Lecture 22

Wednesday, April 13, 2022 8:31 AM

Admin
1. AWOT due Friday.
2. Class cancelled 04/18
3. Projects due 05/04.

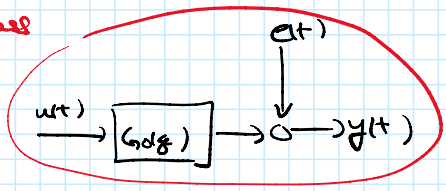
Parameter Covariance and Akaike's Information Criterion

Variance on parameters

Last time

$$\text{var}\left(\hat{G}(e^{i\omega})\right) = \frac{n}{N}\frac{\Phi_{ee}(\omega)}{\Phi_{uu}(\omega)}$$

n: # parameters estimated; N: # data points; evaluating goodness of model.



- True for a linear model + Gaussian noise.

→ Central limit theorem: Average tends to Gaussian for general distribution.

Asymptotic Covariance of the parameter estimate

- True system model: $y(t) = \varphi^T(t)\theta_0 + e(t)$ ← Gaussian noise. (θ_0: parameters of $G(q)$)

$$E[e(t)e(t+\tau)] = \begin{cases} \lambda, & \tau = 0 \\ 0, & \tau \neq 0 \end{cases}$$

Then $\hat{\theta}_N$ minimizes $V_N(\theta, Z^N) = \frac{1}{N}\sum_{t=1}^{N} \varepsilon^2(t,\theta)$, where $\varepsilon(t,\theta) = \hat{y}(t,\theta) - y(t)$ and Z^N: input–output data.

- Can show:

$$\underbrace{\text{cov}(\hat{\theta}_N)}_{\text{Asymptotic variance}} = \frac{\lambda}{N}\left[\underbrace{E\left[\varphi(t)\varphi^T(t)\right]}_{\text{Auto-covariance}}\right]^{-1}$$

$$\frac{1}{N}\sum_{t=1}^{N}\varphi(t)\varphi^T(t)$$

for linear models.

- The asymptotic variance is
  a) Inversely proportional to # of samples
  b) proportion to variance in noise
  c) Inversely related to parameter sensitivity.
     → For a linear model, inversely related to input auto-covariance

- For a general non-linear model,

$$\text{cov}(\hat{\theta}_N) \sim \frac{\lambda}{N}\left[E\left[\psi(t,\theta_0)\psi^T(t,\theta_0)\right]\right]^{-1} \leftarrow$$

$$\text{cov}(\hat{\theta}_N) \sim \frac{\lambda}{N}\left[E\left[\psi(t,\theta_0)\psi^T(t,\theta_0)\right]\right]^{-1} \leftarrow$$

$$\psi(t,\theta_0) = \frac{\partial}{\partial\theta}\left[\varepsilon(t,\theta)\right]\Big|_{\theta=\theta_0} = \frac{\partial}{\partial\theta}\left[\hat{y}(t,\theta) - y(t)\right]\Big|_{\theta=\theta_0}$$

$$= \frac{\partial}{\partial\theta}\left[\hat{y}(t,\theta)\right]\Big|_{\theta=\theta_0}$$

- For a linear model,

$$\hat{y}(t,\theta) = \varphi^T(t)\theta$$

$$\Rightarrow \psi(t,\theta_0) = \frac{\partial\hat{y}}{\partial\theta} = \varphi(t)$$

- For nonlinear model, θ_0 isn't known. → asymptotic variance can't be determined.
- In practice, use empirical estimates of $\hat{\theta}_N$, $\hat{\lambda}_N$ for large N.

Maximum Likelihood Estimation (MLE)

- Metods of parameter estimation.
  - Least squares.
    - → Find a set of parameters to minimize output error.
    - → Requires minimal info. about distribution of data.
  - Artificial Neural Nets.
    - → useful for nonlinear modeling.
    - → little understanding of how each parameter relates to output.
  - Maximum Likelihood Estimates.
    - → Relies on probability density funcns (PDFs); distribution info.
    - → we can estimate parameters. ⇒ Akaike's Information Criterion (AIC).

Principle of Maximum likelihood

unknown process $\Longrightarrow$ $y^N = (y_1, y_2, \ldots y_N)$ observed data.

- Assume a PDF y_i is generated by a stochastic process having

$$f(\mu,\lambda,x) = \frac{1}{\sqrt{2\pi\lambda}} e^{-(x-\mu)^2/2\lambda}$$

⇐ x is random variable associated w/ y_i.

μ = mean

associate w/ y_i.

μ = mean
λ = variance $= \sigma^2$

$$\mu = \frac{1}{N}\sum_{i=1}^{N} y_i \quad ; \quad \lambda = \frac{1}{N}\sum_{i=1}^{N}(y_i - \mu)^2 \Leftarrow \text{WANT THESE USING MLE.}$$

- Look at distribution of data from the MLE perspective.

- Assume N observations $y^N = y_1, \dots y_N$ are independent.
  → write joint probability

$$f(\mu, \lambda \mid x_1, x_2, \dots, x_N) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\lambda}} e^{-(x_i-\mu)^2/2\lambda} \leftarrow$$

- Given generated data, y^N want to find μ, λ that most likely that data.

$$\max f(\mu, \lambda \mid \underbrace{y_1, y_2, \dots y_N}_{\text{known}})$$

$$\hat{\theta} = \begin{bmatrix} \mu \\ \lambda \end{bmatrix} = \underset{\theta}{\operatorname{argmax}}\, f(\mu, \lambda \mid y^N)$$

$$= \underset{\theta}{\operatorname{argmax}} \log f(\mu, \lambda \mid y^N)$$

$$= \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{N} \left(\log \frac{1}{\sqrt{2\pi\lambda}} - \frac{(y_i-\mu)^2}{2\lambda} \right)$$

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

$$\frac{\partial \log f}{\partial \lambda} = -\frac{N}{2}\frac{\partial \log \lambda}{\partial \lambda} - \frac{1}{2}\lambda^{-1}\sum_{i=1}^{N}\frac{(y_i-\mu)^2}{2} \rightarrow \lambda = \frac{1}{N}\sum_{i=1}^{N}(y_i-\mu)^2$$

- MLE gives a probability distribution approach to thinking about optimal parameter estimation.

Information Theory in Parameter Estimation

- maximum likelihood estimate (MLE):

$$\hat{\theta}^{ML}(z^N) = \underset{\theta}{\operatorname{argmax}} \log L(\theta)$$

- likelihood $L(\theta) = \prod_{i=1}^{N} f(\theta; y_1, \dots y_N)$

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

$$\hat{\theta}^{ML}(z^N) = \underset{\theta}{\operatorname{argmax}} \frac{1}{N} \underbrace{\sum_{t=1}^{N} \log f_\varepsilon(\varepsilon, t; \theta)}$$

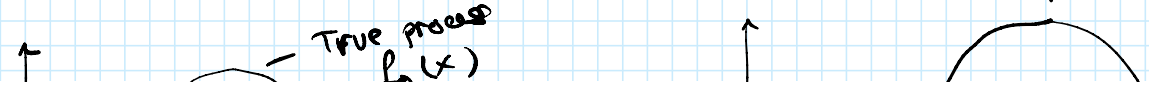
↓ ← related entropy

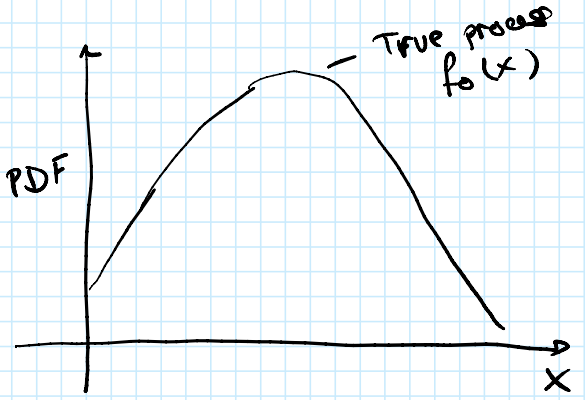
Information Theoretic Approach.

- Model - data agreement is quantified in terms of the amount of information captured by the model.
- A "good" model fully exploits all of the information in the data.
  - Least squares estimate
    - prediction error ε, input data are uncorrelated with each other!
  - Maximum likelihood estimate.
    - The log joint probability is maximized
    - ↓ ± The degree of randomness in prediction error is maximized.
    - ↓
    - The entropy is maximized.
  - We need a unified measure to enable us to compare model structures on the same basis & provide an objective to select an optimal model.
    - ⇒ based on a trade-off between bias in estimate and variance in output error.
      - → Accuracy vs. reliability ← Ljung.

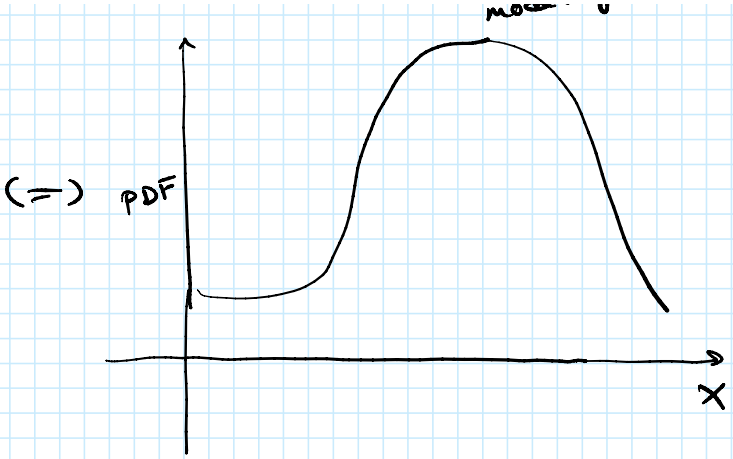
Kullback-Leibler Information distance

- Measure the distance between true system & model!





$(=)$



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

→ To evaluate the distance between true process & model, consider

$$I(f_0, g) = \int_x \underbrace{f_0(x)}_{} \underbrace{\log \frac{f_0(x)}{g(x|\theta)}}_{\equiv \log f_0(x) - \log g(x|\theta)} dx$$

$\equiv \log f_0(x) - \log g(x|\theta)$ ← distance between funcs.

Goal is to minize.