2.160 System Identification, Estimation, and Learning

Lecture Notes No. 24

May 8, 2006

16. Information Theory of System Identification

16.1 Overview

Maximum Likelihood Estimate (MLE)

$$\hat{\theta}^{ML}(Z^N) = \arg\max_{\theta} \log L(\theta) \tag{1}$$

Likelihood function

$$L(\theta) = \Pi f(\theta; y_1, \cdots, y_N) \tag{2}$$

For dynamical systems

$$L(\theta) = \Pi f_{\varepsilon}\left(y(t) - \hat{y}(t|\theta), t; \theta\right) \tag{3}$$

$$\hat{\theta}^{ML}(Z^N) = \arg\max_{\theta} \frac{1}{N}\sum_{t=1}^{N} \underbrace{\log f_{\varepsilon}(\varepsilon, t; \theta)}_{\|} \tag{4}$$

The logarithm of probability

Entropy: $-\sum_i p_i \log p_i \longrightarrow \downarrow$ (5)

Information Theory

Model-data agreement is quantified with regard to the amount of information.

Remember that a "good" model fully exploits all information contained in data.

- Least Square Estimate and the correlation method
  Prediction error and data are orthogonal or uncorrelated to each other.
- Kalman Filter
  The innovation process $e(t)$ is a white noise random process.

- Maximum Likelihood Estimate and the Information Theoretic approach
  The logarithmic joint probability of prediction error is maximized.

  ↓

  The degree of randomness in prediction error is maximized.

  ↓

  The Entropy Maximization Principle

  <u>The Punch Line</u>

In the information theoretic approach, we use "information" as a generic measure for evaluating how much a model fits a given set of data. This unified measure allows us to compare diverse model structures on the same basis, and provides an objective means to select an optimal model based on the trade-off between estimate bias and variance, or accuracy v.s. reliability.

The expected end results

Akaike's Information Criterion

$$\hat{\theta}^{AIC}(Z^N) = \arg\min_{\theta} \frac{1}{N}\left(\sum_{t=1}^{N} \log \frac{1}{L(\theta)} + \dim\theta\right) \qquad (6)$$

Dimension of parameter vector θ
||
The number of parameters
||
The penalty of using many parameters

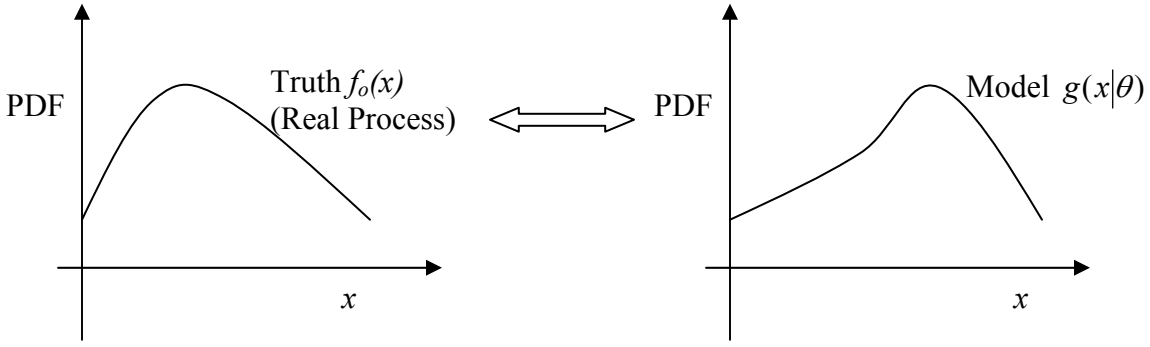
MDL (Minimum Description Length) by Rissanen

A good mode= the shortcut code describing the data

16.2 The Kullback Leibler Information Distance

Measuring the distance between the true system and its model;

How much the model differs from the truth?



If $f_0(x) = g(x|\theta)$ for all x, then the distance must be zero.

To evaluate the overall difference between the model and the truth, consider the following measure:

$$I(f_0, g_m) = \int_{R^N} f_0(x) \log \frac{f_0(x)}{g_m(x(\theta))} dx \qquad (7)$$

Called the Kullback-Leibler information Distance

Properties of $I(f, g)$

1) $I(f, g) \geq 0$ for all PDF f and g

Proof: Since f and g are PDF

$$f(x) \geq 0 \quad \int f(x)dx = 1$$

$$g(x) \geq 0 \quad \int g(x)dx = 1$$

Define

$$h(x) \equiv \frac{g(x)}{f(x)} - 1 \text{ or } \frac{g(x)}{f(x)} = h(x) + 1 \qquad (8)$$

Note that

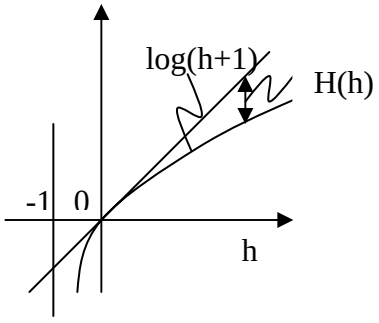
$$\int h(x) f(x)dx = \int \big(g(x) - f(x)\big)dx = 0 \qquad (9)$$

Adding this to $I(f, g) = \int_{R^N} f(x) \log \frac{f(x)}{g(x)} dx$ yields

$$I(f, g) = \int h(x) f(x)dx - \int f(x) \log \frac{g(x)}{f(x)} dx \qquad (10)$$

$$= \int f(x)\underbrace{[h(x) - \log(h(x) + 1)]}_{\| \; H(h)} dx$$

As shown right, $h \geq \log(h + 1)$ for all h>-1. Therefore $H(h) \geq 0$
This implies $I(f, g) \geq 0$ Q.E.D.



2) $I(f, g) = 0$ only when $f(x) \equiv g(x)$ for all x

$$I(f, g) = \int f(x) H(h(x))dx = 0$$

This implies $H(h(x)) \equiv 0$ for all x s.t. $f(x) \neq 0$.

From the above plot, H(h)=0 only when h=0.

Therefore, $H(h(x)) \equiv 0 \implies g(x) \equiv f(x)$ Q.E.D.

Remarks
a) Properties 1) and 2) are fundamental requirements for I(f,g) to be a "distance". However, this distance is not commutable:

$$I(f,g) \neq I(g,f) \tag{11}$$

b) This is equivalent to the negative of Boltzmaun's eutropy when taking mean

$$-\log\frac{f(x)}{g(x)} \rightarrow \int f(x)\log\frac{f(x)}{g(x)}dx$$

Summary so far

Maximum Likelihood Estimate for Dynamical Systems

$$L(\theta) = \prod_{t=1}^{N} f_{\varepsilon}\left(y(t) - \hat{y}(t|\theta), t; \theta\right) \tag{12}$$

We are particularly interested in a simple case where $f_{\varepsilon} = f_{\varepsilon}(\varepsilon)$.

The log likelihood function is then given by

$$\log L(\theta) = \sum_{t=1}^{N} \log f_{\varepsilon}(\varepsilon) \tag{13}$$

Comparing this with LSE where $\varepsilon^2(t)$ is used as a penalty for estimation error ε, the log likelihood for the fixed PDF $f_{\varepsilon}(\varepsilon)$ is interpreted as another way of penalizing error ε. As shown in the figures, the curve of $-\log f_{\varepsilon}$ is very similar to ε^2. Use of $-\log f_{\varepsilon}$, i.e. MLE, has two salient features:

1) MLE provides the best unbiased estimate with the error covariance as small as the Cramer-Rao Lower Bound
2) $-\log f_{\varepsilon}$ pertains to the amount of information and information theory.

16.3 Re-formulating the Kullback-Leibler Distance

Using this information measure, Kullback and Leibler have quantified the distance (disagreement) between a model $g(x|\theta) = L(\theta)$ (for given data x or y) and the truth f(x),

$$I(f, g(\cdot|\theta)) = \int f(x) \log \frac{f(x)}{g(x|\theta)} dx \tag{14}$$

Note that variable *x* has been integrated out, hence *I(f, g)* does not depend on *x*. The greatest difficulty with the K-L distance is that, since f(x) is unknown, it is not computable. It is merely a conceptual measure. Akaike resolved this problem by reformulating the distance measure.

Suppose that observed data y have been sampled with the true PDF f(x). Let $\hat{\theta}(y)$ be the maximum likelihood estimate of θ for given data y. The K-L distance for $\hat{\theta}(y)$ is then,

$$I(f, g(\cdot|\hat{\theta}(y))) = \int f(x) \log \frac{f(x)}{g(x|\hat{\theta}(y))} dx \tag{15}$$

Repeat the sampling many times and compute the average of the above distance:

$$E_y\left[I(f, g(\cdot|\hat{\theta}(y)))\right] = \int f(y) I(f, g(\cdot|\hat{\theta}(y))) dy \tag{16}$$

Instead of minimizing the original K-L distance (14) for obtaining MLE of θ, consider to minimize the average of the K-L distance (16).

The average K-L distance is not only computable, as will be shown next, but also useful for performing so-called "cross validation", an important step needed for assuming the validity of a model.

(16) can be re-written as

$$\begin{aligned} E_y\left[I(f, g(\cdot|\hat{\theta}(y)))\right] &= \int f(y) \left\{ \int \left[f(x) \log f(x) - f(x) \log g(x|\hat{\theta}(y)) \right] dx \right\} dy \\ &= \int \underbrace{f(x) \log f(x) dx}_{\text{Constant Irrelevant to minimization w.r.t. } \theta} - \underbrace{E_y E_x\left[\log g(x|\hat{\theta}(y))\right]}_{\text{T for target}} \end{aligned} \tag{17}$$

Therefore, minimizing $E_y\left[I(f, g(\cdot|\hat{\theta}(y)))\right]$ is equivalent to maximizing T,

$$T = E_y E_x\left[\log g(x|\hat{\theta}(y))\right] = \int f(x) \left[f(x) \log g(x|\hat{\theta}(y)) dx \right] dy \tag{18}$$

16.4 Computation of Target *T*

Consider the second-order Taylor expansion of function $h(\hat{\theta})$

$$h(\hat{\theta}) = h(\theta_0) + \left[\frac{\partial h(\theta)}{\partial \theta}\right]_{\theta_0}^T \left(\hat{\theta} - \theta_0\right) + \frac{1}{2}\left(\hat{\theta} - \theta_0\right)^T \left[\frac{\partial^2 h(\theta)}{\partial \theta^2}\right]_{\theta_0} \left(\hat{\theta} - \theta_0\right) + O_p\left(\frac{1}{N}\right) \qquad (19)$$

Recall the asymptotic variance analysis we did for estimation error covariances as N tends infinity. The error of the above second order approximation is generally on the order of 1/N in average, as N tends to infinity.

Applying this approximation to $\log g\left(x\middle|\hat{\theta}\right)$ yields

$$\log g(x|\hat{\theta}) = \log g(x|\theta_0) + \left[\frac{\partial \log g(x|\theta)}{\partial \theta}\right]_{\theta_0}^T \left(\hat{\theta} - \theta_0\right) + \frac{1}{2}\left(\hat{\theta} - \theta_0\right)^T \left[\frac{\partial^2 \log g(x|\theta))}{\partial \theta^2}\right]_{\theta_0} \left(\hat{\theta} - \theta_0\right) + O_p\left(\frac{1}{N}\right) \qquad (20)$$

Taking expectation of (20) with respect to x, which is uncorrelated with $\hat{\theta}$,

$$E_x[\log g(x|\hat{\theta})] = E_x[\log g(x|\theta_0)] + \underbrace{\left(E_x\left[\frac{\partial \log g(x|\theta)}{\partial \theta}\right]^T\right)}_{\|\ 0}\left(\hat{\theta} - \theta_0\right) + \frac{1}{2}\left(\hat{\theta} - \theta_0\right)^T \underbrace{\left(E_x\left[\frac{\partial^2 \log g(x|\theta))}{\partial \theta^2}\right]\right)}_{\|\ -I(\theta_0)}\left(\hat{\theta} - \theta_0\right) \qquad (21)$$

Since θ_0 is the optimal one minimizing $E_x[\log\, g(x|\hat{\theta})]$, its derivative must be zero.

Note that $I(\theta_0)$ is equivalent to the Fisher Information Matrix, if the model with the optimal θ_0 is the true system; $g(x|\theta_0) = f(x)$.

In the following derivation, we do not assume $g(x|\theta_0) = f(x)$.

The third term $\left(\hat{\theta} - \theta_0\right)^T I(\theta_0)\left(\hat{\theta} - \theta_0\right)$ can be rewritten by using the formula

$$a^T B a = Trace[B \cdot aa^T] \qquad a \in R^{n\times 1} \qquad B \in R^{n\times n} \qquad (22)$$

Prove this. Not difficult.

Then (21) reduces to

$$E_x[\log g(x|\hat{\theta})] = E_x[\log g(x|\theta_0)] - \frac{1}{2} Trace\left[I(\theta_0)(\hat{\theta} - \theta_0)(\hat{\theta} - \theta_0)^T \right] \quad (23)$$

Taking expectation of (23) w.r.t. y yields.

$$E_{\hat{\theta}} E_x[\log g(x|\hat{\theta})] = E_x[\log g(x|\theta_0)] - \frac{1}{2} Trace\left[I(\theta_0) E_{\hat{\theta}}\left[(\hat{\theta} - \theta_0)(\hat{\theta} - \theta_0)^T \right] \right] \quad (24)$$

Note that, since y is involved only in $\hat{\theta}$, taking expectation w.r.t. y is equivalent to that of $\hat{\theta}$.

Therefore,

$$T = E_x[\log g(x|\theta_0)] - \frac{1}{2} Trace\left[I(\theta_0) \sum \right] \quad (25)$$

This shows the trade-off between the model-data agreement accuracy and the magnitude of error covariance, which we want to optimize.

(25) is a function of θ_0, however. Next we need to replace it by $\hat{\theta}$. To this end, we take the Taylor expansion of $\log g(x|\theta_0)$ about $\hat{\theta}(x)$, treating x as sample data,

$$\log g(x|\theta_0) = \log g(x|\hat{\theta}) + \underbrace{\left[\frac{\partial \log g(x|\hat{\theta})}{\partial \theta} \right]^T}_{\| \; 0} (\theta_0 - \hat{\theta}) + \frac{1}{2}(\theta_0 - \hat{\theta})^T \underbrace{\left[\frac{\partial^2 \log g(x|\hat{\theta}))}{\partial \theta^2} \right]}_{\| \; \hat{I}(\hat{\theta})} (\theta_0 - \hat{\theta}) + O_p\left(\frac{1}{N} \right) \quad (26)$$

Since $\hat{\theta}(x)$ is MLE

Taking expectation w.r.t. x (Note $\hat{\theta}(x)$ depends on x)

$$E_x[\log g(x|\theta_0)] = E_x[\log g(x|\hat{\theta})] - \frac{1}{2} Trace\left[I(\theta_0) \sum \right] \quad (27)$$

Substituting (27) into (25) yields

$$T = E_x[\log g(x|\hat{\theta})] - \frac{1}{2} Trace\left[I(\theta_0) \sum \right] \quad (28)$$

If the true system is involved in the model set, $g(x|\theta_0) = f(x)$, then $I(\theta_0)$ becomes the Fisher Information Matrix, which is the inverse of the error covariance matrix $\sum^{-1}$, since MLE provides the lower bound of Cramer and Rao.

Therefore $I(\theta_0) \sum =$ Identity Matrix.

$Trace\left[I(\theta_0)\sum\ \right] = d$: the number of parameters

$$T = E_x[\log g(x|\hat{\theta})] - d \tag{29}$$

16.5 Akaike's Information Criterion (AIC)

Akaike, 1973.

The unbiased estimator that minimizes the average K-L distance $E_y\left[I(f, g(\cdot|\hat{\theta}(y)))\right]$ for a "good" model, i.e. $g(x|\theta_0) = f(x)$, and large sample data is given by

$$\hat{\theta}^{AIC} = -2\log\left[L\left(\hat{\theta}|y\right)\right] + 2d \tag{30}$$

where *d* is the number of parameters, θ.

If $g(x|\theta_0) = f(x)$ is not assumed, *d* must be replaced by $Trace\left[I(\theta_0)\sum\ \right]$ (Takeuchi, 1976).

For a sufficiently large number of data Ex[ ] is redundant,

$$\hat{\theta}^{AIC} = \arg\min_{\hat{\theta}}\left(\left[-\log g\left(x|\theta\right)\right] + d\right) \tag{31}$$

The Principle of Parsimony

The trade-off can be made by minimizing the AIC for both $\hat{\theta}$ and model structure $g(x|\theta)$ and d….. Determining the model structure from data.

$$AIC = -2\log\{L[\hat{\theta}|y]\} + 2d \tag{32}$$

The same as $g(x|\hat{\theta})$

The multiplier (-2) comes from a historical reason.

$(-2)\log\left[\dfrac{Likelihood1}{Likelihood2}\right] =$ Asymptotically chi-squares x^2