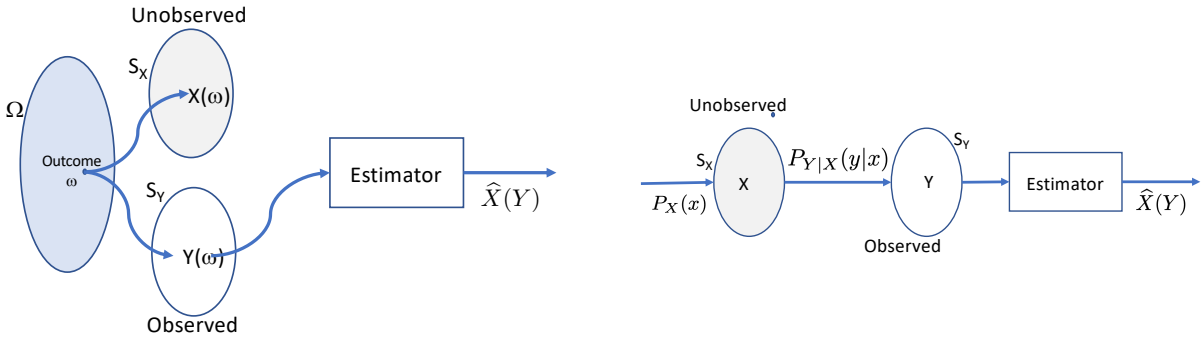
Chapter 7

Estimation

7.1 Introduction

In this chapter we consider the problem of estimating or inferring the values of unknown variables based on observation of a related set of random variables. A simple model of the estimation situation we are considering is depicted in Figure 7.1(a). An experiment generates pairs of random variables X, Y. We observe one of the two random variables, Y. Based on the observed value $Y = y$, we want to estimate the unobserved variable X by using an estimation rule $\widehat{x}(y)$. This model can be extended to cases where $\underline{X}, \underline{Y}$ are random vectors, so that several random variables are observed, and several unknown random variables are to be estimated.



(a) Jointly distributed random variables

(b) Sequential experiment

Figure 7.1: Different Views of Estimation Problem.

Assuming X, Y are discrete random variables, the probabilistic description of the variables X, Y is summarized by the joint probability mass function $P_{X,Y}(x, y)$, which we factor using the product rule as $P_X(x)P_{Y|X}(y|x)$. The second term in this factorization is the likelihood function, which captures the statistical relationship of how Y varies depending on the value of X. We can view this experiment as a sequential experiment, where the unobserved variable X is generated first, with probability law $P_X(x)$. Depending on the value of X, the observed variable Y is generated with probability law $P_{Y|X}(y|x)$. Figure 7.1(b) shows this sequential model, which is the one that we use in discussing estimation problems in this chapter.

This model has two components:

1. A model of the experiment that generates the unobserved random variable X, described by either $P_X(x)$ if X is discrete, or $f_X(x)$ if X is continuous.
2. A model of the observed random variable Y, represented by the conditional probability mass function $P_{Y|X}(y|x)$, if Y is discrete, or the conditional probability density function $f_{Y|X}(y|x)$ if y is continuous.

This model captures the essential elements of many problems in engineering and science, including: finding the location of a target based on radar observations, estimating the heart rate of a patient from electrical measurements, discerning O^+ density in the atmosphere from brightness measurements, and estimating depth in a scene from apparent motion in video.

The goal of estimation is to obtain an estimation rule that maps each observed value $y \in R_Y$ to a corresponding estimated value $\widehat{x}(y) \in \Re$. In some cases, we restrict the estimated value to be in R_X. The rule $\widehat{x}(y)$ is a function from R_Y into $\Re$. This is similar to the decision rules of the previous chapter. In hypothesis testing, a decision rule $D(y)$ mapped observations $Y = y$ into a discrete choice of hypothesis; the choice of decision rule depended on which criteria was used to design that decision rule $D(y)$. We will follow similar approaches for designing the estimation rule.

An important random variable in estimation is the estimation error $X - \widehat{x}(Y)$. This error is a random variable that depends on both X and Y. An estimator $\widehat{x}(y)$ is called **unbiased** in the Bayesian sense (or simply unbiased in the rest of this chapter) if

$$\mathbb{E}[X - \widehat{x}(Y)] = 0.$$

This implies that the error $X - \widehat{x}(Y)$ is an orthogonal random variable to the constant random variable 1. The bias of an estimator is known as $B = \mathbb{E}[X - \widehat{x}(Y)]$.

We note that there is a different concept of unbiased estimator in statistics, where X is not viewed as a random variable, but instead as an unknown constant. In statistics, an estimator $\widehat{x}(y)$ is called **unbiased** if

$$\mathbb{E}[\widehat{x}(Y)|X] = X \text{ for all values of } X.$$

This is a stronger requirement for unbiased estimation. In the remainder of this chapter we use the weaker concept of unbiased estimator in the Bayesian sense.

Another important statistic of an estimator is its mean-square error. The **mean-square error (MSE)** of an estimator is $\text{MSE} = \mathbb{E}\big[(X - \widehat{x}(Y))^2\big]$. We will use these statistics to design and characterize the performance of estimation rules.

7.2 Maximum Likelihood and Maximum A Posteriori Estimation

As was the case for hypothesis testing problems, we refer to the conditional distributions $P_{Y|X}(y|x)$ as likelihoods, because they are probability mass functions over Y, but they are general functions of $X = x$. Since we observe the value of $Y = y$, we are more interested in the properties of $P_{Y|X}(y|x)$ as functions of x, hence we use the term likelihood. For continuous random variables Y, the same applies to $f_{Y|X}(y|x)$, which are densities over Y, but general functions over $X = x$.

We define a **maximum likelihood estimator (ML)** $\widehat{x}_{ML}(y)$ as follows:

$$\widehat{x}_{ML}(y) \in \operatorname*{argmax}_{x \in R_X} P_{Y|X}(y|x), \quad Y \text{ discrete},$$
$$\widehat{x}_{ML}(y) \in \operatorname*{argmax}_{x \in R_X} f_{Y|X}(y|x), \quad Y \text{ continuous},$$

Since it is possible that the likelihood functions have multiple global maxima as a function of x, we use the set notation above to indicate that the ML estimator selects one of the global maxima of the likelihood functions.

The maximum likelihood estimator selects a value of $x \in R_X$ that maximizes the likelihood that the observation $Y = y$ was obtained, hence its name. It is similar to the maximum likelihood decision rule for hypothesis testing. The main difference is that, in binary hypothesis testing, selecting the maximum of two numbers is a straightforward operation. In contrast, selecting the maximum of a continuum of numbers (in case X is continuous) requires the use of optimization techniques involving calculus.

Similar to the ML estimator, we define a **maximum a posteriori (MAP)** estimator $\widehat{x}_{MAP}(y)$ as follows:

$$\begin{aligned}
\widehat{x}_{MAP}(y) &\in \operatorname*{argmax}_{x\in R_X} P(x)P_{Y|X}(y|x), \quad X,Y \text{ discrete},\\
\widehat{x}_{MAP}(y) &\in \operatorname*{argmax}_{x\in R_X} f(x)f_{Y|X}(y|x), \quad X,Y \text{ jointly continuous},\\
\widehat{x}_{MAP}(y) &\in \operatorname*{argmax}_{x\in R_X} f(x)P_{Y|X}(y|x), \quad Y \text{ discrete, } X \text{ continuous},\\
\widehat{x}_{MAP}(y) &\in \operatorname*{argmax}_{x\in R_X} P(x)f_{Y|X}(y|x), \quad Y \text{ continuous, } X \text{ discrete},
\end{aligned}$$

where we have added some extra cases to allow for the possibility that one of X, Y is discrete, while the other is continuous. For instance, in speech recognition, the features of a sound we hear (Y) have continuous values, but the set of possible phonemes (X) that generate that sound is discrete (around 64 phonemes in English). Similarly, in digital communications decoding, the transmitted symbols X are discrete, but the measured waveforms Y are continuous.

Example 7.1

The number of customers arriving at a service station when they open in the morning is modeled as a Binomial(4,0.5) random variable. No other customers arrive that day. Given that X customers arrive, the time Y hours to service their requests is modeled as an exponential random variable with parameter $\lambda(X) = \frac{1}{1+X}$. We come in the next day, and observe that $Y = 2$ for the previous day. We want to estimate the actual number of customers that arrived the previous day.

Note that this is a problem with continuous-valued measurements Y but discrete unknown X. Let's find both the ML and the MAP estimate of X given $Y = 2$. From the problem description, we know that $R_X = \{0, 1, 2, 3, 4\}$, and the likelihood function is known from the properties of exponential random variables, as

$$f_{Y|X}(y|x) = \begin{cases} \frac{1}{1+x}e^{-\frac{y}{1+x}} & y \geq 0 \\ 0 & \text{elsewhere.} \end{cases}$$

For $Y = 2$, the ML estimator will be

$$\widehat{x}_{ML}(y) \in \operatorname*{argmax}_{x\in\{0,1,2,3,4\}} f_{Y|X}(2|x) = \frac{1}{1+x}e^{-\frac{2}{1+x}}.$$

To find the maximum, we enumerate the values for each x :

x :	0	1	2	3	4
	0.135	0.184	0.171	0.152	0.134

Based on these numbers, $\widehat{x}_{ML}(2) = 1$.

From the problem description, we know $P_X(x) = \binom{4}{x}0.5^4$. Hence, the MAP estimator is

$$\widehat{x}_{MAP}(y) \in \operatorname*{argmax}_{x\in\{0,1,2,3,4\}} P_X(x)f_{Y|X}(2|x) = \binom{4}{x}0.5^4\frac{1}{1+x}e^{-\frac{2}{1+x}}.$$

To find the maximum, we enumerate the values for each x :

x :	0	1	2	3	4
	0.008	0.046	0.064	0.038	0.008

and we get that $\widehat{x}_{MAP}(2) = 2$. The difference arises because the prior probability that $X = 2$ is higher than that of $X = 1$.

Could we find the form of the estimators for arbitrary measurements Y? We do this for the ML estimator. In this case, it is possible, since all X does is decrease the service rate as X increases. Thus, for small Y, the best estimate is likely to be $X = 0$, and for large Y, it will be $X = 4$. The plot of the different densities $f_{Y|X}(y|x)$ is show in Figure 7.2 below. The figure illustrates when the different curves are maximal, and we can find those intervals by solving for the intersection points of the curves.

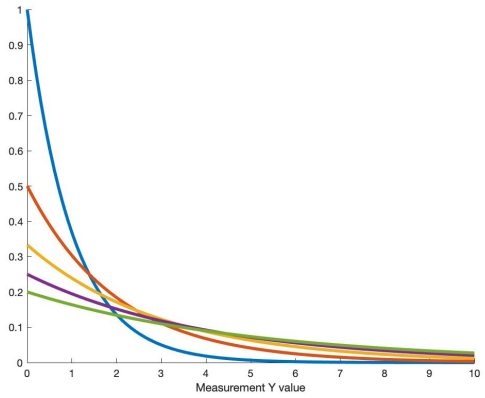


Figure 7.2: Plots of the different densities for different values of X.

We need to find the value of y where the likelihood function $f_{Y|X}(y|0) = f_{Y|X}(y|1)$, which is

$$e^{-y} = \frac{1}{2}e^{-y/2} \iff e^{y/2} = 2 \iff y = 2\ln(2).$$

Similarly, the value of y where the likelihood function $f_{Y|X}(y|1) = f_{Y|X}(y|2)$ is

$$\frac{1}{2}e^{-y/2} = \frac{1}{3}e^{-y/3} \iff e^{y/6} = \frac{3}{2} \iff y = 6\ln(\frac{3}{2}).$$

Similarly, the value of y where the likelihood function $f_{Y|X}(y|2) = f_{Y|X}(y|3)$ is

$$\frac{1}{3}e^{-y/3} = \frac{1}{4}e^{-y/4} \iff e^{y/12} = \frac{4}{3} \iff y = 12\ln(\frac{4}{3}).$$

Finally, the value of y where the likelihood function $f_{Y|X}(y|3) = f_{Y|X}(y|4)$ is

$$\frac{1}{4}e^{-y/4} = \frac{1}{5}e^{-y/5} \iff e^{y/20} = \frac{5}{4} \iff y = 20\ln(\frac{4}{3}).$$

Hence, the ML estimator is

$$\widehat{x}_{ML}(y) = \begin{cases} 0, & y \in [0, 2\ln(2)], \\ 1, & y \in (2\ln(2), 6\ln(3/2)], \\ 2, & y \in (6\ln(3/2), 12\ln(4/3)], \\ 3, & y \in (12\ln(3/4)], 20\ln(5/4)] \\ 4, & y > 20\ln(5/4). \end{cases}$$

Example 7.2
One of the most useful applications of estimation is in estimating the parameters of an unknown probability distribution from observed samples. Let X be a random variable in [0,1], with density $f_X(x) = 2x, x \in [0,1]; 0$ otherwise. Given $X = x$, let Y be a Binomial(N, x) random variable. This corresponds to the following scenario. We have a coin with unknown probability of heads X, as a number between 0 and 1. To estimate X, we flip this coin N times and count the number of heads (Y). Now we want to estimate the original unknown probability X given the number of heads observed (Y out of N).

A quick estimator might be the fraction of heads: $\widehat{x}(y) = \frac{y}{N}$. What is the maximum likelihood estimator? From the problem description,

$$P_{Y|X}(y|x) = \binom{N}{y}(x)^y(1-x)^{N-y}.$$

$$\widehat{x}_{ML}(y) = \underset{x\in[0,1]}{\operatorname{argmax}} \binom{N}{y}(x)^y(1-x)^{N-y}.$$

To simplify this, we maximize the log-likelihood, which has the maximum in the same locations as the likelihood, because the logarithm is a monotone increasing function for positive numbers (e.g. likelihoods).

$$\widehat{x}_{ML}(y) = \underset{x\in[0,1]}{\operatorname{argmax}} \ln(\binom{N}{y}) + y\ln(x) + (N-y)\ln(1-x).$$

To maximize, take the derivative with respect to x and set it equal to 0, as

$$\frac{d}{dx}\Big(\ln(\binom{N}{y}) + y\ln(x) + (N-y)\ln(1-x)\Big) = \frac{y}{x} - \frac{N-y}{1-x} = 0.$$

Solving for x will give us the estimator $\widehat{x}_{ML}(y) = \frac{y}{N}$, which can be verified by substituting into the above equation and checking it solves it. Thus, the ML estimator is the fraction of heads out of N trials.

What is the MAP estimator?

$$\widehat{x}_{MAP}(y) = \underset{x\in[0,1]}{\operatorname{argmax}}\, 2x\binom{N}{y}(x)^y(1-x)^{N-y}.$$

Taking logarithms,

$$\widehat{x}_{MAP}(y) = \underset{x\in[0,1]}{\operatorname{argmax}}\, \ln(2) + \ln(x) + \ln(\binom{N}{y}) + y\ln(x) + (N-y)\ln(1-x).$$

Differentiating with respect to x:

$$\frac{d}{dx}\Big(\ln(2) + \ln(x) + \ln(\binom{N}{y}) + y\ln(x) + (N-y)\ln(1-x).\Big) = \frac{y+1}{x} - \frac{N-y}{1-x} = 0.$$

Solving,

$$(y+1)(1-x) - (N-y)x = 0 \iff y+1-x-Nx = 0 \iff x = \frac{y+1}{N+1},$$

so $\widehat{x}_{MAP}(y) = \frac{y+1}{N+1}$, which is a little larger than the ML estimator.

Are these estimates unbiased? Note the following:

$$\mathbb{E}[Y] = \mathbb{E}\big[\mathbb{E}[Y|X]\big] = \mathbb{E}[NX] = N\mathbb{E}[X],$$

because Y is distributed as a Binomial(N, X) random variable. Then, for the ML estimator,

$$\mathbb{E}[X - \frac{Y}{N}] = \mathbb{E}[X] - \frac{\mathbb{E}[Y]}{N} = \mathbb{E}[X] - \mathbb{E}[X] = 0,$$

so the ML estimator is unbiased. For the MAP estimator,

$$\mathbb{E}[X - \frac{Y+1}{N+1}] = \mathbb{E}[X] - \frac{N\mathbb{E}[X]+1}{N+1} = \frac{E[X]-1}{N+1} \neq 0,$$

so the MAP estimator is biased.

Example 7.3
We have a receiver, at a distance X meters from a transmitter. The transmitter transmits a signal with power 100, and the signal decays as $\frac{1}{X^2}$ to reach the receiver so the nominal received signal $S = \frac{100}{X^2}$. The receiver measures the signal strength in decibels, and the signal in decibels has some noise in it. The measured signal Y is given as

$$Y = 40 - 40\log_{10}(X) + W$$

where W is a Gaussian random variable with mean 0 and variance 4, independent of X. The prior distribution of X is

$$f_X(x) = \begin{cases} \frac{2x}{10^6} & 0 < x \leq 10^3, \\ 0 & \text{otherwise.} \end{cases}$$

Compute the ML and MAP estimators of X given observations $Y = y$.

From the problem description,

$$f_{Y|X}(y|x) = \frac{1}{2\sqrt{2\pi}} e^{-\frac{(y+40\log_{10}(x))^2}{8}}.$$

The ML estimator is

$$\widehat{x}_{ML}(y) = \underset{x\in(0,1000]}{\operatorname{argmax}}\, \frac{1}{2\sqrt{2\pi}} e^{-\frac{(y-40+40\log_{10}(x))^2}{8}}.$$

As was the case with detection problems, it is often easier to maximize the log-likelihood, since the maximum of the log-likelihood is in the same location as the maximum of the likelihood. Thus,

$$\widehat{x}_{ML}(y) = \underset{x\in(0,1000]}{\operatorname{argmax}} \ln\Big(\frac{1}{2\sqrt{2\pi}} e^{-\frac{(y-40+40\log_{10}(x))^2}{8}}\Big) = C - \frac{(y-40+40\log_{10}(x))^2}{8},$$

where C is a constant that does not depend on x, and so it won't affect the location of the maximizing x.

We try to find the maximum by differentiating and setting the derivative equal to 0:

$$\frac{d}{dx}(y-40+40\log_{10}(x))^2 = 2((y-40+40\log_{10}(x))\frac{40}{x\ln(10)} = 0.$$

Eliminating constants yields the following equation:

$$(y-40+40\log_{10}(x)) = 0 \iff x = 10^{\frac{40-y}{40}}.$$

Note that this is not the ML estimator yet, because it is possible that this value of X is greater than 1000. If it is, the best estimate is to set $x = 1000$. The ML estimator is thus

$$\widehat{x}_{ML}(y) = \begin{cases} 10^{\frac{40-y}{40}}, & y >= -80, \\ 10^3 & y < -80. \end{cases}$$

What about the MAP estimator? It is

$$\widehat{x}_{MAP}(y) = \underset{x\in(0,1000]}{\operatorname{argmax}} \frac{2x}{10^6}\frac{1}{2\sqrt{2\pi}} e^{-\frac{(y-40+40\log_{10}(x))^2}{8}}.$$

Taking logarithms as before yields

$$\widehat{x}_{MAP}(y) = \underset{x\in(0,1000]}{\operatorname{argmax}} \ln\Big(\frac{2x}{10^6}\frac{1}{2\sqrt{2\pi}} e^{-\frac{(y-40+40\log_{10}(x))^2}{8}}\Big) = C_1 - \frac{(y-40+40\log_{10}(x))^2}{8} + \ln(x).$$

Differentiating with respect to x, multiplying by -1 and setting it to 0 yields

$$2((y-40+40\log_{10}(x))\frac{40}{x\ln(10)} - \frac{1}{x} = 0 \iff y-40+40\log_{10}(x) - \frac{\ln(10)}{80} = 0.$$

We see the effect of the a priori information on the MAP estimator. It increases the estimated distance. The ML estimator assumes that X is uniformly distributed in $(0, 10^3)$ with a density that does not depend on X. The MAP estimator has more probability for larger values of X. Thus,

$$\widehat{x}_{MAP}(y) = \begin{cases} 10^{\frac{40+\frac{\ln(10)}{80}-y}{40}}, & y >= -80 + \frac{\ln(10)}{80}, \\ 10^3 & \text{elsewhere.} \end{cases}$$

Example 7.4
Assume that X, Y are joint Gaussian random variables, with means μ_X, μ_Y and variances σ_X^2, σ_Y^2 respectively, and with given covariance $\mathrm{Cov}[X,Y]$. Using the results of Chapter 5.4, we know that we can write the joint density of X, Y as

$$f_{X,Y}(x,y) = f_{X|Y}(x|y)f_Y(y)$$

where the conditional density $f_{X|Y}(x|y)$ is Gaussian, with mean $\mu_X + \frac{\mathrm{Cov}[X,Y]}{\mathrm{Var}[Y]}(y-\mu_Y)$, and variance $\sigma_X^2 - \frac{\mathrm{Cov}[X,Y]^2}{\sigma_Y^2}$.

Then, the MAP estimator of X given Y is

$$\widehat{x}_{MAP}(y) \in \arg\max_{x\in(-\infty,\infty)} f_{X|Y}(x|y) = \mu_X + \frac{\mathrm{Cov}[X,Y]}{\mathrm{Var}[Y]}(y-\mu_Y),$$

because the maximum of a Gaussian density is at its mean. Hence, in this Gaussian case,

$$\widehat{x}_{MAP}(y) = \mathbb{E}[X \mid Y = y].$$

Let us close this section by summarizing what we have learned about ML and MAP estimates estimates:

- ML estimates are equal to MAP estimates when the marginal density or distribution of the unobserved variable X does not depend on x.
- The ML estimate is the maximizing value for the likelihood: $\text{argmax}_x\, f_{Y|X}(y \mid x)$.
- The MAP estimate is the conditional mode: $\text{argmax}_x\, f_{X|Y}(x|y)$.
- The MAP and ML estimates may be biased.
- The MAP and ML estimates may not be unique.
- In general, the MAP and ML estimates are nonlinear functions of the observation.
- For jointly Gaussian problems, the MAP estimate is the same as the conditional mean, and in this case is a linear estimate and MMSE.
- In general, finding the MAP or ML estimate requires finding the maximum of the conditional density or likelihood, which may be a difficult problem.

7.3 Minimum Mean Square Error Estimation

For any estimator $\widehat{x}(y)$, the error $X - \widehat{x}(Y)$ is a random variable. The mean square error is MSE $= \mathbb{E}[(X-\widehat{x}(Y))^2]$. We want to find the estimator $\widehat{x}(y)$ that results in the minimum mean-square error (MMSE). We refer to this estimator as $\widehat{x}_{MMSE}(y)$. As before, assume we are given either a PMF $P_X(x)$ or PDF $f_X(x)$, depending on whether X is discrete or continuous as a random variable. Furthermore, assume we know the likelihoods $P_{Y|X}(y|x)$ or $f_{Y|X}(y|x)$, depending on whether Y is discrete or continuous.

Consider first the special case where Y is discrete and $P_{Y|X}(0|x) = 1$ for all $x \in R_X$. In simple words, the measurement $Y = 0$ always happen, no matter what X is. In this case, any estimator must be a constant: $\widehat{x}(0) = a$. What is the best choice of constant a to minimize the mean square error? Define $\mu_X = \mathbb{E}[X]$. Consider the following identity:

$$\begin{aligned}\mathbb{E}[(X-a)^2] &= \mathbb{E}[(X-\mu_X+\mu_X-a)^2] = \mathbb{E}[(X-\mu_X)^2 + 2(X-\mu_X)(\mu_X-a) + (\mu_X-a)^2] \\ &= \mathbb{E}[(X-\mu_X)^2] + 2\mathbb{E}[(X-\mu_X)](\mu_X-a) + (\mu_X-a)^2 \\ &= \mathsf{Var}[X] + (\mu_X-a)^2\end{aligned}$$

Since the last term is non-negative, and is zero when $a = \mu_X$, this means that $\widehat{x}_{MMSE}(0) = \mu_X = \mathbb{E}[X]$.

Let's now derive the MMSE estimator for general forms of likelihood. We will use the Law of Total Expectation, with iterated expectations, as follows:

$$\mathbb{E}[(X-\widehat{x}(Y))^2] = \mathbb{E}\Big[\mathbb{E}[(X-\widehat{x}(Y))^2|Y]\Big]$$

Let's focus on the inner expectation:

$$\begin{aligned}\mathbb{E}[(X-\widehat{x}(Y))^2|Y] &= \mathbb{E}[X^2 - 2X\widehat{x}(Y) + (\widehat{x}(Y))^2|Y] \\ &= \mathbb{E}[X^2|Y] - 2\mathbb{E}[X|Y]\widehat{x}(Y) + (\widehat{x}(Y))^2 \quad \text{[functions of } Y \text{ are conditionally constant]} \\ &= \mathbb{E}[X^2|Y] - (\mathbb{E}[X|Y])^2 + (\mathbb{E}[X|Y])^2 - 2\mathbb{E}[X|Y]\widehat{x}(Y) + (\widehat{x}(Y))^2 \quad \text{[add, subtract same]} \\ &= \mathsf{Var}[X|Y] + (\mathbb{E}[X|Y] - \widehat{x}(Y))^2 \quad \text{[factor into a square]}\end{aligned}$$

The last term is non-negative, and is zero only when $\widehat{x}(Y) = \mathbb{E}[X|Y]$. Any other estimator will have a larger conditional mean square error, and thus will also have a larger unconditional MSE. Thus, the MMSE estimator is

$$\widehat{x}_{MMSE}(y) = \mathbb{E}[X|Y = y].$$

Note that our derivation of this did not depend on whether X or Y was continuous or discrete. This is a strong result: the estimator that results in the smallest mean square error is the conditional mean of X given observation of Y. Note also that, when X is discrete, $\widehat{x}_{MMSE}(y)$ is a real number, and may not belong to R_X.

The MMSE estimator has some interesting properties, discussed below:

- The MMSE estimator is unbiased. That is, $\mathbb{E}\big[X - \mathbb{E}[X|Y]\big] = 0$. This follows from the Law of Total Expectation, because

$$\mathbb{E}\Big[X - \mathbb{E}[X|Y]\Big] = \mathbb{E}[X] - \mathbb{E}\Big[\mathbb{E}[X|Y]\Big] = \mathbb{E}[X] - \mathbb{E}[X] = 0.$$

- The error $X - \mathbb{E}[X|Y]$ is orthogonal to any random variable $Z = g(Y)$, for any bounded function Y. Again, this is a function of the Law of Total Expectation,(as

$$\mathbb{E}[(X - \mathbb{E}[X|Y])g(Y)] = \mathbb{E}\Big[\mathbb{E}[(X - \mathbb{E}[X|Y])g(Y) \mid Y]\Big] = \mathbb{E}\Big[\mathbb{E}[X - \mathbb{E}[X|Y] \mid Y]g(Y)\Big],$$

  because $g(Y)$ is known if Y is observed, Then, $\mathbb{E}[X - \mathbb{E}[X|Y] \mid Y] = \mathbb{E}[X - \mathbb{E}[X|Y]|Y] = \mathbb{E}[X|Y] - \mathbb{E}[X|Y] = 0$. Substituting into the above equation yields the orthogonality property.

- The estimator $\widehat{x}_{MMSE}(Y)$ is orthogonal to the error $X - \mathbb{E}[X|Y]$, by the above property, because it is a function of Y.

The main limitation in computing the MMSE estimator is that one needs to compute the expected value of X given $Y = y$. This can be hard to do. Furthermore, for discrete X, the MMSE estimate may not be in the range R_X. We revisit a couple of our earlier examples to illustrate this.

Example 7.5
Consider Example 7.2, where

$$P_X(x) = \binom{4}{x} 0.5^4,$$

$$f_{Y|X}(y|x) = \begin{cases} \frac{1}{1+x} e^{-\frac{y}{1+x}} & y \geq 0 \\ 0 & \text{elsewhere.} \end{cases}$$

Then, using Bayes' Rule, we get

$$P_{X|Y}(x|y) = \frac{\frac{1}{1+x} e^{-\frac{y}{1+x}} \binom{4}{x} 0.5^4}{f_Y(y)}$$

where $f_Y(y) = \sum_{x=0}^{4} \frac{1}{1+x} e^{-\frac{y}{1+x}} \binom{4}{x} 0.5^4$. Thus,

$$\begin{aligned} \widehat{x}_{MMSE}(y) = \mathbb{E}[X|Y=y] &= \frac{1}{f_Y(y)} \sum_{x=0}^{4} \frac{x}{1+x} e^{-\frac{y}{1+x}} \binom{4}{x} 0.5^4 \\ &= \frac{\frac{1}{2}\binom{4}{1}e^{-y/2} + \frac{2}{3}\binom{4}{2}e^{-y/3} + \frac{3}{4}\binom{4}{3}e^{-y/4} + \frac{4}{5}\binom{4}{4}e^{-y/5}}{e^{-y} + \frac{1}{2}\binom{4}{1}e^{-y/2} + \frac{1}{3}\binom{4}{2}e^{-y/3} + \frac{1}{4}\binom{4}{3}e^{-y/4} + \frac{1}{5}\binom{4}{4}e^{-y/5}}. \end{aligned}$$

Note the complexity of the MMSE estimator. Fortunately, since X only takes five possible values, we are able to write the terms in each sum. For the observation $Y = 2$, the MMSE estimator is $\widehat{x}_{MMSE}(2) \approx 1.95$, which is different from $\widehat{x}_{ML}(2) = 1$ and $\widehat{x}_{MP}(2) = 2$. this also highlights that the MMSE estimate of a discrete random variable can be a real number, whereas the MAP and ML estimates are restricted to elements in R_X, which are integers.

Example 7.6
Consider Example 7.3, where

$$f_X(x) = \begin{cases} \frac{2x}{10^6} & 0 < x \leq 10^3, \\ 0 & \text{otherwise.} \end{cases}$$

$$f_{Y|X}(y|x) = \frac{1}{2\sqrt{2\pi}} e^{-\frac{(y-40+40\log_{10}(x))^2}{8}}.$$

We use Bayes' Rule to compute:

$$f_{X|Y}(x|y) = \begin{cases} \frac{\frac{2x}{10^6}\frac{1}{2\sqrt{2\pi}}e^{-\frac{(y+40\log_{10}(x))^2}{8}}}{f_Y(y)} & 0 < x \le 10^3, \\ 0 & \text{otherwise.} \end{cases}$$

where

$$f_Y(y) = \int_0^{1000} \frac{2x}{10^6}\frac{1}{2\sqrt{2\pi}} e^{-\frac{(y-40+40\log_{10}(x))^2}{8}}\, dx$$

Then,

$$\widehat{x}_{MMSE}(y) = \frac{\int_0^{1000} x\frac{2x}{10^6}\frac{1}{2\sqrt{2\pi}} e^{-\frac{(y-40+40\log_{10}(x))^2}{8}}\, dx}{\int_0^{1000} \frac{2x}{10^6}\frac{1}{2\sqrt{2\pi}} e^{-\frac{(y-40+40\log_{10}(x))^2}{8}}\, dx}.$$

Although we can write the integrals for computing $\widehat{x}_{MMSE}(y)$, these integrals are hard to compute exactly, illustrating the computational complexity of computing the MMSE estimators. In contrast, the MAP and ML estimators were easy to compute in closed form.

Example 7.7

Here we consider a simple example of a joint density where we can compute the conditional expected value needed. Let X, Y be jointly continuous random variables with joint PDF given by

$$f_{X,Y}(x,y) = \begin{cases} 2(x+y) & 0 \le x \le y \le 1 \\ 0 & \text{otherwise.} \end{cases}$$

Then,

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\, dx = \begin{cases} \int_0^y 2(x+y)\, dx = 3y^2, & y \in [0,1] \\ 0 & \text{otherwise,} \end{cases}$$

and the conditional density is

$$f_{X,Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \begin{cases} \frac{2(x+y)}{3y^2} & 0 \le x \le y \le 1 \\ 0 & \text{otherwise.} \end{cases}$$

Hence,

$$\widehat{x}_{MMSE}(y) = \mathbb{E}[X|Y=y] = \frac{1}{3y^2}\int_0^y x2(x+y)\, dx = \frac{1}{3y^2}(\frac{2}{3}y^3 + y^3) = \frac{5}{9}y.$$

In this case, the integrals involved simple polynomials, so we could compute the needed expectations. The MMSE estimator turns out to be a linear function of y. We can also compute the MAP estimator as $\widehat{x}_{MAP}(y) \in \operatorname{argmax}_{x\in[0,y]} f_{X,Y}(x,y) = \operatorname{argmax}_{x\in[0,y]} 2(x+y)$. Hence, $\widehat{x}_{MAP}(y) = y$ for all $y \in [0,1]$. This illustrates how biased the MAP estimator can be.

When the estimator takes on this simple form, we can compute the mean square error exactly. In this case, the error is $X - \frac{5}{9}Y$. Then,

$$\mathbb{E}[Y] = \int_0^1 y(3y^2)\, dy = \frac{3}{4}; \qquad f_X(x) = \begin{cases} \int_x^1 2(x+y)\, dy = 2(x)(1-x) + 1 - x^2 = 1 + 2x - 3x^2 & x \in [0,1] \\ 0 & \text{otherwise.} \end{cases}$$

$$\mathbb{E}[X] = \int_0^1 (x + 2x^2 - 3x^3)\, dx = \frac{1}{2} + \frac{2}{3} - \frac{3}{4} = \frac{5}{12},$$

$$\mathbb{E}[X - \frac{5}{9}Y] = \mathbb{E}[X] - \frac{5}{9}\mathbb{E}[Y] = \frac{5}{12} - \frac{5}{9}\cdot\frac{3}{4} = 0;$$

which we expected because the MMSE estimator is unbiased. To compute the MMSE, we need to compute some additional expectations:

$$\mathbb{E}[Y^2] = \int_0^1 y^2(3y^2)\, dy = \frac{3}{5}; \quad \mathbb{E}[X^2] = \int_0^1 x^2(1 + 2x - 3x^2)\, dx = \frac{1}{3} + \frac{1}{2} - \frac{3}{5} = \frac{7}{30}$$

$$\mathbb{E}[XY] = \int_0^1 \Big(\int_0^y xy2(x+y)\, dx\Big)\, dy = \int_0^1 \frac{5}{3}y^4\, dy = \frac{1}{3}$$

$$\begin{aligned} \text{MMSE} &= \mathbb{E}[(X - \frac{5}{9}Y)^2] = \mathbb{E}[X^2] - \frac{10}{9}\mathbb{E}[XY] + \frac{25}{81}\mathbb{E}[Y^2] \\ &= \frac{7}{30} - \frac{10}{27} + \frac{5}{27} = \frac{13}{270} \end{aligned}$$

It is useful to compare this to the MSE of the MAP estimator, which is

$$MSE_{MAP} = \mathbb{E}[(X-Y)^2] = \mathbb{E}[X^2] - 2\mathbb{E}[XY] + \mathbb{E}[Y^2] = \frac{3}{5} + \frac{7}{30} - 2\frac{1}{3} = \frac{5}{30} = \frac{1}{6},$$

which is much larger than the MMSE.

There is one special case where the MMSE is easy to compute: the case where X, Y are joint Gaussian random variables. In this case, the conditional expected value of X given Y was derived in Chapter 5.4 as

$$\widehat{x}_{MMSE}(y) = \mathbb{E}[X|Y=y] = \mathbb{E}[X] + \frac{\mathsf{Cov}[X,Y]}{\mathsf{Var}[Y]}(y - \mathbb{E}[Y]),$$

with mean square error given by

$$\mathbb{E}\left[(X - \mathbb{E}[X|Y])^2\right] = \mathsf{Var}[X] - \frac{\mathsf{Cov}[X,Y]^2}{\mathsf{Var}[Y]}.$$

In this case, both the MMSE estimator and the minimum mean square error are given in terms of the first- and second-order statistics of X, Y, so no complex integrals need to be computed.

Example 7.8

We measure distance using acoustic echoes by measuring the travel time of a sound pulse. Let X be the unknown distance to an object, which we model as a Gaussian random variable with mean 10^3 meters and standard deviation 100 meters. Assuming the speed of sound is 300 meters/second (to simplify computation), the round trip time of a pulse from a sensor to the object is $2X/300$ seconds. However, our cheap timer is not perfectly accurate, so we model the measurement Y as

$$Y = \frac{1}{150}X + W$$

where W is independent of X, Gaussian, with zero-mean and standard deviation 0.2 seconds.

From the above discussion, we can easily compute the first- and second-order statistics of X, Y as follows:

$$\begin{aligned}
\mathbb{E}[X] &= 1000; \qquad \mathsf{Var}[X] = 10,000; \\
\mathbb{E}[Y] &= \frac{1}{150}\mathbb{E}[X] + \mathbb{E}[W] = \frac{1000}{150} = \frac{20}{3}; \\
\mathsf{Var}[Y] &= \frac{1}{(150)^2}\mathsf{Var}[X] + \mathsf{Var}[W] = \frac{10000}{22500} + \frac{1}{25} = \frac{4}{9} + \frac{1}{25} = \frac{109}{225} \quad X, W \text{ are uncorrelated.} \\
\mathsf{Cov}[X,Y] &= \mathsf{Cov}[X, \frac{1}{150}X + W] = \frac{1}{150}\mathsf{Var}[X] + \mathsf{Cov}[X,W] = \frac{1}{150}\mathsf{Var}[X] = \frac{10000}{150} = \frac{200}{3}.
\end{aligned}$$

Then, the MMSE estimator of the distance X given the measurement $Y = y$ is

$$\widehat{x}_{MMSE}(y) = 1000 + \frac{\frac{200}{3}}{\frac{109}{225}}(y - \frac{20}{3}) = 1000 + \frac{15000}{109}(y - \frac{20}{3}),$$

with MMSE given by

$$MMSE = 10000 - \frac{\frac{40000}{9}}{\frac{109}{225}} = 10000 - \frac{1,000,000}{109} \approx 826 \text{ meters}^2.$$

Thus, the measurement cut down the standard deviation of the location to under 30 meters.

Example 7.9

In this example we wish to estimate X by observing a related random variable Y, where the random variables X and Y are jointly distributed with the density shown in Figure 7.3. This density is uniform over the depicted diamond shaped region. Note that this characterization provides all the information to find both a prior model for X (i.e. the marginal distribution $f_X(x)$) as well as the relationship between Y and X as given by $f_{Y|X}(y \mid x)$.

To find the MMSE estimate for this problem the quantity we need to find is the conditional density $f_{X|Y}(x \mid y)$. We find $f_{X|Y}(x \mid y)$ almost by inspection: restrict the joint density to the slice $Y = y$, and rescale so it normalizes to a probability. Recall that $f_{X|Y}(x \mid y)$ will be a slice of the joint density $f_{X,Y}(x,y)$ parallel to the x-axis and scaled to have unit area. This conditional density is shown on the right in Figure 7.3 for any nontrivial value of y. Since the original density is "flat," each slice will be flat, so all we really need to determine are the edges. The height follows from the constraint that the density has unit area.

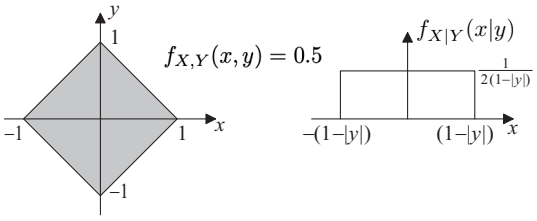


Figure 7.3: MMSE Example

Now, given this density it is easy to see that $\widehat{x}_{MMSE}(y) = \mathbb{E}[x \mid y] = 0$. In this case, the MSE is

$$MSE = E[X^2] = \mathsf{Var}[X] = 4\int_0^1 \left(\int_0^{1-x} x^2 \frac{1}{2}\, dy\right) dx = 2\int_0^1 (x^2 - x^3)\, dx = \frac{1}{6}.$$

Example 7.10
Suppose X and Y are related by the following joint density function:

$$f_{X,Y}(x,y) = \begin{cases} 10x & 0 \le x \le y^2,\ 0 \le y \le 1 \\ 0 & \text{otherwise.} \end{cases}$$

To find the MMSE estimate for this problem we need to find the conditional density $f_{X|Y}(x \mid y)$. By integrating $f_{X,Y}(x,y)$ with respect to y we can find the marginal density for y:

$$f_Y(y) = \begin{cases} \int_0^{y^2} 10x\, dx = 5y^4, 0 \le y \le 1, \\ 0 \qquad\qquad\qquad\qquad\qquad \text{otherwise.} \end{cases}$$

Now we can use Bayes' Rule to obtain the conditional density:

$$f_{X|Y}(x \mid y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \begin{cases} \frac{10x}{5y^4} & 0 \le x \le y^2 \le 1 \\ 0 & \text{otherwise.} \end{cases}$$

The mean of the conditional density is now found as:

$$\mathbb{E}[X \mid Y = y] = \int_0^{y^2} \frac{2x^2}{y^4}\, dx = \frac{2}{3}y^2.$$

Thus $\widehat{x}_{MMSE}(y) = \frac{2}{3}y^2$. Note that this estimate is a *nonlinear* function of y in this case.

Next let us find the conditional variance $\mathsf{Var}[X|Y = y]$.

$$\mathbb{E}[X^2|Y = y] = \int_0^{y^2} \frac{2x^3}{y^4}\, dx = \frac{1}{2}y^4.$$

$$\mathsf{Var}[X|Y = y] = \mathbb{E}[X^2|Y = y] - (\mathbb{E}[X|Y = y])^2 = \frac{1}{18}y^4.$$

Finally, the minimum mean square error is obtained as:

$$\mathsf{MMSE} = \mathbb{E}\Big[\mathsf{Var}[X|Y]\Big] = \int_0^1 \frac{1}{18}y^4 \cdot 5y^4\, dy = \frac{5}{162}.$$

Let us close by summarizing the properties of MMSE estimators.

- The MMSE estimator is the conditional mean $E[X \mid Y]$.
- The MMSE estimator is always unbiased.
- The MMSE estimator error is orthogonal to any random variable that is a function of the observation Y.

- In general, the MMSE estimator is a nonlinear function of the observation Y, and can be hard to compute.
- For jointly Gaussian problems only, the MMSE estimator is linear in Y and the conditional variance is independent of the observation Y.

7.4 Linear Least Squares Estimation

As noted in the previous section, the MMSE estimator $\mathbb{E}[X|Y]$ is often a complex nonlinear function that is hard to compute. To find a simpler estimator, we want to restrict the estimators to have a restricted functional form. In linear least squares estimation, we restrict the estimator to be of the form $\widehat{x}(y) = ay + b$, for some constants a, b. The linear least squares estimator (LLSE) is the estimator of this form that minimizes the mean square error $\mathbb{E}[(X - aY - b)^2]$. For estimators of this form,

$$\begin{aligned}\mathbb{E}[(X-aY-b)^2] &= \mathbb{E}[(X^2 + a^2Y^2 + b^2 - 2aXY - 2bX + 2abY)] \\ &= \mathbb{E}[X^2] + a^2\mathbb{E}[Y^2] + b^2 - 2a\mathbb{E}[X,Y] - 2b\mathbb{E}[X] + 2ab\mathbb{E}[Y] \\ &= (\mathbb{E}[X])^2 + \mathsf{Var}[X] + a^2(\mathbb{E}[Y])^2 + a^2\mathsf{Var}[Y] + b^2 - 2a\mathsf{Cov}[X,Y] - 2a\mathbb{E}[X]\mathbb{E}[Y] - 2b\mathbb{E}[X] + 2ab\mathbb{E}[Y]\end{aligned}$$

We are going to manipulate this expression by adding and subtracting some terms to complete squares. This will help us identify the values of a, b that result in minimum mean square error. We highlight in red terms that we add and subtract to help us complete the squares.

$$\begin{aligned}\mathbb{E}[(X-aY-b)^2] &= (\mathbb{E}[X])^2 + \mathsf{Var}[X] + a^2(\mathbb{E}[Y])^2 + a^2\mathsf{Var}[Y] + b^2 \\ &\quad - 2a\mathsf{Cov}[X,Y] - 2a\mathbb{E}[X]\mathbb{E}[Y] - 2b\mathbb{E}[X] + 2ab\mathbb{E}[Y] \\ &= (b - \mathbb{E}[X] + a\mathbb{E}[Y])^2 + \mathsf{Var}[X] + a^2\mathsf{Var}[Y] - 2a\mathsf{Cov}[X,Y] \\ &= (b - \mathbb{E}[X] + a\mathbb{E}[Y])^2 + \mathsf{Var}[X] + a^2\mathsf{Var}[Y] - 2a\mathsf{Cov}[X,Y] + \frac{\mathsf{Cov}[X,Y]^2}{\mathsf{Var}[Y]} - \frac{\mathsf{Cov}[X,Y]^2}{\mathsf{Var}[Y]} \\ &= (b - \mathbb{E}[X] + a\mathbb{E}[Y])^2 + \mathsf{Var}[X] + \mathsf{Var}[Y]\Big(a - \frac{\mathsf{Cov}[X,Y]}{\mathsf{Var}[Y]}\Big)^2 - \frac{\mathsf{Cov}[X,Y]^2}{\mathsf{Var}[Y]}\end{aligned}$$

The values of a and b that minimize the mean square error are now obvious. Note that a, b are only present in the two quadratic terms in the right hand side of the equation. Those quadratic terms are non-negative, and are zero only when $b^* = \mathbb{E}[X] - a\mathbb{E}[Y]$ and $a^* = \frac{\mathsf{Cov}[X,Y]}{\mathsf{Var}[Y]}$. With these choices, we get the minimum mean square error for the linear estimator as

$$\mathbb{E}[(X - a^*Y - b^*)^2] = \mathsf{Var}[X] - \frac{\mathsf{Cov}[X,Y]^2}{\mathsf{Var}[Y]}.$$

The linear estimator that achieves this error is the LLSE estimator, given by:

$$\widehat{x}_{LLSE}(Y) = \mathbb{E}[X] - \frac{\mathsf{Cov}[X,Y]}{\mathsf{Var}[Y]}\mathbb{E}[Y] + \frac{\mathsf{Cov}[X,Y]}{\mathsf{Var}[Y]}Y = \mathbb{E}[X] + \frac{\mathsf{Cov}[X,Y]}{\mathsf{Var}[Y]}(Y - \mathbb{E}[Y]).$$

This estimator is easy to compute, as it only depends on the first- and second-order statistics of X, Y. It is not necessary to know the joint distributions of X, Y.

The LLSE estimator has several nice properties that we discuss next:

- The LLSE estimator $\widehat{x}_{LLSE}(Y)$ is unbiased, as can be seen from:

$$\mathbb{E}[X - \widehat{x}_{LLSE}(Y)] = \mathbb{E}\left[X - \mathbb{E}[X] - \frac{\mathsf{Cov}[X,Y]}{\mathsf{Var}[Y]}(Y - \mathbb{E}[Y])\right] = \mathbb{E}[X] - \mathbb{E}[X] - \frac{\mathsf{Cov}[X,Y]}{\mathsf{Var}[Y]}(\mathbb{E}[Y] - \mathbb{E}[Y]) = 0.$$

- The error in the LLSE estimator is orthogonal to the observations Y. Again, we show this by direct computation, highlighting in red where we add and subtract equal terms. We also highlight in blue terms that evaluate to zero, so that the logic is clear for how the equations simplify.

$$\begin{aligned}
\mathbb{E}[(X-\widehat{x}_{LLSE}(Y))Y] &= \mathbb{E}\left[\left(X-\mathbb{E}[X]-\frac{\mathsf{Cov}[X,Y]}{\mathsf{Var}[Y]}(Y-\mathbb{E}[Y])\right)Y\right] \\
&= \mathbb{E}\Big[(X-\mathbb{E}[X])Y\Big] - \frac{\mathsf{Cov}[X,Y]}{\mathsf{Var}[Y]}\mathbb{E}\Big[(Y-\mathbb{E}[Y])Y\Big] \\
&= \mathbb{E}\Big[(X-\mathbb{E}[X])Y\Big] - \mathbb{E}\Big[(X-\mathbb{E}[X])\mathbb{E}[Y]\Big] + \mathbb{E}\Big[(X-\mathbb{E}[X])\mathbb{E}[Y]\Big] - \frac{\mathsf{Cov}[X,Y]}{\mathsf{Var}[Y]}\mathbb{E}\Big[(Y-\mathbb{E}[Y])Y\Big] \\
&= \mathbb{E}\Big[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])\Big] + \mathbb{E}\Big[(X-\mathbb{E}[X])\Big]\mathbb{E}[Y] - \frac{\mathsf{Cov}[X,Y]}{\mathsf{Var}[Y]}\mathbb{E}\Big[(Y-\mathbb{E}[Y])Y\Big] \\
&= \mathsf{Cov}[X,Y] - \frac{\mathsf{Cov}[X,Y]}{\mathsf{Var}[Y]}\mathbb{E}\Big[(Y-\mathbb{E}[Y])Y - (Y-\mathbb{E}[Y])\mathbb{E}[Y] + (Y-\mathbb{E}[Y])\mathbb{E}[Y]\Big] \\
&= \mathsf{Cov}[X,Y] - \frac{\mathsf{Cov}[X,Y]}{\mathsf{Var}[Y]}\mathbb{E}\Big[(Y-\mathbb{E}[Y])(Y-\mathbb{E}[Y])\Big] - \frac{\mathsf{Cov}[X,Y]}{\mathsf{Var}[Y]}\mathbb{E}\Big[(Y-\mathbb{E}[Y])\Big]\mathbb{E}[Y] \\
&= \mathsf{Cov}[X,Y] - \frac{\mathsf{Cov}[X,Y]}{\mathsf{Var}[Y]}\mathsf{Var}[Y] = \mathsf{Cov}[X,Y] - \mathsf{Cov}[X,Y] = 0
\end{aligned}$$

  The terms in blue above are 0 because $\mathbb{E}\big[(X-\mathbb{E}[X])\big] = 0, \mathbb{E}\big[(Y-\mathbb{E}[Y])\big] = 0$. As a consequence, the LLSE error $X-\widehat{x}_{LLSE}(Y)$ is also orthogonal to any linear function of Y. In this derivation, we have also shown that $\mathbb{E}\Big[(X-\mathbb{E}[X])Y\Big] = \mathsf{Cov}[X,Y]$, and $\mathbb{E}\Big[(Y-\mathbb{E}[Y])Y\Big] = \mathsf{Var}[Y]$.

- The mean-square error of the LLSE estimator is no smaller than the mean square error for the MMSE estimator. The MMSE estimator mean-square error is the smallest among all the nonlinear estimators, whereas the LLSE estimator mean-square error is the smallest among all the linear estimators only. However, if the MMSE estimator is a linear function of Y, then the mean square errors of the LLSE and MMSE estimators are the same, and the estimators are also equal. Thus, for jointly Gaussian X, Y, the MMSE and LLSE and MAP are equal and have the same mean-square error.
- One interpretation for the LLSE estimator is that, given the first- and second-order statistics for X, Y, it approximates the joint density of X, Y as a Gaussian density with these statistics. The MMSE estimator for this Gaussian problem is the same as the LLSE estimator.

This orthogonality property is depicted in Figure 7.4. The idea is that the optimal estimate is that linear function of the data which has no correlation with the error. Intuitively, if correlation remained between the error and the estimate, there would remain information in the error of help in estimating X that we should have extracted. Note that this geometric condition implies that the error is orthogonal (i.e. uncorrelated with) both the data itself (which is obviously a trivial function of the data) as well as the LLSE estimate (which is clearly a linear function of the data).

We can derive the LLSE estimator from the following properties: We want an unbiased estimator, that is orthogonal to the observations. These properties can be derived directly from the properties of projections, which we have not emphasized in our development. Given these two properties, the unbiased property implies

$$\mathbb{E}[(X-aY-b)] = 0 \iff b = \mathbb{E}[X] - a\mathbb{E}[Y].$$

Furthermore, the orthogonality property implies

$$\mathbb{E}[(X-aY-b)Y] = 0 \iff E[(X-E[X]) - a[Y-E[Y])Y] = 0 \text{ (substitute } b \text{ in.)}$$

$$\mathbb{E}[(X-E[X]) - a[Y-E[Y])Y] = \mathsf{Cov}[X,Y] - a\mathsf{Var}[Y] = 0 \iff a = \frac{\mathsf{Cov}[X,Y]}{\mathsf{Var}[Y]}$$

Hence,

$$\widehat{x}_{LLSE}(y) = \frac{\mathsf{Cov}[X,Y]}{\mathsf{Var}[Y]}Y + \mathbb{E}[X] - \frac{\mathsf{Cov}[X,Y]}{\mathsf{Var}[Y]}\mathbb{E}[Y] = \mathbb{E}[X] + \frac{\mathsf{Cov}[X,Y]}{\mathsf{Var}[Y]}(Y-\mathbb{E}[Y]).$$

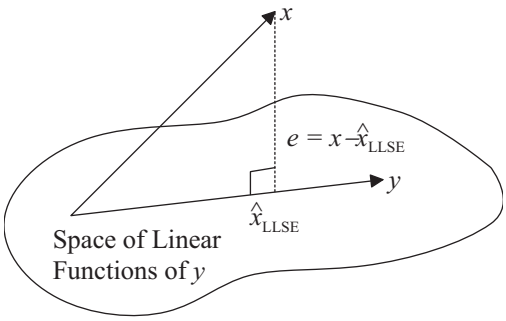


Figure 7.4: Illustration of the projection theorem for LLSE.

We can use the orthogonality property also to derive an expression for the variance of the error. Since the error $X - \widehat{x}_{LLSE}(Y)$ is orthogonal to any linear function of Y, it is also orthogonal to the estimate $\widehat{x}_{LLSE}(y)$. Hence,

$$\mathsf{Var}[X] = \mathsf{Var}[X - \widehat{x}_{LLSE}(Y)] + \mathsf{Var}[\widehat{x}_{LLSE}(Y)].$$

Since $\widehat{x}_{LLSE}(Y) = \mathbb{E}[X] + \frac{\mathsf{Cov}[X,Y]}{\mathsf{Var}[Y]}(Y - \mathbb{E}[Y])$ is a scaled and shifted version of Y, its variance is

$$\mathsf{Var}[\widehat{x}_{LLSE}(Y)] = \Big(\frac{\mathsf{Cov}[X,Y]}{\mathsf{Var}[Y]}\Big)^2 \mathsf{Var}[Y] = \frac{\mathsf{Cov}[X,Y]^2}{\mathsf{Var}[Y]}.$$

Hence,

$$\mathsf{Var}[X - \widehat{x}_{LLSE}(Y)] = E[(X - \widehat{x}_{LLSE}(Y))^2] = \mathsf{Var}[X] - \frac{\mathsf{Cov}[X,Y]^2}{\mathsf{Var}[Y]}.$$

Example 7.11
For this example let us revisit the problem of Example 7.10. We need the second order quantities $\mathbb{E}[X]$, $\mathbb{E}[Y]$, $\mathsf{Cov}[X,Y[$, $\mathsf{Var}[X]$, and $\mathsf{Var}[Y]$, which we compute as:

$$\mathbb{E}[Y] = \int_{-\infty}^{\infty} y f_Y(y)\, dy = \int_0^1 5y^5\, dy = \frac{5}{6}$$

$$\mathbb{E}[X] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x f_{X,Y}(x,y)\, dx\, dy = \int_0^1 \int_0^{y^2} x 10x\, dx\, dy == \frac{10}{21}$$

$$\mathsf{Var}[Y] = \int_{-\infty}^{\infty} y^2\, f_Y(y)\, dy - \mathbb{E}[Y]^2 = \int_0^1 5y^6\, dy - \left(\frac{5}{6}\right)^2 = \frac{5}{252}$$

$$\mathsf{Var}[X] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x^2\, f_X(x)\, dx\, dy - \mathbb{E}[X]^2 = \int_0^1\int_0^{y^2} x^2\, 10x\, dx\, dy - \left(\frac{10}{21}\right)^2 = \frac{5}{18} - \left(\frac{10}{21}\right)^2 = \frac{5}{98}$$

$$\mathsf{Cov}[X,Y] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy\, f_{X,Y}(x,y)\, dx\, dy - \mathbb{E}[X]\mathbb{E}[Y] = \int_0^1\int_0^{y^2} 10x^2 y\; dx\, dy - \frac{5}{6}\frac{10}{21} = .\frac{5}{12} - \frac{5}{6}\frac{10}{21} = \frac{5}{252}$$

Thus we obtain for the LLSE:

$$\widehat{x}_{\mathsf{LLSE}}(y) = \mathbb{E}[X] + \frac{\mathsf{Cov}[X,Y]}{\mathsf{Var}[Y]}\,(y - \mathbb{E}[Y]) = \frac{10}{21} + \frac{5/252}{5/252}\left(y - \frac{5}{6}\right) = y - \frac{5}{14}$$

as before. Using the formula for the MSE we obtain

$$\mathsf{MSE} = \mathsf{Var}[X] - \frac{\mathsf{Cov}[X,Y]^2}{\mathsf{Var}[Y]} = \frac{5}{98} - \frac{(5/252)^2}{5/252} = \frac{55}{1764} = 0.0312$$

Note that this MSE is worse than that obtained by the optimal MMSE estimator of Example 7.10 – but not much worse.

Example 7.12
Consider a simple example where a temperature sensor measures the true temperature T, which is assumed to be a Gaussian random variable, with mean 27 and variance 9. The measurement observed is modeled as

$$Y = T + V$$

where V is measurement error, independent of T, corresponding to quantization noise. V is assumed to be uniformly distributed on [-3,3]. Hence, $\mathbb{E}[V] = 0, \mathsf{Var}[V] = 3$. This is a common model for measurements, where part of the measurement models the relationship between the unobserved variable T and the measurement Y, and the other part represents the errors in the measurement process.

The goal is to generate the LLSE estimate of T based on observation of Y. We compute the needed statistics below:

$$\begin{aligned}\mathbb{E}[T] = 27;\ \mathbb{E}[Y] = \mathbb{E}[T] + \mathbb{E}[V] = 27;\ \mathsf{Var}[T] = 9;\\ \mathsf{Var}[Y] = \mathsf{Var}[T] + \mathsf{Var}[V] + 2\mathsf{Cov}[T,V] = \mathsf{Var}[T] + \mathsf{Var}[V] = 9+3 = 12\ (T,V \text{ independent.})\\ \mathsf{Cov}[T,Y] = \mathsf{Cov}[T,T+V] = \mathsf{Var}[T] + \mathsf{Cov}[T,V] = \mathsf{Var}[T] = 9;\end{aligned}$$

With these statistics, we have

$$\widehat{T}_{LLSE}(y) = \mathbb{E}[T] + \frac{\mathsf{Cov}[T,Y]}{\mathsf{Var}[Y]}(y - \mathbb{E}[Y]) = 27 + \frac{9}{12}(y-27) = \frac{1}{4}\cdot 27 + \frac{3}{4}\cdot y.$$

The MSE error is

$$MSE = \mathsf{Var}[T] - \frac{\mathsf{Cov}[T,Y]^2}{\mathsf{Var}[T]} = 9 - \frac{81}{12} = \frac{9}{4}.$$

Example 7.13
Let's revisit the example of estimating probabilities using multiple trials. Let P be a uniform random variable on [0,1]. Let measurement Y be a Binomial(N,P) random variable, corresponding to the total number of successes in N independent trials with probability of success P for each trial. From previous examples, we know

$$\widehat{x}_{ML}(y) = \widehat{x}_{MAP}(y) = \frac{y}{N}.$$

As discussed previously, this is an unbiased estimate.

Can we compute the MMSE estimate for this simple case? The conditional density of P given observation $Y = y$ is given by

$$f_{P|Y}(p|y) = \frac{P_{Y|P}(y|p)f_P(p)}{P_Y(y)} = \frac{\binom{N}{y}p^y(1-p)^{N-y}}{\binom{N}{y}\int_0^1 q^y(1-q)^{N-y}\,dq}.$$

Evaluating the denominator is not easy, but we can do the integral as follows, using repeated integration by parts:

$$\begin{aligned}&\int_0^1 q^y(1-q)^{N-y}\,dq = \frac{N-y}{y+1}\int_0^1 q^{y+1}(1-q)^{N-y-1}\,dq\\ &\Rightarrow \int_0^1 q^y(1-q)^{N-y}\,dq = \frac{(N-y)!k!}{N!}\int_0^1 q^N\,dq = \frac{1}{N+1}\cdot\frac{1}{\binom{N}{y}}\end{aligned}$$

Thus, $P_Y(y) = \frac{1}{N+1}, y \in \{0,1,\ldots,N\}$, which is a discrete uniform distribution. Intuitively, this makes sense. Since we have no information about P, every value of Y is equally likely when averaged over all the possible P.

The conditional density of P given $Y = y$ is

$$f_{P|Y}(p|y) = (N+1)\binom{N}{y}p^y(1-p)^{N-y}.$$

The MMSE estimator is

$$\widehat{P}_{MMSE}(y) = \int_0^1 p f_{P|Y}(p|y)\,dp = \int_0^1 (N+1)p\binom{N}{y}p^y(1-p)^{N-y}\,dp.$$

This has another difficult integral, but we can again evaluate it using repeated integration by parts as:

$$\begin{aligned}&\int_0^1 p^{y+1}(1-p)^{N-y}\,dp = \frac{N-y}{y+2}\int_0^1 p^{y+2}(1-p)^{N-y-1}\,dp\\ &\Rightarrow \int_0^1 p^{y+1}(1-p)^{N-y}\,dp = \frac{(N-y)!(y+1)!}{(N+1)!}\int_0^1 p^{N+1}\,dp = \frac{1}{N+2}\cdot\frac{1}{\binom{N+1}{y+1}}\end{aligned}$$

Thus,

$$\widehat{P}_{MMSE}(y) = (N+1) \cdot \binom{N}{y} \cdot \frac{1}{N+2} \cdot \frac{1}{\binom{N+1}{y+1}} = \frac{y+1}{N+2}.$$

Note this is different from the MAP estimator, but it is also unbiased. Since it is linear, this is also the LLSE estimator! We can verify this through computation, as

$$\mathbb{E}[P] = 0.5;\ \mathbb{E}[Y] = 0.5N;\ \mathsf{Var}[P] = \frac{1}{12};\ \mathsf{Var}[Y] = \frac{N(N+2)}{12}$$
$$\mathsf{Cov}[Y,P] = \mathbb{E}\big[\mathbb{E}[(Y-0.5N)(P-0.5)|P]\big] = \mathbb{E}\big[\mathbb{E}[(Y-0.5N)|P](P-0.5)\big]$$
$$= \mathbb{E}\big[N(P-0.5)^2\big] = N\int_0^1 (p-0.5)^2\,dp = \frac{N}{12}$$

Then,

$$\widehat{P}_{LLSE}(y) = (N+1) \cdot \binom{N}{y} \cdot \frac{1}{N+2} \cdot \frac{1}{\binom{N+1}{y+1}} = \frac{y+1}{N+2} = \widehat{P}_{MMSE}(y).$$

The resulting MSE for the LLSE and MMSE estimator is

$$MSE = Var[P] - \frac{Cov[Y,P]^2}{Var[Y]} = \frac{1}{12} - \frac{N^2}{144} \cdot \frac{12}{N(N+2)} = \frac{1}{12}(1 - \frac{N}{N+2}) = \frac{1}{6(N+2)}.$$

The MSE for the MAP estimator is

$$MSE = E[(P - \frac{Y}{N})^2 = \mathsf{Var}[P] - 2\frac{\mathsf{Cov}[P,Y]}{N} + \frac{\mathsf{Var}[Y]}{N^2} = \frac{1}{12} - \frac{2}{12} + \frac{N+2}{12N} = \frac{1}{6N}.$$

Note the MAP MSE is slightly larger than the MMSE and LLSE MSE.

Let us close by summarizing the properties of LLSE estimates:

- The LLSE estimate is the minimum MSE estimate over all *linear* functions of the data.
- The LLSE estimate is always unbiased.
- The associated error covariances satisfy is at least as large as the error covariance of the MMSE estimate.
- The LLSE estimate equals the MMSE estimate for the jointly Gaussian case.
- The LLSE estimate only requires knowledge of second-order properties.

7.5 Estimation for Random Vectors

We conclude this chapter by discussing how the estimation concepts extend to random vectors. This extension is critical for many applications, including statistics and data science. In this case, the unobserved variables and the observed variables can both be vectors. The model is as follows. We suppose that we have a random vector $\underline{Z}$ that can be partitioned into two subvectors as $\underline{Z} = \begin{bmatrix} \underline{X} \\ \underline{Y} \end{bmatrix}$, where $\underline{Y}$ is the observation vector and $\underline{X}$ is the unobserved vector. Let $X \in \Re^n, Y \in \Re^m$. The objective in estimation is now to construct an estimator $\underline{\widehat{x}}(\underline{y})$ that will estimate the unobserved vector $\underline{X}$ based on the observed values $\underline{Y} = \underline{y}$. Note that the vector estimate is a vector composed of the estimates for each of the components of X, so that

$$\underline{\widehat{x}}(\underline{y}) = \begin{bmatrix} \widehat{x}_1(\underline{y}) \\ \vdots \\ \widehat{x}_n(\underline{y}) \end{bmatrix}.$$

We begin with the statistical description of the random vectors. Assuming the random vectors are continuous valued, with joint densities, we will have a joint density

$$f_{\underline{X},\underline{Y}}(\underline{x},\underline{y}) = f_{\underline{Y}|\underline{X}}(\underline{y}|\underline{x})f_{\underline{X}}(\underline{x})$$

where we have shown the factorization of the joint density into a conditional density for the observed variables, given the unobserved variables, and a marginal density for the unobserved variables.

The mean square error of an estimator of the vector $\underline{X}$ given $\underline{Y}$ is defined as

$$MSE = \mathbb{E}[\sum_{k=1}^{n}(X_i - \widehat{x}_i(\underline{Y}))^2] = \mathbb{E}[(\underline{X} - \underline{\widehat{x}}(\underline{y}))^T(\underline{X} - \underline{\widehat{x}}(\underline{y}))].$$

This is a common metric that will be used for evaluating the quality of estimators.

7.5.1 ML and MAP estimation for random vectors

Given this statistical description, we extend our estimation concepts to random vectors. The **maximum likelihood estimate** (ML) is given by

$$\underline{\widehat{x}}_{ML}(\underline{y}) \in \arg\max_{\underline{x}} f_{\underline{Y}|\underline{X}}(\underline{y}|\underline{x})$$

where the maximization is now over vector arguments, and often requires iterative search algorithms. The main difficulty in this estimation is that the likelihood function $f_{\underline{Y}|\underline{X}}(\underline{y}|\underline{x})$ can have many local maxima, which makes the search for a global maximum a difficult combinatorial problem. However, for some special cases, it will be possible to find global maxima, as we will illustrate with examples.

Similarly, the **maximum a posteriori** (MAP) estimate is given by

$$\underline{\widehat{x}}_{MAP}(\underline{y}) \in \arg\max_{\underline{x}} f_{\underline{Y}|\underline{X}}(\underline{y}|\underline{x})f_{\underline{X}}(\underline{x}).$$

The MAP estimator includes the additional information on the prior density of the unobserved variables $\underline{X}$. The optimization still has difficulties with local maxima. Nevertheless, specifying a prior distribution for $\underline{X}$ serves as a regularization term that guides the optimization towards specific regions in the search space space where the maxima are expected to be, and makes the solution less sensitive to measurement errors.

Example 7.14
One of the most interesting applications of vector estimation is for estimating the parameters of the distribution of a random variable, given many observation samples. For instance, assume you are measuring the delays in your favorite transportation mechanism, the Green Line. This delay T is modeled as an exponential random variable with parameter λ, but the parameter λ is unknown. To estimate λ, we observe the sample delays over many days. We assume the delays in different days are independent, with the same underlying distribution for T. Let T_k denote the delay measured at day k, where $k = 1, 2, \ldots, N$.

To formulate this in the form of a vector estimation problem, let $X = \lambda$ be the unobserved variable. Our observed vector $\underline{Y}$ is given as

$$\underline{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_N \end{bmatrix}.$$

With this choice of variables, we need the statistical description. Note that, given $\lambda = x$, the conditional density of Y_k is an exponential density:

$$f_{Y_k|X}(y_k|x) = \begin{cases} xe^{-xy_k} & y_k \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

Furthermore, the fact that each of the Y_k represent an independent sample of T yields the following expression for the joint conditional density of $\underline{Y}$ given X:

$$f_{\underline{Y}|X}(\underline{y}|x) = \prod_{k=1}^{N} f_{Y_k|X}(y_k|x).$$

With this information, we can now obtain the ML estimate of X, given observation of the vector $\underline{Y}$, as

$$\widehat{x}_{ML}(\underline{y}) \in \arg\max_{x\geq 0} f_{\underline{Y}|X}(\underline{y}|x) = \arg\max_{x\geq 0} \ln\left(f_{\underline{Y}|X}(\underline{y}|x)\right)$$

$$\ln\left(f_{\underline{Y}|X}(\underline{y}|x)\right) = \sum_{k=1}^{N} \ln\left(f_{Y_k|x}(y_k|x)\right) = N\ln(x) - \sum_{k=1}^{N} y_k x$$

To solve the maximization problem, we differentiate with respect to x and find where the derivative is zero: as long as that happens for positive x, it will be the maximum.

$$\frac{d}{dx}\left(N\ln(x) - \sum_{k=1}^{N} y_k x\right) = \frac{N}{x} - \sum_{k=1}^{N} y_k = 0 \iff x = \frac{N}{\sum_{k=1}^{N} y_k}.$$

Note that the value of x where the derivative vanishes is unique and non-negative. Therefore, the ML estimator is $\widehat{x}_{ML}(\underline{y}) = \frac{N}{\sum_{k=1}^{N} y_k}$. This is an intuitive estimator, as it says that the estimate of the rate λ is the inverse of the average delay.

What if we had some prior information on X, and we wanted to generate the MAP estimator? Assume that we know that the rate is distributed uniformly in $[0.1, 1.1]$, corresponding to average delays between 0.909 and 10 minutes. Thus,

$$f_X(x) = \begin{cases} 1 & x \in [0.1, 1.1], \\ 0 & \text{otherwise.} \end{cases}$$

In this case, the MAP estimate is given by

$$\widehat{x}_{MAP}(\underline{y}) \in \arg\max_{x\geq 0} f_{\underline{Y}|X}(\underline{y}|x) f_X(x) = \arg\max_{x\in[0.1,1.1]} \ln\left(f_{\underline{Y}|X}(\underline{y}|x)\right) = \arg\max_{x\in[0.1,1.1]} N\ln(x) - \sum_{k=1}^{N} y_k x.$$

The prior information resulted in restricting the search to the interval $x \in [0.1, 1.1]$, because the product of the two densities is zero outside of this interval, and cannot be a maximum. Proceeding as above, the point where the derivative with respect to zero vanishes is $x = \frac{N}{\sum_{k=1}^{N} y_k}$. However, this value of x may not be in $[0.1, 1.1]$, in which case we find the closest value in the interval where the function is maximized (the log-likelihood has a unique maximum and is continuous), so that the MAP estimator is

$$\widehat{x}_{MAP}(\underline{y}) = \begin{cases} 0.1 & \frac{N}{\sum_{k=1}^{N} y_k} < 0.1, \\ 1.1 & \frac{N}{\sum_{k=1}^{N} y_k} > 1.1, \\ \frac{N}{\sum_{k=1}^{N} y_k} & \text{otherwise.} \end{cases}$$

Example 7.15
Let U be a Gaussian random variable, with unknown mean m and variance v. We are interested in estimating the mean and variance of U. We collect N independent samples of U, where sample k is denoted as Y_k.

When posed as an estimation problem, our unobserved variables are m and v. Let $\underline{X} = \begin{bmatrix} m \\ v \end{bmatrix} \equiv \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$. The observed vector is

$$\underline{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_N \end{bmatrix}.$$

As before, we need the likelihood of the observed vector given the unobserved vector. From our Gaussian assumptions, we have

$$f_{Y_k|\underline{X}}(y_k|\underline{x}) = \frac{1}{\sqrt{2\pi x_2}} e^{-\frac{(y_k - x_1)^2}{2x_2}}.$$

Furthermore, under the assumption of independent sampling, we have

$$f_{\underline{Y}|\underline{X}}(\underline{y}|\underline{x}) = \prod_{k=1}^{N} f_{Y_k|\underline{X}}(y_k|\underline{x}).$$

Thus, the log-likelihood is given as

$$\ln\left(f_{\underline{Y}|\underline{X}}(\underline{y}|\underline{x})\right) = \sum_{k=1}^{N} \ln\left(f_{Y_k|x}(y_k|\underline{x})\right) = -\frac{N}{2}\ln(2\pi x_2) - \sum_{k=1}^{N} \frac{(y_k - x_1)^2}{2x_2}.$$

We wish to maximize the above over any real-valued x_1, and for $x_2 \geq 0$, as x_2 represents the unknown covariance. Hence, the ML estimate is

$$\widehat{\underline{x}}_{ML}(\underline{y}) \in \arg \max_{x_2 \geq 0, x_1} \left(-\frac{N}{2} \ln(2\pi x_2) - \sum_{k=1}^{N} \frac{(y_k - x_1)^2}{2x_2} \right).$$

To maximize, take the partial derivative of the above with respect to x_1 and x_2 and set both equal to zero:

$$\frac{\partial}{\partial x_1} \left(-\frac{N}{2} \ln(2\pi x_2) - \sum_{k=1}^{N} \frac{(y_k - x_1)^2}{2x_2} \right) = \sum_{k=1}^{N} \frac{(y_k - x_1)}{x_2} = 0 \iff x_1 = \frac{\sum_{k=1}^{N} y_k}{N}.$$

$$\frac{\partial}{\partial x_2} \left(-\frac{N}{2} \ln(2\pi x_2) - \sum_{k=1}^{N} \frac{(y_k - x_1)^2}{2x_2} \right) = -\frac{N}{2x_2} + \sum_{k=1}^{N} \frac{(y_k - x_1)^2}{2x_2^2} = 0 \iff x_2 = \frac{\sum_{k=1}^{N} (y_k - x_1)^2}{N}.$$

Note that the solution for x_2 is always non-negative, satisfying the constraints. The ML estimate is thus

$$\widehat{m}_{ML} = \frac{\sum_{k=1}^{N} y_k}{N}; \quad \widehat{v}_{ML} = \frac{\sum_{k=1}^{N} (y_k - \widehat{m}_{ML})^2}{N}.$$

What if we had some prior information on m and v? Assume that, apriori, we knew that m was Gaussian, with mean 0, variance v, and v was uniform in [1,5]. This implies

$$f_{\underline{X}}(\underline{x}) = f_{X_1|X_2}(x_1|x_2) f_{X_2}(x_2) = \begin{cases} \frac{1}{\sqrt{2\pi x_2}} e^{-\frac{x_1^2}{2x_2}} \cdot \frac{1}{4} & x_2 \in [1,5], x_1 \in \Re, \\ 0 & \text{otherwise..} \end{cases}$$

The MAP estimator is now obtained as

$$\widehat{\underline{x}}_{MAP}(\underline{y}) \in \arg \max_{x_2 > 0, x_1} \left(f_{\underline{Y}|\underline{X}}(\underline{y}|\underline{x}) f_{\underline{X}}(\underline{x}) \right) = \arg \max_{x_2 \in [1,5], x_1} \left(f_{\underline{Y}|\underline{X}}(\underline{y}|\underline{x}) f_{\underline{X}}(\underline{x}) \right).$$

because the joint density is zero when $x_2 \notin [1,5]$ and hence cannot be maximal there. Taking logarithms, we get

$$\ln \left(f_{\underline{Y}|\underline{X}}(\underline{y}|\underline{x}) f_{\underline{X}}(\underline{x}) \right) = \left(-\frac{N}{2} \ln(2\pi x_2) - \sum_{k=1}^{N} \frac{(y_k - x_1)^2}{2x_2} \right) - \frac{1}{2} \ln(2\pi x_2) - \ln(4) - \frac{x_1^2}{2x_2}.$$

Differentiating with respect to x_1 and x_2 yields:

$$\frac{\partial}{\partial x_1} \left(-\frac{N}{2} \ln(2\pi x_2) - \sum_{k=1}^{N} \frac{(y_k - x_1)^2}{2x_2} - \frac{1}{2} \ln(2\pi) - \ln(4) - \frac{x_1^2}{2x_2} \right) = \sum_{k=1}^{N} \frac{(y_k - x_1)}{x_2} - \frac{x_1}{x_2} = 0.$$

$$\frac{\partial}{\partial x_2} \left(-\frac{N}{2} \ln(2\pi x_2) - \sum_{k=1}^{N} \frac{(y_k - x_1)^2}{2x_2} \right) = -\frac{N}{2x_2} + \sum_{k=1}^{N} \frac{(y_k - x_1)^2}{2x_2^2} - \frac{1}{2x_2} = 0.$$

The equations are easily solved as:

$$x_1 = \frac{\sum_{k=1}^{N} y_k}{N+1}.$$

For the second equation, we have

$$-(N+1)x_2 + \sum_{k=1}^{N} (y_k - x_1)^2 = 0 \iff x_2 = \frac{\sum_{k=1}^{N} (y_k - x_1)^2}{N+1}.$$

Taking into account the constraints on the optimization, the MAP estimator is

$$\widehat{m}_{MAP} = \frac{\sum_{k=1}^{N} y_k}{N+1}; \quad \widehat{v}_{MAP} = \begin{cases} 1 & \frac{\sum_{k=1}^{N} (y_k - \widehat{m}_{ML})^2}{N+1} < 1, \\ 5 & \frac{\sum_{k=1}^{N} (y_k - \widehat{m}_{ML})^2}{N+1} > 5, \\ \frac{\sum_{k=1}^{N} (y_k - \widehat{m}_{ML})^2}{N+1} & \text{otherwise.} \end{cases}$$

7.5.2 MMSE and LLSE estimation for random vectors

The MMSE estimator of $\underline{X}$, based on observations of $\underline{Y}$, is given by

$$\widehat{\underline{x}}_{MMSE}(\underline{y}) = \begin{bmatrix} \mathbb{E}[X_1|\underline{Y}] \\ \vdots \\ \mathbb{E}[X_n|\underline{Y}] \end{bmatrix} = \begin{bmatrix} \int \cdots \int x_1 f_{\underline{X}|\underline{Y}}(\underline{x}|\underline{y})\, dx_1 \\ \vdots \\ \int \cdots \int x_n f_{\underline{X}|\underline{Y}}(\underline{x}|\underline{y})\, dx_n \end{bmatrix} = \mathbb{E}[\underline{X}|\underline{Y}].$$

Computation of this estimate requires the conditional density $f_{\underline{X}|\underline{Y}}((\underline{x}|\underline{y})$, obtained by Bayes' Rule as

$$f_{\underline{X}|\underline{Y}}((\underline{x}|\underline{y}) = \frac{f_{\underline{Y}|\underline{X}}(\underline{y}|\underline{x}) f_{\underline{X}}(\underline{x})}{f_{\underline{Y}}(\underline{y})}.$$

Computing the denominator in Bayes' Rule requires a multidimensional integral that is usually very difficult to evaluate; this limits our ability to compute MMSE estimators for general distributions.

There is a special case where the MMSE solution can be computed efficiently: when $\underline{X}$ and $\underline{Y}$ are jointly Gaussian random vectors. As we derived in 5, the conditional expected value $\mathbb{E}[\underline{X}|\underline{Y}]$ is a linear function of the observed value $\underline{Y} = \underline{y}$, and the MMSE estimator will be the same as the LLSE estimator. We will discuss the LLSE estimator for random vectors below.

We assume the random vector $\underline{Z} = \begin{bmatrix} \underline{X} \\ \underline{Y} \end{bmatrix}$ has first-order statistics given by

$$\mathbb{E}[\underline{Z}] = \begin{bmatrix} \underline{\mu}_{\underline{X}} \\ \underline{\mu}_{\underline{Y}} \end{bmatrix}.$$

Let n_X, n_Y be the dimensions of the random vectors $\underline{X}, \underline{Y}$. The random vector $\underline{Z}$ has covariance matrix $\Sigma_{\underline{Z}}$, which can be partitioned along the dimensions of $\underline{X}, \underline{Y}$ as follows:

$$\Sigma_{\underline{Z}} = \begin{bmatrix} \Sigma_{\underline{X}} & \Sigma_{\underline{X},\underline{Y}} \\ \Sigma_{\underline{Y},\underline{X}} & \Sigma_{\underline{Y}} \end{bmatrix}.$$

Note that $\Sigma_{\underline{X}}$ is an $n_X \times n_X$ matrix, which is the covariance matrix for the unobserved vector $\underline{X}$. $\Sigma_{\underline{Y}}$ is an $n_Y \times n_Y$ matrix, which is the covariance matrix for the observed vector $\underline{Y}$. The matrix $\Sigma_{\underline{X},\underline{Y}}$ is an $n_X \times n_Y$ matrix known as the cross-covariance between $\underline{X}$ and $\underline{Y}$, and is defined as

$$\Sigma_{\underline{X},\underline{Y}} = \mathbb{E}\Big[(\underline{X} - \mathbb{E}[\underline{X}])(\underline{Y} - \mathbb{E}[\underline{Y}])^T\Big].$$

A linear estimator of $\underline{X}$ based on $\underline{Y}$ is an estimator of the form $\widehat{\underline{x}}(\underline{y}) = \mathbf{A}\underline{y} + \underline{b}$ for a constant matrix $\mathbf{A}$ of dimension $n_X \times n_Y$, and $\underline{b}$ is a constant vector of dimension n_X.

We want to find the best linear estimator of $\underline{X}$ given observation $\underline{Y}$, where best is the estimator that yields the smallest least squares error. Rather than posing this as an optimization problem, we will derive this estimator using the orthogonality and unbiased properties of the LLSE estimator, which can be established using the principles of best approximation. Specifically, we seek a linear estimator that is unbiased, and where the estimation error is orthogonal to the observations.

The first condition is an unbiased estimator, which requires

$$\mathbb{E}[\underline{X} - \mathbf{A}\underline{Y} - \underline{b}] = 0 = \underline{\mu}_{\underline{X}} - \mathbf{A}\underline{\mu}_{\underline{Y}} - \underline{b} \iff \underline{b} = \underline{\mu}_{\underline{X}} - \mathbf{A}\underline{\mu}_{\underline{Y}}.$$

The next condition is that the estimation error must be orthogonal to the measurements. The error is a vector, $\underline{e} = \underline{X} - \mathbf{A}\underline{Y} - \underline{b}$. Orthogonality requires that every component of the error is orthogonal to every

component of the measurement vector:

$$\begin{aligned}\mathbb{E}\big[(\underline{X}-\mathbf{A}\underline{Y}-\underline{b})\underline{Y}^T\big] &= 0 = \mathbb{E}\big[\underline{X}\underline{Y}^T-\mathbf{A}\underline{Y}\underline{Y}^T-\underline{b}\underline{Y}^T\big]\\ &= \mathbb{E}\big[(\underline{X}-\mathbb{E}[\underline{X}])\underline{Y}^T-\mathbf{A}(\underline{Y}-\mathbb{E}[\underline{Y}])\underline{Y}^T\big] \quad \text{(substituting for } b,)\\ &= \mathbb{E}\big[(\underline{X}-\mathbb{E}[\underline{X}])\underline{Y}^T\big]-\mathbf{A}\mathbb{E}\big[(\underline{Y}-\mathbb{E}[\underline{Y}])\underline{Y}^T\big]\\ &= \Sigma_{\underline{X},\underline{Y}}-\mathbf{A}\Sigma_{\underline{Y}}=0 \iff \mathbf{A}=\Sigma_{\underline{X},\underline{Y}}\Sigma_{\underline{Y}}^{-1}\end{aligned}$$

Thus, we have our LLSE estimator for random vectors:

$$\widehat{\underline{x}}_{LLSE}(\underline{y})=\mathbb{E}[\underline{X}]+\Sigma_{\underline{X},\underline{Y}}\Sigma_{\underline{Y}}^{-1}(\underline{Y}-\mathbb{E}[\underline{Y}]).$$

This is very similar to the structure of the scalar LLSE estimator we discussed in the previous section. The main difference is that, in dealing with vectors, we must take matrix inverses and preserve the order of the operations when we take constants out of the expectations.

We can also use orthogonality to derive an expression for the covariance of the estimation error. We know the estimator is orthogonal to the error vector, because the estimator is a linear function of the measurement Y. Hence, $\underline{X}=\underline{e}+\widehat{\underline{x}}_{LLSE}(\underline{y})$ is the sum of two uncorrelated vectors. Thus,

$$\Sigma_{\underline{X}}=\Sigma_{\underline{e}}+\Sigma_{\widehat{\underline{x}}_{LLSE}(\underline{Y})}.$$

We also know that $\widehat{\underline{x}}_{LLSE}(\underline{Y})$ is a linear transformation of $\underline{Y}$, so

$$\Sigma_{\widehat{\underline{x}}_{LLSE}(\underline{Y})}=\Sigma_{\underline{X},\underline{Y}}\Sigma_{\underline{Y}}^{-1}\Sigma_{\underline{Y}}\Sigma_{\underline{Y}}^{-1}\Sigma_{\underline{X},\underline{Y}}^T=\Sigma_{\underline{X},\underline{Y}}\Sigma_{\underline{Y}}^{-1}\Sigma_{\underline{X},\underline{Y}}^T.$$

Therefore,

$$\Sigma_{\underline{e}}=\Sigma_{\underline{X}}-\Sigma_{\underline{X},\underline{Y}}\Sigma_{\underline{Y}}^{-1}\Sigma_{\underline{X},\underline{Y}}^T.$$

We illustrate these results with an example:

Example 7.16
Let $\underline{X}$ be a random, two-dimensional vector with statistics $\mathbb{E}[\underline{X}]=\underline{0}, \Sigma_{\underline{X}}=\begin{bmatrix}1 & -0.9\\ -0.9 & 1\end{bmatrix}$.

Let $\underline{W}$ also be a two-dimensional vector, uncorrelated with $\underline{X}$, with statistics $\mathbb{E}[\underline{W}]=0, \Sigma_{\underline{W}}=\begin{bmatrix}0.1 & 0\\ 0 & 0.1\end{bmatrix}$.

Define the observation vector $\underline{Y}=\underline{X}+\underline{W}$. Then, the first and second order statistics of $\underline{X},\underline{Y}$ are:

$$\mathbb{E}[Y]=0;\ \Sigma_{\underline{Y}}=\Sigma_{\underline{X}}+\Sigma_{ullW}=\begin{bmatrix}1.1 & -0.9\\ -0.9 & 1.1\end{bmatrix}$$
$$\Sigma_{\underline{X},\underline{Y}}=\mathsf{Cov}[\underline{X},\underline{Y}]=\mathsf{Cov}[\underline{X},\underline{X}]+\mathsf{Cov}[\underline{X},\underline{W}]=\mathsf{Cov}[\underline{X},\underline{X}]=\Sigma_{\underline{X}}$$

Note that $\Sigma_{\underline{Y}}^{-1}=\begin{bmatrix}2.75 & 2.25\\ 2.25 & 2.75\end{bmatrix}$. With this, the LLSE estimator is

$$\widehat{\underline{x}}_{LLSE}(\underline{y})=\underline{\mu}_{\underline{X}}+\Sigma_{\underline{X},\underline{Y}}\Sigma_{\underline{Y}}^{-1}(\underline{y}-\underline{\mu}_{\underline{Y}})=\begin{bmatrix}1 & -0.9\\ -0.9 & 1\end{bmatrix}\begin{bmatrix}2.75 & 2.25\\ 2.25 & 2.75\end{bmatrix}\underline{y}=\begin{bmatrix}0.725 & -0.225\\ -0.225 & 0.725\end{bmatrix}\underline{y}$$

The covariance of the estimation error $\underline{e}$ is

$$\Sigma_{\underline{e}}=\begin{bmatrix}1 & -0.9\\ -0.9 & 1\end{bmatrix}-\begin{bmatrix}1 & -0.9\\ -0.9 & 1\end{bmatrix}\begin{bmatrix}2.75 & 2.25\\ 2.25 & 2.75\end{bmatrix}\begin{bmatrix}1 & -0.9\\ -0.9 & 1\end{bmatrix}=\begin{bmatrix}0.0725 & -0.0225\\ -0.0225 & 0.0725\end{bmatrix}$$