01_Least Squares

1. Squared Error Criterion and the Method of Least Squares

Squared Error Criterion

$$\mathscr{L}(x)=\mathbf{e}^T\mathbf{e}=(\mathbf{y}-\mathbf{H}x)^T(\mathbf{y}-\mathbf{H}\mathbf{x})=\mathbf{y}^T\mathbf{y}-x^T\mathbf{H}^T\mathbf{y}-\mathbf{y}^T\mathbf{H}x+x^T\mathbf{H}^T\mathbf{H}x$$

where:

- $\mathbf{y}$: measurement
- x: True value (sometimes we call **parameter** that need to be estimated)
- $\mathbf{H}$: model parameter (something that we already know)

Least Square Solution

$$\hat{x}_{\text{LS}}=\left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{y}$$

Proof:

To minimize this, we can compute the partial derivative with respect to our parameter, set to 0, and solve for an extremum

$$\begin{aligned}\frac{\partial\mathscr{L}}{\partial x}\Big|_{x=\hat{x}}&=-\mathbf{y}^T\mathbf{H}-\mathbf{y}^T\mathbf{H}+2\hat{x}^T\mathbf{H}^T\mathbf{H}=0\\&-2\mathbf{y}^T\mathbf{H}+2\hat{x}^T\mathbf{H}^T\mathbf{H}=0\end{aligned}$$

Note:

It can only be solved when $\mathbf{H}^T\mathbf{H}$ is invertible. Which means measurements are no less than unknown parameters.

Weighted Least Square

Sometimes the measurements comes from **different measurement equipmen**t and they have different accuracy. We need find ways to paid different trust on different measurement data.

We use the **variance** of different equipment to estimate the noise the equipment may take.

$$\mathbb{E}\left[v_i^2\right] = \sigma_i^2, \quad (i = 1, \ldots, m) \quad \mathbf{R} = \mathbb{E}\left[\mathbf{v}\mathbf{v}^T\right] = \begin{bmatrix} \sigma_1^2 & 0 & \\ & \ddots & \\ 0 & \sigma_m^2 & \end{bmatrix}$$

Weighted Least Square Criterion

Then we will define our criterion as:

$$\begin{aligned} \mathscr{L}_{\text{WLS}}(\mathbf{x}) &= \mathbf{e}^T\mathbf{R}^{-1}\mathbf{e} \\ &= \frac{e_1^2}{\sigma_1^2} + \frac{e_2^2}{\sigma_2^2} + \ldots + \frac{e_m^2}{\sigma_m^2} \end{aligned}$$

This is rationale, because by use $\mathbf{R}^{-1}$, **The higher the expected noise, the lower the weight we place on the measurement**

Weighted Least Square Solution

Then we can follow the same process of the normal least square one, then we will get the solution:

$$\hat{\mathbf{x}} = \left(\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{R}^{-1}\mathbf{y}$$

2. Recursive Least Squares

By Least Square Method, we can estimate parameter based on data. However, in real life, the data we obtained are always stream data.

In order to analyze stream data, we want to compute $\hat{\mathbf{x}_\mathbf{k}}$ as a function of $\mathbf{y}_\mathbf{k}$ and $\hat{\mathbf{x}}_{k-1}$.

One solution is to use **<u>linear recursive update</u>**

$$\hat{\mathbf{x}_\mathbf{k}} = \hat{\mathbf{x}}_{k-1} + \mathbf{K}_k(\mathbf{y}_k - \mathbf{H}_k\hat{\mathbf{x}}_{k-1})$$

The $\mathbf{K}$ can be updated by:

$$\mathbf{K}_k = \mathbf{P}_{k-1}\mathbf{H}_k^T\left(\mathbf{H}_k\mathbf{P}_{k-1}\mathbf{H}_k^T + \mathbf{R}_k\right)^{-1}$$

where $\mathbf{P}_k$ can be updated by:

$$\mathbf{P}_k = (\mathbf{1} - \mathbf{K}_k\mathbf{H}_k)\mathbf{P}_{k-1}(\mathbf{1} - \mathbf{K}_k\mathbf{H}_k)^T + \mathbf{K}_k\mathbf{R}_k\mathbf{K}_k^T$$

Proof:

We always wish to **minimize the expected value of the sum of squared errors** of our current estimate at time step k, that is:

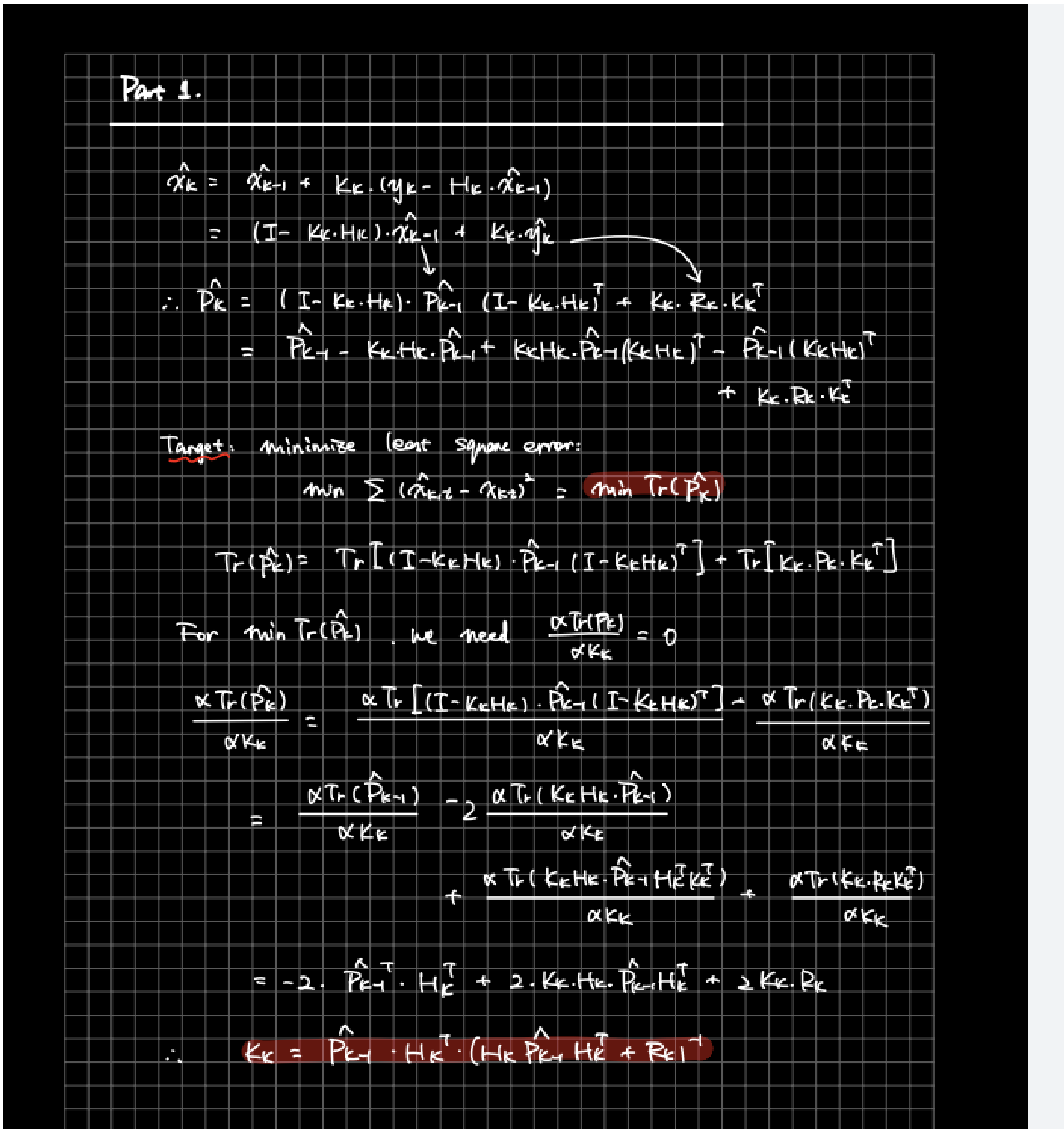
$$\begin{aligned} \min \mathscr{L}_{\text{RLS}} &= \min \mathbb{E}\left[\left(x_{1k} - \hat{x}_{1k}\right)^2 + \ldots + \left(x_{nk} - \hat{x}_{nk}\right)^2\right] \\ &= \min \operatorname{Trace}\left(\mathbf{P}_k\right) \end{aligned}$$

Property

By using the recursive updating of $\mathbf{K}$, we can simplify $\mathbf{P}_k$ to:

$$\begin{aligned} \mathbf{P}_k &= \mathbf{P}_{k-1} - \mathbf{K}_k \mathbf{H}_k \mathbf{P}_{k-1} \\ &= \left(\mathbf{1} - \mathbf{K}_k \mathbf{H}_k\right) \mathbf{P}_{k-1} \end{aligned}$$

That is our covariance **shrink** with each measurement



Part 1.

$$\hat{x}_k = \hat{x}_{k-1} + K_k \cdot (y_k - H_k \cdot \hat{x}_{k-1})$$

$$= (I - K_k \cdot H_k) \cdot \hat{x}_{k-1} + K_k \cdot \hat{y}_k$$

$$\therefore \hat{P}_k = (I - K_k \cdot H_k) \cdot \hat{P}_{k-1} (I - K_k \cdot H_k)^T + K_k \cdot R_k \cdot K_k^T$$

$$= \hat{P}_{k-1} - K_k H_k \cdot \hat{P}_{k-1} + K_k H_k \cdot \hat{P}_{k-1} (K_k H_k)^T - \hat{P}_{k-1} (K_k H_k)^T + K_k \cdot R_k \cdot K_k^T$$

Target: minimize least square error:

$$\min \sum (\hat{x}_{k,t} - x_{k,t})^2 = \min Tr(\hat{P}_k)$$

$$Tr(\hat{P}_k) = Tr\left[(I - K_k H_k) \cdot \hat{P}_{k-1} (I - K_k H_k)^T\right] + Tr\left[K_k \cdot P_k \cdot K_k^T\right]$$

For $\min Tr(\hat{P}_k)$, we need $\frac{\partial Tr(P_k)}{\partial K_k} = 0$

$$\frac{\partial Tr(\hat{P}_k)}{\partial K_k} = \frac{\partial Tr\left[(I - K_k H_k) \cdot \hat{P}_{k-1} (I - K_k H_k)^T\right]}{\partial K_k} + \frac{\partial Tr(K_k \cdot P_k \cdot K_k^T)}{\partial K_k}$$

$$= \frac{\partial Tr(\hat{P}_{k-1})}{\partial K_k} - 2 \frac{\partial Tr(K_k H_k \cdot \hat{P}_{k-1})}{\partial K_k} + \frac{\partial Tr(K_k H_k \cdot \hat{P}_{k-1} H_k^T K_k^T)}{\partial K_k} + \frac{\partial Tr(K_k \cdot R_k K_k^T)}{\partial K_k}$$

$$= -2 \cdot \hat{P}_{k-1}^T \cdot H_k^T + 2 \cdot K_k \cdot H_k \cdot \hat{P}_{k-1} H_k^T + 2 K_k \cdot R_k$$

$$\therefore K_k = \hat{P}_{k-1} \cdot H_k^T \cdot (H_k \hat{P}_{k-1} H_k^T + R_k)^{-1}$$

https://s3-us-west-2.amazonaws.com/secure.notion-static.com/e698a26c-35c2-47da-ab7f-e4e006b828e5/Least_Square.pdf

3. Least Squares and the Method of Maximum Likelihood

We have introduce the least square method, but a big problem is, why we choose to use square error criterion, but not other criterion like squared root?

That is because **LS and WLS produce the same estimates as maximum likelihood assuming Gaussian noise**

Assumption

1. Measurement Model: $y = Hx + v$
2. Noise Model: $v \sim \mathcal{N}\left(0, \sigma^2\right)$

Then $p(y \mid x) = \mathscr{N}\left(Hx, \sigma^2\right)$

Maximum Likelihood Solution

$$\begin{aligned} \hat{x}_{\mathrm{MLE}} &= \operatorname{argmax}_x p(\mathbf{y} \mid x) \\ &= \operatorname{argmax}_x \log p(\mathbf{y} \mid x) \\ &= \operatorname{argmin}_x -(\log p(\mathbf{y} \mid x)) \\ &= \operatorname{argmin}_x \frac{1}{2\sigma^2}\left(\left(y_1 - Hx\right)^2 + \ldots + \left(y_m - Hx\right)^2\right) \end{aligned}$$

That is

$$\hat{x}_{\mathrm{MLE}} = \hat{x}_{\mathrm{LS}} = \operatorname{argmin}_x \mathscr{L}_{\mathrm{LS}}(x) = \operatorname{argmin}_x \mathscr{L}_{\mathrm{MLE}}(x)$$

Central Limit Theorem

Theorem: Central Limit Theorem

Suppose $\{X_1, \ldots, X_n, \ldots\}$ is a sequence of i.i.d. random variables with $\mathbb{E}\left[X_i\right] = \mu$ and $\operatorname{Var}\left[X_i\right] = \sigma^2 < \infty$. Then as $n$ approaches infinity, the random variables $\sqrt{n}\left(\bar{X}n - \mu\right)$ converge in distribution to a normal distribution $\mathcal{N}\left(0, \sigma^2\right)$

$$\sqrt{n}\left(\bar{X}n - \mu\right) \xrightarrow{d} \mathcal{N}\left(0, \sigma^2\right).$$

Note:

- the **central limit theorem (CLT)** establishes that, in many situations, when independent random variables are summed up, their properly normalized sum tends toward a normal distribution (informally a bell curve) even if the original variables themselves are not normally distributed.
- CLT told us, the sum of different errors will tend to be normal distribution
  - when the error is not normal distribution, maybe we can average multiple measurement to approximate a normal distribution

Summary

- Least Square and Weighted Least Square
- Recursive Least Square for stream data
- Least Square corresponding to maximum likelihood when the noise is gaussian distribution