Or equivalently

$$\hat{\theta}_n = \arg\min_{\theta \in \Lambda}\left[-\sum_{k=1}^{n} \log f[Y_k - S_k(\theta)]\right]$$

and the likelihood equation is thus

$$\sum_{k=1}^{n} S_k'(\hat{\theta}_n)\,\psi[Y_k - S_k(\hat{\theta}_n)] = 0$$

where $\psi \triangleq -f'/f$, $f'(x) \triangleq df(x)/dx$

$$S_k'(\theta) \triangleq \partial S_k(\theta)/\partial\theta .$$

When f is a $\mathcal{N}(0, \sigma^2)$, then

$$\hat{\theta}_n = \arg\left[\min_{\theta \in \Lambda} \sum_{k=1}^{n} [Y_k - S_k(\theta)]^2\right] \qquad (*)$$

and $$\sum_{k=1}^{n} S_k'(\hat{\theta}_n)\,[Y_k - S_k(\hat{\theta}_n)] = 0. \qquad (**)$$

The particular estimator (*) is sometimes known as the <u>least-square estimator</u> of θ. Since it chooses that the value of θ for which $\{S_k(\theta)\}_{k=1}^{n}$ is the least-square fit to the data. That is, it chooses θ to minimize the sum of the squared errors between the data and the signal that arises from that choice of θ.

Solutions to the likelihood equation (**) can have asymptotic properties similar to those for MLEs in i.i.d. models, but may depend on the time variation of the signal. A simple example will be the case when the signal becomes identically zero (or independent of θ) after some finite number of samples, it would be unrealistic to expect consistency.

Consider the observation model $Y_k = S_k(\theta) + N_k$ from the beginning of this section. Then the likelihood equation (**) becomes

$$\sum_{k=1}^{n} s_k'(\hat{\theta}_n)\, N_k + \sum_{k=1}^{n} s_k'(\hat{\theta}_n)\left[s_k(\theta) - s_k(\hat{\theta}_n)\right] = 0. \quad (***)$$

To analyse the behavior of $\hat{\theta}_n$, let us consider for each $\theta' \in \Lambda$ the sequence of random variables

$$J_n(\theta;\theta') \triangleq \sum_{k=1}^{n} s_k'(\theta') N_k + \sum_{k=1}^{n} s_k'(\theta')\left[s_k(\theta) - s_k(\theta')\right].$$

Clearly, in the absense of noise, i.e., $N_k = 0$, $\hat{\theta}_n = \theta$ is a solution to the likelihood equation (***).

Let

$$K_n(\theta;\theta') \triangleq \sum_{k=1}^{n} s_k'(\theta')\left[s_k(\theta) - s_k(\theta')\right],$$

and $d_n > 0$ be a sequence of constants. Then,

$$\frac{1}{d_n} J_n(\theta;\theta') \sim \mathcal{N}\left(\frac{1}{d_n} K_n(\theta;\theta'),\ \frac{\sigma^2}{d_n^2} \sum_{k=1}^{n} [s_k'(\theta')]^2\right).$$

Thus, for given $\theta, \theta' \in \Lambda$, $\frac{1}{d_n} J_n(\theta;\theta')$ converges in probability to a constant if and only if

$$\lim_{n\to\infty} \frac{1}{d_n^2} \sum_{k=1}^{n} [s_k'(\theta)]^2 = 0 \qquad (****)$$

and $\lim_{n\to\infty} \frac{1}{d_n} K_n(\theta;\theta')$ exists. (*****)

* Proposition 4.5.1: Consistency of Least Squares

Suppose that there exists a sequence of scalars $\{d_n\}_{n=1}^{\infty}$ such that the two properties (***) and (****) hold for all $\theta' \in \Lambda$. Suppose further that $s_k(\theta')$, $s_k'(\theta')$, and

$$J(\theta;\theta') \triangleq \lim_{n\to\infty} \frac{1}{d_n} K_n(\theta;\theta')$$

are all continuous functions of θ', and that $J(\theta,\theta')$ has a unique root at $\theta'=\theta$. Then, with probability tending to 1, the likelihood equation (**) has a sequence of roots converging in probability to θ: if $\hat{\theta}_n$ is the unique root of (**) for each n, then $\hat{\theta}_n \to \theta$ in probability (i.p.).

The proof of this result is the same as before.

Let us see an example: Consider the problem of signal amplitude estimation (see Example 4.4.2), in which

$$s_k(\theta) = \theta s_k, \quad k=1,2,\cdots,n,$$

for a known sequence $\{s_k\}_{k=1}^{n}$. In this case, $s_k'(\theta) = s_k$. Thus

$$\sum_{k=1}^{n} [s_k'(\theta)]^2 = \sum_{k=1}^{n} s_k^2$$

and

$$K_n(\theta;\theta') = (\theta-\theta')\sum_{k=1}^{n} s_k^2 .$$

A sufficient condition for consistency following from the proposition is the existence of a sequence $\{d_n\}_{n=1}^{\infty}$ such that

$$0 < \lim_{n\to\infty} \frac{1}{d_n}\sum_{k=1}^{n} s_k^2 < \infty .$$

* Asymptotic normality can also be assured for the above least-squares estimate under regularity conditions on the signal sequence. If $s_k(\theta)$ has third derivatives, the likelihood equation can be expanded in a Taylor series about θ, to give

$$\sum_{k=1}^{n} s_k'(\theta)\,[Y_k - s_k(\theta)]$$

$$+(\hat{\theta}_n-\theta)\sum_{k=1}^{n}[s_k''(\theta)[Y_k-s_k(\theta)] - [s_k'(\theta)]^2]$$

$$+\frac{1}{2}(\hat{\theta}_n-\theta)^2\sum_{k=1}^{n}[s_k'''(\bar{\theta}_n)[Y_k-s_k(\bar{\theta}_k)]-3s_k''(\bar{\theta}_n)s_k'(\bar{\theta}_n)]$$

$$=0$$

where $\bar{\theta}_n$ is between θ and $\hat{\theta}_n$. After rearranging, we have

$$\hat{\theta}_n-\theta = \frac{-\sum_{k=1}^{n} s_k'(\theta)\,N_k}{\sum_{k=1}^{n} s_k''(\theta)N_k - \sum_{k=1}^{n}[s_k'(\theta)]^2 + \frac{1}{2}(\hat{\theta}_n-\theta)\sum_{k=1}^{n} z_k(\bar{\theta}_n)}$$

where

$$z_k(\theta') \triangleq [s_k'''(\theta')[N_k + s_k(\theta) - s_k(\theta')]] - 3s_k''(\theta')\,s_k'(\theta')].$$

* **Proposition 4.5.2: Asymptotic Normality of Least Squares**

Suppose that we have the model of $Y_k = S_k(\theta) + N_k$, $k=1, \cdots, n$, with $\mathcal{N}(0, \sigma^2)$ noise, and $\{\hat{\theta}_n\}_{n=1}^{\infty}$ is a consistent sequence of least-squares estimates of θ. Suppose further that the following regularity condition hold:

1) There exists a function M such that $|Z_k(\theta')| \le M(N_k)$ uniformly in θ', and $E_\theta\{M(N_k)\} < \infty$. [The existence of therelevant derivatives of $S_k(\theta)$ is also assumed].

2) $\lim\limits_{n\to\infty} \frac{1}{n} \sum\limits_{k=1}^{n} [s_k'(\theta)]^2 > 0$

3) $\lim\limits_{n\to\infty} \sum_{k=1}^{n} [s_k''(\theta)]^2 \Big/ \Big[\sum\limits_{k=1}^{n} [s_k'(\theta)]^2\Big]^2 = 0$.

Then,

$$\Big[\sum_{k=1}^{n} [s_k'(\theta)]^2\Big]^{\frac{1}{2}} (\hat{\theta}_n - \theta) \to \mathcal{N}(0, \sigma^2)$$

in distribution.

The proof is similar to that for the i.i.d. case. Note that Fisher's information is

$$I_\theta = \sum_{k=1}^{n} [s_k'(\theta)]^2 / \sigma^2 .$$

Thus, in the same sense as in the i.i.d. case, the least-squares estimate is asymptotically efficient.

The signal-amplitude estimation problem, $S_k(\theta) = \theta S_k$, again provides a straightforward example. In this case, $Z_k(\theta') = 0$ and $S_k''(\theta) = 0$. Thus, the only condition needed for asymptotic normality is

$$\lim_{n\to\infty} \sum_{k=1}^{n} s_k^2 / n > 0.$$

A less obvious example is given below.

* Example 4.5.1: Identification of a First-order Linear System

An important class of applications of parameter estimation problem is in the contest of <u>system identification</u>, in which we wish to infer the structure of some input/output system by putting in an input and observing the output.

One of the simplest possible identification problem is that of identifying a stable first-order time-invariant linear system:

$$S_k(\theta) = \theta S_{k-1}(\theta) + u_k, \quad k = 1, 2, \cdots, n, \quad (*)$$

where $|\theta| < 1$ and $\{u_k\}_{k=1}^{n}$ is the known input sequence. The observation of the system output is usually corrupted by measurement noise. Assuming that this noise is i.i.d., the estimation of θ is a problem in the form

$$Y_k = S_k(\theta) + N_k, \quad k = 1, 2, \cdots, n,$$

Assume that in (*), the system is initially at rest, i.e., $s_0(\theta) = 0$. Then

$$s_k(\theta) = \sum_{l=1}^{n} \theta^{k-1} u_l .$$

Whether or not θ can be identified (as $n \to \infty$) depends on the input sequence $\{u_k\}_{k=1}^{n}$.

Consider, for example, a constant input signal $u_k = 1$ for all $k \geq 1$. The output is then

$$s_k(\theta) = \sum_{l=1}^{k} \theta^{l-1} = \frac{1-\theta^k}{1-\theta}$$

and $s_k'(\theta) = \dfrac{(1-\theta^k) - k\theta^{k-1}(1-\theta)}{1-\theta^2}$.

This implies $\displaystyle\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} [s_k'(\theta)]^2 = \frac{(2-\theta)^2}{(1-\theta)^4}$

and

$$\lim_{n\to\infty} \left[\frac{1}{n} \sum_{k=1}^{n} [s_k'(\theta') [s_k(\theta) - s_k(\theta')] \right] = \frac{(2-\theta')(\theta-\theta')}{(1-\theta')^2(1-\theta)}$$

that has a unique root $\theta' = \theta$.

The relevant quantities are continuous for $|\theta'| < 1$. Thus, the conditions of Prop 4.5.1 are satisfied with $d_n = n$. Thus, we have a consistent sequence of roots to the likelihood equation.

For the constant input signal, Prop. 4.5.2 can not be applied directly to this model with $\Lambda = (-1, 1)$

because $Z_k(\theta')$ cannot be uniformly bounded on this set. However, if we assume θ is bounded away from unity, i.e., if we take $\Lambda = (-1, \theta_u)$ with $\theta_u < 1$, then the regularity conditions of Prop. 4.4.4 do hold, and asymptotic normality and efficiency of the consistent roots of the likelihood equation follow. Note that the asymptotic variance of $\sqrt{n}(\hat{\theta}_n - \theta)$ in this case is $\sigma^2 (1-\theta)^4 / (2-\theta)^2$.

* Proposition 4.5.3: Consistency and Asymptotic Normality of Least-Squares with Non-Gaussian Noise

Propositions 4.5.1 and 4.5.2 remain valid if the assumption $N_k \sim \mathcal{N}(0, \sigma^2)$ is replaced by the assumption $E\{N_k\} = 0$ and $E\{N_k^2\} = \sigma^2 < \infty$.

Note, however, that this result does not imply that least squares is asymptotically efficient when the noise is not Gaussian, since Fisher's information is no longer given by

$$I_\theta = \sum_{k=1}^{n} [s_k'(\theta)]^2 / \sigma^2$$

in the non-Gaussian case.

§4.5.3: Robust Estimation of Signal Parameters

Consider again the model $Y_k = S_k(\theta) + N_k$, $k=1,2,\cdots n$, in which we have noted the MLEs are asymptotically optimum in the sense of minimum asymptotic variance.

As we discussed in Section 3.5, statistical models are only approximately valid in practice, and an important question is whether or not procedures designed for a particular model are robust, i.e., whether their performance is insensitive to small changes in the model.

Consider, for example, a nominal model in which the noise samples have the $\mathcal{N}(0,1)$ distribution. Then, within regularity, and assuming that

$$e_\theta \triangleq \lim_{n\to\infty} \sum_{k=1}^{n} [S_k'(\theta)]^2/n$$

exists and is positive, the least-squares estimate is asymptotically $\mathcal{N}(\theta, \frac{1}{ne_\theta})$.

Suppose, however, that the actual statistical behavior of the noise is described by a pdf that is only approximately $\mathcal{N}(0,1)$. For example, suppose that the noise density f is of the form

$$f(x) = (1-\varepsilon)\frac{1}{\sqrt{2\pi}} e^{-x^2/2} + \varepsilon h(x), \quad x\in\mathbb{R}, \quad (*)$$

where $h(x)$ is an arbitrary density, symmetric about zero, and with variance

$$\sigma_h^2 \triangleq \int_{-\infty}^{\infty} x^2 h(x)\,dx$$

finite but not bounded. Then, by Prop. 4.5.3, the least-squares estimate will have asymptotic variance

$$V_h^2 \sim \frac{(1-\varepsilon)+\varepsilon\sigma_h^2}{n e_0}.$$

Note that V_h^2 can be arbitrarily large for any $\varepsilon>0$ since σ_h^2 is not bounded. In particular, the worst case asymptotic variance over the class of densities (*) is

$$\sup_h [(1-\varepsilon)+\varepsilon\sigma_h^2] = \infty$$

for any $\varepsilon>0$.

This points to a lack of robustness of the least-squares estimate for situations in which a small fraction of the noise samples may come from a high variance distribution. This may happen, for example, in radar measurements, in which very high-variance impulsive interference may be present in a small fraction ε of the measurements. Observations that are improbably large for a given nominal model are sometimes termed <u>outliers</u>. As in the signal detection problems treated in §3.5, an alternative to asymptotic variance at a nomial model is needed as a design criterion for such situations.

Suppose that the noise density f is an even symmetric function. Consider estimates of θ of the form:

$$\sum_{k=1}^{n} s_k'(\hat{\theta}_n)\,\psi[Y_k - s_k(\hat{\theta}_n)] = 0 \qquad (*)$$

where ψ is a general odd-symmetric function. With $\psi(x) = x$, $(*)$ gives the least-squares estimate, and with $\psi(x) = -f'(x)/f(x)$, $(*)$ gives the MLE. Estimates of this form are known as M-estimates.

Assuming that $0 < e_\theta < \infty$ and within regularity on ψ, f, and $\{s_k(\theta)\}_{k=1}^{\infty}$, it can be shown, using the techniques developed above, that M-estimates are consistent and asymptotically $\mathcal{N}(\theta, V(\psi, f)/(n e_\theta)]$, where

$$V(\psi, f) \triangleq \int \psi^2 f \Big/ \left(\int \psi' f\right)^2$$

with $\psi'(x) = d\psi(x)/dx$.

In view of these properties, one possible way of designing a robust estimator for an uncertainty class $\mathcal{F}$ of noise densities is to seek a function ψ that minimizes the worst case M-estimates variance,

$\sup_{f\in\mathcal{F}} V(\psi, f)$:

$$\min_{\psi} \sup_{f\in\mathcal{F}} V(\psi; f). \qquad (**)$$

This problem has been studied by Huber (1981) for general sets $\mathcal{F}$. Within appropriate conditions, its solution is basically as follows.

Consider the functional

$$I(f) \triangleq \int [(f')^2/f],$$

Let f_L be a density in $\mathcal{F}$ that minimizes $I(f)$ over $\mathcal{F}$:

$$I(f_L) = \min_{f\in\mathcal{F}} I(f).$$

Then, the M-estimate with ψ-function

$$\psi_R(x) = -f_L'(x)/f_L(x)$$

Solves $(**)$.

Note that for any f,

$$V(\psi, f)\big|_{\psi = -f'f} = \frac{1}{I(f)}$$

so that $[n \ell_\theta I(f)]^{-1}$ is the asymptotic variance of the MLE in our model with given f. Fisher's information here is $n \ell_\theta I(f)$. Thus, f_L is the member of $\mathcal{F}$ whose corresponding optimum estimate (the MLE) has the worst optimum performance. For this reason, f_L can be considered a least-favorable density, and the robust M-estimate is the best estimate for this least-favorable model.

The problem $\min_{f \in \mathcal{F}} I(f)$ has been solved for a number of uncertainty models $\mathcal{F}$. For example, for the ε-contaminated $N(0,1)$ model described before, the least favorable density is given by

$$f_L(x) = \begin{cases} (1-\varepsilon)\frac{1}{\sqrt{2\pi}} e^{-x^2/2}, & \text{if } |x| \leq k' \\ (1-\varepsilon) e^{-k'(|x|-k')} \frac{1}{\sqrt{2\pi}} e^{-(k')^2/2}, & \text{if } |x| > k', \end{cases}$$

where k' is a constant given by the solution to

$$(1-\varepsilon)^{-1} = 2\Phi(k') - 1 + \frac{1}{k'}\left(\frac{2}{\pi}\right)^{1/2} e^{-(k')^2/2}.$$

The corresponding robust ψ function is

$$\psi_k(x) = \begin{cases} x, & \text{if } |x| \leq k' \\ k' \operatorname{sgn}(x) & \text{if } |x| > k'. \end{cases}$$

Thus, as in the analogous hypothesis testing problem, robustness is brought about by limiting the effects of outliers.

§4.5.4. Recursive Parameter Estimation

Athough MLE have good performance, they have the disadvantages of being cumbersome to compute. For example, with n i.i.d. samples drawn from the density f_θ, computation of the MLE requires the maximization of the function $\sum_{k=1}^{n} \log f_\theta(y_k)$.

Unless the maximizing can be found as a closed-form function of $\underline{y}$, an iterative technique must be used to find $\hat{\theta}_{ML}(\underline{y})$. This requires the storage and simultaneous manipulation of all n samples, a task that is undesirable if n is very large. It is thus sometimes desirable to consider alternatives to maximum likelihood that can be implemented in a recursive or sequential manner so that the contribution of each sample to the estimate is computed as the sample is taken.

Consider a consistent sequence $\{\hat{\theta}_n\}_{n=1}^{\infty}$ solving the likelihood equation $\sum_{k=1}^{n} \psi(Y_k;\hat{\theta}_n) = 0$

with $\psi(Y_k;\theta) = \partial \log f_\theta(Y_k)/\partial\theta$, as before. Since $\{\hat{\theta}_n\}_{n=1}^{\infty}$ is consistent, the difference, $\hat{\theta}_n - \hat{\theta}_{n-1}$ converges to zero as $n \to \infty$. Thus, the above equation can be approximated by expanding about $\hat{\theta}_{n-1}$ to give

$$\sum_{k=1}^{n} \psi(Y_k;\hat{\theta}_{n-1}) + (\hat{\theta}_n - \hat{\theta}_{n-1}) \sum_{k=1}^{n} \psi'(Y_k;\hat{\theta}_{n-1}) \sim 0$$

with $\psi'(Y_k;\theta) = \partial\psi(Y_k;\theta)/\partial\theta$. Thus,

$$\tilde{\theta}_n \sim \tilde{\theta}_{n-1} - \frac{\Sigma_{k=1}^{n} \psi(Y_k;\hat{\theta}_{n-1})}{\Sigma_{k=1}^{n} \psi'(Y_k;\hat{\theta}_{n-1})} .$$

Since $\hat{\theta}_{n-1}$ solves $\Sigma_{k=1}^{n-1} \psi(Y_k;\hat{\theta}_{n-1}) = 0$,

$$\tilde{\theta}_n \sim \tilde{\theta}_{n-1} - \frac{\psi(Y_n;\hat{\theta}_{n-1})}{\Sigma_{k=1}^{n} \psi'(Y_k;\hat{\theta}_{n-1})} .$$

Now, the weak law of large numbers implies that

$$-\frac{1}{n}\sum_{k=1}^{n} \psi'(Y_k;\theta) \rightarrow i_\theta \text{ in probability}$$

where $i_\theta = -E_\theta\{\psi'(Y_k;\theta)\} = E\{\psi^2(Y_k;\theta)\}$ is Fisher's information per sample. Since $\tilde{\theta}_{n-1} \rightarrow \theta$, we can approximate

$$\frac{1}{n}\sum_{k=1}^{n} \psi'(Y_k;\tilde{\theta}_{n-1}) \sim i_{\hat{\theta}_{n-1}} .$$

Thus, $$\tilde{\theta}_n \sim \tilde{\theta}_{n-1} + \frac{\psi(Y_n;\hat{\theta}_{n-1})}{n\, i_{\hat{\theta}_{n-1}}} . \qquad (*)$$

This is an asymptotic recursive equation for $\hat{\theta}_n$, since $\hat{\theta}_n$ is computed from $\hat{\theta}_{n-1}$ and Y_n only.

It turns out that the recursion

$$\hat{\theta}_n = \hat{\theta}_{n-1} + \frac{\psi(Y_n;\hat{\theta}_{n-1})}{n\, i_{\hat{\theta}_{n-1}}} , \quad n=1,2,\cdots$$

(with $\hat{\theta}_0$ arbitrary) suggested by (*) has the same desirable asymptotic properties (i.e., consistency and efficiency) as the MLE within regularity on the model.

This recursion is an example of a more general class of recursive parameter estimation algorithms known as stochastic approximation algorithm. Because of the recursive nature. it has on-line or real-time applications.

Exercises

7. Suppose Θ is uniformly distributed on the interval (0, 1) and that we observe $Y = N + \Theta$ where N is a random variable, independent of Θ, with density

$$p_N(n) = \begin{cases} e^{-n}, & n \geq 0 \\ 0, & n < 0. \end{cases}$$

Find $\hat{\theta}_{MMSE}, \hat{\theta}_{ABS}$, and $\hat{\theta}_{MAP}$.

8. (a) Consider the observation model of Exercise 7 but with the prior of Exercises 5 (i.e., N and Θ both have the unit exponential distribution). Find the MMSE and MMAE estimates of Θ based on Y.

(b) Find the minimum mean-squared error for (a).

(c) Consider now the observation model

$$Y_k = N_k + \Theta, \quad k = 1, \ldots, n,$$

where $N_1, N_2, \ldots, N_n$, and Θ are i.i.d random variables with the unit exponential distribution. Find the MAP estimate of Θ based on $Y_1, Y_2, \ldots, Y_n$.

9. Repeat Exercise 1 for the situation in which we have a sequence of observations $Y_1, Y_2, \ldots, Y_n$, that are conditionally i.i.d. with the given pdf p_θ given $\Theta = \theta$.

10. Derive Eq. (IV.B.47).

11. Suppose that we observe a sequence

$$Y_k = X_k + N_k, \quad k = 1, \ldots, n$$

where $N_1, \ldots, N_n$ is a sequence of independent Gaussian random variables, each with zero mean and variance σ^2, and $X_1, \ldots, X_n$ are defined by the equations

$$X_0 = \Theta$$

$$X_k = \alpha X_{k-1}, \quad k = 1, \ldots, n$$

where α is known and Θ is a Gaussian random parameter with zero mean and variance q^2.

(a) Assuming that Θ and $\underline{N}$ are independent, find the MMSE estimate of Θ based on $Y_1, \ldots, Y_n$.

(b) For each $n = 1, 2, \ldots$, let $\hat{\theta}_n$ denote the MMSE estimate of Θ based on $Y_1, \ldots, Y_n$. Show that $\hat{\theta}_n$ can be computed recursively by

$$\hat{\theta}_n = K_n^{-1}[K_{n-1}\hat{\theta}_{n-1} + \alpha^n y_n], \quad n = 1, 2, \ldots,$$

where $\hat{\theta}_0 = 0$ and the coefficients K_n are defined by

$$K_0 = \sigma^2/q^2 \quad \text{and} \quad K_n = K_{n-1} + \alpha^{2n}, \quad n = 1, 2, \ldots.$$

Draw a block diagram of this implementation.

(c) Find an expression for the mean-squared error

$$e_n = E\{(\hat{\theta}_n - \Theta)^2\}, \quad n = 1, 2, \ldots.$$

What happens when $n \to \infty; q^2 \to \infty; \sigma^2 \to 0; \alpha < 1; \alpha = 1; \alpha > 1$?

12. Suppose θ is a nonrandom parameter satisfying $\theta > 1$. Suppose further that, given $\theta, Y_1, Y_2, \ldots, Y_n$ are i.i.d. observations with each density

$$f_\theta(y) = \begin{cases} (\theta - 1)y^{-\theta}, & y \geq 1 \\ 0, & y \leq 1. \end{cases}$$

Find a sufficient statistic for θ that has a complete family of distributions. Justify your answer.

13. Suppose we toss a coin n independent times and define an observation sequence

$$Y_k = \begin{cases} 1 & \text{if the } k\text{th outcome is heads} \\ 0 & \text{if the } k\text{th outcome is tails} \end{cases}$$

$k = 1, 2, \ldots, n$. Let $\theta = P(Y_k = 1), k = 1, \ldots, n$.

(a) Find an MVUE of θ.

(b) Find the ML estimate of θ. Find its bias and variance.

(c) Compute the Cramér-Rao lower bound and compare with results from (a) and (b).

14. Derive Eq. (IV.D.22).

15. Suppose Y is Poisson. Find the ML estimate of its rate. Compute the bias, variance, and Cramér-Rao lower bound.

20. Consider the observation model

$$Y_k = \theta^{1/2} s_k R_k + N_k, \quad k = 1, 2, \ldots, n$$

where $s_1, s_2, \ldots, s_n$ is a known signal, $N_1, N_2, \ldots, N_n, R_1, R_2, \ldots, R_n$ are i.i.d. $\mathcal{N}(0, 1)$ random variables, and $\theta \geq 0$ is an unknown parameter.

(a) Find the likelihood equation for estimating θ from $Y_1, Y_2, \ldots, Y_n$.

(b) Find the Cramér-Rao lower bound on the variance of unbiased estimates of θ.

(c) Suppose $s_1, s_2, \ldots, s_n$ is a sequence of $+1$'s and -1's. Find the MLE of θ explicitly.

(d) Compute the bias and variance of your estimate from (c), and compare the latter with the Cramér-Rao lower bound.

Chapter 5. Elements of Signal Estimation

§5.1 Introduction

What we talked about in the previous chapter was for static parameters, i.e., parameters do not change with time. In many applications, the parameters we are interested in estimating do change with time. These timedependent parameters are called signals and the parameter estimation is then known as signal estimation or tracking.

One example is in radar systems to track targets as they move through the radar's scanning area. In this case, the radar needs to estimate the position of the target (and perhaps its velocity) at successive times. Since the targets of interest are usually moving and the position measurements are noisy, this is a signal estimation problem.

Another application is analog communications, in which analog information, e.g. audio or video, is transmitted by modulating the amplitude, freq. or phase of a sinusoidal carrier. The receiver is to determine the transmitted information with as high a fidelity as possible on the basis of a noisy observation of the received waveform.

Again, since the transmitted information is time Uarying, this is signal estimation.

§5.2. Kalman — Bucy Filtering

* Model:

Many time-varying ph sical problems/phenomena of interest can be modeled as obeying equation of the form

$$\underline{X}_{n+1} = \underline{f}_n(\underline{X}_n, \underline{u}_n), \quad n = 0, 1, \dots, \qquad (5.2.1)$$

whore $\underline{X}_0, \underline{X}_1, \cdots$ is a sequence of vectors in $\mathbb{R}^m$ representing the phenomenon under study; $\underline{u}_0, \underline{u}_1, \cdots$ is a sequence of vectors in $\mathbb{R}^s$ "acting" on $\{\underline{X}_n\}_{n=1}^{\infty}$; $\underline{f}_0, \underline{f}_1, \cdots$ is a sequence of functions (or, in other words, a time-varying function), each of which maps $\mathbb{R}^m \times \mathbb{R}^s$ to $\mathbb{R}^m$.

Equation (5.2.1) is an example of a dynamical system: $\underline{X}_n$ representing the state of the system at time n,
$\underline{u}_n$ representing the input to the system at time n.

A dynamical system is a system having the property that for any fixed times l and k, $\underline{X}_l$

is determined completely from the state at time k, i.e., $\underline{X}_k$ and the inputs from time k up through $l-1$, i.e., $\{\underline{U}_n\}_{n=k}^{l-1}$.

The complete determination of $\{\underline{X}_n\}_{n=1}^{\infty}$ in (5.2.1) requires not only the input sequence $\underline{U}_n$ but also the initial condition $\underline{X}_0$. If the input sequence or the initial condition is random, the states $\underline{X}_0$, $\underline{X}_1, \ldots$ form a sequence of random vectors and the model (5.2.1) is referred to as stochastic system.

Equation (5.2.1) describes the evolution of the states of a system, so it is usually known as the state equation of the system. The system may also have an output sequence $\underline{Z}_0, \underline{Z}_1, \cdots$ of vectors in $\mathbb{R}^k$ given by the output equations:

$$\underline{Z}_n = \underline{h}_n(\underline{X}_n), \quad n=0,1,\cdots \qquad (5.2.2)$$

where $\underline{h}_n$ maps $\mathbb{R}^m$ to $\mathbb{R}^k$. Thus, the overall system is a mapping from the initial condition $\underline{X}_0$ and input sequence $\{\underline{U}_n\}_{n=0}^{\infty}$ to the output sequence $\{\underline{Z}_n\}_{n=1}^{\infty}$.

* Example 5.2.1: One-Dimensional Motion.

Suppose that we wish to model the one-dimensional motion of a particle that is subjected to an acceleration A_t for $t \geq 0$. Note that the position, P_t, and velocity, V_t, of the particle at each time t satisfy the equation

$$V_t = \frac{dP_t}{dt} \quad \text{and} \quad A_t = \frac{dV_t}{dt}.$$

Assume that we look at the position of the particle every Ts seconds, and we wish to write a model of the form (5.2.1) and (5.2.2) describing the particle's motion from observation time to observation time. Assuming that Ts is small, a Taylor series approximation gives

$$P_{(n+1)T_s} \cong P_{nT_s} + T_s V_{nT_s}$$

and $V_{(n+1)T_s} \cong V_{nT_s} + T_s A_{nT_s}$.

From these equations, we find that two states are needed to describe the motion of the particle, namely, position and velocity:

Define $z_n = x_{1,n} = P_{nT_s}$,

$x_{2,n} = V_{nT_s}$,

$u_n = A_{nT_s}$.

Then, the motion can be described approximately by the state equation:

$$\underline{x}_{n+1} = \mathbb{F}\,\underline{x}_n + \mathbb{G}\,u_n, \quad n=0,1,\cdots$$

and the output equation

$$z_n = \mathbb{H}\,\underline{x}_n, \quad n=0,1,\cdots,$$

where $\underline{Z}_n = \begin{pmatrix} z_{1,n} \\ z_{2,n} \end{pmatrix}$,

$$\mathbb{F} = \begin{pmatrix} 1 & T_s \\ 0 & 1 \end{pmatrix}_{2\times2}$$

$$\mathbb{G} = \begin{pmatrix} 0 \\ T_s \end{pmatrix}_{2\times1}$$

$$\mathbb{H} = (1,\ 0)_{1\times2}.$$

Thus, in this case, $m=2$, $s=1$, $k=1$, and $\underline{f}_n$ and $\underline{h}_n$ are:

$$\underline{f}_n(\underline{z}, \underline{u}) = \mathbb{F}\,\underline{z} + \mathbb{G}\,\underline{u}$$

$$\underline{h}_n(\underline{z}) = \mathbb{H}\,\underline{z}.$$

* Estimation Problem: We observe the output of a stochastic system in the presence of observation noise (or measurement noise) upto some time, say t, and we wish to estimate the state of the system at some time u. That is, we have an observation sequence

$$\underline{Y}_n = \underline{Z}_n + \underline{V}_n, \quad n=0,1,\cdots,t, \qquad (5.2.11)$$

for which we wish to estimate $\underline{z}_u$, where the sequence $\underline{V}_0, \underline{V}_1, \cdots$ represents measurement noise, and (5.2.11) is sometimes known as the measurement equation. If $u=t$, this estimation problem known as the filtering problem;

If $u<t$, it is known as the smoothing problem;

If $u > t$, it is known as the *prediction problem.*

The term *state estimation* is applied to all such problems.

State estimation problems occur in many applications. One of them is *track-while-scan* (TWS) radar. Radar measurements of the position of a target are made on each scan of a scanning radar. These measurements are noisy observations of a stochastic system. The radar on each scan would like to estimate the current position of the target and also to predict the position the target will occupy on the next scan. At each scanning time, t, then, a TWS radar estimates states at $u=t$ and $u=t+1$ based on the past observation record of the position of the target.

Other applications of state estimation arise in automatic control systems such as those for aircraft flight control or chemical process control. In flight control, the states of interest are the positional coordinates of the aircraft and also the attitudenal coordinates (roll, pitch, and yaw) describing the angular orientations of the aircraft. The state equation in this case describes the dynamics of the aircraft, and the inputs may consist of both control forces and random forces (such as turbulence) operating on the aircraft.

In chemical process control, the states may be quantities such as temperatures and concentrations of various chemicals, and the state equation describes the dynamics of the chemical reactions involved.

If we adopt the mean-norm-squared-error performance measure $E\{\|\hat{\underline{x}}_u - \underline{x}_u\|^2\}$ for state estimation $\hat{\underline{x}}_u$ in the model above, we know from Chapter 4, case 4.2.4, that the optimum estimate is the conditional mean

$$\hat{\underline{x}}_u = E\{\underline{x}_u \mid \underline{y}_0, \dots, \underline{y}_t\}. \qquad (5.2.12)$$

This will be the same as Chapter 4. However, we are usually interested in producing estimates in real time and as t increases, the data set grows linearly with t, and the above conditional mean estimates will not be practical unless the system model has a structure that makes (5.2.12) computationally efficient. So, we need more structures/restrications on the model (5.2.1), (5.2.2) and (5.2.11).

One such structure/restrication is that the system is a linear stochastic system, i.e., the state and observation equation are of the form:

$$\underline{X}_{n+1} = \mathbb{F}_n \underline{X}_n + \mathbb{G}_n \underline{U}_n, \quad n=0,1,\ldots, \qquad (5.2.13a)$$

$$\underline{Y}_n = \mathbb{H}_n \underline{X}_n + \underline{V}_n, \quad n=0,1,\ldots \qquad (5.2.13b)$$

where, for each n, $\mathbb{F}_n$, $\mathbb{G}_n$, and $\mathbb{H}_n$ are matrices of appropriate dimensions ($m \times m$, $m \times s$, and $k \times m$, respectively).

The above linear model is appropriate for many applications. For example, the one-dimensional model in Example 5.2.1. Also, many non-linear systems can be approximated by linear systems, after using Taylor series expansion of the nonlinearities $\underline{f}_n$.

A further assumption to simplify the estimate (5.2.12) is that the input sequence $\{U_n\}_{n=0}^{\infty}$ and the observation noise $\{V_n\}_{n=0}^{\infty}$ are independent sequences of independent zero-mean Gaussian random vectors. It is also convenient to assume that the initial condition $\underline{X}_0$ is a Gaussian random vector independent of $\{\underline{U}_n\}_{n=0}^{\infty}$ and $\{\underline{V}_n\}_{n=0}^{\infty}$. Some of these assumptions can be relaxed but the Gaussian assumption is crucial, which is reasonable in many applications, such as thermal noise in sensor electronics, etc.

* Proposition 5.2.1: The Discrete-Time Kalman-Bucy Filter

For the linear stochastic system (5.2.13) with $\{\underline{U}_n\}_{n=0}^{\infty}$ and $\{\underline{V}_n\}_{n=0}^{\infty}$ being independent sequences of independent zero-mean Gaussian vectors independent of the Gaussian initial condition $\underline{X}_0$, the estimates $\hat{\underline{X}}_{t|t} \triangleq E\{\underline{X}_t | \underline{Y}_0^t\}$ and $\hat{\underline{X}}_{t+1|t} \triangleq E\{\underline{X}_{t+1} | \underline{Y}_0^t\}$ are given recursively by the following equations:

$$\hat{\underline{X}}_{t|t} = \hat{\underline{X}}_{t|t-1} + \mathbb{K}_t (\underline{Y}_t - \mathbb{H}_t \hat{\underline{X}}_{t|t-1}), \quad t=0,1,\ldots, \qquad (5.2.14a)$$

and

$$\hat{\underline{X}}_{t+1|t} = \mathbb{F}_t \hat{\underline{X}}_{t|t}, \quad t=0,1,\ldots \qquad (5.2.14b)$$

with the initialization $\hat{\underline{X}}_{0|-1} = \underline{m}_0 \triangleq E\{\underline{X}_0\}$,

where

$$\mathbb{K}_t = \Sigma_{t|t-1} \mathbb{H}_t^T (\mathbb{H}_t \Sigma_{t|t-1} \mathbb{H}_t^T + \mathbb{R}_t)^{-1} \qquad (5.2.15)$$

with $\Sigma_{t|t-1} \triangleq \mathrm{Cov}(\underline{X}_t | \underline{Y}_0^{t-1})$ and

$\mathbb{R}_t \triangleq \mathrm{Cov}(\underline{V}_t)$,

$\underline{Y}_a^b$ denotes the set $\underline{Y}_a, \ldots, \underline{Y}_b$ for $b \geq a$.

Note that, since $\hat{\underline{X}}_{t|t-1} = E\{\underline{X}_t | \underline{Y}_0^{t-1}\}$, $\Sigma_{t|t-1}$ is the covariance matrix of the prediction error, $\underline{X}_t - \hat{\underline{X}}_{t|t-1}$, conditioned on $\underline{Y}_0^{t-1}$. This matrix can be computed jointly with the filtering error covariance, $\Sigma_{t|t} \triangleq Cov(\underline{X}_t | \underline{Y}_0^t)$ from the following recursion:

$$\Sigma_{t|t} = \Sigma_{t|t-1} - \mathbb{K}_t \mathbb{H}_t \Sigma_{t|t-1}, \quad t=0,1,\cdots, \quad (5.2.16a)$$

$$\Sigma_{t+1|t} = \mathbb{F}_t \Sigma_{t|t} \mathbb{F}_t^T + \mathbb{G}_t \mathbb{Q}_t \mathbb{G}_t^T, \quad t=0,1,\cdots, \quad (5.2.16b)$$

with the initialization $\Sigma_{0|-1} = \Sigma_0 \triangleq Cov(\underline{X}_0)$, where $\mathbb{Q}_t$ is the covariance matrix of the tth state input, $\mathbb{Q}_t \triangleq Cov(\underline{U}_t)$.

Proof: To prove the proposition, we first show (5.2.14b) and (5.2.16b) directly, and then prove (5.2.14a) and (5.2.16a) by induction.

To prove (5.2.14b), from the state equation,

$$\begin{aligned}\hat{\underline{X}}_{t+1|t} &= E\{\underline{X}_{t+1} | \underline{Y}_0^t\} = E\{\mathbb{F}_t \underline{X}_t + \mathbb{G}_t \underline{U}_t | \underline{Y}_0^t\} \\ &= \mathbb{F}_t E\{\underline{X}_t | \underline{Y}_0^t\} + \mathbb{G}_t E\{\underline{U}_t | \underline{Y}_0^t\} \\ &= \mathbb{F}_t \hat{\underline{X}}_{t|t} + \mathbb{G}_t E\{\underline{U}_t | \underline{Y}_0^t\}.\end{aligned}$$

Note that $\underline{Y}_0^t$ is determined by $\underline{X}_0^t$ and $\underline{V}_0^t$ or in turn by $\underline{X}_0$, $\underline{U}_0^{t-1}$, and $\underline{V}_0^t$, all of which are independent of $\underline{U}_t$. Thus,

$$E\{\underline{U}_t | \underline{Y}_0^t\} = E\{\underline{U}_t\} = \underline{0}.$$

Therefore, (5.2.14b) holds.

Similarly,

$$\Sigma_{t+1|t} = \mathrm{Cov}(\underline{X}_{t+1} \,|\, \underline{Y}_0^t)$$
$$= \mathrm{Cov}(\mathbb{F}_t \underline{X}_t + \mathbb{G}_t \underline{U}_t \,|\, \underline{Y}_0^t)$$
$$= \mathrm{Cov}(\mathbb{F}_t \underline{X}_t \,|\, \underline{Y}_0^t) + \mathrm{Cov}(\mathbb{G}_t \underline{U}_t \,|\, \underline{Y}_0^t)$$
$$= \mathrm{Cov}(\mathbb{F}_t \underline{X}_t \,|\, \underline{Y}_0^t) + \mathrm{Cov}(\mathbb{G}_t \underline{U}_t),$$

Since $\underline{U}_t$ is independent of $\underline{X}_t$ and $\underline{Y}_0^t$.

Using the property $\mathrm{Cov}(A\underline{X}) = A\,\mathrm{Cov}(\underline{X})\,A^T$, and the definition $\Sigma_{t|t}$ and $\mathbb{Q}_t$, we have

$$\Sigma_{t+1|t} = \mathbb{F}_t \,\mathrm{Cov}(\underline{X}_t | \underline{Y}_0^t)\, \mathbb{F}_t^T + \mathbb{G}_t \,\mathrm{Cov}(\underline{U}_t)\, \mathbb{G}_t^T$$
$$= \mathbb{F}_t \Sigma_{t|t} \mathbb{F}_t^T + \mathbb{G}_t \mathbb{Q}_t \mathbb{G}_t^T,$$

which is (5.2.16b).

We next use induction to show (5.2.14a) and (5.2.16a). For $t=0$, $\underline{Y}_0 = \mathbb{H}_0 \underline{X}_0 + \underline{V}_0$. Since $\underline{X}_0$ and $\underline{V}_0$ are independent Gaussian vectors, we see that the estimation of $\underline{X}_0$ from $\underline{Y}_0$ fits the linear estimation model discussed as Example 4.2.3. In particular, since $\underline{X}_0 \sim \mathcal{N}(\underline{m}_0, \Sigma_0)$ and $\underline{V}_0 \sim \mathcal{N}(\underline{0}, \mathbb{R}_0)$, from (4.2.53), we have

$$\hat{\underline{X}}_{0|0} \triangleq E\{\underline{X}_0 \,|\, \underline{Y}_0\}$$

$$= \underline{m}_0 + \Sigma_0 \mathbb{H}_0^T (\mathbb{H}_0 \Sigma_0 \mathbb{H}_0^T + \mathbb{R}_0)^{-1} (\underline{Y}_0 - \mathbb{H}_0 \underline{m}_0)$$

$$= \hat{\underline{X}}_{0|-1} + \mathbb{K}_0 (\underline{Y}_0 - \mathbb{H}_0 \hat{\underline{X}}_{0|-1})$$

which is (5.2.14a) when $t=0$. The error covariance is given from (4.2.54) as

$$\Sigma_{0|0} = \Sigma_0 - \Sigma_0 \mathbb{H}_0^T (\mathbb{H}_0 \Sigma_0 \mathbb{H}_0^T + \mathbb{R}_0)^{-1} \mathbb{H}_0 \Sigma_0$$

$$= \Sigma_{0|-1} - \mathbb{K}_0 \mathbb{H}_0 \Sigma_{0|-1},$$

which is (5.2.16a) for $t=0$.

To complete the proof, we now assume that (5.2.14a) and (5.2.16a) hold for $t = t_0 - 1$. Note that $\underline{X}_{t_0}$ and $\underline{Y}_0^{t_0-1}$ are derived by linear transformations of the Gaussian vectors $\underline{X}_0$, $\underline{U}_0^{t_0-1}$, and $\underline{V}_0^{t_0-1}$. This implies that $\underline{X}_{t_0}$ and $\underline{Y}_0^{t_0-1}$ are jointly Gaussian and thus that $\underline{X}_{t_0}$ is conditionally Gaussian given $\underline{Y}_0^{t_0-1}$, and the conditional distribution of $\underline{X}_{t_0}$ given $\underline{Y}_0^{t_0-1}$ is $N(\hat{\underline{X}}_{t_0|t_0-1}, \Sigma_{t_0|t_0-1})$.

Also note that $\underline{V}_{t_0}$ is Gaussian and independent of $\underline{Y}_0^{t_0-1}$, so it is also conditionally Gaussian given $\underline{Y}_0^{t_0-1}$ with distribution $N(\underline{0}, \mathbb{R}_{t_0})$. Since $\underline{V}_{t_0}$ is independent of all $\underline{X}_0$, $\underline{V}_0^{t_0-1}$, and $\underline{U}_0^{t_0-1}$, it is conditionally independent of $\underline{X}_{t_0}$ given $\underline{Y}_0^{t_0-1}$. Thus, given $\underline{Y}_0^{t_0-1}$,

$\underline{Y}_{t_0} = \mathbb{H}_{t_0}\underline{X}_{t_0} + \underline{V}_{t_0}$ is a Gaussian linear equation of the form disscussed in Example 4.2.3.

Let us compute the conditional expectation of $\underline{X}_{t_0}$ given $\underline{Y}_{t_0}$ under the above conditional model given $\underline{Y}_0^{t_0-1}$, it is $\hat{\underline{X}}_{t_0|t_0}$. Thus, from (4.2.53),

$$\hat{\underline{X}}_{t_0|t_0} = \hat{\underline{X}}_{t_0|t_0-1} + \Sigma_{t_0|t_0-1}\mathbb{H}_{t_0}^T(\mathbb{H}_{t_0}\Sigma_{t_0|t_0-1}\mathbb{H}_{t_0}^T + \mathbb{R}_{t_0})^{-1} \times (\underline{Y}_{t_0} - \mathbb{H}_{t_0}\hat{\underline{X}}_{t_0|t_0-1}),$$

which is (5.2.14a) for $t = t_0$ with the definition $\mathbb{K}_{t_0}$.

Similarly, by applying (4.2.54) and the argument above, we arrive at (5.2.16a) with $t = t_0$. This proves the proposition □.

The estimator structure described by Prop. 5.2.1 is known as the discrete-time Kalman-Bucy filter, because it is the discrete-time version of a continuous-time recursive state estimator developed principally by R.E. Kalman and R.S. Bucy in the late 1950s.

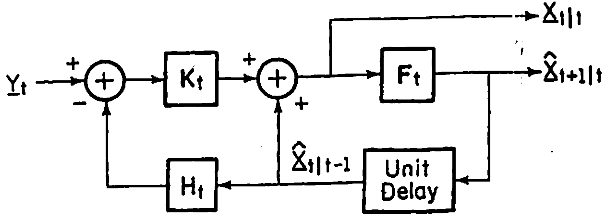

FIGURE V.B.1. The discrete-time Kalman-Bucy filter

In the Kalman-Bucy filter, although the estimators $\underline{\hat{X}}_{t+1|t}$ or $\underline{\hat{X}}_{t|t}$ depends on all the data $\underline{Y}_0^t$, they are computed at each stage from only the latest observation $\underline{Y}_t$, and the previous prediction $\underline{\hat{X}}_{t|t-1}$. Thus, rather than having to store the $(t+1)$ k-dimensional vectors $\underline{Y}_0^t$ (and hence having a linearly growing memory and computational burden), we need only to store and update the single m-vector $\underline{\hat{X}}_{t|t-1}$. All other parts of the estimator (including the Kalman gain matrix $\mathbb{K}_t$) are determined completely from the parameters of the model and are independent of the data.

The recursions (5.2.14) and (5.2.16) each consists of two basic steps: The first of these steps is the measurement update, (5.2.14a) and (5.2.16a), which provides the means of updating the estimate and covariance of $\underline{X}_t$ given $\underline{Y}_0^{t-1}$ to incorporate the new observation $\underline{Y}_t$

The second basic step is the time update, (5.2.14b), (5.2.16b), which provides the means for projecting the state estimate and covariance based on the observation $\underline{Y}_0^t$ to the next time $(t+1)$ before the $(t+1)$st measurement is taken.

Consider the measurement update equation

(5.2.14a) further: The estimate $\hat{\underline{X}}_{t|t}$, which is the best estimate of $\underline{X}_t$ based on $\underline{Y}_0^t$, can be viewed as the combination of the best estimate of $\underline{X}_t$ based on the past prediction $\hat{\underline{X}}_{t|t-1}$, and a correction term, $\mathbb{K}_t(\underline{Y}_t - \mathbb{H}_t\hat{\underline{X}}_{t|t-1})$. The vector $\underline{I}_t \triangleq (\underline{Y}_t - \mathbb{H}_t\hat{\underline{X}}_{t|t-1})$ appearing in the correction term has an interesting interpretation. Since $\underline{Y}_t = \mathbb{H}_t\underline{X}_t + \underline{V}_t$, we have

$$\begin{aligned}\hat{\underline{Y}}_{t|t-1} \triangleq E\{\underline{Y}_t|\underline{Y}_0^{t-1}\} &= \mathbb{H}_t E\{\underline{X}_t|\underline{Y}_0^{t-1}\} \\ &\quad + E\{\underline{V}_t|\underline{Y}_0^{t-1}\} \\ &= \mathbb{H}_t\hat{\underline{X}}_{t|t-1},\end{aligned}$$

where V_t is independent of $\underline{Y}_0^{t-1}$ and thus, $E\{\underline{V}_t|\underline{Y}_0^{t-1}\}=0$. Thus, $\underline{I}_t = \underline{Y}_t - \hat{\underline{Y}}_{t|t-1}$ represents an error signal, i.e., the error in the prediction of $\underline{Y}_t$ from its past $\underline{Y}_0^{t-1}$. This error is sometimes known as the (prediction) *residual* or the *innovation*. The latter term comes from the fact that we can write $\underline{Y}_t$ as

$$\underline{Y}_t = \hat{\underline{Y}}_{t|t-1} + \underline{I}_t,$$

with the interpretation that $\hat{\underline{Y}}_{t|t-1}$ is the part of $\underline{Y}_t$ that can be predicted from the past, and $\underline{I}_t$ is the part of $\underline{Y}_t$ that can not be predicted. Thus, $\underline{I}_t$ contains the *new* information that is gained by taking the tth observation; hence the term "innovation."

It can be shown that the innovation sequence $\{\underline{I}_t\}_{t=0}^{\infty}$ is a sequence of independent zero-mean Gaussian random vectors. First, $\{\underline{I}_t\}$ is a Gaussian sequence since $\{\underline{Y}_t\}$ is a Gaussian sequence and $\{\underline{I}_t\}$ is obtained from linear transformations of $\{\underline{Y}_t\}$.

$$E\{\underline{I}_t\} = E\{\underline{Y}_t - E\{\underline{Y}_t|\underline{Y}_0^{t-1}\}\}$$

$$= E\{\underline{Y}_t\} - E\{\underline{Y}_t\} = 0$$

$(E(Y) = E(E(Y|\underline{Z})))$.

Thus, $Cov(\underline{I}_t, \underline{I}_s) = E\{\underline{I}_t\underline{I}_s^T\}$.

Assuming that $s<t$, we have

$$E\{\underline{I}_t\underline{I}_s^T\} = E\{E\{\underline{I}_t\underline{I}_s^T|\underline{Y}_0^s\}\}$$

$$= E\{E\{I_t|\underline{Y}_0^s\}\underline{I}_s^T\},$$

where $\underline{I}_s$ is constant given $\underline{Y}_0^s$.

Note $E\{\underline{I}_t|\underline{Y}_0^s\} = E\{\underline{Y}_t|\underline{Y}_0^s\}$
$- E\{E\{\underline{Y}_t|\underline{Y}_0^{t-1}\}|\underline{Y}_0^s\}$
$= E\{\underline{Y}_t|\underline{Y}_0^s\} - E\{\underline{Y}_t|\underline{Y}_0^s\} = 0.$

Thus, $E\{\underline{I}_t\underline{I}_s^T\} = 0$, i.e., $Cov(\underline{I}_t, \underline{I}_s) = 0$.

$\{\underline{Y}_t\}$ consists of a part, $\hat{\underline{Y}}_{t|t-1}$, completely dependent on the past, and a part, $\underline{I}_t$, completely independent of the past. This implies that the innovations sequence provides a set of independent observations that is equivantent to the original set $\{Y_t\}_{t=0}^{\infty}$. Thus, the formation of the innovation sequence is a pre whitening operation as discussed in Chapter 3.

* Example 5.2.2: The Time-Invariant Single-Variable Case

The simplest model with which the Kalman filter can be illustrated is the one-dimensional ($m=k=1$) case in which all parameters of the model are independent of time:

$$X_{n+1} = f X_n + U_n, \quad n=0,1,\cdots$$

$$Y_n = h X_n + V_n, \quad n=0,1,\cdots$$

where $\{U_n\}_{n=0}^{\infty}$ and $\{V_n\}_{n=0}^{\infty}$ are independent sequences of i.i.d. $N(0,q)$ and $N(0,r)$ random variables, respectively, $X_0 \sim N(m_0, \Sigma_0)$, where f, h, q, r, m_0, and Σ_0 are all scalars. The estimation recursions for this case are

$$\hat{X}_{t+1|t} = f \hat{X}_{t|t}, \quad t=0,1,\cdots$$

$$\hat{X}_{t|t} = \hat{X}_{t|t-1} + K_t (Y_t - h \hat{X}_{t|t-1}), \quad t=0,1,\cdots$$

with $$K_t = \frac{\Sigma_{t|t-1} h}{(h^2 \Sigma_{t|t-1} + r)} = \frac{1}{h} \frac{\Sigma_{t|t-1}}{\Sigma_{t|t-1} + r/h^2}$$

$$\Sigma_{0|-1} = \Sigma_0, \quad \hat{X}_{0|-1} = m_0.$$

Note that $\Sigma_{t|t-1}$ is the MSE incurred in the estimation of X_t from Y_0^{t-1}, and the ratio r/h^2 is a measure of the "noiseness" of the observations

$$\frac{Y_t}{h} = \gamma_t + \frac{V_t}{h}$$

is an equivalent measurement to Y_t (assuming $h \neq 0$), the variance of V_t/h is r/h^2.

If the previous prediction of γ_t is of much higher quality than the current observation, i.e, $\Sigma_{t|t-1} << r/h^2$, then the gain $K_t \cong 0$ and $\hat{\gamma}_{t|t} \cong \hat{\gamma}_{t|t-1}$. That is, in this case we trust our previous estimate of γ_t much more than we trust our observation, so we retain the former estimate.

In the opposite situation in which the previous estimate is much noiser than our observation, i.e, $\Sigma_{t|t-1} >> r/h^2$, the Kalman gain $K_t \cong 1/h$, and $\hat{\gamma}_{t|t} \cong Y_t/h$. In this case, we simply ignore the previous measurements and invert the current measurement equation.

It is interesting to compare the measurement update here with the Bayesian estimation of signal amplitude as discussed in Example 4.2.2. In particular, we can write the measurement update equation as

$$\hat{\gamma}_{t|t} = \frac{v^2 d^2 \hat{\theta}_1 + \mu}{v^2 d^2 + 1} \qquad (5.2.30)$$

where we have identified $\hat{\theta}_1 = Y_t/h$, $\mu = \hat{\gamma}_{t|t-1}$, $v^2 = \Sigma_{t|t-1}$, and $d^2 = h^2/r$.

Comparing (5.2.30) with

$$\hat{\theta}_{MMSE}(\underline{y}) = \frac{s^T \Sigma^{-1} y + \mu/\nu^2}{d^2 + 1/\nu^2} = \frac{\nu^2 d^2 \hat{\theta}_1(\underline{y}) + \mu}{\nu^2 d^2 + 1}$$

where $\hat{\theta}_1(\underline{y}) = \underline{s}^T \Sigma^{-1} \underline{y} / d^2$, we see that the distribution of X_t conditioned on Y_0^{t-1} can be intepreted as a prior distribution for X_t, $N(\hat{X}_{t|t-1}, \Sigma_{t|t-1})$, and and the update balances this prior knowledge with the knowledge gained by the observation Y_t, according to the value of $\nu^2 d^2$.

For this scalar time-invariant model, the time and measurement updates for the estimation covariance become

$$\Sigma_{t+1|t} = f^2 \Sigma_{t|t} + q$$

$$\Sigma_{t|t} = \frac{\Sigma_{t|t-1}}{\frac{h^2}{r}\Sigma_{t|t-1} + 1}$$

or

$$\Sigma_{t+1|t} = \frac{f^2 \Sigma_{t|t-1}}{h^2 \Sigma_{t|t-1}/r + 1} + q, \quad t = 0, 1, \cdots \qquad (5.2.32)$$

$$\Sigma_{0|-1} = \bar{\Sigma}_0 .$$

A natural question is whether the sequence generated by (5.2.32) approaches a constant as t increases. If so, the Kalman gain approaches a constant also, and the Kalman-Bucy filter becomes time-invariant asymptotically in t.

If $\Sigma_{t+1|t}$ does approach a constant, say Σ_∞, then

Σ_∞ must satisfy

$$\Sigma_\infty = \frac{f^2 \Sigma_\infty}{h^2 \Sigma_\infty / r + 1} + q \qquad (5.2.33)$$

$$\Rightarrow \Sigma_\infty = \frac{1}{2}\left\{\left[\frac{r}{h^2}(1-f^2) - q\right]^2 + \frac{4rq}{h^2}\right\}^{1/2} - \frac{r}{2h^2}(1-f^2) + q.$$

Combining (5.2.32) and (5.2.33),

$$|\Sigma_{t+1|t} - \Sigma_\infty| = f^2 \left| \frac{\Sigma_{t|t-1}}{h^2 \Sigma_{t|t-1}/r + 1} - \frac{\Sigma_\infty}{h^2 \Sigma_\infty / r + 1} \right|$$

$$\overset{(*)}{\leq} f^2 |\Sigma_{t|t-1} - \Sigma_\infty|, \quad t = 0, 1, \cdots$$

(To see (*), let $g(x) = \frac{x}{ax+1}$ with $a = \frac{h^2}{r}$,

$|g(x) - g(y)| = |x-y|\,|g'(\xi)|$, mid-value theorem

for some ξ between x and y. But

$$|g'(\xi)| = \frac{1}{|a\xi+1|^2} \leq 1 \quad \text{when } a\xi > 0.$$

Thus, $|\Sigma_{t+1|t} - \Sigma_\infty| \leq f^{2(t+1)} |\Sigma_0 - \Sigma_\infty|$.

If $|f|<1$, $\Sigma_{t+1|t} \to \Sigma_\infty$ as $t \to \infty$.

Note that $|f|<1$ is also the condition for asymptotic stability of the original system.

* Example 5.2.3: Track-While-Scan (TWS) Radar

A radar scans an air field,
keeps track of the trajectories of targets by
processing position measurements taken
once each scan,
predicts the positions the targets will occupy
on the next scan.

Since the maneuver strategies of the targets are usually unknown to the radar, one way of modeling target motion is ~~to~~ assume that the targets have random accelerations. A simple model is to assume that the accelerations are i.i.d. from scan to scan and are Gaussian.

For simplicity, assume the target motion is one-dimensional, which leads to a state/measurement model of the form described in Example 5.2.1:

$$\begin{pmatrix} P_{n+1} \\ V_{n+1} \end{pmatrix} = \begin{pmatrix} 1 & T_s \\ 0 & 1 \end{pmatrix} \begin{pmatrix} P_n \\ V_n \end{pmatrix} + \begin{pmatrix} 0 \\ T_s \end{pmatrix} A_n \qquad (5.2.37)$$

$$Y_n = (1 \;\; 0) \begin{pmatrix} P_n \\ V_n \end{pmatrix} + \mathcal{E}_n$$

Where P_n and V_n represent the target position and velocity, respectively, on the nth scan, T_s is the time the radar takes to complete each scan, A_n is the target acceleration during the nth scanning period, Y_n is the position measurement at the nth sighting, and ε_n is the error in this measurement.

In general, the problem is 3-dimensional. To track in all three dimensions we would have a six state, three-measurement model. If the accelerations and measurement noises in the three dimensions are independent of one another, the three dimensions can be tracked independently.

Assuming that all statistics are Gaussian and time-invariant, the optimum tracker/predictor equations are

$$\begin{pmatrix} \hat{P}_{t+1|t} \\ \hat{V}_{t+1|t} \end{pmatrix} = \begin{pmatrix} \hat{P}_{t|t} + T_s \hat{V}_{t|t} \\ \hat{V}_{t|t} \end{pmatrix}$$

and

$$\begin{pmatrix} \hat{P}_{t|t} \\ \hat{V}_{t|t} \end{pmatrix} = \begin{pmatrix} \hat{P}_{t|t-1} \\ \hat{V}_{t|t-1} \end{pmatrix} + \begin{pmatrix} K_{t,1} \\ K_{t,2} \end{pmatrix} (Y_t - \hat{P}_{t|t-1}),$$

where

$$\begin{pmatrix} K_{t,1} \\ K_{t,2} \end{pmatrix} = \begin{pmatrix} \Sigma_{t|t-1}(1,1)/(\Sigma_{t|t-1}(1,1)+r) \\ \Sigma_{t|t-1}(2,1)/(\Sigma_{t|t-1}(1,1)+r) \end{pmatrix}$$

$\Sigma_{t|t-1}(k,l)$ is the (k,l)th component of $\Sigma_{t|t-1}$, r is the variance of the measurement noise, $\Sigma_{t|t-1}$ can be computed by (5.2.16).

change Kalman Gains to take parameters to tune for fast convergence

$$\Rightarrow \begin{pmatrix} \hat{P}_{t|t} \\ \hat{V}_{t|t} \end{pmatrix} = \begin{pmatrix} \hat{P}_{t|t-1} \\ \hat{V}_{t|t-1} \end{pmatrix} + \begin{pmatrix} \alpha \\ \beta/T_s \end{pmatrix} (Y_t - \hat{P}_{t|t-1})$$

$\leftarrow$ α & β are parameters to tune to converge faster!

* Returning to the Kalman-Bucy filter in Prop 5.2.1, the couple recursions in each of (5.2.14) and (5.2.16) can be separated:

$$\hat{\underline{X}}_{t+1|t} = \mathbb{F}_t \hat{\underline{X}}_{t|t-1} + \mathbb{F}_t \mathbb{K}_t \underline{I}_t, \quad t=0,1,\ldots \qquad (5.2.42a)$$

and $$\Sigma_{t+1|t} = \mathbb{F}_t \Sigma_{t|t-1} \mathbb{F}_t^T - \mathbb{F}_t \mathbb{K}_t \mathbb{H}_t \Sigma_{t|t-1} \mathbb{F}_t^T + \mathbb{G}_t \mathbb{Q}_t \mathbb{G}_t^T, \quad t=0,1,\ldots, \qquad (5.2.42b)$$

The prediction filter (5.2.42a) is a linear stochastic system driven by the innovation sequence. This system has the same dynamics, i.e., $\mathbb{F}_t$'s, as the system we are trying to track. To track $\underline{X}_t$, we are building a system comprising a duplicate of the dynamics that govern $\underline{X}_t$ and then driving it with the innovations through the matrix sequence $\mathbb{F}_t \mathbb{K}_t$.

The covariance update (5.2.42b) is a dynamical system with a matrix state. This system is nonlinear since $\mathbb{K}_t$ depends on $\Sigma_{t|t-1}$. This equation is known as a (discrete-time) Riccati equation. Similar to the scalar case of Example 5.2.2, a sufficient (not necessary) condition for $\Sigma_{t+1|t}$ to converge to a steady state is that all eigenvalues of $\mathbb{F}$ have less than unit magnitude (in the time-invariant case).

* All of the assumptions in the above are made to derive the Kalman-Bucy filter. Some of them may be too restrictive to use. For example, the independence assumptions on the input and noise sequences $\{\underline{U}_k\}_{k=0}^{\infty}$ and $\{\underline{V}_k\}_{k=0}$ can be relaxed by modeling these processes as themselves being derived from linear stochastic systems driven by independent sequences as shown below from an example.

Example 5.2.4: TWS Radar with Dependent Acceleration Sequences

In the previous example, it is not realistic to assume that the target acceleration is independent from scan to scan. A simple but useful model for acceleration is

$$A_{n+1} = \rho A_n + W_n, \quad n=0,1,\ldots, \qquad (5.2.43)$$

with a Gaussian initial condition A_0 and an i.i.d. Gaussian input sequence $\{W_n\}_{n=0}^{\infty}$, where ρ is a paramater satisfying $0 \le \rho < 1$. Note that if $\rho = 0$, there is no dependence in the acceleration sequence, while large ρ means strong correlation between acceleration in different scans.

With accelerations satisfying (5.2.43), the model (5.2.37) no longer satisfies the assumption required for the Kalman-Bucy filter.

However, we may change the model to include the acceleration dynamics (5.2.43) by treating the acceleration as a state rather than as an input:

$$\begin{pmatrix} P_{n+1} \\ V_{n+1} \\ A_{n+1} \end{pmatrix} = \begin{pmatrix} 1 & T_s & 0 \\ 0 & 1 & T_s \\ 0 & 0 & \rho \end{pmatrix} \begin{pmatrix} P_n \\ V_n \\ A_n \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} W_n, \quad n=0,1,\cdots \qquad (5.2.44a)$$

$$Y_n = (1\,0\,0) \begin{pmatrix} P_n \\ V_n \\ A_n \end{pmatrix} + \varepsilon_n, \quad n=0,1,\cdots \qquad (5.2.44b)$$

$$\Rightarrow \begin{pmatrix} \hat{P}_{t+1|t} \\ \hat{V}_{t+1|t} \\ \hat{A}_{t+1|t} \end{pmatrix} = \begin{pmatrix} \hat{P}_{t|t} + T_s \hat{V}_{t|t} \\ \hat{V}_{t|t} + T_s \hat{A}_{t|t} \\ \rho \hat{A}_{t|t} \end{pmatrix}$$

and
$$\begin{pmatrix} \hat{P}_{t|t} \\ \hat{V}_{t|t} \\ \hat{A}_{t|t} \end{pmatrix} = \begin{pmatrix} \hat{P}_{t|t-1} \\ \hat{V}_{t|t-1} \\ \hat{A}_{t|t-1} \end{pmatrix} + \begin{pmatrix} K_{t,1} \\ K_{t,2} \\ K_{t,3} \end{pmatrix} (Y_t - \hat{P}_{t|t-1})$$

where
$$\begin{pmatrix} K_{t,1} \\ K_{t,2} \\ K_{t,3} \end{pmatrix} = \begin{pmatrix} \Sigma_{t|t-1}(1,1)/(\Sigma_{t|t-1}(1,1) + r) \\ \Sigma_{t|t-1}(2,1)/(\Sigma_{t|t-1}(1,1) + r) \\ \Sigma_{t|t-1}(3,1)/(\Sigma_{t|t-1}(1,1) + r) \end{pmatrix}.$$

In practical tracking systems, the gain vector $(K_{t,1}, K_{t,2}, K_{t,3})^T$ is sometimes replaced by

$(\alpha, \beta/T_s, \gamma/T_s^2)^T$ for some parameters,

$\alpha-\beta-\gamma$ called $\alpha-\beta-\gamma$ tracker. These three parameters α, β, γ are chosen to a given desired performance.

Remark: The Gaussian assumption in the Kalman-Bucy filter can be removed if one wants to optimize among all linear estimators as we shall discuss next.

§5.3 Linear Estimation

If we only consider linear estimations, the Gaussian statistics assumption can be removed, when the second order statistics is given. Then the theory is known as Wiener-Kolmogorov filtering (Wiener Filtering) → Our SSP class.

Suppose that we have two sequences of random variables $\{Y_n\}_{n=-\infty}^{\infty}$ and $\{Z_n\}_{n=-\infty}^{\infty}$. We observe Y_n some set of times $a \le n \le b$ and we wish to estimate Z_t from these observations for some particular time t. Of course, the optimum estimator (in the MMSE sense) is the conditional mean $\hat{Z} = E\{Z_t | Y_a^b)$, and the computation of this estimate has been discussed previously.

However, if the number of observations $(b-a+1)$ is large, this computation can be quite cumbersome unless the problem exhibits special structure (as in the Kalman-Bucy model). Furthermore, the determinati of the conditional mean generally requires knowledge of the joint distribution of the variables $X_t, Y_a, \dots, Y_b$. Knowledge that may be impractical (or impossible) to obtain in practice.

One way to simplify the problem is to constrain the estimators to be considered to be some of computationally convenient form, and then to minimiz the MSE over the constrained class. One such constraint that is quite useful is the linear constraint, in which we consider estimates $\hat{X}_t$ of the form

$$\hat{X}_t = \sum_{n=a}^{b} h_{t,n} Y_n + C_t$$

where $h_{t,a}, \dots, h_{t,b}$ and C_t are scalars. As we shall see below, this constraint also solves the second problem of having to specify the joint distributi of all variables, since only knowlegde of second-order statistics will be needed to optimize over linear estimates.

⇒ Statistical Signal Processing (SSP)