System Identification

6.435

SET 5

- Least Squares
- Statistical Properties

Munther A. Dahleh

Least Squares

- Linear regressions
- LS Estimates: Statistical properties
- Bias, variance, covariance
- Noise-variance estimation
- Introduction to model structure determination:

  Statistical analysis and hypothesis testing

- $y(t) = \Phi^T(t)\theta$ ← unknown parameters

  $y(t)$: measured; $\Phi^T(t)$: known $(1 \times n)$

- Multivariable $\quad y : p \times 1, \quad \Phi^T : p \times n, \quad \theta : n \times 1$

- Examples:

  – $$y(t) = a_o + a_1 t + \ldots + a_r t^r$$

  $$\Phi^T(t) = \begin{pmatrix} 1 & t & \ldots & t^r \end{pmatrix} \qquad \theta = \begin{pmatrix} a_o \\ \vdots \\ a_r \end{pmatrix}$$

  – $$y(t) = h_o(t) + h_1 u(t-1) + \ldots + h_{\mu-1} u(t-\mu+1)$$

  $$\Phi^T(t) = \begin{pmatrix} u(t) & u(t-1) & \ldots & u(t-\mu+1) \end{pmatrix}$$

  $$\theta = \begin{pmatrix} h_o & h_1 & \ldots & h_{\mu-1} \end{pmatrix}$$

- In a matrix form

$$Y = \begin{pmatrix} y(1) \\ \vdots \\ y(N) \end{pmatrix} \quad \Phi = \begin{pmatrix} \Phi^T(1) \\ \vdots \\ \Phi^T(N) \end{pmatrix} \qquad Y = \Phi\theta$$

- In general, this system is perturbed by noise

$$Y = \Phi\theta + e$$

e – noise, stochastic

- For N > n, the system is overdetermined.

- Define $\varepsilon = \begin{pmatrix} \varepsilon(1) \\ \vdots \\ \varepsilon(N) \end{pmatrix}$, then $\varepsilon = Y - \Phi\theta$

- LS: $\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \frac{1}{2}||\varepsilon||_2^2 = \underset{\theta}{\operatorname{argmin}} V(\theta)$

Solution to LS

- $\widehat{\theta} = \left(\Phi^T \Phi\right)^{-1} \left(\Phi^T Y\right)$

- $V\left(\widehat{\theta}\right) = \frac{1}{2}\left[y^T y - y^T \Phi \left(\Phi^T \Phi\right)^{-1} \Phi^T y\right]$

- Proof: Standard pseudo-inverse formula

  $V\left(\widehat{\theta}\right)$: by substitution

- Equivalently

$$\hat{\theta} = \left(\sum_{t=1}^{N} \Phi(t)\Phi^T(t) \right)^{-1} \left(\sum_{t=1}^{N} \Phi(t)y(t) \right)$$

- Example

$$y(t) = b \quad \Rightarrow \quad \Phi^T(t) = 1 \quad , \quad \theta = b$$

$$\hat{\theta} = \frac{1}{N} \sum_{t=1}^{N} y(t)$$

Analysis of LS

- $Y = \Phi\theta + e \qquad e = \begin{pmatrix} e(1) \\ \vdots \\ e(N) \end{pmatrix}$

- Assume that $E\left(ee^T\right) = \lambda^2 I \quad , \quad E(e) = 0$

  $\therefore$ white noise, zero mean, variance λ^2.

- Theorem:

  1) $\widehat{\theta}$: is unbiased estimate of θ

  2) Cov $\left(\widehat{\theta}\right) = E\left(\widehat{\theta}-\theta\right)\left(\widehat{\theta}-\theta\right)^T = \lambda^2\left(\Phi^T\Phi\right)^{-1}$

  3) an unbiased estimate of λ^2 is given by

$$s^2 = \frac{2V\left(\widehat{\theta}\right)}{N-n}$$

- Basic assumption: Φ is fixed.

- Proof:

  1) $\widehat{\theta} = \left(\Phi^T\Phi\right)^{-1}\left(\Phi^T\Phi\theta + \Phi^T e\right)$

$$= \theta + \left(\Phi^T\Phi\right)^{-1}\Phi^T e.$$

$$E\widehat{\theta} = \theta + \left(\Phi^T\Phi\right)^{-1}\Phi^T Ee = \theta$$

2) $$E\left(\widehat{\theta}-\theta\right)\left(\widehat{\theta}-\theta\right)^T = E\left(\Phi^T\Phi\right)^{-1}\Phi^T ee^T \Phi\left(\Phi^T\Phi\right)^{-1}$$
$$= \left(\Phi^T\Phi\right)^{-1}\Phi^T E\left(ee^T\right)\Phi\left(\Phi^T\Phi\right)^{-1}$$
$$= \lambda^2\left(\Phi^T\Phi\right)^{-1}$$

3) $$Es^2 = \frac{2V\left(\widehat{\theta}\right)}{N-n} = \frac{1}{N-n}E\left(Y^T\left(I-\Phi\left(\Phi^T\Phi\right)^{-1}\Phi^T\right)Y\right)$$
$$= \frac{1}{N-n}E\left(e^T\left(I-\Phi\left(\Phi^T\Phi\right)^{-1}\Phi^T\right)e\right)$$
$$= \frac{1}{N-n}E\ \mathrm{tr}\left(e^T\left(I_N-\Phi\left(\Phi^T\Phi\right)^{-1}\Phi^T\right)e\right)$$
$$= \frac{1}{N-n}E\ \mathrm{tr}\left(I_N-\Phi\left(\Phi^T\Phi\right)^{-1}\Phi^T\right)\lambda^2$$
$$= \frac{1}{N-n}E\ \mathrm{tr}\left(I_N-\underbrace{\left(\Phi^T\Phi\right)^{-1}\Phi^T\Phi}_{I_N}\right)\lambda^2$$
$$= \frac{1}{N-n}(N-n)\lambda^2 = \lambda^2$$

Best Linear Unbiased Estimate

- $Y = \Phi\theta + e$

  $E\left(ee^T\right) = R$ correlated noise

- Analysis of LS estimate:

  1) $E\left(\widehat{\theta}\right) = \theta$

  2) $\text{Cov}\left(\widehat{\theta}\right) = \left(\Phi^T\Phi\right)^{-1}\Phi^T R\Phi\left(\Phi^T\Phi\right)^{-1}$

- Consider general linear estimators:

  $\widehat{\theta} = Z^T Y$

- Want to find Z such that the estimate is unbiased and $\text{Cov}\left(\widehat{\theta}\right)$ is minimized.

Solution of BLUE

- Solution: $Z^* = R^{-1}\Phi\left(\Phi^T R^{-1}\Phi\right)^{-1}$
- $\text{Cov}_{Z^*}\left(\hat{\theta}\right) = \left(\Phi^T R^{-1}\Phi\right)^{-1} \leq \text{Cov}_Z\left(\hat{\theta}\right)$ for any unbiased estimate.
- Proof:

$$\hat{\theta} = Z^T Y = Z^T \Phi \theta_o + Z^T e$$

$$E\left(\hat{\theta}\right) = \theta_o \quad \Rightarrow \quad Z^T \Phi = I$$

$$\text{Cov}_Z\left(\hat{\theta}\right) = E\left(Z^T Y - \theta_o\right)\left(Z^T Y - \theta_o\right)^T = Z^T R Z$$

$$\begin{aligned}\text{Cov}_{Z^*}\left(\hat{\theta}\right) &= \left(\Phi^T R^{-1}\Phi\right)^{-1}\Phi^T R^{-1} R R^{-1}\Phi\left(\Phi^T R^{-1}\Phi\right)^{-1} \\ &= \left(\Phi^T R^{-1}\Phi\right)^{-1}\Phi^T R^{-1}\Phi\left(\Phi^T R^{-1}\Phi\right)^{-1} \\ &= \left(\Phi^T R^{-1}\Phi\right)^{-1}\end{aligned}$$

- To show that $\text{Cov}_{Z^*}\left(\hat{\theta}\right) \leq \text{Cov}_Z\left(\hat{\theta}\right)$

$$
\begin{aligned}
\text{Cov}_Z\left(\hat{\theta}\right) - \text{Cov}_{Z^*}\left(\hat{\theta}\right) &= Z^T R Z - \left(\Phi^T R^{-1} \Phi\right)^{-1} \\
&= Z^T R Z - Z^T \Phi \left(\Phi^T R^{-1} \Phi\right)^{-1} \Phi^T Z \\
&= Z^T \left[R - \Phi \left(\Phi^T R^{-1} \Phi\right)^{-1} \Phi^T\right] Z
\end{aligned}
$$

However,

$$
\begin{aligned}
&R - \Phi \left(\Phi^T R^{-1} \Phi\right)^{-1} \Phi^T \\
&\quad = \left(R - \Phi \left(\Phi^T R^{-1} \Phi\right)^{-1} \Phi^T\right) R^{-1} \left(R - \Phi \left(\Phi^T R^{-1} \Phi\right)^{-1} \Phi^T\right) \\
&\quad \geq 0
\end{aligned}
$$

result follows.

- If $R = \lambda^2 \quad \Rightarrow \quad Z^* = \Phi \left(\Phi^T \Phi\right)^{-1}$ and is equal to the least squares estimate. Hence LS is the BLUE when e is white.

- Can nonlinear estimates help? Not if the distribution of e is Gaussian

- BLUE of $A\theta_o$ for any constant matrix is $A\hat{\theta}$.

- Example:

$$y(t) = b_o + e(t) \qquad Ee^2(t) = \lambda t^2$$

$$\Phi = \begin{pmatrix} \vdots \end{pmatrix} \qquad R = \begin{pmatrix} \lambda_1^2 & & \\ & \ddots & \\ & & \lambda_N^2 \end{pmatrix}$$

BLUE of $b_o = \theta$ is

$$\hat{\theta} = \frac{1}{\sum_{j=1}^{N} \left(\frac{1}{\lambda_j^2}\right)} \sum_{i=1}^{N} \frac{1}{\lambda_i^2} y(i)$$

Maximum Likelihood Estimate

$y = \Phi\theta_o + e$ $\qquad \theta_o$ is unknown

$$P(y|\theta_o) = P(\Phi\theta_o + e|\theta_o)$$

Suppose: Maximum likelihood estimate

$$\hat{\theta}_{ML} = \arg \max_{\theta} P(y|\theta)$$

The parameter that makes y the "most likely event".

Result: Suppose that y is a random variable with a distribution that depends on θ_o . Let $L(y,\theta)$ be the likelihood function, and $\hat{\theta}(y)$ is an unbiased estimate of θ. Then:

$$\text{Cov}\left(\hat{\theta}\right) \geq \left[E\left(\frac{\partial \log L}{\partial \theta}\right)^T \left(\frac{\partial \log L}{\partial \theta}\right)\right]^{-1} = -\left[E\frac{\partial^2 \log L}{\partial \theta^2}\right]^{-1}$$

Least Squares

$$y = \Phi\theta_o + e \qquad e \sim N\left(0, \lambda^2 I\right)$$

$$L\left(Y, \theta, \lambda^2\right) = \frac{1}{(2\pi)^{\frac{N}{2}}\left(\det \lambda^2 I_N\right)^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(Y - \Phi\theta)^T \left(\lambda^2 I_N\right)^{-1} (Y - \Phi\theta)\right)$$

$$\log L = -\frac{1}{2\lambda^2}(Y - \Phi\theta)^T (Y - \Phi\theta) - \frac{N}{2}\log 2\pi - \frac{N}{2}\log \lambda^2$$

Differentiate:

$$\frac{\partial \log L}{\partial \theta} = +\frac{1}{\lambda^2}(y - \Phi\theta)^T \Phi$$

$$\frac{\partial \log L}{\partial \lambda^2} = \frac{1}{2\lambda^4}(Y - \Phi\theta)^T (Y - \Phi\theta) - \frac{N}{2\lambda^2}$$

$$\frac{\partial^2 \log L}{\partial \theta^2} = -\frac{1}{\lambda^2}\Phi^T\Phi$$

$$\frac{\partial}{\partial \theta}\left(\frac{\partial \log L}{\partial \lambda^2}\right) = -\frac{1}{2\lambda^4}(Y - \Phi\theta)^T\Phi \quad \Rightarrow \quad E = 0$$

$$\frac{\partial^2 \log L}{\partial \left(\lambda^2\right)^2} = -\frac{1}{\lambda^6}(Y - \Phi\theta)^T(Y - \Phi\theta) + \frac{N}{2\lambda^4}$$

$$E\left(\frac{\partial^2 \log L}{\partial \left(\lambda^2\right)^2}\right) = -\frac{1}{\lambda^6}N\lambda^2 + \frac{N}{2\lambda^4}$$

$$= -\frac{N}{\lambda^4} + \frac{N}{2\lambda^4} = -\frac{1}{2}\frac{N}{\lambda^4}$$

$$\text{Cov}\left(\widehat{\theta}\right) \geq J^{-1} \qquad \widehat{\widehat{\theta}} = \begin{pmatrix} \widehat{\theta} \\ \lambda^2 \end{pmatrix}$$

$$J = E \begin{pmatrix} \frac{1}{\lambda^2}\Phi^T\Phi & 0 \\ 0 & \frac{N}{2\lambda^4} \end{pmatrix}$$

$$\text{Cov}\left(\widehat{\theta}\right) \geq \left(\frac{1}{\lambda^2}\Phi^T\Phi\right)^{-1} = \lambda^2\left(\Phi^T\Phi\right)^{-1}$$

$$\text{Cov}\left(\lambda^2\right) \geq \frac{2\lambda^4}{N}$$

- Remarks:
  - LS estimate for $\widehat{\theta}$ is efficient, i.e. $\text{Cov}\left(\widehat{\theta}\right) = J^{-1}$ = lower bound (termed efficient)
  - LS estimate of λ^2 (guess) is asymptotically efficient.

To show that:

$$\text{Var}\left(s^2\right) = E\left(s^2 - \lambda^2\right)^2 = E\left(s^2\right)^2 - \lambda^4 \quad \text{(assuming Gaussian dist)}$$

$$\text{and } s^2 = \frac{e^T P e}{N-n} \qquad P = \left(1 - \Phi\left(\Phi^T \Phi\right)^{-1} \Phi^T\right)$$

$$P^2 = P$$

$$\text{Var}\left(s^2\right) = \frac{2\lambda^4}{N-n} > \frac{2\lambda^4}{N} = \text{Cramer-Rao bdd.}$$

- Result: BLUE for Gaussian distribution is best estimate over Nonlinear estimators.