Lecture 10
Estimation of Random Variables

Instructor Name: John Lipor

Recommended Reading: Pishro-Nik: 9.1.0 - 9.1.6, 8.2.2; Gubner: 8.4 - 8.6

In Lecture 5, we saw how the likelihood and posterior distributions can be used to decide between two hypotheses, resulting in the maximum likelihood (ML) and maximum a posteriori (MAP) detection rules. In the machine learning world, deciding between a finite number of hypotheses is known as *classification*. What if we instead wish to estimate an actual parameter (called *regression* in the machine learning world) instead of deciding among a few possible options?

Example 1. The canonical example of estimation is the case where we observe a noisy random variable $Y_i = X + Z_i$, where X is a parameter of interest, and Z_i is zero-mean noise.

1 Minimum Mean-Squared Error (MMSE) Estimation

One approach to estimation is to minimize the mean-squared error (MSE) between the estimate and the true value, i.e., to minimize

$$\mathrm{MSE}(\hat{X}) = \mathbb{E}\left[\left(\hat{X} - X\right)^2\right].$$

First, what is random in the above expression? Since we are estimating the RV X, we know that piece is random. Note also that if our observations are at all useful, $\hat{X}$ will be a function of the Y_i's, so the expectation above is a joint expectation.

1.1 MMSE Estimation With No Observations

Suppose we wish to estimate X without obtaining any observations Y_i. What is the best choice? In this case, the expectation when computing the MSE is only over X, so we can compute

$$\begin{aligned} \mathbb{E}\left[\left(\hat{X} - X\right)^2\right] &= \mathbb{E}\left[\left(\hat{X} - \mathbb{E}[X] + \mathbb{E}[X] - X\right)^2\right] \\ &= \mathbb{E}\left[\left(\hat{X} - \mathbb{E}[X]\right)^2\right] + \mathbb{E}\left[(\mathbb{E}[X] - X)^2\right] + 2\mathbb{E}\left[\left(\hat{X} - \mathbb{E}[X]\right)(\mathbb{E}[X] - X)\right]. \end{aligned}$$

Note that $\hat{X}$ and $\mathbb{E}[X]$ are both deterministic, so the last term above becomes

$$2\left(\hat{X} - \mathbb{E}[X]\right)\mathbb{E}\left[\mathbb{E}[X] - X\right] = 2\left(\hat{X} - \mathbb{E}[X]\right)(\mathbb{E}[X] - \mathbb{E}[X]) = 0.$$

Therefore, we see the following **extremely important** breakdown of MSE

$$\mathrm{MSE}(\hat{X}) = \underbrace{\left(\hat{X} - \mathbb{E}[X]\right)^2}_{\text{bias squared}} + \underbrace{\mathbb{E}\left[(X - \mathbb{E}[X])^2\right]}_{\text{variance}}.$$

In other words, the MSE can be decomposed into the sum of squared bias and variance. In machine learning, we spend a great deal of time developing estimators that balance these two terms. In estimation theory, we have no control over the variance, so the $\hat{X}$ that minimizes the above is

$$\hat{X} = \mathbb{E}[X]. \tag{1}$$

1.2 MMSE Estimation With Observations

Now suppose we observe the RV Y and want to incorporate this information. To do this, we first establish a broader fact known as the *orthogonality principle*.

Theorem 1 (Orthogonality Principle)**.** Let $\hat{X} = g(Y)$ be an estimator of X. If

$$\mathbb{E}\left[h(Y)\left(X - g(Y)\right)\right] = 0$$

for all functions h, then

$$\mathbb{E}\left[\left(X - g(Y)\right)^2\right] \leq \mathbb{E}\left[\left(X - h(Y)\right)^2\right],$$

i.e., $g(Y)$ is the MMSE estimator.

Proof. We again use the "add and subtract" trick to rewrite the MSE

$$\begin{aligned}
\mathbb{E}_{XY}\left[\left(X - h(Y)\right)^2\right] &= \mathbb{E}_{XY}\left[\left(X - g(Y) + g(Y) - h(Y)\right)^2\right] \\
&= \mathbb{E}_{XY}\left[\left(X - g(Y)\right)^2\right] + \mathbb{E}_{XY}\left[\left(g(Y) - h(Y)\right)^2\right] - 2\mathbb{E}_{XY}\left[\left(X - g(Y)\right)\left(h(Y) - g(Y)\right)\right].
\end{aligned}$$

Since $h(Y) - g(Y)$ is a function of Y, the final term above is zero by assumption of the theorem. This gives

$$\begin{aligned}
\mathbb{E}_{XY}\left[\left(X - h(Y)\right)^2\right] &= \mathbb{E}_{XY}\left[\left(X - g(Y)\right)^2\right] + \mathbb{E}_{XY}\left[\left(g(Y) - h(Y)\right)^2\right] \\
&\geq \mathbb{E}_{XY}\left[\left(X - g(Y)\right)^2\right]
\end{aligned}$$

as desired, since the square of anything is nonnegative. □

To gain some intuition for this theorem, we can think of X and Y as vectors in $\mathbb{R}^2$. In this case, we're trying to find the closest vector to X that lies in the direction of Y. In this case, we want to "project" X on to the span of Y, so the residual $X - g(Y)$ should be orthogonal to Y and any function of Y.

With the orthogonality principle in mind, we now present a function $g(Y)$ that satisfies the orthogonality principle and is therefore the MMSE estimator.

Theorem 2. Let $g(Y) = \mathbb{E}\left[X \mid Y\right]$. Then for all functions h,

$$\mathbb{E}\left[h(Y)\left(X - g(Y)\right)\right] = 0.$$

Proof. Using the law of total probability,

$$\begin{aligned}
\mathbb{E}_{XY}\left[h(Y)\left(X - g(Y)\right)\right] &= \mathbb{E}_{Y}\left[\mathbb{E}_{X|Y}\left[h(Y)\left(X - g(Y)\right)\right]\right] \\
&= \int_Y \mathbb{E}_{X|Y}\left[h(Y)\left(X - g(Y)\right)\right] f_Y(y)dy \\
&= \int_Y h(Y)\left(\mathbb{E}_{X|Y}\left[X \mid Y = y\right] - g(y)\right) f_Y(y)dy \\
&= \int_Y h(Y)\left(\mathbb{E}_{X|Y}\left[X \mid Y = y\right] - \mathbb{E}_{X|Y}\left[X \mid Y = y\right]\right) f_Y(y)dy \\
&= 0.
\end{aligned}$$

□

Combining the above with the orthogonality principle, we see that

$$\hat{X}_{MMSE} = \mathbb{E}\left[X \mid Y\right]. \tag{2}$$

1.3 Linear MMSE Estimation

Sometimes finding $\mathbb{E}[X \mid Y]$ is too difficult, since it may require knowing the joint distribution of X and Y, which I've mentioned can be impractical. One way to overcome this difficulty is to limit ourselves to simple estimators. One such restriction is to require that $\hat{X}$ be a linear (actually affine) function of Y, i.e.,

$$\hat{X} = aY + b$$

for some constants $a, b \in \mathbb{R}$. Let $\mu_X = \mathbb{E}[X]$ and $\mu_Y = \mathbb{E}[Y]$. To find a linear estimator, we plug this form of $\hat{X}$ into the MSE equation to see that

$$\begin{aligned} \mathbb{E}\left[\left(\hat{X} - X\right)^2\right] &= \mathbb{E}\left[(X - (aY + b))^2\right] \\ &= \mathbb{E}\left[(X - \mu_x - a(Y - \mu_Y) + (\mu_x - a\mu_Y - b))^2\right] \\ &= \mathbb{E}\left[(X - \mu_X - a(Y - \mu_Y))^2\right] + (\mu_x - a\mu_Y - b)^2 + 2\left(\mu_x - a\mu_Y - b\right)\mathbb{E}\left[X - \mu_X - a(Y - \mu_Y)\right]. \end{aligned}$$

Note that

$$\mathbb{E}\left[X - \mu_X - a(Y - \mu_Y)\right] = \mathbb{E}[X - \mu_X] - a\mathbb{E}[Y - \mu_Y] = 0,$$

so

$$\mathbb{E}\left[\left(\hat{X} - X\right)^2\right] = \mathbb{E}\left[(X - \mu_X - a(Y - \mu_Y))^2\right] + (\mu_x - a\mu_Y - b)^2.$$

Our goal now is to minimize the above over a and b. We begin by minimizing b, which we can do by making the second term zero, resulting in

$$b = \mu_X - a\mu_Y.$$

To find a, let $\bar{X} = X - \mu_X$ and $\bar{Y} = Y - \mu_Y$. We wish to minimize (over a)

$$\mathbb{E}\left[\left(\bar{X} - a\bar{Y}\right)^2\right] = \mathbb{E}\left[\bar{X}^2 + a^2\bar{Y}^2 - 2a\bar{X}\bar{Y}\right].$$

We can ignore the first term since it does not depend on a, so we wish to solve

$$\min_a a^2\mathbb{E}\left[\bar{Y}^2\right] - 2a\mathbb{E}\left[\bar{X}\bar{Y}\right].$$

Differentiating and setting to zero, we see that

$$a = \frac{\mathbb{E}\left[\bar{X}\bar{Y}\right]}{\mathbb{E}\left[\bar{Y}^2\right]} = \frac{\text{cov}(X, Y)}{\text{var}(Y)}.$$

Putting this all together gives the linear MMSE estimator

$$\hat{X}_{LMMSE} = \frac{\text{cov}(X, Y)}{\text{var}(Y)}\left(Y - \mu_Y\right) + \mu_X. \quad (3)$$

2 Minimum Absolute Error (MAE) Estimation

There is no reason to be restricted to the MSE as a cost function for our estimator. Another obvious function to consider is the absolute error

$$\text{AE}\left(\hat{X}\right) = \left|\hat{X} - X\right|.$$

As you showed on Homework 3, the optimal MAE estimator is the median of the resulting distribution

$$\hat{X}_{MAE} = \underset{x}{\text{median}}\, f_{X|Y}\left(x \mid y\right). \quad (4)$$

3 Maximum Likelihood (ML) Estimation

The ML estimator is probably the most used and has some nice properties that are discussed in future courses. It is defined as

$$\hat{X}_{ML} = \arg\max_x f_{Y|X}(y \mid x). \tag{5}$$

4 Maximum a Posteriori (MAP) Estimation

Just as with detection, we can define the MAP estimator, which is the same as ML if we have uniform priors on X.

$$\hat{X}_{MAP} = \arg\max_x f_{X|Y}(x \mid y) = \arg\max_x f_{Y|X}(y \mid x) f_X(x). \tag{6}$$