CH5350: Applied Time-Series Analysis

Arun K. Tangirala

Department of Chemical Engineering, IIT Madras

Fisher's Information and Properties of Estimators

Learning Goals

In this lecture, we shall learn the following concepts / topics:

- Goodness of estimators
- Fisher information
- Bias and Variance
- Efficiency and C-R Inequality
- Mean Square Error and MMSE
- Consistency
- Distribution of estimates

Goodness of estimators

In the previous lecture, we discussed the quality of an estimator and preliminaries on a few measures of goodness (or the performance of an estimator), e.g., bias, variance.

Of particular interest are the efficiency (related to variance) and consistency (concerned with convergence of estimates).

Metrics characterize the estimator; however, remember that **a fundamental requirement for obtaining a good estimate is that the data should be informative** (with respect to the parameters).

Fisher information

Fisher introduced the notion of information in a data through a series of works by and some existing results. Intuitively, larger the information index is, the "better" the estimator is.

The Fisher information (FI) (Fisher, 1922, 1950) is based on the **likelihood function** of the given data.

The likelihood function stems from the notion of conditional probability, i.e., the probability of observing an event within the vicinity of given data.

Likelihood function

The probability of obtaining data within the vicinity of $\mathbf{y}_N$ is given by (with some abuse of notation)

$$\Pr(\mathbf{y}_N < \mathbf{Y} < \mathbf{y}_N + d\mathbf{y}_N) = f(\mathbf{y}_N|\boldsymbol{\theta})d\mathbf{y}_N \propto f(\mathbf{y}_N|\boldsymbol{\theta}) \tag{1}$$

For a given $\mathbf{y}_N$, the probability is solely a function of $\boldsymbol{\theta}$. Fisher's argument (and the likes of it) rests on the **maximum likelihood** premise that

Among all possible values of $\boldsymbol{\theta}$, the one that maximizes the probability, i.e., the one that renders the event most likely is the winner!

Likelihood function

The likelihood function (of $\boldsymbol{\theta}$) is, therefore (for continuous RVs), defined as

$$\boxed{l(\boldsymbol{\theta}, \mathbf{y}) = f(\mathbf{y}; \boldsymbol{\theta})} \quad (\text{or } f(\mathbf{y}|\boldsymbol{\theta})) \tag{2}$$

where $\mathbf{y}$ is the vector of N observations.

- The fundamental difference between $l(\boldsymbol{\theta}|\mathbf{y})$ and $f(\mathbf{y}|\boldsymbol{\theta})$ is that the former is a function of a *deterministic* vector $\boldsymbol{\theta}$, while the latter is a function of the *random* vector $\mathbf{y}$ (given $\boldsymbol{\theta}$).
- Likelihood function belongs to the world of statistics while the p.d.f. belongs to the world of **probability**!

Fisher information . . . contd.

Fisher's information quantifies "how informative" a vector of observations is about a parameter θ (or $\boldsymbol{\theta}$). It rests on the following quantities (assume **single parameter**):

$$l(\theta, \mathbf{y}) = f(\mathbf{y}; \theta) \ \ (\text{or } f(\mathbf{y}|\theta)) \qquad \text{(likelihood function)} \tag{3}$$

$$L(\theta, \mathbf{y}) = \ln l(\theta, \mathbf{y}) \qquad \text{(log-likelihood function)} \tag{4}$$

$$S(\theta; \mathbf{y}) = \frac{\partial}{\partial \theta} \ln f(\mathbf{y}; \theta) = \frac{\partial}{\partial \theta} L(\theta, \mathbf{y}) \qquad \text{(score function)} \tag{5}$$

where $\mathbf{y}$ is the set of observations and θ is the parameter to be estimated.

Further **assume that the p.d.f. is regular** $\Longrightarrow$ (i) $\partial L/\partial \theta$ exists and is finite and (ii) the operations of integration w.r.t. y and differentiation w.r.t. θ can be interchanged.

Fisher information . . . contd.

FI measures the variability in sensitivity of likelihood, i.e., the score function, across the outcome space (of $\mathbf{y}$).

The **Fisher information** of a parameter θ in $\mathbf{y}$ is defined as

$$I(\theta) = \text{var}(S) = E\left(\left(\frac{\partial L}{\partial \theta}\right)^2\right) \tag{6}$$

Under the regularity assumption, it can be shown that

$$\mu_S = E(S|\theta) = 0, \qquad \text{var}(S|\theta) = E(S^2) = E\left(\left(\frac{\partial L(\mathbf{y}, \theta)}{\partial \theta}\right)^2\right) \tag{7}$$

Fisher information . . . contd.

Since

$$E\left(\left(\frac{\partial L}{\partial \theta}\right)^2\right) = -E\left(\frac{\partial^2 L}{\partial \theta^2}\right) \tag{8}$$

the information can also be computed as

$$\boxed{I(\theta) = -E\left(\frac{\partial^2 L}{\partial \theta^2}\right) = -E\left(\frac{\partial S}{\partial \theta}\right)} \tag{9}$$

Example 1: Information about mean and variance

Consider the case of estimating mean μ and variance σ^2 of a random signal.

Mean and variance

Given that a stationary signal $y[k] \sim \mathcal{N}(\mu, \sigma^2)$, determine (i) $I(\mu)$ and (ii) $I(\sigma^2)$ in a single observation.

1. The log-likelihood function (assuming σ^2 is known) is

$$L(\mu; Y) = \ln f(y|\mu) = -\frac{1}{2}\ln(2\pi\sigma^2) - \frac{1}{2}\frac{(y-\mu)^2}{\sigma^2} \tag{10}$$

Example 1 . . . contd.

The Fisher information on $\theta = \mu$ using (9) is then

$$I(\mu) = -E\left(\frac{\partial^2 L}{\partial \theta^2}\right) = \frac{1}{\sigma^2} \tag{11}$$

Thus, we have a meaningful result. As the variance (spread of possible outcomes) decreases, the information on μ in a *single sample* increases.

Example 1 . . . contd.

2. Now, $\theta = \sigma^2$. The information contained in a single observation is

$$I(\sigma^2) = -E\left(\frac{\partial^2 L}{\partial \theta^2}\right) = -E(\frac{1}{2\sigma^4} - \frac{(y-\mu)^2}{\sigma^6}) = \frac{1}{2\sigma^4} \tag{12}$$

Example 1 . . . contd.

3. On the other hand, if the parameter of interest is the standard deviation $\theta = \sigma$, the information contained is

$$I(\sigma) = -E\left(\frac{\partial^2 L}{\partial \theta^2}\right) = -E(\frac{1}{\sigma^2} - 3\frac{(y-\mu)^2}{\sigma^4}) = \frac{2}{\sigma^2} \tag{13}$$

Thus, $I(\sigma^2) \neq (I(\sigma))^2$. The *information is not commutative with respect to a functional of the parameter* $\phi(\theta)$.

In general, the FI $I(\phi(\theta))$ is related to $I(\theta)$ through

$$\boxed{I(\theta) = \left(\frac{d\phi}{d\theta}\right)^2 I(\phi(\theta))} \tag{14}$$

Fisher information: General case

Generalizing (9) to the case of $p \times 1$ parameter vector $\boldsymbol{\theta}$ contained in N observations, the **information matrix** results:

$$\mathbf{I}_{ij}(\boldsymbol{\theta}) = \text{cov}(S_i, S_j) = E(S_i(\mathbf{Y}_N)S_j(\mathbf{Y}_N)) = -E\left(\frac{\partial^2}{\partial\theta_i\partial\theta_j}L(\boldsymbol{\theta};\mathbf{y}_N)\right) \quad i,j = 1,\cdots,p \tag{15}$$

where S_i is the i^{th} score statistic,

$$S_i = \frac{\partial}{\partial\theta_i}\ln f(Y_N|\boldsymbol{\theta}) \tag{16}$$

where $f(Y_n|\boldsymbol{\theta})$ is the joint p.d.f. of the N observations $\mathbf{y}$.

Think: What do the off-diagonal elements signify?

Example 2: Estimating μ, σ^2 from N observations

Information in N observations

Compute the information contained in N samples of a GWN process $y[k] \sim \mathcal{N}(\mu, \sigma^2)$ w.r.t.: (i) $\theta = \mu$ and σ^2 is known, (ii) $\theta = \sigma^2$ and (iii) $\theta = \begin{bmatrix} \mu & \sigma^2 \end{bmatrix}^T$.

Solution: For all the three cases,

$$f(\mathbf{Y}_N|(\mu, \sigma^2)) = \prod_{k=0}^{N-1} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y[k]-\mu)^2}{2\sigma^2}\right)$$

Example 2: . . . contd.

Information in N observations

1. Constructing the log-likelihood from $f(\theta; \mathbf{y}_N)$ gives

$$S(\theta; \mathbf{y}_N) = \frac{\sum_{k=0}^{N-1} (y[k] - \mu)}{\sigma^2}$$

Applying (15), $I(\mu) = -E\left(\frac{\partial S}{\partial \theta}\right) = \frac{N}{\sigma^2}$

Example 2 . . . contd.

2. For this case, $S(\theta; \mathbf{y}_N) = -\frac{N}{2\sigma^2} + \frac{1}{2\sigma^2}\sum_{k=0}^{N-1}(y[k] - \mu)^2$

   Applying (15), $I(\sigma^2) = -\frac{\partial S}{\partial \theta} = \frac{N}{2\sigma^4}$

3. Denote $\theta_1 = \mu$ and $\theta_2 = \sigma^2$, $\boldsymbol{\theta} = \begin{bmatrix}\theta_1 & \theta_2\end{bmatrix}^T$ The log-likelihood function is

$$L(\boldsymbol{\theta}; \mathbf{y}_N) = c - \frac{N}{2}\ln\theta_2 - \frac{1}{2\theta_2}\sum_{k=0}^{N-1}(y[k] - \theta_1)^2 \tag{17}$$

Example 2 . . . contd.

The information matrix is thus

$$\mathbf{I}(\boldsymbol{\theta}) = -E\left(\begin{bmatrix} \frac{\partial^2 L}{\partial \theta_1^2} & \frac{\partial^2 L}{\partial \theta_1 \partial \theta_2} \\ \frac{\partial^2 L}{\partial \theta_2 \partial \theta_1} & \frac{\partial^2 L}{\partial \theta_2^2} \end{bmatrix}\right) = \begin{bmatrix} \frac{N}{\sigma^2} & 0 \\ 0 & \frac{N}{2\sigma^4} \end{bmatrix} \tag{18}$$

Thus, the estimates of mean and variance of a WN process do not affect each other, *i.e.*, these parameters can be estimated individually.

Remarks

- The Fisher information is a localized version (in the parameter space) of the more general **Kullback-Leibler information** (KLI) in the vicinity of the true parameters. The KLI measures the information loss incurred in approximating a true probability distribution with a model distribution.
- *Information* is leveraged on two factors: (i) the number and type of unknown(s) that have to be estimated and (ii) how these unknown(s) enter the *model*. Implications of these results are felt in model estimation and in input design.
- From the examples, we learn that by increasing the sample size, the increase in information is proportional. However, this is not the case when the observations are correlated. In fact, for that case $I_N(\theta) < NI_1(\theta)$.