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

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

- **nonlinear model**: model functions $g_k(\theta, S) = 0\,,\ k = 1, \ldots, n$
- example: sinusoid: $s_k = A\,\cos(\omega k + \phi)\,,\quad \theta = \omega$

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

  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 $v_k \not\equiv 0$ (noisy observations) and/or g_k (model description) approximate, then $g_k(\theta, Y) = e_k(\theta) \not\equiv 0 \quad , \quad e_k$ = 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 $\widehat{\theta}$ to minimize the sum of squared errors $E^TE = \sum_{k=1}^{n} e_k^2$
- Let $G(\theta, Y) = [g_1(\theta, Y) \cdots g_n(\theta, Y)]^T$, then

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

  estimator $\widehat{\theta}(Y)$ = function of the observations Y
- remark: LS can be formulated without any statistical context!
- **model linear in the parameters**:

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

- example cont'd: let $\theta = 2\cos\omega \quad \Rightarrow \begin{cases} f_k(Y) &= y_k + y_{k-2} \\ C_k(Y) &= y_{k-1} \end{cases}$
- 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: $\widehat{\theta}_{LS} = \arg\min_{\theta} [F(Y) - H(Y)\,\theta]^T\ [F(Y) - H(Y)\,\theta] = \widehat{\theta}_{LS}(Y)$

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 $\Rightarrow$ $\widehat{\theta} = H^{-1}F$ = unique solution

  (no averaging of errors though)

- $n < m$: *underdetermined* case

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

- assume henceforth: $n > m$, rank$(H) = m$

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

LS: Solution

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

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

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

$$\Rightarrow \quad \widehat{\theta}_{LS} = \left(H^T(Y)\,H(Y)\right)^{-1} H^T(Y)\,F(Y) = \widehat{\theta}_{LS}(Y)$$

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

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

LS: Linear Model

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

- $\widehat{\theta}_{LS} = \left(H^T\,H\right)^{-1} H^T\,Y$

- 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: $\min_\theta E^T W\, E\ ,\quad E = [e_1 \cdots e_n]^T$

  $\widehat{\theta}_{WLS} = \arg\min_\theta G^T(\theta, Y) W\, G(\theta, Y)\ ,\quad W = W^T > 0$ weighting matrix

- LS: $W = I\ ,\quad \Rightarrow\quad E^T E = \sum_{k=1}^{n} e_k^2$

model linear in parameters

- WLS: $\min_\theta \xi_{WLS}(\theta, Y) \;=\; \min_\theta \left[F(Y) - H(Y)\,\theta\right]^T W\ \left[F(Y) - H(Y)\,\theta\right]$

- $\dfrac{\partial \xi_{WLS}}{\partial \theta} = -2H^T(Y)\, W\ \left[F(Y) - H(Y)\,\theta\right] = 0$

  $$\Rightarrow\quad \widehat{\theta}_{WLS} = \left(H^T(Y)\, W\, H(Y)\right)^{-1} H^T(Y)\, W\, F(Y) = \widehat{\theta}_{WLS}(Y)$$

- Hessian = $\dfrac{\partial}{\partial \theta}\left(\dfrac{\partial \xi_{WLS}}{\partial \theta}\right)^T = 2H^T(Y)\, W\, H(Y) > 0$

  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: $Y \;=\; H\,\theta + V \qquad (F(Y), H(Y), E) = (Y, H, V)$

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

LS estimates:

- parameters: $\widehat{\theta} = (H^T H)^{-1} H^T F$
- signal: $\widehat{S} = H\,\widehat{\theta} = P_H\,F \;\;,\quad P_H = H(H^T H)^{-1} H^T$

  projection of F/Y on the *signal subspace* = column space of H
- error/noise: $\widehat{E} = F - \widehat{S} = F - H\,\widehat{\theta} = P_H^\perp\,F \;\;,\quad P_H^\perp = I - P_H$

  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 ($P_H^\perp$) : $P_H\,H = H \;,\quad P_H^\perp\,H = 0$

  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 = signal + 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;

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

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

$$\theta^T \;=\; \left[\alpha_{1\,1} \cdots \alpha_{1\,m_1} \;\; \alpha_{2\,1} \cdots \alpha_{m_0\,m_{m_0}} \right]$$

  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 $s_k - 2\, s_{k-1} + s_{k-2} = 0$ and $s_k = a\,k + b$ (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:

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

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

Application 2: Filter Design

IIR filter design in the time domain

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

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

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

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

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

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

- consider $\mathrm{A}(q)\, y_k = \mathrm{B}(q)\, \delta_{k0} + \underbrace{\mathrm{A}(q)\, v_k}_{e_k}$ 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

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

$\mathcal{A}$ 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: $\widehat{\theta}_{LS} = (H^T H)^{-1}\, H^T Y \;=\; \arg\min_{\theta} E^T E$

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 $e_k = \mathrm{A}(q)\,v_k$).

- 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

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

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

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

- Initialization: e.g. $\widehat{\theta}^{(0)}_{WLS} = 0$ so that $\widehat{\theta}^{(1)}_{WLS} = \widehat{\theta}_{LS} \quad (\mathcal{A}(0) = I \;\Rightarrow\; W^{(1)} = I)$.

- Note: $\quad V^T V = E^T W E\;$: 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 $\mathrm{B}(z) = C(z)\,B,\;\; C(z) = [1\; z^{-1} \cdots z^{-p}]$, $\theta = B = [b_0\; b_1 \cdots b_p]^T$
- We wish to fit the frequency response $\mathrm{B}(e^{j2\pi f})$ to a desired response y_i at frequency f_i , $\;i = 1, \ldots, n'$:

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

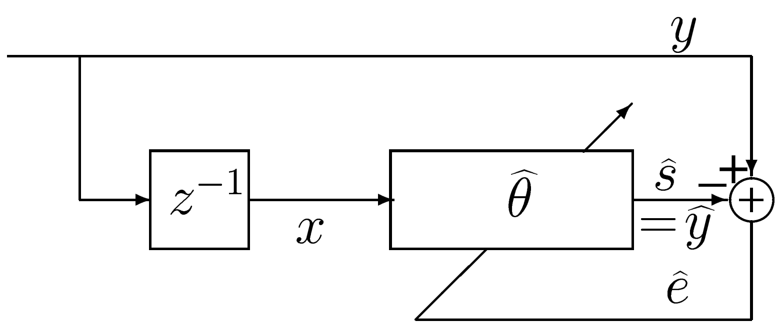
  where v_i here is clearly not noise but approximation error.

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

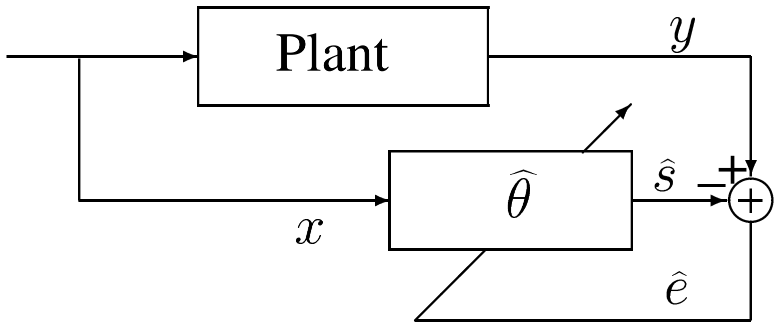
$$Y' = \begin{bmatrix} y_1 \\ \vdots \\ y_{n'} \end{bmatrix},\; H' = \begin{bmatrix} C(e^{j2\pi f_1}) \\ \vdots \\ C(e^{j2\pi f_{n'}}) \end{bmatrix},\; Y = \begin{bmatrix} \Re Y' \\ \Im Y' \end{bmatrix},\; H = \begin{bmatrix} \Re H' \\ \Im H' \end{bmatrix}$$

- 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 $W = \text{blockdiag}\{W', W'\}$, $W' = \text{diag}\{w_1, \ldots, w_{n'}\}$ can be introduced to put a higher weight $w_i > 0$ at frequencies f_i where a tighter fit is desired ($V^T W V = V'^H W' V' = \Sigma_{i=1}^{n'} w_i\,|v_i|^2$ where $V^H = (V^*)^T$).

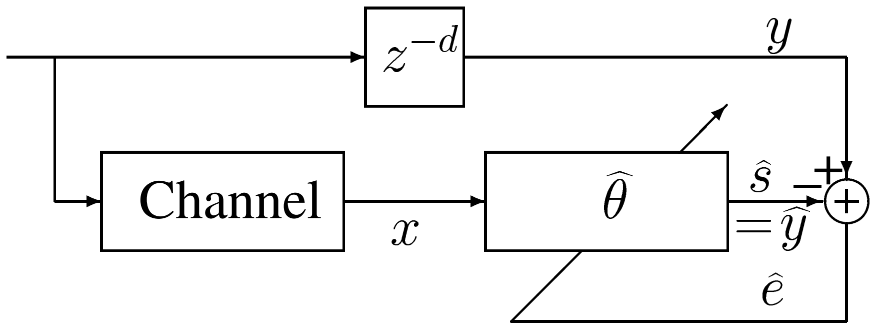
Application 3: Adaptive Filtering



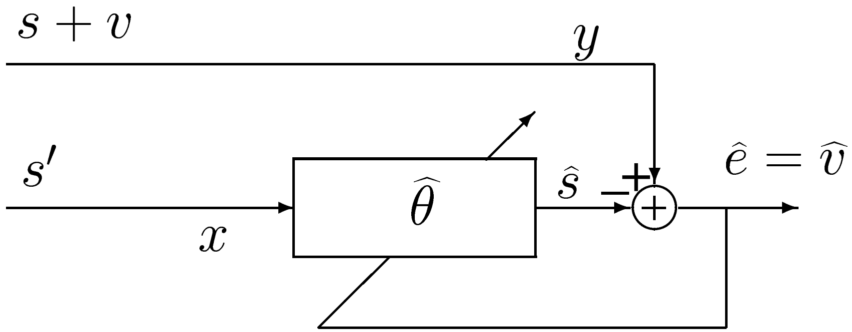
prediction, spectral estimation, whitening



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

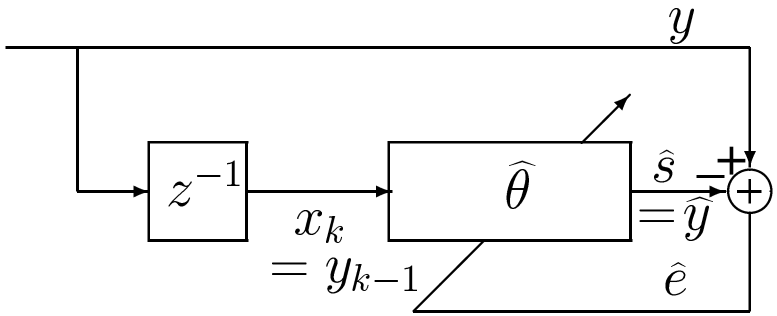
$$H = H(x_{2-m:n}) = \begin{bmatrix} x_1 & x_0 & \cdots & x_{2-m} \\ x_2 & x_1 & \cdots & x_{3-m} \\ \vdots & \vdots & \ddots & \vdots \\ x_n & x_{n-1} & \cdots & x_{n-m+1} \end{bmatrix} , \quad \theta = B = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_p \end{bmatrix} = \begin{bmatrix} b_0 \\ b_1 \\ \vdots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

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 $x_k = y_{k-1} \quad \Rightarrow \quad x_k$ noisy also

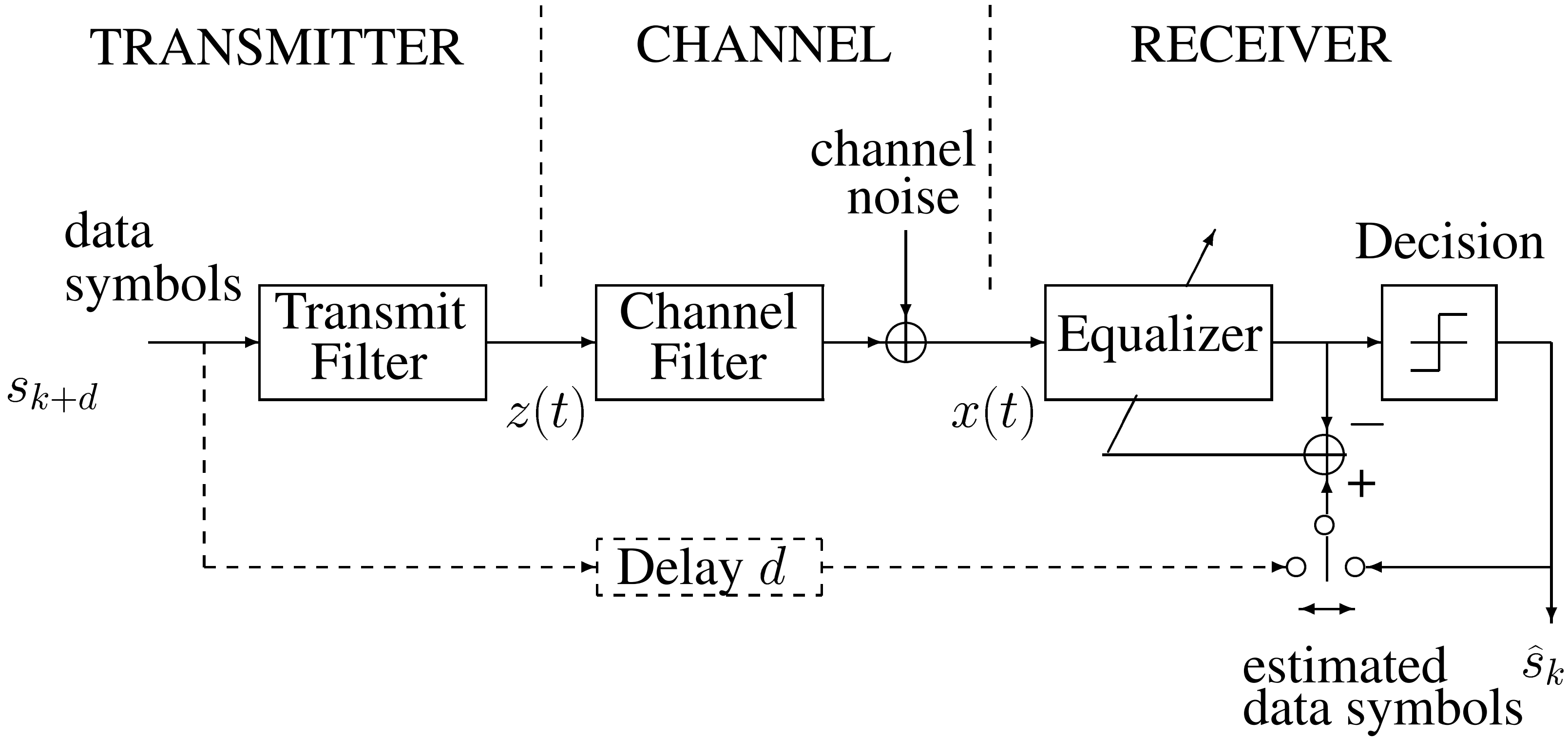
- **prediction** = s_k, e.g. stock market (multidimensional signals though)

- **whitening**: make prediction error e_k as white as possible (unpredictible part): used in signal coding (e_k easier to quantize then y_k)

- **spectral estimation/modeling**: when prediction error e_k becomes white (uncorrelated), θ contains all the spectral (correlation) information of y_k

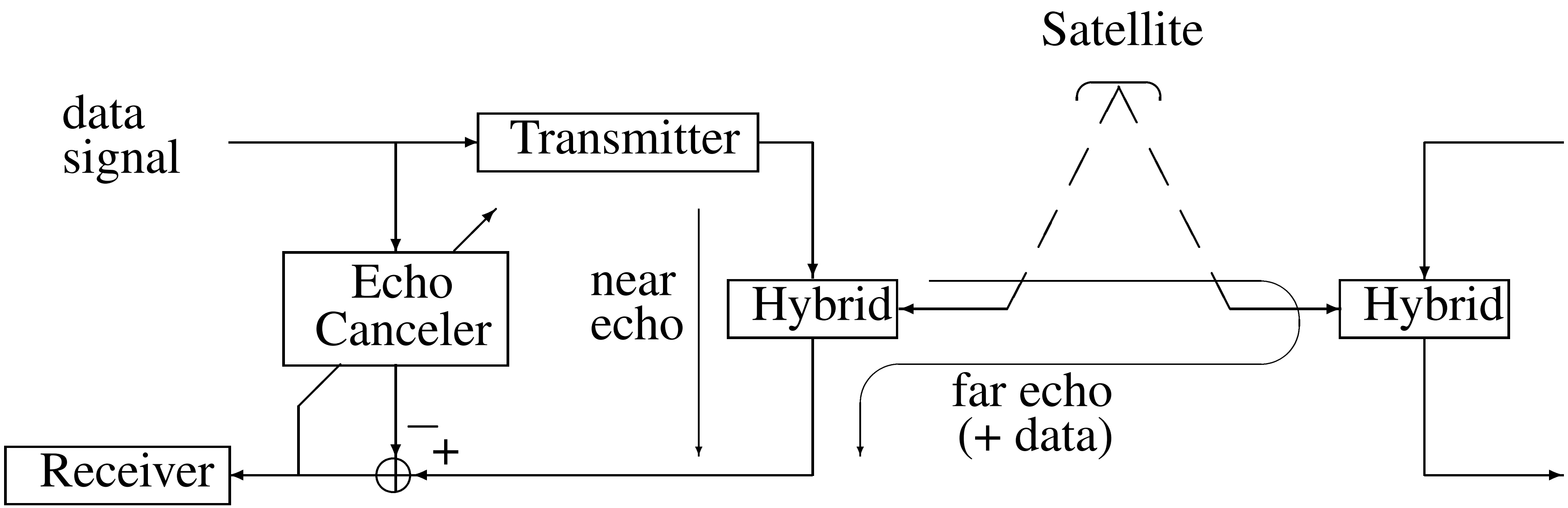
Application 3: Adaptive Filtering (4)

- **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:



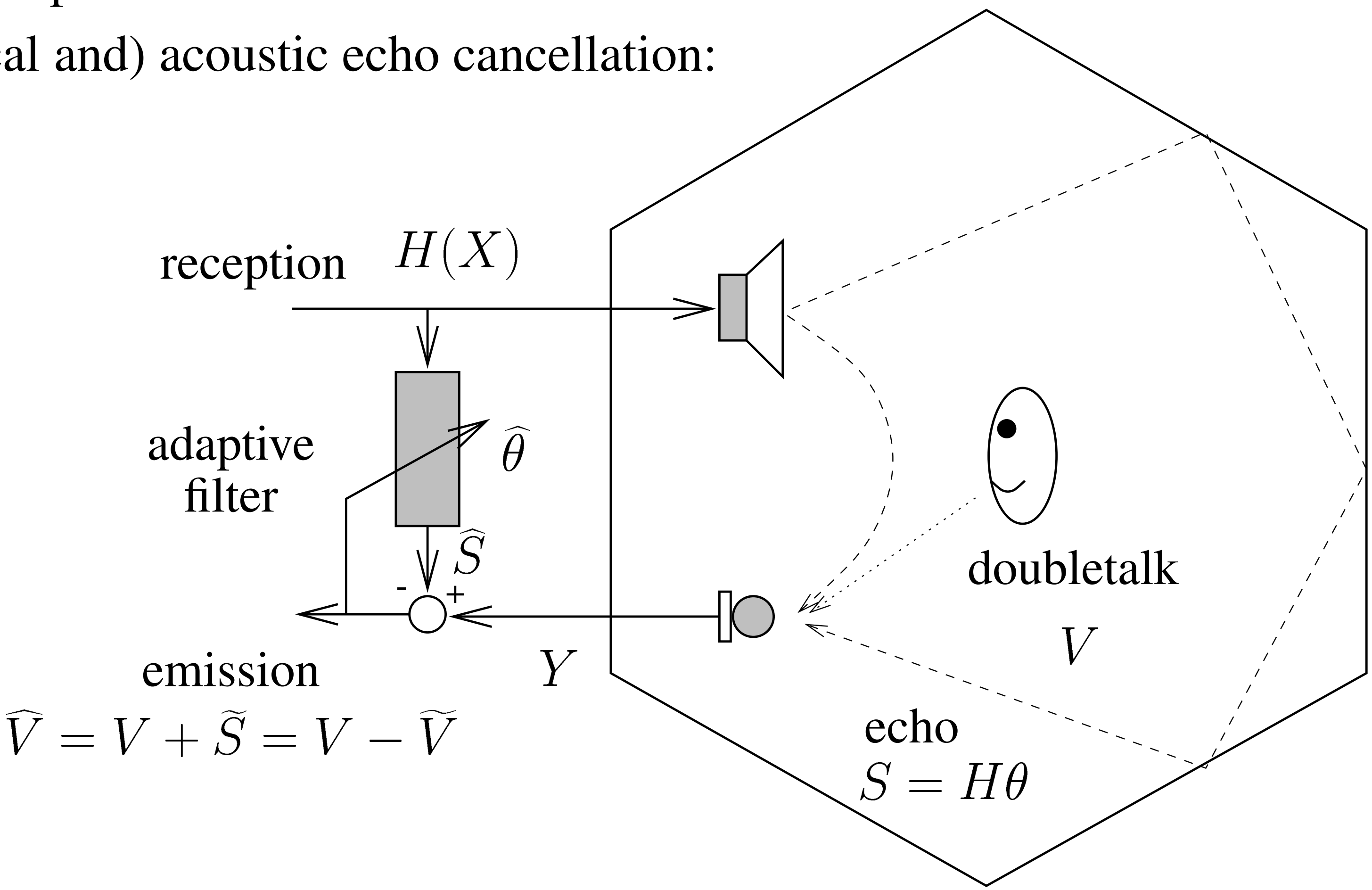
Application 3: Adaptive Filtering (5)

- **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:

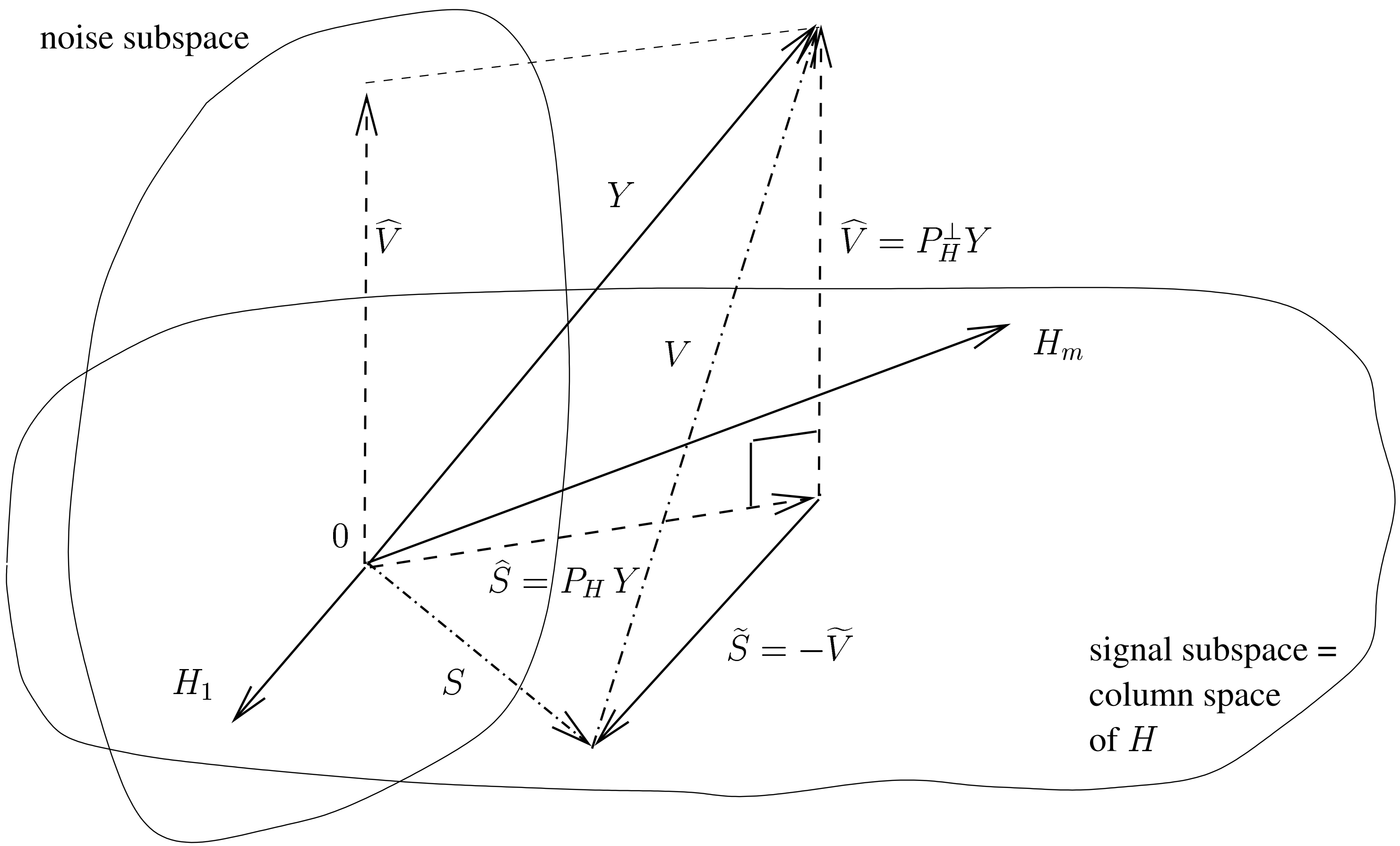


Orthogonality Principle of LS

- we found that $\widehat{\theta}_{LS}$ satisfies **orthogonality conditions of LS**

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

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: $m_X = E\,X \;\;,\quad m_Y = E\,Y$ (E = Expectation)
- correlation matrix: $R_{XY} = E\,XY^T \;\;,\quad R_{XX} = E\,XX^T$
- 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):

$$
\begin{aligned}
E\,\|X\|^2 &= \text{tr}\{\,E\,\|X\|^2\} = E\,\text{tr}\{\,\|X\|^2\} = E\,\text{tr}\{\,X^TX\} \\
&= E\,\text{tr}\{\,X\,X^T\} = \text{tr}\{\,E\,X\,X^T\} = \text{tr}\,\{\,R_{XX}\}
\end{aligned}
$$

- notation: $\begin{cases} \theta = \widehat{\theta} + \widetilde{\theta} \\ S = \widehat{S} + \widetilde{S} \\ V = \widehat{V} + \widetilde{V} \end{cases}$

Performance Analysis of LS in the Linear Model

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

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

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

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

$$b_{\widehat{\theta}}(\theta) \;=\; -E\widetilde{\theta} \;=\; E\left(\widehat{\theta}(Y) - \theta\right) \;=\; E\widehat{\theta}(Y) \;-\; \theta$$

unbiased estimator: $b_{\widehat{\theta}}(\theta) = 0$, $\forall \theta \in \Theta$ (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** = tr $\{R_{\widetilde{\theta}\widetilde{\theta}}\} = E\|\widetilde{\theta}\|_2^2$, $\;R_{\widetilde{\theta}\widetilde{\theta}} = E\,\widetilde{\theta}\,\widetilde{\theta}^T$ = estimation error correlation matrix

$$R_{\widetilde{\theta}\widetilde{\theta}} \;=\; E(\widehat{\theta}-\theta)(\widehat{\theta}-\theta)^T \;=\; E[\underline{\widehat{\theta}\;(-E\widehat{\theta}} + \underline{E\widehat{\theta})-\theta}][\underline{\widehat{\theta}\;(-E\widehat{\theta}} + \underline{E\widehat{\theta})-\theta}]^T$$

$$=\; E(\widehat{\theta}-E\widehat{\theta})(\widehat{\theta}-E\widehat{\theta})^T + (E\widehat{\theta}-\theta)(E\widehat{\theta}-\theta)^T \;=\; C_{\widehat{\theta}\widehat{\theta}} + b_{\widehat{\theta}}(\theta)b_{\widehat{\theta}}^T(\theta) = C_{\widetilde{\theta}\widetilde{\theta}} + b_{\widehat{\theta}}(\theta)b_{\widehat{\theta}}^T(\theta)$$

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

- (mean square) **consistency**: if MSE$(n) \stackrel{n\to\infty}{\longrightarrow} 0$, then $\widehat{\theta} \stackrel{n\to\infty}{\longrightarrow} \theta$ (in mean square)

Performance Analysis of LS in the Linear Model (2)

- No statistical information (about V) needed to derive $\widehat{\theta}_{WLS}$. 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: $\widehat{\theta}_{WLS} - \theta = \left(H^T W\, H\right)^{-1} H^T W\ (H\,\theta + V) - \theta = \left(H^T W\, H\right)^{-1} H^T W\, V$
- $b_{WLS} = E\,\widehat{\theta}_{WLS} - \theta = \left(H^T W\, H\right)^{-1} H^T W\ E\,V = 0$: unbiased if $E\,V = 0$
- $C_{\tilde{\theta}\tilde{\theta}}(W) = C_{\hat{\theta}\hat{\theta}}(W) = \left(H^T W\, H\right)^{-1} H^T W\, C_{VV}\, W\, H\ \left(H^T W\, H\right)^{-1}$
- optimal weighting: $W = C_{VV}^{-1}$: $\quad C_{\tilde{\theta}\tilde{\theta}}(W) \geq C_{\tilde{\theta}\tilde{\theta}}(C_{VV}^{-1}) = \left(H^T C_{VV}^{-1} H\right)^{-1}$
- LS: $\quad C_{\tilde{\theta}\tilde{\theta}} = C_{\tilde{\theta}\tilde{\theta}}(I) = \left(H^T H\right)^{-1} H^T C_{VV} H\ \left(H^T H\right)^{-1}$
- white noise: $C_{VV} = \sigma_v^2\, I_n \quad \Rightarrow \quad$ WLS^{opt} = LS and $C_{\tilde{\theta}\tilde{\theta}} = \sigma_v^2\ \left(H^T H\right)^{-1}$
- (W)LS in general consistent: $\widehat{\theta} \to \theta$ as $\frac{n}{m} \to \infty$

Performance Analysis of LS in the Linear Model (3)

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

$$\widehat{S} = H\widehat{\theta}_{LS} = P_H Y = S + P_H V \quad \Rightarrow \quad \widetilde{S} = S - \widehat{S} = -P_H V$$

  * Hence, $E\,\widehat{S} = S$: unbiased if $E\,V = 0$.
  * $C_{\widetilde{S}\widetilde{S}} = P_H C_{VV} P_H = \sigma_v^2 P_H \quad \Rightarrow \quad E\|\widetilde{S}\|^2 = \text{tr}\{C_{\widetilde{S}\widetilde{S}}\} = \sigma_v^2\,\text{tr}\{P_H\} = m\,\sigma_v^2$ remains finite!
  * Even $C_{\tilde{s}_k\tilde{s}_k} = \sigma^2_{\tilde{s}_k} = \sigma_v^2\,[P_H]_{kk}$ ($= \sigma_v^2\,\frac{m}{n}$ on the avg.) $\stackrel{\frac{n}{m}\to\infty}{\longrightarrow} 0$: $\hat{s}_k$ consistent.
    $\frac{1}{n}\sum_{k=1}^{n}[P_H]_{kk} = \frac{1}{n}\text{tr}\{P_H\} = \frac{1}{n}\text{tr}\{H(H^TH)^{-1}H^T\} = \frac{1}{n}\text{tr}\{(H^TH)^{-1}H^TH\} = \frac{1}{n}\text{tr}\{I_m\} = \frac{m}{n}$
- **noise component**:

$$\widehat{V} = Y - H\widehat{\theta}_{LS} = P_H^{\perp} Y = P_H^{\perp} V \quad \Rightarrow \quad \widetilde{V} = V - \widehat{V} = P_H V$$

  * Hence, $E\,\widehat{V} = 0$: unbiased if $E\,V = 0$ (case of a "random parameter").
  * $C_{\widetilde{V}\widetilde{V}} = C_{\widetilde{S}\widetilde{S}} \quad \Rightarrow \quad E\|\widetilde{V}\|^2 = m\,\sigma_v^2$ remains finite also!
  * Furthermore $C_{\tilde{v}_k\tilde{v}_k} = \sigma^2_{\tilde{v}_k} = \sigma^2_{\tilde{s}_k} \stackrel{\frac{n}{m}\to\infty}{\longrightarrow} 0$: $\widehat{v}_k$ consistent also. (SNR $= \frac{\sigma^2_{v_k}}{\sigma^2_{\tilde{v}_k}} = \frac{n}{m}$)
- observe: $R_{\widehat{S}\widehat{V}} = E\,\widehat{S}\widehat{V}^T = P_H C_{VV} P_H^{\perp} = \sigma_v^2 P_H P_H^{\perp} = 0$: a posteriori signal $\widehat{S} = S + P_H V$ and noise components $\widehat{V} = P_H^{\perp} V$ 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: $\frac{1}{n}H^TH = \frac{1}{n}\sum_{i=1}^{n} X_iX_i^T \overset{n\to\infty}{\longrightarrow} E\,X_iX_i^T \;=\; R_{XX}\;\; (m \times m)$

  $\Rightarrow$ approximation: $H^TH \;\approx\; n\,R_{XX}$
- 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 $C_{\tilde{\theta}\tilde{\theta}} = \left(H^TH\right)^{-1} H^TC_{VV}H\,\left(H^TH\right)^{-1} \approx \frac{\sigma_v^2}{n}\,R_{XX}^{-1}$ ($\Rightarrow$ consistency)
- Is LS criterion = $\|\widehat{V}\| \;=\; Y^TP_H^{\perp}Y$ a good indicator of estimation quality? ($\widehat{V} = Y - H\widehat{\theta}$ = LS error)

$$
\begin{aligned}
E\,\|\widehat{V}\|^2 \;&=\; E\,Y^TP_H^{\perp}Y \;=\; E\,V^TP_H^{\perp}V \;=\; E\,V^TV - E\left\{V^T\,P_H\,V\right\}\\
&=\; E\,\sum_{i=1}^{n} v_i^2 - \mathrm{tr}\{E\,P_H\,VV^T\} \;=\; n\sigma_v^2 - \mathrm{tr}\{E\,P_H\,C_{VV}\}\\
&=\; n\sigma_v^2 - \mathrm{tr}\{E\,(H^TH)^{-1}\,H^TC_{VV}H\} \overset{\text{LLN}}{\approx} n\sigma_v^2 - \mathrm{tr}\left\{(E\,H^TH)^{-1}\,E\,H^TC_{VV}H\right\}\\
&=\; n\sigma_v^2 - \mathrm{tr}\left\{(n\,R_{XX})^{-1}\,n\sigma_v^2R_{XX}\right\} \;=\; n\sigma_v^2 - \sigma_v^2\,\mathrm{tr}\left\{I_m\right\} \;=\; (n-m)\,\sigma_v^2
\end{aligned}
$$

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

Perf Analysis of LS in FIR System Identification (2)

- *white noise case*:
$$\begin{aligned} E\,\|\widehat{V}\|^2 &= E\,V^T P_H^{\perp} V = \mathrm{tr}\left\{P_H^{\perp}\,E\,V V^T\right\} \\ &= \sigma_v^2\,\mathrm{tr}\left\{P_H^{\perp}\right\} = \sigma_v^2\,\mathrm{tr}\left\{I_n - P_H\right\} = \sigma_v^2\,(n{-}m) \end{aligned}$$

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

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

  A posteriori SNRs:
$$\begin{aligned} \mathrm{SNR}_{\widehat{S}} &= \frac{E\,\|S\|^2}{E\,\|\widetilde{S}\|^2} = \frac{n\,E\,s_i^2}{m\sigma_v^2} = \frac{n}{m}\,\mathrm{SNR}_Y \\ \mathrm{SNR}_{\widehat{V}} &= \frac{E\,\|V\|^2}{E\,\|\widetilde{V}\|^2} = \frac{n\,\sigma_v^2}{m\,\sigma_v^2} = \frac{n}{m} \qquad \text{indep. of } \mathrm{SNR}_Y\ ! \end{aligned}$$

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

Perf Analysis of LS in FIR System Identification (3)

- *cross validation*: to get an idea of estimation quality, try estimate $\widehat{\theta}(Y)$ 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**:

$$\widehat{S}' = H'\widehat{\theta}_{LS} = H'(H^T H)^{-1} H^T Y = S' + H'(H^T H)^{-1} H^T V$$
$$\Rightarrow \quad \widetilde{S}' = S' - \widehat{S}' = -H'(H^T H)^{-1} H^T V$$

  * can show $E\,\|\widetilde{S}'\|^2 \approx \frac{n'}{n} m \sigma_v^2$
  * hence $\mathrm{SNR}_{\widehat{S}'} = \dfrac{E\,\|S'\|^2}{E\,\|\widetilde{S}'\|^2} = \dfrac{n}{m}\mathrm{SNR}_Y$ as before

- **noise component**:

$$\widehat{V}' = Y' - H'\widehat{\theta}_{LS} = V' + \widetilde{S}' \quad \Rightarrow \quad \widetilde{S}' = -\widetilde{V}'$$

  * $\mathrm{SNR}_{\widetilde{V}'} = \dfrac{E\,\|V'\|^2}{E\,\|\widetilde{V}'\|^2} = \dfrac{n}{m}$ but $E\,\|\widehat{V}'\|^2 = n'\,\sigma_v^2\,(1 + \dfrac{m}{n}) > E\,\|V'\|^2$ now
  * this time also $R_{\widetilde{V}'V'} = 0$ whereas $R_{\widetilde{V}V} = P_H\, R_{VV} \neq 0$ before
  * to predict performance from $\widehat{V}$: $\frac{1}{n'}\,\|\widehat{V}'\|^2 \approx \frac{n+m}{n-m}\,\frac{1}{n}\,\|\widehat{V}\|^2$ (Akaike's FPEC)

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

WLS: Performance Analysis

- No statistical information (about V) needed to derive $\widehat{\theta}_{WLS}$. 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: $\widehat{\theta}_{WLS} - \theta = \left(H^T W\, H\right)^{-1} H^T W\, \left(H\,\theta + V\right) - \theta = \left(H^T W\, H\right)^{-1} H^T W\, V$
- $E\,\widehat{\theta}_{WLS} - \theta = \left(H^T W\, H\right)^{-1} H^T W\, E\,V = 0$: unbiased
- $C_{\tilde{\theta}\tilde{\theta}}(W) = C_{\widehat{\theta}\widehat{\theta}}(W) = \left(H^T W\, H\right)^{-1} H^T W\, C_{VV}\, W\, H\, \left(H^T W\, H\right)^{-1}$
- optimal weighting: $W = C_{VV}^{-1}$: $C_{\tilde{\theta}\tilde{\theta}}(W) \geq C_{\tilde{\theta}\tilde{\theta}}(C_{VV}^{-1}) = \left(H^T C_{VV}^{-1} H\right)^{-1}$
- 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 $T = \begin{bmatrix} I_r \\ 0 \end{bmatrix}: \;\; \phi = \theta_{1:r} = \overline{\theta}_r$

$$S = H\,\theta = [\overline{H}_r \;\; \underline{H}_r] \begin{bmatrix} \overline{\theta}_r \\ \underline{\theta}_r \end{bmatrix} = \overline{H}_r \overline{\theta}_r + \underline{H}_r \underline{\theta}_r$$

- reduced-rank LS: $\;\widehat{\overline{\theta}}_r = \arg\min_{\overline{\theta}_r} \|Y - \overline{H}_r\,\overline{\theta}_r\|^2 = (\overline{H}_r^T \overline{H}_r)^{-1} \overline{H}_r^T Y$

$$\widehat{S} = \widehat{S}_r = \overline{H}_r\,\widehat{\overline{\theta}}_r = P_{\overline{H}_r} Y \quad , \quad \widehat{V} = \widehat{V}_r = Y - \widehat{S}_r = P^{\perp}_{\overline{H}_r} Y$$

-

$$\begin{aligned} \widehat{\overline{\theta}}_r &= (\overline{H}_r^T \overline{H}_r)^{-1} \overline{H}_r^T (\overline{H}_r\,\overline{\theta}_r + \underline{H}_r\,\underline{\theta}_r + V) \\ &= \overline{\theta}_r + (\overline{H}_r^T \overline{H}_r)^{-1} \overline{H}_r^T (\underline{H}_r\,\underline{\theta}_r + V) = \overline{\theta}_r - \widetilde{\overline{\theta}}_r \end{aligned}$$

$$\widehat{\theta} = \begin{bmatrix} \widehat{\overline{\theta}}_r \\ 0 \end{bmatrix} \; , \; \widetilde{\theta} = \begin{bmatrix} \widetilde{\overline{\theta}}_r \\ \underline{\theta}_r \end{bmatrix}$$

Rank Reduction in the Linear Model (2)

- **estimator bias and variance**

$$b_{\widehat{\theta}}(\theta) = \begin{bmatrix} (\overline{H}_r^T\overline{H}_r)^{-1}\overline{H}_r^T\underline{H}_r\,\underline{\theta}_r \\ -\underline{\theta}_r \end{bmatrix} \;,\;\; C_{\widetilde{\theta}\widetilde{\theta}} = \begin{bmatrix} C_{\widetilde{\overline{\theta}}_r\widetilde{\overline{\theta}}_r} & 0 \\ 0 & 0 \end{bmatrix}$$

$$C_{\widetilde{\overline{\theta}}_r\widetilde{\overline{\theta}}_r} = (\overline{H}_r^T\overline{H}_r)^{-1}\overline{H}_r^T C_{VV}\overline{H}_r(\overline{H}_r^T\overline{H}_r)^{-1} \;\;,\;\; R_{\widetilde{\theta}\widetilde{\theta}} = C_{\widetilde{\theta}\widetilde{\theta}} + b_{\widehat{\theta}}b_{\widehat{\theta}}^T$$

- **signal component**

$$\widetilde{S} = S - \widehat{S}_r = \overline{H}_r\overline{\theta}_r + \underline{H}_r\underline{\theta}_r - (\overline{H}_r\overline{\theta}_r + P_{\overline{H}_r}\underline{H}_r\underline{\theta}_r + P_{\overline{H}_r}V) = P^{\perp}_{\overline{H}_r}\underline{H}_r\underline{\theta}_r - P_{\overline{H}_r}V$$

bias : $b_{\widehat{S}_r}(\theta) = -E\,\widetilde{S} = -P^{\perp}_{\overline{H}_r}\underline{H}_r\underline{\theta}_r \neq 0$: biased !

$$R_{\widetilde{S}\widetilde{S}} = C_{\widetilde{S}\widetilde{S}} + b_{\widehat{S}_r\widehat{S}_r}b^T_{\widehat{S}_r\widehat{S}_r} = P_{\overline{H}_r}C_{VV}P_{\overline{H}_r}(P^{\perp}_{\overline{H}_r}\underline{H}_r\underline{\theta}_r)(P^{\perp}_{\overline{H}_r}\underline{H}_r\underline{\theta}_r)^T$$

- **noise component**

$$\widetilde{V} = V - \widehat{V}_r = V - (P^{\perp}_{\overline{H}_r}\underline{H}_r\underline{\theta}_r + P^{\perp}_{\overline{H}_r}V) = -P^{\perp}_{\overline{H}_r}\underline{H}_r\underline{\theta}_r + P_{\overline{H}_r}V = -\widetilde{S}$$

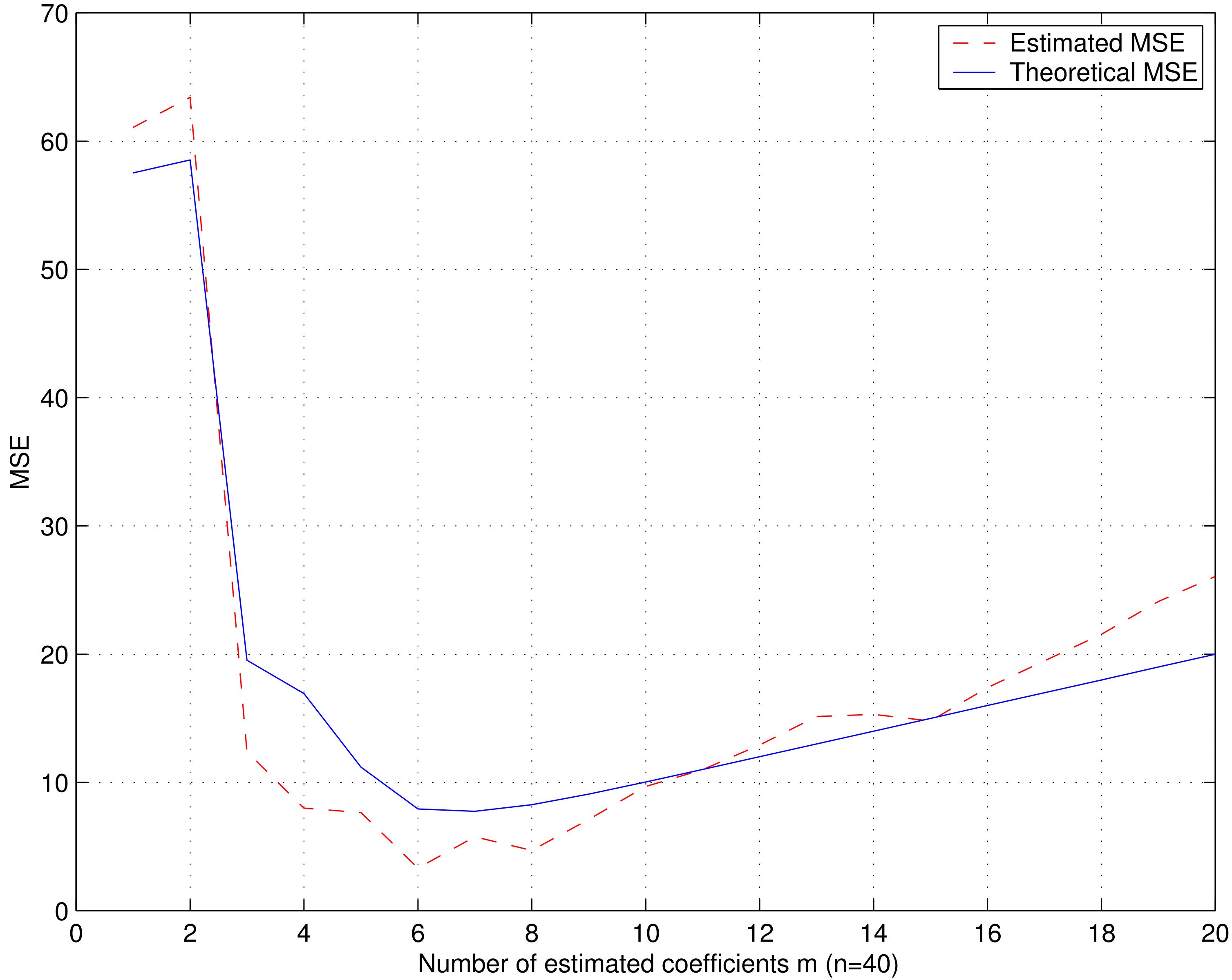
$$\text{SNR}_{\widehat{V}_r} = \frac{E\,\|V\|^2}{E\,\|\widetilde{V}\|^2} = \frac{n\sigma_v^2}{\|P^{\perp}_{\overline{H}_r}\underline{H}_r\underline{\theta}_r\|^2 \;+\; r\sigma_v^2}$$

Rank Reduction in FIR System Identification

- Assume now H filled with samples of x_k, being white noise.
- $$\begin{aligned}\mathrm{SNR}_{\widehat{V}_r} &= \frac{n\sigma_v^2}{E_X\,\|P^{\perp}_{\underline{H}_r}\underline{H}_r\underline{\theta}_r\|^2 \;+\; r\sigma_v^2} = \frac{n\sigma_v^2}{|\mathrm{bias}|^2 \;+\; r\sigma_v^2}\\ &= \frac{1}{\frac{\sigma_x^2}{\sigma_v^2}\|\underline{\theta}_r\|^2+\frac{r}{n}} = \frac{1}{\mathrm{SNR}_Y\,\frac{\|\theta_r\|^2|}{\|\theta\|^2}+\frac{r}{n}}\end{aligned}$$
- to maximize $\mathrm{SNR}_{\widehat{V}_r}$, need to minimize $|\mathrm{bias}|^2 \;+\; r\sigma_v^2$
- we have $E\,\|\widehat{V}_m\|^2 = (n{-}m)\,\sigma_v^2,\quad E\,\|\widehat{V}_r\|^2 = |\mathrm{bias}|^2 + (n{-}r)\,\sigma_v^2$
- Hence can estimate

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

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:
  - $\widehat{\theta}_{MMSE}$ preferable
  - $\widehat{\theta}_{MAP}$ easier to calculate
  - $\widehat{\theta}_{LMMSE}$ 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