_m^T. \tag{11}$$

We will see below that $H \in \mathbb{R}^{m\times m}$ in (11) can be chosen to be an upper Hessenberg matrix, and $X \in \mathbb{R}^{n\times m}$ can be chosen to have orthogonal columns, all of the same Euclidean norm. With these choices of H and X, equation (11) is closely related to the Arnoldi decomposition of A obtained by application of m steps of the Arnoldi process. Specifically, m steps of the Arnoldi process applied to the matrix A with initial vector v_1 of Euclidean norm one yields

Copyright © 2001 John Wiley & Sons, Ltd.

Prepared using nlaauth.cls

the Arnoldi decomposition

$$AV_m - V_m H_m = \eta_m v_{m+1} e_m^T, \tag{12}$$

where $V_m = [v_1, v_2, \ldots, v_m] \in \mathbb{R}^{n\times m}$ and $V_m^T V_m = I_m$, $H_m \in \mathbb{R}^{m\times m}$ is an unreduced upper Hessenberg matrix, and $v_{m+1} \in \mathbb{R}^n$ satisfies $v_{m+1}^T v_{m+1} = 1$ and $V_m^T v_{m+1} = 0$. Another step of the Arnoldi process would give the matrix V_{m+1}, whose last column is v_{m+1}. We note for future reference that the vectors $v_1, v_2, \ldots, v_{m+1}$ determined by (12) form an orthonormal basis of the Krylov subspace

$$\mathbb{K}_{m+1}(A, v_1) = \text{span}\{v_1, Av_1, \ldots, A^m v_1\}. \tag{13}$$

For certain initial vectors v_1, the Arnoldi process may break down before the decomposition (12) has been determined. These (rare) cases allow certain simplifications and have to be treated separately. For notational simplicity, we assume throughout this paper that m is chosen small enough so that the Arnoldi decomposition (12) exists with $\eta_m \neq 0$.

The similar form of the equations (11) and (12) suggests that the Arnoldi process may be applied to determine a solution of (11). This observation is the basis of the solution method for (11) proposed by Datta and Saad [11]. The following results form the basis of their method. Throughout this paper $\|\cdot\|$ denotes the Euclidean vector norm or the associated induced matrix norm.

Theorem 2.1. *Let the matrix $H_m = [h_{i,j}]_{i,j=1}^m \in \mathbb{R}^{m\times m}$ be an unreduced upper Hessenberg matrix with spectrum $\lambda(H_m)$, and let $\{\mu_j\}_{j=1}^m$ be a set of distinct complex numbers, such that $\{\mu_j\}_{j=1}^m \cap \lambda(H_m) = \emptyset$. Introduce the quantities*

$$s = \prod_{j=1}^{m} (H_m - \mu_j I)e_1, \qquad \alpha = \prod_{j=1}^{m-1} h_{j+1,j}^{-1}. \tag{14}$$

Then the upper Hessenberg matrix $H_m - \alpha s e_m^T$ has the spectrum $\lambda(H_m - \alpha s e_m^T) = \{\mu_j\}_{j=1}^m$. If the set $\{\mu_j\}_{j=1}^m$ is invariant under complex conjugation, then the matrix $H_m - \alpha s e_m^T$ is real.

Proof. Slight modifications of this theorem are formulated by Datta [6] and Datta and Saad [11]. The proof presented in [6] can be modified to show the theorem. □

Lemma 2.1. *Let the matrices A, V_m and H_m, vectors v_1 and v_{m+1} and scalar η_m be those in the Arnoldi decomposition (12). Assume that $c \in \mathbb{K}_{m+1}(A, v_1)$ and $v_{m+1}^T c \neq 0$. Then there are a vector $f \in \mathbb{R}^m$ and scalar $\beta_m \in \mathbb{R}$, such that*

$$AV_m - V_m(H_m - f e_m^T) = \beta_m c e_m^T. \tag{15}$$

Proof. Let $\beta_m = \eta_m / v_{m+1}^T c$ and let the vector $f \in \mathbb{R}^m$ satisfy

$$\beta_m c = V_m f + \beta_m (v_{m+1}^T c) v_{m+1} = V_m f + \eta_m v_{m+1}. \tag{16}$$

The Arnoldi decomposition (12) and formula (16) yield

$$\begin{aligned} AV_m - V_m(H_m - f e_m^T) &= \eta_m v_{m+1} e_m^T + V_m f e_m^T \\ &= \eta_m v_{m+1} e_m^T + (\beta_m c - \eta_m v_{m+1}) e_m^T. \end{aligned}$$

This shows (15). □

Copyright © 2001 John Wiley & Sons, Ltd.
Prepared using nlaauth.cls

Note that the matrix $H_m - fe_m^T$ in (15) is of upper Hessenberg form. Therefore Lemma 2.1 shows that if we determine the vector v_1 so that $c \in \mathbb{K}_{m+1}(A, v_1)$, then the equation (15) is of the form (11) up to the scaling factor β_m. The following lemmata and theorem show how such a Krylov subspace can be determined and that $\lambda(H_m - fe_m^T) = \{\mu_j\}_{j=1}^m$.

Lemma 2.2. *Let the matrices A, V_m and H_m and vector v_1 be those of the Arnoldi decomposition (12). Let p be a polynomial of degree less than m. Then*

$$p(A)v_1 = V_m p(H_m)e_1. \tag{17}$$

Proof. It is easy to show that

$$A^j v_1 = V_m H_m^j e_1, \qquad 0 \le j < m,$$

and the lemma follows. □

Lemma 2.3. *Let $H_{m+1} = [h_{i,j}]_{i,j=1}^{m+1} \in \mathbb{R}^{(m+1)\times(m+1)}$ be an upper Hessenberg matrix and p a monic polynomial of degree m. Then*

$$e_{m+1}^T p(H_{m+1})e_1 = \prod_{j=1}^{m} h_{j+1,j}. \tag{18}$$

Proof. The result can be shown by induction. □

Theorem 2.2. *Let $A \in \mathbb{R}^{n\times n}$ and $c \in \mathbb{R}^n$ be defined by (11). Let $\{\mu_j\}_{j=1}^m$ be a set of m distinct complex numbers, such that $\{\mu_j\}_{j=1}^m \cap \lambda(A) = \emptyset$. Introduce the monic polynomial*

$$p_m(t) = \prod_{j=1}^{m}(t - \mu_j) \tag{19}$$

and let x be the unique solution of the linear system of equations

$$p_m(A)x = c. \tag{20}$$

Let V_m, H_m, v_{m+1} and η_m be determined by the Arnoldi decomposition (12) with initial vector $v_1 = x/\|x\|$. Define $\beta_m = \eta_m / v_{m+1}^T c$ and $f = \beta_m V_m^T c$. Then

$$c \in \mathbb{K}_{m+1}(A, v_1) \tag{21}$$

and

$$\lambda(H_m - fe_m^T) = \{\mu_j\}_{j=1}^m. \tag{22}$$

Proof. Equation (21) follows from (19) and (20). In order to show (22), we note that

$$f = \beta_m V_m^T c = \beta_m V_m^T p_m(A)x = \beta_m \|x\| V_m^T p_m(A)v_1. \tag{23}$$

Let $p_{m-1}(t) = \prod_{j=1}^{m-1}(t - \mu_j)$. Substituting $p_m(t) = (t - \mu_m)p_{m-1}(t)$ into the right-hand side of (23) and applying Lemma 2.2 yield

$$\begin{aligned} \beta_m \|x\| V_m^T p_m(A)v_1 &= \beta_m \|x\| V_m^T (A - \mu_m I)p_{m-1}(A)v_1 \\ &= \beta_m \|x\| V_m^T (A - \mu_m I)V_m p_{m-1}(H_m)e_1 \\ &= \beta_m \|x\| (H_m - \mu_m I)p_{m-1}(H_m)e_1 \\ &= \beta_m \|x\| p_m(H_m)e_1 = \beta_m \|x\| s, \end{aligned} \tag{24}$$

Copyright © 2001 John Wiley & Sons, Ltd.

Prepared using nlaauth.cls

where we have used that $H_m = V_m^T A V_m$ and s is defined by (14). It follows from equation (20) that

$$\beta_m \|x\| = \frac{\eta_m}{v_{m+1}^T c}\|x\| = \frac{\eta_m}{v_{m+1}^T p_m(A)x}\|x\| = \frac{\eta_m}{v_{m+1}^T p_m(A)v_1}. \tag{25}$$

Application of Lemmata 2.2 and 2.3 to the right-hand side of (25) yields

$$\begin{aligned} \frac{\eta_m}{v_{m+1}^T p_m(A)v_1} &= \frac{\eta_m}{v_{m+1}^T V_{m+1} p_m(H_{m+1})e_1} = \frac{\eta_m}{e_{m+1}^T p_m(H_{m+1})e_1} \\ &= \frac{\eta_m}{\prod_{j=1}^m h_{j+1,j}} = \frac{1}{\prod_{j=1}^{m-1} h_{j+1,j}}, \end{aligned} \tag{26}$$

where $H_{m+1} \in \mathbb{R}^{(m+1)\times(m+1)}$ is the upper Hessenberg matrix determined by applying $m+1$ steps of the Arnoldi process with initial vector v_1 to the matrix A. The last equality in (26) follows from the fact that $h_{m+1,m}$, the last subdiagonal entry of H_{m+1}, equals η_m.

Equations (25)-(26) show that

$$\alpha = \beta_m \|x\|,$$

where α is defined by (14). It follows from equations (23)-(24) that $f = \alpha s$. Hence, (22) is a consequence of Theorem 2.1. □

In the remainder of this section, we outline the computations required in order to determine the vectors x and f of Theorem 2.2. More details are provided in Sections 3-4. We first turn to the system of equations (20). Introduce the partial fraction decomposition

$$\frac{1}{p_m(t)} = \sum_{j=1}^m \frac{\alpha_j}{t-\mu_j}, \qquad \alpha_j = \frac{1}{p_m'(\mu_j)}. \tag{27}$$

The solution of (20) can be expressed as

$$x = \sum_{j=1}^m x_j, \tag{28}$$

where x_j solves

$$(A - \mu_j I)x_j = \alpha_j c, \qquad 1 \le j \le m. \tag{29}$$

The GMRES method is one of the most popular iterative methods for the solution of large linear systems of equations with a nonsymmetric matrix. Following Datta and Saad [11], we will apply this method to solve the linear systems (29). A recent description of the GMRES method is provided by Saad [19, Chapter 6].

The standard implementation of the GMRES method for the solution of the jth equation of (29) uses an Arnoldi decomposition of the matrix $A - \mu_j I$ of the form

$$(A-\mu_j I)V_\ell = V_\ell H_\ell^{(\mu_j)} + \eta_\ell^{(\mu_j)} v_{\ell+1} e_\ell^T = V_{\ell+1} H_{\ell+1,\ell}^{(\mu_j)}, \qquad V_\ell e_1 = c/\|c\|, \tag{30}$$

where we note that

$$H_\ell^{(\mu_j)} = H_\ell + \mu_j I_\ell \tag{31}$$

Copyright © 2001 John Wiley & Sons, Ltd.

Prepared using nlaauth.cls

and H_ℓ is the upper Hessenberg matrix in the Arnoldi decomposition

$$AV_\ell = V_\ell H_\ell + \eta_\ell v_{\ell+1} e_\ell^T, \qquad V_\ell e_1 = c/\|c\|. \tag{32}$$

The matrix $H_{\ell+1,\ell}^{(\mu_j)} \in \mathbb{C}^{(\ell+1)\times\ell}$ in (30) consists of the first ℓ columns of the upper Hessenberg matrix $H_{\ell+1}^{(\mu_j)} \in \mathbb{C}^{(\ell+1)\times(\ell+1)}$ determined by application of $\ell+1$ steps of the Arnoldi process to the matrix $A - \mu_j I$. The matrix $V_{\ell+1}$ in (30) is independent of the value of μ_j. This fact and the simple form (31) of the matrix $H_\ell^{(\mu_j)}$ made it possible for Datta and Saad [11] to solve the m linear systems (29) by the GMRES method by only applying the Arnoldi process once to compute the Arnoldi decomposition (32) for some ℓ sufficiently large, and then determining the decompositions (30) by modifying the matrix H_ℓ in (30) according to (31).

Using the Arnoldi decomposition (30), the GMRES method determines an approximate solution $x_j^{(\ell)}$ of the jth linear systems of equations (29), such that the associated residual vector

$$r_j^{(\ell)} = \alpha_j c - (A - \mu_j) x_j^{(\ell)} \tag{33}$$

satisfies

$$\begin{aligned} \|r_j^{(\ell)}\| &= \|\alpha_j c - (A-\mu_j I)x_j^{(\ell)}\| = \min_{x\in\mathbb{K}_\ell(A-\mu_j I,c)} \|\alpha_j c - (A-\mu_j I)x\| \\ &= \min_{y\in\mathbb{C}^\ell} \|\alpha_j c - AV_\ell y\| = \min_{y\in\mathbb{C}^\ell} \|\alpha_j c - V_{\ell+1}H_{\ell+1,\ell}^{(\mu)} y\| \\ &= \min_{y\in\mathbb{C}^\ell} \|\alpha_j \|c\| e_1 - H_{\ell+1,\ell}^{(\mu)} y\|. \end{aligned} \tag{34}$$

Denote the solution of the minimization problem (34) by $y_j^{(\ell)}$. Having evaluated this solution, we compute $x_j^{(\ell)} = V_\ell y_j^{(\ell)}$.

The following theorem sheds light on the behavior of the linear system (29) as $\mathrm{Re}(\mu_j)$ decreases. In particular, the theorem shows that the condition number $\kappa(A-\mu_j I) = \|A - \mu_j I\|\|(A-\mu_j I)^{-1}\|$ is not large when $\mathrm{Re}(\mu_j)$ is of large magnitude.

Theorem 2.3. *Let $x_j^{(\ell)}$ be the approximate solution of the jth linear system (29) determined by the GMRES method by solving the minimization problem (34), and let $r_j^{(\ell)}$ be the associated residual vector (33). Then*

$$r_j^{(\ell)} \to 0, \qquad \kappa(A-\mu_j I) \to 1 \quad \textit{and} \quad x_j^{(\ell)} \to x_j \quad \textit{as} \quad |\mu_j| \to \infty. \tag{35}$$

Proof. It follows from equation (34) that

$$\|r_j^{(\ell)}\| \le \|r_j^{(\ell-1)}\| \le \cdots \le \|r_j^{(1)}\|.$$

It is straightforward to bound $\|r_j^{(1)}\|$. Equation (34) with $\ell = 1$ yields that $x_j^{(1)} = \gamma_j c$, where the scalar γ_j solves

$$\|r_j^{(1)}\| = \min_{\gamma\in\mathbb{C}} \|\alpha_j c - (A-\mu_j I)(\gamma c)\|.$$

Thus,

$$\gamma_j = \frac{\alpha_j c^T (A^T - \bar{\mu}_j I)c}{c^T (A^T - \bar{\mu}_j I)(A-\mu_j I)c} \to -\frac{\alpha_j}{\mu_j} \quad \text{as} \quad |\mu_j| \to \infty. \tag{36}$$

Copyright © 2001 John Wiley & Sons, Ltd.

Prepared using nlaauth.cls

Substituting $x_j^{(1)} = \gamma_j c$ into (33) yields, in view of (36), that

$$r_j^{(1)} \to \alpha_j c - (A - \mu_j I)(-\frac{\alpha_j}{\mu_j})c \quad \text{as} \quad |\mu_j| \to \infty, \tag{37}$$

and the right-hand side expression in (37) converges to zero as $|\mu_j|$ increases.

To show that the condition number $\kappa(A - \mu_j I)$ approaches one as $|\mu_j| \to \infty$, observe that

$$\|A - \mu_j I\|\|(A - \mu_j I)^{-1}\| = \|I - \frac{1}{\mu_j}A\|\|(I - \frac{1}{\mu_j}A)^{-1}\| \to 1 \quad \text{as} \quad |\mu_j| \to \infty.$$

Finally,

$$x_j^{(1)} - x_j = -(A - \mu_j I)^{-1} r_j^{(1)} \to 0 \quad \text{as} \quad |\mu_j| \to \infty.$$

□

In applications of the GMRES method, we may, for instance, choose ℓ in the Arnoldi decomposition (32) as the smallest integer, such that the residual errors satisfy the bound

$$\|r_j^{(\ell)}\| \leq \epsilon, \qquad 1 \leq j \leq m, \tag{38}$$

for some specified $\epsilon > 0$.

The theorem suggests that for a fixed vector c and parameter $\ell > 0$, the norm of the residual error $\|r_j^{(\ell)}\|$ and of the error $\|x_j^{(1)} - x_j\|$ decrease as $\mathrm{Re}(\mu_j)$ decreases. Indeed, we have observed this behavior in numerous numerical examples, already for fairly small values of $|\mathrm{Re}(\mu_j)|$. This indicates that in order for the GMRES method to give rapid convergence, we should choose μ_j to have a negative real part of large magnitude.

On the other hand, choosing all the eigenvalues μ_j of the matrix H in (11) with real parts much smaller than the real part of the leftmost eigenvalue of the matrix H_m in the Arnoldi decomposition (12), makes it necessary to move the eigenvalues of H_m a long distance to the μ_j in the eigenvalue assignment problem for the matrix H discussed in Theorem 2.1. The sensitivity of this eigenvalue assignment problem increases the further the eigenvalues have to be moved; see Mehrmann and Xu [17] for bounds for the condition number of the eigenvalue assignment problem. Related results can be also be found in [3, 18]. Let the Hessenberg matrix $H = H_m - \alpha s e_m^T$ have spectral factorization

$$H = W_H \Lambda_H W_H^{-1},$$

with a diagonal matrix Λ_H. Then

$$\kappa(W_H) = \|W_H\|\|W_H^{-1}\|$$

provides an estimate of the sensitivity of the eigenvalue assignment problem, see [3, 18], and is displayed in the numerical examples of Section 4.

Thus, we would like to choose the real part of the μ_j small enough to make the GMRES algorithm converge rapidly and large enough so that the eigenvalue assignment problem discussed in Theorem 2.1 is not too ill-conditioned.

The next section considers a further aspect of the choice of the μ_j; there we discuss how the choice of the μ_j affects the magnitude of the coefficients α_j in the partial fraction representation (27).

Copyright © 2001 John Wiley & Sons, Ltd.

Prepared using nlaauth.cls

3. PARTIAL FRACTION REPRESENTATION

The eigenvalues μ_j of the matrix H in (11) determine the partial fraction representation (27). The location of the eigenvalues determines how accurately the partial fraction can be evaluated in finite precision arithmetic. Close poles can greatly compromise the accuracy of the computed value of $1/p_m(t)$. In addition coefficients α_j of large magnitude in the representation (27) may make it necessary to compute an Arnoldi decomposition (32) with a large value of ℓ in order to satisfy the residual error bound (38) since residual errors $r_j^{(\ell)}$ determined by the GMRES method are proportional to α_j.

Example 3.1. Let $p_2(t) = (t-\epsilon)(t+\epsilon)$ with $|\epsilon| > 0$ tiny. Consider the evaluation of the partial fraction representation

$$\frac{1}{p_2(t)} = \frac{1}{2\varepsilon}\frac{1}{t-\varepsilon} - \frac{1}{2\varepsilon}\frac{1}{t+\varepsilon} \tag{39}$$

in finite precision arithmetic and assume that the computations are carried out with three significant digits. Let $\epsilon = 1/900$. Then evaluation of the representation (39) yields the value 0.00.

Evaluating the product form representation

$$\frac{1}{p_2(t)} = \frac{1}{(t-\varepsilon)(t+\varepsilon)}$$

with the same arithmetic and value of ε yields the value 1.00. In exact arithmetic we have

$$\frac{1}{p_2(1)} = \frac{1}{1-\varepsilon^2} = 1.00000.$$

This example illustrates that the partial fraction representation may yield significantly lower accuracy than the product form representation. The difficulty in the present example stems from loss of significant digits due to partial fraction coefficients of large magnitude and opposite sign. This in turn is caused by the poles being very close. □

Example 3.2. This example shows that the magnitude of the coefficients in the partial fraction representation depends not only on the distance between poles, but also on the distribution of the poles. Let

$$p_m(t) = \prod_{j=1}^{m}(t - \frac{j}{m}).$$

Then the partial fraction coefficients (27) are given by

$$\alpha_\ell = \frac{m^{m-1}}{\prod\limits_{\substack{j=1 \\ j\neq\ell}}^{m}(\ell - j)}. \tag{40}$$

In particular, for $m = 10$, we obtain

$$\alpha_6 = -\alpha_5 \approx 3 \cdot 10^5.$$

Copyright © 2001 John Wiley & Sons, Ltd.

Prepared using nlaauth.cls

Thus, evaluation of the partial fraction representation (27) in finite precision arithmetic may cause severe cancellation of significant digits despite the fact that $1/p_m(t)$ does not have close poles. Moreover, the large magnitude of the partial fraction coefficients may make the use of an Arnoldi decomposition (32) with a large value of ℓ necessary. Note that some of the coefficients (40) are of much larger magnitude than others. This is measured by the quotient

$$q(m) = \frac{\max_{1\leq j\leq m} |\alpha_j|}{\min_{1\leq j\leq m} |\alpha_j|} \tag{41}$$

We have $q(10) = 126$. Note that $q(m)$ is invariant under linear transformation of the zeros μ_j of $p_m(t)$. In particular, the value $q(10)$ is the same whenever the zeros μ_j are equidistant on some interval in the complex plane. □

Example 3.3. Let $p_m(t) = \cos(m \arccos(t/2))$, i.e., $p_m(t)$ is a Chebyshev polynomial of the first kind for the interval $[-2, 2]$. Its zeros are

$$\mu_j = 2\cos(\frac{2j-1}{2m}\pi). \tag{42}$$

Then

$$p'_m(\mu_j) = \frac{(-1)^{j-1}m}{2\sin(\frac{2j-1}{2m}\pi)}$$

yields

$$|\alpha_j| \leq \frac{2}{m}, \qquad 1 \leq j \leq m. \tag{43}$$

The distance between some adjacent zeros μ_j is fairly small for large values of m. For instance,

$$|\mu_1 - \mu_2| = 2|\cos(\frac{\pi}{2m}) - \cos(\frac{3\pi}{2m})| = 4|\sin\frac{\pi}{m}\sin\frac{\pi}{2m}| < \frac{2\pi^2}{m^2}.$$

Nevertheless, the bound (43) shows that $\max_{1\leq j\leq m} |\alpha_j|$ converges to zero as m increases. Moreover, all partial fraction coefficients are of roughly the same order of magnitude. For instance, we have $q(10) = 6.4$, where q is defined by (41). This value is much smaller than the value for equidistant zeros reported in Example 3.2.

Numerical experiments indicate that the partial fraction representation (27) with the zeros (42) does not suffer from severe loss of significant digits when evaluated in finite precision arithmetic. □

Further examples of difficulties of partial fraction representations, as well as a discussion on how to ameliorate these difficulties, can be found in [2].

4. CHOICE OF EIGENVALUES OF H

The examples in Section 3 illustrate that the use of a partial fraction representation can, but must not, lead to severe loss of accuracy when evaluated in finite precision arithmetic. Partial fraction coefficients of large magnitude can give rise to severe loss of accuracy due to cancellation of significant digits during evaluation of the partial fraction representation. Conversely, numerical experiments display little loss of accuracy when evaluating a partial

Copyright © 2001 John Wiley & Sons, Ltd.

Prepared using nlaauth.cls

fraction representation whose coefficients all are of roughly same, not very large magnitude. Example 3.3 suggests that choosing the eigenvalues μ_j of the matrix H in (11) as zeros of a Chebyshev polynomial gives coefficients α_j in the partial fraction representation (27) of roughly the same magnitude, which typically is not very large.

We propose the following selection of eigenvalues $\{\mu_j\}_{j=1}^m$ of the matrix H. In order to secure that the matrices $A-\mu_j I$ are nonsingular, we determine approximations of the eigenvalues of A with smallest real part. The eigenvalues of the upper Hessenberg matrix H_ℓ in the Arnoldi decomposition (32) used for the solution of the linear systems (29) by the GMRES method furnishes approximations of the desired eigenvalues of A. If these approximations are not sufficiently accurate then they can be improved, e.g., by the Implicitly Restarted Arnoldi method proposed by Sorensen [20]. We choose the μ_j to be zeros of the Chebyshev polynomial of the first kind of degree m for the interval $[\tau+i\rho, \tau-i\rho]$, $i=\sqrt{-1}$, where $\tau < \min_{t\in\lambda(A)} \mathrm{Re}(t)$ and $\rho = \max_{t\in\lambda(A)} \mathrm{Im}(t)$. The accurate determination of $\max_{t\in\lambda(A)} \mathrm{Im}(t)$ is not crucial for the performance of the the algorithm below, which computes the solution $\{H, X\}$ of the Sylvester-observer equation

$$AX - XH = \beta_m c e_m^T, \tag{44}$$

where β_m is the same as in Lemma 2.1. This equation is obtained by scaling the equation (11) by the factor β_m. Therefore, if $\{H, X\}$ solves (44), then $\{H, \beta_m^{-1}X\}$ solves (11).

Algorithm 1 Input: A, c, m, ℓ. Output: H, X.

1. Apply ℓ steps of the Arnoldi process with initial vector c to determine an Arnoldi decomposition of the form (32). Compute approximations of the eigenvalues of A with smallest and largest real parts.
2. If the eigenvalue approximations computed in Step 1 are not sufficiently accurate, then compute improved approximations with the Implicitly Restarted Arnoldi method [20].
3. Choose the set of eigenvalues $\{\mu_j\}_{j=1}^m$ of the matrix H in (11). We propose that μ_j be chosen as zeros of Chebyshev polynomials for the interval $[\tau + i\rho, \tau - i\rho]$ introduced above. This set of eigenvalues is invariant under complex conjugation.
4. Solve the systems of equations (29) by the GMRES method as described in Section 3, i.e., use the available Arnoldi decomposition (32) computed in Step 1 to determine the decompositions (30) for $1 \le j \le m$. The latter Arnoldi decompositions are used to compute approximations $x_j^{(\ell)}$ of x_j for $1 \le j \le m$ by the GMRES method, i.e., by solving the minimization problems (34). If the approximations $x_j^{(\ell)}$ are sufficiently accurate, then evaluate the approximate solution

$$x^{(\ell)} = \sum_{j=1}^{m} x_j^{(\ell)}$$

of the linear system of equations (20), cf. the formula (28), otherwise increase ℓ in the Arnoldi decompositions (32) and (30).
5. Apply m steps of the Arnoldi process with initial vector $v_1 = x/\|x\|$ to determine the Arnoldi decomposition (12). This gives the matrix H_m.
6. Compute the vector $f \in \mathbb{R}^m$, such that $\lambda(H_m - f e_m^T) = \{\mu_j\}_{j=1}^m$ by using the formulas of Theorem 2.1 with β_m defined in Lemma 2.1. Let $X = V_m/\beta_m$, $H = H_m - f e_m^T$. □

Copyright © 2001 John Wiley & Sons, Ltd.

Prepared using nlaauth.cls

We remark that the vector f used for the eigenvalue assignment in Step 6 of the algorithm can be computed either by the formulas of Theorem 2.1 or according to Lemma 2.1. When the matrix A is large, we generally only solve the equation (20) approximately, and then these two approaches to compute the vector f are not equivalent. When the system (20) is not solved exactly and we compute f by the formulas of Lemma 2.1, the eigenvalue of the Hessenberg matrix H in general are further away from the μ_j than when the formulas of Theorem 2.1 are applied. We therefore use the latter approach in Algorithm 1.

5. NUMERICAL EXAMPLES

The computations reported in this section were carried out on an Intel Pentium workstation using Matlab 5.3 and floating point arithmetic with 16 significant digits. In all examples, we used the same matrix $A \in \mathbb{R}^{500\times 500}$, which we determined by generating its spectral decomposition. The eigenvalues λ_j were distributed in the disk $\mathbb{D} = \{z : |z+1| = 1\} \subset \mathbb{C}$ as follows. We let $\lambda_j = \rho_j \exp(i\tau_j) - 1$, $i = \sqrt{-1}$, where the ρ_j and τ_j were uniformly distributed in the intervals $[0, 1]$ and $[0, \pi]$, respectively. When $\mathrm{Im}(\lambda_j) > 0$, $\bar{\lambda}_j = \rho_j \exp(-i\tau_j) - 1$ was also chosen as an eigenvalue. We determined the eigenvector matrix W with real or complex conjugate columns, the real and imaginary parts of the entries uniformly distributed in the interval $[0, 1]$. This gave an eigenvector matrix with condition number $\kappa(W) = \|W\|\|W^{-1}\| = 5.1 \cdot 10^4$. The eigenvalues of A are marked by dots in the figures below.

The computations were carried out as described by Algorithm 1 and illustrate the numerical consequences of the choice of eigenvalues μ_j of the Hessenberg matrix H in Step 3 of Algorithm 1.

Example 5.1. Steps 1-2 of Algorithm 1 with $\ell = 20$ gave the approximations

$$\min_{t\in\lambda(A)} \mathrm{Re}(t) = -1.9, \qquad \max_{t\in\lambda(A)} \mathrm{Im}(t) = 0.9.$$

We let $m = 8$ and choose the μ_j to be the zeros of a Chebyshev polynomial of the first kind of degree 8 for the interval $[-1.9 - 0.90i, -1.9 + 0.90i]$, where $i = \sqrt{-1}$. The eigenvalues of A and the μ_j are displayed by Figure 1. The solution of the equation (20) is determined by solving the m systems of equations (29) using the Arnoldi decomposition (32) with $\ell = 20$. Numerical results are shown in Table I under the heading "Nearby μ_j."

For comparison, we repeat the computations of Step 3-6 of Algorithm 1 with the μ_j allocated further to the left in the complex plane. Specifically, we let the μ_j be the zeros of the Chebyshev polynomial of the first kind of degree 8 for the interval $[-5.7 - 0.90i, -5.7 + 0.90i]$. The μ_j are displayed in Figure 2 and the performance of Algorithm 1 is shown in Table I under the heading "Distant μ_j."

Table I. Effect of the choice of zeros on the solution of the Sylvester-observer equation (11)

Quantity	Nearby μ_j	Distant μ_j
$\|AX - XH - \beta_m c\|$	$8.4 \cdot 10^2$	$3.6 \cdot 10^{-3}$
$\kappa(W_H)$	$1.3 \cdot 10^3$	$5.1 \cdot 10^{11}$

Copyright © 2001 John Wiley & Sons, Ltd.

Prepared using nlaauth.cls

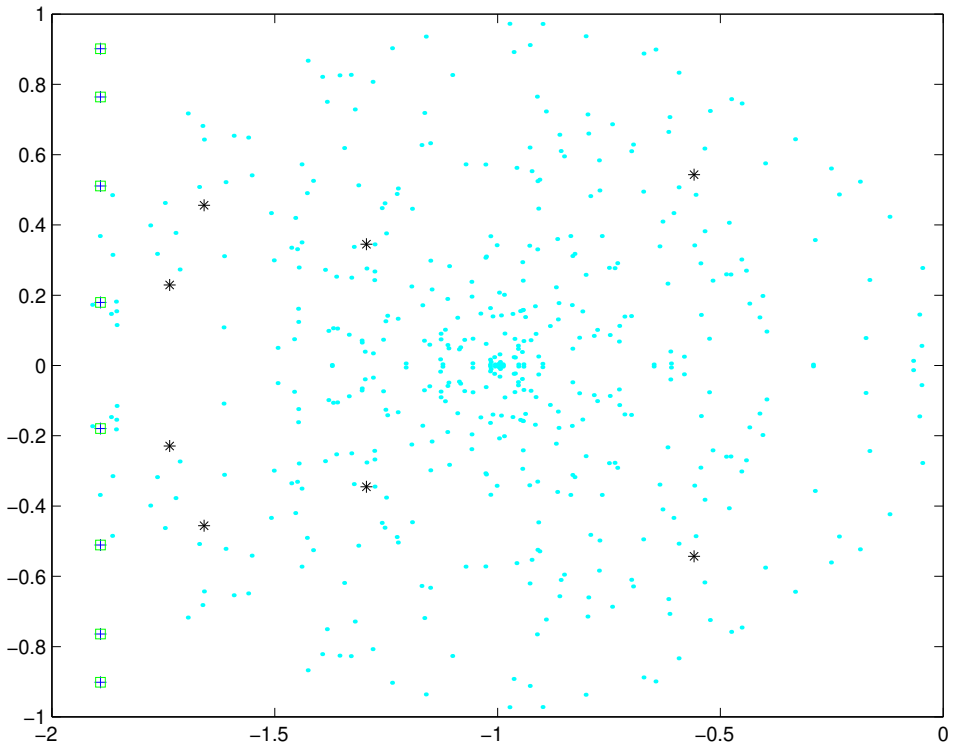

Figure 1. $\cdot$: Eigenvalues of A, $*$: Eigenvalues of H_m, $\square$: μ_j, $+$: Eigenvalues of $H = H_m - fe_m^T$

We see that the set of μ_j further away from the imaginary axis produce substantially lower residual error. However, the condition number of the matrix H is larger for these μ_j and this makes the pole placement problem more sensitive to perturbations. $\square$

Example 5.2. This example illustrates the effect of the distribution of the μ_j on the solution of the Sylvester equation. We let $m = 14$ and solve the equation (20) using the Arnoldi decomposition (32) with $\ell = 50$. The μ_j are chosen to be zeros of the Chebyshev polynomial of the first kind of degree 14 for the interval $[-2.7 - 0.96i, -2.7 + 0.96i]$. Figure 3 shows the spectra of A, H, and H_m, as well as the μ_j. Table II displays the performance of Algorithm 1.

For comparison, the computations were also carried out with the μ_j allocated equidistantly

Table II. Effect of the choice of the μ_j on the solution of the Sylvester-observer equation (11)

Quantity	μ_j equidistant	μ_j zeros of Chebyshev polynomials
$\|AX - XH - \beta_m c\|$	$1.8 \cdot 10^0$	$5.3 \cdot 10^{-2}$
$\max_{1\le i\le m} \|\alpha_i\|$	$1.6 \cdot 10^4$	$8.9 \cdot 10^2$

Copyright © 2001 John Wiley & Sons, Ltd.

Prepared using nlaauth.cls

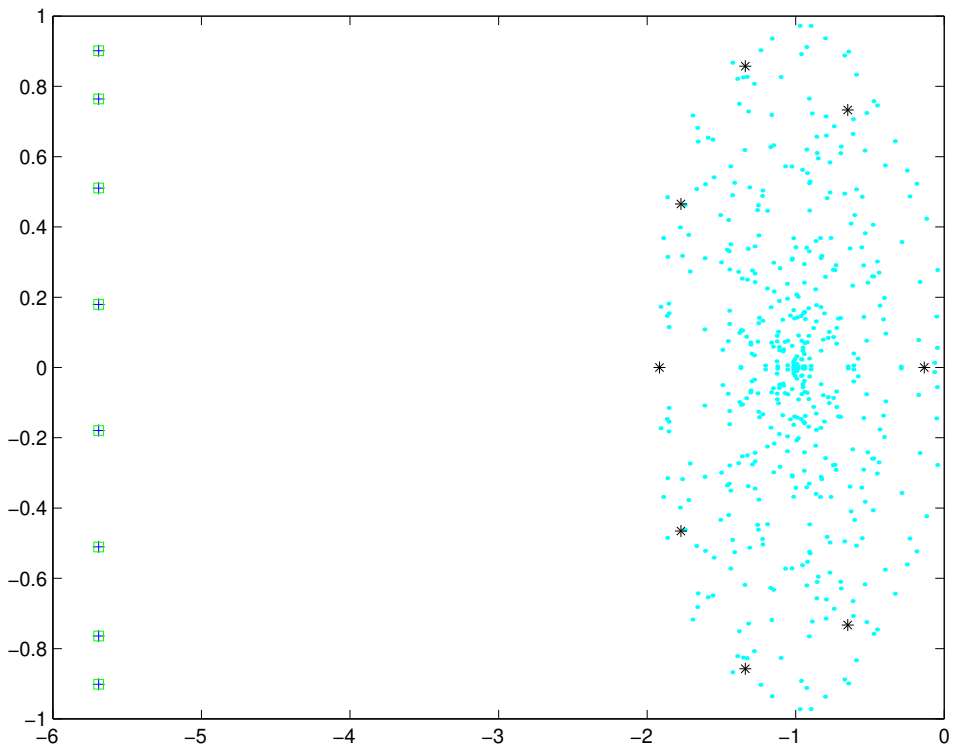

Figure 2. $\cdot$: Eigenvalues of A, $*$: Eigenvalues of H_m, $\square$: μ_j, $+$: Eigenvalues of $H = H_m - fe_m^T$

in the interval $[-2.7 - 0.96i, -2.7 + 0.96i]$; see Figure 4. Numerical results are presented in Table II. Clearly, Algorithm 1 performs better when the μ_j are zeros of Chebyshev polynomials. $\square$

6. CONCLUSION

The performance of the method by Datta and Saad [11] for the solution of the Sylvester-observer equations (11) depends on the choice of the set of eigenvalues $\{\mu_j\}_{j=1}^m$ of the matrix H. Let $\tau_0 = \min_{t\in\lambda(A)} \mathrm{Re}(t)$ and $\rho = \max_{t\in\lambda(A)} \mathrm{Im}(t)$, and assume that $\tau_0 \leq 0$ and $\rho > 0$. We propose to choose the μ_j to be zeros of the Chebyshev polynomial of the first kind of degree m for the interval $[\tau + i\rho, \tau - i\rho]$ for some $\tau < \tau_0$. The smaller value of τ, the more accurately we can solve the linear system (20) for a fixed value of ℓ in the Arnoldi decomposition (32), but the more ill-conditioned the eigenvalue assignment problem of Theorem 2.1. Typically, we seek to choose τ small, but large enough to be able to solve the eigenvalue assignment problem to desired accuracy.

ACKNOWLEDGEMENT

Copyright © 2001 John Wiley & Sons, Ltd.

Prepared using nlaauth.cls

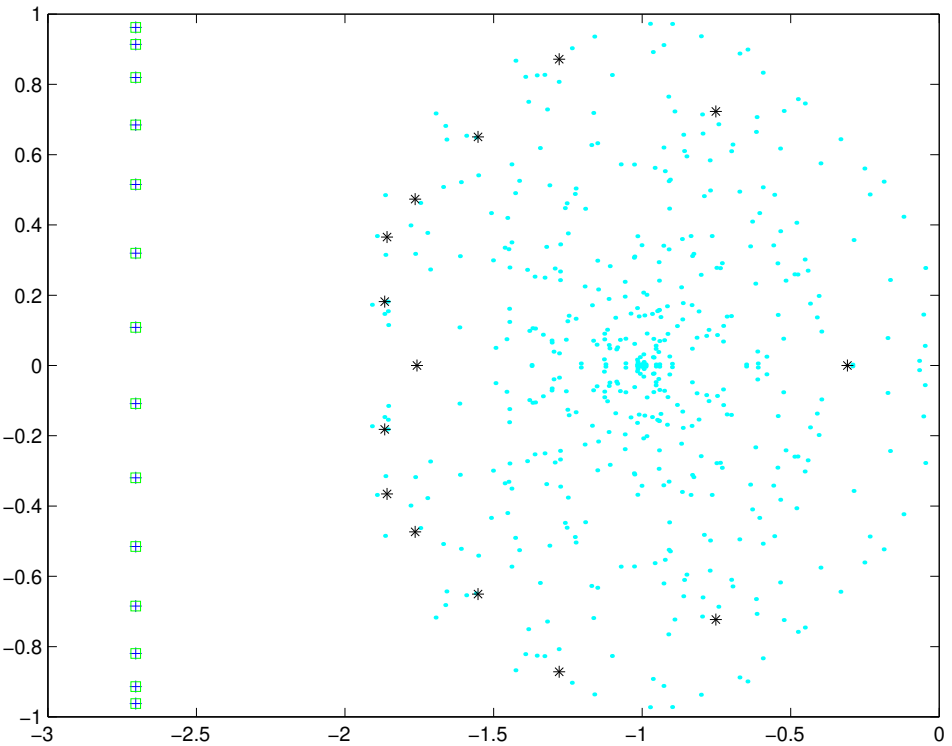

Figure 3. $\cdot$: Eigenvalues of A, $*$: Eigenvalues of H_m, $\square$: μ_j, $+$: Eigenvalues of $H = H_m - fe_m^T$

We would like to thank the referee for comments.

REFERENCES

1. C. Bischof, B. N. Datta and A. Purkyastha, A parallel algorithm for the Sylvester-observer matrix equation, *SIAM Journal on Scientific Computing* 1996; **17**:686–698.
2. D. Calvetti, E. Gallopoulos and L. Reichel, Incomplete partial fractions for parallel evaluation of rational matrix functions, *Journal of Computational and Applied Mathematics* 1995; **59**:349–380.
3. D. Calvetti, B. Lewis and L. Reichel, On the selection of poles in the single input pole placement problem, *Linear Algebra and Its Applications* 1999; **302-303**:331–345.
4. D. Calvetti and L. Reichel, Application of ADI iterative methods to the restoration of noisy images, *SIAM Journal on Matrix Analysis and Applications* 1996; **17**:165–186.
5. D. Calvetti and L. Reichel, Numerical aspects of some solution methods for large Sylvester-observer equations, in *Proceedings of the 36th IEEE Conference on Decision and Control*, IEEE, Piscataway, 1997:4389–4393
6. B. N. Datta, An algorithm to assign eigenvalues in a Hessenberg matrix, *IEEE Transactions on Automatic Control* 1987; **AC-32**:414–417.
7. B. N. Datta, Linear and Numerical Linear Algebra in Control Theory: Some Research Problems, *Linear Algebra and Its Applications* 1994; **198**:755–790.
8. B. N. Datta, Krylov subspace methods in control: an overview, in *Proceedings of the 36th IEEE Conference on Decision and Control*, IEEE, Piscataway, 1997:3844–3848.

Copyright © 2001 John Wiley & Sons, Ltd.

Prepared using nlaauth.cls

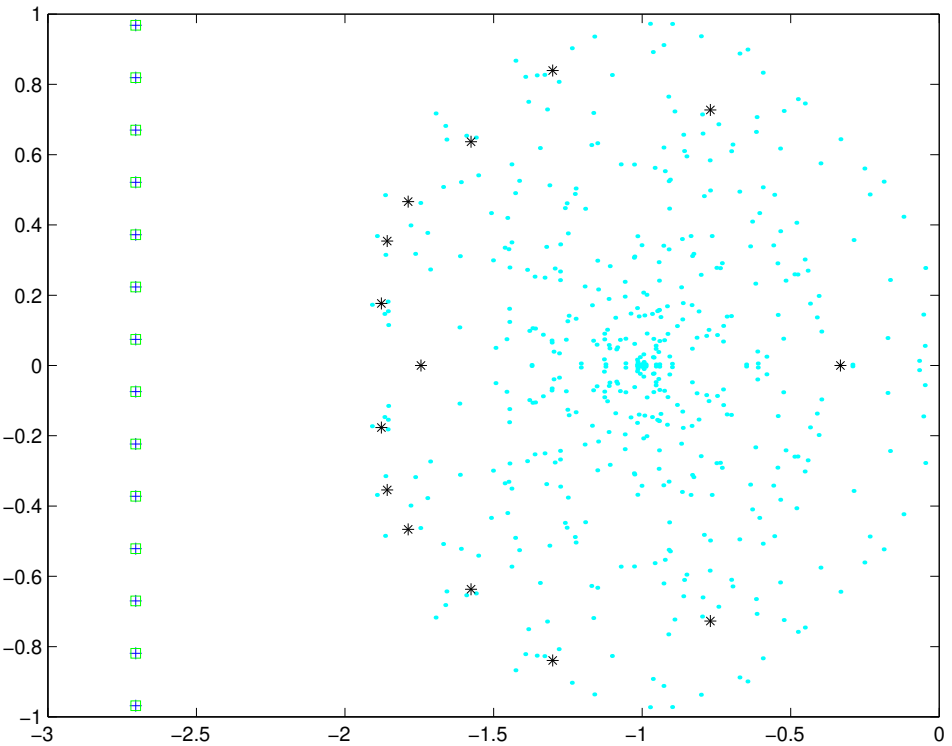

Figure 4. $\cdot$: Eigenvalues of A, $*$: Eigenvalues of H_m, $\square$: μ_j, $+$: Eigenvalues of $H = H_m - fe_m^T$

9. B. N. Datta, *Numerical Methods for Linear Control Systems Design and Analysis*, Academic Press, to appear.
10. B. N. Datta and C. Hetti, Generalized Arnoldi methods for the Sylvester-observer equation and the multi-input pole placement problem, in *Proceedings of the 36th IEEE Conference on Decision and Control*, IEEE, Piscataway, 1997:4379–4383.
11. B. N. Datta and Y. Saad, Arnoldi methods for large Sylvester-like observer matrix equations, and an associated algorithm for partial spectrum assignment, *Linear Algebra and Its Applications* 1991; **154–156**:225–244.
12. C. Hetti, On numerical solutions of the Sylvester-observer equation, and the multi-input eigenvalue assignment problem, Ph.D. thesis, Department of Mathematical Sciences, Northern Illinois University, DeKalb, IL, 1996.
13. R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*, Cambridge Univ. Press, Cambridge, 1991.
14. D. Y. Hu and L. Reichel, Krylov subspace methods for the Sylvester equation, *Linear Algebra and Its Applications* 1992; **172**:283–313.
15. T. Kailath, *Linear Systems*, Prentice-Hall, Englewood Cliffs, 1980.
16. D. G. Luenberger, Observers for multivariable systems, *IEEE Transactions on Automatic Control* 1966; **AC-11**:190–197.
17. V. Mehrmann and H. Xu, An analysis of the pole placement problem I. The single-input case, *Electronic Transactions on Numerical Analysis* 1996; **4**:89–105.
18. V. Mehrmann and H. Xu, Choosing the poles so that the single-input pole placement problem is well conditioned, *SIAM Journal on Matrix Analysis and Applications* 1998; **19**:664–681.
19. Y. Saad, *Iterative Methods for Sparse Linear Systems*, PWS, Boston, 1996.
20. D. C. Sorensen, Implicit application of polynomial filters in a k-step Arnoldi method, *SIAM Journal on Matrix Analysis and Applications*, 1992; **13**:357–385.

Copyright © 2001 John Wiley & Sons, Ltd.

Prepared using nlaauth.cls