2 Statistical Estimation: Basic Concepts

2.1 Probability

We briefly remind some basic notions and notations from probability theory that will be required in this chapter.

The Probability Space:

The basic object in probability theory is the *probability space* $(\Omega, \mathcal{F}, \mathbf{P})$, where
Ω is the sample space (with sample points $\omega \in \Omega$),
$\mathcal{F}$ is the (sigma-field) of possible events $B \in \mathcal{F}$, and
$\mathbf{P}$ is a probability measure, giving the probabilty $\mathbf{P}(B)$ of each possible event.

A (vector-valued) *Random Variable* (RV) X is a mapping

$$X : \Omega \to \mathbb{R}^n \,.$$

X is also required to be *measurable* on $(\Omega, \mathcal{F})$, in the sense that $X^{-1}(A) \in \mathcal{F}$ for any open (or Borel) set A in $\mathbb{R}^n$.

In this course we shall not explicitly define the underlying probability space, but rather define the probability distributions of the RVs of interest.

Distribution and Density:

For an RV $X : \Omega \to \mathbb{R}^n$, the *(cumulative) probability distribution function* (cdf) is defined as

$$F_X(x) = \mathbf{P}(X \leq x) \stackrel{\triangle}{=} \mathbf{P}\{\omega : X(\omega) \leq x\}\,, \quad x \in \mathbb{R}^n\,.$$

The *probability density function* (pdf), if it exists, is given by

$$p_X(x) = \frac{\partial^n F_X(x)}{\partial x_1 \, \ldots \, \partial x_n}\,.$$

The RV's $(X_1\,, \ldots\,, X_k)$ are *independent* if

$$F_{X_1\,,\ldots,\,X_k}(x_1\,, \ldots\,, x_k) = \prod_{k=1}^{K} F_{X_k}(x_k)$$

(and similarly for their densities).

Moments:

The *expected value* (or *mean*) of X:

$$\mu_X \equiv E(X) \stackrel{\triangle}{=} \int_{\mathbb{R}^n} x \, dF_X(x)\,.$$

More generally, for a real function g on $\mathbb{R}^n$,

$$E(g(X)) = \int_{\mathbb{R}^n} g(x) \, dF_X(x)\,.$$

The covariance matrices:

$$\mathrm{cov}(X) = E\{(X - E(X))(X - E(X))^T\}$$

$$\mathrm{cov}(X_1, X_2) = E\{(X_1 - E(X_1))(X_2 - E(X_2))^T\}\,.$$

When X is scalar then $\mathrm{cov}(X)$ is simply its *variance*.
The RV's X_1 and X_2 are *uncorrelated* if $\mathrm{cov}(X_1, X_2) = 0$.

Gaussian RVs:

A (non-degenerate) Gaussian RV X on $\mathbb{R}^n$ has the density

$$f_X(x) = \frac{1}{(2\pi)^{n/2}\,\det(\Sigma)^{1/2}}\, e^{-\frac{1}{2}(x-m)^T\,\Sigma^{-1}(x-m)}\,.$$

It follows that $m = E(X)$, $\Sigma = \mathrm{cov}(X)$. We denote $X \sim N(m, \Sigma)$.

X_1 and X_2 are *jointly* Gaussian if the random vector $(X_1; X_2)$ is Gaussian.

It holds that:

1. X Gaussian $\iff$ all linear combinations $\sum_i a_i X_i$ are Gaussian.
2. X Gaussian $\Rightarrow$ $Y = AX$ is Gaussian.
3. X_1, X_2 jointly Gaussian and uncorrelated
   $\Rightarrow$ X_1, X_2 are independent.

Conditioning:

For two events A, B, with $\mathbf{P}(B) > 0$, define:

$$\mathbf{P}(A|B) = \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(B)}\,.$$

The conditional distribution of X given Y:

$$\begin{aligned} F_{X|Y}(x|y) &= \mathbf{P}(X \le x | Y = y) \\ &\doteq \lim_{\epsilon\to 0} \mathbf{P}(X \le x \,|\, y - \epsilon < Y < y + \epsilon)\,. \end{aligned}$$

The conditional density:

$$p_{X|Y}(x|y) = \frac{\partial^n}{\partial x_1 \,\ldots\, \partial x_n}\, F_{X|Y}(x|y) = \frac{p_{XY}(x,y)}{p_Y(y)}\,.$$

In the following we simply write $p(x|y)$ etc. when no confusion arises.

Conditional Expectation:

$$E(X|Y = y) = \int_{\mathbb{R}^n} x\, p(x|y)\, dx\,.$$

Obviously, this is a function of y : $E(X|Y=y)=g(y)$.
Therefore, $E(X|Y) \stackrel{\triangle}{=} g(Y)$ is an RV, and a function of Y.

Basic properties:

* Smoothing: $E(E(X|Y))=E(X)$.
* Orthogonality principle:
  $E([X-E(X|Y)]\,h(Y))=0$ for every scalar function h.
* $E(X|Y)=E(X)$ if X and Y are independent.

Bayes Rule:

$$p(x|y)=\frac{p(x,y)}{p(y)}=\frac{p(y|x)p(x)}{\int p(y|x)p(x)\,dx}\,.$$

2.2 The Estimation Problem

The basic estimation problem is:

- Compute an estimate for an unknown quantity $x \in \mathcal{X} = \mathbb{R}^n$, based on measurements $y = (y_1, \ldots, y_m)' \in \mathbb{R}^m$.

Obviously, we need a model that relates y to x. For example,

$$y = h(x) + v$$

where h is a known function, and v a "noise" (or error) vector.

- An <u>estimator</u> $\hat{x}$ for x is a function

$$\hat{x} : y \mapsto \hat{x}(y) .$$

- The value of $\hat{x}(y)$ at a specific observed value y is an <u>estimate</u> of x.

Under different statistical assumptions, we have the following major solution concepts:

(i) Deterministic framework:

Here we simply look for x that minimizes the error in $y \simeq h(x)$. The most common criterion is the square norm:

$$\min_x \|y - h(x)\|^2 = \min_x \sum_{i=1}^{m} |y_i - h_i(x)|^2 .$$

This is the well-known (non-linear) <u>least-squares (LS)</u> problem.

(ii) Non-Bayesian framework:

Assume that y is a *random* function of x. For example,
$Y = h(x) + \mathbf{v}$, with $\mathbf{v}$ an RV. More generally, we are given, for each fixed x, the pdf $p(y|x)$ (i.e., $y \sim p(\cdot|x)$).
No statistical assumptions are made on x.
The main solution concept here is the MLE.

(iii) Bayesian framework:

Here we assume that both y and x are RVs with known joint statistics. The main solution concepts here are the MAP estimator and the optimal (MMSE) estimator.

A problem related to estimation is the *regression* problem: given measurements $(x_k, y_k)_{k=1}^{N}$, find a function h that gives the best fit $y_k \simeq h(x_k)$. h is the regressor, or regression function. We shall not consider this problem directly in this course.

2.3 The Bayesian Setting

In the Bayesian setting, we are given:

(*i*) $p_X(x)$ – the *prior* distribution for x.

(*ii*) $p_{Y|X}(y|x)$ – the conditional distribution of Y given $X = x$.

Note that $p(y|x)$ is often specified through an equation such as $Y = h(X, \mathbf{v})$ or $Y = h(X) + \mathbf{v}$, with $\mathbf{v}$ an RV, but this is immaterial for the theory.

We can now compute the posterior probability of X:

$$p(x|y) = \frac{p(y|x)p(x)}{\int p(y|x)p(x)\,dx}\,.$$

Given $p(x|y)$, what would be a reasonable choice for $\hat{x}$?

The two common choices are:

(i) The mean of X according to $p(x|y)$:

$$\hat{x}(y) = E(X|y) \equiv \int x\,p(x|y)\,dx\,.$$

(ii) The most likely value of X according to $p(x|y)$:

$$\hat{x}(y) = \arg\max_x \, p(x|y)$$

The first leads to the MMSE estimator, the second to the MAP estimator.

2.4 The MMSE Estimator

The Mean Square Error (MSE) of as estimator $\hat{x}$ is given by

$$\mathrm{MSE}(\hat{x}) \stackrel{\triangle}{=} E(\|X - \hat{x}(Y)\|^2)\,.$$

The Minimial Mean Square Error (MMSE) estimator, $\hat{x}_{\mathrm{MMSE}}$, is the one that minimizes the MSE.

Theorem: $\hat{x}_{\mathrm{MMSE}}(y) = E(X|Y = y)$.

Remarks:

1. Recall that conditional expectation $E(X|Y)$ satisfies the orthogonality principle (see above). This gives an easy proof of the theorem.
2. The MMSE estimator is *unbiased*: $E(\hat{x}_{\mathrm{MMSE}}(Y)) = E(X)$.
3. The *posterior* MSE is defined (for every y) as:

   $$\mathrm{MSE}\,(\hat{x}|y) = E(\|X - \hat{x}(y)\|^2 \,|\, Y = y)\,.$$

   with minimal value MMSE(y). Note that

   $$\begin{aligned} \mathrm{MSE}(\hat{x}) &= E\Big(E(\|X - \hat{x}(Y)\|^2 \,|Y)\Big) \\ &= \int\limits_y \mathrm{MSE}(\hat{x}|y)p(y)dy\,. \end{aligned}$$

   Since MSE($\hat{x}|y$) can be minimizing for each y separately, it follows that minimizing the MSE is *equivalent* to minimizing the posterior MSE for every y.

Some shortcomings of the MMSE estimator are:
– Hard to compute (except for special cases).
– May be inappropriate for multi-modal distributions.
– Requires the prior $p(x)$, which may not be available.

Example: The Gaussian Case.

Let X and Y be jointly Gaussian RVs with means

$$E(X) = m_X\,, \quad E(Y) = m_Y\,,$$

and covariance matrix

$$\mathrm{cov}\begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY} \end{pmatrix}\,.$$

By direct calculation, the posterior distribution $p_{X|Y=y}$ is Gaussian, with mean

$$m_{X|y} = m_X + \Sigma_{XY}\Sigma_{YY}^{-1}(y - m_Y)\,,$$

and covariance

$$\Sigma_{X|y} = \Sigma_{XX} - \Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX}\,.$$

(If Σ_{YY}^{-1} does not exist, it may be replaced by the pseudo-inverse.) Note that the posterior variance $\Sigma_{X|y}$ does not depend on the actual value y of Y!

It follows immediately that for the Gaussian case,

$$\hat{x}_{\mathrm{MMSE}}(y) \equiv E(X|Y = y) = m_{X|y}\,,$$

and the associated posterior MMSE equals

$$\mathrm{MMSE}(y) = E(\|X - \hat{x}_{\mathrm{MMSE}}(y)\|^2|Y = y) = \mathrm{trace}(\Sigma_{\mathrm{X|y}})\,.$$

Note that here $\hat{x}_{\mathrm{MMSE}}$ is a *linear* function of y. Also, the posterior MMSE does not depend on y.

2.5 The Linear MMSE Estimator

When the MMSE is too complicated we may settle for the best *linear* estimator. Thus, we look for $\hat{x}$ of the form:

$$\hat{x}(y) = Ay + b$$

that minimizes

$$\mathrm{MSE}\,(\hat{x}) = E\Big(\parallel X - \hat{x}(Y) \parallel^2\Big)\,.$$

The solution may be easily obtained by differentiation, and has exactly the same form as the MMSE estimator for the Gaussian case:

$$\hat{x}_{\mathrm{L}}(y) = m_X + \Sigma_{XY}\Sigma_{YY}^{-1}(y - m_Y)\,.$$

Note:

- The LMMSE estimator depends only on the first and second order statistics of X and Y.
- The linear MMSE does *not* minimize the *posterior* MSE, namely $\mathrm{MSE}\,(\hat{x}|y)$. This holds only in the Gaussian case, where the LMMSE and MMSE estimators coincide.
- The orthogonality principle here is:

  $$E\left((X - \hat{x}_{\mathrm{L}}(Y))\,L(Y)^T\right) = 0\,,$$

  for every *linear* function $L(y) = Ay + b$ of y.
- The LMMSE is unbiased: $E(\hat{x}_{\mathrm{L}}(Y)) = E(X)$.

2.6 The MAP Estimator

Still in the Bayesian setting, the MAP (Maximum a-Posteriori) estimator is defined as

$$\hat{x}_{\mathrm{MAP}}(y) \stackrel{\triangle}{=} \arg\max_x \, p(x|y) \, .$$

Noting that

$$p(x|y) = \frac{p(x,y)}{p(y)} = \frac{p(x)p(y|x)}{p(y)} \, ,$$

we obain the equivalent characterizations:

$$\begin{aligned} \hat{x}_{\mathrm{MAP}}(y) &= \arg\max_x \, p(x,y) \\ &= \arg\max_x \, p(x)p(y|x) \, . \end{aligned}$$

Motivation: Find the value of x which has the highest probability according to the posterior $p(x|y)$.

Example: In the Gaussian case, with $p(x|y) \sim N(m_{X|y}, \Sigma_{X|y})$, we have:

$$\hat{x}_{\mathrm{MAP}}(y) = \arg\max_x \, p(x|y) = m_{X|y} \equiv E(X|Y=y) \, .$$

Hence, $\hat{x}_{\mathrm{MAP}} \equiv \hat{x}_{\mathrm{MMSE}}$ for this case.

2.7 Non-Bayesian Setting – The ML Estimator

The MLE is defined in a non-Bayesian setting:

* No prior $p(x)$ is given. In fact, x need not be random.
* The distribution $p(y|x)$ of Y given x is given as before.

The MLE is defined by:

$$\hat{x}_{\mathrm{ML}}(y) = \arg\max_{x \in \mathcal{X}} p(y|x) \,.$$

It is convenient to define the *likelihood function* $L_y(x) = p(y|x)$ and the log-likelihood function $\Lambda_y(x) = \log L_y(x)$, and then we have

$$\hat{x}_{\mathrm{ML}}(y) = \arg\max_{x \in \mathcal{X}} L_y(x) \equiv \arg\max_{x \in \mathcal{X}} \Lambda_y(x) \,.$$

Note:

- Often x is denoted as θ in this context.
- Motivation: The value of x that makes y "most likely".
  This justification is merely heuristic!
- Compared with the MAP estimator:

  $$\hat{x}_{\mathrm{MAP}}(y) = \arg\max_{x} p(x)p(y|x) \,,$$

  we see that the MLE lacks the weighting of $p(y|x)$ by $p(x)$.
- The power of the MLE lies in:
  * its simplicity
  * good asymptotic behavior.

Example 1: Y is exponentially distributed with rate $x > 0$, namely $x = E(Y)^{-1}$. Thus:

$$
\begin{aligned}
& F(y|x) = (1 - e^{-xy})\, 1_{\{y \geq 0\}} \\
& p_{y|x}(y) = x\, e^{-xy}\, 1_{\{y \geq 0\}} \\
& \hat{x}_{\mathrm{ML}}(y) = \arg\max_{x \geq 0}\ x\, e^{-xy} \\
& \frac{d}{dx}\,(x\, e^{-xy}) = 0 \quad \Rightarrow \quad x = y^{-1} \\
& \hat{x}_{\mathrm{ML}}(y) = y^{-1}\,.
\end{aligned}
$$

Example 2 (Gaussian case):

$$
\begin{aligned}
& y = Hx + v \qquad\qquad (y \in \mathbb{R}^m\,,\ x \in \mathbb{R}^n) \\
& v \sim N(0, R_v) \\
& L_y(x) = p(y|x) = \frac{1}{c}\, e^{-\frac{1}{2}\,(y-Hx)^T\, R_v^{-1}(y-Hx)} \\
& \log\, L_y(x) = c_1 - \frac{1}{2}\,(y - Hx)^T\, R_v^{-1}(y - Hx) \\
& \hat{x}_{\mathrm{ML}} = \arg\min_x\, (y - Hx)^T\, R_v^{-1}(y - Hx)\,.
\end{aligned}
$$

This is a (weighted) LS problem! By differentiation,

$$
\begin{aligned}
& H^T R_v^{-1}(y - Hx) = 0\,, \\
& \hat{x}_{\mathrm{ML}} = (H^T R_v^{-1} H)^{-1}\, H^T\, R_v^{-1} y
\end{aligned}
$$

(assuming that $H^T R_v^{-1} H$ is invertible: in particular, $m \geq n$). $\square$

2.8 Bias and Covariance

Since the measurement y is random, the estimate $\hat{X} = \hat{x}(Y)$ is a random variable, and we can relate to its mean and variance.

The conditional mean of $\hat{x}$ is given by

$$\hat{m}(x) \stackrel{\triangle}{=} E(\hat{X}|x) \equiv E(\hat{X}|X = x) = \int \hat{x}(y)\, p(y|x)\, dy$$

The bias $\hat{x}$ is defined as

$$b(x) = E(\hat{X}|x) - x\,.$$

The of estimator $\hat{x}$ is *(conditionally) unbiased* if $b(x) = 0$ for every $x \in \mathcal{X}$.

The *covariance matrix* of $\hat{x}$ is,

$$\mathrm{cov}(\hat{x}|x) = E((\hat{X} - E(\hat{X}|x))(\hat{X} - E(\hat{X}|x)'|X = x)$$

In the scalar case, it follows by orthogonality that

$$\begin{aligned} \mathrm{MSE}(\hat{x}|x) &\equiv E((x - \hat{X})^2|x) = E((x - E(\hat{X}|x) + E(\hat{X}|x) - \hat{X})^2|x) \\ &= \mathrm{cov}(\hat{x}|x) + b(x)^2\,. \end{aligned}$$

Thus, if $\hat{x}$ is conditionally unbiased, $\mathrm{MSE}(\hat{x}|x) = \mathrm{cov}(\hat{x}|x)$.

Similarly, if x is vector-valued, then $\mathrm{MSE}(\hat{x}|x) = \mathit{trace}(\mathrm{cov}(\hat{x}|x)) + ||b(x)||^2$.

In the Bayesian case, we say that $\hat{x}$ is unbiased if $E(\hat{x}(Y)) = E(X)$. Note that the first expectation is both over X and Y.

2.9 The Cramer-Rao Lower Bound (CRLB)

The CRLB gives a lower bound on the MSE of any (unbiased) estimator. For illustration, we mention here the non-Bayesian version, with a scalar parameter x.

Assume that $\hat{x}$ is conditionally unbiased, namely $E_x(\hat{x}(Y)) = x$. (We use here $E_x(\cdot)$ for $E(\cdot|X = x)$). Then

$$MSE(\hat{x}|x) = E_x\{(\hat{x}(Y) - x)^2\} \geq J(x)^{-1}\,,$$

where J is the Fisher information:

$$\begin{aligned} J(x) &\triangleq - E_x\left\{\frac{\partial^2 \ln p(Y|x)}{\partial x^2}\right\} \\ &= E_x\left\{\left(\frac{\partial \ln p(Y|x)}{\partial x}\right)^2\right\}\,. \end{aligned}$$

An (unbiased) estimator that meets the above CRLB is said to be *efficient*.

2.10 Asymptotic Properties of the MLE

Suppose x is estimated based on multiple i.i.d. samples:

$$y = y^n = (y_1, \ldots, y_n), \text{ with } p(y^n|x) = \prod_{i=1}^{n} p_0(y_i|x).$$

For each $n \geq 1$, let $\hat{x}^n$ denote an estimator based on y^n. For example, $\hat{x}^n = \hat{x}^n_{\text{ML}}$.

We consider the asymptotic properties of $\{\hat{x}^n\}$, as $n \to \infty$.

Definitions: The (non-Bayesian) estimator sequence $\{\hat{x}^{(n)}\}$ is termed:

* *Consistent* if: $\lim_{n\to\infty} \hat{x}^n(Y^n) = x$ (w.p. 1).
* *Asymptotically unbiased* if: $\lim_{n\to\infty} E^x(\hat{x}^n(Y^n)) = x$.
* *Asymptotically efficient* if it satisfies the CRLB for $n \to \infty$, in the sense that: $\lim_{n\to\infty} J^n(x) \cdot \text{MSE}(\hat{x}^n) = 1$.
  Here $\text{MSE}(x^n) = E^x(\hat{x}^n(Y^n) - x)^2)$, and J^n is the Fisher information for y^n.

  For i.i.d. observations, $J^n = nJ^{(1)}$.

The ML Estimator $\hat{x}^n_{\text{ML}}$ is both *asymptotically unbiased* and *asymptotically efficient* (under mild technical conditions).