Control
and Cybernetics
VOL. 13 (1984) No. 1-2

A stable computational algorithm for synthesis of state observers

by

P. Hr. PETKOV

N. D. CHRISTOV

Department of Automatics
Higher Institute of Mechanical and
Electrical Engineering
1156 Sofia, Bulgaria

M. M. KONSTANTINOV

Institute of Engineering Cybernetics
and Robotics
Akad. G. Bonchev Str., Bl. 2
1113 Sofia, Bulgaria

Numerically stable algorithms for synthesis of full and reduced order state observers are proposed. First the general problem of synthesis of state observers is posed and the set of all attainable structures of the observer is completely described. A constructive method for obtaining all solutions of the problem is proposed. The computational algorithm is based on the orthogonal canonical form of linear time-invariant systems and involves orthogonal similarity transformations only.

1. Introduction

In the synthesis of state observers as proposed in [1], [2] the following problems remained open:

— description of the set of all observers that are attainable for a given system, and

— development of a general method for obtaining all solutions of the equations for determining the observers' matrices.

A complete solution of these problems was given in [3], [4]. However the general method proposed there is not always suitable for numerical computations since it involves transformations which are not numerically stable. Thus an important problem in the synthesis of state observers remains the creation of numerically stable algorithms for obtaining the observers' matrices.

In the present paper an efficient computational algorithm is proposed for synthesis of full and reduced order state observers. It involves only orthogonal transformations and is numerically stable.

2. Statement of the problem

Consider the completely observable system

$$\dot{x}(t)=Ax(t)+Bu(t)$$
$$y(t)=Cx(t) \tag{1}$$

where $x(t)\in R^n$, $u(t)\in R^m$, $y(t)\in R^r$, and rank $C=r$.

Under the assumption that only the output $y(t)$ is available there arises the problem of synthesis of a state observer

$$\dot{z}(t)=Fz(t)+Gy(t)+Hu(t),\ z(t)\in R^q, \tag{2}$$

where $q=n$ (a full order observer), or $q=n-r$ (a reduced order observer).

The observer matrices satisfy the equations [2]

$$TA-FT=GC, \tag{3}$$

$$H=TB \tag{4}$$

and the relation

$$\text{rank } M(T)=n, \tag{5}$$

where $M(T)=T$ for $q=n$, and $M(T)=\begin{bmatrix} C \\ \hdashline T \end{bmatrix}$ for $q=n-r$.

The matrix $F\in R^{q\cdot q}$ is chosen so that the system

$$\dot{v}(t)=Fv(t) \tag{6}$$

of the dynamic error $v(t)=z(t)-Tx(t)$ has a prescribed dynamics.

The set $E_q\subset R^{q\cdot q}$ of all matrices $F\in R^{q\cdot q}$ such that equations (3), (4) have a solution (T, G, H) satisfying (5), is called the set of attainable matrices of the system (6).

According to this definition the problem of determining all matrices (T, G, H), satisfying (3)—(5) for a $F\in E_q$, is said to be the general problem of synthesis of state observers (GPSO). The corresponding triple (T, G, H) is said to be the solution of GPSO.

It is clear in view of (4) that only the determination of the matrices T, G satisfying (3), (5) is of interest.

3. Solution of the general problem of synthesis of state observers

Let $s\ (n/r\leqq s\leqq n+1-r)$ be the observability index of the pair (C, A): $s=\min\{i:\ \text{rank } D_i=n\}$, where

$$D_i=[C^T, A^T C^T, \ldots, (A^T)^{i-1} C^T].$$

Define the conjugate Kronecker indices $r_1, ..., r_s$ ($r_1 \geqq ... \geqq r_s \geqq 1, r_1 + ... + r_s = n$):

$$r_1 = r,$$
$$r_i = \text{rank } D_i\text{-rank } D_{i-1}, \quad i \in \overline{2, s}.$$

Then the set of Kronecker indices of (C, A) is $\{s_1, ..., s_r\}$, $s_1 \geqq ... \geqq s_r \geqq 1$, $s_1 + ... + s_r = n$, where s_j is the number of $r_i - s$ that are $\geqq j$.

Let t_j be the number of the integers from the set $\{r_2, ..., r_s\}$, which are $\geqq j$: $t_1 \geqq ... ... \geqq t_{r_2} \geqq 1$, $t_1 + ... + t_{r_2} = n - r$.

For each $F \in R^{q \cdot q}$ denote by $f_1, ..., f_k$ the degrees of the invariant polynomials of F ordered by magnitude: $f_1 \geqq ... \geqq f_k \geqq 1$, $f_1 + ... + f_k = q$.

Now the set $E_q \subset R^{q \cdot q}$ can be completely described as follows:

PROPOSITION [3], [4].

1. Let $q = n$. Then $F \in E_n$ iff

$$f_1 + ... + f_i \geqq s_1 + ... + s_i \text{ for each } i = 1, ..., k-1. \tag{7}$$

2. Let $q = n - r$. Then $F \in E_{n-r}$ iff

$$f_1 + ... + f_i \geqq t_1 + ... + t_i \text{ for each } i = 1, ..., k-1. \tag{8}$$

The first part of the proposition is a result dual to the well known theorem of Rosenbrock [5].

As a corollary of part 2 of the proposition one obtains that E_q is the set of cyclic $q \times q$ matrices iff $r_2 = n - r$, and in particular if $r = 1$.

The proof of the proposition is based on a constructive method for determining the matrices T, G satisfying (4), (5) [3], [4]. This method, however, is numerically unstable which restricts the area of its implementation especially for high order and ill-conditioned problems.

In the following section a numerically stable algorithm for synthesis of full and reduced order state observers is described.

4. A numerically stable algorithm for syntheis of state observers

4.1. Full Order Observers

For $q = n$ one can choose $T = I_n$ which corresponds to the so called identical observer. Hence the problem is to determine the matrix G from the equation

$$A - GC = F, \quad F \in E_n. \tag{9}$$

We shall consider the case when the dynamic properties of (6) are determined by the spectrum $s(F)$ of F.

Let $s_D = (s_1, ..., s_a;\ p_1 \pm jq_1, ..., p_b \pm jq_b)$ be the desired spectrum of F, where $s_1, ..., s_a$ are the real, and $p_1 + jq_1, ..., p_b - jq_b$ — the complex conjugate eigenvalues of F, $a+2b=n$. Hence we are going to solve the problem: Given s_D find F and G such that (9) is valid, and and $s(F)=s_D$.

This is a useful modification of the problem considered in Sect. 3. Indeed, the determination of T and G from (3) and (5) for $F \in E_n$ yields a preliminary verification of the fact that $F \in E_n$. If F is chosen with distinct eigenvalues than it is cyclic and surely $F \in E_n$. If however the desired matrix F has multiple eigenvalues than according to the proposition it is necessary to check the condition (7) which requires the determination of the Kronecker indices of (C, A). If finally the matrix F is chosen in companion form than the inclusion $F \in E_n$ is again guaranteed. In this case however large coefficients of the characteristic polynomial (and hence large elements of F) may be obtained which shall deteriorate the solution.

If $r>1$ the above problem does not use the full freedom in the solution G of (9). That is why we shall consider only the case $r=1$. If $r>1$ then using a dyadic injection $G=gd$ ($g \in R^n$, $d^T \in R^r$) the problem is reduced to the case $r=1$.

The pair (A^T, C^T) can be transformed to the orthogonal canonical form $(\tilde{A}, \tilde{C})=$ $=(U^T A^T U, U^T C^T)$, where

$$\tilde{A}=\begin{bmatrix} & a_1 & & \\ \hdashline a_{21} & & a_2 & \\ \hdashline & \ddots & & \\ \hdashline 0 & & a_{n,n-1} & a_n \end{bmatrix}, \quad \tilde{C}=\begin{bmatrix} c_{10} \\ \hdashline 0 \end{bmatrix},$$

$a_i^T \in R^{n-i+1}$, $c_{10} \neq 0$; $a_{i,i-1} \neq 0$, $i \in \overline{2, n}$, and $U \in R^{n \cdot n}$ is an orthogonal matrix. This can be done using $n-1$ Householder reflections [6], [7]. The above reduction is numerically stable in the sense that the computed $(\tilde{A}, \tilde{C})$ is exact for a pair $(A^T+$ $+dA, C^T+dC)$, where

$$\|dA\| \leqq \text{eps}\,(6n^2+\text{const.}\, n)\,\|A\|,$$
$$\|dC\| \leqq \text{eps}\,(6n+\text{const})\,\|C\|,$$

and eps is the relative machine precision of the computer used.

Hence the problem is to find $\tilde{g}$, $\tilde{g}^T=U^T G \in R^n$, such that the matrix $\tilde{A}-\tilde{C}\tilde{g}$ have the prescribed spectrum s_D. Obviously the problem considered is equivalent to pole assignment of linear single-input systems.

The pole assignment algorithm which is further discussed is based on the following idea. Since the matrices $\tilde{A}$ and $\tilde{A}_c=\tilde{A}-\tilde{C}\tilde{g}$ are in Hessenberg form and differ only in their first rows it is possible, by setting a desired pole, to find en eigenvector of the matrix $\tilde{A}_c$ before computing $\tilde{g}$. Using sequences of plane rotations, belonging to the group of orthogonal transformations, all but the first elements of the eigenvector may annihilated. Then by necessity the first column of the transformed matrix $\tilde{A}_c$ will have zero elements below the first one which must be equal to the desired eigenvalue. This gives an equation for the first element of the transformed matrix $\tilde{g}$. The key observation here is that after the transformation the matrices $\tilde{A}$ and $\tilde{A}_c$ remain in Hessenberg which permits to work by the same way at the next step. At each step the algorithm works with a subsystem of decreasing order and the plane rotations are determined by the subsystem eigenvector. Since the subsystem matrices are in Hessenberg form their eigenvectors may be computed by solving triangular systems of linear equations.

The n elements of $\tilde{g}$ can be computed by the following algorithm:

Step 1. The eigenvector $\tilde{v}_1$ of $\tilde{A}_c$, corresponding to s_1, $\tilde{A}_c \tilde{v}_1 = \tilde{v}_1 s_1$, may be determinined from

$$T_1 \bar{v}_1 = \underline{v}_1 s_1 - h_1 , \tag{10}$$

where

$$\tilde{v}_1 = \begin{bmatrix} \bar{v}_1 \\ \hline 1 \end{bmatrix} = \begin{bmatrix} \times \\ \hline \underline{v}_1 \end{bmatrix}; \qquad \bar{v}_1,\ \underline{v}_1 \in R^{n-1},$$

and the matrix $\tilde{A}$ is partitioned as

$$\tilde{A} = \left[\begin{array}{c|c} \times \cdots & \times \\ \hline T_1 & h_1 \end{array}\right]$$

with $T_1 \in R^{(n-1)\cdot(n-1)}$ being a non-singular upper triangular matrix. The last element of $\tilde{v}_1$ is non-zero and hence is chosen equal to 1.

The linear triangular system of equations (10) may be solved by backward substitution. However the elements of the eigenvector may be computed simultaneously with the transformation of this eigenvector exploiting the fact that some of the previous elements are already annihilated. This reduces the number of the computational operations and improves the accuracy of the eigenvector.

After annihilating the elements from n to 2 of $\tilde{v}_1$ by plane rotations the transformed $\tilde{A}_c$ and $\tilde{C}$ are to be in the form

$$Q_1^T \tilde{A}_c Q_1 = \left[\begin{array}{c|c} s_1 & \times \cdots \times \\ \hline 0 & \tilde{A}_c^{(2)} \end{array}\right], \qquad Q_1^T \tilde{C} = \begin{bmatrix} \tilde{c}_1 \\ \tilde{c}_2 \\ 0 \\ \vdots \\ 0 \end{bmatrix},$$

where the $n-1$ plane rotations are accumulated in the matrix Q_1, and $\tilde{A}_c^{(2)} \in \in R^{(n-1)\cdot(n-1)}$ is a Hessenberg matrix. It follows from the complete observability of the pair (C, A) that $\tilde{c}_2 \neq 0$.

Now the element g_1, $\tilde{g}Q_1 = [g_1, \ldots, g_n]$, is determined from

$$\tilde{c}_1 g_1 = a_{11} - s_1 , \tag{11}$$

$$\tilde{c}_2 g_1 = a_{21} , \tag{12}$$

where $Q_1^T \tilde{A} Q_1 = [a_{ij}]$. The equations (11) and (12) are algebraically consistent but in some cases (11) may be zero identity. That is why it is reasonable to determine g_1 from

$$g_1 = (a_{11} - s_1)/\tilde{c}_1 , \quad \text{if} \quad |\tilde{c}_1| \geqq |\tilde{c}_2| ,$$

$$g_1 = a_{21}/\tilde{c}_2 , \quad \text{if} \quad |\tilde{c}_1| < |\tilde{c}_{2r}| .$$

In this way as a result of step 1 one element of the transformed matrix $\tilde{g}$ is obtained and the problem is reduced to a problem of dimension $n-1$.

Steps 2, ..., a. The next $a-1$ elements of $\tilde{g}$ are determined. Every eigenvector is obtained as a solution of a 3-diagonal system of linear equations and the number of necessary plane rotations decreases with 1 at each step.

Let $Q_2, ..., Q_a$ be the transformation matrices at steps 2, ..., a. Denote by $\tilde{A}^{(a+1)} \in R^{(n-a)\cdot(n-a)}$ and $\tilde{A}_c^{(a+1)} \in R^{(n-a)\cdot(n-a)}$ the lower right submatrices of the matrices $Q_a^T ... Q_1^T \tilde{A} Q_1 ... Q_a$ and $Q_a^T ... Q_1^T \tilde{A}_c Q_1 ... Q_a$ resp.

It is clear that using complex plane rotations the above technique may also be applied to determine the elements of g in the case of complex conjugate poles. However it is possible to solve the problem with slightly complicated technique using real arithmetic only. As a result the transformed matrix $\tilde{A}_c^{(a+1)}$ will have 2×2 blocks on its diagonal. This technique is described in following double step.

Steps a+1, a+2. The computation of the real x_1 and the imaginery y_1 parts of the complex eigenvectors x_1+jy_1, x_1-jy_1 of the matrix $\tilde{A}_c^{(a+1)}$, corresponding to the poles p_1+jq_1, p_1-jq_1, may be performed by the equations

$$T_{a+1}\,[\bar{x}_1, \bar{y}_1]=[\underline{x}_1, \underline{y}_1]\,S_1-[h_{a+1}, h_{a+1}],$$

where

$$x_1=\begin{bmatrix}\bar{x}_1\\ \hdashline 1\end{bmatrix}=\begin{bmatrix}\times\\ \hdashline \underline{x}_1\end{bmatrix}\in R^{n-a},$$

$$y_1=\begin{bmatrix}\bar{y}_1\\ \hdashline 1\end{bmatrix}=\begin{bmatrix}\times\\ \hdashline \underline{y}_1\end{bmatrix}\in R^{n-a};\quad S_1=\begin{bmatrix}p_1 & q_1\\ -q_1 & p_1\end{bmatrix}$$

and the matrix $\tilde{A}^{(a+1)}$ is partitioned as

$$\tilde{A}^{(a+1)}=\begin{bmatrix}\times\times\cdots\times & \\ \hdashline T_{a+1} & h_{a+1}\end{bmatrix}$$

with $T_{a+1}\in R^{(n-a-1)\cdot(n-a-1)}$ being non-singular upper triangular matrix.

Steps a+3, a+4, ..., n−1, n. These steps are performed in the same way. As a result all elements of $\tilde{g}$ are determined: $\tilde{g}=[g_1, ..., g_n]\,Q^T$ where $Q=Q_1 Q_2 ...$ $... Q_a Q_{a+1,a+2} ... Q_{n-1,n}$ and Q_{ij} is the transformation matrix at the double step i, j.

Finally one obtains $g=\tilde{g}U^T$.

4.2. Reduced Order Observers

Consider again the case $r=1$, $q=n-1$. We shall carry out the synthesis of a reduced $(n-1)$ th order observer under the assumption that the matrix $F\in E_{n-1}$ has a desired spectrum $\bar{s}_D$. Thus the problem is to determine the matrices $T \in R^{(n-1)\cdot n}$, $G \in R^{n-1}$ and $E \in R^{(n-1)\cdot(n-1)}$ such that

$$TA - FT = GC,$$
$$\text{rank}\begin{bmatrix} T \\ \hdashline C \end{bmatrix} = n, \tag{13}$$
$$s(F) = \bar{s}_D.$$

Since rank $T = n-1$ we shall take G in the form $G = T\tilde{G}$, where $\tilde{G} \in R^n$ is an unknown vector.

Using the algorithm from 4.1 we can determine the matrices $\tilde{g} \in R^{1\cdot n}$, $F_{11} \in R^{(n-1)\cdot(n-1)}$, $F_{12} \in R^{n-1}$ and $F_{22} \in R$ such that

$$(\tilde{A} \times \tilde{C}\tilde{g})\, Q = Q \begin{bmatrix} F_{11} & F_{12} \\ 0 & F_{22} \end{bmatrix},$$

where $s(F_{11}) = \bar{s}_D$, F_{22} is chosen so that $\tilde{g} = [g_1, ..., g_{n-1}, 0]\, Q^T$, and $Q \in R^{n\cdot n}$ is an orthogonal matrix.

Denoting $Q = [Q_1, Q_2]$, $Q_1 \in R^{n\cdot(n-1)}$ one obtains

$$Q_1^T \tilde{A} - F_{11}^T Q_1^T = Q_1^T \tilde{g}\tilde{C}. \tag{14}$$

The comparison of (13) and (14) gives the final solution in the form

$$T = Q_1^T, \quad G = Q_1^T \tilde{g}, \quad F = F_{11}^T.$$

It remains to show that the condition

$$\text{rank}\begin{bmatrix} T \\ \hdashline C \end{bmatrix} = n \tag{15}$$

from (13) is fulfilled.

Suppose that (15) is not valid. Then since rank $T = n-1$ it follows that $\tilde{C} = cT$, where $c^T \in R^{n-1}$. Having in mind that

$$\tilde{A}Q_1 - \tilde{C}\tilde{g}Q_1 = Q_1 F_{11}$$

one obtains

$$\text{rank}\,[\tilde{C}^T, \tilde{A}^T \tilde{C}^T, ..., (\tilde{A}^T)^{n-1} \tilde{C}^T] = \text{rank}\, Q_1\, [c^T, \bar{F}_{11} c^T, ..., (\bar{F}_{11}^T)^{n-1} c^T] \leqq n-1,$$

where $\bar{F}_{11} = F_{11} + c^T \tilde{g} Q_1$. This contradiction with the complete observability of (C, A) shows that (15) holds true.

It must be pointed out that the obtained matrix Q_1 (and hence the matrix T) depends only on the desired spectrum $\bar{s}_D$.

4.3. Numerical Considerations

The algorithm presented in 4.1 (and in slightly modified form in 4.2) has many common with the deflation techniques [8] used to eliminate a known eigenvalue from an eigenvalue problem. One of these techniques is of particular interest here.

If an approximated eigenvector is known it is possible to construct an orthogonal transformation in order to produce a matrix of order one less than the original matrix that does not contain the eigenvalue corresponding to the known eigenvector. This technique is very stable even if the approximated eigenvector is far from the accurate one.

The detailed numerical analysis shows that our algorithm has also very good numerical properties due to the fact that the computation of an eigenvector, its transformation and the determination of the matrix g corresponds to a small residual in the equation for this eigenvector. In this way the subdiagonal elements of the triangular form obtained are negligible which leads to the numerical stability of the algorithm.

It must be pointed out finally that in case of synthesis of reduced order observers the conditioning of the matrix $\begin{bmatrix} T \\ \hdashline C \end{bmatrix}$ depends only on the desired spectrum $\bar{s}_D$ of the system (6) of the dynamic error. In most cases however the matrix $\begin{bmatrix} T \\ \hdashline C \end{bmatrix}$ is well conditioned due to the fact that T contains $n-1$ rows of an orthogonal matrix.

An approximate operation count for the algorithm shows that about $4n^3$ floating point operations must be performed for the transformation of the system matrix. Adding to this figure the number of necessary operations for reducing the system into orthogonal canonical form one can find total of $6n^3$ operations. With respect to the array storage the algorithm requires $2n^2+6n$ working precision words.

5. Conclusions

An efficient computational algorithm for synthesis of full and reduced order state observers is proposed, based on orthogonal transformations only. The algorithm is numerically stable with respect to the determination of the observers' matrices and performs equally well with real and complex, distinct and multiple desired poles of the system of the dynamic error. It is applicable to ill-conditioned and high order problems and may be used for synthesis of state observers in continuous as well as in discrete systems. The algorithm is implemented as a FORTRAN program which is used for solving various problems of different orders.

References

[1] LUENBERGER D. G. Observers for linear multivariable systems. *IEEE Trans. Autom. Contr.*, **AC-11** (1966), 190–197.

[2] LUENBERGER D. G. An introduction to observers. *IEEE Trans. Autom. Contr.*, **AC-16** (1971), 596–602.

[3] KONSTANTINOV M. M., PETKOV P. HR., CHRISTOV N. D. On the general problem of synthesis of state observers. *Systems Sci.*, **6** (1980), 201–209.

[4] KONSTANTINOV M. M., PETKOV P. HR., CHRISTOV N. D. Synthesis of linear systems with desired equivalent form. *J. Comp. & Appl. Math.*, **6** (1980), 27–35.

[5] ROSENBROCK H. H. State space and multivariable theory. London, Nelson, 1970.
[6] KONSTANTINOV M. M., PETKOV P. HR., CHRISTOV N. D. Invariants and canonical forms for linear multivariable systems under the action of orthogonal transformation groups. *Kybernetika*, **17** (1981), 413–424.
[7] KONSTANTINOV M. M., PETKOV P. HR., CHRISTOV N. D. A Schur approach to pole assignment problem. Preprints 8th IFAC Congress, Kyoto, 1981, vol. **11**, 1–6.
[8] WILKINSON J. H. The algebraic eigenvalue problem. Oxford, Clarendon Press, 1965.

Received, October 1983.

Numeryczny stabilny algorytm syntezy obserwatorów stanu

W pracy proponuje się numerycznie stabilny algorytm syntezy obserwatorów stanu o pełnym i zredukowanym rzędzie. Najpierw sformułowano ogólne zadanie syntezy obserwatorów stanu i opisano w pełni zbiór wszystkich osiągalnych struktur obserwatorów stanu. Następnie przedstawiono metodę otrzymywania wszystkich rozwiązań tak postawionego zadania. Algorytm numeryczny oparty jest na ortogonalnej postaci kanonicznej liniowych układów niezmiennych w czasie i wymaga jedynie zastosowania ortogonalnych przekształceń równoważnościowych.

Численный устойчивый алгоритм синтеза наблюдателей состояния

В работе предлагается численно устойчивый алгоритм синтеза наблюдателей состояния полного и редуцированного порядка. Вначале формулируется общая задача синтеза наблюдателей состояния и описывается полностью множество всех достижимых структур наблюдателей состояния. Затем представлен метод получения всех решений так поставленной задачи. Численный алгоритм основан на ортогональном каноническом виде линейных систем, постоянных во времени и требует лишь применения ортогональных эквивалентных преобразований.