Lecture 18: Maximum Likelihood Parameter Estimation

Forms of Parameter Estimation [1, Ch.4]

We continue our discussion of parameter estimation for dynamical systems by first considering the simpler parameter estimation for a static nonlinear model of the form

$$\boldsymbol{y} = \varphi(\boldsymbol{\theta}) + \boldsymbol{v} \tag{1}$$

where $\boldsymbol{\theta}$ is an unknown parameter vector, $\varphi(\boldsymbol{\theta})$ is the (noise-free) output, and $\boldsymbol{y}$ is a measurement that consists of the output corrupted by noise $\boldsymbol{v}$. As a special case of the above nonlinear function we have the linear model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\theta} + \boldsymbol{v} \tag{2}$$

Morreli and Klein [1] discuss three forms parameter estimation models:

1. *The Least-Squares Model*
   (a) $\boldsymbol{\theta}$ is a constant (but unknown) vector of parameters (i.e., it has no p.d.f.)
   (b) $\boldsymbol{v}$ is a random vector of measurement noise which may or may not have a p.d.f. The p.d.f. is not needed for the least-square estimate but can be used to compute a confidence, if available.
   (c) The least-square estimate is

$$\hat{\boldsymbol{\theta}} = \min_{\boldsymbol{\theta}} \sum_{i=1}^{N} [\boldsymbol{y}_i - \varphi(\boldsymbol{\theta})]^2 \tag{3}$$

   for $\boldsymbol{y}_i$ data points.

2. *The Fisher Model*
   (a) $\boldsymbol{\theta}$ is a constant (but unknown) vector of parameters (i.e., it has no p.d.f.)
   (b) $\boldsymbol{v}$ is a random vector with a probability distribution function $p(\boldsymbol{v})$
   (c) The constant vector $\boldsymbol{\theta}$ is viewed as parameter in the p.d.f. that describes the distribution of the N data points $\boldsymbol{Y}_{1:N}$ according to the likelihood function

$$\mathcal{L}(\boldsymbol{Y}_{1:N}; \boldsymbol{\theta}) = p(\boldsymbol{Y}_{1:N}|\boldsymbol{\theta}) \tag{4}$$

   (d) The maximum likelihood estimator (MLE) is

$$\hat{\boldsymbol{\theta}} = \max_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{Y}_{1:N}; \boldsymbol{\theta}) \tag{5}$$

3. *The Bayesian Model*
   (a) $\boldsymbol{\theta}$ is a random vector with a probability distribution $p(\boldsymbol{\theta})$
   (b) $\boldsymbol{v}$ is a random vector with a probability distribution $p(\boldsymbol{v})$
   (c) The p.d.f. of the parameter $\boldsymbol{\theta}$ is related to the data $\boldsymbol{y}$ by Bayes' Rule:

$$p(\boldsymbol{\theta}|\boldsymbol{Y}_{1:N}) = \frac{p(\boldsymbol{Y}_{1:N}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\boldsymbol{Y}_{1:N})} \tag{6}$$

(d) The Bayesian estimator is

$$\hat{\boldsymbol{\theta}} = \max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\boldsymbol{Y}_{1:N}) \tag{7}$$

We've already discussed least-square regression and in this lecture we will introduce maximum likelihood estimation. A future lecture will describe Bayesian estimation.

Maximum Likelihood Estimation of a Static Parameter

The *likelihood function*

$$\mathcal{L}(\boldsymbol{y};\boldsymbol{\theta}) = p(\boldsymbol{y}|\boldsymbol{\theta}) \tag{8}$$

is the probability of the data/measurements $\boldsymbol{y}$ for a particular choice of parameters $\boldsymbol{\theta}$. Now, consider a system that produces a sequence of measurements $\{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N\}$ and let $\boldsymbol{Y}_{1:N} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N\}$ denote the set of N measurements. The likelihood function that considers *all* these measurements is denoted

$$\mathcal{L}(\boldsymbol{Y}_{1:N};\boldsymbol{\theta}) = \mathcal{L}(\{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N\};\boldsymbol{\theta}) = p(\boldsymbol{Y}_{1:N}|\boldsymbol{\theta}) \tag{9}$$

Split the set $\boldsymbol{Y}_{1:N}$ into two subsets:

$$\begin{aligned}\boldsymbol{Y}_{1:N} &= (\boldsymbol{y}_N, \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_{N-1}\}) \\ &= (\boldsymbol{y}_N, \boldsymbol{Y}_{1:N-1})\end{aligned}$$

Aside: Recall the relationship between joint and conditional probabilities

$$p(\boldsymbol{x}, \boldsymbol{y}) = p(\boldsymbol{x}|\boldsymbol{y})p(\boldsymbol{y}) \tag{10}$$

The joint probability $p(\boldsymbol{y}_N, \boldsymbol{Y}_{1:N-1};\boldsymbol{\theta})$ can be re-written as

$$p(\boldsymbol{y}_N, \boldsymbol{Y}_{1:N-1};\boldsymbol{\theta}) = p(\boldsymbol{y}_N|\boldsymbol{Y}_{1:N-1};\boldsymbol{\theta})p(\boldsymbol{Y}_{1:N-1};\boldsymbol{\theta}) \tag{11}$$

or, in terms of likelihood functions, it may be factored further as:

$$\begin{aligned}\mathcal{L}(\boldsymbol{Y}_{1:N};\boldsymbol{\theta}) &= \mathcal{L}(\{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N\};\boldsymbol{\theta}) \\ &= \mathcal{L}(\boldsymbol{y}_N|\boldsymbol{Y}_{1:N-1};\boldsymbol{\theta})\mathcal{L}(\boldsymbol{Y}_{1:N-1};\boldsymbol{\theta}) \\ &= \mathcal{L}(\boldsymbol{y}_N|\boldsymbol{Y}_{1:N-1};\boldsymbol{\theta})\mathcal{L}(\boldsymbol{y}_{N-1}|\boldsymbol{Y}_{1:N-2};\boldsymbol{\theta})\mathcal{L}(\boldsymbol{Y}_{1:N-2};\boldsymbol{\theta}) \\ &\quad\vdots \\ &= \prod_{i=1}^{N} \mathcal{L}(\boldsymbol{y}_i|\boldsymbol{Y}_{1:i-1};\boldsymbol{\theta})\end{aligned}$$

It is advantageous to consider minimizing the negative logarithm of this function, rather than

maximize the likelihood. The maximum likelihood (ML) estimator is

$$\begin{aligned}\hat{\boldsymbol{\theta}} &= \max_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{Y}_{1:N};\boldsymbol{\theta}) \\ &= \max_{\boldsymbol{\theta}} \prod_{i=1}^{N} \mathcal{L}(\boldsymbol{y}_i|\boldsymbol{Y}_{1:i-1};\boldsymbol{\theta}) \\ &= \min_{\boldsymbol{\theta}} \underbrace{\sum_{i=1}^{N} -\ln[\mathcal{L}(\boldsymbol{y}_i|\boldsymbol{Y}_{1:i-1};\boldsymbol{\theta})]}_{=J(\boldsymbol{\theta})}\end{aligned}$$

If we further assume that the noise is uncorrelated in time then the likelihood functions also become independent:

$$\mathcal{L}(\boldsymbol{y}_i|\boldsymbol{Y}_{1:i-1};\boldsymbol{\theta}) = \mathcal{L}(\boldsymbol{y}_i;\boldsymbol{\theta}) \tag{12}$$

and the ML estimator is:

$$\hat{\boldsymbol{\theta}}_{\text{MLE}} = \min_{\boldsymbol{\theta}} \underbrace{\sum_{i=1}^{N} -\ln[\mathcal{L}(\boldsymbol{y}_i;\boldsymbol{\theta})]}_{=J(\boldsymbol{\theta})}$$

The above cost function can be minimized analytically if it is simple enough, otherwise we may resort to numerical methods. One advantage of MLE is that it allows us to consider non-Gaussian distributions, as in the following example.

Example (Rayleigh distribution). The distribution of wave heights in the ocean is commonly modeled using a *Rayleigh* probability distribution that has probability density function (p.d.f.):

$$f(y;\theta) = \frac{y}{\theta^2} e^{-y^2/(2\theta^2)} \tag{13}$$

where θ is a scale parameter that describes the shape of the distribution. The distribution is restricted to the positive real numbers $y \in [0, \infty]$. Now suppose that we obtain a set of data

$$Y = \{y_1, y_2, \ldots, y_N\} \tag{14}$$

that represents wave height measurements and we wish to determine estimate the parameter θ that best matches the data. The negative log-likelihood is

$$\begin{aligned}-\ln\left[\mathcal{L}(y;\theta)\right] &= -\ln\left[\frac{y}{\theta^2} e^{-y^2/(2\theta^2)}\right] \\ &= -\ln\left[y\right] + 2\ln\left[\theta\right] + \left[y^2/(2\theta^2)\right]\end{aligned}$$

The MLE minimizes the cost function

$$\begin{aligned}J(\boldsymbol{\theta}) &= \sum_{i=1}^{N} \left\{-\ln\left[y_i\right] + 2\ln\left[\theta\right] + \left[y_i^2/(2\theta^2)\right]\right\} \\ &= -\sum_{i=1}^{N} \ln\left[y_i\right] + 2N\ln\left[\theta\right] + \frac{1}{2\theta^2}\sum_{i=1}^{N} y_i^2\end{aligned}$$

which considers all of the data. To find the minimum we set the partial derivative to zero

$$0 = \frac{\partial J(\theta)}{\partial \theta} = \frac{\partial}{\partial \theta}\left\{-\sum_{i=1}^{N} \ln\left[y_i\right]\right\} + 2N\frac{\partial}{\partial \theta}\ln\left[\theta\right] + \frac{\partial}{\partial \theta}\frac{1}{2\theta^2}\sum_{i=1}^{N} y_i^2$$

$$0 = 0 + \frac{2N}{\hat{\theta}_{MLE}} - \frac{1}{\hat{\theta}_{MLE}^3}\sum_{i=1}^{N} y_i^2$$

$$2N\hat{\theta}_{MLE}^2 = \sum_{i=1}^{N} y_i^2$$

$$\implies \hat{\theta}_{MLE}^2 = \sqrt{\frac{1}{2N}\sum_{i=1}^{N} y_i^2}$$

In the above example the parameter of interest characterized the probability distribution, however the same approach can be taken for any parameters that appear in the model, or both the model parameters and noise parameters. Consider the following example.

Example (MLE for Scalar Linear Model with Gaussian Noise). Consider a scalar measurement model of the form:

$$\begin{aligned} y &= mx + b + v \\ v &\sim \mathcal{N}(0, \sigma^2) \end{aligned} \tag{15}$$

where m, b, and σ^2 are unknown model parameters. A data set is collected consisting of pairs

$$\{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\} \tag{16}$$

The vector of parameters to be estimated is

$$\hat{\boldsymbol{\theta}} = [\hat{m}, \hat{b}, \hat{\sigma}^2]^{\mathrm{T}} \tag{17}$$

The parameters m and b are unknown but they are deterministic. Thus, the probability of measuring some y_i given an input x_i is equivalent to the probability that the noise v_i takes on the particular value

$$v_i = y_i - (mx_i + b) \tag{18}$$

Since the noise is normally distributed the probability can be expressed using the Gaussian p.d.f. function:

$$\mathcal{L}(y; m, b, \sigma^2) = p(v|m, b, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(\frac{-v^2}{2\sigma^2}\right) \tag{19}$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(\frac{-(z - mx - b)^2}{2\sigma^2}\right) \tag{20}$$

The negative log-likelihood is:

$$-\ln\left[\mathcal{L}(y;m,b,\sigma^2)\right] = -\ln\left[\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(\frac{-(y-mx-b)^2}{2\sigma^2}\right)\right] \tag{21}$$

$$= -\left(\frac{-(y-mx-b)^2}{2\sigma^2}\right) + \ln\sqrt{2\pi\sigma^2} \tag{22}$$

$$= -\left(\frac{-(y-mx-b)^2}{2\sigma^2}\right) + \frac{1}{2}(\ln 2\pi + \ln\sigma^2) \tag{23}$$

$$= -\left(\frac{-(y-mx-b)^2}{2\sigma^2}\right) + \frac{1}{2}\ln 2\pi + \frac{1}{2}\ln\sigma^2 \tag{24}$$

The MLE minimizes the cost function

$$J(\boldsymbol{\theta}) = \sum_{i=1}^{N}\left[-\left(\frac{-(y_i-mx_i-b)^2}{2\sigma^2}\right) + \frac{1}{2}\ln 2\pi + \frac{1}{2}\ln\sigma^2\right] \tag{25}$$

$$= -\sum_{i=1}^{N}\left(\frac{-(y_i-mx_i-b)^2}{2\sigma^2}\right) + \frac{N}{2}\ln 2\pi + \frac{N}{2}\ln\sigma^2 \tag{26}$$

which considers all of the data. To find the minimum we set the partial derivative with respect to each parameter to zero. Begin with the partial with respect to b:

$$0 = \frac{\partial J(\theta)}{\partial b} = -\frac{\partial}{\partial b}\sum_{i=1}^{N}\left(\frac{-(y_i-mx_i-b)^2}{2\sigma^2}\right) \tag{27}$$

$$0 = -\sum_{i=1}^{N}\left(\frac{-2(y_i-\hat{m}x_i-\hat{b})}{2\sigma^2}\right)(-1) \tag{28}$$

$$0 = \sum_{i=1}^{N}(y_i-\hat{m}x_i-\hat{b}) \tag{29}$$

$$\hat{b} = \frac{1}{N}\sum_{i=1}^{N}(y_i-\hat{m}x_i) \tag{30}$$

$$\implies \hat{b} = \bar{y} - \hat{m}\bar{x} \tag{31}$$

where $\bar{x}$ and $\bar{y}$ denote the mean values of the x and y data points (i.e., the *sample mean*). Next, evaluate the partial derivative with respect to m:

$$0 = \frac{\partial J(\theta)}{\partial m} = -\frac{\partial}{\partial m}\sum_{i=1}^{N}\left(\frac{-(y_i-\hat{m}x_i-\hat{b})^2}{2\sigma^2}\right) \tag{32}$$

$$0 = \sum_{i=1}^{N}\left(\frac{2(y_i-\hat{m}x_i-\hat{b})}{2\sigma^2}\right)(-x_i) \tag{33}$$

$$0 = \sum_{i=1}^{N}(y_ix_i-\hat{m}x_i^2-\hat{b}x_i) \tag{34}$$

the substiuting $\hat{b}$

$$0 = \sum_{i=1}^{N}(y_i x_i - \hat{m}x_i^2 - (\bar{z} - \hat{m}\bar{x})x_i) \tag{35}$$

$$0 = \sum_{i=1}^{N} y_i x_i - \hat{m}\sum_{i=1}^{N} x_i^2 - \bar{z}\sum_{i=1}^{N} x_i + \hat{m}\bar{x}\sum_{i=1}^{N} x_i \tag{36}$$

$$0 = \sum_{i=1}^{N} y_i x_i - \hat{m}\sum_{i=1}^{N} x_i^2 - N\bar{z}\bar{x} + N\hat{m}\bar{x}^2 \tag{37}$$

$$\hat{m} = \frac{\sum_{i=1}^{N} y_i x_i - N\bar{z}\bar{x}}{\sum_{i=1}^{N} x_i^2 - N\bar{x}^2} \tag{38}$$

$$\hat{m} = \frac{\sum_{i=1}^{N} y_i x_i - \sum_{i=1}^{N} \bar{z}\bar{x}}{\sum_{i=1}^{N} x_i^2 - \sum_{i=1}^{N} \bar{x}^2} \tag{39}$$

$$\hat{m} = \frac{\sum_{i=1}^{N}(y_i x_i - \bar{z}\bar{x})}{\sum_{i=1}^{N}(x_i^2 - \bar{x}^2)} \tag{40}$$

$$\implies \hat{m} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{z})}{\sum_{i=1}^{N}(x_i - \bar{x})^2} \tag{41}$$

Notice tha the numerator of this expression is equialent to the *sample covariance* between x and y and the denominator is the sample variance of x. Lastly, take the partial derivative with respect to σ^2:

$$0 = \frac{\partial J(\theta)}{\partial \sigma^2} = -\frac{\partial}{\partial \sigma^2}\sum_{i=1}^{N}\left(\frac{-(y_i - mx_i - b)^2}{2\sigma^2}\right) + \frac{\partial}{\partial \sigma^2}\frac{N}{2}\ln\sigma^2 \tag{42}$$

$$0 = \sum_{i=1}^{N}\left(\frac{-(y_i - mx_i - b)^2}{2\sigma^4}\right) + \frac{N}{2\sigma^2} \tag{43}$$

$$0 = -\sum_{i=1}^{N}(y_i - mx_i - b)^2 + N\sigma^2 \tag{44}$$

$$\implies \hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{m}x_i - \hat{b})^2 \tag{45}$$

Together the three equations (31), (41), and (45) form the maximum likelihood estimate.

Additive Gaussian Measurement Noise. The example above which considers scalar Gaussian measurement noise can be extended to a more general vector-valued and nonlinear measurement model of the form

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{\varphi}(\boldsymbol{x};\boldsymbol{\theta}) + \boldsymbol{v} \\ \boldsymbol{v} &\sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{R}) \end{aligned} \tag{46}$$

As before, the probability of observing some data point $\boldsymbol{y}$ is equivalent to the probability that the noise takes on the value

$$\boldsymbol{v} = \boldsymbol{y} - \boldsymbol{\varphi}(\boldsymbol{x};\boldsymbol{\theta}) \tag{47}$$

for a given set of parameters $\boldsymbol{\theta}$. The likelihood function is given by the standard Gaussian multivariate p.d.f. for the probability that $\boldsymbol{v} = \boldsymbol{y} - \varphi(\boldsymbol{x};\boldsymbol{\theta})$. That is,

$$\mathcal{L}(\boldsymbol{y};\boldsymbol{\theta}) = p(\boldsymbol{y}|\boldsymbol{\theta}) = \frac{\exp\left(-\frac{1}{2}(\boldsymbol{y} - \varphi(\boldsymbol{x};\boldsymbol{\theta}))^{\mathrm{T}}\boldsymbol{R}^{-1}(\boldsymbol{y} - \varphi(\boldsymbol{x};\boldsymbol{\theta})\right)}{(2\pi)^{n/2}\sqrt{|\boldsymbol{R}|}} \tag{48}$$

If we compute the negative log likelihood (and ignore parts of the cost function that are not dependent on $\boldsymbol{\theta}$) we end up with the cost function:

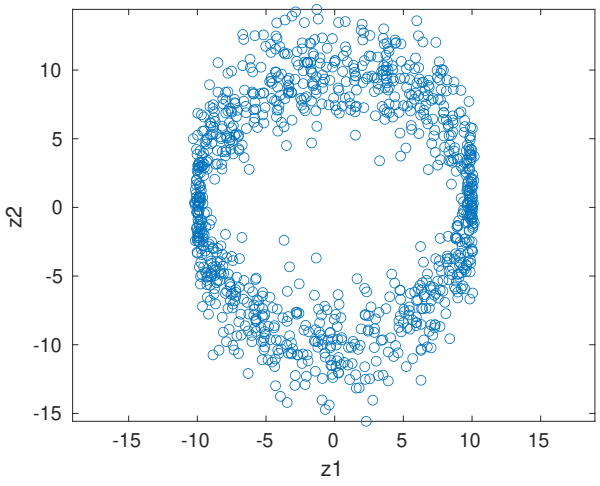
$$J(\boldsymbol{\theta}) = \sum_{i=1}^{N} \frac{1}{2}(\boldsymbol{y}_i - \varphi(\boldsymbol{x};\boldsymbol{\theta}))^{\mathrm{T}}\boldsymbol{R}^{-1}(\boldsymbol{y}_i - \varphi(\boldsymbol{x};\boldsymbol{\theta})) \tag{49}$$

that appears within the exponent in the p.d.f. The optimal $\hat{\boldsymbol{\theta}}$ can be solved for using various optimization procedures such as gradient descent. For low dimensional $\boldsymbol{\theta}$ a coarse optimization may be possible by brute force (i.e., evaluating $J(\boldsymbol{\theta})$ for a grid of $\boldsymbol{\theta}$ values and identifying the minimum).

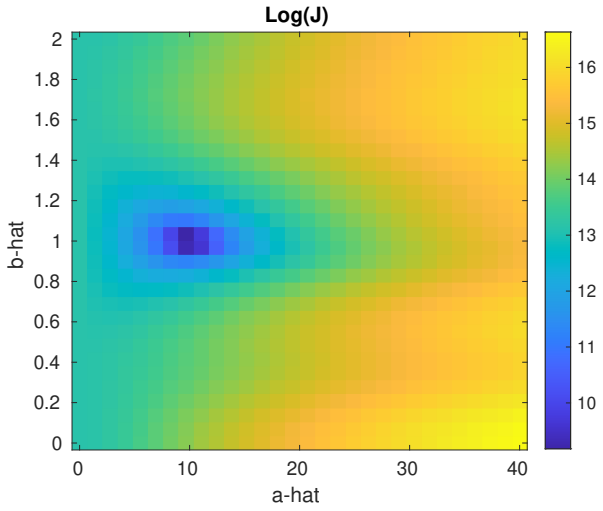
Example. Consider a nonlinear vector-valued measurement model of the form:

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} a\cos bx_1 \\ a\sin bx_1 \end{bmatrix} + \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} \qquad \boldsymbol{v} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{R}) \tag{50}$$

where a and b are unknown model parameters (with true values $a = 10$ and $b = 1$) to be estimated and $\boldsymbol{R} = \mathrm{diag}([\sigma_1^2, \sigma_2^2])$ with known variances $\sigma_1^2 = 0.1$ and $\sigma_2^2 = 2$. An example data set of $N = 1000$ points was generated by varying $x_1 \in [0, 2\pi]$ as shown below:



The data resembles a noisy circle that is "spread out" along the z_2 axis since $\sigma_2 > \sigma_1$. The cost function is defined by computing $J(\boldsymbol{\theta})$ with all N data points and is plotted below for a $M \times M$ search grid with $\hat{a} \in [0, 40]$ and $\hat{b} \in [0, 2]$ with $M = 30$. To emphasize changes in the cost function the logarithm of J is plotted below.



As shown, the cost function has a minimum near the correct values of $a = 10$ and $b = 1$. In this example our actual estimate $\hat{a}$ and $\hat{b}$ will depend on the resolution of the search grid.

Maximum Likelihood Estimation in A Dynamical System

The above examples discuss maximum likelihood estimation for a static system model. How can we extend this to find the parameters of a dynamical system using the output data? The parameters of a dynamic system can appear anywhere in the state equation, measurement equation, or the noise covariance matrices. Maximum likelihood estimation provides an approach for estimating these parameters. The most common methods described in [1, Ch.4] methods are the (a) *filter-error method*, (b) the *output-error method*, and (c) the *equation-error* method. A block diagram of each methods is shown below. (Note: the block diagram is for a particular system "aircraft" but the approach is more general.)

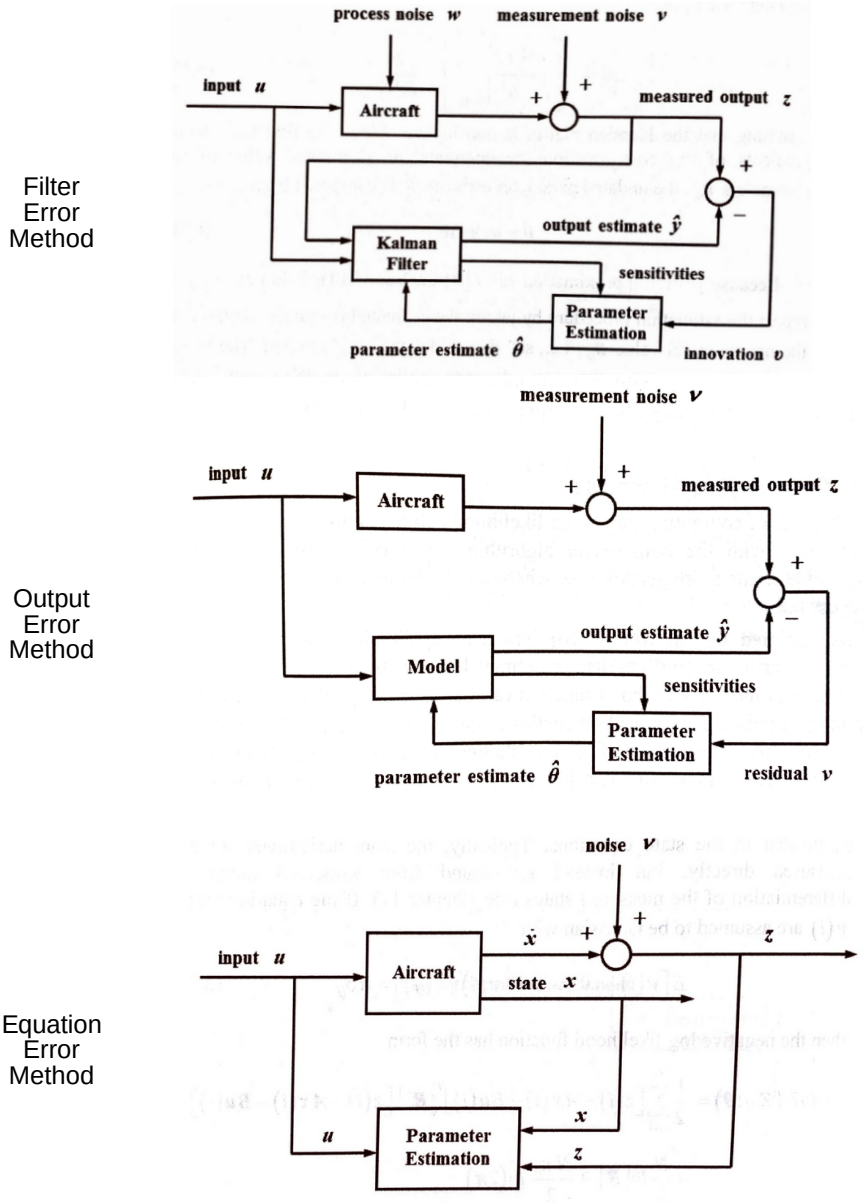


Image source: [1, Ch.4]

The filter-error is the most general form and works in tandem with a Kalman filter, as we will show below. The output-error considers a simplified system model in which the process noise is

ignored. Thus the state equation is entirely deterministic. Lastly, notice that the equation error model has no output/measurement equation. That is, this method assumes all of the states and all of the state derivatives are available. In this lecture we will focus on the filter-error method as described next.

Filter-error Parameter Estimation

Consider a LTI discrete system in the form

$$\boldsymbol{x}_k = \boldsymbol{F}\boldsymbol{x}_{k-1} + \boldsymbol{G}\boldsymbol{u}_{k-1} + \boldsymbol{w}_{k-1} \tag{51}$$

$$\boldsymbol{y}_k = \boldsymbol{H}\boldsymbol{x}_k + \boldsymbol{v}_k \tag{52}$$

$$\boldsymbol{w}_k \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{Q}) \tag{53}$$

$$\boldsymbol{v}_k \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{R}) \tag{54}$$

where the matrices $\boldsymbol{F}$, $\boldsymbol{G}$, $\boldsymbol{Q}$, $\boldsymbol{H}$, and $\boldsymbol{R}$ may be a function of some unknown parameters. Following our approach for static system maximum likelihood parameter estimation, we consider a sequence of noisy measurements $\{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N\}$ and let $\boldsymbol{Y}_{1:N} = [\boldsymbol{y}_1^T, \ldots, \boldsymbol{y}_N^T]^T$ denote the a vector that stacks these N measurements into a single column. The likelihood function that considers *all* these measurements is denoted

$$\mathcal{L}(\boldsymbol{Y}_{1:N}; \boldsymbol{\theta}) = \mathcal{L}([\boldsymbol{y}_1^T, \ldots, \boldsymbol{y}_N^T]^T; \boldsymbol{\theta}) \tag{55}$$

As before, the MLE $\hat{\boldsymbol{\theta}}$ minimizes the cost function of the negative log-likelihood (assuming uncorrleated measurements):

$$J(\boldsymbol{\theta}) = -\ln[\mathcal{L}(\boldsymbol{Y}_{1:N}; \boldsymbol{\theta})] = \sum_{i=1}^{N} -\ln[\mathcal{L}(\boldsymbol{y}_i; \boldsymbol{\theta})]$$

That is, the maximum likelihood (ML) estimator is

$$\hat{\boldsymbol{\theta}} = \min_{\boldsymbol{\theta}} \underbrace{\sum_{i=1}^{N} -\ln[\mathcal{L}(\boldsymbol{y}_i; \boldsymbol{\theta})]}_{=J(\boldsymbol{\theta})}$$

We can view each measurement $\boldsymbol{y}_k$ as a random vector

$$\boldsymbol{y}_k \sim \mathcal{N}(\boldsymbol{y}_{k|k-1}, \boldsymbol{S}_k) \tag{56}$$

with mean given by the measurement $\boldsymbol{y}_{k|k-1}$ predicted from a Kalman filter and the variance given as the innovation $\boldsymbol{S}_k$. That is,

$$\boldsymbol{y}_{k|k-1} = E[\boldsymbol{y}_k] \tag{57}$$

$$\boldsymbol{S}_k = E[(\boldsymbol{y}_k - \boldsymbol{y}_{k|k-1})(\boldsymbol{y}_k - \boldsymbol{y}_{k|k-1})^{\mathrm{T}})] \tag{58}$$

$$= E[\boldsymbol{\epsilon}_k \boldsymbol{\epsilon}_k^{\mathrm{T}}] \tag{59}$$

where

$$\boldsymbol{\epsilon}_k = \boldsymbol{y}_k - \boldsymbol{y}_{k|k-1} \tag{60}$$

is called the *innovation* vector and $\boldsymbol{S}_k$ is the innovation covariance. As the number of measurements increases this assumption is justified and the p.d.f. of the innovation approaches a Gaussian distribution. The quantities (57) and (58) that describe this likelihood are provided by the Kalman filter. Recall that in the Kalman filter we perform the following steps at each iteration:

$$\text{State Prior :} \quad \hat{\boldsymbol{x}}_{k-1|k-1} \tag{61}$$

$$\text{Covariance Prior :} \quad \boldsymbol{P}_{k-1|k-1} \tag{62}$$

$$\text{Current Measurement :} \quad \boldsymbol{y}_k \tag{63}$$

$$\text{State Prediction :} \quad \hat{\boldsymbol{x}}_{k|k-1} = \boldsymbol{F}\hat{\boldsymbol{x}}_{k-1|k-1} + \boldsymbol{G}\boldsymbol{u}_{\boldsymbol{k}-1} \tag{64}$$

$$\text{Covariance Prediction :} \quad \boldsymbol{P}_{k|k-1} = \boldsymbol{F}\boldsymbol{P}_{k-1|k-1}\boldsymbol{F}^T + \boldsymbol{Q} \tag{65}$$

$$\text{Measurement Prediction :} \quad \boldsymbol{y}_{k|k-1} = \boldsymbol{H}\hat{\boldsymbol{x}}_{k|k-1} \tag{66}$$

$$\text{Innovation Covariance :} \quad \boldsymbol{S}_k = \boldsymbol{H}_k\boldsymbol{P}_{k|k-1}\boldsymbol{H}_k^T + \boldsymbol{R}_k \tag{67}$$

$$\text{Kalman Gain :} \quad \boldsymbol{K}_k = \boldsymbol{P}_{k|k-1}\boldsymbol{H}_k^T\boldsymbol{S}_k^{-1} \tag{68}$$

$$\text{State Posterior :} \quad \hat{\boldsymbol{x}}_{k|k} = \hat{\boldsymbol{x}}_{k|k-1} + \boldsymbol{K}_k(\boldsymbol{y}_k - \boldsymbol{y}_{k|k-1}) \tag{69}$$

$$\text{Covariance Posterior :} \quad \boldsymbol{P}_{k|k} = (\boldsymbol{I} - \boldsymbol{K}_k\boldsymbol{H}_k)\boldsymbol{P}_{k|k-1} \tag{70}$$

Thus (57) and (58) are found from (66) and (67). Having characterized the mean and covariance of the Gaussian random vector (56) we may compute the likelihood:

$$\mathcal{L}(\boldsymbol{y}_k;\boldsymbol{\theta}) = p(\boldsymbol{y}|\boldsymbol{\theta}) = \frac{\exp\left(-\frac{1}{2}(\boldsymbol{y}_k - \boldsymbol{y}_{k|k-1})^{\mathrm{T}}\boldsymbol{S}_k^{-1}(\boldsymbol{y}_k - \boldsymbol{y}_{k|k-1})\right)}{(2\pi)^{m/2}\sqrt{|\boldsymbol{S}_k|}} \tag{71}$$

where m is the size of the output. The negative log likelihood (across all N data points) is:

$$-\ln[\mathcal{L}(\boldsymbol{Y}_N;\boldsymbol{\theta})] = -\sum_{i=1}^{N}\left[\left(-\frac{1}{2}(\boldsymbol{y}_k - \boldsymbol{y}_{i|i-1})^{\mathrm{T}}\boldsymbol{S}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{y}_{i|i-1})\right) - \ln\sqrt{|\boldsymbol{S}_i|} - \ln(2\pi)^{m/2}\right] \tag{72}$$

$$= -\sum_{i=1}^{N}\left[\left(-\frac{1}{2}(\boldsymbol{y}_i - \boldsymbol{y}_{i|i-1})^{\mathrm{T}}\boldsymbol{S}_i^{-1}(\boldsymbol{y}_i - \boldsymbol{y}_{i|i-1})\right) - \frac{1}{2}\ln|\boldsymbol{S}_i| - \frac{m}{2}\ln(2\pi)\right] \tag{73}$$

$$= \frac{1}{2}\sum_{i=1}^{N}\left[\boldsymbol{\epsilon}_i^{\mathrm{T}}\boldsymbol{S}_i^{-1}\boldsymbol{\epsilon}_i + \ln|\boldsymbol{S}_i|\right] + \frac{mN}{2}\ln(2\pi) \tag{74}$$

The last term is a constant that does not affect the optimization. Thus the cost function to be minimized is

$$J(\boldsymbol{\theta}) = \sum_{i=1}^{N}\left[\boldsymbol{\epsilon}_i^{\mathrm{T}}\boldsymbol{S}_i^{-1}\boldsymbol{\epsilon}_i + \ln|\boldsymbol{S}_i|\right] \tag{75}$$

where $\boldsymbol{S}_i$ and $\boldsymbol{\epsilon}_i$ depend on the unknown parameters $\boldsymbol{\theta}$. To minimize this cost function we may use several techniques, for example, gradient descent. Implementing gradient descent requires computing the partial derivative of the cost with respect to each parameter. The cost depends on the Kalman filter equations given above. A review of how these partial derivatives can be computed (analytically or numerically) is presented in [2, 11.1].

Maximum Likelihood Estimate via Gradient Descent

One approach is to approximate $J(\boldsymbol{\theta})$ as a Taylor series and use a Newton-Raphson scheme (i.e., gradient descent). Assume that $\boldsymbol{\theta}$ can be expressed as a small perturbation from a nominal estimate (i.e., $\boldsymbol{\theta} = \boldsymbol{\theta}_0 + \Delta\boldsymbol{\theta}$). Then,

$$J(\boldsymbol{\theta}_0 + \Delta\boldsymbol{\theta}) = J(\boldsymbol{\theta}_0) + \Delta\boldsymbol{\theta}^T \frac{\partial J}{\partial \boldsymbol{\theta}}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} + \frac{1}{2}\Delta\boldsymbol{\theta}^T \frac{\partial^2 J}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \Delta\boldsymbol{\theta} + \text{h.o.t.} \tag{76}$$

$$\approx J(\boldsymbol{\theta}_0) + \Delta\boldsymbol{\theta}^T \frac{\partial J}{\partial \boldsymbol{\theta}}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} + \frac{1}{2}\Delta\boldsymbol{\theta}^T \frac{\partial^2 J}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \Delta\boldsymbol{\theta} \tag{77}$$

A necessary condition for $J(\boldsymbol{\theta}_0 + \Delta\boldsymbol{\theta})$ to be a minimum is

$$\frac{\partial}{\partial\Delta\boldsymbol{\theta}}[J(\boldsymbol{\theta}_0 + \Delta\boldsymbol{\theta})] = 0 \tag{78}$$

Applying this necessary condition to (77):

$$0 = \frac{\partial}{\partial\Delta\boldsymbol{\theta}}[J(\boldsymbol{\theta}_0 + \Delta\boldsymbol{\theta})] = \frac{\partial}{\partial\Delta\boldsymbol{\theta}}\left(J(\boldsymbol{\theta}_0) + \Delta\boldsymbol{\theta}^T \frac{\partial J}{\partial \boldsymbol{\theta}}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} + \frac{1}{2}\Delta\boldsymbol{\theta}^T \frac{\partial^2 J}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \Delta\boldsymbol{\theta}\right) \tag{79}$$

$$0 = 0 + \frac{\partial}{\partial\Delta\boldsymbol{\theta}}\left(\Delta\boldsymbol{\theta}^T \frac{\partial J}{\partial \boldsymbol{\theta}}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right) + \frac{\partial}{\partial\Delta\boldsymbol{\theta}}\left(\frac{1}{2}\Delta\boldsymbol{\theta}^T \frac{\partial^2 J}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \Delta\boldsymbol{\theta}\right) \tag{80}$$

$$0 = \frac{\partial J}{\partial \boldsymbol{\theta}}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} + \frac{\partial^2 J}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \Delta\boldsymbol{\theta} \tag{81}$$

Solving this last equation gives a vector of parameter changes

$$-\frac{\partial J}{\partial \boldsymbol{\theta}}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} = \frac{\partial^2 J}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \Delta\boldsymbol{\theta} \tag{82}$$

$$\Delta\boldsymbol{\theta} = -\left[\frac{\partial^2 J}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right]^{-1} \frac{\partial J}{\partial \boldsymbol{\theta}}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \tag{83}$$

which can be implemented as the iterative gradient descent algorithm:

$$\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_0 + \Delta\hat{\boldsymbol{\theta}} \tag{84}$$

References

[1] Eugene A Morelli and Vladislav Klein. *Aircraft System Identification: Theory and Practice*.

[2] Bruce P Gibbs. *Advanced Kalman Filtering, Least-squares and Modeling*. John Wiley & Sons, 2011.

[3] Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.