Lecture 4: Least Squares (LS) Estimation

Background and general solution
Solution in the Gaussian case
Properties
Example

Big picture

general least squares estimation:

- given: jointly distributed x (n-dimensional) & y (m-dimensional)
- goal: find the optimal estimate $\hat{x}$ that minimizes

$$E\left[||x-\hat{x}||^2 \middle| y=y_1\right] = E\left[(x-\hat{x})^T(x-\hat{x}) \middle| y=y_1\right]$$

- *solution*: consider

$$J(z) = E\left[||x-z||^2 \middle| y=y_1\right] = E\left[x^T x \middle| y=y_1\right] - 2z^T E[x|y=y_1] + z^T z$$

which is quadratic in z. For optimal cost,

$$\frac{\partial}{\partial z} J(z) = 0 \Rightarrow z = E[x|y=y_1] \triangleq \hat{x}$$

hence

$$\boxed{\hat{x} = E[x|y=y_1] = \int_{-\infty}^{\infty} x p_{x|y}(x|y_1)\,dx}$$

$$\boxed{J_{\min} = J(\hat{x}) = \mathrm{Tr}\left\{E\left[(x-\hat{x})(x-\hat{x})^T |y=y_1\right]\right\}}$$

Big picture

general least squares estimation:

$$\boxed{\hat{x} = \mathrm{E}[x|y = y_1] = \int_{-\infty}^{\infty} x p_{x|y}(x|y_1)\,\mathrm{d}x}$$

achieves the minimization of

$$\mathrm{E}\left[||x - \hat{x}||^2 \middle| y = y_1 \right]$$

solution concepts:

- the solution holds for any probability distribution in y
- for each y_1, $\mathrm{E}[x|y = y_1]$ is different
- if no specific value of y is given, $\hat{x}$ is a function of the random vector/variable y, written as

$$\hat{x} = \mathrm{E}[x|y]$$

Least square estimation in the Gaussian case

Why Gaussian?

- Gaussian is common in practice:
  - macroscopic random phenomena = superposition of microscopic random effects (Central limit theorem)
- Gaussian distribution has nice properties that make it mathematically feasible to solve many practical problems:
  - pdf is solely determined by the mean and the variance/covariance
  - linear functions of a Gaussian random process are still Gaussian
  - the output of an LTI system is a Gaussian random process if the input is Gaussian
  - if two jointly Gaussian distributed random variables are uncorrelated, then they are independent
  - X_1 and X_2 jointly Gaussian $\Rightarrow X_1|X_2$ and $X_2|X_1$ are also Gaussian

Least square estimation in the Gaussian case

Why Gaussian?

Gaussian and white:

- they are different concepts
- there can be Gaussian white noise, Poisson white noise, etc
- Gaussian white noise is used a lot since it is a good approximation to many practical noises

Least square estimation in the Gaussian case

the solution

problem (re-stated): x, y–Gaussian distributed

$$\text{minimize } \mathrm{E}\left[\left|\left|x-\hat{x}\right|\right|^2 \middle| y\right]$$

solution: $\boxed{\hat{x}=\mathrm{E}\left[x|y\right]=\mathrm{E}\left[x\right]+X_{xy}X_{yy}^{-1}\left(y-\mathrm{E}\left[y\right]\right)}$

properties:

- the estimation is unbiased: $\boxed{\mathrm{E}\left[\hat{x}\right]=\mathrm{E}\left[x\right]}$
- y is Gaussian$\Rightarrow\hat{x}$ is Gaussian; and $x-\hat{x}$ is also Gaussian
- covariance of $\hat{x}$:

$$\mathrm{E}\left[\left(\hat{x}-\mathrm{E}\left[\hat{x}\right]\right)\left(\hat{x}-\mathrm{E}\left[\hat{x}\right]\right)^T\right]=\mathrm{E}\left\{\left(y-\mathrm{E}\left[y\right]\right)\left[X_{xy}X_{yy}^{-1}\left(y-\mathrm{E}\left[y\right]\right)\right]^T\right\}=X_{xy}X_{yy}^{-1}X_{yx}$$

- estimation error $\tilde{x}\triangleq x-\hat{x}$: zero mean and

$$\mathrm{Cov}\left[\tilde{x}\right]=\underbrace{\mathrm{E}\left[\left(x-\mathrm{E}\left[x|y\right]\right)\left(x-\mathrm{E}\left[x|y\right]\right)^T\right]}_{\text{conditional covariance}}=X_{xx}-X_{xy}X_{yy}^{-1}X_{yx}$$

Least square estimation in the Gaussian case

$$\boxed{\hat{x} = \mathrm{E}[x|y] = \mathrm{E}[x] + X_{xy}X_{yy}^{-1}(y - \mathrm{E}[y])}$$

$\mathrm{E}[x|y]$ is a better estimate than $\mathrm{E}[x]$:

- the estimation is unbiased: $\mathrm{E}[\hat{x}] = \mathrm{E}[x]$
- estimation error $\tilde{x} \triangleq x - \hat{x}$: zero mean and

$$\mathrm{Cov}[x - \hat{x}] = X_{xx} - X_{xy}X_{yy}^{-1}X_{yx} \preceq \mathrm{Cov}[x - \mathrm{E}[X]]$$

Properties of least square estimate (Gaussian case)

two random vectors x and y

Property 1:
(i) the estimation error $\tilde{x} = x - \hat{x}$ is uncorrelated with y
(ii) $\tilde{x}$ and $\hat{x}$ are orthogonal:

$$\mathrm{E}\left[(x - \hat{x})^T \hat{x}\right] = 0$$

proof of (i):

$$\begin{aligned}\mathrm{E}\left[\tilde{x}(y - m_y)^T\right] &= \mathrm{E}\left[\left(x - \mathrm{E}[x] - X_{xy}X_{yy}^{-1}(y - m_y)\right)(y - m_y)^T\right] \\ &= X_{xy} - X_{xy}X_{yy}^{-1}X_{yy} = 0\end{aligned}$$

Properties of least square estimate (Gaussian case)

two random vectors x and y

proof of (ii): $\mathrm{E}\left[\tilde{x}^T\hat{x}\right] = \mathrm{E}\left[(x-\hat{x})^T\left(\mathrm{E}[x] + X_{xy}X_{yy}^{-1}(y-m_y)\right)\right] =$
$\mathrm{E}\left[\tilde{x}^T\right]\mathrm{E}[x] + \mathrm{E}\left[(x-\hat{x})^T X_{xy}X_{yy}^{-1}(y-m_y)\right]$ where $\mathrm{E}\left[\tilde{x}^T\right] = 0$ and

$$\begin{aligned}\mathrm{E}\left[(x-\hat{x})^T X_{xy}X_{yy}^{-1}(y-m_y)\right] &= \mathrm{Tr}\left\{\mathrm{E}\left[X_{xy}X_{yy}^{-1}(y-m_y)(x-\hat{x})^T\right]\right\} \\ &= \mathrm{Tr}\left\{X_{xy}X_{yy}^{-1}\mathrm{E}\left[(y-m_y)(x-\hat{x})^T\right]\right\} = 0 \text{ because of (i)}\end{aligned}$$

- note: $\mathrm{Tr}\{BA\} = \mathrm{Tr}\{AB\}$. Consider, e.g. $A = [a, b]$, $B = \begin{bmatrix} c \\ d \end{bmatrix}$

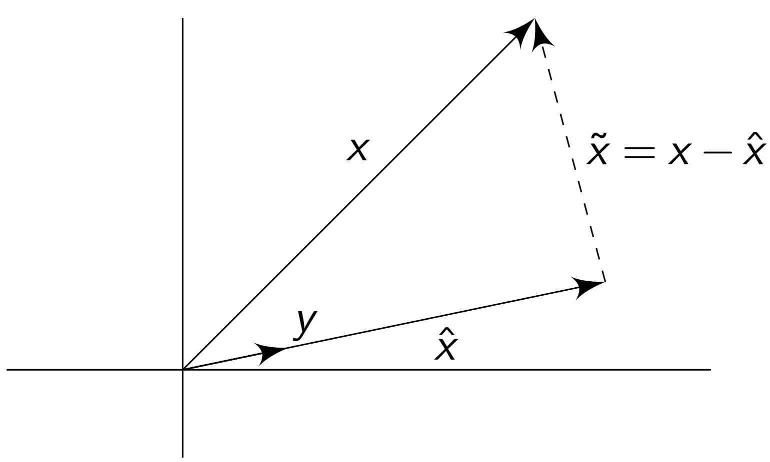
Properties of least square estimate (Gaussian case)

two random vectors x and y

Property 1 (re-stated):
(i) the estimation error $\tilde{x} = x - \hat{x}$ is uncorrelated with y
(ii) $\tilde{x}$ and $\hat{x}$ are orthogonal:

$$\mathrm{E}\left[(x-\hat{x})^T\hat{x}\right] = 0$$

- intuition: least square estimation is a projection



Properties of least square estimate (Gaussian case)

three random vectors x y and z, where y and z are uncorrelated

Property 2: let y and z be Gaussian and uncorrelated, then
(i) the optimal estimate of x is

$$\begin{aligned} \mathrm{E}[x|y,z] &= \mathrm{E}[x] + \overbrace{(\mathrm{E}[x|y] - \mathrm{E}[x])}^{\text{first improvement}} + \overbrace{(\mathrm{E}[x|z] - \mathrm{E}[x])}^{\text{second improvement}} \\ &= \mathrm{E}[x|y] + (\mathrm{E}[x|z] - \mathrm{E}[x]) \end{aligned}$$

Alternatively, let $\boxed{\hat{x}_{|y} \triangleq \mathrm{E}[x|y]}$, $\boxed{\tilde{x}_{|y} \triangleq x - \mathrm{E}[x|y] = x - \hat{x}_{|y}}$, then

$$\boxed{\mathrm{E}[x|y,z] = \mathrm{E}[x|y] + \mathrm{E}\left[\tilde{x}_{|y}|z\right]}$$

(ii) the estimation error covariance is

$$X_{xx} - X_{xy}X_{yy}^{-1}X_{yx} - X_{xz}X_{zz}^{-1}X_{zx} = X_{\tilde{x}\tilde{x}} - X_{xz}X_{zz}^{-1}X_{zx} = \underline{X_{\tilde{x}\tilde{x}} - X_{\tilde{x}z}X_{zz}^{-1}X_{z\tilde{x}}}$$

where $X_{\tilde{x}\tilde{x}} = \mathrm{E}\left[\tilde{x}_{|y}\tilde{x}_{|y}^T\right]$ and $X_{\tilde{x}z} = \mathrm{E}\left[\tilde{x}_{|y}(z - m_z)^T\right]$

Properties of least square estimate (Gaussian case)

three random vectors x y and z, where y and z are uncorrelated

proof of (i): let $w = [y, z]^T$

$$\mathrm{E}[x|w] = \mathrm{E}[x] + \begin{bmatrix} X_{xy} & X_{xz} \end{bmatrix} \begin{bmatrix} X_{yy} & X_{yz} \\ X_{zy} & X_{zz} \end{bmatrix}^{-1} \begin{bmatrix} y - \mathrm{E}[y] \\ z - \mathrm{E}[z] \end{bmatrix}$$

Using $X_{yz} = 0$ yields

$$\begin{aligned} \mathrm{E}[x|w] &= \mathrm{E}[x] + \underbrace{X_{xy}X_{yy}^{-1}(y - \mathrm{E}[y])}_{\mathrm{E}[x|y] - \mathrm{E}[x]} + \underbrace{X_{xz}X_{zz}^{-1}(z - \mathrm{E}[z])}_{\mathrm{E}[x|z] - \mathrm{E}[x]} \\ &= \mathrm{E}[x|y] + \mathrm{E}\left[\left(\hat{x}_{|y} + \tilde{x}_{|y}\right)|z\right] - \mathrm{E}[x] \\ &= \mathrm{E}[x|y] + \mathrm{E}\left[\tilde{x}_{|y}|z\right] \end{aligned}$$

where $\mathrm{E}\left[\hat{x}_{|y}|z\right] = \mathrm{E}[\mathrm{E}[x|y]|z] = \mathrm{E}[x]$ as y and z are independent

Properties of least square estimate (Gaussian case)

three random vectors x y and z, where y and z are uncorrelated

proof of (ii): let $w = [y, z]^T$, the estimation error covariance is

$$X_{xx} - X_{xw}X_{ww}^{-1}X_{wx} = X_{xx} - X_{xy}X_{yy}^{-1}X_{yx} - X_{xz}X_{zz}^{-1}X_{zx}$$

additionally

$$\begin{aligned} X_{xz} &= \mathrm{E}\left[(\underline{x} - \mathrm{E}[x])(z - \mathrm{E}[z])^T\right] = \mathrm{E}\left[\left(\underline{\hat{x}_{|y} + \tilde{x}_{|y}} - \mathrm{E}[x]\right)(z - \mathrm{E}[z])^T\right] \\ &= \mathrm{E}\left[\left(\hat{x}_{|y} - \mathrm{E}[x]\right)(z - \mathrm{E}[z])^T\right] + \mathrm{E}\left[\tilde{x}_{|y}(z - \mathrm{E}[z])^T\right] \end{aligned}$$

but $\hat{x}_{|y} - \mathrm{E}[x]$ is a linear function of y, which is uncorrelated with z, hence $\mathrm{E}\left[\left(\hat{x}_{|y} - \mathrm{E}[x]\right)(z - \mathrm{E}[z])^T\right] = 0$ and $X_{xz} = X_{\tilde{x}_{|y}z}$

Properties of least square estimate (Gaussian case)

three random vectors x y and z, where y and z are uncorrelated

Property 2 (re-stated): let y and z be Gaussian and uncorrelated

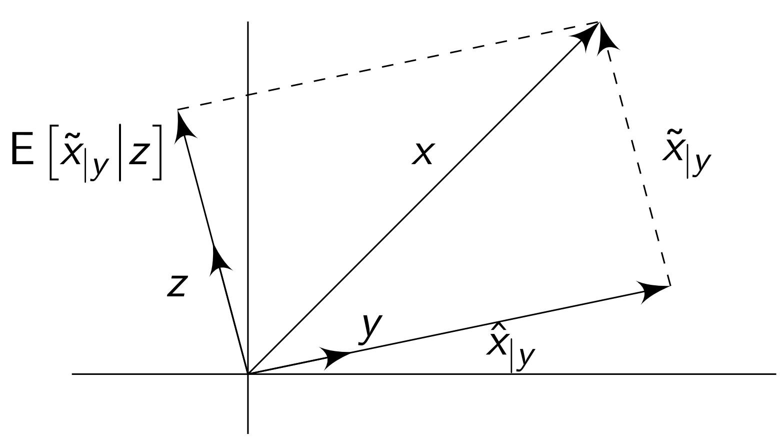
(i) the optimal estimate of x is

$$\mathrm{E}[x|y,z] = \mathrm{E}[x|y] + \mathrm{E}\left[\tilde{x}_{|y}|z\right]$$

(ii) the estimation error covariance is

$$X_{\tilde{x}\tilde{x}} - X_{\tilde{x}z}X_{zz}^{-1}X_{z\tilde{x}}$$

- intuition:



Properties of least square estimate (Gaussian case)

three random vectors x y and z, where y and z are correlated

Property 3: let y and z be Gaussian and correlated, then

(i) the optimal estimate of x is

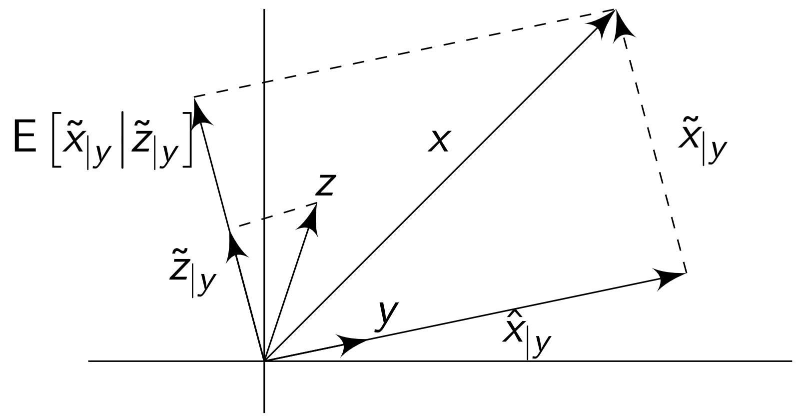
$$\mathrm{E}[x|y,z] = \mathrm{E}[x|y] + \mathrm{E}\left[\tilde{x}_{|y}|\tilde{z}_{|y}\right]$$

where $\tilde{z}_{|y} = z - \hat{z}_{|y} = z - \mathrm{E}[z|y]$ and $\tilde{x}_{|y} = x - \hat{x}_{|y} = x - \mathrm{E}[x|y]$

(ii) the estimation error covariance is

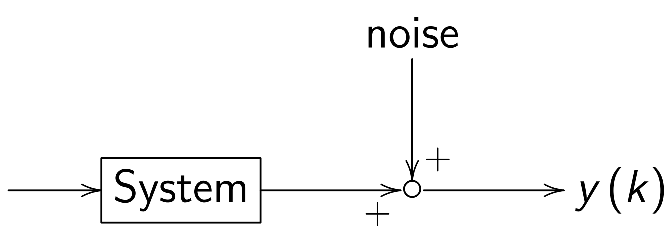
$$X_{\tilde{x}_{|y}\tilde{x}_{|y}} - X_{\tilde{x}_{|y}\tilde{z}_{|y}} X^{-1}_{\tilde{z}_{|y}\tilde{z}_{|y}} X_{\tilde{z}_{|y}\tilde{x}_{|y}}$$

- intuition:



Application of the three properties

Consider



Given $[y(0), y(1), \ldots, y(k)]^T$, we want to estimate the state $x(k)$

- the properties give a recursive way to compute

$$\hat{x}(k) | \{y(0), y(1), \ldots, y(k)\}$$

Example

Consider estimating the velocity x of a motor, with

$$\begin{aligned} \mathrm{E}[x] &= m_x = 10 \text{ rad/s} \\ \mathrm{Var}[x] &= 2 \text{ rad}^2/\text{s}^2 \end{aligned}$$

There are two (tachometer) sensors available:

- $y_1 = x + v_1$: $\mathrm{E}[v_1] = 0$, $\mathrm{E}\left[v_1^2\right] = 1 \text{ rad}^2/\text{s}^2$
- $y_2 = x + v_2$: $\mathrm{E}[v_2] = 0$, $\mathrm{E}\left[v_2^2\right] = 1 \text{ rad}^2/\text{s}^2$

where v_1 and v_2 are independent, Gaussian, $\mathrm{E}[v_1 v_2] = 0$ and x is independent of v_i, $\mathrm{E}[(x - \mathrm{E}[x])\, v_i] = 0$

Example

- best estimate of x using only y_1:

$$\begin{aligned} X_{xy_1} &= \mathrm{E}[(x - m_x)(y_1 - m_{y_1})] = \mathrm{E}[(x - m_x)(x - m_x + v_1)] \\ &= X_{xx} + \mathrm{E}[(x - m_x)\, v_1] = 2 \end{aligned}$$

$$\begin{aligned} X_{y_1 y_1} &= \mathrm{E}[(y_1 - m_{y_1})(y_1 - m_{y_1})] = \mathrm{E}[(x - m_x + v_1)(x - m_x + v_1)] \\ &= X_{xx} + \mathrm{E}\left[v_1^2\right] = 3 \end{aligned}$$

$$\hat{x}_{|y_1} = \mathrm{E}[x] + X_{xy_1} X_{y_1 y_1}^{-1} (y_1 - \mathrm{E}[y_1]) = 10 + \frac{2}{3}(y_1 - 10)$$

- similarly, best estimate of x using only y_2: $\hat{x}_{|y_2} = 10 + \frac{2}{3}(y_2 - 10)$

Example

- best estimate of x using y_1 and y_2 (direct approach): let $y = [y_1, y_2]^T$

$$X_{xy} = \mathrm{E}\left[(x - m_x)\begin{bmatrix} y_1 - m_{y_1} \\ y_2 - m_{y_2} \end{bmatrix}^T\right] = [2, 2]$$

$$X_{yy} = \mathrm{E}\left[\begin{bmatrix} y_1 - m_{y_1} \\ y_2 - m_{y_2} \end{bmatrix}\begin{bmatrix} y_1 - m_{y_1} & y_2 - m_{y_2} \end{bmatrix}\right] = \begin{bmatrix} 3 & 2 \\ 2 & 3 \end{bmatrix}$$

$$\hat{x}_{|y} = \mathrm{E}[x] + X_{xy} X_{yy}^{-1}(y - m_y) = 10 + [2, 2]\begin{bmatrix} 3 & 2 \\ 2 & 3 \end{bmatrix}^{-1}\begin{bmatrix} y_1 - 10 \\ y_2 - 10 \end{bmatrix}$$

- note: X_{yy}^{-1} is expensive to compute at high dimensions

Example

- best estimate of x using y_1 and y_2 (alternative approach using Property 3):

$$\mathrm{E}[x|y_1, y_2] = \mathrm{E}[x|y_1] + \mathrm{E}\left[\tilde{x}_{|y_1}|\tilde{y}_{2|y_1}\right]$$

which involves just the scalar computations:

$$\mathrm{E}[x|y_1] = 10 + \frac{2}{3}(y_1 - 10),\ \tilde{x}_{|y_1} = x - \mathrm{E}[x|y_1] = \frac{1}{3}(x - 10) + \frac{2}{3}v_1$$

$$\tilde{y}_{2|y_1} = y_2 - \mathrm{E}[y_2|y_1] = y_2 - \left[\mathrm{E}[y_2] + X_{y_2 y_1}\frac{1}{X_{y_1 y_1}}(y_1 - m_{y_1})\right] = (y_2 - 10) - \frac{2}{3}(y_1 - 10)$$

$$X_{\tilde{x}_{|y_1}\tilde{y}_{2|y_1}} = \mathrm{E}\left[\left(\frac{1}{3}(x - 10) + \frac{2}{3}v_1\right)\left((y_2 - 10) - \frac{2}{3}(y_1 - 10)\right)^T\right] = \frac{1}{9}\mathrm{Var}[x] + \frac{4}{9}\mathrm{Var}[v_1] = \frac{2}{3}$$

$$X_{\tilde{y}_{2|y_1}\tilde{y}_{2|y_1}} = \frac{1}{9}\mathrm{Var}[x] + \mathrm{Var}[v_2] + \frac{4}{9}\mathrm{Var}[v_1] = \frac{5}{3}$$

$$\mathrm{E}\left[\tilde{x}_{|y_1}|\tilde{y}_{2|y_1}\right] = \mathrm{E}\left[\tilde{x}_{|y_1}\right] + X_{\tilde{x}_{|y_1}\tilde{y}_{2|y_1}}\frac{1}{X_{\tilde{y}_{2|y_1}\tilde{y}_{2|y_1}}}\left[\tilde{y}_{2|y_1} - \mathrm{E}\left[\tilde{y}_{2|y_1}\right]\right]$$

$$= 10 + \frac{2}{5}(y_1 - 10) + \frac{2}{5}(y_2 - 10)$$

Summary

1. Big picture
   $\hat{x} = \mathrm{E}[x|y]$ minimizes $J = \mathrm{E}\left[||x-\hat{x}||^2 \middle| y\right]$

2. Solution in the Gaussian case
   Why Gaussian?
   $\hat{x} = \mathrm{E}[x|y] = \mathrm{E}[x] + X_{xy}X_{yy}^{-1}(y - \mathrm{E}[y])$

3. Properties of least square estimate (Gaussian case)
   two random vectors x and y
   three random vectors x y and z: y and z are uncorrelated
   three random vectors x y and z: y and z are correlated

* Appendix: trace of a matrix

- the trace of a $n \times n$ matrix is given by $\mathrm{Tr}(A) = \sum_{i=1}^{n} a_{ii}$
- trace is the matrix inner product:

$$\langle A, B\rangle = \mathrm{Tr}\left(A^T B\right) = \mathrm{Tr}\left(B^T A\right) = \langle B, A\rangle \tag{1}$$

- take a three-column example: write the matrices in the column vector form $B = [b_1, b_2, b_3]$, $A = [a_1, a_2, a_3]$, then,

$$A^T B = \begin{bmatrix} a_1^T b_1 & * & * \\ * & a_2^T b_2 & * \\ * & * & a_3^T b_3 \end{bmatrix} \tag{2}$$

$$\mathrm{Tr}\left(A^T B\right) = a_1^T b_1 + a_2^T b_2 + a_3^T b_3 = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix}^T \cdot \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} \tag{3}$$

  which is the inner product of the two long stacked vectors.
- we frequently use the inner-product equality $\langle A, B\rangle = \langle B, A\rangle$