CHAPTER 3

# Filtering theory

The stochastic state space model introduced in Section 2.4 is an internal model: its states $x_k$ are not observed directly but do contribute to the observed outputs $y_k$ as specified by the observation equation in (2.4.3). It is natural then to consider the problem of forming 'best estimates' of the state $x_k$ give the available data $(y_0, y_1, \ldots, y_k)$. This procedure is known as *filtering*. There are at least three situations in which filtering is required. Firstly, it may be an end in itself: this is the case when, as often happens, the state variables $x_k^i$ represent important physical quantities in a system which we need to know as accurately as possible even though they cannot be measured directly. Secondly, if we wish to control systems described by state space models then the natural class of controls to consider is that of state feedback controls where the control variable $u_k$ takes the form $u_k = u(k, x_k)$. If $x_k$ is not 'known' then in some circumstances it can be replaced by a best estimate $\hat{x}_k$ produced by filtering; this topic is described at length in Chapter 6. Finally, filtering is relevant when we wish to replace the state space model by an 'equivalent' external model; see section 3.4 below.

Initially we will consider the filtering or estimation problem in a more general setting than that described above, specializing to state space models later. The general problem may be described as follows: one observes the values of random variables $Y_1, \ldots, Y_n$ and wishes to 'estimate' the value of another random variable $Y_0$. Here $\bar{Y}^T := (Y_0, Y_1, \ldots, Y_n)$ is a vector random variable with a given joint distribution function $F$. An *estimator* is any function $g(Y)$ of the observed vector $Y^T := (Y_1, \ldots, Y_n)$ and this is to be chosen so as to minimize the *mean square error*

$$\mathscr{E} = E[Y_0 - g(Y)]^2. \tag{3.0.1}$$

We have already seen in Section 1.1 that the function $g$ which minimizes the mean square error is the conditional expectation

$$E[Y_0|Y] = \int_{-\infty}^{\infty} y_0 \, \mathrm{d}F_{Y_0|Y}(y_0|Y).$$

However, this may be hard to compute and in any case we may only know certain parameters of the joint distribution of $\bar{Y}$ rather than the function $F$ itself. For these reasons and others which will emerge later, we are led to study the *linear estimation problem* where the choice of estimators $g$ is limited to linear functions, i.e. those of the form

$$g(Y) = \alpha_1 Y_1 + \alpha_2 Y_2 + \cdots + \alpha_n Y_n. \tag{3.0.2}$$

This is much simpler since we are now just searching for the $n$-vector $\alpha^{\mathrm{T}} = (\alpha_1, \ldots, \alpha_n)$ which mininizes (3.0.1) with $g$ given by (3.0.2). Notice that in this case

$$E[Y_0 - g(Y)]^2 = E \sum_{i,j=0}^{n} \alpha_i \alpha_j Y_i Y_j$$

$$= \sum_{i,j}^{n} \alpha_i \alpha_j E Y_i Y_j$$

where for notational convenience we have defined $\alpha_0 = -1$. Suppose that all the random variables have zero mean. Then $E Y_i Y_j$ is just the $(i,j)$th entry of the covariance matrix $\mathrm{cov}(\bar{Y})$, and this shows that in order to solve the linear estimation problem we only need to know the means ($= 0$) and covariances of the random variables. This is much more reasonable than requiring that the whole joint distribution function be known. (Of course, the theory only applies when all the random variables have finite variance, but this is hardly a restriction in practice.)

The solution of the linear estimation problem in principle is quite straightforward and in fact a formula for $\alpha$ is given in Theorem 3.1.1 below. The key idea is that the best linear estimate can be thought of geometrically as the 'orthogonal projection' of $Y_0$ onto the observations $Y$. Section 3.1 is devoted to explaining this idea and its relation to the conditional expectation mentioned above. What remains is to develop effective ways of calculating this projection. The main application we have in mind is estimating the state vector of the state-space model of Section 2.4 from the output. This problem has a dynamic structure in that the output values $y_0, y_1, \ldots$ are measured at successive instants of time and we wish to 'keep track of' the state vector $x_k$ as it evolves. Thus a *recursive algorithm* is required which will take the estimate at time $k$ and, using the new observation $y_{k+1}$,

update it to give the estimate at time $k+1$. Such recursive estimators, or *filters*, are discussed in general terms in Section 3.2. We then derive in Section 3.3 the *Kalman filter* equations which provide a recursive estimator for the state-space model. Kalman filtering theory is applied in Section 3.4 to derive the *innovations representation* of the state-space model mentioned at the end of Chapter 2.

## 3.1 The geometry of linear estimation

To introduce the geometric picture of linear estimation let us consider the problem introduced above with $n=1$. Thus $(Y_0, Y_1)$ are jointly distributed zero-mean random variables, and we wish to find the number $\alpha$ which minimizes

$$\mathscr{E} = E[Y_0 - \alpha Y_1]^2 = E(Y_0^2) - 2\alpha E(Y_0 Y_1) + \alpha^2 E(Y_1^2).$$

Elementary calculus shows that the right choice is

$$\alpha = \frac{E(Y_0 Y_1)}{E(Y_1^2)} \tag{3.1.1}$$

(provided that $E(Y_1^2) \neq 0$) resulting in a minimum error of

$$\mathscr{E} = E(Y_0^2) - \frac{1}{E(Y_1^2)}(E(Y_1 Y_0))^2.$$

Let $\sigma_0, \sigma_1, \rho$ be the standard deviations and correlation coefficient of $Y_0$, $Y_1$ (see Section 1.1.1). Then the best estimator is

$$\hat{Y}_0 = \alpha Y_1 = \rho \frac{\sigma_0}{\sigma_1} Y_1 \tag{3.1.2}$$

and the error is

$$\tilde{Y}_0 = Y_0 - \alpha Y_1 = \sigma_0 \left( \frac{Y_0}{\sigma_0} - \rho \frac{Y_1}{\sigma_1} \right) \tag{3.1.3}$$

with variance

$$\mathscr{E} = \sigma_0^2 (1 - \rho^2).$$

Now note the crucial fact that *the error $\tilde{Y}$ is uncorrelated with the observed random variable $Y_1$*, i.e.

$$E(\tilde{Y}_0 Y_1) = 0.$$

This is easily seen from (3.1.3). It is also easily seen that the value of $\alpha$

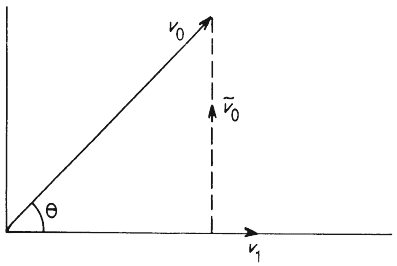


Fig. 3.1

given by (3.1.1) is the only one such that $(Y_0 - \alpha Y_1)$ and $Y_1$ are uncorrelated.

The geometric picture that goes along with this is as follows: Suppose $\mathbf{v}_0$, $\mathbf{v}_1$ are vectors in the plane which have lengths $\sigma_0$, $\sigma_1$ respectively and intersect at an angle $\theta$ such that $\cos\theta = \rho$ (see Fig. 3.1). The vector $\mathbf{v}_0$ can be expressed as the vector sum of its projection $\hat{\mathbf{v}}_0$ on to $\mathbf{v}_1$ and the difference $\tilde{\mathbf{v}}_0 = \mathbf{v}_0 - \hat{\mathbf{v}}_0$ which is orthogonal to $\mathbf{v}_1$. The projection $\hat{\mathbf{v}}_0$ is given by

$$\hat{\mathbf{v}}_0 = \sigma_0 \cos\theta \left( \frac{1}{\sigma_1} \mathbf{v}_1 \right) = \rho \frac{\sigma_0}{\sigma_1} \mathbf{v}_1 \tag{3.1.4}$$

Comparing (3.1.2) and (3.1.4) we see that if the random variables $Y_0$, $Y_1$ are identified with the vectors $\mathbf{v}_0$, $\mathbf{v}_1$ respectively then the best linear estimate $\hat{Y}_0$ corresponds to the projection $\hat{\mathbf{v}}_0$ of $\mathbf{v}_0$ onto $\mathbf{v}_1$. The *inner* (or *dot*) *product* of the vectors $\mathbf{v}_0$ and $\mathbf{v}_1$ is

$$\mathbf{v}_0 \cdot \mathbf{v}_1 = \sigma_0 \sigma_1 \cos\theta = \sigma_0 \sigma_1 \rho = E Y_0 Y_1 = \operatorname{cov}(Y_0, Y_1).$$

Thus the vectors representing the random variables have lengths equal to the standard deviations of the random variables and inner product equal to the covariance. Notice in particular that if $\theta = 0$ or $\theta = \pi$ then the vectors are colinear and $\hat{\mathbf{v}}_0 = \pm \mathbf{v}_0 = \pm (\sigma_0/\sigma_1)\mathbf{v}_1$. Since $\rho = \cos\theta$ the equivalent condition on $\rho$ is that $\rho = \pm 1$. But we already saw in Chapter 1 that if $Y_0$, $Y_1$ have correlation coefficient $\pm 1$ then they are linearly related: $Y_0 = \pm (\sigma_0/\sigma_1) Y_1$. Thus 'linear estimation' can be done with zero error, as the geometric picture indicates.

In order to formalize the above discussion and generalize it to higher dimensions we need to review the geometrical properties of $\mathbb{R}^d$ considered as a vector space. Elements or *vectors* $\mathbf{x}$ of $\mathbb{R}^d$ are $n$-tuples of real numbers $\mathbf{x} = (x_1, x_2, \ldots, x_d)$. Addition and scalar multiplic-

ation are defined componentwise: $\mathbf{x}+\mathbf{y}=(x_1+y_1,\ldots,x_d+y_d)$ and $\alpha\mathbf{x}=(\alpha x_1,\ldots,\alpha x_d)$ for $\alpha\in\mathbb{R}$. The *inner* or *dot* product of two vectors $\mathbf{x}$, $\mathbf{y}$ is defined by

$$\mathbf{x}\cdot\mathbf{y}=\sum_{i=1}^{d}x_iy_i.$$

The vectors $\mathbf{x}$ and $\mathbf{y}$ are *orthogonal* $(\mathbf{x}\perp\mathbf{y})$ if $\mathbf{x}\cdot\mathbf{y}=0$. The *norm* of a vector is

$$\|\mathbf{x}\|=\sqrt{(\mathbf{x}\cdot\mathbf{x})}.$$

For $i=1, 2,\ldots,d$ define

$$\mathbf{z}:=(0,\ldots,0,\ 1,\ 0,\ldots,0)\ (1 \text{ in the } i\text{th position}).$$

These are the *coordinate vectors*. They have the following properties:

(a) They are normalized and mutually orthogonal:

$$\mathbf{z}_i\cdot\mathbf{z}_j=\begin{cases}1 & i=j\\ 0 & i\neq j\end{cases}$$

(b) They form a *basis* for $\mathbb{R}^d$: any $\mathbf{x}\in\mathbb{R}^d$ can be expressed as $\mathbf{x}=\sum_1^d a_i\mathbf{z}_i$ for some coefficients $a_i$.

It is clear from the definitions that the coefficients $a_i$ in (b) are given by $a_i=\mathbf{x}\cdot\mathbf{z}_i$, so that each $\mathbf{x}\in\mathbb{R}^d$ has the representation:

$$\mathbf{x}=\sum_{i=1}^{d}(\mathbf{x}\cdot\mathbf{z}_i)\mathbf{z}_i.$$

Any set of vectors $\mathbf{z}_i$ satisfying (a) and (b) is called an *orthonormal basis* of $\mathbb{R}^d$. A *subspace* $\mathscr{L}$ of $\mathbb{R}^d$ is a subset with the property that if $\mathbf{x}, \mathbf{y}\in\mathscr{L}$ then $\alpha\mathbf{x}+\beta\mathbf{y}\in\mathscr{L}$ for any $\alpha, \beta\in\mathbb{R}$. The subspace *generated* or *spanned* by any collection of vectors $\mathbf{u}_1,\ldots,\mathbf{u}_m$ is denoted by $\mathscr{L}(\mathbf{u}_1,\ldots,\mathbf{u}_m)$ and is the smallest subspace containing the generating vectors. It is easy to see that

$$\mathscr{L}(\mathbf{u}_1,\ldots,\mathbf{u}_m)=\left\{\sum_{i=1}^{m}a_i\mathbf{u}_i:\mathbf{a}=(a_1,\ldots,a_m)\in\mathbb{R}^m\right\}.$$

It is always possible to construct an orthonormal basis $\mathbf{x}_1,\ldots,\mathbf{x}_d$ of $\mathbb{R}^d$ such that $\mathscr{L}(\mathbf{u}_1,\ldots,\mathbf{u}_m)=\mathscr{L}(\mathbf{x}_1,\ldots,\mathbf{x}_k)$ for some $k\leq\min\{d,m\}$. This

can be done by using the *Gram–Schmidt orthogonalization procedure*, which we describe next. Supppose, to avoid triviality, that $\|\mathbf{u}_i\| > 0$ for some $i$ (otherwise $\mathscr{L}(\mathbf{u}_1, \ldots, \mathbf{u}_m) = \{0\}$); we can then assume, permuting indices if necessary, that $\|\mathbf{u}_1\| > 0$. Define

$$\mathbf{x}_1 = \frac{1}{\|\mathbf{u}_1\|}\mathbf{u}_1.$$

Now suppose that orthogonal vectors $\mathbf{x}_1, \ldots, \mathbf{x}_{k(l)}$ have been found for some number $k(l) \leq \min\{d, m\}$ such that $\mathscr{L}(\mathbf{u}_1, \ldots, \mathbf{u}_l) = \mathscr{L}(\mathbf{x}_1, \ldots, \mathbf{x}_{k(l)})$. Define

$$\mathbf{v} := \mathbf{u}_{l+1} - \sum_{i=1}^{k(l)} (\mathbf{u}_{l+1} \cdot \mathbf{x}_i)\mathbf{x}_i.$$

Then $\mathbf{v} \perp \mathbf{x}_i$ for $i = 1, \ldots, k(l)$. If $\|\mathbf{v}\| = 0$, set $k(l+1) := k(l)$; otherwise, set $k(l+1) := k(l) + 1$ and $\mathbf{x}_{k(l+1)} = \mathbf{v}/\|\mathbf{v}\|$. Then $\mathbf{x}_1, \ldots, \mathbf{x}_{k(l+1)}$ are orthonormal and $\mathscr{L}(\mathbf{u}_1, \ldots, \mathbf{u}_{l+1}) = \mathscr{L}(\mathbf{x}_1, \ldots, \mathbf{x}_{k(l+1)})$. Since clearly $\mathscr{L}(\mathbf{u}_1) = \mathscr{L}(\mathbf{x}_1)$ we conclude by induction that $\mathscr{L}(\mathbf{u}_1, \ldots, \mathbf{u}_m) = \mathscr{L}(\mathbf{x}_1, \ldots, \mathbf{x}_{k(m)})$. By construction $k := k(m) \leq m$, and $k \leq d$ since $d$ is the maximum number of linearly independent vectors in $\mathbb{R}^d$. If $k < d$ then orthonormal vectors $\mathbf{x}_{k+1}, \ldots, \mathbf{x}_d$ can be constructed in a similar way so that $\mathbf{x}_1, \ldots, \mathbf{x}_d$ form a basis of $\mathbb{R}^d$. We leave it to the reader to supply the details.

The *orthogonal projection* $\hat{\mathbf{v}}$ of $\mathbf{v} \in \mathbb{R}^d$ onto a subspace $\mathscr{U} := \mathscr{L}(\mathbf{u}_1, \ldots, \mathbf{u}_m)$ is defined by

$$\hat{\mathbf{v}} = \sum_{i=1}^{k} (\mathbf{v} \cdot \mathbf{x}_i)\mathbf{x}_i$$

where $\mathbf{x}_1, \ldots, \mathbf{x}_d$ is an orthonormal basis such that $\mathscr{U} = \mathscr{L}(\mathbf{x}_1, \ldots, \mathbf{x}_k)$. $\hat{\mathbf{v}}$ can be characterized in the following two equivalent ways.

(a) $\hat{\mathbf{v}}$ is the unique vector satisfying

$$\hat{\mathbf{v}} \in \mathscr{U}$$
$$\mathbf{v} - \hat{\mathbf{v}} \perp \mathscr{U}$$

(b) $\hat{\mathbf{v}}$ is the closest point in $\mathscr{U}$ to $\mathbf{v}$, i.e.

$$\|\mathbf{v} - \hat{\mathbf{v}}\| = \min_{\mathbf{u} \in \mathscr{L}} \|\mathbf{v} - \mathbf{u}\|.$$

In (a), $\mathbf{v} - \hat{\mathbf{v}} \perp \mathscr{U}$ means that $(\mathbf{v} - \hat{\mathbf{v}}) \perp \mathbf{u}$ for all $\mathbf{u} \in \mathscr{L}$.

Both (a) and (b) are very easily established using the basis $\mathbf{x}_1, \ldots, \mathbf{x}_d$, but note that the statements themselves do not involve any particular choice of basis. For part (b) we have

$$\mathbf{v} - \mathbf{u} = (v_1 - u_1)\mathbf{x}_1 + \cdots + (v_k - u_k)\mathbf{x}_k + v_{k+1}\mathbf{x}_{k+1} + \cdots + v_d\mathbf{x}_d$$

where $v_i = \mathbf{v}\cdot\mathbf{x}_i$, $u_i = \mathbf{u}\cdot\mathbf{x}_i$. Thus

$$\|\mathbf{v} - \mathbf{u}\|^2 = \sum_{i=1}^{k} (v_i - u_i)^2 + \sum_{i=k+1}^{d} v_i^2$$

and this is clearly minimized by taking $u_i = v_i$, $i \leq k$.

Let us denote $\hat{\mathbf{v}} = \mathscr{P}\mathbf{v}$. Then $\mathscr{P}$ is a *projection operator* which maps each vector in $\mathbb{R}^d$ to its projection onto the subspace $\mathscr{U}$. We note the following properties of the projection operator:

(a) $\mathscr{P}$ is linear: $\mathscr{P}(\alpha\mathbf{v}_1 + \beta\mathbf{v}_2) = \alpha\mathscr{P}\mathbf{v}_1 + \beta\mathscr{P}\mathbf{v}_2$, $\alpha, \beta \in \mathbb{R}$

(b) $\mathscr{P}^2 = \mathscr{P}$ (Here $\mathscr{P}^2\mathbf{v} := \mathscr{P}(\mathscr{P}\mathbf{v})$)

(c) If $\mathscr{U}'$ is a subspace such that $\mathscr{U}' \supset \mathscr{U}$ and $\mathscr{P}'$ is the projection onto $\mathscr{U}'$ then for any $\mathbf{v} \in \mathbb{R}^d$

$$\mathscr{P}\mathbf{v} = \mathscr{P}(\mathscr{P}'\mathbf{v}).$$

The first two of these are evident. For (c), suppose that $\mathscr{U}' = \mathscr{L}(\mathbf{u}_1, \ldots, \mathbf{u}_{m'})$ for some $m' > m$. By means of the Gram–Schmidt procedure we can construct a basis $\mathbf{x}_1, \ldots, \mathbf{x}_d$ and numbers $k, k'$ with $k \leq k' \leq d$ such that $\mathscr{U} = \mathscr{L}(\mathbf{x}_1, \ldots, \mathbf{x}_k)$ and $\mathscr{U}' = \mathscr{L}(\mathbf{x}_1, \ldots, \mathbf{x}_{k'})$. For $\mathbf{v} \in \mathbb{R}^d$, denote $v_i = \mathbf{v}\cdot\mathbf{x}_i$. Then $\mathscr{P}'\mathbf{v} = \sum_1^{k'} v_i\mathbf{x}_i$ so that $\mathscr{P}(\mathscr{P}'\mathbf{v}) = \sum_1^k v_i\mathbf{x}_i = \mathscr{P}\mathbf{v}$.

Now back to random variables. Suppose as before that $\bar{Y} := (Y_0, Y_1, \ldots, Y_n)^{\mathrm{T}}$ is a random $(n+1)$-vector such that for each $i$

$$EY_i = 0, \operatorname{var}(Y_i) < \infty,$$

and denote $Q := \operatorname{cov}(\bar{Y})$. We wish to associate these random variables with vectors $\mathbf{v}_0, \ldots, \mathbf{v}_n$ in such a way that

$$\mathbf{v}_i\cdot\mathbf{v}_j = \operatorname{cov}(Y_i, Y_j) = EY_iY_j$$

More precisely, let $\mathscr{H}$ denote the set of all linear combinations of the random variables $Y_0, \ldots, Y_n$, i.e.

$$\mathscr{H} = \left\{ \sum_{i=0}^{n} \alpha_i Y_i : \alpha = (\alpha_0, \ldots, \alpha_n) \in \mathbb{R}^{n+1} \right\}.$$

We take the function $U, V \to EUV$ as an 'inner product' for $U, V \in \mathscr{H}$. Note that $EUV$ is entirely determined by the covariance matrix $Q$

since if $U, V \in \mathscr{H}$ then $U = \mathbf{a}^{\mathrm{T}}\bar{Y}$ and $V = \mathbf{b}^{\mathrm{T}}\bar{Y}$ for some $\mathbf{a}, \mathbf{b} \in \mathbb{R}^{n+1}$ and then $EUV = \mathbf{a}^{\mathrm{T}}Q\mathbf{b}$. We wish to construct a function $\varphi: \mathscr{H} \to \mathbb{R}^d$ for some integer $d$ with the following properties

(a) $\varphi$ is linear, one-to-one and onto

(b) $\varphi$ is inner product preserving:

$$\varphi(U) \cdot \varphi(V) = EUV. \tag{3.1.5}$$

Such a function $\varphi$ always exists. Recall from Proposition 1.1.3 that by factoring $Q$ in the form $Q = AA^{\mathrm{T}}$ we can express $Y$ in the form

$$\bar{Y} = AZ$$

where $Z^{\mathrm{T}} = (Z_1, \ldots, Z_d)$ is a vector of unit variance uncorrelated random variables and $d \leq n+1$. Now define

$$\varphi(Z_i) := \mathbf{z}_i$$

where $\mathbf{z}_1, \ldots, \mathbf{z}_d$ is the coordinate basis of $\mathbb{R}^d$, and

$$\varphi(\mathbf{a}^{\mathrm{T}}Z) := \sum_{i=1}^{d} a_i \mathbf{z}_i \quad \text{for } \mathbf{a} \in \mathbb{R}^d.$$

Since $\mathscr{H} = \{\mathbf{a}^{\mathrm{T}}Z : \mathbf{a} \in \mathbb{R}^d\}$ this defines $\varphi(U)$ for all $U \in \mathscr{H}$. By construction, $\varphi$ is linear and onto, and an immediate calculation shows that (3.1.5) holds. In particular if we define $\mathbf{v}_i := \varphi(Y_i) = \varphi(\sum_k a_{ik} Z_k)$ then we see that

$$\mathbf{v}_i \cdot \mathbf{v}_j = EY_i \cdot Y_j.$$

To check that $\varphi$ is one-to-one, suppose that $\varphi(U) = \varphi(V)$; then $\varphi(U - V) = \varphi(U) - \varphi(V) = 0$ so that $E(U-V)^2 = \varphi(U-V) \cdot \varphi(U-V) = 0$. Recall by the way that $E(U-V)^2 = 0$ if and only if $P[U = V] = 1$. Thus this theory does not distinguish between equivalent random variables, i.e. if $P[U = V] = 1$ then $\varphi(U) = \varphi(V)$.

The existence of the map $\varphi$ means that the geometrical properties of the space $\mathscr{H}$ with 'inner product' $EUV$ and 'distance' $[E(U-V)^2]^{1/2}$ are identical to those of Euclidean space $\mathbb{R}^d$. To illustrate the utility of this, let $\mathscr{V}$ be the subspace spanned by $\mathbf{v}_1, \ldots, \mathbf{v}_n$ where $\mathbf{v}_i = \varphi(Y_i)$ and let $\hat{\mathbf{v}}_0$ be the projection of $\mathbf{v}_0$ on to $\mathscr{V}$. Then

$$\hat{\mathbf{v}}_0 = \alpha_1 \mathbf{v}_1 + \cdots + \alpha_n \mathbf{v}_n$$

for some constants $\alpha_1, \ldots, \alpha_n$, and the corresponding element of $\mathscr{H}$ is

$$\hat{Y}_0 := \varphi^{-1}(\hat{\mathbf{v}}_0) = \alpha_1 Y_1 + \cdots + \alpha_n Y_n.$$

Now recall that $\hat{\mathbf{v}}_0$ is the closest point in $\mathscr{V}$ to $\mathbf{v}_0$:

$$\|\hat{\mathbf{v}}_0 - \mathbf{v}_0\| = \min_{\mathbf{u}\in V} \|\mathbf{v}_0 - \mathbf{u}\|.$$

It follows from (3.1.5) that $E(U-V)^2 = \|\boldsymbol{\varphi}(U) - \boldsymbol{\varphi}(V)\|^2$; thus $\hat{Y}_0$ satisfies

$$E(Y_0 - \hat{Y}_0)^2 = \min_U E(Y_0 - U)^2$$

where the minimum is taken over all linear combinations $U = \sum_1^n \alpha_i Y_i$. But this means that $\hat{Y}_0$ solves the linear estimation problem. $\hat{Y}_0$ has the property that $E[(Y_0 - \hat{Y}_0)Y_i] = 0$, $i = 1, 2, \ldots, n$, i.e. the error $Y_0 - \hat{Y}_0$ is uncorrelated with the observed random variables $Y_1, \ldots, Y_n$ just as in the scalar case.

We can dispense with explicit mention of the map $\boldsymbol{\varphi}$ and Euclidean space $\mathbb{R}^d$. Just think of the random variables as 'vectors' with lengths equal to their standard deviations and 'inner product' given by the covariance. Thus two random variables are 'orthogonal' (and we write $U \perp V$) if they are uncorrelated, and the best linear estimator $\hat{Y}_0$ is the 'projection' of $Y_0$ onto the 'subspace' $\mathscr{L}(Y_1, \ldots, Y_n)$ spanned by $Y_1, \ldots, Y_n$.

Let us summarize the results we have obtained. At the same time we generalize to the vector case, replacing $Y_0$ by a $p$-vector $X$.

*Theorem* 3.1.1

Let $X$ and $Y$ be random $p$- and $n$-vectors respectively, all components having zero mean and finite variance. (Here, $Y^{\mathrm{T}} = (Y_1, \ldots, Y_n)$.) Then for each $j = 1, \ldots, p$ there is a unique (up to equivalence) random variable $\hat{X}_j$ such that:

(a) $\hat{X}_j \in \mathscr{L}(Y)$
(b) $X_j - \hat{X}_j \perp \mathscr{L}(Y)$.

$\hat{X}^{\mathrm{T}} := (\hat{X}_1, \ldots, \hat{X}_p)$ is the minimum mean-square error estimate of $X$ given $Y$, i.e. for any $\beta \in \mathbb{R}^p$

$$E[(\beta^{\mathrm{T}}(X - \hat{X})]^2 = \min_{U\in\mathscr{L}(Y)} E[\alpha^{\mathrm{T}}X - U]^2.$$

If $\mathrm{cov}(Y)$ is non-singular then $\hat{X}$ is given by

$$\hat{X} = E[XY^{\mathrm{T}}][\mathrm{cov}(Y)]^{-1}Y. \tag{3.1.6}$$

REMARK By a slight abuse of terminology, $\hat{X}$ is referred to as the 'projection of $X$ onto $\mathscr{L}(Y)$'.

PROOF Only the last part remains to be established. By definition, $\hat{X} = AY$ for some $p \times n$ matrix $A$. Using the orthogonality relation (b) we see that for any $\beta \in \mathbb{R}^p$, $\gamma \in \mathbb{R}^n$

$$E[\beta^{\mathrm{T}}(X - AY)(\gamma^{\mathrm{T}}Y)] = 0$$

i.e.

$$\beta^{\mathrm{T}}[E(XY^{\mathrm{T}} - AYY^{\mathrm{T}})]\gamma = 0.$$

This implies that

$$E[XY^{\mathrm{T}}] - AE[YY^{\mathrm{T}}] = 0$$

and hence that $A = E[XY^{\mathrm{T}}][\mathrm{cov}(Y)]^{-1}$ if $\mathrm{cov}(Y)$ is non-singular. If $\mathrm{cov}(Y)$ is singular then some components of $Y$ are linearly related and it may be possible to express $\hat{X}$ in several different but equivalent ways.

### *Random variables with non-zero mean*

Let us consider the same problem as above (with scalar $Y_0$) but supposing now that the random variables have possibly non-zero means

$$EY_i = m_i \qquad i = 0, 1, \ldots, m.$$

This situation easily reduces to the zero-mean case. Rather than a linear estimator, it is preferable now to use an *affine* (linear + constant) estimator:

$$\hat{Y}_0 = \alpha_1 Y_1 + \cdots + \alpha_n Y_n + \beta.$$

We have to choose $\alpha_1, \ldots, \alpha_n, \beta$ to minimize $E[Y_0 - \hat{Y}_0]^2$. Minimization can be carried out over these coefficients in any order, so let us fix $\alpha_1, \ldots, \alpha_n$ and minimize first over $\beta$. Define

$$U = Y_0 - \alpha_1 Y_1 - \cdots - \alpha_n Y_n.$$

Then

$$E[Y_0 - \hat{Y}_0]^2 = E[U - \beta]^2$$

It was shown in Proposition 1.1.1 that this is minimized by taking

$$\beta = EU = m_0 - \alpha_1 m_1 - \cdots - \alpha_n m_n.$$

Incidentally, this justifies our previous implicit choice $\beta = 0$ for the zero-mean case. With the above choice of $\beta$ we see that

$$E[Y_0 - \hat{Y}_0]^2 = E[Y_0^c - (\alpha_1 Y_1^c + \cdots + \alpha_n Y_n^c)]^2 \tag{3.1.7}$$

where $Y_i^c$ is the 'centered' random variable $Y_i^c = Y_i - m_i$. We now have to choose $\alpha_1, \ldots, \alpha_n$ to minimize (3.1.7), but this is the zero-mean problem that was solved before. Let $P$ be the covariance matrix of $Y$, now given by

$$P_{ij} = E[(Y_i - m_i)(Y_j - m_j)].$$

If $P$ is non-singular, then from Theorem 3.1.1

$$\hat{Y}_0 = (Y - m)^{\mathrm{T}} P^{-1} E[(Y - m)(Y_0 - m_0)] + m_0 \tag{3.1.8}$$

where $m^{\mathrm{T}} = (m_1, \ldots, m_n)$. Notice that *the error* $Y_0 - \hat{Y}_0$ *always has zero mean.*

To get the geometric picture for this case we adopt the rather artificial, but convenient, stratagem of adjoining to the observations another random variable denoted $\mathbb{1}$ which takes on the value 1 with probability one (thus no new 'information' has been added). $\hat{Y}_0$ can then be regarded as a *linear* (no longer affine) combination of the observations:

$$\hat{Y}_0 = \beta \mathbb{1} + \alpha_1 Y_1 + \cdots + \alpha_n Y_n$$

As before, random variables $U, V$ are regarded as vectors with inner product $EUV$, but note that this is *not* now the covariance, which is $E(U - EU)(V - EV)$. Now $U \perp \mathbb{1}$ if $E(\mathbb{1}U) = EU = 0$ and thus if we express $U$ as

$$U = (EU)\mathbb{1} + U^{\mathrm{c}}$$

then the first term on the right is the projection of $U$ onto the one-dimensional subspace spanned by the random-variable $\mathbb{1}$. Thus the random variables $\mathbb{1}, Y_0, \ldots, Y_n$ form a vector space of dimension $k \leq n + 2$ consisting of a $(k-1)$-dimensional subspace of zero-mean random variables (spanned, in fact, by $Y_0^{\mathrm{c}}, \ldots, Y_n^{\mathrm{c}}$) and a 1-dimensional subspace spanned by $\mathbb{1}$. The best estimate of $Y_0$ is the sum of its projection into $\mathscr{L}(Y_1^{\mathrm{c}}, \ldots, Y_n^{\mathrm{c}})$ and its projection onto $\mathscr{L}(\mathbb{1})$ and these projections are the two terms on the right of (3.1.8), respectively.

***The normal case***

As pointed out at the beginning of this chapter, only means and covariances are required to calculate best linear estimators. If we suppose that the random variables involved are jointly normally distributed then we get the following result strengthening Theorem 3.1.1.

*Theorem* 3.1.2

Let $X$ and $Y$ be as in Theorem 3.1.1 but with possibly non-zero means and suppose that $X$ and $Y$ are jointly normally distributed. Then the best affine estimate of $X$ given $Y$ coincides with the conditional expectation $E[X|Y]$.

PROOF Consider first the zero mean case. Since $\hat{X} = AY$ for some matrix $A$, the random variables $(X, \hat{X}, Y)$ are jointly normally distributed, and $(X_i - \hat{X}_i)$ is uncorrelated with and hence independent of $Y_j$ for each $i, j$. Using the properties of conditional expectation given in Proposition 1.1.4 we see that, with $\tilde{X} = X - \hat{X}$,

$$\begin{aligned} E[X|Y] &= E[\hat{X} + \tilde{X}|Y] \\ &= \hat{X} + E[\tilde{X}|Y] \\ &= \hat{X} + E\tilde{X} = \hat{X}. \end{aligned}$$

If $X, Y$ have non-zero means $m_X, m_Y$, write $X^c = X - m_X$, $Y^c = Y - m_Y$. Then

$$E[X|Y] = E[X^c + m_X|Y] = m_X + E[X^c|Y].$$

It follows from Proposition 1.1.7 that $E[X^c|Y] = E[X^c|Y^c]$ and the latter expression coincides with the best linear estimator. This completes the proof. □

This result shows that in the normal case $\hat{X}$ is the best estimate of $X$ not only in the class of affine functions $AY + b$ but also in the class of all finite-variance functions $g(Y)$. It also shows that the *conditional distribution* of $X$ given $Y$ is normal with mean $\hat{X}$ and covariance $\mathrm{cov}(X - \hat{X})$. This follows from the fact that $X = \hat{X} + \tilde{X}$ where $\hat{X}$ is a function of $Y$ and $\tilde{X}$ is independent of $Y$. We have thus, somewhat belatedly, completed the proof of Proposition 1.1.7(e) of Chapter 1.

### 3.2 Recursive estimation

The idea of recursive estimation arises when random variables $Y_1, Y_2, \ldots$ are observed sequentially and we wish to process them in real time to form successive best estimates of an unobserved random variable $Y_0$. At time $n$ we can form the best linear estimate $\hat{Y}_{0,n}$ of $Y_0$ given $Y_1, \ldots, Y_n$ by using formula (3.1.9) (supposing that all means are zero and that the covariance matrix $P_n = \mathrm{cov}(Y_1, \ldots, Y_n)$ is non-singular). Note that this involves inverting the $n \times n$ matrix $P_n$. At the next time instant we have one more observation, $Y_{n+1}$. How are we to compute $\hat{Y}_{0,n+1}$? The most obvious way would be to apply the same formula again. However, if we do this successively for $n = 1, 2, 3, \ldots$, then:

(a) It is necessary to store the entire observation record as this becomes available; and,
(b) At each time $n$, an $n \times n$ matrix must be inverted.

Obviously, the computational effort required to do this becomes massive even for moderate $n$. Is it really necessary, at each stage, to throw away the results of all previous calculations, or is there some method by which $\hat{Y}_{0,n}$ can be updated using the new observation $Y_{n+1}$ to give $\hat{Y}_{0,n+1}$? The simplest form such an updating could take is as follows:

$$\hat{Y}_{0,n+1} = a_n \hat{Y}_{0,n} + b_n Y_{n+1} \tag{3.2.1}$$

i.e. the next estimate is a linear combination of the current estimate and the next observation. Only in special cases will a formula such as (3.2.1) be possible, but these include important applications such as the Kalman filter discussed in Section 3.3.

In this section we discuss the general relation between successive estimates. In view of later applications it is convenient to deal from the outset with the vector case. Thus suppose $x$ is an $n$-vector random variable and $y_1, y_2, \ldots$ are $r$-vectors of observed random variables.[†] All random variables will be taken to have zero mean and finite variance, and to avoid difficulties with non-uniqueness it will be supposed that the covariance matrix of the $rk$-vector $y^k = \mathrm{col}\{y_1, y_2, \ldots y_k\}$ is non-singular for each $k$.

[†] In accordance with the established notational conventions of Kalman filtering theory these are denoted by lower-case letters.

Denote by $\mathscr{L}(y^k)$ the linear subspace spanned by the observations up to time $k$, and by $\hat{x}_k$ the best linear estimate of $x$ given $y^k$, i.e. the projection of $x$ onto $\mathscr{L}(y^k)$. (Recall the notational conventions for projection of vector r.v.s introduced in Section 3.1).

$\mathscr{L}(y^{k-1})$ is a subspace of $\mathscr{L}(y^k)$. Let $\hat{y}_{k|k-1}$ be the projection of $y_k$ onto $\mathscr{L}(y^{k-1})$ and $\tilde{y}_{k|k-1}$ the error: $\tilde{y}_{k|k-1} = y_k - \hat{y}_{k|k-1}$. The random variables $\{\tilde{y}^i_{k|k-1}, i = 1, 2, \ldots, r\}$ span the orthogonal complement of $\mathscr{L}(y^{k-1})$ in $\mathscr{L}(y^k)$, so that any r.v. $Z$ in $\mathscr{L}(y^k)$ has a unique orthogonal decomposition

$$Z = Z_1 + Z_2$$

where $Z_1 \in \mathscr{L}(y^{k-1})$ and $Z_2$ is a linear combination of $\{\tilde{y}^i_{k|k-1}, i = 1, \ldots, r\}$. Take in particular $Z = \hat{x}^i_k$; then we claim that $Z_1 = \hat{x}^i_{k-1}$. Indeed, let $\tilde{x}^i_k = x^i - \hat{x}^i_k$ be the estimation error at time $k$. Then

$$x^i = \hat{x}^i_k + \tilde{x}^i_k = Z_1 + (Z_2 + \tilde{x}^i_k)$$

where $Z_1 \in \mathscr{L}(y^{k-1})$ and $(Z_2 + \tilde{x}^i_k) \perp \mathscr{L}(y^{k-1})$. But we also have

$$x^i = \hat{x}^i_{k-1} + \tilde{x}^i_{k-1}$$

and again $\hat{x}^i_{k-1} \in \mathscr{L}(y^{k-1})$, $\tilde{x}^i_{k-1} \perp \mathscr{L}(y^{k-1})$. Since such orthogonal decompositions are unique, it must be the case that $Z_1 = \hat{x}^i_{k-1}$, as claimed. As to $Z_2$, this is the projection of $\hat{x}^i_k$ onto $\mathscr{L}(\tilde{y}_{k|k-1})$ and this is the same as the projection of $x^i$ onto $\mathscr{L}(\tilde{y}_{k|k-1})$ since $\mathscr{L}(\tilde{y}_{k|k-1}) \subset \mathscr{L}(y^k)$, But this projection can be calculated using formula (3.1.9) again. Collecting the above results we see that $\hat{x}_k$ can be written in the form

$$\hat{x}_k = \hat{x}_{k-1} + E[x\tilde{y}^{\mathrm{T}}_{k|k-1}](E[\tilde{y}_{k|k-1}\tilde{y}^{\mathrm{T}}_{k|k-1}])^{-1}(y_k - \hat{y}_{k|k-1}). \tag{3.2.2}$$

In general this is *not* a recursive formula for $\hat{x}_k$, since $\hat{y}_{\mu|k-1}$ depends on $y_1, \ldots, y_{k-1}$. It *is* a recursive formula precisely when this dependence factors through $\hat{x}_{k-1}$. Let us examine an important example where this occurs.

*Example* 3.2.1

Suppose

$$y_k = Hx + z_k$$

where $H$ is an $r \times n$ matrix and $z_1, z_2, \ldots$ is a sequence of mutually uncorrelated random variables with zero mean and common cova-

riance $\text{cov}(z_k) = N > 0$. We also suppose $x$ and $z_k$ are uncorrelated for each $k$. Thus $y_k$ represents a sequence of 'measurements' of $x$ with uncorrelated measurement errors $z_k$. Let $P$ be the covariance matrix of $x$.

In this example $\hat{y}_{k|k-1}$ is the projection of $y_k = Hx + z_k$ onto $\mathscr{L}(y^{k-1})$ and this is the same as the projection of $Hx$ onto $\mathscr{L}(y^{k-1})$ since $z_k \perp \mathscr{L}(y^{k-1})$. Thus

$$\hat{y}_{k|k-1} = H\hat{x}_{k-1}$$

and (3.2.2) becomes

$$\hat{x}_k = \hat{x}_{k-1} + K(k)(y_k - H\hat{x}_{k-1}) \tag{3.2.3}$$

where $K(k)$ denotes the matrix coefficient in (3.2.2). This is a recursive formula for $\hat{x}_k$ and it only remains to calculate $K(k)$. We will do this in two ways: the 'slick' way specifically adapted to this problem, and by use of a general technique which will be useful in connection with the Kalman filter in the next section.

The slick way is to notice that the $y_k$ are *interchangeable*, in that if

$$\hat{x}_k = A_1 y_1 + \cdots + A_k y_k$$

then all the $A_i$ must be the same, since the correlation structure of the random variables would be completely unchanged if any two observations $y_i$ and $y_j$ were permuted. Denote by $\bar{y}_k$ the sample mean

$$\bar{y}_k = \frac{1}{k}\sum_{i=1}^{k} y_i = Hx + \frac{1}{k}\sum_{i=1}^{k} z_i = Hx - \bar{z}_k.$$

The noise sample mean $\bar{z}_k$ has covariance $N/k$ and our contention is that

$$\hat{x}_k = A\bar{y}_k$$

for some $n \times r$ matrix $A$. The orthogonality condition is

$$x - \hat{x}_k = (I - AH)x - A\bar{z}_k \perp y_i = Hx + z_i \qquad i = 1, \ldots, k.$$

Since $x$ is uncorrelated with $z_i$ and $\bar{z}_k$, this is equivalent to requiring that

$$(I - AH)E[xx^{\mathrm{T}}]H^{\mathrm{T}} - AE[\bar{z}_k z_i^{\mathrm{T}}] = 0.$$

Now $E[xx^{\mathrm{T}}] = \text{cov}(x) = P$ and $E[\bar{z}_k z_i^{\mathrm{T}}] = N/k$ since the $z_j$ are mutually uncorrelated. The fact that this expression is independent of $i$

confirms the 'interchangeability' argument. Thus the orthogonality requirement is:

$$(I - AH)PH^{\mathrm{T}} = \frac{1}{k}AN$$

and hence $A$ is given by

$$A = PH^{\mathrm{T}}\left[ HPH^{\mathrm{T}} + \frac{1}{k}N \right]^{-1}.$$

Notice that $(HPH^{\mathrm{T}} + (1/k)N)$ is non-singular, since by assumption $N > 0$. Thus

$$\hat{x}_k = A\bar{y}_k = \frac{1}{k}PH^{\mathrm{T}}\left[ HPH^{\mathrm{T}} + \frac{1}{k}N \right]^{-1}\left( \sum_{i=1}^{k} y_k \right). \tag{3.2.4}$$

Comparing this with (3.2.3), we see that the coefficient of $y_k$ is $K(k)$ and hence

$$K(k) = \frac{1}{k}PH^{\mathrm{T}}\left[ HPH^{\mathrm{T}} + \frac{1}{k}N \right]^{-1}. \tag{3.2.5}$$

The more general method of obtaining this result is to calculate $K(k)$ from the expression for it in (3.2.2). Now

$$\begin{aligned} \tilde{y}_{k|k-1} = y_k - \hat{y}_{k|k-1} &= (Hx + z_k) - H\hat{x}_{k-1} \\ &= H\tilde{x}_{k-1} + z_k \end{aligned} \tag{3.2.6}$$

where $\tilde{x}_{k-1} = x - \hat{x}_{k-1}$ is the error at time $k-1$. Thus

$$\begin{aligned} E[x\tilde{y}_{k|k-1}^{\mathrm{T}}] &= E[x\tilde{x}_{k-1}^{\mathrm{T}}]H^{\mathrm{T}} \\ &= E[\tilde{x}_{k-1}\tilde{x}_{k-1}^{\mathrm{T}}]H^{\mathrm{T}} \end{aligned}$$

since $x = \hat{x}_{k-1} + \tilde{x}_{k-1}$ and $\hat{x}_{k-1} \perp \tilde{x}_{k-1}$. We denote $P(k-1) = \mathrm{cov}(\tilde{x}_{k-1})$ (the error covariance at time $k-1$). Similarly,

$$\begin{aligned} E[\tilde{y}_{k|k-1}\tilde{y}_{k|k-1}^{\mathrm{T}}] &= E[(H\tilde{x}_{k-1} + z_k)(H\tilde{x}_{k-1} + z_k)^{\mathrm{T}}] \\ &= HP(k-1)H^{\mathrm{T}} + N. \end{aligned}$$

This is non-singular since $N > 0$, and hence

$$K(k) = P(k-1)H^{\mathrm{T}}[HP(k-1)H^{\mathrm{T}} + N]^{-1}$$

It remains to calculate $P(k-1)$. Subtracting $x$ from both sides of

(3.2.2) and using (3.2.6) gives

$$\begin{aligned}\tilde{x}_k &= \tilde{x}_{k-1} - K(k)(H\tilde{x}_{k-1} + z_k)\\ &= (I - K(k)H)\tilde{x}_{k-1} + K(k)z_k.\end{aligned}$$

The two terms in this expression are orthogonal since $\tilde{x}_{k-1} \in \mathscr{L}(x, z_1, \ldots, z_{k-1})$. Thus

$$P(k) = E[\tilde{x}_k \tilde{x}_k^{\mathrm{T}}] = (I - K(k)H)P(k-1)(I - K(k)H)^{\mathrm{T}} + K(k)NK(k). \tag{3.2.7}$$

Now substitute for $K(k)$ from (3.2.5). After a little algebra one finds that (3.2.7) becomes simply

$$P(k) = P(k-1) - P(k-1)H^{\mathrm{T}}[HP(k-1)H^{\mathrm{T}} + N]^{-1}HP(k-1). \tag{3.2.8}$$

Together with the initial condition $P(0) = P = \mathrm{cov}(x)$ this provides a recursive algorithm for generating $P(1), P(2), \ldots$ and hence $K(k)$ from (3.2.5). In this example one can in fact obtain a closed-form expression for $P(k)$ from (3.2.4). Indeed, subtracting $x$ from both sides of (3.2.4) and using the fact that $\bar{y}_k = Hx + \bar{z}_k$ we see that

$$\begin{aligned}\tilde{x}_k = &\left(I - PH^{\mathrm{T}}\left[HPH^{\mathrm{T}} + \frac{1}{k}N\right]^{-1}H\right)x\\ &+ PH^{\mathrm{T}}\left[HPH^{\mathrm{T}} + \frac{1}{k}N\right]^{-1}\bar{z}_k.\end{aligned}$$

Again, the two terms on the right-hand side are orthogonal, and calculating the sum of their covariances we find that

$$P(k) = P - PH^{\mathrm{T}}\left[HPH^{\mathrm{T}} + \frac{1}{k}N\right]^{-1}HP.$$

Some laborious algebra confirms that this indeed satisfies (3.2.8).

In this example the recursive estimator (3.2.3) offers no advantages over the non-recursive form (3.2.4): in either case the main computational task at each stage is to invert an $r \times r$ matrix, so the general problem of having to invert matrices of growing dimensions has been avoided. The storage requirements are also similar: in (3.2.4) one requires the sample mean $\bar{y}_k$ at each stage and this can be updated as follows:

$$\bar{y}_k = \left(\frac{k-1}{k}\right)\bar{y}_{k-1} + \frac{1}{k}y_k.$$

Thus in neither case is it necessary to store the complete observation record. In more general problems such as the Kalman filter considered below, it is usually not possible to obtain simple closed-form expressions for the estimator, but the recursive solution may still be viable. From the implementation point of view this is perfectly satisfactory. The coefficient matrices $K(k)$ can be computed in advance and then the 'data processing' consists of on-line implementation of the very simple algorithm (3.2.3).

## 3.3 The Kalman filter

The Kalman filter is a recursive algorithm for estimating the state $x_k$ of a state-space model given the values of the observed outputs $y^{k-1}(=y_0, y_1, y_2, \ldots, y_{k-1})$. The equations describing the model are

$$x_{k+1} = A(k)x_k + B(k)u_k + C(k)w_k \tag{3.3.1}$$

$$y_k = H(k)x_k + G(k)w_k. \tag{3.3.2}$$

Here, $\{w_k\}$ is an $l$-vector white-noise process with unit covariance ($Ew_k w_k^{\mathrm{T}} = I_l$) and the initial random variable $x_0$ is uncorrelated with $\{w_k\}$, with known mean and covariance $m_0$, $P_0$ respectively. The coefficient matrices $A(k)$, etc., may be time-varying, as indicated by their dependence on $k$ in (3.3.1), (3.3.2). The model is, in this respect, more general than that of Section 2.4. We assume that

$$G(k)G^{\mathrm{T}}(k) > 0 \tag{3.3.3}$$

(in particular this implies that $l \geq r$, $r$ being the dimension of $y_k$). If this were not the case then there would exist vectors $\lambda$ such that $\lambda^{\mathrm{T}}G(k) = 0$, so that, from (3.3.2),

$$\lambda^{\mathrm{T}} y_k = \lambda^{\mathrm{T}} H(k)x_k$$

i.e. certain linear combinations of components of $x_k$ could be measured *exactly*. Thus (3.3.3) says essentially that all observations and linear combinations of observations are 'noisy'.

The sequence $u_k$ is the $m$-vector control input. In this section we suppose that this is a *deterministic sequence*. In future sections we shall wish to consider feedback controls, where $u_k$ depends on the observed outputs $y^k$, but this presents a more delicate situation, consideration of which we defer to Section 6.3 below.

The example considered in the preceding section is a special case of