System Identification

6.435

SET 7

– Parameter Estimation Methods

– Minimum Prediction Error Paradigm

– Maximum Likelihood

Munther A. Dahleh

Parameters Estimation Methods

Paradigm: Pick the parameters that minimize a scalar function of the prediction error. "Min Prediction Error Paradigm".

Set-Up: $m^* = R(m) = \{m(\theta) | \theta \in D_m\}$, where $\boldsymbol{m}$ is a model structure.

$$Y = T(q, \theta) \begin{bmatrix} u \\ e \end{bmatrix}$$

$$\hat{Y} = W(q, \theta) Z \qquad Z = \begin{bmatrix} u \\ y \end{bmatrix}$$

$\varepsilon = y - \hat{y} \triangleq$ Prediction error

Data $$Z^N = \left(u(1) \;\; u(2) \;\; y(1) \;\; y(2) \;\; \ldots \right) = \left[\, u \;\; y \, \right]$$

$$\widehat{\theta} : Z^N \longrightarrow D_m$$

$$V_N(\cdot,\cdot) : \Re^d \times \Re^N \longrightarrow \Re \qquad \left(V_N\left(\theta, Z^N\right) \right)$$

$$V_N\left(\theta, Z^N\right) = \frac{1}{N} \sum_{t=1}^{N} l(\varepsilon(t,\theta))$$

<u>Problem</u>: Minimize the prediction error.

$$\widehat{\theta} = \underset{\widehat{\theta} \in D_m}{\text{argmin}} \; V_N\left(\theta, Z^N\right) = \underset{\widehat{\theta} \in D_m}{\text{argmin}} \; \frac{1}{N} \sum_{t=1}^{N} l(\varepsilon(t,\theta))$$

Points to consider:

1) The choice of $l(\cdot)$ is arbitrary. Typical choices

Typical choices $||\cdot||_p \quad (p = 2, \infty, 1)$.

2) $\varepsilon(t,\theta)$ may or may not correspond to a linear regression:

ARX $\quad \varepsilon(t,\theta) = y(t) - \Phi^T(t)\theta \quad \Phi$ is indep of θ

ARMAX $\quad \varepsilon(t,\theta) = y(t) - \Phi^T(t,\theta)\theta$ pseudo linear

3) Practical considerations

– Choice of l ; Robustness

– Filtering the data (equivalently $\varepsilon(t,\theta)$).

Quadratic Criterion

$$V_N\left(\theta, Z^N\right) = \frac{1}{N}\sum_{t=1}^{N} \varepsilon^2(t,\theta)$$

Recall: $\varepsilon(t,\theta) = y - \hat{y}$

$$= H^{-1}(q,\theta)(y - G(q,\theta)u)$$

Let $E_N\left(\frac{2\pi}{N}k, \theta\right) \quad k = 0, \ldots, N-1$ to be the $\boldsymbol{N}$-point DFT of $\varepsilon(t,\theta)$:

$$E_N\left(\frac{2\pi}{N}k, \theta\right) = \frac{1}{\sqrt{N}}\sum_{t=1}^{N} \varepsilon(t,\theta)e^{-i\frac{2\pi}{N}tk}$$

Then:

$$V_N\left(\theta, Z^N\right) = \frac{1}{2N}\sum_{k=0}^{N-1}\left|E_N\left(\frac{2\pi}{N}k, \theta\right)\right|^2$$

However

$$E_N(\omega,\theta) = H^{-1}\left(e^{iw},\theta\right)\left(Y_N(\omega) - G\left(e^{iw},\theta\right)U_N(\omega)\right) + R_N(\omega)$$

$$|R_N(\omega)| \leq \frac{C}{\sqrt{N}}$$

Then

$$V_N\left(\theta, Z^N\right) = \frac{1}{2N}\sum_{k=0}^{N-1}\left|H^{-1}\left(e^{i\frac{2\pi}{N}k},\theta\right)\right|^{-2}\left|Y_N\left(\frac{2\pi}{N}k\right) - G\left(e^{i\frac{2\pi}{N}k},\theta\right)U_N\left(\frac{2\pi}{N}k\right)\right|^2$$

$\Rightarrow$

$$V_N\left(\theta, Z^N\right) = \frac{1}{2N}\sum_{k=0}^{N-1}\frac{\left|U_N\left(\frac{2\pi}{N}k\right)\right|^2}{\left|H\left(e^{i\frac{2\pi}{N}k},\theta\right)\right|^2}\left|\hat{\hat{G}}\left(e^{i\frac{2\pi}{N}k}\right) - G\left(e^{i\frac{2\pi}{N}k},\theta\right)\right|^2$$

$\downarrow$ as $N \to \infty$

$$V_N\left(\theta, Z^N\right) \cong \frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{1}{2}\left|\hat{\hat{G}}\left(e^{iw},\theta\right) - G\left(e^{iw},\theta\right)\right|^2 Q_N(\omega,\theta)dw$$

where $\quad Q_N(\omega,\theta) = \dfrac{|U_N(\omega)|^2}{|H(e^{i\omega},\theta)|^2}$

Strong relationship to spectral estimation: What is the best approximation of $\hat{\hat{G}}\left(e^{iw},\theta\right)$ that lies in $\boldsymbol{m}$.

We will do more with this formula.

MIMO Systems

Define $Q_N\left(\theta, Z^N\right) = \frac{1}{N}\sum_{t=1}^{N} \varepsilon(t,\theta)\varepsilon^T(t,\theta)$

$$V_N\left(\theta, Z^N\right) = h\left(Q_N\left(\theta, Z^N\right)\right)$$

Examples of h: $\text{tr}\ (\cdot)$

$\text{tr}\ (\cdot\Lambda^{-1})$ for some weighting matrix Λ .

Maximum Likelihood

Assumption:

$$m(\theta) : \hat{y}(t|\theta) \triangleq g\left(t, Z^{(t-1)}, \theta\right)$$

$$\varepsilon(t,\theta) = y(t) - \hat{y}(t|\theta)$$

$\varepsilon(t,\theta)$ are independent and have the PDF $f_e(x,t,\theta)$!

Comment: $\varepsilon(t,\theta)$ will have the distribution of the noise when $\theta = \theta_o =$ actual value. The above is an approximation.

$$y(t) = \hat{y}(t|\theta) + \varepsilon(t,\theta) \qquad Z^N = \left(U^N, Y^N\right)$$

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

$$\log f_y\left(\theta, Y^N\right) = \sum_{t=1}^{N} \log f_e\left(y(t) - \hat{y}(t|\theta)\right)$$

Define

$$l(\varepsilon, t, \theta) = -\log f_e(\varepsilon, t, \theta)$$

Then

$$\hat{\theta}_{ML} = \underset{\theta}{\text{argmax}}\ f_y\left(\theta, Y^N\right) = \underset{\theta}{\text{argmax}}\ \frac{1}{N} \log f_y\left(\theta, Y^N\right)$$

$$= \underset{\theta}{\text{argmin}}\ \frac{1}{N} \sum_{t=1}^{N} l(\varepsilon(t), \theta, t) \equiv \text{ min prediction error}$$

Example: f_e is Gaussian

$$-\log f_e(\varepsilon, t, \theta) = \text{const} + \frac{1}{2}\log\lambda + \frac{1}{2}\frac{\varepsilon^2}{\lambda}$$

λ known ⟶ Quadratic objective

λ unknown ⟶ Parameterized norm criterion; $l(\varepsilon, t, \theta)$!

Information matrix:

$$\frac{d}{d\theta}\log f_y = \sum_{t=1}^{N} \frac{d}{d\theta}\log f_e(\varepsilon, \theta)$$

$$= \sum_{t=1}^{N} \frac{d}{d\theta} l(\varepsilon, \theta, t)$$

Suppose l depends on ε only. Then.

$$\frac{d}{d\theta}\log f_y = \sum_{t=1}^{N} l'(\varepsilon)\frac{d\varepsilon}{d\theta} = -\sum_{t=1}^{N} l'(\varepsilon)\Psi(t,\theta)$$

where $\Psi(t,\theta) = \frac{d}{d\theta}\widehat{y}(t,\theta) = -\frac{d}{d\theta}\varepsilon(t,\theta)$

Evaluate the information matrix

$$M_N = E\left(\frac{d}{d\theta}L\left(\frac{d}{d\theta}L\right)^T\right)\Bigg|_{\varepsilon=e_o}$$

$$= E\sum_{s,t=1}^{N} l'(e_o)\Psi(t,\theta_o)l'(e_o(s))\Psi^T(s,\theta_o)$$

$$= \sum_{t=1}^{N} E\left(l'(e_o(t))\right)^2 \cdot E\Psi(t,\theta_o)\Psi^T(t,\theta_o)$$

$e_o(t), e_o(s)$ are indep.

$$l_o'(x) = (\log f_e(x))' = \frac{f_e'(x)}{f_e(x)}$$

$$E\left(l'(e_o(t))\right)^2 = \int \frac{[f_e'(x)]^2}{[f_e(x)]^2} f_e(x)dx$$

$$= \int \frac{[f_e'(x)]^2}{f_e(x)} dx = \frac{1}{\kappa}$$

If e_o is Gaussian with variance $\lambda_o \Rightarrow \kappa = \lambda_o$, it follows that

$$M_N = \frac{1}{\kappa_o} \sum_{t=1}^{N} E\left(\Psi(t,\theta_o)\Psi^T(t,\theta_o)\right)$$

The Cramer-Rao Bound: for any unbiased estimator $\widehat{\theta}$

$$\text{Cov}\left(\widehat{\theta}\right) \geq \kappa_o \sum_{t=1}^{N} E\left(\Psi(t,\theta_o)\Psi^T(t,\theta_o)\right)^{-1}$$

Exact Likelihood Function

Example: $y(t) + ay(t-1) = e(t)$

$e(t)$ is WN with PDF f_e

Need to find P_Y ?

$$P(Y_1, \ldots, Y_N) = P(Y_2, \ldots, Y_N | Y_1) P(Y_1)$$

$$= P(Y_3, \ldots, Y_N | Y_2, Y_1) P(Y_2 | Y_1) P(Y_1)$$

$$= \prod_{k=2}^{N} P(Y(k) | Y(k-1), \ldots, Y(1)) P(Y(1))$$

$$= \prod_{k=2}^{N} P(e(k)) P(Y(1))$$

$$L\left(Y^N,\theta\right) = \frac{1}{N}\sum_{k=2}^{N} \log f_e(y(t)+ay(t-1)) + \frac{1}{N}\log P(Y(1))$$

$\searrow$ 0 as $N \to \infty$

Approximate Likelihood

$$\hat{Y} = -ay(t-1)$$

$$f_y\left(\theta, Y^N\right) = \prod_{t=1}^{N} f_e(y(t)+ay(t-1))$$

$$L\left(\theta, Y^N\right) = \frac{1}{N}\sum \log f_e(y(t)+ay(t-1))$$