Estimation

Chapter Intended Learning Outcomes:

(i) Understand the basics of deterministic parameter estimation models

(ii) Able to apply probability and random variables to formulate the maximum likelihood estimators for Gaussian disturbance scenarios

(iii) Able to compute the maximum likelihood and least squares estimates for linear models

Estimation with Probability Models

Estimation refers to finding the parameters of interest under uncertainty.

Examples include:

- Computing the mean and covariance of random variables such as Examples 2.27, 3.7, 3.11, 3.15, 4.8 and 4.11
- Estimating the population intention on a certain proposition based on a sampled population such as Example 3.16
- Estimating unknown constant/deterministic values from a set of noisy measurements such as Example 3.18

When the uncertainty is characterized by known probability models, optimal estimation may be attained.

Deterministic Parameter Estimation Models

A generic model for estimating an unknown constant $x \in \mathbb{R}$ is:

$$\boldsymbol{r} = \boldsymbol{f}(x) + \boldsymbol{w} \quad (6.1)$$

where $\boldsymbol{r} = [r_1 \ \cdots \ r_N]^T \in \mathbb{R}^N$ is observation vector, signal $\boldsymbol{f}(x) = [f_1(x) \ \cdots \ f_N(x)]^T \in \mathbb{R}^N$ is a known function of x and $\boldsymbol{w} = [w_1 \ \cdots \ w_N]^T \in \mathbb{R}^N$ is noise vector.

The signal component $\boldsymbol{f}(x)$ is deterministic while the noise component is random specified by a probability model.

In estimating multiple parameters $\boldsymbol{x} = [x_1 \ \cdots \ x_M]^T \in \mathbb{R}^M$, (6.1) is generalized to

$$\boldsymbol{r} = \boldsymbol{f}(\boldsymbol{x}) + \boldsymbol{w} \quad (6.2)$$

In this model, the parameters are unknown deterministic values. The estimation problem is to find x or $\boldsymbol{x}$ given $\mathbf{r}$.

Example 6.1

Suggest four examples for estimation models in (6.1) or (6.2).

1. Estimation of a DC voltage A from a single observation r:

$$r = A + w, \quad w \sim \mathcal{N}(0, \sigma^2)$$

This example can be extended to N observations $r_1, \cdots, r_N$:

$$r_n = A + w_n, \quad n = 1, \cdots, N, \quad w_n \sim \mathcal{N}(0, \sigma^2)$$

or in vector form:

$$\boldsymbol{r} = A\mathbf{1}_N + \boldsymbol{w}, \quad \boldsymbol{w} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{C})$$

where $\mathbf{1}_N = [1 \cdots 1]^T \in \mathbb{R}^N$.

In both cases, we can write $\boldsymbol{f}(x) = [x \ \cdots \ x]^T = \mathbf{1}x$.

2. Polynomial fitting using N observation pairs $\{(x_n, y_n)\}_{n=1}^{N}$:

$$y_n = ax_n^2 + bx_n + c + w_n, \quad n = 1, \cdots, N, \quad w_n \sim \mathcal{N}(0, \sigma^2)$$

or

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{\theta} + \boldsymbol{w}, \quad \boldsymbol{w} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{C})$$

where

$$\boldsymbol{y} = [y_1 \ \cdots \ y_N]^T$$

$$\boldsymbol{\theta} = [a \ b \ c]^T$$

$$\boldsymbol{A} = \begin{bmatrix} x_1^2 & x_1 & 1 \\ \vdots & \vdots & \vdots \\ x_N^2 & x_N & 1 \end{bmatrix}$$

Here $\boldsymbol{f}(\boldsymbol{\theta}) = \boldsymbol{A}\boldsymbol{\theta}$.

3. Estimating $\boldsymbol{x} = [x_1 \ \cdots \ x_M]^T$ from N linear equations:

$$y_n = a_{n1}x_1 + a_{n2}x_2 + \cdots a_{nM}x_M + w_n, \quad n = 1, \cdots, N, \quad w_n \sim \mathcal{N}(0, \sigma^2)$$

or

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{w}, \quad \boldsymbol{w} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{C})$$

where

$$\boldsymbol{y} = [y_1 \ \cdots \ y_N]^T$$

$$\boldsymbol{x} = [x_1 \ \cdots \ x_M]^T$$

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1M} \\ \vdots & \vdots & \vdots & \vdots \\ a_{N1} & a_{N2} & \cdots & a_{NM} \end{bmatrix}$$

Here $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{x}$.

4. Estimating the amplitude, frequency and phase of a sinusoidal signal:

$$r_n = A\cos(\omega n + \phi) + w_n, \quad n = 1, \cdots, N, \quad w_n \sim \mathcal{N}(0, \sigma^2)$$

or

$$\boldsymbol{r} = \boldsymbol{f}(\boldsymbol{x}) + \boldsymbol{w}, \quad \boldsymbol{x} = [A \ \omega \ \phi]^T, \quad \boldsymbol{w} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{C})$$

Here, $\boldsymbol{f}(\boldsymbol{x})$ is a cosine function parameterized by $\boldsymbol{x}$.

This is a non-linear model as $\boldsymbol{f}(\boldsymbol{x})$ cannot be expressed in the form of $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{x}$.

On the other hand, the first 3 examples correspond to linear model because the functions can be written as $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{x}$.

Maximum Likelihood Estimation

When the probability density function (PDF) or probability mass function (PMF) of $\boldsymbol{r}$ is known, then $\boldsymbol{x}$ can be estimated by maximizing the likelihood function.

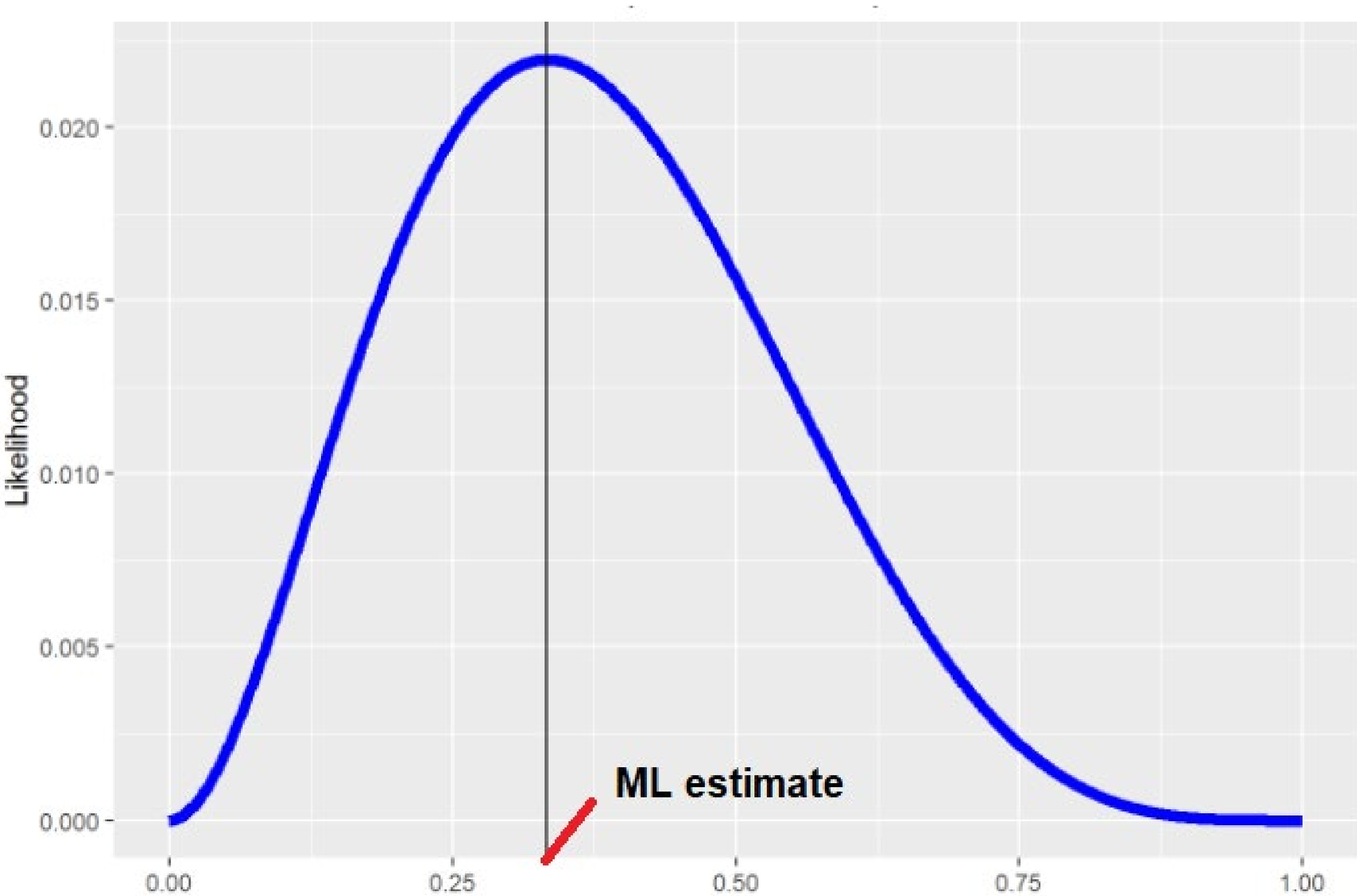
The likelihood function is the PDF or PMF parameterized by $\boldsymbol{x}$, denoted by $p(\boldsymbol{r};\boldsymbol{x})$, which is similar to the notation in (5.5).

The maximum likelihood (ML) estimate is given by

$$\hat{\boldsymbol{x}}_{\mathrm{ML}} = \arg\max_{\tilde{\boldsymbol{x}}} p(\boldsymbol{r};\tilde{\boldsymbol{x}}) \tag{6.3}$$

That is, $\hat{\boldsymbol{x}}_{\mathrm{ML}}$ is equal to the variable vector $\tilde{\boldsymbol{x}}$ which gives the maximum value of $p(\boldsymbol{r};\tilde{\boldsymbol{x}})$.

Given the likelihood function, the ML estimate for a scalar parameter is illustrated as follows.



Source: [Chapter 2 Beyond Least Squares: Using Likelihoods | Beyond Multiple Linear Regression (bookdown.org)]

Example 6.2

Consider a single measurement r which contains a constant A embedded in zero-mean Gaussian noise $w \sim \mathcal{N}(0, \sigma^2)$:

$$r = A + w, \quad p(w) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2} w^2}$$

Determine the ML estimate of A.

The likelihood function is the PDF parameterized by A:

$$p(r; A) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(r-A)^2} \Rightarrow r \sim \mathcal{N}(A, \sigma^2)$$

The ML estimate of A is:

$$\hat{A}_{\mathrm{ML}} = \arg\max_{\tilde{A}} p(r; \tilde{A}) = \arg\max_{\tilde{A}} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}\left(r-\tilde{A}\right)^2}$$

Clearly, $\hat{A}_{\mathrm{ML}} = r$ as r is a Gaussian random variable with $\mu = A$.

A formal calculation is given as follows. Maximizing $p(r;\tilde{A})$ is equivalent to maximizing its logarithm value $\ln(p(r;\tilde{A}))$, i.e., the value of $\tilde{A}$ maximizes $p(r;\tilde{A})$ also maximizes $\ln(p(r;\tilde{A}))$.

We have:

$$\begin{aligned}\hat{A}_{\mathrm{ML}} &= \arg\max_{\tilde{A}} -\ln\sqrt{2\pi\sigma^2} - \frac{1}{2\sigma^2}\left(r-\tilde{A}\right)^2 = \arg\max_{\tilde{A}} -\frac{1}{2\sigma^2}\left(r-\tilde{A}\right)^2 \\ &= \arg\min_{\tilde{A}} \frac{1}{2\sigma^2}\left(r-\tilde{A}\right)^2 = \arg\min_{\tilde{A}}\left(r-\tilde{A}\right)^2 = r\end{aligned}$$

Note also:

$$\left.\frac{d\left(r-\tilde{A}\right)^2}{d\tilde{A}}\right|_{\tilde{A}=\hat{A}_{\mathrm{ML}}} = 2\left(r-\hat{A}_{\mathrm{ML}}\right)(-1) = 0 \Rightarrow \hat{A}_{\mathrm{ML}} = r$$

<u>Example 6.3</u>

Repeat Example 6.2 with N measurements:

$$r_n = A + w_n, \quad n = 1, \cdots, N$$

or

$$\boldsymbol{r} = A\mathbf{1}_N + \boldsymbol{w}$$

where $\boldsymbol{r} = [r_1 \ \cdots \ r_N]^T$ and $\boldsymbol{w} = [w_1 \ \cdots \ w_N]^T$ with independent and identically distributed (IID) $w_n \sim \mathcal{N}(0, \sigma^2)$.

Using (3.38), we have

$$p(\boldsymbol{r}; A) = \frac{1}{(2\pi)^{N/2}\sigma^N} e^{-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(r_n - A)^2}$$

Maximizing $p(\boldsymbol{r}; \tilde{A})$ is equivalent to maximizing its logarithm value $\ln(p(\boldsymbol{r}; \tilde{A}))$, resulting in minimizing:

$$\hat{A}_{\rm ML} = \arg\min_{\tilde{A}} \sum_{n=1}^{N} (r_n - \tilde{A})^2$$

We perform differentiation with respect to $\tilde{A}$ and set the resultant expression to zero:

$$\left.\frac{d\sum_{n=1}^{N}(r_n - \tilde{A})^2}{d\tilde{A}}\right|_{\tilde{A}=\hat{A}_{\rm ML}} = \sum_{n=1}^{N} 2\left(r_n - \hat{A}_{\rm ML}\right)(-1) = 0$$

$$\Rightarrow \sum_{n=1}^{N} r_n = \sum_{n=1}^{N} \hat{A}_{\rm ML} = N\hat{A}_{\rm ML}$$

$$\Rightarrow \hat{A}_{\rm ML} = \frac{1}{N}\sum_{n=1}^{N} r_n$$

Hence the ML estimate of A is simply the average.

Least Squares Solution for Linear Model

Many science and engineering problems can be boiled down to estimating $\boldsymbol{x} \in \mathbb{R}^m$ from a system of linear noisy equations:

$$
\begin{aligned}
y_1 &= a_{11}x_1 + a_{12}x_2 + \cdots a_{1m}x_m + w_1 \\
y_2 &= a_{21}x_1 + a_{22}x_2 + \cdots a_{2m}x_m + w_2 \\
\cdots & \quad \cdots \quad \cdots \quad \cdots \quad \cdots \\
y_n &= a_{n1}x_1 + a_{n2}x_2 + \cdots a_{nm}x_m + w_n
\end{aligned}
$$

Or in compact matrix form:

$$
\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{w}, \quad \boldsymbol{y} \in \mathbb{R}^n, \ \boldsymbol{A} \in \mathbb{R}^{n \times m}, \ n \geq m \tag{6.4}
$$

A standard approach to solve for $\boldsymbol{x}$ is least squares (LS), whose idea is to minimize the sum of squared errors.

The LS cost function is:

$$J(\tilde{\boldsymbol{x}}) = (\boldsymbol{y} - \boldsymbol{A}\tilde{\boldsymbol{x}})^T \boldsymbol{W} (\boldsymbol{y} - \boldsymbol{A}\tilde{\boldsymbol{x}}) \tag{6.5}$$

where $\boldsymbol{W} \in \mathbb{R}^{n \times n}$ is a symmetric weighting matrix, i.e., $\boldsymbol{W} = \boldsymbol{W}^T$, whose purpose is to rely more on data with small noise, and rely less on data with large noise.

When $\boldsymbol{W}$ is a diagonal matrix such as $\boldsymbol{W} = \boldsymbol{I}_n$ or

$$\boldsymbol{W} = \mathrm{diag}(\alpha_1, \cdots, \alpha_n) = \begin{bmatrix} \alpha_1 & 0 & \cdots & 0 \\ 0 & \alpha_2 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & \alpha_n \end{bmatrix}$$

$J(\tilde{\boldsymbol{x}})$ can be easily written in scalar form:

$$\begin{aligned} J(\tilde{\boldsymbol{x}}) &= (\boldsymbol{y}-\boldsymbol{A}\tilde{\boldsymbol{x}})^T\boldsymbol{I}_n(\boldsymbol{y}-\boldsymbol{A}\tilde{\boldsymbol{x}}) = (\boldsymbol{y}-\boldsymbol{A}\tilde{\boldsymbol{x}})^T(\boldsymbol{y}-\boldsymbol{A}\tilde{\boldsymbol{x}}) \\ &= \sum_{i=1}^{n}(y_i - a_{i1}\tilde{x}_1 - a_{i2}\tilde{x}_2 \cdots - a_{im}\tilde{x}_m)^2 \\ &= \sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{m} a_{ij}\tilde{x}_j\right)^2 \end{aligned}$$

or

$$\begin{aligned} J(\tilde{\boldsymbol{x}}) &= (\boldsymbol{y}-\boldsymbol{A}\tilde{\boldsymbol{x}})^T\mathrm{diag}(\alpha_1,\cdots,\alpha_n)(\boldsymbol{y}-\boldsymbol{A}\tilde{\boldsymbol{x}}) \\ &= \sum_{i=1}^{n}\alpha_i(y_i - a_{i1}\tilde{x}_1 - a_{i2}\tilde{x}_2 \cdots - a_{im}\tilde{x}_m)^2 \\ &= \sum_{i=1}^{n}\alpha_i\left(y_i - \sum_{j=1}^{m} a_{ij}\tilde{x}_j\right)^2 \end{aligned}$$

The LS estimate is given by:

$$\hat{\boldsymbol{x}} = \arg\min_{\tilde{\boldsymbol{x}}} (\boldsymbol{y} - \boldsymbol{A}\tilde{\boldsymbol{x}})^T \boldsymbol{W} (\boldsymbol{y} - \boldsymbol{A}\tilde{\boldsymbol{x}}) \qquad (6.6)$$

Expanding (6.5) yields:

$$\begin{aligned} J(\tilde{\boldsymbol{x}}) &= (\boldsymbol{y} - \boldsymbol{A}\tilde{\boldsymbol{x}})^T \boldsymbol{W} (\boldsymbol{y} - \boldsymbol{A}\tilde{\boldsymbol{x}}) = (\boldsymbol{y}^T - \tilde{\boldsymbol{x}}^T \boldsymbol{A}^T) \boldsymbol{W} (\boldsymbol{y} - \boldsymbol{A}\tilde{\boldsymbol{x}}) \\ &= (\boldsymbol{y}^T \boldsymbol{W} - \tilde{\boldsymbol{x}}^T \boldsymbol{A}^T \boldsymbol{W})(\boldsymbol{y} - \boldsymbol{A}\tilde{\boldsymbol{x}}) \\ &= \boldsymbol{y}^T \boldsymbol{W} \boldsymbol{y} - 2\tilde{\boldsymbol{x}}^T \boldsymbol{A}^T \boldsymbol{W} \boldsymbol{y} + \tilde{\boldsymbol{x}}^T (\boldsymbol{A}^T \boldsymbol{W} \boldsymbol{A}) \tilde{\boldsymbol{x}} \end{aligned}$$

where each of them is a scalar.

Note that

$$\boldsymbol{y}\boldsymbol{W}\boldsymbol{A}\tilde{\boldsymbol{x}} = (\boldsymbol{y}\boldsymbol{W}\boldsymbol{A}\tilde{\boldsymbol{x}})^T = \tilde{\boldsymbol{x}}^T \boldsymbol{A}^T \boldsymbol{W}^T \boldsymbol{y} = \tilde{\boldsymbol{x}}^T \boldsymbol{A}^T \boldsymbol{W} \boldsymbol{y}$$

The required vector differentiation rules are:

$$\frac{d\boldsymbol{x}^T\boldsymbol{a}}{d\boldsymbol{x}} = \frac{d\boldsymbol{a}^T\boldsymbol{x}}{d\boldsymbol{x}} = \boldsymbol{a}$$

$$\frac{d\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}}{d\boldsymbol{x}} = 2\boldsymbol{A}\boldsymbol{x}, \ \boldsymbol{A} = \boldsymbol{A}^T$$

Differentiating $J(\tilde{\boldsymbol{x}})$ with respect to $\tilde{\boldsymbol{x}}$ and setting the resultant expression to zero, we obtain:

$$-2\boldsymbol{A}^T\boldsymbol{W}\boldsymbol{y} + 2\boldsymbol{A}^T\boldsymbol{W}\boldsymbol{A}\hat{\boldsymbol{x}}_{\mathrm{LS}} = \boldsymbol{0} \Rightarrow \boldsymbol{A}^T\boldsymbol{W}\boldsymbol{A}\hat{\boldsymbol{x}}_{\mathrm{LS}} = \boldsymbol{A}^T\boldsymbol{W}\boldsymbol{y}$$

The LS estimate is thus:

$$\hat{\boldsymbol{x}}_{\mathrm{LS}} = (\boldsymbol{A}^T\boldsymbol{W}\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{W}\boldsymbol{y} \tag{6.7}$$

Without the information of weighting matrix, we may just use $\boldsymbol{W} = \boldsymbol{I}_n$, leading to

$$\hat{\boldsymbol{x}}_{\mathrm{LS}} = (\boldsymbol{A}^T\boldsymbol{I}_n\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{I}_n\boldsymbol{y} = (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{y} \tag{6.8}$$

When $\boldsymbol{w}$ in (6.4) is jointly Gaussian distributed such that $\boldsymbol{w} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{C})$, we can write using (3.37):

$$p(\boldsymbol{y};\boldsymbol{x}) = \frac{1}{(2\pi)^{N/2}|\boldsymbol{C}|^{1/2}} e^{-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x})^T \boldsymbol{C}^{-1}(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x})} \quad (6.9)$$

Since $\mathbb{E}\{\boldsymbol{w}\} = \mathbf{0}$, the covariance matrix has the form of:

$$\boldsymbol{C} = \mathbb{E}\{\boldsymbol{w}\boldsymbol{w}^T\} = \begin{bmatrix} \mathbb{E}\{w_1^2\} & \mathbb{E}\{w_2 w_1\} & \cdots & \mathbb{E}\{w_n w_1\} \\ \mathbb{E}\{w_1 w_2\} & \mathbb{E}\{w_2^2\} & \cdots & \mathbb{E}\{w_n w_2\} \\ \vdots & \ddots & \ddots & \vdots \\ \mathbb{E}\{w_1 w_n\} & \cdots & \mathbb{E}\{w_{n-1} w_1\} & \mathbb{E}\{w_n^2\} \end{bmatrix}$$

Based on (6.7) and (6.9), the ML estimate of $\boldsymbol{x}$ for the general linear model is then

$$\begin{aligned}\hat{\boldsymbol{x}}_{\mathrm{ML}} &= \arg\max_{\tilde{\boldsymbol{x}}} p(\boldsymbol{y};\tilde{\boldsymbol{x}}) = \arg\max_{\tilde{\boldsymbol{x}}} \frac{1}{(2\pi)^{N/2}|\boldsymbol{C}|^{1/2}} e^{-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{A}\tilde{\boldsymbol{x}})^T\boldsymbol{C}^{-1}(\boldsymbol{y}-\boldsymbol{A}\tilde{\boldsymbol{x}})} \\ &= \arg\min_{\tilde{\boldsymbol{x}}} (\boldsymbol{y}-\boldsymbol{A}\tilde{\boldsymbol{x}})^T \boldsymbol{C}^{-1} (\boldsymbol{y}-\boldsymbol{A}\tilde{\boldsymbol{x}}) \\ &= (\boldsymbol{A}^T\boldsymbol{C}^{-1}\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{C}^{-1}\boldsymbol{y} \end{aligned} \quad (6.10)$$

That is, when the noise in the linear model is jointly Gaussian distributed with zero mean, the ML estimate is identical to the LS estimate of (6.7) with $\boldsymbol{W} = \boldsymbol{C}^{-1}$.

It can be shown that

$$\mathbb{E}\{\hat{\boldsymbol{x}}_{\mathrm{ML}}\} = \boldsymbol{x} \quad (6.11)$$

$$\mathrm{var}(\hat{x}_{m,\mathrm{ML}}) = \left[(\boldsymbol{A}^T\boldsymbol{C}^{-1}\boldsymbol{A})^{-1}\right]_{m,m} \quad (6.12)$$

That is, the ML estimate is unbiased and the variance of the estimate of x_m is the mth diagonal element of $(\boldsymbol{A}^T\boldsymbol{C}^{-1}\boldsymbol{A})^{-1}$.

Example 6.4

Show that the LS solution for Example 6.3 is also the ML estimate.

Recall the measurement model:

$$\boldsymbol{r} = \mathbf{1}_N A + \boldsymbol{w} \qquad \text{or} \qquad r_n = A + w_n, \quad n = 1, \cdots, N$$

Using (3.38) again, the weighting matrix is

$$\boldsymbol{W} = \boldsymbol{C}^{-1} = \sigma^{-2} \boldsymbol{I}_N$$

The LS cost function is:

$$\begin{aligned} J(\tilde{A}) &= (\boldsymbol{r} - \mathbf{1}_N \tilde{A})^T \cdot \sigma^{-2} \boldsymbol{I}_N \cdot (\boldsymbol{r} - \mathbf{1}_N \tilde{A}) = \sigma^{-2} (\boldsymbol{r} - \mathbf{1}_N \tilde{A})^T (\boldsymbol{r} - \mathbf{1}_N \tilde{A}) \\ &= \sum_{n=1}^{N} \sigma^{-2} \left(r_n - \tilde{A} \right)^2 \end{aligned}$$

We apply (6.7) by replacing $\boldsymbol{y}$, $\boldsymbol{A}$ and $\boldsymbol{x}$ by $\boldsymbol{r}$, $\mathbf{1}_N$ and A, respectively:

$$\hat{A}_{\mathrm{LS}} = (\mathbf{1}_N^T(\sigma^{-2}\boldsymbol{I}_N)\mathbf{1}_N)^{-1}\mathbf{1}_N^T(\sigma^{-2}\boldsymbol{I}_N)\boldsymbol{r} = (\mathbf{1}_N^T\mathbf{1}_N)^{-1}\mathbf{1}_N^T\boldsymbol{r} = \frac{1}{N}\sum_{n=1}^{N} r_n = \hat{A}_{\mathrm{ML}}$$

Note that scalar differentiation can be applied to achieve the same result as in Example 6.3, but (6.7)-(6.8) have a compact form realized by matrix operations.

According to (6.11) and (6.12), we have:

$$\mathbb{E}\{\hat{A}_{\mathrm{ML}}\} = A$$

$$\mathrm{var}\left(\hat{A}_{\mathrm{ML}}\right) = (\mathbf{1}_N^T(\sigma^{-2}\boldsymbol{I}_N)\mathbf{1}_N)^{-1} = \sigma^2(\mathbf{1}_N^T\mathbf{1}_N)^{-1} = \sigma^2 N^{-1}$$

which align with the calculation in Example 3.18.

Example 6.5

Given 2 measurements:

$$\boldsymbol{r} = A\mathbf{1}_2 + \boldsymbol{w}, \ \boldsymbol{r} = [r_1 \ r_2]^T, \ \boldsymbol{w} = [w_1 \ w_2]^T, \ \boldsymbol{w} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{C})$$

or

$$r_n = A + w_n, \quad n = 1, 2$$

where

$$\boldsymbol{C} = \mathbb{E}\{\boldsymbol{w}\boldsymbol{w}^T\} = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$$

Determine the ML estimate of A. Perform a MATLAB simulation to compare with the estimate based on average using $\sigma_1^2 = 0.1$ and $\sigma_2^2 = 10$.

Clearly, $w_1 \sim \mathcal{N}(0, \sigma_1^2)$ and $w_2 \sim \mathcal{N}(0, \sigma_2^2)$ are independent. We have

$$\boldsymbol{C} = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix} \Rightarrow \boldsymbol{C}^{-1} = \frac{1}{\sigma_1^2 \sigma_2^2} \begin{bmatrix} \sigma_2^2 & 0 \\ 0 & \sigma_1^2 \end{bmatrix} = \begin{bmatrix} \dfrac{1}{\sigma_1^2} & 0 \\ 0 & \dfrac{1}{\sigma_2^2} \end{bmatrix}$$

The ML estimate is computed using (6.10) as:

$$\hat{A}_{\mathrm{ML}} = \frac{\mathbf{1}_2^T \boldsymbol{C}^{-1} \boldsymbol{r}}{\mathbf{1}_2^T \boldsymbol{C}^{-1} \mathbf{1}_2} = \frac{\dfrac{r_1}{\sigma_1^2} + \dfrac{r_2}{\sigma_2^2}}{\dfrac{1}{\sigma_1^2} + \dfrac{1}{\sigma_2^2}}$$

The measurement with smaller noise power will have a larger weight in computing $\hat{A}_{\mathrm{ML}}$, e.g., if $\sigma_1^2 < \sigma_2^2$, then r_1 dominates, and vice versa. Also, if $\sigma_2^2 \to \infty$, then $\hat{A}_{\mathrm{ML}} \to r_1$.

According to (6.11) and (6.12), we have:

$$\mathbb{E}\{\hat{A}_{\mathrm{ML}}\} = A$$

$$\mathrm{var}\left(\hat{A}_{\mathrm{ML}}\right) = \left(\mathbf{1}_2^T \boldsymbol{C}^{-1} \mathbf{1}_2\right)^{-1} = \left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}\right)^{-1} = \frac{\sigma_1^2 \sigma_2^2}{\sigma_1^2 + \sigma_2^2}$$

The estimate based on averaging is:

$$\hat{A}_{\mathrm{AV}} = \frac{r_1 + r_2}{2}$$

Suppose now $\sigma_1^2 = 0.1$ and $\sigma_2^2 = 10$. We generate 10000 sets of $\boldsymbol{r}$ with $A = 10$.

```
>>W=[randn(1,10000)*sqrt(0.1);randn(1,10000)*sqrt(10)];
%10000 columns of noise vectors
```

```
>> mean(W(1,:).*W(1,:))
ans = 0.1005
>> mean(W(2,:).*W(2,:))
ans = 9.8157
>> mean(W(1,:).*W(2,:))
ans = 0.0076
```

The empirical covariance matrix is:

$$\hat{C} = \begin{bmatrix} 0.1005 & 0.0076 \\ 0.0076 & 9.8157 \end{bmatrix} \approx \begin{bmatrix} 0.1 & 0 \\ 0 & 10 \end{bmatrix}$$

```
>> A=10*ones(2,10000); %10000 columns of [10 10]
>> R=A+W; %10000 columns of measurement vectors
>> m=mean(R); %vector contains 10000 average
>> mean(m) % empirical mean estimate
ans =  10.0082
>> var(m,1) % empirical variance
ans = 2.4828
```

```
>> o=(R(1,:)/0.1+R(2,:)/10)/10.1; %ML estimates
>> mean(o)
ans = 10.0015
>> var(o,1)
ans = 0.0996
```

We can see that in the mean sense, both give unbiased estimation but the ML solution provides much smaller variance.

Note that

$$\mathrm{var}\left(\hat{A}_{\mathrm{ML}}\right) = \frac{0.1 \cdot 10}{0.1 + 10} = 0.099$$

is also validated.

What is the mean square error of the estimate?

Example 6.6

Given N noisy measurements of the form:

$$y_n = \alpha n + \beta + w_n, \quad n = 1, \cdots, N, \quad w_n \sim \mathcal{N}(0, \sigma^2)$$

where $\{w_n\}$ are IID. Compute the ML estimates of α and β as well as their variances.

It is clear that this straight line fitting problem is a linear model. According to Example 6.1, we can write:

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{w}$$

where

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}, \quad \boldsymbol{x} = \begin{bmatrix} \theta \\ \beta \end{bmatrix}, \quad \boldsymbol{A} = \begin{bmatrix} 1 & 1 \\ 2 & 1 \\ \vdots & \vdots \\ N & 1 \end{bmatrix}$$

As the noise is IID, $\boldsymbol{C}^{-1} = \sigma^{-2}\boldsymbol{I}_N$. Using (6.10) and (6.12):

$$\hat{\boldsymbol{x}}_{\mathrm{ML}} = \begin{bmatrix} \hat{\alpha}_{\mathrm{ML}} \\ \hat{\beta}_{\mathrm{ML}} \end{bmatrix} = (\boldsymbol{A}^T(\sigma^{-2}\boldsymbol{I}_N)\boldsymbol{A})^{-1}\boldsymbol{A}^T(\sigma^{-2}\boldsymbol{I}_N)\boldsymbol{y} = (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{y}$$

$$= \left(\begin{bmatrix} 1 & 1 \\ 2 & 1 \\ \vdots & \vdots \\ \vdots & \vdots \\ N & 1 \end{bmatrix}^T \begin{bmatrix} 1 & 1 \\ 2 & 1 \\ \vdots & \vdots \\ \vdots & \vdots \\ N & 1 \end{bmatrix} \right)^{-1} \begin{bmatrix} 1 & 1 \\ 2 & 1 \\ \vdots & \vdots \\ \vdots & \vdots \\ N & 1 \end{bmatrix}^T \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}$$

$$\mathrm{var}\left(\hat{\alpha}_{\mathrm{ML}}\right) = \sigma^2 \left[(\boldsymbol{A}^T\boldsymbol{A})^{-1}\right]_{1,1}$$

$$\mathrm{var}\left(\hat{\beta}_{\mathrm{ML}}\right) = \sigma^2 \left[(\boldsymbol{A}^T\boldsymbol{A})^{-1}\right]_{2,2}$$

We can use MATLAB to verify the results, e.g., by setting $N = 5$, $\alpha = 2$, $\beta = 1$, and $\sigma^2 = 0.01$.

```
>> n=1:5; %data length is 5
y=(2.*n+1).'; %noise-free y
Y = repmat(y,1,10000)+0.1.*randn(5,10000); %add noise
A=[1 2 3 4 5; 1 1 1 1 1].'
R=inv(A.'*A)*A.'*Y;
mean(R,2) %compute the mean
ans = 2.0003
      0.9991
>> var(R(1,:),1)
ans = 0.0010
>> var(R(2,:),1)
ans = 0.0111
>> 0.01.*inv(A.'*A)
ans = 0.0010   -0.0030
     -0.0030    0.0110
```

Hence the unbiasedness of the ML solution as well as their variances are validated.

The ML estimator can also be applied to discrete PMF such as binomial distribution.

Suppose we obtain $0 \leq r \leq n$ successes out of n independent trials and assume the probability of success is same for each trial, say, θ.

Now we want to find the most probable value of θ. The corresponding likelihood is then:

$$p(r;\theta) = C(n,r)\theta^r(1-\theta)^{n-r}$$

Note that here r is the measurement which depends on θ, and the ML estimate $\hat{\theta}$ is

$$\hat{\theta} = \arg\max_{\theta} C(n,r)\theta^r(1-\theta)^{n-r}$$

This is equivalent to finding the maximum of

$$r \ln \theta + (n-r)\ln(1-\theta) \Rightarrow r \cdot \frac{1}{\hat{\theta}} + (n-r) \cdot \frac{1}{1-\hat{\theta}} \cdot (-1) = 0 \Rightarrow \hat{\theta} = \frac{r}{n}$$

This aligns with the binomial PMF that given n and p, the most probable value of r is $r = np$ because $\mathbb{E}\{r\} = np$.

For example, we consider flipping a coin 10 times and obtain 7 heads (H), and let $P(\mathrm{H}) = \theta$. The corresponding likelihood is:

$$C(10,7)\theta^7(1-\theta)^3$$

The ML estimate of $P(\mathrm{H})$ is then:

$$\hat{\theta} = \frac{7}{10}$$

References:

1. S. Kay, *Fundamentals of Statistical Signal Processing, Volume I: Estimation Theory*, Prentice Hall, 1993
2. M. Stamp, *Introduction to Machine Learning with Applications in Information Security*, Chapman & Hall/CRC, 2017