Lecture #1:
Stochastic Least Squares (LS) Estimation

Gustaf Hendeby
gustaf.hendeby@liu.se

Version: 2023-11-11

We will start off by studying the least squares estimator.

Theorem 1. *Let X and Y be two jointly distributed random variables. The least squares (minimum variance, MV) estimator $\hat{X}$ of X given Y is*

$$\hat{X} = \mathsf{E}_{X|Y}\left[X\right]$$

and $\hat{x} = \mathsf{E}_{X|Y=y}\left[X\right]$ is the matching (deterministic) estimate.

Proof. Let $\tilde{X} = X - \hat{X} = X - h(Y)$, then we seek the function $h(\cdot)$ that solves

$$\min_{h(\cdot)} \mathsf{E}[\tilde{X}^*\tilde{X}],$$

where A^* is the Hermitian transpose of A.

Note: $\mathsf{E}_{XY}\left[g^*(Y)X\right] = \mathsf{E}_X\left[\mathsf{E}_{X|Y}\left[g^*(Y)X\right]\right] = E_Y\left[g^*(Y)\,\mathsf{E}_{X|Y}\left[X\right]\right]$

$$\begin{aligned}
\mathsf{E}\left[\tilde{X}^*\tilde{X}\right] &= \mathsf{E}\left[(X-\hat{X})^*(X-\hat{X})\right] = \mathsf{E}\Big[\big(X - \mathsf{E}_{X|Y}\left[X\right] + \mathsf{E}_{X|Y}\left[X\right] - \hat{X}\big)^*\big(X - \mathsf{E}_{X|Y}\left[X\right] + \mathsf{E}_{X|Y}\left[X\right] - \hat{X}\big)\Big] \\
&= \Big/ g(Y) = \mathsf{E}_{X|Y}\left[X\right] - \hat{X} \Big/ = \mathsf{E}\Big[\big(X - \mathsf{E}_{X|Y}\left[X\right] + g(Y)\big)^*\big(X - \mathsf{E}_{X|Y}\left[X\right] + g(Y)\big)\Big] \\
&= \mathsf{E}\Big[\big(X - \mathsf{E}_{X|Y}\left[X\right]\big)^*\big(X - \mathsf{E}_{X|Y}\left[X\right]\big)\Big] + \mathsf{E}\Big[g^*(Y)g(Y)\Big] + 2\,\mathrm{Re}\,\underbrace{\mathsf{E}\Big[g^*(Y)\big(X - E_{X|Y}\left[X\right]\big)\Big]}_{=\mathsf{E}_Y\left[g^*(Y)\,\mathsf{E}_{X|Y}\left[X - \mathsf{E}_{X|Y}[X]\right]\right]=0} \\
&= \mathsf{E}\Big[\big(X - \mathsf{E}_{X|Y}\left[X\right]\big)^*\big(X - \mathsf{E}_{X|Y}\left[X\right]\big)\Big] + \mathsf{E}\Big[\big(\mathsf{E}_{X|Y}\left[X\right] - \hat{X}\big)^*\big(\mathsf{E}_{X|Y}\left[X\right] - \hat{X}\big)\Big]
\end{aligned}$$

This expression is minimized by $\hat{X} = \mathsf{E}[X|Y]$, as both terms are positive and only the second depends on $\hat{X}$. □

1 Conditional vs Unconditional MSE

There are two possibilities versions of MSE:

Unconditional:

$$\mathsf{E}\left[\left(X-\hat{X}(Y)\right)^*\left(X-\hat{X}(Y)\right)\right] \quad \text{(expectation of both } X \text{ and } Y\text{)}$$

Conditional:

$$\mathsf{E}_{X|Y=y}\left[(X-\hat{X}(Y))^*(X-\hat{X}(Y))\right] = \mathsf{E}_{X|Y=y}\left[X^*X\right] - \hat{x}^*\hat{x}$$

Generally the unconditional and the conditional MSE are not the same, but in the Gaussian case they are! (See HW #1.)

2 Unbiasedness

Unbiasedness is defined as the estimator/estimate that on average assumes the true value, that is

Estimator:

$$\mathsf{E}_{XY}\left[X-\hat{X}(Y)\right] = \mathsf{E}_X\left[X\right] - \underbrace{\mathsf{E}_Y\left[\mathsf{E}_{X|Y}\left[X\right]\right]}_{\mathsf{E}_X[X]} = 0$$

Estimate:

$$\mathsf{E}_{X|Y}\left[X-\hat{x}\right] = \mathsf{E}_{X|Y=y}\left[X\right] - \hat{x} = 0$$

3 Problem with Explicit Solutions

- The estimate $\hat{x}$ is often a complicated function of the observations $y_1, \ldots, y_N$.
- Knowledge of the joint pdf is needed.

4 Linear LS estimation

If our estimate $\hat{x}$ of X is restricted to be a linear function (including affine mappings), then $\hat{x}$ only depends on the first and second order moments of X and Y. If X and Y are jointly Gaussian, then the conditional mean $\mathsf{E}_{Y|Y}[X]$ is a linear function of the observations.

4.1 Gaussian case (Part of HW #1)

If

$$\begin{pmatrix} x \\ y \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{pmatrix}\right)$$

then the conditional pdf $f_{X|Y}(x|y)$ is Gaussian with

mean: $\mu_{x|y} = \mu_x + \Sigma_{xy}\Sigma_{yy}^{-1}(y-\mu_y)$

covariance: $\Sigma_{x|y} = \Sigma_{xx} - \Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx}$

which yields

$$\mathsf{E}_{X|Y=y}\left[X\right] = \mu_x + \Sigma_{xy}\Sigma_{yy}^{-1}(y-\mu_y)$$

4.2 Linear LSE (derivation scalar case)

Let $\hat{X} = aY + b$ and find the a and b that minimizes the MSE $\mathsf{E}\big[(X-\hat{X})^2\big]$.

$$\frac{\partial\,\mathsf{E}\big[(X-\hat{X})^2\big]}{\partial a} = 2\,\mathsf{E}\left[(X-aY-b)\frac{\partial(X-\hat{X})}{\partial a}\right] = -2\big(\mathsf{E}\big[XY\big] - a\,\mathsf{E}\big[Y^2\big] - b\,\mathsf{E}\big[Y\big]\big) = 0$$

$$\frac{\partial\,\mathsf{E}\big[(X-\hat{X})^2\big]}{\partial b} = -2\,\mathsf{E}\big[X-aY-b\big] = 0$$

yielding

$$a = \frac{\mathsf{E}\big[XY\big]-\mu_x\mu_y}{\mathsf{E}\big[Y^2\big]-\mu_y^2} = \frac{\sigma_{xy}^2}{\sigma_y^2}$$

$$b = \mu_x - \frac{\sigma_{xy}^2}{\sigma_y^2}\mu_y$$

that is

$$\hat{X} = \mu_x + \frac{\sigma_{xy}^2}{\sigma_y^2}(Y-\mu_y).$$

- **Note:** To compute the estimator only requires the first and second order moments.
- The vector case follows analogous [1, Sec. 3.2.8]:

$$\hat{X} = \mu_x + \Sigma_{xy}\Sigma_{yy}^{-1}(Y-\mu_y).$$

- **Generally we will assume μ_x, $\mu_y = 0$ without loss of generality.**

4.3 Genereic Solution and Comutational Complexity

If

$$\hat{X}_N = \sum_{i=0}^{N} a_{N,i}^T y_{N-i},$$

where N is the number of observations and the zero mean assumption is used, then the coefficients

$$A_N = \begin{pmatrix} a_{N,0} & \dots & a_{N,N} \end{pmatrix}$$

are the solution to $A_N R_y = R_{xy}$. This is a standard problem to slow, but

- Requires $\mathscr{O}(N^3)$ operations to compute.
- Often a recursive solution is desired, *i.e.*, $A_N \to A_{N+1}$ and $\hat{x}_N \to \hat{x}_{N+1}$.

If we impose a structure on R_y we can reduce the complexity and find a recursive update.

4.3.1 Example:

Assume $X(t)$ to be a *wide-sense stationary* (wss.) process, where the mean is $\mu_x = \mathsf{E}\big[X(t)\big] = 0$ and the auto correlation function $r_{xx}(\tau) = e^{-\alpha|\tau|}$.

Predict $X(3T+\Delta)$, $\Delta > 0$ given $X(T)$, $X(2T)$, and $X(3T)$. The estimator (predictor) should be a llse. Given that the mean $\mu_x = 0$ it follows that

$$\hat{X} = ay, \qquad\qquad y = \begin{pmatrix} X(T) & X(2T) & X(3T) \end{pmatrix}^T$$

where $a = \Sigma_{xy}\Sigma_{yy}^{-1}$

$$\Sigma_{xy} = \mathsf{E}[X(3T+\Delta)y^T] = e^{-\alpha\Delta}\begin{pmatrix} e^{-\alpha 2T} & e^{-\alpha T} & 1 \end{pmatrix}$$
$$\Sigma_{yy} = \begin{pmatrix} 1 & e^{-\alpha T} & e^{-2\alpha T} \\ e^{-\alpha T} & 1 & e^{-\alpha T} \\ e^{-2\alpha T} & e^{-\alpha T} & 1 \end{pmatrix}$$

assuming that $\alpha T \neq 0$ it follows that

$$a = \Sigma_{xy}\Sigma_{yy}^{-1} \iff a\Sigma_{yy} = \Sigma_{xy}$$
$$a = \begin{pmatrix} 0 & 0 & e^{-\alpha\Delta} \end{pmatrix} \iff a \begin{pmatrix} 1 & e^{-\alpha T} & e^{-2\alpha T} \\ e^{-\alpha T} & 1 & e^{-\alpha T} \\ e^{-2\alpha T} & e^{-\alpha T} & 1 \end{pmatrix} = e^{-\alpha\Delta}\begin{pmatrix} e^{-2\alpha T} & e^{-\alpha T} & 1 \end{pmatrix}$$
$$\hat{X} = \begin{pmatrix} 0 & 0 & e^{-\alpha\Delta} \end{pmatrix} y = e^{-\alpha\Delta}X(3T)$$

MSE:

$$\mathsf{E}\Big[\big(X(3T+\Delta) - \hat{X}\big)^2\Big] = \Sigma_{xx} - \Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx} = r_{xx}(0) - a\Sigma_{xy}^T = 1 - e^{-2\alpha\Delta}$$

> **Note:**
>
> -
> $$\Delta \to \infty \Rightarrow \begin{cases} a \to 0 \\ \hat{x} \to 0 & \text{zero-mean process} \\ \text{MSE} \to 1 \end{cases}$$
>
> - The estimator/estimate depends only on $X(3T)$, since $X(t)$ is a Markov process.

References

[1] Thomas Kailath, Ali H. Sayed, and Babak Hassibi. *Linear Estimation*. Prentice-Hall, Inc, 2000. ISBN 0-13-022464-2.