Lecture 22

Last time:

$$\hat{x} = \frac{\dfrac{\hat{x}_1}{\sigma_{\hat{x}}^2} + \displaystyle\sum_{k=N_1+1}^{N} \frac{z_k}{\sigma_k^2}}{\dfrac{1}{\sigma_{\hat{x}}^2} + \displaystyle\sum_{k=N_1+1}^{N} \frac{1}{\sigma_k^2}}$$

The estimate of x based on N data points can then be made without reprocessing the first N_1 points. Their effect can be included simply by starting with a pseudo observation which is equal to the estimate based on the first N_1 points having a variance equal to the variance of the estimate based on N_1.

The same is true of the variance of the estimate based on N.

$$\frac{1}{\sigma_{\hat{x}}^2} = \frac{1}{\sigma_{\hat{x}_1}^2} + \sum_{k=N_1+1}^{N} \frac{1}{\sigma_k^2}$$

A priori information about x can be included in exactly this way whether or not it was derived from previous measurements. Whatever the source of the prior information, it can be expressed as an a priori distribution $f(x)$, or at least as an expected value and a variance. Take the expected value as a pseudo observation, σ_0^2, and accumulate this data with the actual data using the standard formulae. With the prior information included as a pseudo observation, the least squares estimate is formed just as if there were no prior information. The result, for normal variables at least, is identical to the estimators based on the conditional distribution of x.

Bayes' rule can be used to form the distribution $f\left(\underline{x} \mid \underline{z}_1, \underline{z}_2, ..., \underline{z}_N\right)$ starting from the original a priori distribution $f(x)$

$f(x)$

$$f(x \mid z_1) = \frac{f(x) f\left(z_1 \mid x\right)}{\int f(u) f\left(z_1 \mid u\right) du}$$

$$f(x \mid z_1, z_2) = \frac{f\left(x \mid z_1\right) f\left(z_2 \mid x\right)}{\int f\left(u \mid z_1\right) f\left(z_2 \mid u\right) du}$$

etc.

if the measurements are conditionally independent. Two disadvantages relative to the previous method:

- More computation unless you know each conditional density is going to be normal
- Must provide $f(x)$ - the a priori distribution. This is both the advantage and disadvantage of this method.

Other estimators include the effect of a priori information directly. Several estimators are based on the conditional probability distribution of x given the values of the observations. In this approach, we think of x as a random variable having some distribution. This troubles some people since we know x is in fact fixed at some value throughout all the experiment. However, the fact that we do not know what the value is is expressed in terms of a distribution of possible values for x. The extent of our a priori knowledge is reflected in the variance of the a priori distribution we assign.

Having an a priori distribution for x, and the values of the observations, we can in principle – and often in fact – calculate the conditional distribution of x given the observations. This is in fact the a posteriori distribution, $f\left(x \mid z_1,...,z_N\right)$. This distribution expresses the probability density for various values of x given the values of the observations and the a priori distribution. Having this distribution, one can define a number of reasonable estimates. One is the minimum variance estimate – that value $\hat{x}$ which minimizes the error variance.

$$\text{Var} = \int_{-\infty}^{\infty} \left(\hat{x} - x\right)^2 f\left(x \mid z_1,...,z_N\right) dx$$

But a derivative of this variance with respect to $\hat{x}$ shows the minimizing value of $\hat{x}$ to be

$$\hat{x} = \int_{-\infty}^{\infty} x f\left(x \mid z_1,...,z_N\right) dx$$

which is the conditional mean – the mean of the conditional distribution of x.

Another reasonable estimate of x based on this conditional distribution is the value at the maximum probability density. This can perfectly well be called the "maximum likelihood" estimate though it is not necessarily the same as the maximum likelihood estimate we have just derived. Schweppe calls it the MAP ("maximum a posteriori probability") estimator. The two are related as follows:

The first is the x which maximizes

$$f\left(\underline{z}_1,...,\underline{z}_N \mid \underline{x}\right) = \frac{f\left(\underline{z}_1,...,\underline{z}_N,\underline{x}\right)}{f(\underline{x})}$$

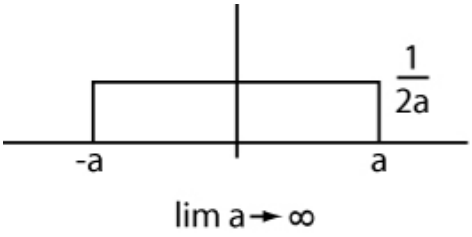
The second is the x which maximizes

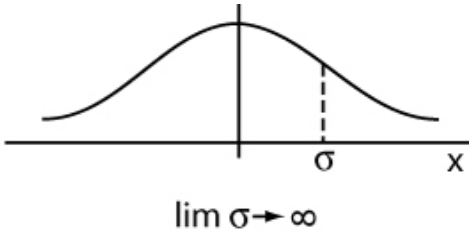
$$f\left(\underline{x} \mid \underline{z}_1,...,\underline{z}_N\right) = \frac{f\left(\underline{x},\underline{z}_1,...,\underline{z}_N\right)}{f\left(\underline{z}_1,...,\underline{z}_N\right)}$$

$$f\left(\underline{z}_1,...,\underline{z}_N \mid \underline{x}\right) = \frac{f\left(\underline{z}_1,...,\underline{z}_N,\underline{x}\right)}{f(\underline{x})}$$

$$= f\left(\underline{x} \mid \underline{z}_1,...,\underline{z}_N\right)\frac{f\left(\underline{z}_1,...,\underline{z}_N\right)}{f(\underline{x})}$$

The a priori distribution $f(\underline{x})$, and the distribution of x is also involved in the joint distribution of the observations. Only if the distribution of x is flat for all x can we guarantee that these two functions will have maxima at the same value of x. But a flat distribution for x for all x is exactly the case when there is no a priori knowledge about x. In that case all values of x have equal probability density which is in fact an infinitesimal. We can consider $f(x)$ in that case to be the limit of almost any convenient distribution as the variance $\to \infty$.





If, on the other hand, we do have some prior information about x based on some previous measurements or on physical reason, $f(x)$ will have some finite shape – often a normal shape – and the x which maximizes $f\left(z_1,...,z_N \mid x\right)$ will not maximize $f\left(x \mid z_1,...,z_N\right)$. In this case the latter choice of x is to be preferred since it is the most probable value of x based on the a priori distribution of x and the values of the observations, whereas the former depends only on the observations.

We just note in passing that for a normal $f\left(x \mid z_1,...,z_N\right)$ or any other distribution which is symmetric about the maximum point, the most probable value is equal to the conditional mean which is the minimum variance value of $\hat{x}$. In the absence of prior information, this is also the maximum likelihood estimate, which is the least weighted squares estimate. This we found to be a linear combination of the data. So in the case of a normal conditional distribution with no prior information, the optimum linear estimate based on the data is the minimum variance estimator. No nonlinear operation on the data can give a smaller variance estimate.

For non-normal distributions, one often defines a linear estimator and optimizes it based on minimum mean squared error. This gives the same estimate as that derived here. But in that case it may be that some nonlinear operation on the data could do better.

Also note that if $f\left(x \mid z_1,...,z_N\right)$ is known to be normal, as if all noises are normal, the initial distribution is normal, and only linear operations are involved, then the distribution is completely defined by its mean and variance – and only these parameters need be computed. But in any other case, updating a general distribution requires the entire distribution- and the estimation problem becomes infinite in dimension.

Estimators based on the conditional distribution of x are thus very satisfying theoretically, but are more complicated to derive than a simple least squares fit. We have found the least squares estimate to equal these other estimators if there is no a priori information. Fortunately it is possible to include a priori information in the least squares format in such a way as to recast the problem in the form of an equivalent estimation problem with no prior information. This is done by introducing the a priori information in the form of an additional pseudo observation. That this can be done is due to the fact that the least squares estimate is a linear combination of the observations – and thus it is possible to group the observations in any way.

Recursive formulation

(The other method is called batch processing.)

Suppose we had the estimate $\hat{x}_0$ based on any prior information and all measurements already taken, and its variance σ_0^2. Then we took just one more measurement and wished to obtain an improved estimate of x immediately. The formula says

$$\hat{x}_1 = \frac{\dfrac{\hat{x}_0}{\sigma_0^2} + \dfrac{z}{\sigma_z^2}}{\dfrac{1}{\sigma_0^2} + \dfrac{1}{\sigma_z^2}} = \frac{\sigma_z^2}{\sigma_z^2 + \sigma_0^2}\hat{x}_0 + \frac{\sigma_0^2}{\sigma_z^2 + \sigma_0^2} z$$

$$= \hat{x}_0 + \frac{\sigma_0^2}{\sigma_0^2 + \sigma_z^2}\left(z - \hat{x}_0\right)$$

$$= \hat{x}_0 + k\left(z - \hat{x}_0\right)$$

New estimate = Old estimate + Gain(Measurement residual)

This is the form of the modern recursive estimator. If this is formulated directly in the case where several parameters are being estimated, the $\left(\sigma_0^2+\sigma_z^2\right)^{-1}$ is a matrix inversion. However, if only one scalar measurement is being processed, the inversion is a scalar reciprocal.

$$\frac{1}{\sigma_1^2}=\frac{1}{\sigma_0^2}+\frac{1}{\sigma_z^2}=\frac{\sigma_z^2+\sigma_0^2}{\sigma_0^2\sigma_z^2}$$

$$\sigma_1^2=\sigma_0^2\sigma_z^2\left(\sigma_0^2+\sigma_z^2\right)^{-1}$$

$$=\left(1-k\right)\sigma_0^2$$

$$k\equiv\sigma_0^2\left(\sigma_0^2+\sigma_z^2\right)^{-1}$$

Extensions to this simple problem:

- Several estimated parameters instead of one
  - No conceptual difficulty
  - Get vector and matrix operations instead of scalar
- Dynamic parameters instead of static
  - No real difficulty if they obey a set of linear differential equations
- Several simultaneous measurements instead of one
  - No difficulty if they are linearly related to $\underline{x}$
- Biased measurement noise
  - Estimate the bias
- Correlated measurement noise
  - Estimate the noise (different form of filter) or work with independent measurement differences
- Non-normal noises
  - Makes maximum likelihood difficult
  - Requires full distribution rather than mean and variance
- Nonlinear system constraints or measurements
  - Makes things very difficult
  - Requires more information than mean and variance

Statistics in State Space Formulation

In the state space formulation we depend on the concept of the shaping filter exclusively. Even if the input statistics are time varying, we suppose the input to be generated by passing white noise through a suitable time-varying shaping filter. In the non-stationary case it is not clear how to generate the shaping system directly.

Non-stationary, multiple-input, multiple-output case

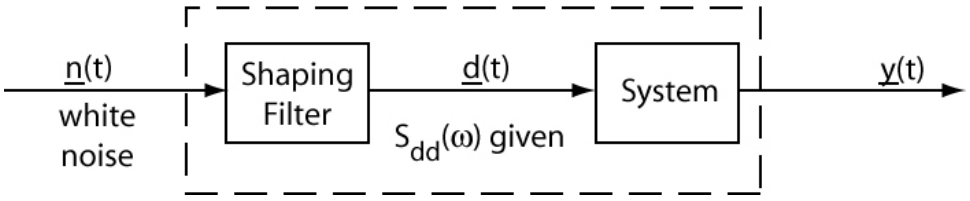
Note that this model may represent the shaping filter for a random process alone, or a system driven by white noise, or a system driven by correlated noise with the shaping filter included.

Shaping filter:

$$\dot{\underline{v}} = A\underline{v} + B\underline{n}$$
$$\underline{d} = C\underline{v}$$

Transfer function (state space model):

$$\dot{\underline{x}} = F\underline{x} + \underline{d}$$
$$\underline{y} = H\underline{x}$$



Augmented system:

$$\underline{x}' = \begin{bmatrix} \underline{x} \\ \underline{v} \end{bmatrix}$$

$$\dot{\underline{x}}' = \underbrace{\begin{bmatrix} F & GC \\ 0 & A \end{bmatrix}}_{A'} \underline{x}' + \underbrace{\begin{bmatrix} 0 \\ B \end{bmatrix}}_{B'} \underline{n} \equiv A'\underline{x}' + B'\underline{n}'$$

$$\underline{y} = \underbrace{\begin{bmatrix} H & 0 \end{bmatrix}}_{C'} \underline{x}' \equiv C'\underline{x}'$$

We will drop the primes and treat the input noise as white.

$$\dot{\underline{x}}(t) = A(t)\underline{x}(t) + B(t)\underline{n}(t)$$
$$\underline{y}(t) = C(t)\underline{x}(t)$$

We will allow the possibility of a biased input noise.

$\underline{n}(t)$ is defined.

$$\overline{\left[\underline{n}(t_1) - \overline{\underline{n}(t_1)}\right]\left[\underline{n}(t_2) - \overline{\underline{n}(t_2)}\right]^T} = N(t_1)\delta(t_2 - t_1)$$

The elements of $N(t)$ are the intensities, or spectral densities, of the components of $n(t)$. Never call it the variance of the white noise!

Propagation of the mean

$$\dot{\overline{\underline{x}(t)}} = A(t)\overline{\underline{x}(t)} + B(t)\overline{\underline{n}(t)}$$
$$\overline{\underline{y}(t)} = C(t)\overline{\underline{x}(t)}$$

So the mean is propagated directly by the system dynamics.

Propagation of the covariance matrix

$$\overline{\left[\underline{y}(t) - \overline{\underline{y}(t)}\right]\left[\underline{y}(t) - \overline{\underline{y}(t)}\right]^T} = C(t)\overline{\left[\underline{x}(t) - \overline{\underline{x}(t)}\right]\left[\underline{x}(t) - \overline{\underline{x}(t)}\right]^T}C^T$$
$$= C(t)X(t)C(t)^T$$

So we require the covariance matrix for the full state variable $\underline{x}(t)$.

Can we derive a differential equation for the covariance matrix in the same way as we did in the error propagation section?

$$X(t) = E\left[\underline{x}(t) - \overline{\underline{x}(t)}\right]\left[\underline{x}(t) - \overline{\underline{x}(t)}\right]^T$$

For simplicity, define:

$$\tilde{\underline{x}}(t) = \underline{x}(t) - \overline{\underline{x}(t)}$$
$$\tilde{\underline{n}}(t) = \underline{n}(t) - \overline{\underline{n}(t)}$$
$$\dot{\tilde{\underline{x}}}(t) = A(t)\tilde{\underline{x}}(t) + B(t)\tilde{\underline{n}}(t)$$

$$X(t) = \overline{\tilde{\underline{x}}(t)\tilde{\underline{x}}(t)^T}$$
$$\dot{X}(t) = \overline{\dot{\tilde{\underline{x}}}(t)\tilde{\underline{x}}(t)^T} + \overline{\tilde{\underline{x}}(t)\dot{\tilde{\underline{x}}}(t)^T}$$
$$= \overline{\left[A(t)\tilde{\underline{x}}(t) + B(t)\tilde{\underline{n}}(t)\right]\tilde{\underline{x}}(t)^T} + \overline{\tilde{\underline{x}}(t)\left[\tilde{\underline{x}}(t)^T A(t)^T + \tilde{\underline{n}}(t)^T B(t)^T\right]}$$
$$= A(t)\overline{\tilde{\underline{x}}(t)\tilde{\underline{x}}(t)^T} + B(t)\overline{\tilde{\underline{n}}(t)\tilde{\underline{x}}(t)^T} + \overline{\tilde{\underline{x}}(t)\tilde{\underline{x}}(t)^T}A(t)^T + \overline{\tilde{\underline{x}}(t)\tilde{\underline{n}}(t)^T}B(t)^T$$
$$= A(t)X(t) + X(t)A(t)^T + B(t)\overline{\tilde{\underline{n}}(t)\tilde{\underline{x}}(t)^T} + \overline{\tilde{\underline{x}}(t)\tilde{\underline{n}}(t)^T}B(t)^T$$

What is the correlation function between $\tilde{\underline{n}}(t)$ and $\tilde{\underline{x}}(t)$? It is not zero even though $\tilde{\underline{n}}(t)$ drives $\tilde{\underline{x}}(t)$ and not $\tilde{\underline{x}}(t)$ directly. Because $\tilde{\underline{n}}(t)$ is a white noise, and thus is infinite in magnitude almost everywhere, it has a non-negligible effect on $\tilde{\underline{x}}(t)$ at the same point in time. But we need a more sophisticated calculus i.e., Ito calculus, to figure out what it is.

So we cannot derive a differential equation for $X(t)$ just by differentiating its definition.

Instead, we can use a more round-about procedure. The response of the differential equation for $\underline{\tilde{x}}(t)$ can be expressed as

$$\underline{\tilde{x}}(t) = \Phi(t,t_0)\underline{\tilde{x}}(t_0) + \int_{t_0}^{t} \phi(t,\tau)B(\tau)\underline{\tilde{n}}(\tau)d\tau$$

where $\Phi(t,\tau)$ is the transition matrix for the linear system with system matrix $A(t)$. It satisfies the system homogeneous equation

$$\frac{d}{dt}\Phi(t,\tau) = A(t)\Phi(t,\tau)$$
$$\Phi(\tau,\tau) = I$$

This approach works in this case because we can write down the form of the solution to a set of linear differential equations. We cannot write down the form of the solution to nonlinear differential equations so we cannot use this approach in the case of nonlinear systems. However, the Ito calculus does apply to nonlinear stochastic differential equations.