On the minimal stable observer problem

INT. J. CONTROL, 1980, VOL. 32, NO. 1, 17–30

On the minimal stable observer problem

J. M. SCHUMACHER†

We give a general formulation of the minimal stable observer problem, and show its equivalence to a minimization problem concerning certain pairs of invariant subspaces in the system state space. From this, we obtain a complete solution of the problem of finding a minimal stable observer that estimates the complete state.

1. Introduction

Since the original work of Luenberger (1964, 1966, 1971), observer theory has been a highly active field of research. One of the main points of interest is to improve on Luenberger's value for the minimal observer order ($n-p$, where n is the state-space dimension and p is the output-space dimension) by relaxing the observer requirements. For instance, one may require only that a certain given function of the state be estimated by the observer, instead of the state itself; or one may, rather than prescribe the poles that determine the error dynamics precisely (fixed-pole problem), require only that these poles should be situated in some pre-assigned part of the complex plane (stable observer problem). On the other hand, there have also been extensions; one of these consists of assuming that disturbances are present of which no measurement is available to the observer.

There is a wealth of papers on all of these subjects. Some of the recent contributions to the problem of finding a stable observer that estimates a function of the state are those by Sirisena (1979), Scott and Anderson (1978); one may see the references cited in these papers for the older work in the field. An attempt to capitalize only on the change from a fixed-pole to a stable observer problem, without dropping the requirement that the complete state should be estimated, can be found in Wonham (1979, § 3.7). The observer problem with unmeasured disturbances is one of the first problems in control theory to which the geometric approach has been applied; see Basile and Marro (1969). For a recent result in this field, see Bhattacharyya (1978). These references could be supplemented by many more, but we do not aim here to give an overview of the large field of observer theory.

So, even within the limits of the geometric approach, a whole variety of observer problems exists that are attacked by different methods. However, the distinct observer problems are not so sharply separated in practical situations, because usually there are modelling alternatives. For instance, one may take a certain set of variables into the model as a disturbance, or, alternatively, one may try to set up dynamic equations for these variables and incorporate them into the state. What is the complete state in a certain model may be only a function of the state in an enlarged model. One may have a certain freedom in carrying out measurements, and in this case it is of interest to know what can

Received 19 December 1979.

† Mathematics Department, Vrije Universiteit, De Boelelaan 1081, 1081 HV Amsterdam, The Netherlands.

0020-7179/80/3201 0017 $02-00 © 1980 Taylor & Francis Ltd

be gained—in terms of observer order, or observer fastness—from installing extra measuring devices.

The state of affairs calls for a unified framework for observer problems ; this will enable one to effectively compare different modelling alternatives. Such a unified framework is also desirable from a theoretical point of view, since the relations between the different observer problems are not all cleared up. To give a simple example of this, consider a stable and undisturbed system of which the full state is to be estimated. It requires very little theory to see that a stable observer for this system is one which at all time gives zero as an estimate of the state, and this obviously requires no observer dynamics. Nevertheless, the theory of estimating a function of the state (as presented for instance in Wonham and Morse (1972)) always gives $n-p$ as the observer order (where n is the number of state variables and p is the number of measurements), when specialized to the case of the identity function (which corresponds to estimating the complete state). The problem of constructing a stable full state observer is considered in Wonham (1979), as said before, and though Wonham's method allows him to improve on the value $n-p$, it does not solve the ' minimal detector problem ' (as he calls it) effectively.

The present paper aims to make further steps towards the creation of a unified framework as referred to above. It has two contributions : first, it is shown that all observer properties can be formulated in terms of properties of certain pairs of invariant subspaces of the system state space ; then, this formulation is used to give a complete solution of the minimal detector problem. It is not attempted here to solve the general minimization problem, although we do obtain an upper bound which may be useful in many cases ; neither does the present paper concern itself with robustness questions.

Pairs of invariant subspaces have been used before by the author in work on disturbance decoupling and on compensator synthesis (Schumacher 1979 a, b). In fact, there are some results in this article that are actually specializations of theorems in the earlier papers ; still, for reasons of self-containment and because some simplifications were possible, the proofs of these results have been included.

In § 2, we formulate the general minimal stable observer problem. The equivalence of this problem to a minimization problem concerning pairs of invariant subspaces of the original state space is proved in §§ 3 and 4 ; § 3 derives necessary conditions for a pair to correspond to a stable observer, and § 4 provides the counterpart to this by constructing a stable observer from a suitable pair. In § 5, we use the theory to solve the minimal detector problem.

2. Problem formulation

We consider a linear, finite-dimensional, deterministic system given by the following equations

$$\left.\begin{aligned} \dot{x}(t) &= Ax(t) + Bu(t) \\ y(t) &= Cx(t) \\ r(t) &= Hu(t) \\ z(t) &= Dx(t) \end{aligned}\right\} \tag{1}$$

Here, $x(t) \in \mathscr{X}$ is the *state* of the system, $u(t) \in \mathscr{U}$ is the *disturbance*, $y(t) \in \mathscr{Y}$ is the *observation of the state*, $r(t) \in \mathscr{R}$ is the *observation of the disturbance*, and $z(t) \in \mathscr{Z}$

is the *variable to be estimated.* To this system, we want to add an *observer*, that is, a new dynamical system of the following form

$$\left.\begin{aligned}\dot{w}(t) &= Nw(t)+My(t)+Lr(t)\\ \hat{z}(t) &= Sw(t)+Ty(t)\end{aligned}\right\}\tag{2}$$

The *order* of the observer is the dimension of the space $\mathscr{W}$ in which $w(t)$ runs; $\hat{z}(t)\in\mathscr{Z}$ is the *estimate* produced by the observer. A *Kalman observer* is one for which $T=0$ (so there is no direct feedthrough from the observation of the state to the output). We shall refer to the general observer of the form (2) as a *Luenberger observer*. The equations (1) and (2) give rise to a composite system which has $\mathscr{X}\oplus\mathscr{W}$ for its state space

$$\left.\begin{aligned}\frac{d}{dt}\begin{pmatrix}x\\ w\end{pmatrix}(t) &= \begin{pmatrix}A & O\\ MC & N\end{pmatrix}\begin{pmatrix}x\\ w\end{pmatrix}(t)+\begin{pmatrix}B\\ LH\end{pmatrix}u(t)\\ e(t)=z(t)-\hat{z}(t) &= (D-TC \quad -S)\begin{pmatrix}x\\ w\end{pmatrix}(t)\end{aligned}\right\}\tag{3}$$

Here, we have introduced the *error* $e(t)=z(t)-\hat{z}(t)\in\mathscr{Z}$. We shall write the *composite system matrix* as A_c, so

$$A_c=\begin{pmatrix}A & O\\ MC & N\end{pmatrix}\tag{4}$$

An important subspace of $\mathscr{X}\oplus\mathscr{W}$ which will be used very often in the sequel is the largest A_c-invariant subspace of Ker $(D-TC \quad -S)$; we shall denote it by $\mathscr{V}$ (which is fixed for a given observer). This subspace contains the initial values $\begin{pmatrix}x(0)\\ w(0)\end{pmatrix}$ from which the composite system will move in such a way that the error is zero for all time, so we shall refer to it as the observer's *errorless subspace*.

$\mathscr{V}$ is naturally related to two subspaces of the original state space, which will be denoted by $\mathscr{V}_i$ and $\mathscr{V}_p$ (i for intersection, p for projection):

$$\mathscr{V}_i=\left\{x\in\mathscr{X}\;\middle|\;\begin{pmatrix}x\\ 0\end{pmatrix}\in\mathscr{V}\right\}\tag{5}$$

$$\mathscr{V}_p=\left\{x\in\mathscr{X}\;\middle|\;\exists w\in\mathscr{W}:\;\begin{pmatrix}x\\ w\end{pmatrix}\in\mathscr{V}\right\}\tag{6}$$

It is the pair $(\mathscr{V}_i, \mathscr{V}_p)$ that will be our main object of study in this article.

We now want to define when an observer of the form (2) will be called *stable*. For this goal, let us first fix some (symmetric) part of the complex plane that we want to associate with the word 'stable'; write $\mathbb{C}_g$ for this part of $\mathbb{C}$, and $\mathbb{C}_b$ for its complement $\mathbb{C}\backslash\mathbb{C}_g$. (Of course, the open left half-plane is the usual choice for $\mathbb{C}_g$, but making any other choice will not change the theory below. So we allow ourselves this generality while retaining the use of the short phrase 'stable'. It may be noted, at this point, that a 'fixed-pole' problem is obtained by making $\mathbb{C}_g$ sufficiently small.) For an operator A acting on a linear space $\mathscr{X}$, we write $\mathscr{X}_g(A)$ for the subspace of $\mathscr{X}$ spanned by the (generalized) eigenvectors of A associated with eigenvalues in $\mathbb{C}_g$, and define $\mathscr{X}_b(A)$ analogously. We recall that, if $\mathscr{V}$ is an A-invariant subspace of $\mathscr{X}$, the operator $\bar{A}$ induced by A on $\mathscr{X}/\mathscr{V}$ is stable if and only if $\mathscr{X}_b(A) \subset \mathscr{V}$. For more elaborate information on these points, one should consult Wonham (1979, § 2.3, § 4.4).

Now let us formulate the stability requirements. Because we do not want to make any assumptions about the behaviour of $u(t)$ (except for the usual weak regularity hypotheses, for instance piecewise continuity), the disturbance must be decoupled from the error. This means

$$\left\langle A_c \,\middle|\, \operatorname{Ran}\begin{pmatrix} B \\ LH \end{pmatrix} \right\rangle \subset \operatorname{Ker}\,(D - TC \quad -S) \tag{7}$$

Of course, this is equivalent to

$$\left\langle A_c \,\middle|\, \operatorname{Ran}\begin{pmatrix} B \\ LH \end{pmatrix} \right\rangle \subset \mathscr{V} \tag{8}$$

Then, the exponents appearing in the dynamics of the error $e(t)$ will be precisely the points in the spectrum of the operator $\overline{A_c}$ induced by A_c on the factor space $(\mathscr{X} \oplus \mathscr{W})/\mathscr{V}$. So, as our second stability requirement we get

$$\sigma(\overline{A_c}) \subset \mathbb{C}_g. \tag{9}$$

which is the same as

$$(\mathscr{X} \oplus \mathscr{W})_b(A_c) \subset \mathscr{V} \tag{10}$$

So we formulate the general minimal stable observer problem as follows. Given a system (1), find an observer of lowest possible order such that its errorless subspace $\mathscr{V}$ contains both $\left\langle A_c \,\middle|\, \operatorname{Ran}\begin{pmatrix} B \\ LH \end{pmatrix} \right\rangle$ and $(\mathscr{X} \oplus \mathscr{W})_b(A_c)$.

3. Necessary conditions

Our aim in this section is to collect some properties which are necessary for a system of the form (2) to constitute an observer for the system (1). Since we are interested in minimal observers, we shall only consider observers that are *observable* in the sense that there is no N-invariant subspace in Ker S other than $\mathscr{O}$. (If Ker S does have a non-trivial N-invariant subspace, then this subspace can be factored out to produce an observer of lower order satisfying the same stability requirements as the original one; so minimal observers must be observable.)

We shall write $\mathscr{B} = \operatorname{Ran} B$ and $\mathscr{B}_0 = B\,[\operatorname{Ker} H]$. A subspace $\mathscr{T}$ of $\mathscr{X}$ is called (C, A)-invariant if there exists G such that $(A + GC)\mathscr{T} \subset \mathscr{T}$, or, alternatively, if $A(\mathscr{T} \cap \operatorname{Ker} C) \subset \mathscr{T}$.

Theorem 1

If the system (2) is an observable observer for the system (1), and if $\mathscr{V}$ is its errorless subspace, then the following conditions hold :

(i) $\mathscr{V}_{\mathrm{i}}$ is (C, A)-invariant and $\mathscr{V}_{\mathrm{p}}$ is A-invariant ;

(ii) $\mathscr{V}_{\mathrm{i}} \cap \operatorname{Ker} C \subset \operatorname{Ker} D$ (for a Luenberger observer) ;

(ii)′ $\mathscr{V}_{\mathrm{i}} \subset \operatorname{Ker} D$ (for a Kalman observer) ;

(iii) $\mathscr{V}_{\mathrm{i}} \supset \mathscr{B}_0$ and $\mathscr{V}_{\mathrm{p}} \supset \mathscr{B}$;

(iv) $\dim \mathscr{V}_{\mathrm{p}} = \dim \mathscr{V}$;

(v) there exists $G : \mathscr{Y} \to \mathscr{X}$ such that $(A+GC)\mathscr{V}_{\mathrm{i}} \subset \mathscr{V}_{\mathrm{i}}$, Ran $G \subset \mathscr{V}_{\mathrm{p}}$ and $\sigma(\overline{A+GC}) \subset \mathbb{C}_{\mathrm{g}}$, where $\overline{A+GC}$ is the induced operator on $\mathscr{X}/\mathscr{V}_{\mathrm{i}}$.

Proof of (i)

We know that $\mathscr{V}$ is $\begin{pmatrix} A & O \\ MC & N \end{pmatrix}$-invariant. Take $x \in \mathscr{V}_{\mathrm{i}} \cap \operatorname{Ker} C$; then

$$\begin{pmatrix} x \\ 0 \end{pmatrix} \in \mathscr{V} \quad \text{and} \quad \begin{pmatrix} Ax \\ 0 \end{pmatrix} = \begin{pmatrix} A & O \\ MC & N \end{pmatrix} \begin{pmatrix} x \\ 0 \end{pmatrix} \in \mathscr{V}$$

so $Ax \in \mathscr{V}_{\mathrm{i}}$. Next take $x \in \mathscr{V}_{\mathrm{p}}$; then there exists $w \in \mathscr{W}$ such that $\begin{pmatrix} x \\ w \end{pmatrix} \in \mathscr{V}$, and so

$$\begin{pmatrix} Ax \\ MCx + Nw \end{pmatrix} = \begin{pmatrix} A & O \\ MC & N \end{pmatrix} \begin{pmatrix} x \\ w \end{pmatrix} \in \mathscr{V}$$

which proves $Ax \in \mathscr{V}_{\mathrm{p}}$. □

Proof of (ii)–(ii)′

From $\mathscr{V} \subset \operatorname{Ker}(D - TC \quad -S)$ it follows that $\mathscr{V}_{\mathrm{i}} \subset \operatorname{Ker}(D - TC)$. So if $T = 0$ (Kalman observer), then $\mathscr{V}_{\mathrm{i}} \subset \operatorname{Ker} D$. In the general case, $Dx = (D - TC)x = 0$ for all $x \in \mathscr{V}_{\mathrm{i}} \cap \operatorname{Ker} C$, so $\mathscr{V}_{\mathrm{i}} \cap \operatorname{Ker} C \subset \operatorname{Ker} D$. □

Proof of (iii)

We know that $\mathscr{V}$ must contain Ran $\begin{pmatrix} B \\ LH \end{pmatrix}$; so for all $u \in \mathscr{U}$,

$$\begin{pmatrix} Bu \\ LHu \end{pmatrix} \in \mathscr{V}$$

We get immediately that $\mathscr{B} \subset \mathscr{V}_{\mathrm{p}}$ and $\mathscr{B}_0 \subset \mathscr{V}_{\mathrm{i}}$. □

Proof of (iv)

Write

$$\mathscr{W}_0 = \left\{ w \in \mathscr{W} \;\middle|\; \begin{pmatrix} 0 \\ w \end{pmatrix} \in \mathscr{V} \right\}$$

Clearly, $\dim \mathscr{V} = \dim \mathscr{V}_{\mathrm{p}} + \dim \mathscr{W}_0$, so we have to prove that $\mathscr{W}_0 = \mathcal{O}$. To this purpose, note that $w \in \mathscr{W}_0$ implies

$$\begin{pmatrix} 0 \\ Nw \end{pmatrix} = \begin{pmatrix} A & O \\ MC & N \end{pmatrix} \begin{pmatrix} 0 \\ w \end{pmatrix} \in \mathscr{V}$$

so $\mathscr{W}_0$ is N-invariant ; moreover,

$$(D - TC \quad -S) \begin{pmatrix} 0 \\ w \end{pmatrix} = 0$$

for $w \in \mathscr{W}_0$, so $\mathscr{W}_0 \subset \operatorname{Ker} S$. The result now follows from the assumption of observability. □

The proof of (v) is considerably harder than the foregoing ones, and we insert a few lemmas which will be needed. First, we derive a simple but useful formula for the error poles of an observable observer.

Lemma 1

If (2) constitutes an observable observer for (1), then the following relation (in which $\bar{A}$ is the operator induced by A on the factor space $\mathscr{X}/\mathscr{V}_{\mathrm{p}}$) holds for the error poles :

$$\sigma(\overline{A_{\mathrm{c}}}) = \sigma(\bar{A}) \cup \sigma(N) \tag{11}$$

Proof

Let us write P for the natural projection of $\mathscr{V}$ onto $\mathscr{V}_{\mathrm{p}}$. We have just seen that $\dim \mathscr{V} = \dim \mathscr{V}_{\mathrm{p}}$, so P is an isomorphism. Moreover, the following diagram commutes :

$$\begin{tikzcd} \mathscr{V} \arrow[r, "A_{\mathrm{c}}"] \arrow[d, "P"'] & \mathscr{V} \arrow[d, "P"] \\ \mathscr{V}_{\mathrm{p}} \arrow[r, "A"] & \mathscr{V}_{\mathrm{p}} \end{tikzcd} \tag{12}$$

So the restrictions of A to $\mathscr{V}_{\mathrm{p}}$ and of A_{c} to $\mathscr{V}$ are similar, which implies

$$\sigma(A_{\mathrm{c}} | \mathscr{V}) = \sigma(A | \mathscr{V}_{\mathrm{p}}) \tag{13}$$

On the other hand, we have from the block-triangular form of the matrix A_{c}

$$\sigma(A_{\mathrm{c}}) = \sigma(A) \cup \sigma(N) \tag{14}$$

Using the formula $\sigma(A) = \sigma(\bar{A}) \cup \sigma(A | \mathscr{V}_{\mathrm{p}})$, we now obtain (11) by ' subtracting ' (13) from (14). □

Our second lemma is essentially a standard result in pole placement, but formulated in a somewhat unusual way.

Lemma 2

Suppose $\mathscr{T}$ is a subspace of a linear space $\mathscr{X}$, and A_0 is a linear mapping from $\mathscr{T}$ to $\mathscr{X}$. We can extend A_0 to a stable mapping A defined on all of $\mathscr{X}$ if and only if $A_0 x = \lambda x$ for $x \in \mathscr{T}$, $\lambda \in \mathbb{C}_{\mathrm{b}}$ implies $x = 0$.

Proof

The 'only if' part is trivial, so let us assume that the hypothesis holds. Extend A_0 to a mapping $\tilde{A}: \mathscr{X} \to \mathscr{X}$ in an arbitrary way. Introduce a new space $\mathscr{Y}$ and a mapping $C: \mathscr{X} \to \mathscr{Y}$ such that $\mathscr{T} = \mathrm{Ker}\ C$. The hypothesis then implies that the restriction of $\tilde{A}$ to the largest $\tilde{A}$-invariant subspace of Ker C is stable. This means that the pair $(C, \tilde{A})$ is detectable; so there exists $G: \mathscr{Y} \to \mathscr{X}$ such that $\tilde{A} + GC$ is stable. (Cf. Wonham 1979, § 3.6.) Now just take $A = \tilde{A} + GC$. □

We use this lemma in the proof of the following one.

Lemma 3

Suppose we have a linear space $\mathscr{X}$, linear mappings $A: \mathscr{X} \to \mathscr{X}$ and $C: \mathscr{X} \to \mathscr{Y}$, and a (C, A)-invariant subspace $\mathscr{T}$ of $\mathscr{X}$. Suppose further that $R: \mathscr{X} \to \mathscr{R}$ is a surjection with $\mathrm{Ker}\ R = \mathscr{T}$. Then there exists G with $(A + GC)\mathscr{T} \subset \mathscr{T}$ such that the induced operator $\overline{A + GC}$ on $\mathscr{X}/\mathscr{T}$ is stable if and only if $RAx = \lambda Rx$ for $x \in \mathrm{Ker}\ C$, $\lambda \in \mathbb{C}_b$ implies $Rx = 0$.

Proof

'if': Write $\mathscr{R}_0 = R\,[\mathrm{Ker}\ C]$. Introduce a mapping $N_0: \mathscr{R}_0 \to \mathscr{R}$ by

$$N_0 Rx = RAx \quad (x \in \mathrm{Ker}\ C) \tag{15}$$

(This definition is correct because $Rx = 0$ implies $RAx = 0$ for $x \in \mathrm{Ker}\ C$, Ker R being (C, A)-invariant.) By the hypothesis and Lemma 2, we can extend N_0 to a stable mapping N defined on all of $\mathscr{R}$. Because $Cx = 0$ implies $(NR - RA)x = 0$, there exists $M: \mathscr{Y} \to \mathscr{R}$ such that $MC = NR - RA$. If we now let S be any mapping from $\mathscr{R}$ to $\mathscr{X}$ such that $RS = I$, and define $G = SM$, then $NR = R(A + GC)$ and the following diagram commutes

$$\begin{array}{ccc} \mathscr{X} & \xrightarrow{A + GC} & \mathscr{X} \\ R \downarrow & N & \downarrow R \\ \mathscr{R} & \longrightarrow & \mathscr{R} \end{array} \tag{16}$$

From this, we see that $\mathrm{Ker}\ R = \mathscr{T}$ is $(A + GC)$-invariant and $\sigma(\overline{A + GC}) = \sigma(N) \subset \mathbb{C}_g$.

'only if': We now assume that there exists G with $(A + GC)\mathscr{T} \subset \mathscr{T}$ and $\sigma(\overline{A + GC}) \subset \mathbb{C}_g$. If $RAx = \lambda Rx$ for $x \in \mathrm{Ker}\ C$, then also $R(A + GC)x = \lambda Rx$ so that $\lambda \in \sigma(\overline{A + GC}) \subset \mathbb{C}_g$ or $Rx = 0$. □

We are now finally ready to complete the proof of the theorem.

Proof of (v)

We have proved in (iv) that $\dim \mathscr{V}_p = \dim \mathscr{V}$, so a (unique) mapping $R: \mathscr{V}_p \to \mathscr{W}$ exists such that

$$\mathscr{V} = \left\{ \begin{pmatrix} x \\ Rx \end{pmatrix} \middle| \, x \in \mathscr{V}_p \right\} \tag{17}$$

Note that Ker $R = \mathscr{V}_{\mathrm{i}}$. Because $\mathscr{V}$ is A_e-invariant,

$$\begin{pmatrix} A & O \\ MC & N \end{pmatrix}\begin{pmatrix} x \\ Rx \end{pmatrix} = \begin{pmatrix} Ax \\ (MC+NR)x \end{pmatrix} \in \mathscr{V}$$

for all $x \in \mathscr{V}_{\mathrm{p}}$; in particular, it follows that

$$(MC+NR)x = RAx \quad \text{for all } x \in \mathscr{V}_{\mathrm{p}} \tag{18}$$

Thus if $RAx = \lambda Rx$ for some $x \in \mathscr{V}_{\mathrm{p}} \cap \operatorname{Ker} C$, and $Rx \neq 0$, then $NRx = \lambda Rx$ so that $\begin{pmatrix} 0 \\ Rx \end{pmatrix} \notin \mathscr{V}$ is an eigenvector of A_e with the eigenvalue λ, which implies $\lambda \in \mathbb{C}_{\mathrm{g}}$.

We now want to apply Lemma 3, with $\mathscr{V}_{\mathrm{p}}$ playing the role of $\mathscr{X}$ in the formulation of the lemma. So we consider R as a mapping from $\mathscr{V}_{\mathrm{p}}$ to Ran R, and we let A_0 and C_0 denote the restrictions of A and C, respectively, to $\mathscr{V}_{\mathrm{p}}$. Then Ker $C_0 = \mathscr{V}_{\mathrm{p}} \cap \operatorname{Ker} C$ and Ker R is (C_0, A_0)-invariant ; moreover, as we just have seen, $RA_0x = \lambda Rx$ for $x \in \operatorname{Ker} C_0$, $\lambda \in \mathbb{C}_{\mathrm{b}}$ implies $Rx = 0$. So we can indeed apply the lemma and obtain a mapping $G_0 : C\mathscr{V}_{\mathrm{p}} \to \mathscr{V}_{\mathrm{p}}$ such that $(A_0 + G_0C_0)\mathscr{V}_{\mathrm{i}} \subset \mathscr{V}_{\mathrm{i}}$ and $\sigma(\overline{A_0 + G_0C_0}) \subset \mathbb{C}_{\mathrm{g}}$ (where $\overline{A_0 + G_0C_0}$ is the induced operator on $\mathscr{V}_{\mathrm{p}}/\mathscr{V}_{\mathrm{i}}$).

Now define G on $\mathscr{Y}$ as the extension of G_0 by 0. Then $\mathscr{V}_{\mathrm{i}}$ is $(A+GC)$-invariant, and Ran $G \subset \mathscr{V}_{\mathrm{p}}$. Moreover, we have

$$\sigma(\overline{A+GC}) = \sigma(\bar{A}) \cup \sigma(\overline{A_0 + G_0C_0}) \tag{19}$$

($\overline{A+GC}$ on $\mathscr{X}/\mathscr{V}_{\mathrm{i}}$, $\bar{A}$ on $\mathscr{X}/\mathscr{V}_{\mathrm{p}}$). Since $\sigma(\bar{A}) \subset \mathbb{C}_{\mathrm{g}}$ by Lemma 1, we are finished. □

4. Observer construction

Roughly speaking, what we want to do in this section is to show that the necessary conditions of Theorem 1 are also sufficient. That is, we shall construct an observer when a pair of subspaces $(\mathscr{V}_1, \mathscr{V}_2)$ in $\mathscr{X}$ is given, $\mathscr{V}_1$ having the properties of $\mathscr{V}_{\mathrm{i}}$, and $\mathscr{V}_2$ those of $\mathscr{V}_{\mathrm{p}}$. This will lead to an algebraic formulation, in terms of the original state-space, for the minimal stable observer problem.

Theorem 2

Suppose we have, given the system (1), and a pair of subspaces of $\mathscr{X}$, $(\mathscr{V}_1, \mathscr{V}_2)$ having the following properties

(i) $\mathscr{V}_1 \subset \mathscr{V}_2$, $\mathscr{V}_1$ is (C, A)-invariant and $\mathscr{V}_2$ is A-invariant ;
(ii) $\mathscr{V}_1 \cap \operatorname{Ker} C \subset \operatorname{Ker} D$, or
(ii)′ $\mathscr{V}_1 \subset \operatorname{Ker} D$;
(iii) $\mathscr{V}_1 \supset \mathscr{B}_0$ and $\mathscr{V}_2 \supset \mathscr{B}$;
(iv) there exists $G : \mathscr{Y} \to \mathscr{X}$ such that $(A+GC)\mathscr{V}_1 \subset \mathscr{V}_1$, Ran $G \subset \mathscr{V}_2$ and $\sigma(\overline{A+GC}) \subset \mathbb{C}_{\mathrm{g}}$, where $\overline{A+GC}$ is the induced operator on $\mathscr{X}/\mathscr{V}_1$.

Then we can construct an observer for (1) (a Kalman observer, if (ii)′ holds) of order dim $\mathscr{V}_2 -$ dim $\mathscr{V}_1$.

Proof

The observer is constructed as follows. Let $\mathscr{W}$ be any linear space of dimension $\dim \mathscr{V}_2 - \dim \mathscr{V}_1$, and let R be a linear mapping from $\mathscr{V}_2$ onto $\mathscr{W}$ with $\operatorname{Ker} R = \mathscr{V}_1$. Write

$$\tilde{\mathscr{V}} = \left\{ \begin{pmatrix} x \\ Rx \end{pmatrix} \middle| x \in \mathscr{V}_2 \right\}$$

Let $G : \mathscr{Y} \to \mathscr{X}$ be a mapping that satisfies the requirement of (iv), take

$$M = -RG \tag{20}$$

and define N by

$$NRx = R(A + GC)x \quad \text{for all } x \in \mathscr{V}_2 \tag{21}$$

For T, we do the following construction. Let $\{v_1, \ldots, v_r\}$ be a basis of $\mathscr{V}_1$ such that $\{v_1, \ldots, v_s\}$ is a basis for $\mathscr{V}_1 \cap \operatorname{Ker} C$. Because the elements $Cv_{s+1}, \ldots, Cv_r$ of Ker C are linearly independent, we can choose $T : \mathscr{Y} \to \mathscr{X}$ such that $TCv_i = Dv_i$ $(i = s+1, \ldots, r)$. (Note that this relation holds automatically for $i = 1, \ldots, s$, since $TCv_i = 0 = Dv_i$ in this case ; so we get $\mathscr{V}_1 \subset \operatorname{Ker}(D - TC)$.) When (ii)′ is given, this construction procedure allows us to take $T = 0$, so that a Kalman observer is obtained. We now have $\operatorname{Ker} R = \mathscr{V}_1 \subset \operatorname{Ker}(D - TC)$. Therefore, we can take S such that

$$SRx = (D - TC)x \quad \text{for all } x \in \mathscr{V}_2 \tag{22}$$

Finally, because $B[\operatorname{Ker} H] \subset \operatorname{Ker} R$ (or, put in another way, $\operatorname{Ker} H \subset \operatorname{Ker} RB$), L may be chosen such that

$$LH = RB \tag{23}$$

We have specified all the elements of the system (2), and now we must see whether the so-constructed system is indeed an observer. It follows from (20) and (21) that

$$(MC + NR)x = RAx \quad \text{for all } x \in \mathscr{V}_2 \tag{24}$$

so that

$$\begin{pmatrix} A & O \\ MC & N \end{pmatrix} \begin{pmatrix} x \\ Rx \end{pmatrix} = \begin{pmatrix} Ax \\ RAx \end{pmatrix} \quad \text{for all } x \in \mathscr{V}_2 \tag{25}$$

So we see that $\tilde{\mathscr{V}}$ is A_e-invariant. It follows immediately from (22) that $\tilde{\mathscr{V}} \subset \operatorname{Ker}(D - TC \quad -S)$, and from (23) that $\operatorname{Ran} \begin{pmatrix} B \\ LH \end{pmatrix} \subset \tilde{\mathscr{V}}$.

Let $\mathscr{V}$ be the largest A_e-invariant subspace of $\operatorname{Ker}(D - TC \quad -S)$; then it follows that $\tilde{\mathscr{V}} \subset \mathscr{V}$ and therefore also $\operatorname{Ran} \begin{pmatrix} B \\ LH \end{pmatrix} \subset \mathscr{V}$. So this satisfies the first stability requirement (8). For the second stability requirement (9), we note first of all that the spectrum of the operator induced by A_e on $\mathscr{X}/\mathscr{V}$ is part of the spectrum of the operator induced by A_e on $\mathscr{X}/\tilde{\mathscr{V}}$; so it suffices to prove that the latter (which we shall write as $\sigma(\overline{A}_e)$) is in $\mathbb{C}_g$. It is seen from the proof of Lemma 1 that the formula (11) holds in this situation. From the facts that Ran $G \subset \mathscr{V}_2$ and that $\mathscr{V}_2$ is A-invariant it follows that $\mathscr{V}_2$ is $(A+GC)$-invariant ; moreover, we have the formula

$$\sigma(\overline{A+GC}) = \sigma(\overline{A+GC|\mathscr{V}_2}) \cup \sigma(\bar{A}) \tag{26}$$

But it follows from (21) that

$$\sigma(N) = \sigma(\overline{A+GC|\mathscr{V}_2}) \tag{27}$$

Taking this in connection with (11), we get

$$\sigma(\overline{A_0}) = \sigma(\overline{A+GC}) \tag{28}$$

Because $\sigma(\overline{A+GC}) \subset \mathbb{C}_g$, the proof is finished. □

As a corollary, we obtain the following formulation for the minimal stable observer problem.

Theorem 3

The minimal stable observer problem is equivalent to the problem of finding, among all pairs of subspaces $(\mathscr{V}_1, \mathscr{V}_2)$ satisfying the requirements (i)–(iv) given in Theorem 2, one for which the value of $\dim \mathscr{V}_2 - \dim \mathscr{V}_1$ is minimal.

Proof

Let us call a pair $(\mathscr{V}_1, \mathscr{V}_2)$ with the required properties a 'minimal pair' if $\dim \mathscr{V}_2 - \dim \mathscr{V}_1$ is minimal. Suppose we have such a minimal pair $(\mathscr{V}_1, \mathscr{V}_2)$, then we can use it for the construction of an observer, by the method of the proof of Theorem 2. We want to show that this observer will be minimal. Suppose, to the contrary, that the minimal observer order is less than $\dim \mathscr{V}_2 - \dim \mathscr{V}_1$. Let $\mathscr{V}$ be the errorless subspace of a minimal observer ; then we get

$$\begin{aligned}\dim \mathscr{V}_2 - \dim \mathscr{V}_1 &\leqslant \dim \mathscr{V}_p - \dim \mathscr{V}_i = \dim \mathscr{V} - \dim \mathscr{V}_i \\ &= \dim \left\{ w \in \mathscr{W} \,\middle|\, \exists x \in \mathscr{X} : \begin{pmatrix} x \\ w \end{pmatrix} \in \mathscr{V} \right\} \\ &\leqslant \dim \mathscr{W} < \dim \mathscr{V}_2 - \dim \mathscr{V}_1, \quad \text{a contradiction}\end{aligned}$$

On the other hand, if we have a minimal observer with its associated errorless subspace $\mathscr{V}$, then the pair $(\mathscr{V}_i, \mathscr{V}_p)$ must be minimal. We prove this again by contradiction. If $(\mathscr{V}_i, \mathscr{V}_p)$ is not minimal, let $(\mathscr{V}_1, \mathscr{V}_2)$ be a pair that is. As above, we get $\dim \mathscr{W} \geqslant \dim \mathscr{V}_p - \dim \mathscr{V}_i > \dim \mathscr{V}_2 - \dim \mathscr{V}_1$. But by Theorem 2, we know that there exist observers of order $\dim \mathscr{V}_2 - \dim \mathscr{V}_1$; so the original one could not be minimal. □

Let us point out there is a procedure available to find out, given a pair $(\mathscr{V}_1, \mathscr{V}_2)$ (with $\mathscr{V}_1 \subset \mathscr{V}_2$, $\mathscr{V}_1$ (C, A)-invariant and $\mathscr{V}_2$ A-invariant), whether condition (iv) of Theorem 2 holds for this pair ; also, if condition (iv) holds, then there is a constructive way to find a suitable G. We can see this as follows. From (26), it is clear that we must have $\mathscr{X}_b(A) \subset \mathscr{V}_2$ for condition (iv) to hold. If this is checked, then we consider the restrictions A_0 and C_0 of A and C to $\mathscr{V}_2$, and forget about everything outside $\mathscr{V}_2$. Thus, $\mathscr{V}_1$ is considered as a (C_0, A_0)-invariant subspace of $\mathscr{V}_2$, and we ask ourselves whether there exists G_0 such that $(A_0+G_0C_0)\mathscr{V}_1 \subset \mathscr{V}_1$ and $\sigma(\overline{A_0+G_0C_0}) \subset \mathbb{C}_g$. (If such a G_0 is found, then a suitable G can be obtained from it by extending G_0 by 0 to a mapping defined on all of $\mathscr{Y}$.) By dualization, the question is of the following type: given a linear space $\mathscr{X}$ and an (A, B)-invariant subspace $\mathscr{V}$ of $\mathscr{X}$, does there exist a mapping F such that $(A+BF)\mathscr{V} \subset \mathscr{V}$ and $\sigma(A=BF|\mathscr{V}) \subset \mathbb{C}_g$? A solution procedure for this question, and a construction for F if it exists, is given in Wonham (1979, § 5.4).

One does the following. Take any F_1 such that $(A+BF_1)\mathscr{V} \subset \mathscr{V}$. A necessary and sufficient condition for the existence of an F as required is now that the induced mapping $\overline{A+BF_1|\mathscr{V}}$ on $\mathscr{V}/\langle A+BF_1|\mathscr{B}\cap\mathscr{V}\rangle$ is stable. If this condition holds, we may find (by a standard pole placement procedure) a mapping $F_2 : \langle A+BF_1|\mathscr{B}\cap\mathscr{V}\rangle \to \{u\in\mathscr{U}|Bu\in\mathscr{B}\cap\mathscr{V}\}$ such that the mapping $A+B(F_1+F_2)$ on $\langle A+BF_1|\mathscr{B}\cap\mathscr{V}\rangle$ is stable. Extend F_2 by 0 to a mapping (also denoted by F_2) defined on all of $\mathscr{X}$; then $F=F_1+F_2$ has the required properties.

The requirements (i)–(iv) of Theorem 2 need not be compatible. We shall not work out the precise necessary and sufficient conditions here; one may refer to Bhattacharyya (1978) for a result on a closely related problem. The usual approach to observer problems, in the case where $\mathscr{B}_0=\mathscr{O}$, is to take $\mathscr{V}_2=\mathscr{X}$; one can then dualize to obtain a 'stable cover problem' (Wonham and Morse 1972). Of course, this choice of $\mathscr{V}_2$ need not be optimal. This is especially clear when $\mathscr{B}=\mathscr{O}$; we shall consider this situation in the next section.

5. The minimal detector problem

In this section, we consider the special problem of observing all of the state variables of a given system. So, in the notation of (1), $\mathscr{Z}=\mathscr{X}$, $D=I$ and $z(t)=x(t)$. This gives Ker $D=\mathscr{O}$, so the conditions (ii) and (ii)′ of Theorem 2 reduce to $\mathscr{V}_1\cap\text{Ker } C=\mathscr{O}$ and $\mathscr{V}_1=\mathscr{O}$, respectively. To avoid some complications, we shall assume $\mathscr{B}_0=\mathscr{O}$ (which is, as we see, necessary anyhow if we want to construct a Kalman observer).

This problem is approached in Wonham (1979) (where the name 'minimal detector problem' is used) by a procedure of state reduction followed by an application of Luenberger observer theory. This approach leads to an unsolved algebraic problem. Moreover, it does not give optimal results in all cases, as we shall see in an example below.

Our result is the following.

Theorem 4

There exists an observable observer for the full state of the system (1) with completely measurable disturbance (i.e. $\mathscr{B}_0=\mathscr{O}$) if and only if the pair (C, A) is detectable. If this condition holds, then the minimal observer order is given by

$$\dim\,((\mathscr{X}_b(A)+\langle A|\mathscr{B}\rangle)\cap\text{Ker } C) \tag{29}$$

for a Luenberger observer, and by

$$\dim\,(\mathscr{X}_b(A)+\langle A|\mathscr{B}\rangle) \tag{30}$$

for a Kalman observer.

Proof

We first show that the detectability of the pair (C, A) follows from the necessary conditions of Theorem 1. By (v), there exists $G: \mathscr{Y} \to \mathscr{X}$ such that $(A+GC)\mathscr{V}_1 \subset \mathscr{V}_1$ and $\sigma(\overline{A+GC}) \subset \mathbb{C}_g$, where $\mathscr{V}_1$ is such that $\mathscr{V}_1 \cap \operatorname{Ker} C = \mathscr{O}$. It is clear that there exists $G_1: \mathscr{Y} \to \mathscr{X}$ with $\operatorname{Ran} G_1 \subset \mathscr{V}_1$ such that $\sigma(A+(G+G_1)C|\mathscr{V}_1) \subset \mathbb{C}_g$. Thus, the pair (C, A) must be detectable.

Now let us suppose that (C, A) is detectable. A Kalman observer of order as given by (30) is found as follows. Take $\mathscr{V}_1 = \mathscr{O}$ and $\mathscr{V}_2 = \mathscr{X}_b(A) + \langle A|\mathscr{B}\rangle$; this clearly satisfies the conditions (i), (ii)′ and (iii) of Theorem 2. To see that (iv) is also satisfied, consider the restrictions A_0 and C_0 of A and C to $\mathscr{V}_2$. Because (C, A) is detectable, (C_0, A_0) is detectable too; so there exists $G_0: C\mathscr{V}_2 \to \mathscr{V}_2$ such that $A_0 + G_0C_0$ is stable. Extend G_0 by 0 to a mapping G defined on all of $\mathscr{Y}$; then $\operatorname{Ran} G \subset \mathscr{V}_2$ and

$$\sigma(A+GC) = \sigma(A_0+G_0C_0) \cup \sigma(\bar{A}) \subset \mathbb{C}_g$$

(where $\bar{A}$ on $\mathscr{X}/\mathscr{V}_2$ is stable because $\mathscr{X}_b(A) \subset \mathscr{V}_2$).

Next, we show how to find a Luenberger observer of the order indicated in (29). Again, we take $\mathscr{V}_2 = \mathscr{X}_b(A) + \langle A|\mathscr{B}\rangle$ and consider the restrictions A_0 and C_0 of A and C to $\mathscr{V}_2$. By Lemma 5 in Schumacher (1979 b) (a result closely related to Theorem 3.3 in Wonham (1979)), we can find a subspace $\mathscr{T}$ of $\mathscr{V}_2$ such that $\mathscr{V}_2 = \mathscr{T} \oplus (\mathscr{V}_2 \cap \operatorname{Ker} C)$, and a mapping $G_0: C\mathscr{V}_2 \to \mathscr{V}_2$ such that $(A_0 + G_0C_0)\mathscr{T} \subset \mathscr{T}$ and $\sigma(\overline{A_0+G_0C_0}) \subset \mathbb{C}_g$. Taking $\mathscr{V}_1 = \mathscr{T}$ and following the extension procedure once again, we see that the conditions (i)–(iv) of Theorem 2 are satisfied.

It remains to prove that the values given by (29) and (30) are indeed minimal. If a pair $(\mathscr{V}_1, \mathscr{V}_2)$ satisfies the conditions (i), (ii)′, (iii) and (iv) of Theorem 2, then $\mathscr{V}_1 = \mathscr{O}$ and $\mathscr{V}_2 \supset \mathscr{X}_b(A) + \langle A|\mathscr{B}\rangle$, so $\dim \mathscr{V}_2 - \dim \mathscr{V}_1 \geqslant \dim(\mathscr{X}_b(A) + \langle A|\mathscr{B}\rangle)$. By Theorem 3, this gives the result for Kalman observers. For Luenberger observers, we get: $\mathscr{V}_1 \cap \operatorname{Ker} C = \mathscr{O}$ and $\mathscr{V}_2 \supset \mathscr{X}_b(A) + \langle A|\mathscr{B}\rangle$. It follows that

$$\begin{aligned}\dim \mathscr{V}_2 - \dim \mathscr{V}_1 &\geqslant \dim(\mathscr{V}_2 \cap \operatorname{Ker} C) - \dim(\mathscr{V}_1 \cap \operatorname{Ker} C)\\ &\geqslant \dim((\mathscr{X}_b(A) + \langle A|\mathscr{B}\rangle) \cap \operatorname{Ker} C)\end{aligned}$$

and we obtain the desired result using Theorem 3 again. □

It should be emphasized that all the procedures we have used are constructive.

Remark

The formula (29) giving the minimal order of a Luenberger observer can be viewed as a description of the conflict between three design purposes, concerning (i) fastness of the observer (determined by $\mathbb{C}_g$), (ii) number of measurements (equal to dim Ran C) and (iii) order of the observer (given by (29)). Since

$$\begin{aligned}&\dim((\mathscr{X}_b(A) + \langle A|\mathscr{B}\rangle) \cap \operatorname{Ker} C)\\ &\qquad = \dim(\mathscr{X}_b(A) + \langle A|\mathscr{B}\rangle) - \dim C(\mathscr{X}_b(A) + \langle A|\mathscr{B}\rangle)\end{aligned}$$

it is clear that, whatever choice is made for dim Ran C and for $\mathbb{C}_g$, it will never be optimal to have $C(\mathscr{X}_b(A) + \langle A|\mathscr{B}\rangle) \neq \operatorname{Ran} C$.

Let us assume that $C(\mathscr{X}_b(A)+\langle A|\mathscr{B}\rangle)=\text{Ran } C$; then we can prove a special property for the minimal Luenberger observer constructed in the proof of Theorem 4. Consider the mapping $(I-TC \quad -S)$ which takes the extended state $\begin{pmatrix} x \\ w \end{pmatrix}(t)$ to the error $e(t)$. The range of this mapping contains Ker C, for $(I-TC \quad -S)\begin{pmatrix} x \\ 0 \end{pmatrix}=x$ if $x\in\text{Ker } C$; by definition, its kernel contains the errorless subspace $\mathscr{V}$ of the observer. So we get:

$$\begin{aligned}\dim(\mathscr{X}\oplus\mathscr{W}) &\geqslant \dim \text{Ker } C+\dim \mathscr{V}=\dim \text{Ker } C+\dim \mathscr{V}_2 \\ &=\dim \text{Ker } C+\dim \text{Ran } C+\dim \mathscr{W}=\dim \mathscr{X}+\dim \mathscr{W}\end{aligned}$$

which shows that in fact Ker $(I-TC \quad -S)=\mathscr{V}$ and Ran $(I-TC \quad -S)=\text{Ker } C$. From the latter result, we see that $I-TC$ is a projection on $\mathscr{X}$, and also that $e(t)\in\text{Ker } C$ for all t and for all initial values $x(0)$ and $w(0)$. If we write $C\hat{x}(t)=\hat{y}(t)$, this may be re-formulated as: $\hat{y}(t)=y(t)$ for all t. We might say that *the output is reproduced by the observer.*

This property can be used as a check on the computations when a minimal Luenberger observer is constructed. In Schumacher (1978), it has been used as a starting point for proving the ' Luenberger ' part of Theorem 4.

Example

We take the following example from Wonham (1979, § 3.7)

$$A=\left(\begin{array}{ccc|cc} -1 & 1 & 0 & & \\ 0 & -1 & 1 & & \bigcirc \\ 0 & 0 & -1 & & \\ \hline & & & 0 & 1 \\ & \bigcirc & & 0 & 0 \end{array}\right),\quad C=\begin{pmatrix} 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 1 & 0 \end{pmatrix} \tag{31}$$

There is no disturbance present ($B=0$), and we are asked to construct a minimal order stable Luenberger observer for the full state of the system, where ' stable ' has the usual meaning: $\mathbb{C}_g=\{z\in\mathbb{C}\,|\,\text{Re } z<0\}$. Our theory indicates that a zero-order observer should be possible. Indeed, following the construction procedures leads to the following estimate, which uses no extra dynamics:

$$\begin{pmatrix} \hat{x}_1 \\ \hat{x}_2 \\ \hat{x}_3 \\ \hat{x}_4 \\ \hat{x}_5 \end{pmatrix}=\begin{pmatrix} 0 \\ 0 \\ 0 \\ y_2 \\ y_1 \end{pmatrix}=\begin{pmatrix} 0 \\ 0 \\ 0 \\ x_2+x_3+x_4 \\ x_5 \end{pmatrix} \tag{32}$$

(The quickest way to see that this does the job is perhaps to note that the kernel of the mapping which takes $x(t)$ to $e(t)$ (that is, the subspace of those x for which $\hat{x}=x$) is precisely $\mathscr{X}_b(A)$.) We have $C\mathscr{X}_b(A)=\mathscr{Y}$ in this example, so the ' output reproduction ' property must hold, and it is easily checked that it does.

It is shown in Wonham (1979) that the method put forward there can only lead to a stable observer of order two; consequently, we conclude that this method does not, in general, produce minimal stable observers in the sense of the present paper.

6. Conclusions

We have presented a general formulation of the minimal stable observer problem, and showed its equivalence to an algebraic problem phrased in terms of the original state-space. Using this as a basis, we were able to solve the special problem of finding a minimal observer that estimates the complete state of the given system. While we did not solve the general problem, the solution of the special problem provides an upper bound that may be useful in many practical cases. It also opens perspectives for generalization to infinite-dimensional systems.

References

Basile, G., and Marro, G., 1969, *J. optim. Theory Applic.*, **3**, 410.
Bhattacharyya, S. P., 1978, *I.E.E.E. Trans. autom. Control*, **23**, 483.
Luenberger, D. G., 1964, *I.E.E.E. Trans. milit. Electron.*, **8**, 74 ; 1966, *I.E.E.E. Trans. autom. Control*, **11**, 190 ; 1971, *I.E.E.E. Trans. autom. Control*, **16**, 596.
Schumacher, J. M., 1978, Report 87 Wisk. Sem. V.U. ; 1979 a, Report 110 Wisk. Sem. V.U. ; 1979 b, Report 113 Wisk. Sem. V.U.
Scott, R. W., and Anderson, B. D. O., 1978, *Automatica*, **14**, 481.
Sirisena, H. R., 1979, *Int. J. Control*, **29**, 235.
Wonham, W. M., 1979, *Linear Multivariable Control : a Geometric Approach* (New York : Springer-Verlag).
Wonham, W. M., and Morse, A. S., 1972, *Automatica*, **8**, 93.