Chapter 3: Least Squares Estimation

Least Squares: There are a few different ways to formulate the least squares (LS) problem

Our formulation: Suppose we have some model $\check{\bar{z}} = h(\bar{x})$

- $\bar{x}$ is an N-element input vector
- $h(\bar{x})$ is a model relating $\bar{x}$ to $\bar{y}$
- $\check{\bar{z}}$ is an M-element output vector (or "measurement vector")

Typically, we know the model $h(\bar{x})$, we have a noisy (i.e., erroneous) measurement of the output vector $\check{\bar{z}}$ ($\bar{z}$), and we are attempting to estimate the unknown input vector $\bar{x}$ ($\hat{x}$)

- If there is no noise on $\bar{z}$, and our model $h(\bar{x})$ is perfect, then there will be some $\hat{x}$ that exactly satisfies $\bar{z} = h(\hat{x})$ where $\hat{x} = \bar{x}$

Never true in practice! We cannot perfectly estimate $\hat{x} = \bar{x}$

Normally, noise on $\bar{z}$ is modeled as additive $\rightarrow \bar{z} = h(\bar{x}) + \bar{v}$ ($\bar{v}$: M-element RV)

Adding noise means our estimate of $\bar{x}$ ($\hat{x}$) will have errors such that $h(\hat{x}) \neq \bar{z}$

- Our prediction of $\bar{z}$ based on the model ($h(\hat{x})$) and our estimate of $\bar{x}$ ($\hat{x}$) will never match our observed values of $\bar{z}$

- The best we can do is try to minimize the error
  - Goal: select $\hat{x}$ that gives $h(\hat{x})$ as close as possible to $\bar{z}$

Error between observed measurements ($\bar{z}$) and our prediction of the measurements ($h(\hat{x})$):

$$\bar{\varepsilon}_z = \bar{z} - h(\hat{x})$$

→ $\bar{\varepsilon}_z$ is $M \times 1$ vector

→ $\bar{\varepsilon}_z$ is called the measurement residual

We want to try to minimize $\bar{\varepsilon}_z$ based on some metric.

One way: define a cost function J as the sum of the squares of the elements of $\bar{\varepsilon}_z$

$$\rightarrow J = \varepsilon_{z1}^2 + \varepsilon_{z2}^2 + \dots + \varepsilon_{zM}^2 = \sum_{i=1}^{M} \varepsilon_{zi}^2 = \bar{\varepsilon}_z^T \bar{\varepsilon}_z$$

→ The goal of LS estimation is to find the estimate $\hat{x}$ that minimizes J (hence, "least squares")

Ordinary Least Squares (OLS):

Often, we will assume that the model $h(\bar{x})$ is linear

$$\rightarrow z_1 = H_{11} x_1 + \dots + H_{1N} x_N$$

$$\vdots$$

$$z_M = H_{M1} x_1 + \dots + H_{MN} x_N$$

$$\rightarrow \bar{z} = H\bar{x}$$

$\bar{z}$: $M \times 1$ H: $M \times N$ $\bar{x}$: $N \times 1$

With noise: $\bar{z} = H\bar{x} + \bar{v}$ → $\bar{\varepsilon}_z = \bar{z} - H\hat{x}$ → $J = \bar{\varepsilon}_z^T \bar{\varepsilon}_z$

We want to find $\hat{x}$ to minimize J - how?

→ Just like single-variable minimum → solve for $\hat{x}$ that makes the derivative of J wrt $\hat{x}$ equal to 0

Plug in $\bar{\varepsilon}_z = \bar{z} - H\hat{x} \rightarrow J = (\bar{z} - H\hat{x})^T(\bar{z} - H\hat{x})$

$$= \bar{z}^T\bar{z} - \hat{x}^T H^T \bar{z} - \bar{z}^T H\hat{x} + \hat{x}^T H^T H\hat{x}$$

$(H\hat{x})^T = \hat{x}^T H^T$ — Vector quadratic

Partial derivative wrt $\hat{x}$: $\frac{\partial J}{\partial \hat{x}} = \frac{\partial}{\partial \hat{x}}\left(\bar{z}^T\bar{z} - \hat{x}^T H^T\bar{z} - \bar{z}^T H\hat{x} + \hat{x}^T H^T H\hat{x}\right)$

$= 0$; Must transpose: $= -\bar{z}^T H$; $= -\bar{z}^T H$; $= \hat{x}^T H^T H + \hat{x}^T H^T H = 2\hat{x}^T H^T H$

$$= -\bar{z}^T H - \bar{z}^T H + 2\hat{x}^T H^T H = 2\hat{x}^T H^T H - 2\bar{z}^T H$$

Set equal to zero & solve for $\hat{x}$: $\frac{\partial J}{\partial \hat{x}} = 0 = 2\hat{x}^T H^T H - 2\bar{z}^T H$

$\rightarrow \hat{x}^T H^T H = \bar{z}^T H$ → Transpose both sides

$\rightarrow H^T H\hat{x} = H^T\bar{z}$

$\rightarrow \boxed{\hat{x} = (H^T H)^{-1} H^T\bar{z} = H^L\bar{z}}$ → OLS Estimator

→ H^L is called the left pseudo inverse of H

→ Requires that $M \geq N$ (more equations than unknowns) and H must be full-rank

(Note: must also show $\frac{\partial^2 J}{\partial \hat{x}^2}$ is positive semidefinite to prove that $\hat{x}$ is a local minimum of J and not another stationary point - see book)

Simple Example:

Measuring resistance x of an unknown resistance

Obtain M noisy measurements $\bar{z} = [z_1 \ldots z_M]^T$

$$\rightarrow z_1 = x + v_1$$
$$\vdots$$
$$z_M = x + v_M$$

$$\rightarrow \underbrace{\begin{bmatrix} z_1 \\ \vdots \\ z_M \end{bmatrix}}_{M\times 1} = \underbrace{\begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}}_{M\times 1} \underbrace{x}_{1\times 1} + \underbrace{\begin{bmatrix} v_1 \\ \vdots \\ v_M \end{bmatrix}}_{M\times 1}$$

$\bar{z}$ $\quad H \quad \bar{x} \quad \bar{v}$

$$\text{OLS} \rightarrow \hat{x} = (H^T H)^{-1} H^T \bar{z}$$

$$= \underbrace{\left(\underbrace{[1 \cdots 1]}_{} \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \right)^{-1}}_{= M \;\Rightarrow\; = \frac{1}{M}} \underbrace{[1 \cdots 1] \begin{bmatrix} z_1 \\ \vdots \\ z_M \end{bmatrix}}_{= \sum_{i=1}^{M} z_i} = \boxed{\frac{1}{M} \sum_{i=1}^{M} z_i} \quad \text{(take the average)}$$

Weighted Least Squares (WLS):

OLS assumes all our measurements in $\bar{z}$ are equally reliable

→ The elements of $\bar{v}$ $(v_1, v_2, \ldots)$ assumed to have equal variance

Suppose instead that each element of $\bar{v}$ has a unique variance AND is zero-mean:

$$\rightarrow \sigma_i^2 = E\{v_i^2\}$$

(zero-mean)

Assuming each element of $\bar{v}$ is independent of the others (i.e., $E\{v_i v_j\} = 0$ for $i \neq j$), then $\bar{v}$ has a covariance matrix R:

$$R = \text{cov}\{\bar{v}\} = E\{\bar{v}\bar{v}^T\} = \begin{bmatrix} \sigma_1^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_M^2 \end{bmatrix}$$

Note: Because $\check{z}$ is deterministic (but unknown) and $\bar{z} = \check{z} + \bar{v}$, it can be shown that $\text{cov}\{\bar{z}\} = \text{cov}\{\bar{v}\} = R$

We can define a new weighted cost function to minimize:

$$\rightarrow J = \frac{\varepsilon_{z1}^2}{\sigma_1^2} + \cdots + \frac{\varepsilon_{zM}^2}{\sigma_M^2} = \sum_{i=1}^{M} \frac{\varepsilon_{zi}^2}{\sigma_i^2} = \bar{\varepsilon}_z^T R^{-1} \bar{\varepsilon}_z$$

→ Cost function takes into account our confidence in our estimates
→ Lower variance estimates are weighted more in the cost function and higher variance estimates are weighted less

Again we want to minimize J:

$$\rightarrow J = \hat{\varepsilon}_z^T R^{-1} \hat{\varepsilon}_z = (\bar{z} - H\hat{x})^T R^{-1} (\bar{z} - H\hat{x})$$

$$= \bar{z}^T R^{-1} \bar{z} - \hat{x}^T H^T R^{-1} \bar{z} - \bar{z}^T R^{-1} H \hat{x} + \hat{x}^T H^T R^{-1} H \hat{x}$$

Vector quadratic

Differentiate wrt $\hat{x}$ and set to zero to find minimum:

$$\rightarrow \frac{\partial J}{\partial \hat{x}} = \frac{\partial}{\partial \hat{x}} \left(\bar{z}^T R^{-1} \bar{z} - \hat{x}^T H^T R^{-1} \bar{z} - \bar{z}^T R^{-1} H \hat{x} + \hat{x}^T H^T R^{-1} H \hat{x} \right)$$

$= 0$

Must transpose:
$= -\bar{z}^T (R^{-1})^T H$
$= -\bar{z}^T R^{-1} H$
$(R^{-1})^T = R^{-1}$

$= -\bar{z}^T R^{-1} H$

$= \hat{x}^T H^T R^{-1} H + \hat{x}^T H^T (R^{-1})^T H$
$= 2\hat{x}^T H^T R^{-1} H$

$$= 2\hat{x}^T H^T R^{-1} H - 2\bar{z}^T R^{-1} H = 0$$

$$\rightarrow \hat{x}^T H^T R^{-1} H = \bar{z}^T R^{-1} H \quad \rightarrow \text{Transpose both sides}$$

$$\rightarrow H^T R^{-1} H \hat{x} = H^T R^{-1} \bar{z}$$

$$\rightarrow \boxed{\hat{x} = (H^T R^{-1} H)^{-1} H^T R^{-1} \bar{z}} \rightarrow \text{WLS Estimator}$$

WLS assumes elements of $\hat{z}$ are independent (i.e., R is diagonal). However, it can be shown that this estimator minimizes the error if elements of $\bar{z}$ are correlated (i.e., non-diagonal R). This is called generalized least squares (GLS) or the best linear unbiased estimator (BLUE).

Recursive Least Squares (RLS)

Problem: we may not get all our measurements $\bar{z}$ at once, and we may want to get estimates $\hat{x}$ before all our measurements arrive

Prior formulation: $\bar{z} = H\bar{x} \rightarrow \bar{z} = H\bar{x} + \bar{v}$

$\bar{z}$: Mx1, H: MxN, $\bar{x}$: Nx1, $\bar{v}$: Mx1

→ We have M noisy observations/measurements relating to $\bar{x}$ through a linear model H

→ These M measurements may consist of a "vector of vectors"

Example: 2-D position

$\bar{x} = \begin{bmatrix} x \\ y \end{bmatrix}$ and $M = 2k$ where k is the number of noisy pos. estimates

$$\rightarrow \bar{z} = \begin{bmatrix} \bar{p}_1 \\ \vdots \\ \bar{p}_k \end{bmatrix} = \begin{bmatrix} x_1 \\ y_1 \\ x_2 \\ y_2 \\ \vdots \\ x_k \\ y_k \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \\ & \vdots \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} v_{x1} \\ v_{y1} \\ \vdots \\ v_{xk} \\ v_{yk} \end{bmatrix}$$

$\bar{z}$: Mx1, H: Mx2, $\bar{x}$: 2x1, $\bar{v}$: Mx1

Every time we get a new measurement, M increases and we must solve OLS/GLS from scratch. For very large values of M, this will become very computationally taxing!

Recursive Formulation: Suppose instead that we already have an estimate of $\bar{x}$ ($\hat{x}$) based on $(k-1)$ measurements $\bar{z}_1, \bar{z}_2, \ldots, \bar{z}_{k-1}$, and we get a new measurement $\bar{z}_k$

Goal: Now want to recompute $\hat{x}$ ($\hat{x}_k$) based only on $\bar{z}_k$ and previous $\hat{x}$ ($\hat{x}_{k-1}$)

Example: 2-D position

$$\bar{z}_k = \begin{bmatrix} x_k \\ y_k \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} v_{xk} \\ v_{yk} \end{bmatrix}$$

$\bar{z}_k$ | H_k | $\bar{x}$ | $\bar{v}_k$

H_k is constant in this problem, but in general may change w/ k

→ Measurements are $M \times 1$ and $\hat{x}/\bar{x}$ is $N \times 1$

→ Each new measurement is still $M \times 1$, but M doesn't grow w/ k now

→ Computational cost of adding new measurements never increases

→ In this problem, $M = 2$ always instead of $M = 2k$

Assume at time index k we have a new measurement $\bar{z}_k$ and a previous estime $\hat{x}_{k-1}$

→ A linear recursive estimator can be written as:

$$\bar{z}_k = H_k \bar{x} + \bar{v}_k$$

$$\hat{x}_k = \hat{x}_{k-1} + K_k\left(\bar{z}_k - H_k \hat{x}_{k-1}\right)$$

Updated estimate of $\bar{x}$

Previous estimate of $\bar{x}$

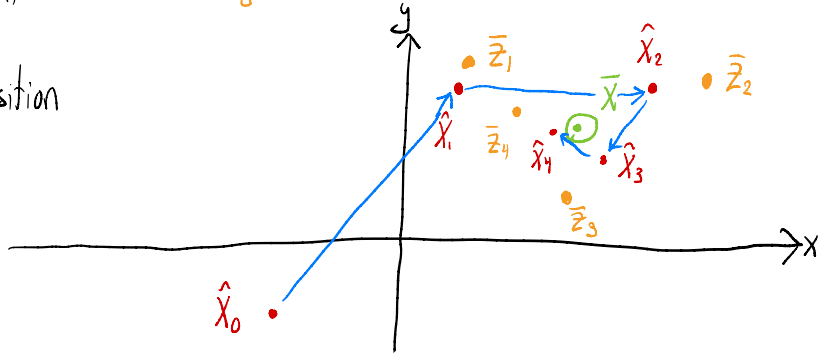
Estimator gain matrix
→ This matrix essentially "weights" the new measurement against our previous estimate to determine how much adjustment is needed, based on our confidence in either one

Correction term
→ This terms describes how different our new measurement $\bar{z}_k$ is from what it "should" be assuming $\hat{x}_{k-1} = \bar{x}$

Assumptions:
- $\bar{x}$ is constant (i.e., does not change with k)
- Recursive → start with some initial estimate $\hat{x}_0$
- Noise $\bar{v}_k$ is white w/ covariance R_k → $\bar{v}_k$ is measurement noise, R_k is measurement/observation covariance

In English: The correction term compares the new measurement ($\bar{z}_k$) to a prediction of $\bar{z}_k$ based on the previous estime $\hat{x}_{k-1}$. The larger the difference between the two (i.e., the more our new measurement disagrees with our previous estimate), the larger the correction of the previous estimate will be when computing the new estimate $\hat{x}_k$. The correction gain matrix K_k weights the confidence of the new measurement against the confidence of the previous estimate to determine how large of a correction is required. If our confidence is the previous estimate is high, the correction gain will be smaller. If our confidence in the new measurement is high, the correction gain will be larger.

Example: 2-D position



Observations:

- As more measurements are added, our confidence in $\hat{x}_{k-1}$ increases and the estimator gain goes down, reducing the size of correction
  → If we ever have perfect confidence in $\hat{x}_{k-1}$, the estimator gain $K_k = 0$ and we won't change our estimate
- The closer our measurement $\bar{z}_k$ is to the prior estimate $\hat{x}_{k-1}$, the smaller the correction term, reducing the size of correction
  → If our measurement $\bar{z}_k$ ever perfectly matches the prediction $H_k \hat{x}_{k-1}$, the correction term = 0 and we won't change our estimate

Note: in the 2-D position example, $\bar{X}$ and $\bar{Z}_k$ represent the same thing (position vs. position measurement), so we can plot in same space.
This won't be true in general (i.e., H_k not always an identity matrix).
This situation is harder to vizualize but the math is unaffected.

Back to model:

$$\bar{Z}_k = H_k \bar{X} + \bar{V}_k$$

$$\hat{X}_k = \hat{X}_{k-1} + K_k \left(\bar{Z}_k - H_k \hat{X}_{k-1} \right)$$ → Note: this estimator is unbiased for any K_k value (proof in book)

→ How do we calculate the gain matrix K_k?

→ The criteria we will use is to try to minimize the variance of the errors of $\hat{X}_k$

$$\bar{\varepsilon}_{x,k} = \bar{X} - \hat{X}_k \longrightarrow P_k = \text{cov}\{\bar{\varepsilon}_{x,k}\} = E\{\bar{\varepsilon}_{x,k} \bar{\varepsilon}_{x,k}^T\}$$

→ P_k is the estimation error covariance (covariance of $\bar{\varepsilon}_{x,k}$ and $\hat{x}_0$!)

→ Although we can't estimate the errors, we can estimate their covariance!

Remember that the diagonal of a covariance matrix are variances of elements

Suppose we want to minimize the sum of the variances of $\bar{\varepsilon}_{x,k}$

→ New cost function: $J_k = \underbrace{E\{\varepsilon_{x1,k}^2\}}_{E\{\varepsilon_{xi,k}\} = 0 \text{ b/c estimator unbiased}} + \dots + E\{\varepsilon_{xN,k}^2\} = Tr(P_k)$

$$= E\{Tr(\bar{\varepsilon}_{x,k} \bar{\varepsilon}_{x,k}^T)\}$$

Let's find an expression for $\bar{\varepsilon}_{x,k}$:

$$\bar{\varepsilon}_{x,k} = \bar{X} - \hat{X}_k = \underbrace{\bar{X} - \hat{X}_{k-1}}_{= \bar{\varepsilon}_{x,k-1}} - K_k \left(\underbrace{\bar{Z}_k}_{= H_k \bar{X} + V_k} - H_k \hat{X}_{k-1} \right)$$

$$= \bar{\varepsilon}_{x,k-1} - K_k \left(H_k \bar{x} + \bar{v}_k - H_k \hat{x}_{k-1} \right)$$

$$= \bar{\varepsilon}_{x,k-1} - K_k H_k \underbrace{\left(\bar{x} - \hat{x}_{k-1} \right)}_{= \bar{\varepsilon}_{x,k-1}} - K_k \bar{v}_k$$

$$= \underline{\left(I - K_k H_k \right) \bar{\varepsilon}_{x,k-1} - K_k \bar{v}_k}$$

→ We can express the error of the current estimate $\hat{x}_k$ in terms of the noise on $\bar{z}_k$ and the error of the previous estimate $\hat{x}_{k-1}$

→ This means we can express P_k as a function of P_{k-1} and R_k, and thus solve for an optimal K_k that takes into account our confidence in our previous estimate $\hat{x}_{k-1}$ vs. our confidence in our measurements $\bar{z}_k$

→ Expression for P_k:

$$P_k = E\left\{ \bar{\varepsilon}_{x,k} \bar{\varepsilon}_{x,k}^T \right\}$$

$$= E\left\{ \left(\left(I - K_k H_k \right) \bar{\varepsilon}_{x,k-1} - K_k \bar{v}_k \right) \left(\left(I - K_k H_k \right) \bar{\varepsilon}_{x,k-1} - K_k \bar{v}_k \right)^T \right\}$$

$$= \left(I - K_k H_k \right) E\left\{ \bar{\varepsilon}_{x,k-1} \bar{\varepsilon}_{x,k-1} \right\} \left(I - K_k H_k \right)^T - K_k E\left\{ \bar{v}_k \bar{\varepsilon}_{x,k-1}^T \right\} \left(I - K_k H_k \right)^T$$

$E\left\{ \bar{\varepsilon}_{x,k-1} \bar{\varepsilon}_{x,k-1} \right\} = P_{k-1}$

$E\left\{ \bar{v}_k \bar{\varepsilon}_{x,k-1}^T \right\}$: Uncorrelated and zero-mean → $= 0$

$$- \left(I - K_k H_k \right) E\left\{ \bar{\varepsilon}_{x,k-1} \bar{v}_k^T \right\} K_k^T + K_k \underbrace{E\left\{ \bar{v}_k \bar{v}_k^T \right\}}_{= R_k} K_k^T$$

$E\left\{ \bar{\varepsilon}_{x,k-1} \bar{v}_k^T \right\}$: Uncorrelated and zero-mean → $= 0$

$$\rightarrow P_k = \left(I - K_k H_k \right) P_{k-1} \left(I - K_k H_k \right)^T + K_k R_k K_k^T$$

→ Now we can solve for optimal K_k by computing $\frac{d\,\mathrm{Tr}(P_k)}{dK_k} = \frac{dJ_k}{dK_k}$ and setting $= 0$

Recall: $\frac{\partial \text{Tr}(ABA^T)}{\partial A} = 2AB$ if B symmetric

Put that together with some chain rule magic and we get:

$$\rightarrow \frac{\partial J_k}{\partial K_k} = 2(I - K_k H_k) P_{k-1}(-H_k^T) + 2K_k R_k = 0$$

$$\rightarrow K_k R_k = (I - K_k H_k) P_{k-1} H_k^T = P_{k-1} H_k^T - K_k H_k P_{k-1} H_k^T$$

$$\rightarrow K_k R_k + K_k H_k P_{k-1} H_k^T = K_k \left(R_k + H_k P_{k-1} H_k^T\right) = P_{k-1} H_k^T$$

$$\rightarrow \boxed{K_k = P_{k-1} H_k^T \left(R_k + H_k P_{k-1} H_k^T\right)^{-1}}$$

→ Computing K_k requires knowledge of P_{k-1}

→ Just like how we need an initial estimate $\hat{X}_0$, we need an initial covariance of that estimate P_0

Recursive Least Squares:

1. Initialize estimator

$$\rightarrow \boxed{\hat{X}_0 = E\{\bar{X}\} \quad \text{and} \quad P_0 = E\{(\bar{X} - \hat{X}_0)(\bar{X} - \hat{X}_0)^T\}}$$

Selecting these values is often more art than science...

Rules of thumb: If we are confident in our estimate of $\hat{X}_0$, then P_0 will be close to 0.

If we have no clue about our estimate $\hat{X}_0$, then $P_0 \rightarrow \infty I$.

2. For $k = 1, 2, \ldots$

a.) Obtain measurements $\bar{Z}_k$ w/ model $\bar{Z}_k = H_k \bar{X} + \bar{V}_k$

- We assume $\bar{v}_k$ is zero-mean w/ covariance R_k
- We assume $\bar{v}_k$ is white (i.e., $\bar{v}_i$ is independent of $\bar{v}_j$ unless $i = j$)

b.) Update estimate $\hat{x}_k$ and estimation error covariance P_k

$$\rightarrow K_k = P_{k-1} H_k^T \left(H_k P_{k-1} H_k^T + R_k \right)^{-1}$$

$$\hat{x}_k = \hat{x}_{k-1} + K_k \left(\bar{z}_k - H_k \hat{x}_{k-1} \right)$$

$$P_k = \left(I - K_k H_k \right) P_{k-1} \left(I - K_k H_k \right)^T + K_k R_k K_k^T$$

Other ways of calculating K_k and P_k (derived in book):

- $K_k = P_k H_k^T R_k^{-1}$ (requires computing P_k first)
- $P_k = \left(P_{k-1}^{-1} + H_k^T R_k^{-1} H_k \right)^{-1}$

  $= \left(I - K_k H_k \right) P_{k-1}$ (requires computing K_k first)