

# Linear Model

Aleksandar Dogandžić

2017-02-16

## Contents

Introduction	1
Polynomial curve fitting	1
Sinusoidal amplitude and phase estimation	2
Problem Formulation and MVU Estimation	3
MVU linear model estimator	3
MVU Estimation for Colored Noise	5
BLUE	6
Gauss-Markov theorem	6
LS Approach to Estimation	9
Orthogonality principle	11
Computational Aspects	12
Applications of Least Squares	13
$H$ known up to $\theta$	20
Linearly constrained least squares	20
Underdetermined systems	21

READING: §4 and 6 in the textbook and (Hero 2015, §5.6 and 5.7).

## Introduction

We start with a couple of examples.

### Polynomial curve fitting

MODEL a continuous-time (CT) signal  $X(t)$  as a polynomial of degree  $p - 1$  in additive noise:

$$X(t) = \theta_1 + \theta_2 t + \cdots + \theta_p t^{p-1} + W(t).$$

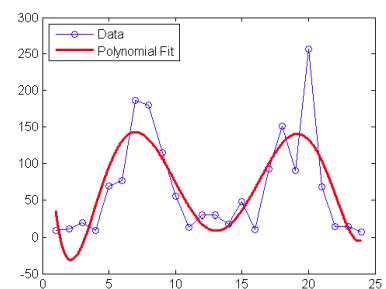


Figure 1: Polynomial fit.

We observe  $\mathbf{x} = (x[n])_{n=0}^{N-1} = (x(t_n))_{n=0}^{N-1}$ . Define

$$\begin{aligned} \mathbf{W} &= [W(t_0), \dots, W(t_{N-1})]^\top \\ \boldsymbol{\theta} &= [\theta_1, \dots, \theta_p]^\top \\ H &= \begin{bmatrix} 1 & t_0 & \dots & t_0^{p-1} \\ 1 & t_1 & \dots & t_1^{p-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & t_{N-1} & \dots & t_{N-1}^{p-1} \end{bmatrix}_{N \times p} \end{aligned} \quad .$$

The measurement model is then

$$\mathbf{X} = H\boldsymbol{\theta} + \mathbf{W} \quad (1)$$

where  $H$  is known and  $\boldsymbol{\theta}$  is the parameter vector to be estimated.

Sinusoidal amplitude and phase estimation

MODEL the measured signal  $x(t)$  as a superposition of  $p/2$  sinusoids that have known frequencies but unknown amplitudes and phases:

$$X(t) = \sum_{k=1}^{p/2} r_k \sin(\omega_k t + \phi_k) + W(t).$$

This model is linear in  $r_k$  but nonlinear in  $\phi_k$ . However, we can rewrite it as

$$X(t) = \sum_{k=1}^{p/2} [a_k \cos(\omega_k t) + b_k \sin(\omega_k t)] + W(t)$$

and we get the model (1), with a different (trigonometric)  $H$ .


For  $p/2 = 2$  sinusoids:

$$H = \begin{bmatrix} \cos(\omega_1 t_0) & \cos(\omega_2 t_0) & \sin(\omega_1 t_0) & \sin(\omega_2 t_0) \\ \cos(\omega_1 t_1) & \cos(\omega_2 t_1) & \sin(\omega_1 t_1) & \sin(\omega_2 t_1) \\ \vdots & \vdots & \vdots & \vdots \\ \cos(\omega_1 t_{N-1}) & \cos(\omega_2 t_{N-1}) & \sin(\omega_1 t_{N-1}) & \sin(\omega_2 t_{N-1}) \end{bmatrix}$$

and

$$\boldsymbol{\theta} = [a_1, \dots, a_{p/2}, b_1, \dots, b_{p/2}]^\top.$$

Once we compute an estimate  $\hat{\boldsymbol{\theta}}$  of  $\boldsymbol{\theta}$ ,  $\hat{r}_k$  and  $\hat{\phi}_k$  are obtained using the simple conversion from rectangular to polar coordinates.

 NOTE: Even if  $\hat{\boldsymbol{\theta}}$  is a minimum-variance unbiased (MVU) estimator,  $(\hat{r}_k)_{k=1}^{p/2}$  and  $(\hat{\phi}_k)_{k=1}^{p/2}$  will only be *asymptotically MVU*<sup>1</sup>, as we will see later.

<sup>1</sup> for large number of measurements  $N$

## Problem Formulation and MVU Estimation

CONSIDER the model (1) where  $H$  is a known *deterministic*  $N \times p$  matrix with

$$N \geq p. \quad (2)$$

We wish to estimate the parameter vector  $\theta$ .

Assume that  $W$  is white Gaussian noise:

$$W \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_N)$$

where  $\sigma^2$  is the noise variance. If  $\sigma^2$  is known, the likelihood function of  $\theta$  for the measurement vector  $x$  is

$$f_{X|\Theta}(x|\theta) = \frac{1}{\sqrt{\det(2\pi\sigma^2 I_N)}} \exp\left[-\frac{1}{2\sigma^2}(x - H\theta)^\top(x - H\theta)\right].$$

In this case, the *identifiability condition* [(10) in intro]:

$$f_{X|\Theta}(\cdot|\theta_1) = f_{X|\Theta}(\cdot|\theta_2) \iff \theta_1 = \theta_2$$

reduces to

$$H\theta_1 = H\theta_2 \iff \theta_1 = \theta_2.$$

To satisfy this condition, we assume that  $H$  has full rank  $p$ .<sup>2</sup>

<sup>2</sup> We can handle the case where the model is not identifiable; then, the MVU estimate of  $\theta$  will not be unique.

MVU linear model estimator

FOR the linear model

$$X = H\theta + W \quad (3)$$

where the known  $N \times p$  matrix  $H$  has full rank  $p$  and

$$W \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_N)$$

the MVU estimator of  $\theta$  is given by

$$\begin{aligned} \hat{\theta} &= \hat{\theta}(X) \\ &= (H^\top H)^{-1} H^\top X. \end{aligned} \quad (4)$$

The covariance matrix of  $\hat{\theta}$  given  $\theta$  attains the Cramér-Rao bound (CRB) for all  $\theta \in \mathbb{R}^p$  and is given by

$$\begin{aligned} \text{cov}_{X|\Theta}(\hat{\theta}|\theta) &= \text{MSE}\{\hat{\theta}\} \\ &= \mathbb{E}_{X|\Theta}[(\hat{\theta} - \theta)(\hat{\theta} - \theta)^\top | \theta] \\ &= \sigma^2 (H^\top H)^{-1}. \end{aligned}$$

*Proof:* The measurement model probability density function (pdf) is  $f_{X|\Theta}(\mathbf{x} | \boldsymbol{\theta}) = \mathcal{N}(\mathbf{x} | H\boldsymbol{\theta}, \sigma^2 I_N)$  with score function

$$\frac{\partial}{\partial \boldsymbol{\theta}} \ln f_{X|\Theta}(\mathbf{x} | \boldsymbol{\theta}) = \frac{1}{\sigma^2} \underbrace{H^\top H}_{\mathcal{I}(\boldsymbol{\theta})} \underbrace{[(H^\top H)^{-1} H^\top \mathbf{x} - \boldsymbol{\theta}]}_{\hat{\boldsymbol{\theta}}(\mathbf{x})}$$

which fits the CRB tightness condition in (7) of handout `multipar_gauss_crb`.  $\square$

*An alternative proof:* Verifying the unbiasedness of  $\hat{\boldsymbol{\theta}}$  and the covariance matrix expression for  $\text{cov}_{X|\Theta}(\hat{\boldsymbol{\theta}} | \boldsymbol{\theta})$  proves the result. For the above model,

$$\text{CRB}(\boldsymbol{\theta}) = \mathcal{I}^{-1}(\boldsymbol{\theta})$$

and the Fisher information matrix (FIM)  $\mathcal{I}(\boldsymbol{\theta})$  for  $\boldsymbol{\theta}$  is computed using the general Gaussian FIM expression (13) from handout `multipar_gauss_crb`:

$$[\mathcal{I}(\boldsymbol{\theta})]_{i,k} = \frac{1}{\sigma^2} \frac{\partial \boldsymbol{\mu}^\top(\boldsymbol{\theta})}{\partial \theta_i} \frac{\partial \boldsymbol{\mu}(\boldsymbol{\theta})}{\partial \theta_k}$$

where  $\boldsymbol{\mu}(\boldsymbol{\theta}) = H\boldsymbol{\theta}$ . Now,

$$\frac{\partial \boldsymbol{\mu}(\boldsymbol{\theta})}{\partial \theta_i} = \frac{\partial (H\boldsymbol{\theta})}{\partial \theta_i} = \mathbf{h}_i \quad \text{\textit{i}th column of } H$$

which implies

$$\mathcal{I}(\boldsymbol{\theta}) = \frac{1}{\sigma^2} H^\top H \quad (5a)$$

$$\text{CRB}(\boldsymbol{\theta}) = \sigma^2 (H^\top H)^{-1}. \quad (5b)$$

$\square$

#### \* COMMENTS:

- Since the joint FIM and CRB for

$$\boldsymbol{\rho} = \begin{bmatrix} \boldsymbol{\theta} \\ \sigma^2 \end{bmatrix}$$

are block-diagonal matrices,  $\boldsymbol{\theta}$  and  $\sigma^2$  are decoupled:  $\text{CRB}(\boldsymbol{\theta})$  is the same regardless of whether or not  $\sigma^2$  is known. To be more precise,  $\text{CRB}(\boldsymbol{\theta})$  in (5b) is CRB for  $\boldsymbol{\theta}$  assuming that  $\sigma^2$  is known and the full CRB for  $\boldsymbol{\rho}$  when both  $\boldsymbol{\theta}$  and  $\sigma^2$  are *unknown* is

$$\text{CRB}_{\boldsymbol{\rho}, \boldsymbol{\rho}}(\boldsymbol{\rho}) = \begin{bmatrix} \text{same as (5b)} & \\ \text{CRB}_{\boldsymbol{\theta}, \boldsymbol{\theta}}(\boldsymbol{\rho}) & \mathbf{0}_{p \times 1} \\ \mathbf{0}_{1 \times p} & \text{CRB}_{\sigma^2, \sigma^2}(\boldsymbol{\rho}) \end{bmatrix}.$$

Therefore,  $\hat{\boldsymbol{\theta}}$  in (4) is the MVU estimator of  $\boldsymbol{\theta}$  regardless of whether or not  $\sigma^2$  is known.

- $\hat{\theta}$  in (4) coincides with the least-squares (LS) estimator of  $\theta$ :<sup>3</sup>

$$\hat{\theta} = \arg \min_{\theta} \|x - H\theta\|_2^2$$

which can be shown by differentiating  $\|x - H\theta\|_2^2$  with respect to  $\theta$  and setting the result to zero or by completing the squares.<sup>4</sup> The solution in Theorem 3 is numerically not sound as given. Later in this handout, we will see a geometric interpretation of the LS approach and a numerically stable solution via QR factorization.

<sup>3</sup> Matlab has the backslash command for computing the LS solution

$$\theta = H \backslash x;$$

<sup>4</sup> Here,

$$\|a\|_2^2 \triangleq a^T a \quad \ell_2 \text{ (Euclidean) norm}$$

for an arbitrary real-valued vector  $a$ .

## MVU Estimation for Colored Noise

SUPPOSE that the noise is colored:

$$W \sim \mathcal{N}(0, \sigma^2 C)$$

where  $C \neq I$  is a known positive definite matrix. Here,  $\sigma^2$  can be an unknown parameter or a known constant. We can use *prewhitening* to get back to the white-noise case. Compute the Cholesky factorization of  $C^{-1}$ :

$$C^{-1} = D^T D$$

(Any other square-root factorization could be used as well.)

Now, define the transformed measurement model:

$$\underbrace{DX}_{X^{\text{transf}}} = \underbrace{DH}_{H^{\text{transf}}} \theta + \underbrace{DW}_{W^{\text{transf}}}.$$

Clearly,

$$\begin{aligned} \text{cov}_{W^{\text{transf}}}(W^{\text{transf}}) &= \text{cov}_W(DW D^T) \\ &= D(\sigma^2 C)D^T \\ &= \sigma^2 D(D^T D)^{-1} D^T = \sigma^2 I \end{aligned}$$

and

$$W^{\text{transf}} \sim \mathcal{N}(0, \sigma^2 I)$$

reducing the problem to the white-noise case.

For colored Gaussian noise with covariance matrix  $\sigma^2 C$  where  $C$  is a known positive-definite matrix, the MVU estimator of  $\theta$  is

$$\hat{\theta} = (H^T C^{-1} H)^{-1} H^T C^{-1} x.$$

The covariance matrix of  $\hat{\theta} = \hat{\theta}(X)$  attains the CRB for  $\theta$  and is given by

$$\begin{aligned} \text{cov}_{X|\theta}(\hat{\theta} | \theta) &= \text{MSE}\{\hat{\theta}\} \\ &= \text{CRB}(\theta) \\ &= \sigma^2 (H^T C^{-1} H)^{-1}. \end{aligned}$$

Matlab: `D = inv(chol(C))';`

## BLUE

CONSIDER the linear model (3) where  $\mathbf{W}$  has zero mean and a positive-definite covariance matrix

$$\mathbf{C} = \text{cov}_{\mathbf{W}}(\mathbf{W}) = \mathbb{E}_{\mathbf{W}}(\mathbf{W}\mathbf{W}^{\top}).$$

We look for the best linear unbiased estimator (BLUE) of  $\boldsymbol{\theta}$ . Hence, we restrict our estimators  $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(X)$  to be

- *linear*, i.e.,

$$\hat{\boldsymbol{\theta}}(x) = \mathbf{A}^{\top} \mathbf{x}$$

and

- *unbiased*, i.e.,

$$\mathbb{E}_{X|\boldsymbol{\theta}}[\hat{\boldsymbol{\theta}}(X) | \boldsymbol{\theta}] = \boldsymbol{\theta}$$

and search for a  $\hat{\boldsymbol{\theta}}$  that minimizes the estimator covariance matrix  $\text{cov}_{X|\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}} | \boldsymbol{\theta})$ .

## Gauss-Markov theorem

THE BLUE of  $\boldsymbol{\theta}$  is

$$\begin{aligned} \hat{\boldsymbol{\theta}}_{\text{BLUE}} &= \hat{\boldsymbol{\theta}}_{\text{BLUE}}(x) \\ &= (\mathbf{H}^{\top} \mathbf{C}^{-1} \mathbf{H})^{-1} \mathbf{H}^{\top} \mathbf{C}^{-1} \mathbf{x} \end{aligned} \quad (6)$$

and its covariance matrix is

$$\text{cov}_{X|\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}_{\text{BLUE}} | \boldsymbol{\theta}) = (\mathbf{H}^{\top} \mathbf{C}^{-1} \mathbf{H})^{-1}. \quad (7)$$

which holds *independently* of the distribution of the noise  $\mathbf{W}$ : All we impose on  $\mathbf{W}$  is that it has zero mean and known positive-definite covariance matrix  $\mathbf{C}$ .

 NOTE:

- it is easy to generalize the above result and show that (6) is BLUE in the more general case where  $\text{cov}_{\mathbf{W}}(\mathbf{W}) = \sigma^2 \mathbf{C}$ ,  $\mathbf{C}$  is known and positive definite, and  $\sigma^2$  is *unknown*.
- The estimate  $\hat{\boldsymbol{\theta}}$  is statistically efficient (attains the CRB) if  $\mathbf{W}$  is Gaussian, but it is not efficient in general. For non-Gaussian measurement models, there might exist a better nonlinear estimate.

*Proof:* Linear estimates of  $\boldsymbol{\theta}$ :

$$\hat{\boldsymbol{\theta}}(X) = \mathbf{A}^{\top} \mathbf{X}. \quad (8)$$

To simplify the notation, we use  $\mathbb{E}_{\mathbf{X}}[\cdot]$  instead of  $\mathbb{E}_{X|\boldsymbol{\theta}}[\cdot | \boldsymbol{\theta}]$ .

The unbiasedness condition of  $\hat{\theta}$  in (8) implies:

$$\begin{aligned}
 \theta &= E_X[\hat{\theta}(X)] \\
 &= E_X(A^T X) \\
 &= E_W[A^T(H\theta + W)] \\
 &= A^T H \theta
 \end{aligned} \tag{9}$$

see (3)

yielding

$$A^T H = I. \tag{10}$$

Now, use (8) and (10) to compute  $\text{cov}_X(\hat{\theta} | \theta)$  for an arbitrary linear unbiased estimator  $\hat{\theta}$  of  $\theta$ :

$$\begin{aligned}
 \text{cov}_X(\hat{\theta} | \theta) &= \text{cov}_W[A^T(H\theta + W)] \\
 &= A^T \text{cov}_W(W) A
 \end{aligned} \tag{11}$$

see (3)

and

$$\begin{aligned}
 \text{cov}_X(\hat{\theta}_{\text{BLUE}}) &= (H^T C^{-1} H)^{-1} H^T C^{-1} C C^{-1} H (H^T C^{-1} H)^{-1} \\
 &= (H^T C^{-1} H)^{-1}.
 \end{aligned} \tag{12}$$

To prove that  $\hat{\theta}_{\text{BLUE}}$  has the smallest covariance matrix within the family of linear unbiased estimators  $\hat{\theta}$  in (8) satisfying (10), we show that

$$\text{cov}_X(\hat{\theta}_{\text{BLUE}}) \leq \text{cov}_X(\hat{\theta})$$

as follows:

$$\begin{aligned}
 \text{cov}_X[\hat{\theta}(X)] - \text{cov}_X[\hat{\theta}_{\text{BLUE}}(X)] &= A^T C A - (H^T C^{-1} H)^{-1} \\
 &\stackrel{A^T H = I}{=} A^T C A - A^T H (H^T C^{-1} H)^{-1} H^T A \\
 &= A^T [C - H (H^T C^{-1} H)^{-1} H^T] A \\
 &= A^T [C - H (H^T C^{-1} H)^{-1} H^T] C^{-1} [C - H (H^T C^{-1} H)^{-1} H^T] A
 \end{aligned}$$

which is always positive semidefinite. □

\* EXAMPLE. Estimate DC level in colored noise.

Example 4.4 in the textbook.

$$(X[n])_{n=0}^{N-1} = a + W[n]$$

where  $W = (W[n])_{n=0}^{N-1}$  is colored noise vector with zero mean and known covariance matrix

$$\text{cov}_W(W) = C.$$

Hence, substituting  $H = \mathbf{h} = \mathbf{1} = [1, 1, \dots, 1]^T$  into (6) and (7) yields BLUE of  $a$

$$\begin{aligned}
 \hat{a} &= (\mathbf{h}^T C^{-1} \mathbf{h})^{-1} \mathbf{h}^T C^{-1} X \\
 &= \frac{\mathbf{1}^T C^{-1} X}{\mathbf{1}^T C^{-1} \mathbf{1}}
 \end{aligned} \tag{13}$$

and its variance given  $a$  is

$$\text{var}_{X|A}(\hat{a} | a) = \frac{1}{\mathbf{1}^\top C^{-1} \mathbf{1}}.$$

Consider the Cholesky factorization

$$C^{-1} = D^\top D.$$

Then, BLUE of  $a$  becomes

$$\hat{a} = \frac{\mathbf{1}^\top D^\top D \mathbf{X}}{\mathbf{1}^\top D^\top D \mathbf{1}} = \frac{(D\mathbf{1})^\top \overbrace{D\mathbf{x}}^{\mathbf{x}^{\text{transf}}}}{\mathbf{1}^\top D^\top D \mathbf{1}} = \sum_{n=0}^{N-1} d_n x^{\text{transf}}[n]$$

where

$$d_n = \frac{[D\mathbf{1}]_n}{\mathbf{1}^\top D^\top D \mathbf{1}}.$$

\* **EXAMPLE.** Sometimes, BLUE is completely wrong. For example, consider  $(X[n])_{n=0}^{N-1}$  white Gaussian noise with unknown variance  $\sigma^2$ . The MVU estimator of  $\sigma^2$  is

$$\hat{\sigma}^2 = \hat{\sigma}^2(X) = \frac{1}{N} \sum_{n=0}^{N-1} X^2[n].$$

On the other hand,

$$\hat{\sigma}_{\text{BLUE}}^2 = \hat{\sigma}_{\text{BLUE}}^2(X) = \sum_{n=0}^{N-1} a_n X[n].$$


For an estimator  $\hat{\sigma}^2$  to be unbiased, we need

$$\mathbb{E}_{X|\sigma^2}(\hat{\sigma}^2 | \sigma^2) = \sigma^2$$

but

$$\mathbb{E}_{X|\sigma^2}(\hat{\sigma}_{\text{BLUE}}^2 | \sigma^2) = \sum_{n=0}^{N-1} a_n \mathbb{E}_{X|\sigma^2}(x[n] | \sigma^2) = 0.$$

It is impossible to find  $a_n$ s to make  $\hat{\sigma}_{\text{BLUE}}^2$  unbiased.

 **NOTE:** Although BLUE is not suitable for this problem, *transforming the data*

$$Y[n] = X^2[n]$$

and designing BLUE for  $Y[n]$  would produce a viable estimator of  $\sigma^2$ .



## LS Approach to Estimation

CONSIDER the measurement model (1):

$$\mathbf{x} = H\boldsymbol{\theta} + \mathbf{W} \quad (14)$$

where  $\mathbf{x} = [x[0], \dots, x[N-1]]^T$  is the vector of measurements,  $H$  is a known *regression vector matrix* and  $\mathbf{W}$  is the *error* vector.

When  $H$  is a full-rank matrix and, therefore,  $H^T H$  is nonsingular, a common approach is to find the LS solution, i.e., the vector  $\boldsymbol{\theta}$  that minimizes the squared Euclidean norm of the error vector  $\mathbf{x} - H\boldsymbol{\theta}$ :

$$\min_{\boldsymbol{\theta}} \|\mathbf{x} - H\boldsymbol{\theta}\|_2^2$$

has closed form:

$$\hat{\boldsymbol{\theta}} = (H^T H)^{-1} H^T \mathbf{x}. \quad (15)$$

☞ FOR  $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$ , the error  $\mathbf{x} - H\hat{\boldsymbol{\theta}}$  is orthogonal to the columns of  $H$ :

$$\begin{aligned} H^T [\mathbf{x} - H \underbrace{(H^T H)^{-1} H^T \mathbf{x}}_{\hat{\boldsymbol{\theta}}}] &= H^T \mathbf{x} - H^T H (H^T H)^{-1} H^T \mathbf{x} \\ &= \mathbf{0} \end{aligned}$$

with geometric interpretation depicted in Fig. 3.

*Proof of (15) using completion of squares.* If  $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}} + \boldsymbol{\epsilon}$ , then

$$\begin{aligned} \|\mathbf{x} - H\boldsymbol{\theta}\|_2^2 &= \|\mathbf{x} - H\hat{\boldsymbol{\theta}} - H\boldsymbol{\epsilon}\|_2^2 \\ &= (\mathbf{x} - H\hat{\boldsymbol{\theta}})^T (\mathbf{x} - H\hat{\boldsymbol{\theta}}) - \underbrace{\boldsymbol{\epsilon}^T H^T (\mathbf{x} - H\hat{\boldsymbol{\theta}})}_0 \\ &\quad - \underbrace{(\mathbf{x} - H\hat{\boldsymbol{\theta}})^T H}_{0} \boldsymbol{\epsilon} + \boldsymbol{\epsilon}^T H^T H \boldsymbol{\epsilon} \\ &= \|\mathbf{x} - H\hat{\boldsymbol{\theta}}\|_2^2 + \|H\boldsymbol{\epsilon}\|_2^2 \\ &\geq \|\mathbf{x} - H\hat{\boldsymbol{\theta}}\|_2^2 \end{aligned} \quad (16)$$

thus  $\hat{\boldsymbol{\theta}}$  is optimal.  $\square$

The best LS approximation  $\hat{\mathbf{x}}$  to  $\mathbf{x}$  is given by

$$\hat{\mathbf{x}} = H\hat{\boldsymbol{\theta}} = H(H^T H)^{-1} H^T \mathbf{x} = P_H \mathbf{x}$$

where

$$P_H = H(H^T H)^{-1} H^T$$

is called *projection matrix*, with properties:

We can also use weighted least squares, which allows us to assign different weights to measurements. For example, if

$$C = E_{\mathbf{W}}(\mathbf{W} \mathbf{W}^T)$$

is known, we could use

$$\hat{\boldsymbol{\theta}} = \arg \min_{\boldsymbol{\theta}} (\mathbf{x} - H\boldsymbol{\theta})^T C^{-1} (\mathbf{x} - H\boldsymbol{\theta})$$

orthogonality principle

use orthogonality (16)

•

$$P_H \mathbf{h} = \mathbf{h}$$

if the vector  $\mathbf{h}$  belongs to the column space of  $H$  and

•

$$P_H \mathbf{a} = \mathbf{0}$$

if  $\mathbf{a}$  is orthogonal to the column space of  $H$ .

NOTE:  $P_H = H(H^\top H)^{-1}H^\top$  is the projection matrix onto the column space of  $H$ , and  $P_H^\perp = I - P_H$  is the complementary projection matrix.

Define  $H = [\mathbf{h}_1 \cdots \mathbf{h}_p]$  and rewrite (14) as

$$\mathbf{X} = \sum_{k=1}^p \mathbf{h}_k \theta_k + \mathbf{W}.$$

The *signal part*  $H\boldsymbol{\theta}$  of the  $N$ -vector  $\mathbf{X}$  is confined to the  $p$ -dimensional subspace spanned by  $[\mathbf{h}_1 \cdots \mathbf{h}_p]$ , the *signal subspace*. The signal estimate

$$\hat{\mathbf{x}} = H\hat{\boldsymbol{\theta}} = \underbrace{H(H^\top H)^{-1}H^\top}_{P_H} \mathbf{x} = P_H \mathbf{x}$$

is the *orthogonal projection* of  $\mathbf{x}$  onto  $\text{span}(H)$ , the column space of  $H$ . The error

$$\hat{\mathbf{w}} = \mathbf{x} - H(H^\top H)^{-1}H^\top \mathbf{x}$$

is the *orthogonal projection* of  $\mathbf{x}$  onto the *orthogonal complement* of  $\text{span}(H)$ .

\* A real-valued square matrix  $P$  is a projection matrix if and only if it satisfies

$$P^\top = P \quad (17a)$$

symmetry

$$P^2 = P. \quad (17b)$$

idempotency

The minimum LS error is

$$\begin{aligned} \min \|\mathbf{x} - H\boldsymbol{\theta}\|_2^2 &= \|\mathbf{x} - H\hat{\boldsymbol{\theta}}\|_2^2 \\ &= \|[I - H(H^\top H)^{-1}H^\top]\mathbf{x}\|_2^2 \\ &= \|(I - P_H)\mathbf{x}\|_2^2 \\ &= \|P_H^\perp \mathbf{x}\|_2^2 \\ &= \mathbf{x}^\top P_H^\perp \mathbf{x} \end{aligned} \quad (18)$$

where  $P_H^\perp = I - P_H$  is the projection matrix onto the subspace orthogonal to the column space of  $H$ .<sup>5</sup>

<sup>5</sup> verify that  $P_H$  and  $P_H^\perp$  satisfy (17a) and (17b)

# Orthogonality principle

THE VECTOR  $\hat{\mathbf{x}} = H\hat{\boldsymbol{\theta}}$  is the orthogonal projection of  $\mathbf{x}$  onto the subspace  $\text{span}(H)$  spanned by the columns of  $H$ , see Fig. 2. This is true even if  $H$  is not of full column rank.<sup>6</sup>

*Proof: Pythagoras.*

$$\mathbf{x} - \hat{\mathbf{x}} \perp \text{span}(H)$$

equivalent to

$$\mathbf{x} - H\hat{\boldsymbol{\theta}} \perp \mathbf{h}_i, \quad \forall i$$

where  $\mathbf{h}_i$  are the columns of  $H$ , equivalent to

$$H^T(\mathbf{x} - H\hat{\boldsymbol{\theta}}) = \mathbf{0}.$$

<sup>6</sup> row rank and column rank are the same; referring to column rank emphasizes that we refer to the number of linearly independent columns

the same as (16)

□

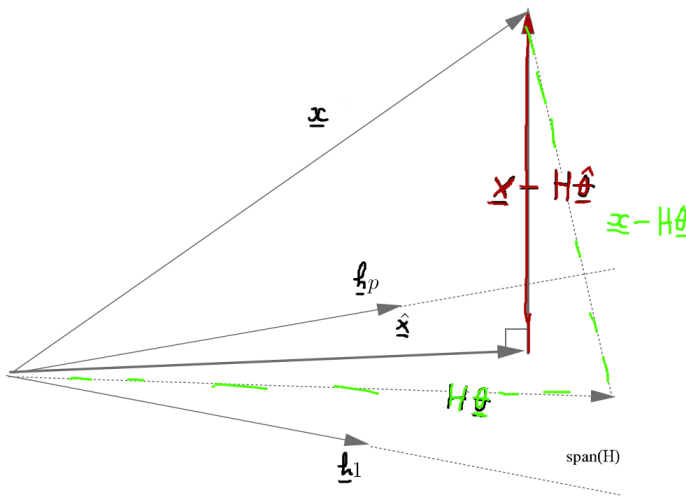


Figure 2: Least squares: A geometric interpretation.

EXAMPLE. Consider  $N = 3$  and  $p = 2$ . Obviously,  $(\mathbf{x} - \hat{\mathbf{x}}) \perp \{\mathbf{h}_1, \mathbf{h}_2\}$ . The *orthogonality principle* states that the minimum error is orthogonal to the columns of  $H$  (called regressors in statistics).

In general,

$$\mathbf{x} - \hat{\mathbf{x}} \perp \text{span}(H) \Leftrightarrow \mathbf{x} - \hat{\mathbf{x}} \perp \mathbf{h}_j, \forall \mathbf{h}_j \Leftrightarrow H^T(\mathbf{x} - H\hat{\boldsymbol{\theta}}) = \mathbf{0}.$$

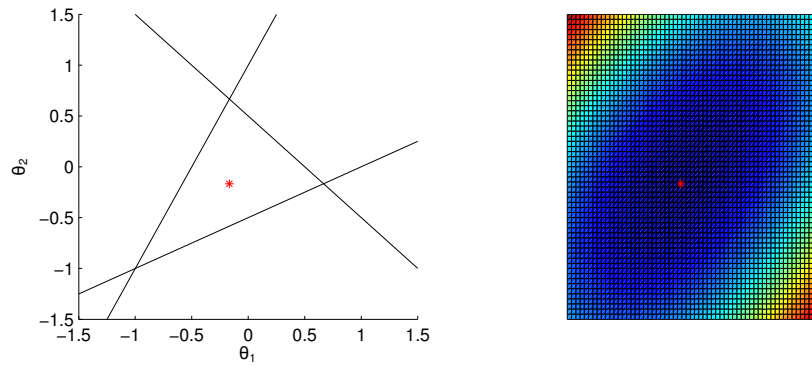
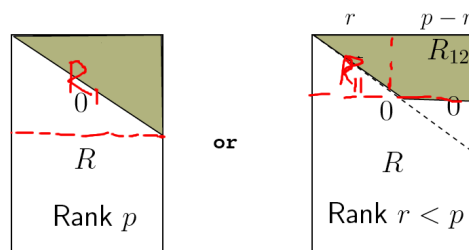


Figure 3: Least squares: An illustration.

### Computational Aspects

$$H_{N \times p} = Q_{N \times N} R_{N \times p}$$

$$= \begin{array}{|c|c|} \hline Q_1 & Q_2 \\ \hline \end{array} \begin{array}{|c|} \hline \begin{array}{c} 0 \\ R \end{array} \\ \hline \end{array}$$

 Figure 4: QR decomposition of a matrix  $H$ .

 Figure 5:  $R$  is upper triangular.

CONSIDER QR decomposition of a matrix  $H$ , illustrated in Fig. 4, where

- $Q$  has orthonormal columns:  $Q^T Q = I$  (and rows, i.e.  $Q Q^T = I$ ) and
- $R$  is upper triangular and may not have full rank, see Fig. 5. Then,  $H$  does not have full rank.

\* FULL-RANK  $H$ :

$$\begin{aligned}\|x - H\theta\|_2^2 &= \|Q^\top x - R\theta\|_2^2 \\ &= \|Q_1^\top x - R_1\theta\|_2^2 + \|Q_2^\top x\|_2^2\end{aligned}$$

which yields

$$\hat{\theta} = R_1^{-1} Q_1^\top x.$$

solve by backsubstitution from

$$R_1 \hat{\theta} = Q_1^\top x$$

\* COMMENTS:

- $Q^\top x$  yields coordinates of  $x$  on columns of  $Q$ .
- $\hat{x} = Q_1 Q_1^\top x = Px = H(H^\top H)^{-1} H^\top x$ . Here, the projection matrix  $P$  is also known as the *hat matrix* because it puts the “hat” on  $\theta$ .
- When  $H$  does not have full rank, i.e.,  $\text{rank } H = r < p$ , we need to solve

$$Q_{11}^\top x = R_{11}\theta_1 + R_{12}\theta_2$$

where  $Q_{11}$  has  $r$  columns; see Fig. 5(right). There are infinitely many solutions: to get one, arbitrarily set  $\theta_2 = \mathbf{0}_{(p-r) \times 1}$  and solve for  $\theta_1$ :

$$\theta_1 = R_{11}^{-1} Q_{11}^\top x.$$

solve by backsubstitution from

$$R_{11}\theta_1 = Q_{11}^\top x.$$

Here,

$$\hat{x} = Q_{11} Q_{11}^\top x$$

is still well defined and unique.

## Applications of Least Squares

DATA FITTING. Find best estimate of a signal given noisy measurements.

Machine learning, see Fig. 6.

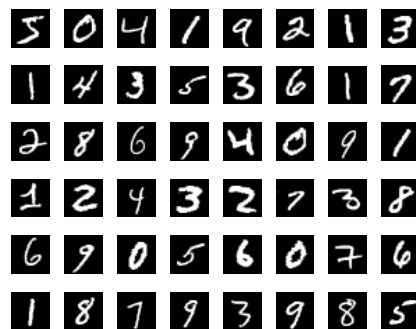


Figure 6: Character recognition.

\* EXAMPLES from (Boyd and Vandenberghe 2016, §13).

# Model

- ▶ choose *model*  $\hat{f} : \mathbf{R}^n \rightarrow \mathbf{R}$ , a *guess* or *approximation* of  $f$ , based on some observed data

$$(x_1, y_1), \dots, (x_N, y_N)$$

called *observations*, *examples*, *samples*, or *measurements*

- ▶ model form:

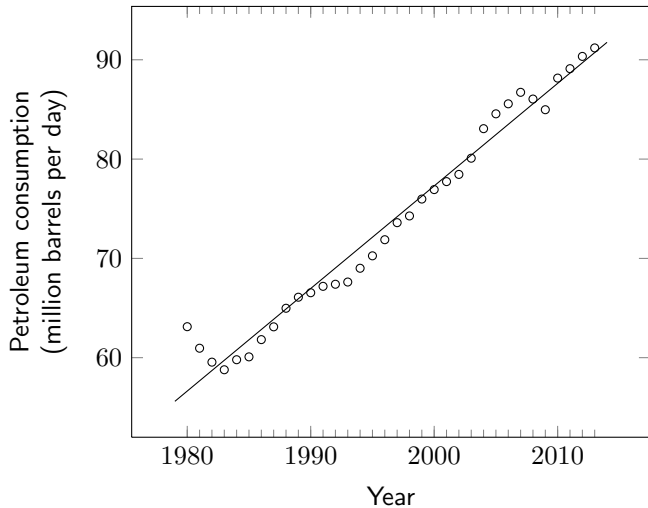
$$\hat{f}(x) = \theta_1 f_1(x) + \dots + \theta_p f_p(x)$$

- ▶  $f_i : \mathbf{R}^n \rightarrow \mathbf{R}$  are *basis functions* that we choose
- ▶  $\theta_i$  are *model parameters* that we choose
- ▶  $\hat{y}_i = \hat{f}(x_i)$  is (the model's) *prediction* of  $y_i$
- ▶ we'd like  $\hat{y}_i \approx y_i$ , i.e., model is consistent with observed data

## Time series trend

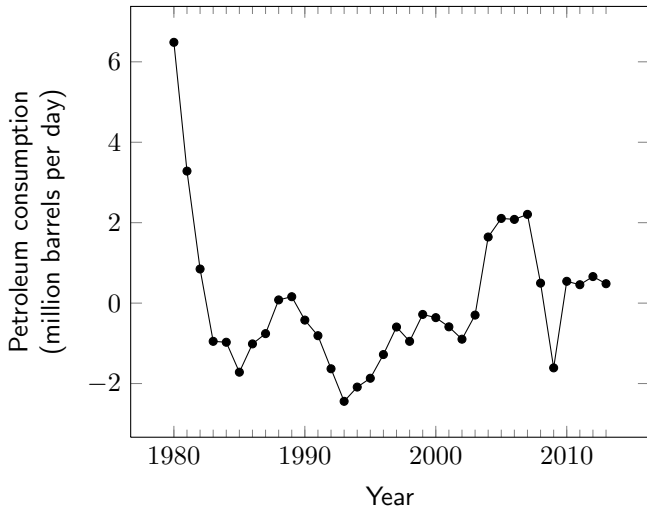
- ▶  $y_i$  is value of quantity at time  $x_i$
- ▶ common case:  $x_i = i$
- ▶  $\hat{y} = \hat{\theta}_1 + \hat{\theta}_2 x$  is called *trend line*
- ▶  $y - \hat{y}$  is called *de-trended time series*
- ▶  $\hat{\theta}_2$  is *trend coefficient*

## World petroleum consumption





## World petroleum consumption, de-trended



## Polynomial fit

- ▶  $f_i(x) = x^{i-1}$ ,  $i = 1, \dots, p$
- ▶ model is a polynomial of degree less than  $p$

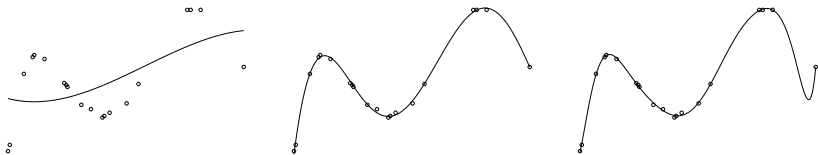
$$\hat{f}(x) = \theta_1 + \theta_2 x + \dots + \theta_p x^{p-1}$$

- ▶  $H$  is a Vandermonde matrix

$$H = \begin{bmatrix} 1 & x_1 & \dots & x_1^{p-1} \\ 1 & x_2 & \dots & x_2^{p-1} \\ \vdots & \vdots & & \vdots \\ 1 & x_N & \dots & x_N^{p-1} \end{bmatrix}$$

## Examples

- ▶  $N = 20$  data points
- ▶ fits of degree  $p - 1 = 3, 5$ , and 10



$H$  known up to  $\theta$

CONSIDER

$$\min_{\theta, s} \|x - H(\theta)s\|_2^2.$$

For a fixed  $\theta$ , we know

$$\begin{aligned}\hat{s}(\theta) &= [H^\top(\theta)H(\theta)]^{-1}H^\top(\theta)x \\ &= \arg \min_s \|x - H(\theta)s\|_2^2.\end{aligned}\tag{15}$$

Substitute this back into the LS criterion:

$$\begin{aligned}\min_{\theta, s} \|x - H(\theta)s\|_2^2 &= \min_{\theta} \|x - H(\theta)\hat{s}(\theta)\|_2^2 \\ &= \max_{\theta} x^\top P_H(\theta)x\end{aligned}\tag{19}$$

where

$$P_H(\theta) = H(\theta)[H^\top(\theta)H(\theta)]^{-1}H^\top(\theta).$$

Then, the LS estimates

$$(\hat{\theta}, \hat{s}) = \arg \min_{\theta, s} \|x - H(\theta)s\|_2^2$$

are given by

$$\hat{\theta} = \arg \max_{\theta} x^\top P_H(\theta)x\tag{20a}$$

$$\hat{s} = [H^\top(\hat{\theta})H(\hat{\theta})]^{-1}H^\top(\hat{\theta})x.\tag{20b}$$

Linearly constrained least squares

CONSIDER

$$\min \|C\theta - d\|_2^2 \quad \text{subject to } x = H\theta.\tag{21}$$

The equations  $x = H\theta$  determine an affine subspace. We wish to minimize  $\|C\theta - d\|_2^2$  on that subspace.

✱ **EXAMPLE:** If  $C = I$  and  $d = 0$ , solving linearly constrained least squares finds the minimum-norm solution of  $\theta = H\theta$ .

We can solve this problem by writing optimality conditions (setting gradients to zero and using Lagrange multipliers), but it is messy.

Easier to use QR decomposition instead. Consider the QR decomposition of the (tall) matrix  $H^\top$ :

$$H^\top = QR = \begin{bmatrix} Q_1 & Q_2 \end{bmatrix} \begin{bmatrix} R_1 \\ 0 \end{bmatrix}$$

and define

$$\beta = Q^\top \theta$$

partitioned as

$$\begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} Q_1^\top \\ Q_2^\top \end{bmatrix} \theta = \begin{bmatrix} Q_1^\top \theta \\ Q_2^\top \theta \end{bmatrix}.$$

Note that

$$\theta = Q\beta = Q_1\beta_1 + Q_2\beta_2.$$

Now, the equation  $\theta = H\theta$  becomes

$$\theta = R^\top \underbrace{Q^\top \theta}_{\beta} = R_1^\top \beta_1. \quad (22)$$

$\beta_2$  does not appear in the constraint.

Since  $R_1^\top$  is invertible, we can solve (22) to obtain

$$\beta_1 = R_1^{-T} \theta.$$

$$R_1^{-T} = (R_1^\top)^{-1}$$

The objective function to be minimized is

$$\begin{aligned} \|C\theta - d\|_2^2 &= \|CQ_1\beta_1 + CQ_2\beta_2 - d\|_2^2 \\ &= \|\underbrace{CQ_2}_{\tilde{H}} \beta_2 - \underbrace{(d - CQ_1\beta_1)}_{\tilde{d}}\|_2^2 \end{aligned}$$

which is an *unconstrained LS problem* in  $\beta_2$ .

Solve for  $\beta_1$  first, then for  $\beta_2$ , then compute

$$\theta = Q_1\beta_1 + Q_2\beta_2.$$

## Underdetermined systems

MORE unknowns than equations.

$$\theta = H\theta$$

with  $H \in \mathbb{R}^{N \times p}$ ,  $\theta \in \mathbb{R}^{N \times 1}$ ,  $\theta \in \mathbb{R}^{p \times 1}$ .

The matrix  $H$  is fat (i.e.,  $N < p$ ). We assume the rows of  $H$  are linearly independent (i.e.,  $H$  has full rank equal to  $N$ ).

Many solutions, a whole subspace. We wish to make the solution *unique*. One way to make the solution unique: linearly constrained least squares.

☞ RECENTLY, solving  $\min \|\theta\|$  subject to  $x = H\theta$  has attracted significant attention for cases where  $\|\theta\|$  is not  $\ell_2$ .

\* EXAMPLE: Consider

$$x = \frac{1}{5} \begin{bmatrix} 1 & 2 \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} \quad (23)$$

a line with slope  $-0.5$  in the  $(\theta_1, \theta_2)$  plane, see Fig. 8.

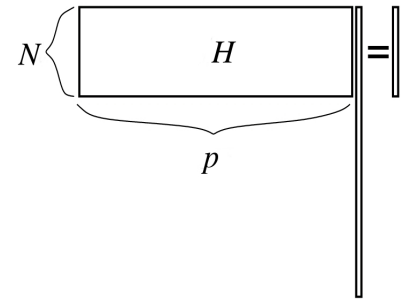


Figure 7: Underdetermined system.

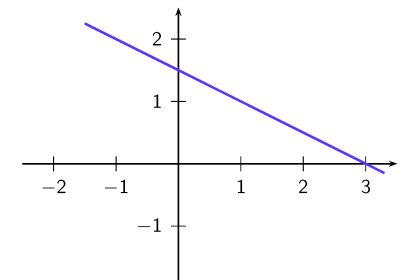


Figure 8: Set of  $\theta$  satisfying (23).

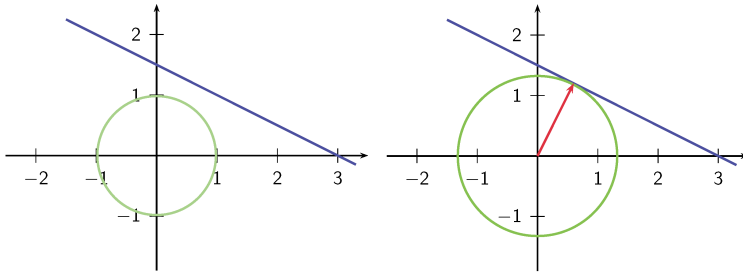


Figure 9: Keep expanding the  $\ell_2$  ball until we hit the line  $\mathbf{x} = \mathbf{H}\boldsymbol{\theta}$ .

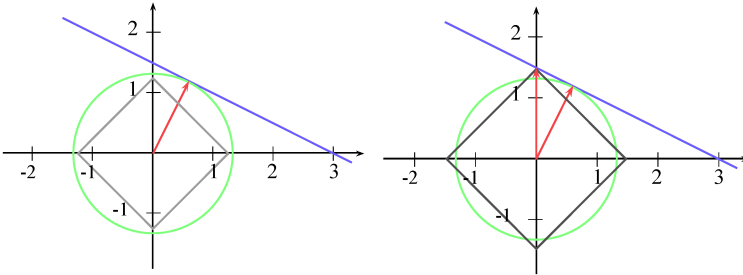


Figure 10: Keep expanding the  $\ell_1$  ball until we hit the line  $\mathbf{x} = \mathbf{H}\boldsymbol{\theta}$ .

Suppose we wish to find  $\boldsymbol{\theta} = [\theta_1, \theta_2]^\top$  that satisfies the above equation and has smallest  $\ell_1$  or  $\ell_2$  norm, see Figs. 9 and 10.

The solution to

$$\min \|\boldsymbol{\theta}\|_1 \quad \text{subject to } \mathbf{x} = \mathbf{H}\boldsymbol{\theta} \quad (24)$$

is sparse!

\* TOMOGRAPHY using regularized LS: (Boyd and Vandenberghe 2016, §15.3.4).

## Acronyms

*BLUE* best linear unbiased estimator. 6–8

*CRB* Cramér-Rao bound. 3–6

*CT* continuous-time. 1

*FIM* Fisher information matrix. 4

*LS* least-squares. 5, 9, 10, 20, 21, 23

*MVU* minimum-variance unbiased. 2–5, 8

*pdf* probability density function. 4

## References

- Boyd, Stephen and Lieven Vandenberghe (2016). *Vectors, Matrices, and Least Squares*. URL: <http://stanford.edu/class/ee103/mma.pdf> (visited on 08/23/2016) (cit. on pp. 13, 22).
- Hero, Alfred O. (2015). *Statistical Methods for Signal Processing*. Lecture notes. Univ. Michigan, Ann Arbor, MI (cit. on p. 1).