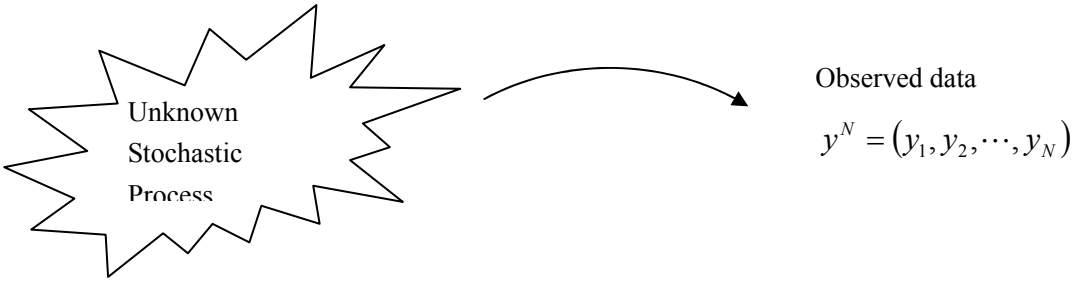
2.160 System Identification, Estimation, and Learning
Lecture Notes No. 20
May 3, 2006

15. Maximum Likelihood

15.1 Principle

Consider an unknown stochastic process



Assume that each observed datum y_i is generated by an assumed stochastic process having a PDF:

$$f(m, \lambda; x) = \frac{1}{\sqrt{2\pi\lambda}} e^{-\frac{(x-m)^2}{\lambda}} \tag{1}$$

where m is mean, λ is variance, and x is the random variable associated with y_i. We know that mean m and variance λ are determined by

$$m = \frac{1}{N}\sum_{i=1}^{N} y_i \qquad \lambda = \frac{1}{N}\sum_{i=1}^{N} (y_i - m)^2 \tag{2}$$

Let us now obtain the same parameter values, m and λ, based on a different principle: Maximum Likelihood.

Assuming that the N observations $y_1, y_2, \cdots, y_N$ are stochastically independent, consider the joint probability associated with the *N* observations:

$$f(m, \lambda; x_1, x_2, \cdots x_N) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\lambda}} e^{-\frac{(x_i-m)^2}{\lambda}} \tag{3}$$

Now, once $y_1, y_2, \cdots, y_N$ have been observed (have taken specific values), what parameter values, m and λ, provide the highest probability in $f(m, \lambda; x_1, x_2, \cdots x_N)$? In other words, what values of m and λ are most likely the case? This means that we maximize the following functional with respect to m and λ:

$$\underset{m,\lambda}{Max}\, f(m, \lambda; y_1, y_2, \cdots y_N) \tag{4}$$

Note that $y_1, y_2, \cdots, y_N$ are known values. Therefore, $f(m, \lambda; y_1, y_2, \cdots y_N)$ is a function of m and λ only. Using our standard notation, this can be rewritten as

$$\hat{\theta} = \arg\max_{\theta} f(m, \lambda; y^N) \tag{5}$$

where $\hat{\theta}$ is estimate of m and λ. Maximizing $f(m, \lambda; y^N)$ is equivalent to maximizing $\log f(\quad)$,

$$\begin{aligned} \hat{\theta} &= \arg\max_{\theta} \left[\log f(m, \lambda; y^N)\right] \\ &= \arg\max_{\theta} \sum \left(\log \frac{1}{\sqrt{2\pi\lambda}} - \frac{(y_i - m)^2}{2\lambda} \right) \end{aligned} \tag{6}$$

Taking derivatives and setting them to zero,

$$\frac{\partial \log f}{\partial m} = \sum_{i=1}^{N} \frac{1}{\lambda}(y_i - m) = 0 \tag{7}$$

$$\Longrightarrow \qquad m = \frac{1}{N}\sum_{i=1}^{N} y_i \tag{9}$$

$$\frac{\partial \log f}{\partial \lambda} = N \frac{d \log \pi^{-\frac{1}{2}}}{d\lambda} - \frac{d}{d\lambda}\lambda^{-1} \sum_{i=1}^{N} \frac{(y_i - m)^2}{2} = 0 \tag{8}$$

$$\Longrightarrow \qquad \lambda = \frac{1}{N}\sum_{i=1}^{N} (y_i - m)^2 \tag{10}$$

The above results $m = \frac{1}{N}\sum_{i=1}^{N} y_i$ and $\lambda = \frac{1}{N}\sum_{i=1}^{N} (y_i - m)^2$ provide a stochastic process model that is most likely to generate the observed data y^N. And these agree with (2)

This Maximum Likelihood Estimate (MLE) is formally stated as follows.

Maximum Likelihood Estimate

Consider a joint probability density function with parameter vector θ as a stochastic model of an unknown process:

$$f(\theta; x_1, x_2, \cdots x_N) \tag{11}$$

Given observed data $y_1, y_2, \cdots, y_N$ form a deterministic function of θ, called the likelihood function:

$$L(\theta) = f(\theta; y_1, y_2, \cdots y_N) \tag{12}$$

Determine parameter vector θ so that this likelihood function becomes maximum.

$$\hat{\theta} = \arg\max_{\theta} L(\theta) \tag{13}$$

This Maximum Likelihood Estimator was found by Ronald A. Fisher in 1912, when he was an undergraduate junior at Cambridge University. He was 22 years old.

Don't be disappointed. You are still young!

15.2 Likelihood Function for Probabilistic Models of Dynamic Systems

Consider a model set in predictor form

$$M(\theta): \qquad \hat{y}(t|\theta) = g(t, z^{t-1}; \theta) \tag{14}$$

and an associated PDF of prediction error

$$f_\varepsilon(x, t; \theta) \tag{15}$$

$$\varepsilon(t; \theta) = y(t) - \hat{y}(t|\theta) \tag{16}$$

Assume that the random process $\varepsilon(t)$ is independent in t, and consider the joint probability density function:

$$f(\theta; x_1, x_2, \cdots x_N) = \prod_{t=1}^{N} f_\varepsilon(x_t, t; \theta) \tag{17}$$

Substituting observed data $y_1, y_2, \cdots, y_N$ into (17) yields a likelihood function of parameter θ,

$$L(\theta) = \prod_{t=1}^{N} f_\varepsilon(y(t) - \hat{y}(t|\theta), t; \theta) \tag{18}$$

Maximizing this function is equivalent to maximizing its logarithmic function,

$$\frac{1}{N} \log L(\theta) = \frac{1}{N} \sum_{t=1}^{N} \log f_\varepsilon(\varepsilon, t; \theta) \tag{19}$$

Replacing $\log f_\varepsilon$ by $l(\varepsilon, \theta, t) = -\log f_\varepsilon(\varepsilon, t; \theta)$,

$$\hat{\theta}^{ML}(Z^N) = \arg\max_\theta \frac{1}{N} \sum_{t=1}^{N} l(\varepsilon, \theta, t) \tag{20}$$

The above minimization is for a deterministic function, and is analogous to the prediction-error method with the least square criterion.

$$\hat{\theta}^{LS}(z^N) = \arg\max_\theta \frac{1}{N} \sum_{t=1}^{N} \frac{1}{2} \varepsilon^2(t, \theta) \tag{21}$$

$$\Big\downarrow \quad \longleftarrow \text{ Replacing } \frac{1}{2}\varepsilon^2(t, \theta) = l(\varepsilon, \theta, t)$$

$$\hat{\theta}^{LS}(z^N) = \arg\max_\theta \frac{1}{N} \sum_{t=1}^{N} l(\varepsilon, \theta, t)$$

Therefore, the Maximum Likelihood Estimate is a special case of the prediction-error method. Function $l(\varepsilon,\theta,t)$ may take a different form depending on the assumed PDF of the process under consideration.

15.3 The Cramer-Rao Lower Bound

Consider the mean square error matrix of an estimate $\hat{\theta}_N$ compared to its true value θ_0,

$$P = E\left[\left(\hat{\theta}(y^N)-\theta_0\right)\left(\hat{\theta}(y^N)-\theta_0\right)^T\right] \tag{22}$$

This mean square error matrix provides the overall quality of an estimate. Our goal is to minimizing a lower limit to this matrix. The limit, called the Cramer-Rao Lower Bound, gives the best performance that we can expect among many unbiased estimators.

Theorem (***The Cramer-Rao Lower Bound***)

Let $\hat{\theta}(y^N)$ be an unbiased estimate of θ;

$$E[\hat{\theta}(y^N)] = \theta_0 \tag{23}$$

where the PDF of y^N is $f_y\left(\theta_0; y^N\right)$. Then

$$E\left[\left(\hat{\theta}(y^N)-\theta_0\right)\left(\hat{\theta}(y^N)-\theta_0\right)^T\right] \geq M^{-1} \tag{24}$$

where M is the ***Fisher Information Matrix*** given by

$$M = E\left[\frac{d}{d\theta}\log f_y(\theta;x^N)\right]\left[\frac{d}{d\theta}\log f_y(\theta;x^N)\right]^T\Big|_{\theta_0} \qquad (25-a)$$

$$= -E\left[\frac{d^2}{d\theta^2}\log f_y(\theta;x^N)\right]\Big|_{\theta_0} \qquad (25-b)$$

Proof

From (23)

$$\theta_0 = \int_{R^N}\hat{\theta}\left(x^N\right)f_y(\theta_0;x^N)dx^N \tag{26}$$

where dummy vector x^N is a $N\times 1$ vector in space R^N.

Differentiating (26), which is a $d\times 1$ vector equation, with respect to $\theta_0 \in R^{d\times 1}$ yields

$$I = \int_{R^N} \hat{\theta}\left(x^N\right)\left[\frac{d}{d\theta_0} f_y(\theta_0; x^N)\right]^T dx^N \tag{27}$$

where I is the $d \times d$ identity matrix,

$$I = \begin{bmatrix} 1 & & 0 \\ & 1 & \\ & \ddots & \\ 0 & & 1 \end{bmatrix} \in R^{d \times d}$$

Using the following relationship

$$\frac{d}{d\theta} \log f(\theta) = \frac{1}{f(\theta)} \frac{d}{d\theta} f(\theta) \tag{28}$$

(27) can be written as

$$\begin{aligned} I &= \int_{R^N} \hat{\theta}\left(x^N\right)\left[\frac{d}{d\theta_0} \log f_y(\theta_0; x^N)\right]^T f_y(\theta_0; x^N) dx^N \\ &= E\left[\hat{\theta}\left(x^N\right)\left[\frac{d}{d\theta_0} \log f_y(\theta_0; x^N)\right]^T\right] \end{aligned} \tag{29}$$

Similarly, differentiating the identity: $1 = \int_{R^N} f_y(\theta_0; x^N) dx^N$ yields

$$0 = \int_{R^N} \left[\frac{d}{d\theta_0} f_y(\theta_0; x^N)\right]^T dx^N = \int_{R^N} \left[\frac{d}{d\theta_0} \log f_y(\theta_0; x^N)\right]^T f_y(\theta_0; x^N) dx^N$$

$$\therefore \quad 0 = E\left[\frac{d}{d\theta_0} \log f_y(\theta_0; x^N)\right]^T \tag{30}$$

Multiplying (30) by θ_0 and subtracting it from (29),

$$E\left\{\underbrace{\left[\hat{\theta}(y^N) - \theta_0\right]}_{\alpha \in R^{d \times 1}} \quad \underbrace{\left[\frac{d}{d\theta_0} \log f_y(\theta_0; x^N)\right]^T}_{\beta^T \in R^{1 \times d}}\right\} = I \tag{31}$$

$$E\left(\alpha\beta^T\right) = I$$

Consider $2d \times 1$ vector $\begin{pmatrix} \alpha \\ \beta \end{pmatrix}$,

$$E\left[\begin{pmatrix}\alpha\\ \beta\end{pmatrix}\quad\begin{pmatrix}\alpha^T & \beta^T\end{pmatrix}\right]=\begin{bmatrix}E\alpha\alpha^T & E\alpha\beta^T\\ E\beta\alpha^T & E\beta\beta^T\end{bmatrix}\geq 0 \qquad \text{Positive Semi-definite} \quad (32)$$

Note $E\alpha\beta^T = I$, and $E\beta\alpha^T = E\left(\alpha\beta^T\right)^T = I$

$$\therefore \quad E\left(\alpha\alpha^T\right)\cdot E\left(\beta\beta^T\right) - I^2 \geq 0 \tag{33}$$

$$E\left(\alpha\alpha^T\right) \geq \left[E\left(\beta\beta^T\right)\right]^{-1} \tag{34}$$

This gives the first equation, (25)-(a). Differentiating the transpose of (30) yields

$$\begin{aligned} 0 = &\int_{R^N}\left[\frac{d^2}{d\theta_0^{\,2}}\log f_y(\theta_0;x^N)\right]f_y(\theta_0;x^N)dx^N \\ &+\int_{R^N}\left[\frac{d}{d\theta_0}\log f_y(\theta_0;x^N)\right]\left[\frac{d}{d\theta_0}f_y(\theta_0;x^N)\right]^T dx^N \end{aligned} \tag{35}$$

Rewriting this in Expectation form yields (25)-(b). Q.E.D.

15.4 Best Unbiased Estimators for Dynamical Systems.

The Cramer-Rao Lower Bound, given by the Fisher Information Matrix M, is the best performance we can expect among all unbiased estimates with regard to error covariance. Interestingly, Maximum Likelihood Estimate provides this best error covariance performance for large sample data. The following is to formulate the problem for proving this. The complete proof, however, is left for your Extra-Credit Problem.

Based on the logarithmic likelihood function, we can compute the Fisher Information Matrix, the inverse of which provides the lower bound of the error covariance. To simplify the problem, we assume that the PDF of prediction error ε is a function of ε only. Namely,

$$\begin{aligned} &\frac{\partial f_\varepsilon}{\partial\theta}=0 \qquad \frac{\partial f_\varepsilon}{\partial t}=0 \\ &f_\varepsilon(\varepsilon,t,\theta)\xrightarrow{\text{Reduces to}} f_\varepsilon(\varepsilon) \end{aligned} \tag{36}$$

Note, however, that the total derivative of f_ε w.r.t. θ is not zero, since ε depends on θ; $\varepsilon(t;\theta) = y(t) - \hat{y}(t|\theta)$. Denoting $l_0(\varepsilon) \equiv -\log f_\varepsilon(\varepsilon)$, we can re-write (19), that is the log likelihood function of a dynamical system is given by

$$\log L(\theta) = \sum \log f_\varepsilon(\varepsilon) = -\sum_{t=1}^{N} l_0\left(\varepsilon(t,0)\right) \tag{37}$$

Differentiating this w.r.t. θ

$$\frac{d}{d\theta}\log L(\theta) = -\sum_{t=1}^{N}\frac{dl_0}{d\varepsilon}\frac{d\varepsilon}{d\theta} = \sum_{t=1}^{N} l_0'[\varepsilon(t,\theta)]\psi(t,\theta) \tag{38}$$

Note that $\frac{d\varepsilon}{d\theta} = -\frac{d\hat{y}}{d\theta} = -\psi$. Replacing $\frac{d}{d\theta}\log f_y$ in (25) by the above equation yields the Fisher Information Matrix M, evaluated at the true parameter value θ_0. With this θ_0, we can assume that the prediction error sequence $\varepsilon(t,\theta_0)$ are generated as a sequence of independent random variables $\varepsilon(t,\theta_0) = e_0(t)$ whose PDF is $f_\varepsilon(x)$. This leads to the Fisher Information Matrix given by

$$M = \frac{1}{k_0}\sum_{t=1}^{N} E\left[\psi(t,\theta_0)\psi^T(t,\theta_0)\right] \qquad \frac{1}{k_0} \equiv \int_{-\infty}^{\infty}\frac{[f_e'(x)]}{[f_e(x)]}dx$$

on the other hand, as N tends to infinity, the prediction-error method provides the asymptotic variance P_θ, as discussed last time. From (20) we can treat Maximum Likelihood Estimate as a special case of the prediction-error method. Therefore, the previous result on asymptotic variance applies to MLE.

Extra Credit Problem

Obtain the asymptotic variance of MLE for dynamical systems, and show that it agrees with the Fisher Information Matrix. Namely, Maximum Likelihood Estimate provides the best unbiased estimate for large *N*.