

Statistical Signal Processing

Lecture 5a

chapter 1: parameter estimation: deterministic parameters

simplified estimators: BLUE, method of moments, (W)LS:

- problem formulation and solution
- linear model
- applications of the linear model
- interpretations of the LS solution
- performance analysis: bias, MSE, consistency
- acoustic echo cancellation demo, part 1
- model order reduction
- acoustic echo cancellation demo, part 2

Least-Squares (LS) Problem Formulation

- Consider n' data (signal) samples S that depend on m parameters θ

$$S = \begin{bmatrix} s_1 \\ \vdots \\ s_{n'} \end{bmatrix}, \quad \theta = \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_m \end{bmatrix}, \quad Y = \begin{bmatrix} y_1 \\ \vdots \\ y_{n'} \end{bmatrix}, \quad V = \begin{bmatrix} v_1 \\ \vdots \\ v_{n'} \end{bmatrix}, \quad E = \begin{bmatrix} e_1 \\ \vdots \\ e_n \end{bmatrix}$$

- **nonlinear model:** model functions $g_k(\theta, S) = 0$, $k = 1, \dots, n$

- **example: sinusoid:** $s_k = A \cos(\omega k + \phi)$, $\theta = \omega$

can show: $s_k - 2 \cos \omega s_{k-1} + s_{k-2} = g_k(\theta, S) = 0 \quad \Rightarrow \quad n' = n+2$

indeed, characteristic equation associated with the difference equation:

$$z^2 - 2 \cos \omega z + 1 = 0 \Rightarrow z = e^{\pm j\omega} \Rightarrow s_k = \frac{Ae^{j\phi}}{2} e^{j\omega k} + \frac{Ae^{-j\phi}}{2} e^{-j\omega k} = A \cos(\omega k + \phi)$$

- **observed data:** $y_k = s_k + v_k$, $v_k =$ measurement/observation noise
- if $v_k \neq 0$ (noisy observations) and/or g_k (model description) approximate, then $g_k(\theta, Y) = e_k(\theta) \neq 0$, $e_k =$ equation error
- **LS method:** introduced by Gauss in 18th century for the estimation of the parameters of elliptical orbits of planets from noisy observations.

LS Estimation

- **LS strategy:** adjust $\hat{\theta}$ to minimize the sum of squared errors $E^T E = \sum_{k=1}^n e_k^2$

- Let $G(\theta, Y) = [g_1(\theta, Y) \cdots g_n(\theta, Y)]^T$, then

$$\hat{\theta}_{LS} = \arg \min_{\hat{\theta}} G^T(\hat{\theta}, Y) G(\hat{\theta}, Y) = \arg \min_{\hat{\theta}} \sum_{k=1}^n g_k^2(\hat{\theta}, Y) = \hat{\theta}_{LS}(Y)$$

estimator $\hat{\theta}(Y)$ = function of the observations Y

- remark: LS can be formulated without any statistical context!

- **model linear in the parameters:**

$$g_k(\theta, Y) = f_k(Y) - C_k(Y) \theta, \quad f_k(Y) : 1 \times 1, \quad C_k(Y) : 1 \times m, \quad \theta : m \times 1$$

- example cont'd: let $\theta = 2 \cos \omega \Rightarrow \begin{cases} f_k(Y) = y_k + y_{k-2} \\ C_k(Y) = y_{k-1} \end{cases}$

- Let $F(Y) = \begin{bmatrix} f_1(Y) \\ \vdots \\ f_n(Y) \end{bmatrix} : n \times 1, \quad H(Y) = \begin{bmatrix} C_1(Y) \\ \vdots \\ C_n(Y) \end{bmatrix} : n \times m$

- LS: $\hat{\theta}_{LS} = \arg \min_{\theta} [F(Y) - H(Y) \theta]^T [F(Y) - H(Y) \theta] = \hat{\theta}_{LS}(Y)$

LS: Discussion

- $F(Y) - H(Y) \theta = \begin{bmatrix} f_1 \\ \vdots \\ f_n \end{bmatrix} - \begin{bmatrix} C_1 \\ \vdots \\ C_n \end{bmatrix} \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_m \end{bmatrix} = \begin{bmatrix} f_1 \\ \vdots \\ f_n \end{bmatrix} - [H_1 \cdots H_m] \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_m \end{bmatrix} = \begin{bmatrix} e_1 \\ \vdots \\ e_n \end{bmatrix} = E$

n equations, m unknowns θ (if try to make $E = 0$)

- $n > m$: *overdetermined* case

exact fit impossible \Rightarrow least-squares fit

(assume: H = full rank = full column rank \Rightarrow unique solution)

- $n = m$: *exactly determined* case

if H = full rank $\Rightarrow H^{-1}$ exists $\Rightarrow \hat{\theta} = H^{-1} F$ = unique solution

(no averaging of errors though)

- $n < m$: *underdetermined* case

∞^{m-n} solutions exist, there is a unique solution of minimum norm $\|\hat{\theta}\|$

- assume henceforth: $n > m$, $\text{rank}(H) = m$

then parameters *identifiable*: θ can be found exactly if optimal $E(\theta) = 0$

LS: Solution

- LS: $\hat{\theta}(Y) = \arg \min_{\theta} \xi_{LS}(\theta, Y)$

$$\begin{aligned} \xi_{LS}(\theta, Y) &= \|F(Y) - H(Y) \theta\|_2^2 \\ &= [F(Y) - H(Y) \theta]^T [F(Y) - H(Y) \theta] \\ &= [F^T(Y) - \theta^T H^T(Y)] [F(Y) - H(Y) \theta] \end{aligned}$$

- $\frac{\partial \xi_{LS}}{\partial \theta} = -2H^T(Y) [F(Y) - H(Y) \theta] = 0 \Rightarrow H^T(Y) H(Y) \theta = H^T(Y) F(Y)$

$$\Rightarrow \hat{\theta}_{LS} = (H^T(Y) H(Y))^{-1} H^T(Y) F(Y) = \hat{\theta}_{LS}(Y)$$

- Hessian = $\frac{\partial}{\partial \theta} \left(\frac{\partial \xi_{LS}}{\partial \theta} \right)^T = 2H^T(Y) H(Y) > 0$ since $H(Y)$ full column rank (constant w.r.t. θ)

\Rightarrow extremum = minimum, only one \Rightarrow global one

LS: Linear Model

- $\left. \begin{array}{l} F(Y) = Y \\ H(Y) = H \end{array} \right\} \rightarrow \left\{ \begin{array}{l} y_k = C_k \theta + v_k, \quad k = 1, \dots, n \quad v_k = \text{error} \\ Y = H \theta + V \\ \quad = \sum_{i=1}^m H_i \theta_i + V \end{array} \right\} \begin{array}{l} H \theta = S = \text{signal component} \\ V = \text{noise} \end{array}$

- $\hat{\theta}_{LS} = (H^T H)^{-1} H^T Y$

- example 1: amplitude and phase estimation of a noisy sinusoid (ω known)

$$\begin{aligned} y_k &= A \cos(\omega k + \phi) + v_k \\ &= A \cos \phi \cos(\omega k) - A \sin \phi \sin(\omega k) + v_k \\ &= \underbrace{[\cos(\omega k) \quad \sin(\omega k)]}_{C_k} \underbrace{\begin{bmatrix} A \cos \phi \\ -A \sin \phi \end{bmatrix}}_{\theta} + v_k \end{aligned}$$

- example 2: line fitting

$$y_k = a x_k + b + v_k = \underbrace{[x_k \quad 1]}_{C_k} \underbrace{\begin{bmatrix} a \\ b \end{bmatrix}}_{\theta} + v_k$$



Weighted Least-Squares (WLS)

non-linear model

- WLS: $\min_{\theta} E^T W E$, $E = [e_1 \cdots e_n]^T$
 $\hat{\theta}_{WLS} = \arg \min_{\theta} G^T(\theta, Y) W G(\theta, Y)$, $W = W^T > 0$ weighting matrix
- LS: $W = I$, $\Rightarrow E^T E = \sum_{k=1}^n e_k^2$

model linear in parameters

- WLS: $\min_{\theta} \xi_{WLS}(\theta, Y) = \min_{\theta} [F(Y) - H(Y) \theta]^T W [F(Y) - H(Y) \theta]$
- $\frac{\partial \xi_{WLS}}{\partial \theta} = -2H^T(Y) W [F(Y) - H(Y) \theta] = 0$
 $\Rightarrow \hat{\theta}_{WLS} = (H^T(Y) W H(Y))^{-1} H^T(Y) W F(Y) = \hat{\theta}_{WLS}(Y)$
- Hessian = $\frac{\partial}{\partial \theta} \left(\frac{\partial \xi_{WLS}}{\partial \theta} \right)^T = 2H^T(Y) W H(Y) > 0$
 since $W > 0$ and $H(Y)$ full column rank
 \Rightarrow extremum = minimum, only one \Rightarrow global one



3 Quantities of Potential Interest

model linear in parameters: $F(Y) = H(Y) \theta + E$

linear model: $Y = H \theta + V$ $(F(Y), H(Y), E) = (Y, H, V)$

- 3 quantities:
- parameters: θ
 - signal: $S = H \theta$
 - error/noise: $E = F(Y) - H(Y) \theta$ or $V = Y - H \theta$

LS estimates:

- parameters: $\hat{\theta} = (H^T H)^{-1} H^T F$
- signal: $\hat{S} = H \hat{\theta} = P_H F$, $P_H = H(H^T H)^{-1} H^T$
 projection of F/Y on the *signal subspace* = column space of H
- error/noise: $\hat{E} = F - \hat{S} = F - H \hat{\theta} = P_H^\perp F$, $P_H^\perp = I - P_H$
 projection of F/Y on the *noise subspace* = orthogonal complement of column space of H
 P = projection matrix if $P = P^T$ (symmetric) and $P P = P$ (idempotent)
 eigenvectors/values of P_H (P_H^\perp): $P_H H = H$, $P_H^\perp H = 0$
 basis vectors of signal subspace, corresponding to eigenvalue 1 (0),
 basis vectors of noise subspace, corresponding to eigenvalue 0 (1).



Applications

- linear model
 1. polynomial curve fitting / modal analysis
 2. filter design
- model linear in parameters
 3. optimal/adaptive filtering



Application 1: Polynomial Curve fitting/Modal Analysis

- measurements $y_k = \text{signal} + \text{noise}$
 signal is a linear combination of known basis functions $h_i(k)$ (*modes*)

$$y_k = s_k + v_k = \sum_{i=1}^m \theta_i h_i(k) + v_k = c_k^T \theta + v_k$$

where $c_k^T = [h_1(k) \cdots h_m(k)]$. The linear combination coefficients θ_i are the parameters.

- typical signal model: solution of a homogenous difference equation with constant coefficients;

$$s_k = \sum_{i=1}^{m_0} \left(\sum_{j=1}^{m_i} \alpha_{ij} k^{j-1} \right) \lambda_i^k = c_k^T \theta$$

$$c_k^T = [k^0 \lambda_1^k \cdots k^{m_1-1} \lambda_1^k \quad k^0 \lambda_2^k \cdots k^{m_{m_0}-1} \lambda_{m_0}^k]$$

$$\theta^T = [\alpha_{11} \cdots \alpha_{1 m_1} \quad \alpha_{2 1} \cdots \alpha_{m_0 m_{m_0}}]$$

for m_0 distinct roots λ_i with multiplicity m_i .

Applic. 1: Polynomial Curve fitting/Modal Analysis (2)

- The signal s_k is the solution of the following difference equation

$$\prod_{i=1}^{m_0} (1 - \lambda_i q^{-1})^{m_i} s_k = 0$$

where q^{-1} is the delay operator: $q^{-1}s_k = s_{k-1}$ (q^{-1} transforms to a multiplication by z^{-1} when taking the z -transform). The total order of the difference equation is $m = \sum_{i=1}^{m_0} m_i$.

- particular case 1: $m_0 = 1$ root and $\lambda_1 = 1$: s_k is a polynomial function of k .
In particular, if $m_1 = 1$, then $(1 - q^{-1})s_k = s_k - s_{k-1} = 0$ and $s_k \equiv b$ is a constant.
If $m_1 = 2$, then $s_k - 2s_{k-1} + s_{k-2} = 0$ and $s_k = ak + b$ (example 1 above).
- particular case 2: m_0 even, λ_i on the unit circle ($\lambda_i = e^{j\omega_i}$) and occurring in complex conjugate pairs, and $m_i = 1, \forall i$. Useful reparameterization:

$$s_k = \sum_{i=1}^{m_0/2} (\alpha_i e^{j\omega_i k} + \alpha_i^* e^{-j\omega_i k}) = \sum_{i=1}^{m_0/2} (a_i \cos(\omega_i k) + b_i \sin(\omega_i k)) .$$

(see example 2 above: $m_0 = 2$)

Application 2: Filter Design

IIR filter design in the time domain

- IIR model transfer function: $\frac{B(z)}{A(z)} = \frac{b_0 + b_1 z^{-1} + \dots + b_p z^{-p}}{1 + a_1 z^{-1} + \dots + a_r z^{-r}}$

parameters $\theta = [a_1 \dots a_r \ b_0 \ b_1 \dots b_p]^T$, $m = p + q + 1$

- IIR model impulse response: $s_k = \frac{B(q)}{A(q)} \delta_{k0}$

Kronecker delta: $\delta_{ij} = \begin{cases} 1 & , \ i = j \\ 0 & , \ i \neq j \end{cases}$

- target impulse response (causal, truncated): $y_k = s_k + v_k$, $k = 0, 1, \dots, n$

error $v_k = y_k - \frac{B(q)}{A(q)} \delta_{k0}$ nonlinear in parameters θ

- consider $A(q) y_k = B(q) \delta_{k0} + \underbrace{A(q) v_k}_{e_k}$ or $e_k = y_k + \sum_{i=1}^r a_i y_{k-i} - b_k$

error e_k linear in the parameters

($b_k = 0, k > p$)

Application 2: Filter Design (2)

- with $Y = [y_0 \ y_1 \ \cdots \ y_n]^T$, $E = [e_0 \ e_1 \ \cdots \ e_n]^T$, $B = [b_0 \ b_1 \ \cdots \ b_n]^T$, we can write

$$E = \mathcal{A}Y - B = Y - H\theta, \quad H = [-\mathcal{Y} \ \mathcal{I}]$$

where

$$\mathcal{A}(\theta) = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ a_1 & \ddots & \ddots & \\ \vdots & \ddots & \ddots & \ddots \\ a_r & & \ddots & \ddots \\ \vdots & \ddots & & \ddots & 0 \\ 0 & \cdots & a_r & \cdots & a_1 & 1 \end{bmatrix}, \quad \mathcal{Y} = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ y_0 & 0 & \cdots & 0 \\ y_1 & y_0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ y_{n-1} & y_{n-2} & \cdots & y_{n-r} \end{bmatrix}, \quad \mathcal{I} = \begin{bmatrix} I_{p+1} \\ 0 \end{bmatrix}$$

\mathcal{A} and \mathcal{Y} are Toeplitz (elements along a diagonal are the same), hence they are specified by their first row and column; they are also lower triangular, and \mathcal{A} is banded (limited number of non-zero diagonals).

For filtering with \mathcal{A} : Toeplitzness corresponds to time-invariance, triangularity to causality and bandedness to FIR.

- Strictly speaking: model linear in parameters: $F = Y$ but $H(Y)$ depends on Y .
- LS solution: $\hat{\theta}_{LS} = (H^T H)^{-1} H^T Y = \arg \min_{\theta} E^T E$

Application 2: Filter Design (3)

- Assume now that we insist on obtaining the LS solution in the *output error* V rather than the *equation error* $E = \mathcal{A}V$ (corresponding to $e_k = \mathcal{A}(q) v_k$).
- Observe that we have

$$V = \mathcal{A}^{-1} E = \mathcal{A}^{-1} (Y - H\theta)$$

- We can obtain the LS solution $\arg \min_{\theta} V^T V$ iteratively as follows. Note

$$V^T V = \|\mathcal{A}^{-1} (Y - H\theta)\|_2^2 = (Y - H\theta)^T (\mathcal{A} \mathcal{A}^T)^{-1} (Y - H\theta)$$

Hence the solution $\hat{\theta}^{(i)}$ at iteration i can be obtained as

$$\hat{\theta}_{WLS}^{(i)} = \left(H^T W^{(i)} H \right)^{-1} H^T W^{(i)} Y \text{ where } W^{(i)} = \left(\mathcal{A}(\hat{\theta}^{(i-1)}) \mathcal{A}^T(\hat{\theta}^{(i-1)}) \right)^{-1}$$

- Initialization: e.g. $\hat{\theta}_{WLS}^{(0)} = 0$ so that $\hat{\theta}_{WLS}^{(1)} = \hat{\theta}_{LS}$ ($\mathcal{A}(0) = I \Rightarrow W^{(1)} = I$).
- Note: $V^T V = E^T W E$: the LS problem in the output error V corresponds to a WLS problem in the equation error E .
- known as Steiglitz-McBride iterations

Application 2: Filter Design (4)

FIR filter design in the frequency domain

- FIR filter $B(z) = C(z) B$, $C(z) = [1 \ z^{-1} \dots z^{-p}]$, $\theta = B = [b_0 \ b_1 \dots b_p]^T$
- We wish to fit the frequency response $B(e^{j2\pi f})$ to a desired response y_i at frequency f_i , $i = 1, \dots, n'$:

$$y_i = C(e^{j2\pi f_i}) \theta + v_i, \quad i = 1, \dots, n'$$

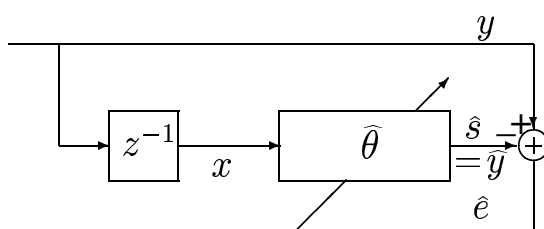
where v_i here is clearly not noise but approximation error.

- Then $\hat{\theta}_{LS} = (H^T H)^{-1} H^T Y$ where

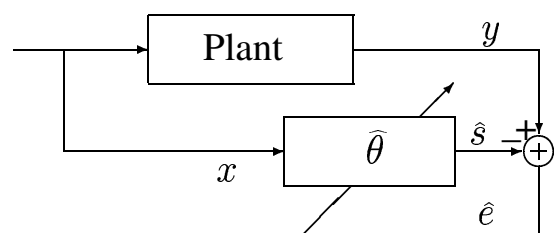
$$Y' = \begin{bmatrix} y_1 \\ \vdots \\ y_{n'} \end{bmatrix}, \quad H' = \begin{bmatrix} C(e^{j2\pi f_1}) \\ \vdots \\ C(e^{j2\pi f_{n'}}) \end{bmatrix}, \quad Y = \begin{bmatrix} \Re Y' \\ \Im Y' \end{bmatrix}, \quad H = \begin{bmatrix} \Re H' \\ \Im H' \end{bmatrix}$$

- For the design of a filter with real coefficients $\theta = B$, the distribution of the frequency points f_i can be limited to the normalized frequency interval $[0, \frac{1}{2}]$.
- A weighting matrix $W = \text{blockdiag}\{W', W'\}$, $W' = \text{diag}\{w_1, \dots, w_{n'}\}$ can be introduced to put a higher weight $w_i > 0$ at frequencies f_i where a tighter fit is desired ($V^T W V = V'^H W' V' = \sum_{i=1}^{n'} w_i |v_i|^2$ where $V^H = (V^*)^T$).

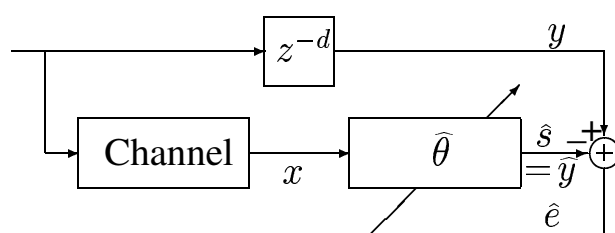
Application 3: Adaptive Filtering



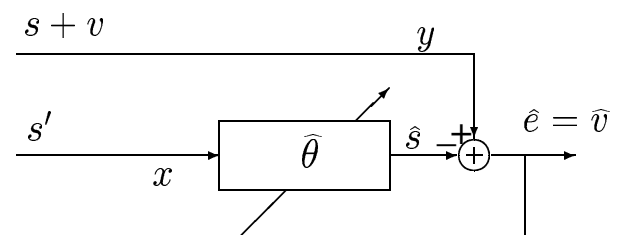
prediction, spectral estimation, whitening



system identification



equalization, deconvolution



interference canceling

Application 3: Adaptive Filtering (2)

- adaptive filtering terminology: y_k = desired-response signal, x_k = filter input
- strictly speaking: adaptive filtering = application of model linear in parameters because H contains signal
- adaptive filtering cases:

I. single-channel FIR filtering (4 cases): previous figure with $\theta = B$ ($m = p+1$) = FIR filter impulse response: $y_{1:n} = [y_1 \cdots y_n]^T = H\theta + V$ with

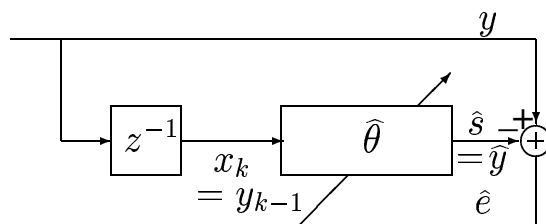
$$H = H(x_{2-m:n}) = \begin{bmatrix} x_1 & x_0 & \cdots & x_{2-m} \\ x_2 & x_1 & \cdots & x_{3-m} \\ \vdots & \vdots & \ddots & \vdots \\ x_n & x_{n-1} & \cdots & x_{n-m+1} \end{bmatrix}, \quad \theta = B = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_p \end{bmatrix} = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_{m-1} \end{bmatrix}, \quad Y = \begin{bmatrix} y_{1:n} \\ x_{2-m:n} \end{bmatrix}$$

H is Toeplitz. $E = V$ in this case.

II. multichannel applications: (combinations of:)

- * IIR filters formulated as multichannel FIR filters
- * multirate FIR filters
- * vector input signals: spatial filtering (beamforming)/spatiotemporal filtering of multiple sensor (antennas/sensors) signals
- * other multidimensional signals (images)

Application 3: Adaptive Filtering (3)



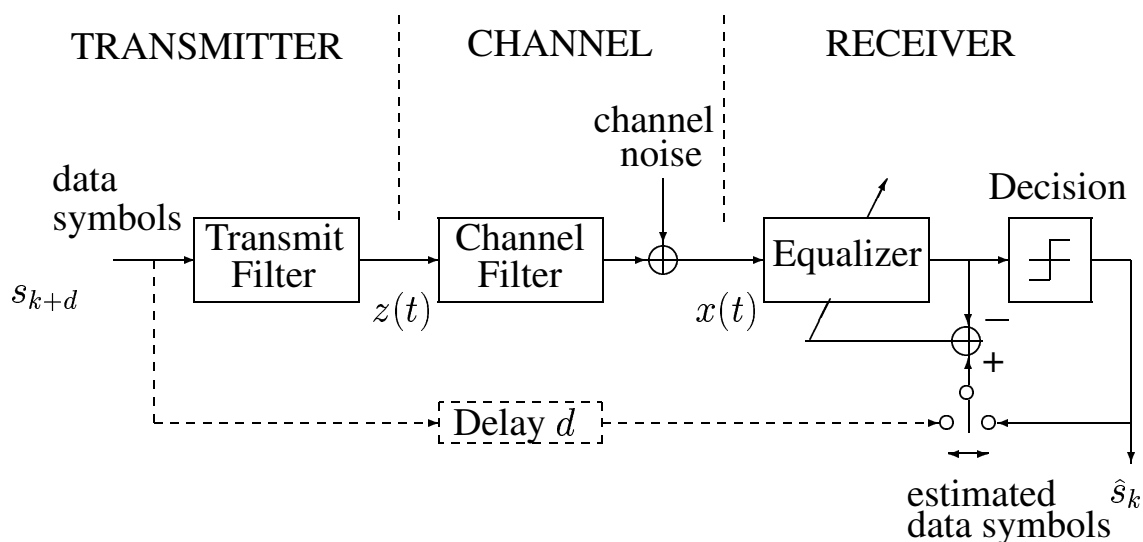
prediction, spectral estimation, whitening

- here $x_k = y_{k-1} \Rightarrow x_k$ noisy also
- **prediction** = s_k , e.g. stock market (multidimensional signals though)
- **whitening**: make prediction error e_k as white as possible (unpredictable part): used in signal coding (e_k easier to quantize than y_k)
- **spectral estimation/modeling**: when prediction error e_k becomes white (uncorrelated), θ contains all the spectral (correlation) information of y_k

Application 3: Adaptive Filtering (4)

• equalization, deconvolution:

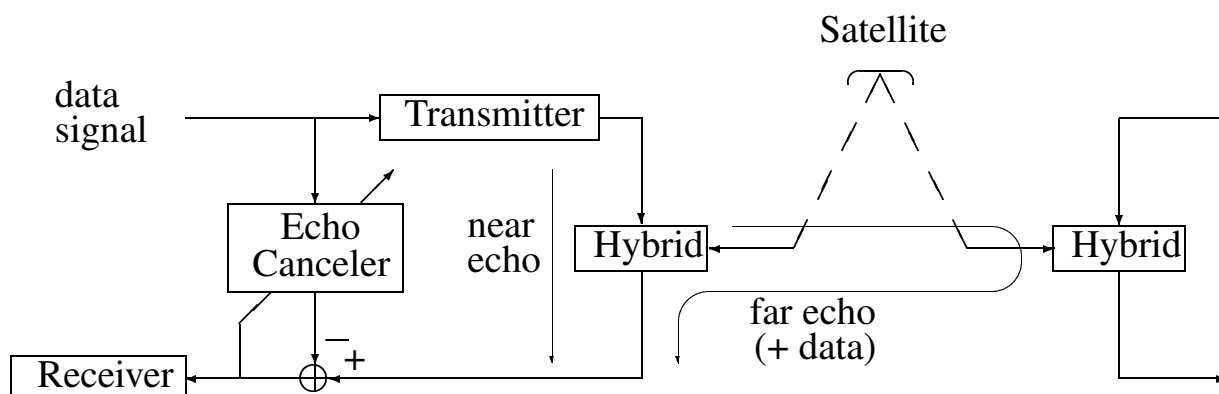
- s_k of interest here (transmitted symbols, original image/object)
- the noise is here situated at the filter input x_k instead of at the filter output y_k
- recovery of original image from a blurred version
- reconstruction of 3D object from 2D images
- channel equalization in communications:



Application 3: Adaptive Filtering (5)

• interference cancellation: $e_k = v_k$ signal of interest, corrupted by unmeasurable noise s_k , which is correlated with the measurable noise $s'_k = x_k$ applications:

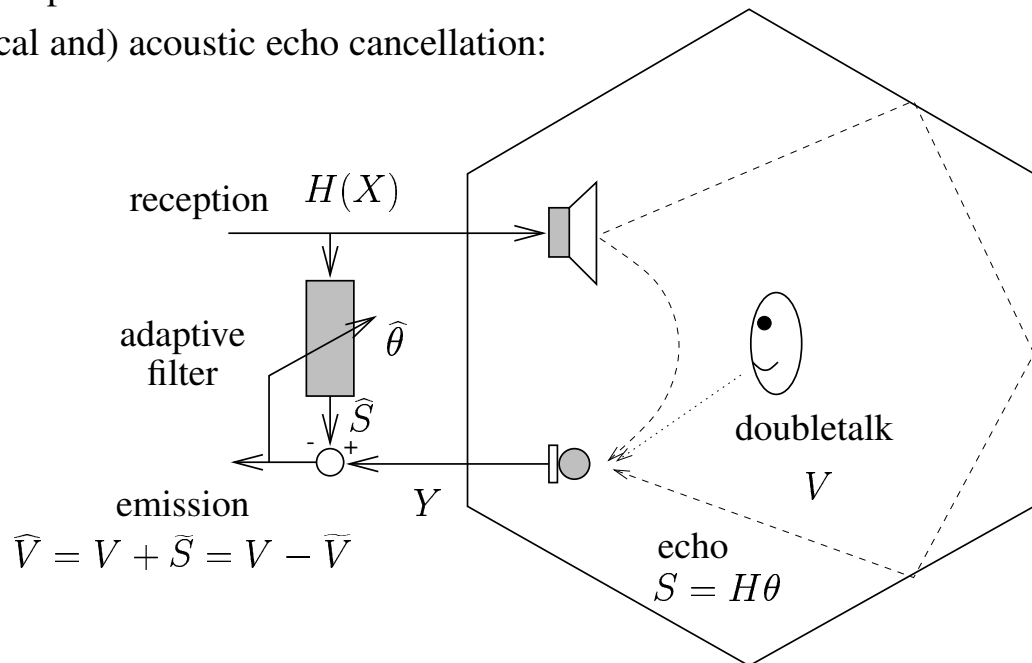
- acoustic (motor) noise reduction for handsfree telephony systems in cars
- fan/air conditioning system noise reduction in teleconferencing systems
- 50 Hz interference in electrocardiography
- interference from other users in mobile communications
- electrical echo cancellation in telephone lines (voiceband modems/xDSL):



Application 3: Adaptive Filtering (6)

- **system identification:** θ (filter) of interest, examples:

- channel identification
- automatic control
- seismic exploration
- (electrical and) acoustic echo cancellation:



Orthogonality Principle of LS

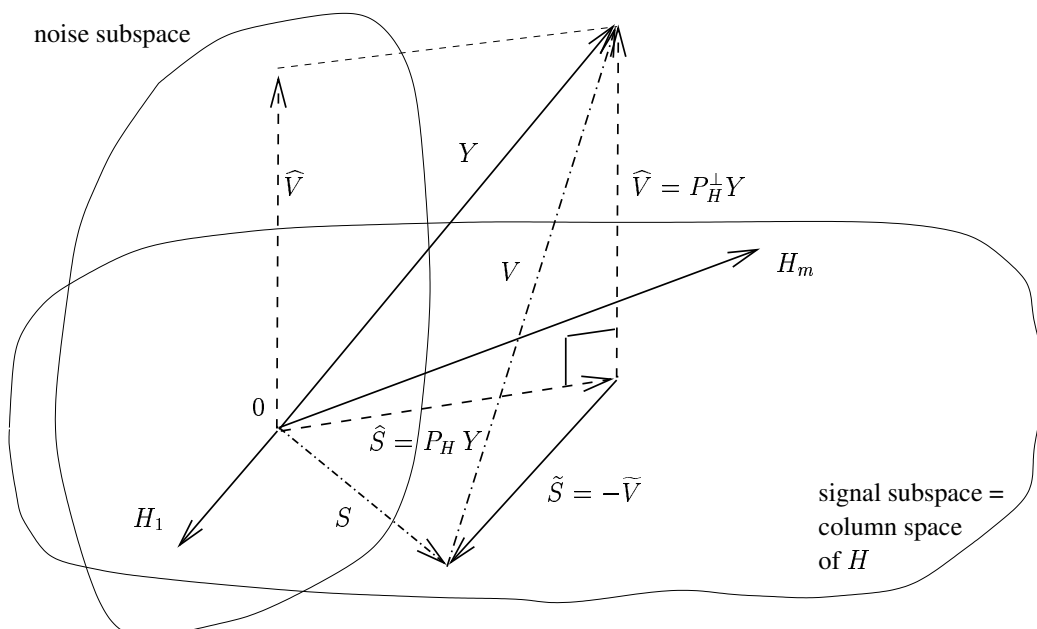
- we found that $\hat{\theta}_{LS}$ satisfies

orthogonality conditions of LS

$$H^T (Y - H\hat{\theta}_{LS}) = H^T \hat{V}_{LS} = 0 \Leftrightarrow H_i^T \hat{V}_{LS} = 0, \quad i = 1, \dots, m$$

the smallest fitting error is orthogonal to the signal subspace (column space of H)

linear model notation assumed here





Correlation and Covariance Matrices

- random vectors X and Y
- mean: $m_X = E X$, $m_Y = E Y$ (E = Expectation)
- correlation matrix: $R_{XY} = E XY^T$, $R_{XX} = E XX^T$
- covariance matrix:

$$C_{XY} = R_{X-m_X, Y-m_Y} = E (X - m_X)(Y - m_Y)^T = R_{XY} - m_X m_Y^T$$
- vector power (mean square value):

$$\begin{aligned} E \|X\|^2 &= \text{tr} \{ E \|X\|^2 \} = E \text{tr} \{ \|X\|^2 \} = E \text{tr} \{ X^T X \} \\ &= E \text{tr} \{ X X^T \} = \text{tr} \{ E X X^T \} = \text{tr} \{ R_{XX} \} \end{aligned}$$
- notation: $\begin{cases} \theta = \hat{\theta} + \tilde{\theta} \\ S = \hat{S} + \tilde{S} \\ V = \hat{V} + \tilde{V} \end{cases}$



Performance Analysis of LS in the Linear Model

- *a priori* and *a posteriori* decompositions of Y :

$$Y = \underbrace{S + V}_{\text{a priori decomposition}} = \underbrace{\hat{S} + \hat{V}}_{\text{a posteriori decomposition}}$$

where $\hat{S} \perp \hat{V} : \hat{S}^T \hat{V} = \hat{\theta}^T H^T \hat{V} = 0$

- estimator **bias** : average deviation from the true parameter (E = Expectation)

$$b_{\hat{\theta}}(\theta) = -E\tilde{\theta} = E(\hat{\theta}(Y) - \theta) = E\hat{\theta}(Y) - \theta$$

unbiased estimator: $b_{\hat{\theta}}(\theta) = 0, \forall \theta \in \Theta$ (set of possible values for θ)

Unbiasedness is a weak property: estimator can be correct on the average, but with large deviations (large MSE). Also, good estimators exist that are biased.

- **MSE** = $\text{tr} \{ R_{\tilde{\theta}\tilde{\theta}} \} = E \|\tilde{\theta}\|_2^2$, $R_{\tilde{\theta}\tilde{\theta}} = E \tilde{\theta} \tilde{\theta}^T$ = estimation error correlation matrix

$$\begin{aligned} R_{\tilde{\theta}\tilde{\theta}} &= E(\hat{\theta} - \theta)(\hat{\theta} - \theta)^T = E[\hat{\theta}(-E\hat{\theta} + E\hat{\theta}) - \theta][\hat{\theta}(-E\hat{\theta} + E\hat{\theta}) - \theta]^T \\ &= E(\hat{\theta} - E\hat{\theta})(\hat{\theta} - E\hat{\theta})^T + (E\hat{\theta} - \theta)(E\hat{\theta} - \theta)^T = C_{\hat{\theta}\hat{\theta}} + b_{\hat{\theta}}(\theta)b_{\hat{\theta}}^T(\theta) = C_{\tilde{\theta}\tilde{\theta}} + b_{\hat{\theta}}(\theta)b_{\hat{\theta}}^T(\theta) \end{aligned}$$

$$\text{tr} \{ R_{\tilde{\theta}\tilde{\theta}} \} = \text{tr} \{ C_{\tilde{\theta}\tilde{\theta}} \} + \|b_{\hat{\theta}}\|^2: \text{Mean Squared Error} = \text{variance} + \text{bias squared}$$

- (mean square) **consistency**: if $\text{MSE}(n) \xrightarrow{n \rightarrow \infty} 0$, then $\hat{\theta} \xrightarrow{n \rightarrow \infty} \theta$ (in mean square)

Performance Analysis of LS in the Linear Model (2)

- No statistical information (about V) needed to derive $\hat{\theta}_{WLS}$. However, in order to evaluate its performance (for the linear model), we need to introduce a stochastic context: V random with
$$\begin{cases} E V = 0 \\ E V V^T = C_{VV} \end{cases}$$
- note: $\hat{\theta}_{WLS} - \theta = (H^T W H)^{-1} H^T W (H \theta + V) - \theta = (H^T W H)^{-1} H^T W V$
- $b_{WLS} = E \hat{\theta}_{WLS} - \theta = (H^T W H)^{-1} H^T W E V = 0$: unbiased if $E V = 0$
- $C_{\hat{\theta}\hat{\theta}}(W) = C_{\hat{\theta}\hat{\theta}}(W) = (H^T W H)^{-1} H^T W C_{VV} W H (H^T W H)^{-1}$
- optimal weighting: $W = C_{VV}^{-1}$: $C_{\hat{\theta}\hat{\theta}}(W) \geq C_{\hat{\theta}\hat{\theta}}(C_{VV}^{-1}) = (H^T C_{VV}^{-1} H)^{-1}$
- LS: $C_{\hat{\theta}\hat{\theta}} = C_{\hat{\theta}\hat{\theta}}(I) = (H^T H)^{-1} H^T C_{VV} H (H^T H)^{-1}$
- white noise: $C_{VV} = \sigma_v^2 I_n \Rightarrow WLS^{opt} = LS$ and $C_{\hat{\theta}\hat{\theta}} = \sigma_v^2 (H^T H)^{-1}$
- (W)LS in general consistent: $\hat{\theta} \rightarrow \theta$ as $\frac{n}{m} \rightarrow \infty$

Performance Analysis of LS in the Linear Model (3)

- consider LS and white noise ($C_{VV} = \sigma_v^2 I$)
- **signal component:**

$$\hat{S} = H \hat{\theta}_{LS} = P_H Y = S + P_H V \Rightarrow \tilde{S} = S - \hat{S} = -P_H V$$
 - * Hence, $E \hat{S} = S$: unbiased if $E V = 0$.
 - * $C_{\tilde{S}\tilde{S}} = P_H C_{VV} P_H = \sigma_v^2 P_H \Rightarrow E \|\tilde{S}\|^2 = \text{tr} \{C_{\tilde{S}\tilde{S}}\} = \sigma_v^2 \text{tr} \{P_H\} = m \sigma_v^2$ remains finite!
 - * Even $C_{\hat{s}_k \hat{s}_k} = \sigma_{\hat{s}_k}^2 = \sigma_v^2 [P_H]_{kk} (= \sigma_v^2 \frac{m}{n} \text{ on the avg.}) \xrightarrow{\frac{n}{m} \rightarrow \infty} 0$: \hat{s}_k consistent.

$$\frac{1}{n} \sum_{k=1}^n [P_H]_{kk} = \frac{1}{n} \text{tr} \{P_H\} = \frac{1}{n} \text{tr} \{H (H^T H)^{-1} H^T\} = \frac{1}{n} \text{tr} \{(H^T H)^{-1} H^T H\} = \frac{1}{n} \text{tr} \{I_m\} = \frac{m}{n}$$
- **noise component:**

$$\tilde{V} = Y - H \hat{\theta}_{LS} = P_H^\perp Y = P_H^\perp V \Rightarrow \tilde{V} = V - \hat{V} = P_H V$$
 - * Hence, $E \tilde{V} = 0$: unbiased if $E V = 0$ (case of a “random parameter”).
 - * $C_{\tilde{V}\tilde{V}} = C_{\tilde{S}\tilde{S}} \Rightarrow E \|\tilde{V}\|^2 = m \sigma_v^2$ remains finite also!
 - * Furthermore $C_{\tilde{v}_k \tilde{v}_k} = \sigma_{\tilde{v}_k}^2 = \sigma_{\hat{s}_k}^2 \xrightarrow{\frac{n}{m} \rightarrow \infty} 0$: \tilde{v}_k consistent also. ($\text{SNR} = \frac{\sigma_{\tilde{v}_k}^2}{\sigma_{\hat{s}_k}^2} = \frac{n}{m}$)
- observe: $R_{\hat{S}\tilde{V}} = E \hat{S} \tilde{V}^T = P_H C_{VV} P_H^\perp = \sigma_v^2 P_H P_H^\perp = 0$: a posteriori signal $\hat{S} = S + P_H V$ and noise components $\tilde{V} = P_H^\perp V$ are uncorrelated

Perf Analysis of LS in FIR System Identification

- recall: $H = [X_1 \ X_2 \ \cdots \ X_n]^T$, $X_i = [x_i \ x_{i-1} \ \cdots \ x_{i-m+1}]^T$
- linear model: H deterministic \rightarrow model linear in parameters: H can be stochast.
- law of large numbers: $\frac{1}{n} H^T H = \frac{1}{n} \sum_{i=1}^n X_i X_i^T \xrightarrow{n \rightarrow \infty} E X_i X_i^T = R_{XX} \ (m \times m)$
 \Rightarrow approximation: $H^T H \approx n R_{XX}$
- observe: if x_k and v_k are independent and at least one of them is white noise ($R_{XX} = \sigma_x^2 I$ and/or $R_{VV} = \sigma_v^2 I$), then $E H^T R_{VV} H = n \sigma_v^2 R_{XX}$
- hence $C_{\hat{\theta}\hat{\theta}} = (H^T H)^{-1} H^T C_{VV} H (H^T H)^{-1} \approx \frac{\sigma_v^2}{n} R_{XX}^{-1} \ (\Rightarrow \text{consistency})$
- Is LS criterion $= \|\hat{V}\| = Y^T P_H^\perp Y$ a good indicator of estimation quality? ($\hat{V} = Y - H\hat{\theta}$ = LS error)

$$\begin{aligned} E \|\hat{V}\|^2 &= E Y^T P_H^\perp Y = E V^T P_H^\perp V = E V^T V - E \{V^T P_H V\} \\ &= E \sum_{i=1}^n v_i^2 - \text{tr}\{E P_H V V^T\} = n \sigma_v^2 - \text{tr}\{E P_H C_{VV}\} \\ &= n \sigma_v^2 - \text{tr}\{E (H^T H)^{-1} H^T C_{VV} H\} \stackrel{\text{LLN}}{\approx} n \sigma_v^2 - \text{tr}\{(E H^T H)^{-1} E H^T C_{VV} H\} \\ &= n \sigma_v^2 - \text{tr}\{(n R_{XX})^{-1} n \sigma_v^2 R_{XX}\} = n \sigma_v^2 - \sigma_v^2 \text{tr}\{I_m\} = (n - m) \sigma_v^2 \end{aligned}$$

hence $E \|\hat{V}\|^2 \rightarrow 0$ as $m \nearrow n$ (or $n \searrow m$). Extreme case: $n = m \Rightarrow \hat{V} = 0$.
 But estimation not good at all.

Perf Analysis of LS in FIR System Identification (2)

- white noise case: $E \|\hat{V}\|^2 = E V^T P_H^\perp V = \text{tr}\{P_H^\perp E V V^T\}$
 $= \sigma_v^2 \text{tr}\{P_H^\perp\} = \sigma_v^2 \text{tr}\{I_n - P_H\} = \sigma_v^2 (n - m)$

- “signal” and “noise” parts: $\begin{cases} Y = S + V \\ \hat{S} = S - \bar{S} \\ \hat{V} = V - \bar{V} \end{cases}$

- A priori SNR: $\text{SNR}_Y = \frac{E \|S\|^2}{E \|V\|^2} = \frac{n E s_i^2}{n E v_i^2} = \frac{E (\theta^T X_i)^2}{\sigma_v^2} = \frac{\theta^T R_{XX} \theta}{\sigma_v^2}$

A posteriori SNRs:

$$\begin{aligned} \text{SNR}_{\hat{S}} &= \frac{E \|S\|^2}{E \|\bar{S}\|^2} = \frac{n E s_i^2}{m \sigma_v^2} = \frac{n}{m} \text{SNR}_Y \\ \text{SNR}_{\hat{V}} &= \frac{E \|V\|^2}{E \|\bar{V}\|^2} = \frac{n \sigma_v^2}{m \sigma_v^2} = \frac{n}{m} \quad \text{indep. of } \text{SNR}_Y ! \end{aligned}$$

- For $n = m$: $\text{SNR}_{\hat{S}} = \text{SNR}_Y$ (estimation did not improve SNR!),
 $\text{SNR}_{\hat{V}} = 1 = 0\text{dB}$ (LS error $\hat{V} = 0 \Rightarrow \bar{V} = V$)

Perf Analysis of LS in FIR System Identification (3)

- *cross validation*: to get an idea of estimation quality, try estimate $\hat{\theta}(Y)$ on n' other data $Y' = S' + V'$, $S' = H'\theta$ (independent from Y but identically distributed). In practice: often $n' = 1$ (1 new sample)

- **signal component**:

$$\begin{aligned}\hat{S}' &= H'\hat{\theta}_{LS} = H'(H^T H)^{-1} H^T Y = S' + H'(H^T H)^{-1} H^T V \\ \Rightarrow \bar{S}' &= S' - \hat{S}' = -H'(H^T H)^{-1} H^T V\end{aligned}$$

* can show $E \|\bar{S}'\|^2 \approx \frac{n'}{n} m \sigma_v^2$

* hence $\text{SNR}_{\hat{S}'} = \frac{E \|S'\|^2}{E \|\bar{S}'\|^2} = \frac{n}{m} \text{SNR}_Y$ as before

- **noise component**:

$$\bar{V}' = Y' - H'\hat{\theta}_{LS} = V' + \bar{S}' \Rightarrow \bar{S}' = -\bar{V}'$$

* $\text{SNR}_{\bar{V}'} = \frac{E \|V'\|^2}{E \|\bar{V}'\|^2} = \frac{n}{m}$ but $E \|\bar{V}'\|^2 = n' \sigma_v^2 (1 + \frac{m}{n}) > E \|V'\|^2$ now

* this time also $R_{\bar{V}'V'} = 0$ whereas $R_{\bar{V}V} = P_H R_{VV} \neq 0$ before

* to predict performance from \bar{V} : $\frac{1}{n'} \|\bar{V}'\|^2 \approx \frac{n+m}{n-m} \frac{1}{n} \|\bar{V}\|^2$ (Akaike's FPEC)

- conclusion: need $\frac{n}{m} = \frac{\# \text{ equations}}{\# \text{ unknowns}} \gg 1$ for good quality estimation

WLS: Performance Analysis

- No statistical information (about V) needed to derive $\hat{\theta}_{WLS}$. However, in order to evaluate its performance (for the linear model), we need to introduce a stochastic context: V random with $\begin{cases} E V = 0 \\ E V V^T = C_{VV} \end{cases}$

• note: $\hat{\theta}_{WLS} - \theta = (H^T W H)^{-1} H^T W (H \theta + V) - \theta = (H^T W H)^{-1} H^T W V$

• $E \hat{\theta}_{WLS} - \theta = (H^T W H)^{-1} H^T W E V = 0$: unbiased

• $C_{\hat{\theta}\hat{\theta}}(W) = C_{\hat{\theta}\hat{\theta}}(W) = (H^T W H)^{-1} H^T W C_{VV} W H (H^T W H)^{-1}$

• optimal weighting: $W = C_{VV}^{-1}$: $C_{\hat{\theta}\hat{\theta}}(W) \geq C_{\hat{\theta}\hat{\theta}}(C_{VV}^{-1}) = (H^T C_{VV}^{-1} H)^{-1}$

- Further statistical knowledge and optimality properties:

WLS = ML if $V \sim \mathcal{N}(0, W^{-1})$ and independent of θ

Rank Reduction in the Linear Model

- reparameterize in terms of a reduced set of parameters $\underbrace{\theta}_{m \times 1} = \underbrace{T}_{m \times r} \underbrace{\phi}_{r \times 1}$
- issue of optimal transformation T
- we shall limit analysis to $T = \begin{bmatrix} I_r \\ 0 \end{bmatrix} : \phi = \theta_{1:r} = \bar{\theta}_r$

$$S = H \theta = [\bar{H}_r \quad \underline{H}_r] \begin{bmatrix} \bar{\theta}_r \\ \underline{\theta}_r \end{bmatrix} = \bar{H}_r \bar{\theta}_r + \underline{H}_r \underline{\theta}_r$$

- reduced-rank LS: $\hat{\bar{\theta}}_r = \arg \min_{\bar{\theta}_r} \|Y - \bar{H}_r \bar{\theta}_r\|^2 = (\bar{H}_r^T \bar{H}_r)^{-1} \bar{H}_r^T Y$

$$\hat{S} = \hat{S}_r = \bar{H}_r \hat{\bar{\theta}}_r = P_{\bar{H}_r} Y, \quad \hat{V} = \hat{V}_r = Y - \hat{S}_r = P_{\bar{H}_r}^\perp Y$$

•

$$\begin{aligned} \hat{\bar{\theta}}_r &= (\bar{H}_r^T \bar{H}_r)^{-1} \bar{H}_r^T (\bar{H}_r \bar{\theta}_r + \underline{H}_r \underline{\theta}_r + V) \\ &= \bar{\theta}_r + (\bar{H}_r^T \bar{H}_r)^{-1} \bar{H}_r^T (\underline{H}_r \underline{\theta}_r + V) = \bar{\theta}_r - \bar{\bar{\theta}}_r \\ \hat{\theta} &= \begin{bmatrix} \hat{\bar{\theta}}_r \\ 0 \end{bmatrix}, \quad \bar{\theta} = \begin{bmatrix} \bar{\bar{\theta}}_r \\ \underline{\theta}_r \end{bmatrix} \end{aligned}$$

Rank Reduction in the Linear Model (2)

- estimator bias and variance

$$b_{\hat{\theta}}(\theta) = \begin{bmatrix} (\bar{H}_r^T \bar{H}_r)^{-1} \bar{H}_r^T \underline{H}_r \underline{\theta}_r \\ -\underline{\theta}_r \end{bmatrix}, \quad C_{\hat{\theta}\hat{\theta}} = \begin{bmatrix} C_{\bar{\bar{\theta}}\bar{\bar{\theta}}} & 0 \\ 0 & 0 \end{bmatrix}$$

$$C_{\bar{\bar{\theta}}\bar{\bar{\theta}}} = (\bar{H}_r^T \bar{H}_r)^{-1} \bar{H}_r^T C_{VV} \bar{H}_r (\bar{H}_r^T \bar{H}_r)^{-1}, \quad R_{\hat{\theta}\hat{\theta}} = C_{\hat{\theta}\hat{\theta}} + b_{\hat{\theta}} b_{\hat{\theta}}^T$$

- signal component

$$\tilde{S} = S - \hat{S}_r = \bar{H}_r \bar{\theta}_r + \underline{H}_r \underline{\theta}_r - (\bar{H}_r \bar{\theta}_r + P_{\bar{H}_r} \underline{H}_r \underline{\theta}_r + P_{\bar{H}_r} V) = P_{\bar{H}_r}^\perp \underline{H}_r \underline{\theta}_r - P_{\bar{H}_r} V$$

$$\text{bias : } b_{\hat{S}_r}(\theta) = -E \tilde{S} = -P_{\bar{H}_r}^\perp \underline{H}_r \underline{\theta}_r \neq 0 : \text{biased !}$$

$$R_{\tilde{S}\tilde{S}} = C_{\tilde{S}\tilde{S}} + b_{\hat{S}_r} b_{\hat{S}_r}^T = P_{\bar{H}_r} C_{VV} P_{\bar{H}_r} (P_{\bar{H}_r}^\perp \underline{H}_r \underline{\theta}_r) (P_{\bar{H}_r}^\perp \underline{H}_r \underline{\theta}_r)^T$$

- noise component

$$\tilde{V} = V - \hat{V}_r = V - (P_{\bar{H}_r}^\perp \underline{H}_r \underline{\theta}_r + P_{\bar{H}_r}^\perp V) = -P_{\bar{H}_r}^\perp \underline{H}_r \underline{\theta}_r + P_{\bar{H}_r} V = -\tilde{S}$$

$$\text{SNR}_{\hat{V}_r} = \frac{E \|V\|^2}{E \|\tilde{V}\|^2} = \frac{n\sigma_v^2}{\|P_{\bar{H}_r}^\perp \underline{H}_r \underline{\theta}_r\|^2 + r\sigma_v^2}$$

Rank Reduction in FIR System Identification

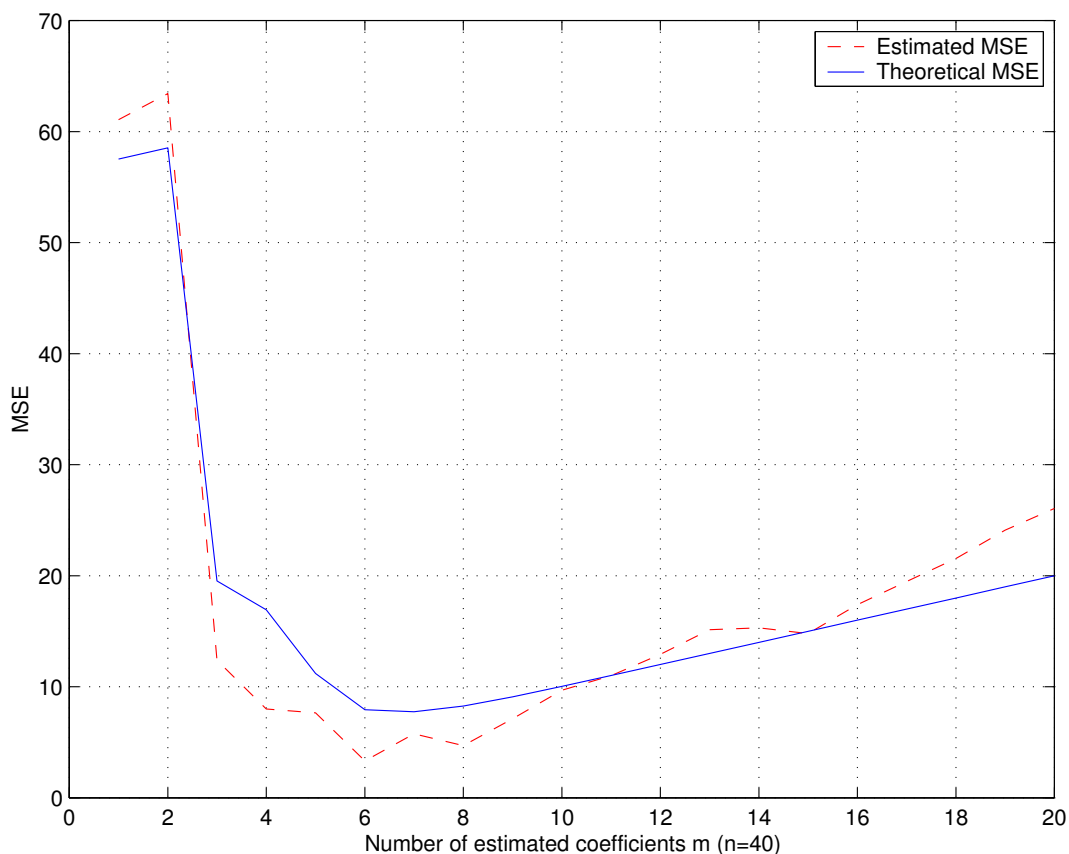
- Assume now H filled with samples of x_k , being white noise.
-

$$\begin{aligned} \text{SNR}_{\widehat{V}_r} &= \frac{n\sigma_v^2}{E_X \|P_{\widehat{H}_r}^\perp \underline{H}_r \underline{\theta}_r\|^2 + r\sigma_v^2} = \frac{n\sigma_v^2}{|\text{bias}|^2 + r\sigma_v^2} \\ &= \frac{1}{\frac{\sigma_x^2}{\sigma_v^2} \|\underline{\theta}_r\|^2 + \frac{r}{n}} = \frac{1}{\text{SNR}_Y \frac{\|\underline{\theta}_r\|^2}{\|\underline{\theta}\|^2} + \frac{r}{n}} \end{aligned}$$

- to maximize $\text{SNR}_{\widehat{V}_r}$, need to minimize $|\text{bias}|^2 + r\sigma_v^2$
- we have $E \|\widehat{V}_m\|^2 = (n-m)\sigma_v^2$, $E \|\widehat{V}_r\|^2 = |\text{bias}|^2 + (n-r)\sigma_v^2$
- Hence can estimate

$$|\text{bias}|^2 + r\sigma_v^2 \approx \|\widehat{V}_r\|^2 - \|\widehat{V}_m\|^2 + (2r-m)\sigma_v^2 \approx \|\widehat{V}_r\|^2 - \|\widehat{V}_m\|^2 + \frac{2r-m}{n-m} \|\widehat{V}_m\|^2$$

Rank Reduction in FIR System Identification (2)



Choice of Estimator

- stochastic (Bayesian) information matrix:

$$J_{stoch} = J_{prior} + E_{\theta} J_{det}(\theta)$$

as $J_{det} \sim n$, J_{det} dominant as $n \gg 1$.

Hence if lots of data \Rightarrow prior of little relevance \Rightarrow deterministic estimation

If little data \Rightarrow need prior (even if invented) to regularize the problem, to avoid singularity of J_{det}

- Bayesian estimation:

- $\hat{\theta}_{MMSE}$ preferable
- $\hat{\theta}_{MAP}$ easier to calculate
- $\hat{\theta}_{LMMSE}$ simple, acceptable if everything \approx Gaussian (model \approx linear)

- deterministic (classical) estimation:

- Maximum Likelihood (ML) if possible
- if ML too complex or if a good initialization is required for an iterative optimization of ML: Least-Squares or Method of Moments
- Linear Gaussian model: all reasonable estimators identical