Chapter 4

Introduction to Estimation Theory

4.1 Concepts of Probabilistic Estimation

The problem we are interest in this lecture is that of estimating the value of an $n-$dimensional vector of parameters $\mathbf{w}$, of a given system, on the basis of p observations taken on these parameters, and stacked in a p dimensional observation vector $\mathbf{z}$. We refer to $\hat{\mathbf{w}}$ as the estimate of the vector of parameters $\mathbf{w}$ under investigation, and we refer to the quantity $\tilde{\mathbf{w}} = \hat{\mathbf{w}} - \mathbf{w}$ as the estimation error. Based on the statistical formulation of the problem, we assume that the observational process is imperfect, and therefore the observations can be considered realizations of a random variable. Analogously, the vector of parameters $\mathbf{w}$ is seen as a quantity belonging to realizations of another random vector.

4.1.1 Bayesian Approach

In Bayesian estimation theory we introduce a functional $\mathcal{J}$ which corresponds to a measure of the "risk" involved in the estimate obtained for the parameter $\mathbf{w}$. That is, we define

$$\begin{aligned} \mathcal{J}(\hat{\mathbf{w}}) &\equiv \mathcal{E}\{J(\tilde{\mathbf{w}})\} \\ &= \int_{-\infty}^{\infty} J(\tilde{\mathbf{w}})\, p_{\mathbf{w}}(\mathbf{w})\, d\mathbf{w} \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} J(\tilde{\mathbf{w}})\, p_{\mathbf{wz}}(\mathbf{w},\mathbf{z})\, d\mathbf{z}\, d\mathbf{w} \end{aligned} \tag{4.1}$$

where $p_{\mathbf{w}}(\mathbf{w})$ is the marginal probability density of $\mathbf{w}$, $p_{\mathbf{wz}}(\mathbf{w},\mathbf{z})$ is the joint probability density of the random variables $\mathbf{w}$ and $\mathbf{z}$, and the function $J(\tilde{\mathbf{w}})$ is the one that provides the risk evaluation criteria, many times referred to as the cost function. The problem of determining an estimate $\hat{\mathbf{w}}$ gets reduced to that of minimizing the risk, or expected cost value, by means of an appropriate choice of the functional $J(\tilde{\mathbf{w}})$. We refer to the value of $\hat{\mathbf{w}}$ providing the minimum as the *optimal estimate*.

In general, the optimal estimate depends on the cost function being employed. Example of

two common cost functions are the quadratic cost function,

$$J = ||\tilde{\mathbf{w}}||_{\mathbf{E}}^2 = \tilde{\mathbf{w}}^T \mathbf{E} \tilde{\mathbf{w}} \,, \tag{4.2}$$

where the $n \times n$ matrix $\mathbf{E}$ is assumed to be non–negative and symmetric; and the uniform cost function,

$$J = \begin{cases} 0 \,, & ||\tilde{\mathbf{w}}|| < \epsilon \\ 1/2\epsilon \,, & ||\tilde{\mathbf{w}}|| \geq \epsilon \end{cases} \,. \tag{4.3}$$

However, for a large class of estimation problems, the resulting estimate is independent of the choice of the cost function.

A desirable property of an estimate is that it be *unbiased*, that is, that its ensemble average equals the ensemble average of the variable of interest. This is expressed mathematically as

$$\mathcal{E}\{\hat{\mathbf{w}}\} = \mathcal{E}\{\mathbf{w}\} \tag{4.4}$$

or in other words, the estimation error is zero: $\mathcal{E}\{\tilde{\mathbf{w}}\} = \mathbf{0}$. Estimates satisfying the equality above are said to be *unconditionally unbiased*, which is more general than being a *conditionally unbiased* estimate, that is obeying

$$\mathcal{E}\{\hat{\mathbf{w}}|\mathbf{w}\} = \mathbf{w} \,. \tag{4.5}$$

4.1.2 Minimum Variance Estimation

The minimum variance estimate, denoted $\hat{\mathbf{w}}_{\text{MV}}$, minimizes the risk function with the cost function given by (4.2). Therefore, the risk function to be minimized is written explicitly as

$$\mathcal{J}_{\text{MV}}(\hat{\mathbf{w}}) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (\mathbf{w} - \hat{\mathbf{w}})^T \mathbf{E} (\mathbf{w} - \hat{\mathbf{w}}) \, p_{\mathbf{wz}}(\mathbf{w}, \mathbf{z}) \, d\mathbf{z} \, d\mathbf{w} \tag{4.6}$$

which, using the definition of conditional probability distribution (1.77), can also be written as

$$\mathcal{J}_{\text{MV}}(\hat{\mathbf{w}}) = \int_{-\infty}^{\infty} \left\{ \int_{-\infty}^{\infty} (\mathbf{w} - \hat{\mathbf{w}})^T \mathbf{E} (\mathbf{w} - \hat{\mathbf{w}}) \, p_{\mathbf{w}|\mathbf{z}}(\mathbf{w}|\mathbf{z}) \, d\mathbf{w} \right\} p_{\mathbf{z}}(\mathbf{z}) \, d\mathbf{z} \,. \tag{4.7}$$

The outer integral does not involve $\hat{\mathbf{w}}$, and since the marginal probability density $p_{\mathbf{z}}(\mathbf{z})$ is always positive, we see that to search for the minimum of $\mathcal{J}_{\text{MV}}$ is equivalent to minimizing the integral in the kernel of the expression above. The kernel can be identified as an expression for the conditional Bayes risk, that is,

$$\mathcal{J}_{\text{MV}}(\hat{\mathbf{w}}|\mathbf{z}) \equiv \int_{-\infty}^{\infty} (\mathbf{w} - \hat{\mathbf{w}})^T \mathbf{E} (\mathbf{w} - \hat{\mathbf{w}}) \, p_{\mathbf{w}|\mathbf{z}}(\mathbf{w}|\mathbf{z}) \, d\mathbf{w} \tag{4.8}$$

which is what we want to minimize with respect to $\hat{\mathbf{w}}$.

Using the definition of differentiation of a *scalar* function $f = f(\mathbf{x})$ of an n–dimensional *vector* $\mathbf{x}$, that is,

$$\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} \equiv \begin{pmatrix} \frac{\partial f(\mathbf{x})}{\partial x_1} \\ \frac{\partial f(\mathbf{x})}{\partial x_2} \\ \vdots \\ \frac{\partial f(\mathbf{x})}{\partial x_n} \end{pmatrix} \tag{4.9}$$

we can show that for a constant n–vector $\mathbf{a}$ we have

$$\frac{\partial \mathbf{a}^T \mathbf{x}}{\partial \mathbf{x}} = \frac{\partial \mathbf{x}^T \mathbf{a}}{\partial \mathbf{x}} = \mathbf{a}\,. \tag{4.10}$$

Moreover, for an $n \times n$ symmetric matrix $\mathbf{A}$ we have

$$\frac{\partial \mathbf{x}^T \mathbf{A} \mathbf{x}}{\partial \mathbf{x}} = 2\mathbf{A}\mathbf{x}\,. \tag{4.11}$$

Applying these rules of differentiation to the minimization of $\mathcal{J}_{\mathrm{MV}}(\hat{\mathbf{w}}|\mathbf{z})$ it follows that

$$\mathbf{0} = \left.\frac{\partial \mathcal{J}_{\mathrm{MV}}(\hat{\mathbf{w}}|\mathbf{z})}{\partial \hat{\mathbf{w}}}\right|_{\hat{\mathbf{w}}=\hat{\mathbf{w}}_{\mathrm{MV}}} = -\ 2\,\mathbf{E} \left.\int_{-\infty}^{\infty} (\mathbf{w} - \hat{\mathbf{w}})\, p_{\mathbf{w}|\mathbf{z}}(\mathbf{w}|\mathbf{z})\, d\mathbf{w}\right|_{\hat{\mathbf{w}}=\hat{\mathbf{w}}_{\mathrm{MV}}} \tag{4.12}$$

and for any $\mathbf{E}$,

$$\hat{\mathbf{w}}_{\mathrm{MV}} \int_{-\infty}^{\infty} p_{\mathbf{w}|\mathbf{z}}(\mathbf{w}|\mathbf{z})\, d\mathbf{w} = \int_{-\infty}^{\infty} \mathbf{w} p_{\mathbf{w}|\mathbf{z}}(\mathbf{w}|\mathbf{z})\, d\mathbf{w} \tag{4.13}$$

since the integral of $p_{\mathbf{w}|\mathbf{z}}$ is unity (because p is a probability density), hence

$$\begin{aligned} \hat{\mathbf{w}}_{\mathrm{MV}}(\mathbf{z}) &= \int_{-\infty}^{\infty} \mathbf{w} p_{\mathbf{w}|\mathbf{z}}(\mathbf{w}|\mathbf{z})\, d\mathbf{w} \\ &= \mathcal{E}\{\mathbf{w}|\mathbf{z}\} \end{aligned} \tag{4.14}$$

This estimate has the desirable property of being unbiased. This can be shown simply as

$$\begin{aligned} \mathcal{E}\{\tilde{\mathbf{w}}\} &= \mathcal{E}\{\mathbf{w} - \hat{\mathbf{w}}_{\mathrm{MV}}\} \\ &= \mathcal{E}\{\mathbf{w} - \mathcal{E}\{\mathbf{w}|\mathbf{z}\}\} \\ &= \mathcal{E}\{\mathbf{w}\} - \mathcal{E}\{\mathcal{E}\{\mathbf{w}|\mathbf{z}\}\} \\ &= \mathcal{E}\{\mathbf{w}\} - \mathcal{E}\{\mathbf{w}\} \\ &= \mathbf{0} \end{aligned} \tag{4.15}$$

where the fourth equality follows from the chain rule for expectation operators in (1.84).

That the solution (4.14) is in fact a minimum of $\mathcal{J}_{\mathrm{MV}}(\hat{\mathbf{w}}|\mathbf{z})$ can be seen by calculating the second derivative of this quantity with respect to $\hat{\mathbf{w}}$, that is,

$$\frac{\partial^2 \mathcal{J}_{\mathrm{MV}}(\hat{\mathbf{w}}|\mathbf{z})}{\partial \hat{\mathbf{w}}^2} = 2\,\mathbf{E} \tag{4.16}$$

and since $\mathbf{E}$ is a non-negative matrix, the second derivative is non-negative, therefore the solution represents a minimum. Notice the extremely important fact that the estimate with minimum error variance (4.14) corresponds to the conditional mean. Substitution of (4.14) in expression (4.6) provides the Bayes risk with minimum error variance.

4.1.3 Maximum *a posteriori* Probability Estimation

Another estimator is defined through the risk function for the uniform cost function (4.3), and can be written explicitly as

$$\mathcal{J}_U(\hat{\mathbf{w}}) = \int_{-\infty}^{\infty} \left\{ \int_{-\infty}^{\infty} J(\tilde{\mathbf{w}})\, p_{\mathbf{w}|\mathbf{z}}(\mathbf{w}|\mathbf{z})\, d\mathbf{w} \right\} p_{\mathbf{z}}(\mathbf{z})\, d\mathbf{z}$$

$$= \int_{-\infty}^{\infty} \left\{ \frac{1}{2\epsilon} \int_{-\infty}^{\hat{\mathbf{w}}-\epsilon} p_{\mathbf{w}|\mathbf{z}}(\mathbf{w}|\mathbf{z})\, d\mathbf{w} + \frac{1}{2\epsilon} \int_{\hat{\mathbf{w}}+\epsilon}^{\infty} p_{\mathbf{w}|\mathbf{z}}(\mathbf{w}|\mathbf{z})\, d\mathbf{w} \right\} p_{\mathbf{z}}(\mathbf{z})\, d\mathbf{z} \tag{4.17}$$

where, some caution is needed in reading the integrals inside the brackets: these are multiple integrals and the notation $\hat{\mathbf{w}} \pm \epsilon$ should be interpreted as $\hat{w}_1 \pm \epsilon$, $\hat{w}_2 \pm \epsilon$, and so on, for each one of the n components of the vector $\hat{\mathbf{w}}$. Since $p_{\mathbf{w}|\mathbf{z}}$ is a probability density function its integral over the whole R^n domain is unity, consequently the Bayes risk function can be written as

$$\mathcal{J}_U(\hat{\mathbf{w}}) = \int_{-\infty}^{\infty} \frac{1}{2\epsilon} \left\{ 1 - \int_{\hat{\mathbf{w}}-\epsilon}^{\hat{\mathbf{w}}+\epsilon} p_{\mathbf{w}|\mathbf{z}}(\mathbf{w}|\mathbf{z})\, d\mathbf{w} \right\} p_{\mathbf{z}}(\mathbf{z})\, d\mathbf{z}\,. \tag{4.18}$$

For the problem of minimizing $\mathcal{J}_U$ with respect to $\hat{\mathbf{w}}$, the first term gives no relevant contribution, thus we can think of minimizing

$$\mathcal{J}_U(\hat{\mathbf{w}}) \sim -(1/2\epsilon) \int_{-\infty}^{\infty} \left\{ \int_{\hat{\mathbf{w}}-\epsilon}^{\hat{\mathbf{w}}+\epsilon} p_{\mathbf{w}|\mathbf{z}}(\mathbf{w}|\mathbf{z})\, d\mathbf{w} \right\} p_{\mathbf{z}}(\mathbf{z})\, d\mathbf{z}\,. \tag{4.19}$$

or yet, we can minimize the conditional Bayes risk

$$\mathcal{J}_U(\hat{\mathbf{w}}|\mathbf{z}) \equiv -(1/2\epsilon) \int_{\hat{\mathbf{w}}-\epsilon}^{\hat{\mathbf{w}}+\epsilon} p_{\mathbf{w}|\mathbf{z}}(\mathbf{w}|\mathbf{z})\, d\mathbf{w} \tag{4.20}$$

since $p_{\mathbf{z}}(\mathbf{z})$ is positive. As $\epsilon \geq 0$ approaches 0, the mean value theorem for integrals[1] can be employed to produce

$$\mathcal{J}_U(\hat{\mathbf{w}}|\mathbf{z}) = -\, p_{\mathbf{w}|\mathbf{z}}(\hat{\mathbf{w}}|\mathbf{z}) \tag{4.21}$$

which can also be obtained by noticing that as ϵ approaches zero the cost function $J(\tilde{\mathbf{w}})$ turns into a common representation for the negative of the delta function, in an n–dimensional space, that is, the cost function becomes

$$J(\tilde{\mathbf{w}}) \to -\prod_{i=1}^{n} \delta(w_i - \hat{w}_i)\,. \tag{4.22}$$

Minimization of $\mathcal{J}_U(\hat{\mathbf{w}}|\mathbf{z})$ is equivalent to maximization of the conditional probability density function $p_{\mathbf{w}|\mathbf{z}}(\hat{\mathbf{w}}|\mathbf{z})$. The value $\hat{\mathbf{w}} = \hat{\mathbf{w}}_{\mathrm{MAP}}$ that maximizes this quantity is known as the maximum *a posteriori* probability (MAP) estimate, and is determined by means of

$$\left. \frac{\partial p_{\mathbf{w}|\mathbf{z}}(\hat{\mathbf{w}}|\mathbf{z})}{\partial \hat{\mathbf{w}}} \right|_{\hat{\mathbf{w}}=\hat{\mathbf{w}}_{\mathrm{MAP}}} = \mathbf{0}\,, \tag{4.23}$$

which is the same as

$$\left. \frac{\partial p_{\mathbf{w}|\mathbf{z}}(\mathbf{w}|\mathbf{z})}{\partial \mathbf{w}} \right|_{\mathbf{w}=\hat{\mathbf{w}}_{\mathrm{MAP}}} = \mathbf{0}\,, \tag{4.24}$$

[1] The mean value theorem for integrals (e.g., Butkov [22]) can be stated as:

$$(1/2\epsilon) \int_{-\epsilon}^{\epsilon} f(x)\, dx = (1/2\epsilon)(2\epsilon) f(\xi) = f(\xi)$$

for $-\epsilon \leq \xi \leq \epsilon$.

since that the variables $\mathbf{w}$ and $\hat{\mathbf{w}}$ play the role of "dummy" derivation variables. Knowing that $p_{\mathbf{w}|\mathbf{z}}$ is really a function of $\mathbf{w}$, we prefer to use (4.24) rather than (4.23) to avoid confusion. The designation *a posteriori* refers to the fact that the estimate is obtained *after* the observations have been collected, that is, probability of $\mathbf{w}$ *given* $\mathbf{z}$. An estimate of this type is briefly described in (1.29), consequently we can identify maximum *a posteriori* probability estimation with *mode* estimation.

To maximize the probability density above is also equivalent to maximize its natural logarithm, $\ln p_{\mathbf{w}|\mathbf{z}}(\mathbf{w}|\mathbf{z})$, with respect to $\mathbf{w}$. Using Bayes rule (1.79) we can write

$$\ln p_{\mathbf{w}|\mathbf{z}}(\mathbf{w}|\mathbf{z}) = \ln[p_{\mathbf{z}|\mathbf{w}}(\mathbf{z}|\mathbf{w})p_{\mathbf{w}}(\mathbf{w})] - \ln p_{\mathbf{z}}(\mathbf{z}) \tag{4.25}$$

and since $p_{\mathbf{z}}(\mathbf{z})$ does not depend on $\mathbf{w}$ the maximum *a posteriori* probability estimate can be obtained by solving either

$$\left.\frac{\partial \ln[p_{\mathbf{z}|\mathbf{w}}(\mathbf{z}|\mathbf{w})p_{\mathbf{w}}(\mathbf{w})]}{\partial \mathbf{w}}\right|_{\mathbf{w}=\hat{\mathbf{w}}_{\mathrm{MAP}}} = \mathbf{0}\,, \tag{4.26}$$

or

$$\left.\frac{\partial p_{\mathbf{z}|\mathbf{w}}(\mathbf{z}|\mathbf{w})p_{\mathbf{w}}(\mathbf{w})}{\partial \mathbf{w}}\right|_{\mathbf{w}=\hat{\mathbf{w}}_{\mathrm{MAP}}} = \mathbf{0}\,. \tag{4.27}$$

In general, the unbiasedness of the estimate is not necessarily guaranteed in this case.

4.1.4 Maximum Likelihood Estimation

In maximum *a posteriori* probability estimation it is necessary to know the probability density of the process of interest, that is $p_{\mathbf{w}}(\mathbf{w})$. In maximum likelihood (ML) estimation, we assume this *a priori* information is unknown. Assuming for the moment that the *a priori* probability distribution is Gaussian, with mean $\boldsymbol{\mu}_{\mathbf{w}}$ and covariance $\mathbf{P}_{\mathbf{w}}$, we have

$$p_{\mathbf{w}}(\mathbf{w}) = \frac{1}{(2\pi)^{n/2}|\mathbf{P}_{\mathbf{w}}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{w}-\boldsymbol{\mu}_{\mathbf{w}})^T\mathbf{P}_{\mathbf{w}}^{-1}(\mathbf{w}-\boldsymbol{\mu}_{\mathbf{w}})\right] \tag{4.28}$$

or yet

$$\ln p_{\mathbf{w}}(\mathbf{w}) = -\ln[(2\pi)^{n/2}|\mathbf{P}_{\mathbf{w}}|^{1/2}] - \frac{1}{2}\left[(\mathbf{w}-\boldsymbol{\mu}_{\mathbf{w}})^T\mathbf{P}_{\mathbf{w}}^{-1}(\mathbf{w}-\boldsymbol{\mu}_{\mathbf{w}})\right]\,. \tag{4.29}$$

Hence,

$$\frac{\partial \ln p_{\mathbf{w}}(\mathbf{w})}{\partial \mathbf{w}} = -\mathbf{P}_{\mathbf{w}}^{-1}(\mathbf{w}-\boldsymbol{\mu}_{\mathbf{w}}) \tag{4.30}$$

which indicates that lack of information about the random variable $\mathbf{w}$ implies infinite variance, $\mathbf{P}_{\mathbf{w}} \to \infty$, or yet $\mathbf{P}_{\mathbf{w}}^{-1} \to \mathbf{0}$. Thus, without a priori knowledge on $\mathbf{w}$ we have

$$\frac{\partial \ln p_{\mathbf{w}}(\mathbf{w})}{\partial \mathbf{w}} = \mathbf{0}\,. \tag{4.31}$$

This is also assumed to be the case even when the probability distribution of $\mathbf{w}$ is not Gaussian.

From (4.24) and (4.25), the maximum likelihood estimate of $\mathbf{w}$ can be obtained by

$$\mathbf{0} = \left[\frac{\partial \ln p_{\mathbf{z}|\mathbf{w}}(\mathbf{z}|\mathbf{w})}{\partial \mathbf{w}} + \frac{\partial \ln p_{\mathbf{w}}(\mathbf{w})}{\partial \mathbf{w}}\right]\Bigg|_{\mathbf{w}=\hat{\mathbf{w}}_{\mathrm{MAP}}} = \left.\frac{\partial \ln p_{\mathbf{z}|\mathbf{w}}(\mathbf{z}|\mathbf{w})}{\partial \mathbf{w}}\right|_{\mathbf{w}=\hat{\mathbf{w}}_{\mathrm{ML}}}, \tag{4.32}$$

or equivalently,

$$\left.\frac{\partial p_{\mathbf{z}|\mathbf{w}}(\mathbf{z}|\mathbf{w})}{\partial \mathbf{w}}\right|_{\mathbf{w}=\hat{\mathbf{w}}_{\mathrm{ML}}} = \mathbf{0}\,. \tag{4.33}$$

The estimate $\hat{\mathbf{w}}_{\mathrm{ML}}$ is sometimes referred to as the most likely estimate. However, because of the assumptions used in obtaining (4.33), this estimate is only reliable under certain conditions (see Jazwinski [84], p. 157). Just as in the case of the MAP estimate, the ML estimate is also a mode estimation, in analogy to (1.29). When we choose to refer to mode estimation, we should always make explicit which conditional probability is being maximized to avoid confusion, this defines whether we are performing MAP or ML estimation. As in MAP estimation, the estimate from ML is not guaranteed to be unbiased.

4.2 Example: Estimation of a Constant Vector

In this section we exemplify the problem of estimation by treating the case of estimating a constant (time independent) vector $\mathbf{w}$ by means of an observational process corrupted by noise, represented by the vector $\mathbf{v}$. We assume that $\mathbf{w}$ and $\mathbf{v}$ are independent and Gaussian distributed: $\mathbf{w} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{P})$, and $\mathbf{v} \sim \mathcal{N}(\mathbf{0}, \mathbf{R})$. Moreover, the observational process is taken to be a linear transformation

$$\mathbf{z} = \mathbf{H}\mathbf{w} + \mathbf{v} \tag{4.34}$$

where $\mathbf{w}$ is an n–vector, $\mathbf{z}$ and $\mathbf{v}$ are m–vectors, and $\mathbf{H}$ is an $m \times n$ matrix, referred to as the observation matrix which accounts, for example, for linear combinations among elements of the vector $\mathbf{w}$. To obtain an estimate based on the methods described in the previous section, we investigate the probability densities of the random variables involved in the observational process.

For the minimum variance estimate we need to determined the *a posteriori* probability density $p_{\mathbf{w}|\mathbf{z}}(\mathbf{w}|\mathbf{z})$, so that we can solve the integral in (4.14). From Bayes rule we have

$$p_{\mathbf{w}|\mathbf{z}}(\mathbf{w}|\mathbf{z}) = \frac{p_{\mathbf{z}|\mathbf{w}}(\mathbf{z}|\mathbf{w})p_{\mathbf{w}}(\mathbf{w})}{p_{\mathbf{z}}(\mathbf{z})} \tag{4.35}$$

and consequently we need to determine each one of the probability densities in this expression.

Since $\mathbf{w}$ is Gaussian, we can readily write

$$p_{\mathbf{w}}(\mathbf{w}) = \frac{1}{(2\pi)^{n/2}|\mathbf{P}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{w}-\boldsymbol{\mu})^T\mathbf{P}^{-1}(\mathbf{w}-\boldsymbol{\mu})\right]. \tag{4.36}$$

Linear transformations of Gaussian distributed variables result in Gaussian distributed variables (e.g., Sage & Melsa [121], pp. 71–72; see also Exercise 4, here). Therefore, the

probability distribution for the observations is given by

$$p_{\mathbf{z}}(\mathbf{z}) \;=\; \frac{1}{(2\pi)^{m/2}|\mathbf{P}_{\mathbf{z}}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{z}-\boldsymbol{\mu}_{\mathbf{z}})^T\mathbf{P}_{\mathbf{z}}^{-1}(\mathbf{z}-\boldsymbol{\mu}_{\mathbf{z}})\right] \tag{4.37}$$

where $\boldsymbol{\mu}_{\mathbf{z}}$ and $\mathbf{P}_{\mathbf{z}}$ correspond to the mean and covariance of the random variable $\mathbf{z}$, respectively. These quantities can be determined by applying the ensemble average operator to (4.34), and using the definition of covariance. Thus,

$$\boldsymbol{\mu}_{\mathbf{z}} \;=\; \mathcal{E}\{\mathbf{H}\mathbf{w}\} + \mathcal{E}\{\mathbf{v}\} \;=\; \mathbf{H}\boldsymbol{\mu} \tag{4.38}$$

and also,

$$\begin{aligned}
\mathbf{P}_{\mathbf{z}} &= \mathcal{E}\{(\mathbf{z}-\boldsymbol{\mu}_{\mathbf{z}})(\mathbf{z}-\boldsymbol{\mu}_{\mathbf{z}})^T\} \\
&= \mathcal{E}\{\,[(\mathbf{H}\mathbf{w}+\mathbf{v})-\mathbf{H}\boldsymbol{\mu})]\,[(\mathbf{H}\mathbf{w}+\mathbf{v})-\mathbf{H}\boldsymbol{\mu})]^T\} \\
&= \mathcal{E}\{\,[(\mathbf{H}\mathbf{w}-\mathbf{H}\boldsymbol{\mu})-\mathbf{v}]\,[(\mathbf{H}\mathbf{w}-\mathbf{H}\boldsymbol{\mu})-\mathbf{v}]^T\} \\
&= \mathbf{H}\mathcal{E}\{(\mathbf{w}-\boldsymbol{\mu})(\mathbf{w}-\boldsymbol{\mu})^T\}\mathbf{H}^T + \mathcal{E}\{\mathbf{v}\mathbf{v}^T\} \\
&\quad + \mathbf{H}\mathcal{E}\{(\mathbf{w}-\boldsymbol{\mu})\mathbf{v}^T\} + \mathcal{E}\{\mathbf{v}(\mathbf{w}-\boldsymbol{\mu})^T\}\mathbf{H}^T\,.
\end{aligned} \tag{4.39}$$

Noticing that $\mathbf{w}$ and $\mathbf{v}$ are independent $\mathcal{E}\{\mathbf{w}\mathbf{v}^T\} = \mathbf{0}$, and that $\mathbf{v}$ has zero mean, it follows that

$$\mathbf{P}_{\mathbf{z}} \;=\; \mathbf{H}\mathbf{P}\mathbf{H}^T + \mathbf{R} \tag{4.40}$$

and consequently, the probability distribution of $\mathbf{z}$ becomes

$$\begin{aligned}
p_{\mathbf{z}}(\mathbf{z}) \;&=\; \frac{1}{(2\pi)^{m/2}|(\mathbf{H}\mathbf{P}\mathbf{H}^T + \mathbf{R})|^{1/2}} \\
&\quad \times \exp\left[-\frac{1}{2}(\mathbf{z}-\mathbf{H}\boldsymbol{\mu})^T(\mathbf{H}\mathbf{P}\mathbf{H}^T + \mathbf{R})^{-1}(\mathbf{z}-\mathbf{H}\boldsymbol{\mu})\right]\,.
\end{aligned} \tag{4.41}$$

It remains for us to determine the conditional probability density $p_{\mathbf{z}|\mathbf{w}}(\mathbf{z}|\mathbf{w})$ explicitly. This distribution is also Gaussian (e.g., Sage & Melsa [121] pp. 73–74), and can be written as

$$p_{\mathbf{z}|\mathbf{w}}(\mathbf{z}|\mathbf{w}) \;=\; \frac{1}{(2\pi)^{m/2}|\mathbf{P}_{\mathbf{z}|\mathbf{w}}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{z}-\boldsymbol{\mu}_{\mathbf{z}|\mathbf{w}})^T\mathbf{P}_{\mathbf{z}|\mathbf{w}}^{-1}(\mathbf{z}-\boldsymbol{\mu}_{\mathbf{z}|\mathbf{w}})\right] \tag{4.42}$$

Analogously to what we have just done to determine $p_{\mathbf{z}}(\mathbf{z})$, we have

$$\boldsymbol{\mu}_{\mathbf{z}|\mathbf{w}} \;=\; \mathcal{E}\{\mathbf{H}\mathbf{w}|\mathbf{w}\} + \mathcal{E}\{\mathbf{v}|\mathbf{w}\} \;=\; \mathbf{H}\mathbf{w} \tag{4.43}$$

and

$$\begin{aligned}
\mathbf{P}_{\mathbf{z}|\mathbf{w}} &= \mathcal{E}\{(\mathbf{z}-\boldsymbol{\mu}_{\mathbf{z}|\mathbf{w}})(\mathbf{z}-\boldsymbol{\mu}_{\mathbf{z}|\mathbf{w}})^T|\mathbf{w}\} \\
&= \mathcal{E}\{\,[(\mathbf{H}\mathbf{w}+\mathbf{v})-\mathbf{H}\mathbf{w})]\,[(\mathbf{H}\mathbf{w}+\mathbf{v})-\mathbf{H}\mathbf{w})]^T\,|\mathbf{w}\} \\
&= \mathcal{E}\{\mathbf{v}\mathbf{v}^T|\mathbf{w}\} \\
&= \mathcal{E}\{\mathbf{v}\mathbf{v}^T\} \\
&= \mathbf{R}\,.
\end{aligned} \tag{4.44}$$

Therefore,

$$p_{\mathbf{z}|\mathbf{w}}(\mathbf{z}|\mathbf{w}) \;=\; \frac{1}{(2\pi)^{m/2}|\mathbf{R}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{z}-\mathbf{H}\mathbf{w})^T\mathbf{R}^{-1}(\mathbf{z}-\mathbf{H}\mathbf{w})\right] \tag{4.45}$$

which is the conditional probability of $\mathbf{z}$ given $\mathbf{w}$.

Combining the results (4.36), (4.41), and (4.45) in Bayes rule (4.35) it follows that the *a posteriori* probability distribution we are interested in takes the form

$$p_{\mathbf{w}|\mathbf{z}}(\mathbf{w}|\mathbf{z}) \;=\; \frac{|\mathbf{H}\mathbf{P}\mathbf{H}^T + \mathbf{R}|^{1/2}}{(2\pi)^{n/2}|\mathbf{P}|^{1/2}|\mathbf{R}|^{1/2}} \exp[-\frac{1}{2}J] \tag{4.46}$$

where J is defined as,

$$\begin{aligned} J(\mathbf{w}) \;\equiv\; & (\mathbf{z}-\mathbf{H}\mathbf{w})^T\mathbf{R}^{-1}(\mathbf{z}-\mathbf{H}\mathbf{w}) + (\mathbf{w}-\boldsymbol{\mu})^T\mathbf{P}^{-1}(\mathbf{w}-\boldsymbol{\mu}) \\ & -(\mathbf{z}-\mathbf{H}\boldsymbol{\mu})^T(\mathbf{H}\mathbf{P}\mathbf{H}^T+\mathbf{R})^{-1}(\mathbf{z}-\mathbf{H}\boldsymbol{\mu}) \end{aligned} \tag{4.47}$$

This quantity J can also be written in the following more compact form:

$$J(\mathbf{w}) = (\mathbf{w}-\hat{\mathbf{w}})^T\mathbf{P}_{\tilde{\mathbf{w}}}^{-1}(\mathbf{w}-\hat{\mathbf{w}}) \tag{4.48}$$

where $\mathbf{P}_{\tilde{\mathbf{w}}}^{-1}$ is given by

$$\mathbf{P}_{\tilde{\mathbf{w}}}^{-1} \;=\; \mathbf{P}^{-1} + \mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}\,, \tag{4.49}$$

the vector $\hat{\mathbf{w}}$ is given by

$$\hat{\mathbf{w}} \;=\; \mathbf{P}_{\tilde{\mathbf{w}}}(\mathbf{H}^T\mathbf{R}^{-1}\mathbf{z} + \mathbf{P}^{-1}\boldsymbol{\mu}) \tag{4.50}$$

and the reason for using the subscript $\tilde{\mathbf{w}}$ for the matrix $\mathbf{P}_{\tilde{\mathbf{w}}}$, indicating a relationship with the estimation error, will soon become clear.

According to (4.14), the minimum variance estimate is given by the conditional mean of the *a posteriori* probability density, that is,

$$\hat{\mathbf{w}}_{\mathrm{MV}} \;=\; \int_{-\infty}^{\infty} \mathbf{w}\, p_{\mathbf{w}|\mathbf{z}}(\mathbf{w}|\mathbf{z})\, d\mathbf{w} \;=\; \hat{\mathbf{w}} \tag{4.51}$$

where the integration can be performed using the approach of moments calculation of the Gaussian distribution (e.g., Maybeck [101]; see also Exercise 3, here).

The maximum *a posteriori* probability estimate (4.24) is the one that maximizes $p_{\mathbf{w}|\mathbf{z}}(\mathbf{w}|\mathbf{z})$ in (4.46), and is easily identified to be

$$\hat{\mathbf{w}}_{\mathrm{MAP}} = \hat{\mathbf{w}}\,. \tag{4.52}$$

Thus we see that the minimum variance estimate coincides with the maximum *a posteriori* probability density estimate.

Let us now return to the reason for using the subscript $\tilde{\mathbf{w}}$ in $\mathbf{P}_{\tilde{\mathbf{w}}}$. For that, remember that we defined the estimation error $\tilde{\mathbf{w}}$ as the difference between the estimate and the actual value taken by the variable of interest, that is,

$$\tilde{\mathbf{w}} \;\equiv\; \hat{\mathbf{w}} \;-\; \mathbf{w}\,. \tag{4.53}$$

We want to show that $\mathbf{P}_{\tilde{\mathbf{w}}}$ is indeed the estimate *error covariance* matrix. To verify this, let us show first that $\boldsymbol{\mu}_{\tilde{\mathbf{w}}} = \mathbf{0}$, that is, that the ensemble mean error estimate is zero for

the minimum variance and MAP estimates. In other words, we want to show that these estimates are *unbiased*. Using (4.50) we have

$$\begin{aligned}\boldsymbol{\mu}_{\tilde{\mathbf{w}}} &= \mathcal{E}\{(\hat{\mathbf{w}} - \mathbf{w})\} \\ &= \mathbf{P}_{\tilde{\mathbf{w}}}(\mathbf{H}^T\mathbf{R}^{-1}\mathcal{E}\{\mathbf{z}\} + \mathbf{P}^{-1}\boldsymbol{\mu}) - \boldsymbol{\mu} \\ &= \mathbf{P}_{\tilde{\mathbf{w}}}(\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H} + \mathbf{P}^{-1})\boldsymbol{\mu} - \boldsymbol{\mu} \end{aligned} \tag{4.54}$$

where we replaced $\mathbf{z}$ from (4.34), and we recall that $\mathbf{v}$ has zero mean. Therefore, using the definition of $\mathbf{P}_{\tilde{\mathbf{w}}}$ in (4.49), it follows that $\boldsymbol{\mu}_{\tilde{\mathbf{w}}} = \mathbf{0}$. Given what we know from (4.15), this result comes as no surprise in the case of the minimum variance estimate (4.51); in case of the MAP estimate this proves that (4.52) does provide an unbiased estimate.

To show that $\mathbf{P}_{\tilde{\mathbf{w}}}$ is the error covariance matrix of the estimate, we observe that $\tilde{\mathbf{w}}$ can be decomposed as

$$\begin{aligned}\mathbf{w} - \hat{\mathbf{w}} &= \mathbf{w} - \mathbf{P}_{\tilde{\mathbf{w}}}\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}\mathbf{w} - \mathbf{P}_{\tilde{\mathbf{w}}}\mathbf{H}^T\mathbf{R}^{-1}\mathbf{v} - \mathbf{P}_{\tilde{\mathbf{w}}}\mathbf{P}^{-1}\boldsymbol{\mu} \\ &= \mathbf{w} - \mathbf{P}_{\tilde{\mathbf{w}}}(\mathbf{P}_{\tilde{\mathbf{w}}}^{-1} - \mathbf{P}^{-1})\mathbf{w} - \mathbf{P}_{\tilde{\mathbf{w}}}\mathbf{H}^T\mathbf{R}^{-1}\mathbf{v} - \mathbf{P}_{\tilde{\mathbf{w}}}\mathbf{P}^{-1}\boldsymbol{\mu} \\ &= \mathbf{P}_{\tilde{\mathbf{w}}}\mathbf{P}^{-1}(\mathbf{w} - \boldsymbol{\mu}) - \mathbf{P}_{\tilde{\mathbf{w}}}\mathbf{H}^T\mathbf{R}^{-1}\mathbf{v}\,. \end{aligned} \tag{4.55}$$

Therefore,

$$\begin{aligned} var\{\tilde{\mathbf{w}}\} = cov\{\tilde{\mathbf{w}}, \tilde{\mathbf{w}}\} &= \mathcal{E}\{(\hat{\mathbf{w}} - \mathbf{w})(\hat{\mathbf{w}} - \mathbf{w})^T\} \\ &= \mathbf{P}_{\tilde{\mathbf{w}}}\mathbf{P}^{-1}\mathcal{E}\{(\mathbf{w} - \boldsymbol{\mu})(\mathbf{w} - \boldsymbol{\mu})^T\}\mathbf{P}^{-1}\mathbf{P}_{\tilde{\mathbf{w}}} \\ &\quad + \mathbf{P}_{\tilde{\mathbf{w}}}\mathbf{H}^T\mathbf{R}^{-1}\mathcal{E}\{\mathbf{v}\mathbf{v}^T\}\mathbf{R}^{-1}\mathbf{H}\mathbf{P}_{\tilde{\mathbf{w}}} \end{aligned} \tag{4.56}$$

where the cross–terms give no contribution since $\mathbf{w}$ and $\mathbf{v}$ are independent, and because $\mathbf{v}$ has zero mean. Using the definition of $\mathbf{P}$ it follows that

$$\begin{aligned} var\{\tilde{\mathbf{w}}\} &= \mathbf{P}_{\tilde{\mathbf{w}}}\mathbf{P}^{-1}\mathbf{P}_{\tilde{\mathbf{w}}} + \mathbf{P}_{\tilde{\mathbf{w}}}\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}\mathbf{P}_{\tilde{\mathbf{w}}} \\ &= \mathbf{P}_{\tilde{\mathbf{w}}}(\mathbf{P}^{-1} + \mathbf{H}^T\mathbf{R}^{-1}\mathbf{H})\mathbf{P}_{\tilde{\mathbf{w}}} \\ &= \mathbf{P}_{\tilde{\mathbf{w}}} \end{aligned} \tag{4.57}$$

where (4.49) was used. This shows that $\mathbf{P}_{\tilde{\mathbf{w}}}$ defined in (4.49) is indeed the estimation error covariance matrix, thus justifying its subscript $\tilde{\mathbf{w}}$. Moreover, it is simple to see that

$$|\mathbf{P}_{\tilde{\mathbf{w}}}| = |\mathbf{H}\mathbf{P}^{-1}\mathbf{H}^T + \mathbf{R}||\mathbf{P}||\mathbf{R}| \tag{4.58}$$

and therefore (4.46) can be written as

$$p_{\mathbf{w}|\mathbf{z}}(\mathbf{w}|\mathbf{z}) = \frac{1}{(2\pi)^{n/2}|\mathbf{P}_{\tilde{\mathbf{w}}}|^{1/2}} \exp[-\frac{1}{2}(\mathbf{w} - \hat{\mathbf{w}})^T\mathbf{P}_{\tilde{\mathbf{w}}}^{-1}(\mathbf{w} - \hat{\mathbf{w}})] \tag{4.59}$$

justifying the rewriting of J from (4.47) to (4.48).

It is now left for us to determine the maximum likelihood estimate (4.33). This can be done by maximizing the probability density $p_{\mathbf{z}|\mathbf{w}}(\mathbf{z}|\mathbf{w})$ in (4.45). Hence,

$$\begin{aligned}\mathbf{0} &= \left.\frac{\partial p_{\mathbf{z}|\mathbf{w}}(\mathbf{z}|\mathbf{w})}{\partial \mathbf{w}}\right|_{\mathbf{w}=\hat{\mathbf{w}}_{\mathrm{ML}}} \\ &= \mathbf{H}^T\mathbf{R}^{-1}(\mathbf{z} - \mathbf{H}\hat{\mathbf{w}}_{\mathrm{ML}}) \end{aligned} \tag{4.60}$$

that is,

$$\hat{\mathbf{w}}_{\mathrm{ML}} = (\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{R}^{-1}\mathbf{z} \tag{4.61}$$

which is, in principle, distinct from the estimates obtained above, following the minimum variance and maximum *a posteriori* probability estimation approaches. Remembering now that in maximum likelihood estimation we assume lack of statistical information regarding the process $\mathbf{w}$, and observing that this means $\mathbf{P}^{-1} = \mathbf{0}$, we see from (4.50) and (4.49) that, in this case,

$$\hat{\mathbf{w}}_{\mathrm{MV}}|_{\mathbf{P}^{-1}=\mathbf{0}} = \hat{\mathbf{w}}_{\mathrm{MAP}}|_{\mathbf{P}^{-1}=\mathbf{0}} = (\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{R}^{-1}\mathbf{z} = \hat{\mathbf{w}}_{\mathrm{ML}} \tag{4.62}$$

and therefore all three estimation approaches produce the same result.

Applying the average operator to (4.61) we have

$$\begin{aligned}
\mathcal{E}\{\hat{\mathbf{w}}_{\mathrm{ML}}\} &= (\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{R}^{-1}\mathcal{E}\{\mathbf{z}\} \\
&= (\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{R}^{-1}(\mathbf{H}\mathcal{E}\{\mathbf{w}\} + \mathcal{E}\{\mathbf{v}\}) \\
&= (\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H}\mathcal{E}\{\mathbf{w}\} \\
&= \mathcal{E}\{\mathbf{w}\}
\end{aligned} \tag{4.63}$$

where we used the fact that $\mathbf{v}$ has mean zero. This shows that the ML estimate is also unbiased.

It is simple to show that the maximum likelihood estimate error covariance is given by

$$var\{\tilde{\mathbf{w}}_{\mathrm{ML}}\} = (\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H})^{-1} \tag{4.64}$$

which is always greater than the error covariance obtained with the minimum variance estimation approach. This makes sense since the minimum variance estimate is that corresponding to the *minimum* of the Bayes risk.

Notice that all estimates above result in a *linear combination* of the observations. Moreover, although in this example all three estimation procedures studied above provide the same estimate this is not always the case. An example in which these estimates do not coincide is given in Exercise 2.

Another remark can be made by noticing that in the maximum *a posteriori* probability estimation context the denominator in (4.35) is not relevant for the maximization of the *a posteriori* probability distribution, as indicated in equations (4.26) and (4.27). This implies that we can derive the result for in (4.52) by minimizing the part of the functional J in (4.47) corresponding only to the probability density functions in the numerator of (4.35). That is, we can define the functional corresponding to these probability densities as

$$J_{\mathrm{MAP}}(\mathbf{w}) \equiv (\mathbf{z} - \mathbf{H}\mathbf{w})^T\mathbf{R}^{-1}(\mathbf{z} - \mathbf{H}\mathbf{w}) + (\mathbf{w} - \boldsymbol{\mu})^T\mathbf{P}^{-1}(\mathbf{w} - \boldsymbol{\mu}) \tag{4.65}$$

and its minimization can be shown to produces the same result as in (4.50) with error variance as in (4.49) — see Exercise 3. Analogously, we can define a cost function related to the *a priori* probability distribution associated with the maximum likelihood estimate, that is,

$$J_{\mathrm{ML}}(\mathbf{w}) \equiv (\mathbf{z} - \mathbf{H}\mathbf{w})^T\mathbf{R}^{-1}(\mathbf{z} - \mathbf{H}\mathbf{w})\,. \tag{4.66}$$

The minimization of J_{ML} gives the estimate in (4.61) with error variance (4.64).

4.3 Least Squares Estimation

All of the estimation methods seen so far, i.e., minimum variance, maximum *a posteriori* probability, and maximum likelihood, require statistical knowledge of part or all the random variables in question. However, when going from minimum variance and MAP to ML we relaxed the statistical assumptions by considering we knew nothing about the statistics of the variable(s) of interest ($\mathbf{w}$, in that case). Relaxing even further the statistical assumptions for the estimation problem takes us in to the situation where we have no statistical information about any of the variables involved in problem. In this extreme case, estimation reduces to the method of finding the least squares fit among the observations.

Let us consider again, as an example, the observational process in the previous section for an n–vector constant $\mathbf{w}$. Let us assume further that several observations are taken about the variable of interest, and that the i–th observation can be written as

$$\mathbf{z}_i = \mathbf{H}_i \mathbf{w} + \mathbf{v}_i \tag{4.67}$$

where $\mathbf{z}_i$, $\mathbf{H}_i$ and $\mathbf{v}_i$ represent an m_i–observation vector, a linear transformation matrix $m_i \times n$ and a m_i–noise vector, respectively. It is important to recognize now that we are assuming we do not know the statistics of the noise $\mathbf{v}_i$, and also that due to lack of statistical information we are not interpreting $\mathbf{w}$ as a random vector.

By collecting the result of k experiments in a long vector, we can write the expression above in the following compact form:

$$\tilde{\mathbf{z}}_k = \tilde{\mathbf{H}}_k \mathbf{w} + \tilde{\mathbf{v}}_k \tag{4.68}$$

where the $\tilde{m}_k$–vector $\tilde{\mathbf{z}}_k$ is defined as:

$$\tilde{\mathbf{z}}_k \equiv [\mathbf{z}_1^T \, \mathbf{z}_2^T \, \cdots \, \mathbf{z}_k^T]^T \tag{4.69}$$

for $\tilde{m}_k = \sum_{i=1}^{k} m_i$, and where

$$\tilde{\mathbf{v}}_k \equiv [\mathbf{v}_1^T \, \mathbf{v}_2^T \, \cdots \, \mathbf{v}_k^T]^T \tag{4.70}$$

and the matrix $\tilde{\mathbf{H}}_k$, of dimension $\tilde{m}_k \times n$, is defined as

$$\tilde{\mathbf{H}}_k \equiv [\mathbf{H}_1^T \, \mathbf{H}_2^T \, \cdots \, \mathbf{H}_k^T]^T \, . \tag{4.71}$$

The problem we want to consider is that of finding an estimate $\hat{\mathbf{w}}_k$ which minimizes the quadratic function $\mathcal{J}$,

$$\mathcal{J}(\hat{\mathbf{w}}_k) = \frac{1}{2}(\tilde{\mathbf{z}}_k - \tilde{\mathbf{H}}_k \hat{\mathbf{w}}_k)^T \tilde{\mathbf{O}}_k^{-1} (\tilde{\mathbf{z}}_k - \tilde{\mathbf{H}}_k \hat{\mathbf{w}}_k) \tag{4.72}$$

which measures the distance between the observations and the estimate. The value that minimizes this function is called the least squares estimate and is denoted by $\hat{\mathbf{w}}_k^{\mathrm{LS}}$. The positive definite and symmetric matrix $\tilde{\mathbf{O}}_k^{-1}$ represents weights attributed to each experiment, and convey a certain degree of confidence regarding the experiment in question.

The estimator function $\mathcal{J}$ is deterministic, therefore the problem of minimizing $\mathcal{J}$ is a common optimization problem, where the solution $\hat{\mathbf{w}}_k^{\mathrm{LS}}$ can be determined by means solving,

$$\left. \frac{\partial \mathcal{J}}{\partial \hat{\mathbf{w}}_k} \right|_{\hat{\mathbf{w}}_k = \hat{\mathbf{w}}_k^{\mathrm{LS}}} = \mathbf{0} \, . \tag{4.73}$$

Then, the differentiation of (4.72) yields

$$\tilde{\mathbf{H}}_k^T \tilde{\mathbf{O}}_k^{-1} (\mathbf{z}_k - \tilde{\mathbf{H}}_k \hat{\mathbf{w}}_k^{\mathrm{LS}}) = \mathbf{0} \tag{4.74}$$

from where it follows that

$$\hat{\mathbf{w}}_k^{\mathrm{LS}} = \mathbf{P}_k \tilde{\mathbf{H}}_k^T \tilde{\mathbf{O}}_k^{-1} \mathbf{z}_k \,, \tag{4.75}$$

which is the estimate for the value of $\mathbf{w}$. For convenience we define a matrix $\mathbf{P}_k$ of dimension $n \times n$ as

$$\mathbf{P}_k \equiv (\tilde{\mathbf{H}}_k^T \tilde{\mathbf{O}}_k^{-1} \tilde{\mathbf{H}}_k)^{-1} \,, \tag{4.76}$$

and assume that the inverse exists. The matrix $\mathbf{P}_k^{-1}$ is sometimes referred to as the Gram matrix. A comparison with the estimate provided by the ML (4.61) method shows certain resemblance, however, since $\mathbf{R}$ and $\mathbf{O}_k$ are not related in any way, this resemblance is purely formal.

Suppose now that an additional experiment was made and it produced a new observation $\mathbf{z}_{k+1}$:

$$\mathbf{z}_{k+1} = \mathbf{H}_{k+1}\mathbf{w} + \mathbf{v}_{k+1}. \tag{4.77}$$

Then, by means of the notation introduced above, we can write

$$\tilde{\mathbf{z}}_{k+1} = \tilde{\mathbf{H}}_{k+1}\mathbf{w} + \tilde{\mathbf{v}}_{k+1} \,, \tag{4.78}$$

where

$$\tilde{\mathbf{z}}_{k+1} = [\tilde{\mathbf{z}}_k^T \; \mathbf{z}_{k+1}^T]^T \,, \quad \tilde{\mathbf{H}}_{k+1} = [\tilde{\mathbf{H}}_k^T \; \mathbf{H}_{k+1}^T]^T \,, \quad \tilde{\mathbf{v}}_{k+1} = [\tilde{\mathbf{v}}_k^T \; \mathbf{v}_{k+1}^T]^T \,. \tag{4.79}$$

Direct use of the minimization procedure just described leads to an estimate including the new observation $\mathbf{z}_{k+1}$, and given by

$$\hat{\mathbf{w}}_{k+1}^{\mathrm{LS}} = \mathbf{P}_{k+1} \tilde{\mathbf{H}}_{k+1}^T \tilde{\mathbf{O}}_{k+1}^{-1} \tilde{\mathbf{z}}_{k+1} \,, \tag{4.80}$$

where $\mathbf{P}_{k+1}$ is defined, in analogy to $\mathbf{P}_k$, as

$$\mathbf{P}_{k+1} \equiv (\tilde{\mathbf{H}}_{k+1}^T \tilde{\mathbf{O}}_{k+1}^{-1} \tilde{\mathbf{H}}_{k+1})^{-1} \,, \tag{4.81}$$

and $\tilde{\mathbf{O}}_{k+1}^{-1}$ is a new weight matrix that takes into account the observation $\mathbf{z}_{k+1}$.

The processing of an extra observation forces us to have to solve the minimization problem completely again. In particular, we have to calculate the inverse of an $n \times n$ matrix for each new observation made. This computational burden can be avoided if we assume that the matrix $\tilde{\mathbf{O}}_{k+1}^{-1}$ can be partitioned in the following manner:

$$\tilde{\mathbf{O}}_{k+1}^{-1} = \begin{bmatrix} \tilde{\mathbf{O}}_k^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{O}_{k+1}^{-1} \end{bmatrix} . \tag{4.82}$$

that is, $\tilde{\mathbf{O}}_{k+1}^{-1}$ is assumed to be a block–diagonal matrix.

With this assumption, we can write the product of the matrices in $\mathbf{P}_{k+1}$ as

$$\begin{aligned} \tilde{\mathbf{H}}_{k+1}^T \tilde{\mathbf{O}}_{k+1}^{-1} \tilde{\mathbf{H}}_{k+1} &= [\tilde{\mathbf{H}}_k^T \; \mathbf{H}_{k+1}^T] \begin{bmatrix} \tilde{\mathbf{O}}_k^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{O}_{k+1}^{-1} \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{H}}_k \\ \mathbf{H}_{k+1} \end{bmatrix} \\ &= \tilde{\mathbf{H}}_k^T \tilde{\mathbf{O}}_k^{-1} \tilde{\mathbf{H}}_k + \mathbf{H}_{k+1}^T \mathbf{O}_{k+1}^{-1} \mathbf{H}_{k+1} \,. \end{aligned} \tag{4.83}$$

Furthermore, using the definitions of the matrices $\mathbf{P}$ given above, we have that

$$\mathbf{P}_{k+1}^{-1} = \mathbf{P}_k^{-1} + \mathbf{H}_{k+1}^T \mathbf{O}_{k+1}^{-1} \mathbf{H}_{k+1} \tag{4.84}$$

or yet, using the Sherman–Morrison–Woodbury formula (e.g., Golub & Van Loan [67], p. 51).

$$\begin{aligned} \mathbf{P}_{k+1} &= (\mathbf{P}_k^{-1} + \mathbf{H}_{k+1}^T \mathbf{O}_{k+1}^{-1} \mathbf{H}_{k+1})^{-1} \\ &= \mathbf{P}_k - \mathbf{P}_k \mathbf{H}_{k+1}^T (\mathbf{H}_{k+1} \mathbf{P}_k \mathbf{H}_{k+1}^T + \mathbf{O}_{k+1})^{-1} \mathbf{H}_{k+1} \mathbf{P}_k \,. \end{aligned} \tag{4.85}$$

Defining a matrix $\mathbf{G}_{k+1}$ as

$$\mathbf{G}_{k+1} \equiv \mathbf{P}_k \mathbf{H}_{k+1}^T (\mathbf{H}_{k+1} \mathbf{P}_k \mathbf{H}_{k+1}^T + \mathbf{O}_{k+1})^{-1} \,, \tag{4.86}$$

we can compactly write

$$\mathbf{P}_{k+1} = (\mathbf{I} - \mathbf{G}_{k+1} \mathbf{H}_{k+1}) \mathbf{P}_k \,. \tag{4.87}$$

Therefore the estimate $\hat{\mathbf{w}}_{k+1}^{\mathrm{LS}}$, which includes the new observation can be re–written as

$$\hat{\mathbf{w}}_{k+1}^{\mathrm{LS}} = (\mathbf{I} - \mathbf{G}_{k+1} \mathbf{H}_{k+1}) \mathbf{P}_k \, \tilde{\mathbf{H}}_{k+1}^T \tilde{\mathbf{O}}_{k+1}^{-1} \tilde{\mathbf{z}}_{k+1} \,. \tag{4.88}$$

Using the matrix partition for $\tilde{\mathbf{O}}_{k+1}^{-1}$, introduced above, we can decompose the expression for the estimate in two terms,

$$\tilde{\mathbf{H}}_{k+1}^T \tilde{\mathbf{O}}_{k+1}^{-1} \tilde{\mathbf{z}}_{k+1} = \tilde{\mathbf{H}}_k^T \tilde{\mathbf{O}}_k^{-1} \tilde{\mathbf{z}}_k + \mathbf{H}_{k+1}^T \mathbf{O}_{k+1}^{-1} \mathbf{z}_{k+1} \,, \tag{4.89}$$

and consequently (4.88) is transformed in

$$\begin{aligned} \hat{\mathbf{w}}_{k+1}^{\mathrm{LS}} &= [\mathbf{I} - \mathbf{G}_{k+1} \mathbf{H}_{k+1}] \mathbf{P}_k (\tilde{\mathbf{H}}_k^T \tilde{\mathbf{O}}_k^{-1} \tilde{\mathbf{z}}_k + \mathbf{H}_{k+1}^T \mathbf{O}_{k+1}^{-1} \mathbf{z}_{k+1}) \,, \\ &= [\mathbf{I} - \mathbf{G}_{k+1} \mathbf{H}_{k+1}] \hat{\mathbf{w}}_k^{\mathrm{LS}} + [\mathbf{I} - \mathbf{G}_{k+1} \mathbf{H}_{k+1}] \mathbf{P}_k \mathbf{H}_{k+1}^T \mathbf{O}_{k+1}^{-1} \mathbf{z}_{k+1} \end{aligned} \tag{4.90}$$

where we used (4.75) to obtain the second equality.

A even better expression for the estimate can be derived if we use the definition for the matrix $\mathbf{G}_{k+1}$. In this case, the coefficient of the last term in the previous expression can be re–written as

$$\begin{aligned} [\mathbf{I} - \mathbf{G}_{k+1} \mathbf{H}_{k+1}] \mathbf{P}_k \mathbf{H}_{k+1}^T \mathbf{O}_{k+1}^{-1} &= [\mathbf{I} - \mathbf{G}_{k+1} \mathbf{H}_{k+1}] \\ &\quad \times \mathbf{G}_{k+1} (\mathbf{H}_{k+1} \mathbf{P}_k \mathbf{H}_{k+1}^T + \mathbf{O}_{k+1}) \mathbf{O}_{k+1}^{-1} \\ &= [\mathbf{I} - \mathbf{G}_{k+1} \mathbf{H}_{k+1}] \mathbf{G}_{k+1} \\ &\quad \times (\mathbf{I} + \mathbf{H}_{k+1} \mathbf{P}_k \mathbf{H}_{k+1}^T \mathbf{O}_{k+1}^{-1}) \\ &= \mathbf{G}_{k+1} [\mathbf{I} + \mathbf{H}_{k+1} \mathbf{P}_k \mathbf{H}_{k+1}^T \mathbf{O}_{k+1}^{-1} \\ &\quad - \mathbf{H}_{k+1} \mathbf{G}_{k+1} (\mathbf{I} + \mathbf{H}_{k+1} \mathbf{P}_k \mathbf{H}_{k+1}^T \mathbf{O}_{k+1}^{-1})] \\ &= \mathbf{G}_{k+1} [\mathbf{I} + \mathbf{H}_{k+1} \mathbf{P}_k \mathbf{H}_{k+1}^T \mathbf{O}_{k+1}^{-1} \\ &\quad - \mathbf{H}_{k+1} \mathbf{P}_k \mathbf{H}_{k+1}^T \mathbf{O}_{k+1}^{-1}] \\ &= \mathbf{G}_{k+1} \,. \end{aligned} \tag{4.91}$$

Thus, the estimate can be placed finally in the form

$$\hat{\mathbf{w}}_{k+1}^{\text{LS}} = \hat{\mathbf{w}}_k^{\text{LS}} + \mathbf{G}_{k+1}(\mathbf{z}_{k+1} - \mathbf{H}_{k+1}\hat{\mathbf{w}}_k^{\text{LS}}) \,, \tag{4.92}$$

where $\mathbf{G}_{k+1}$ is given by (4.86). This expression provides a *recursive* manner of updating the estimate, given a new observation of the variable of interest and the estimate obtained before the new observation had been made. This recursive expression requires inverting an $m_{k+1} \times m_{k+1}$ matrix embedded in the definition of $\mathbf{G}_{k+1}$ in (4.86), rather than the $n \times n$ matrix (4.81), for each new observation becoming available. This represents an enormous computational savings especially for $n \gg m_k$, for all k.

4.4 Relationship between Least Squares and Minimum Variance

The estimates produced by the minimum variance and least squares methods are of fundamental importance in many studies in estimation theory. Consequently, in this section, we explore the relationship between these two estimates.

To simplify this notation let us omit the index k from the previous section, so that the observational process can be written just as in (4.34),

$$\mathbf{z} = \mathbf{H}\mathbf{w} + \mathbf{v} \,, \tag{4.93}$$

Moreover, the estimate of $\mathbf{w}$ provided by the least squares method is written as

$$\hat{\mathbf{w}}_{\text{LS}} = \mathbf{M}\mathbf{z} \,, \tag{4.94}$$

where for convenience we define the $n \times m$ matrix $\mathbf{M}$ as

$$\mathbf{M} = (\mathbf{H}^T\mathbf{O}^{-1}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{O}^{-1} \,. \tag{4.95}$$

Notice that $\mathbf{MH} = \mathbf{I}$ which, assuming the noise $\mathbf{v}$ has zero mean is a way of expressing the fact that the estimate $\hat{\mathbf{w}}_{\text{LS}}$ is unbiased. To see this, we define the error associated to the least squares estimate as

$$\tilde{\mathbf{w}}_{\text{LS}} \equiv \mathbf{w} - \hat{\mathbf{w}}_{\text{LS}} \,, \tag{4.96}$$

where once again we use a *tilde* to indicate an error vector. Application the ensemble average operator, and using (4.93) and (4.94), it follows that

$$\begin{aligned} \mathcal{E}\{\tilde{\mathbf{w}}_{\text{LS}}\} &= \mathcal{E}\{[\mathbf{w} - \mathbf{M}(\mathbf{H}\mathbf{w} + \mathbf{v})]\} \\ &= -\mathbf{M}\mathcal{E}\{\mathbf{v}\} \\ &= \mathbf{0} \,, \end{aligned} \tag{4.97}$$

which justifies the assertion above that the least squares estimate is unbiased.

The least squares estimate error variance can be calculated according to

$$\mathbf{P}_{\tilde{\mathbf{w}}_{\text{LS}}} = \mathcal{E}\{\tilde{\mathbf{w}}_{\text{LS}}\tilde{\mathbf{w}}_{\text{LS}}^T\} = \mathbf{M}\mathcal{E}\{\mathbf{v}\mathbf{v}^T\}\mathbf{M}^T = \mathbf{M}\mathbf{R}\mathbf{M}^T \tag{4.98}$$

where $\mathbf{R}$ is the (co)variance matrix of the noise $\mathbf{v}$, as defined in Section 4.3. Substituting the value of $\mathbf{M}$ as defined above we have

$$\mathbf{P}_{\tilde{\mathbf{w}}_{\mathrm{LS}}} = (\mathbf{H}^T\mathbf{O}^{-1}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{O}^{-1}\mathbf{R}\mathbf{O}^{-1}\mathbf{H}(\mathbf{H}^T\mathbf{O}^{-1}\mathbf{H})^{-1}\,. \tag{4.99}$$

Now remember that, by the procedure of Section 4.3, the linear estimate of minimum variance, with zero mean $\boldsymbol{\mu}_{\mathbf{w}} = \mathbf{0}$ and for which $\mathbf{P}_{\mathbf{w}}^{-1} = \mathbf{0}$, is given by

$$\tilde{\mathbf{w}}_{\mathrm{MV}} = (\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{R}^{-1}\mathbf{z}\,. \tag{4.100}$$

which is the same as that obtained when using the approach of maximum likelihood estimation. As we know, this estimate is also unbiased, and with associated error (co)variance

$$\mathbf{P}_{\tilde{\mathbf{w}}_{\mathrm{MV}}} = (\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H})^{-1}\,, \tag{4.101}$$

as it can be seen in (4.50) and (4.51), and also (4.61) and (4.64), respectively. Therefore, we notice by comparison that the estimate obtained by the least squares method is the same as the one obtained by linear minimum variance when the matrix of weight $\mathbf{O}$ used by the first method is substituted by the noise (co)variance matrix, that is, $\mathbf{O} = \mathbf{R}$.

In general, the weight matrix used in the least squares method is a general positive definite and symmetric matrix, without any statistical meaning; since the estimate provided by the minimum variance approach is that with *minimum variance*, for the linear case, it follows that in general

$$\mathbf{P}_{\tilde{\mathbf{w}}_{\mathrm{LS}}} \geq \mathbf{P}_{\tilde{\mathbf{w}}_{\mathrm{MV}}}\,, \tag{4.102}$$

where the equality holds when $\mathbf{O} = \mathbf{R}$. This inequality is valid even if we do not use the fact that the estimate $\hat{\mathbf{w}}_{\mathrm{MV}}$ is that of minimum variance. To derive this inequality, we can use the following matrix inequality

$$\mathbf{A}^T\mathbf{A} \geq (\mathbf{B}^T\mathbf{A})^T(\mathbf{B}^T\mathbf{B})^{-1}(\mathbf{B}^T\mathbf{A})\,, \tag{4.103}$$

for $\mathbf{A}$ and $\mathbf{B}$, of dimensions $n \times m$, with $n \geq m$, and $\mathbf{B}$ of full rank. This derivation is left as an exercise.

Exercises

1. (Sage & Melsa [121], Problem 6.1) Another example of cost function, aside from those given in the main text, is that defined by the absolute value of the error: $J(\tilde{w}) = |\tilde{w}| = |w - \hat{w}|$, considering the scalar case. Show that in this case, the estimate $\hat{w}_{\mathrm{ABS}}$ that minimizes the Bayes risk is the one for which we have:

$$\int_{-\infty}^{\hat{w}_{\mathrm{ABS}}} p_{w|z}(w|z)\,dw \;=\; \int_{\hat{w}_{\mathrm{ABS}}}^{\infty} p_{w|z}(w|z)\,dw$$

and that consequently, the estimate with minimum absolute value can be determined by solving:

$$\int_{\hat{w}_{\mathrm{ABS}}}^{\infty} p_{w|z}(w|z)\,dw \;=\; \frac{1}{2}$$

for $\hat{w} = \hat{w}_{\text{ABS}}$. In other words, the estimate with minimum absolute value $\hat{w}_{\text{ABS}}$ is the median, as introduced in (1.27). Derive the corresponding modification of the result above for the vector case, if we define the cost function to be

$$J(\tilde{\mathbf{w}}) = \sum_i |\tilde{w}_i|$$

2. Consider the observational process of a binary variable (binary signal), subject to noise (measurement errors). This scalar observation process can be written as

$$z = w + v$$

where w and v are independent, and v is a gaussian noise, represented by $\mathcal{N}(0, \sigma_v^2)$. The signal w follows the binary distribution defined as

$$p_w(w) = 0.5\delta(w-1) + 0.5\delta(w+1)$$

where δ represents the Dirac delta. Then,

   (a) Determine the *a priori* probability density $p_{z|w}(z|w)$.
   (b) Show that the probability density $p_z(z)$ is given by[2]

$$p_z(z) = \frac{1}{2\sqrt{2\pi}\sigma_v}\left\{\exp\left[-\frac{(z-1)^2}{2\sigma_v^2}\right] + \exp\left[-\frac{(z+1)^2}{2\sigma_v^2}\right]\right\}$$

   (c) Show that the maximum *a posteriori* probability estimate is $\hat{w}_{\text{MAP}} = \text{sign}(z)$.
   (d) Show that the minimum variance error estimate is $\hat{w}_{\text{MV}} = \tanh\left(\frac{z}{\sigma_v^2}\right)$.

   In the minimum variance estimation case, what happens when the observations become more accurate?

3. Show that the solution of the minimization of J_{MAP} in (4.65) is given by (4.50) with error estimate (4.49).

4. Writing a few terms for the traces in the expressions below, verify that:

   (a) $\frac{d[\text{Tr}(\mathbf{AB})]}{d\mathbf{A}} = \mathbf{B}^T$, where $\mathbf{AB}$ is symmetric
   (b) $\frac{d[\text{Tr}(\mathbf{ACA}^T)]}{d\mathbf{A}} = 2\mathbf{AC}$, where $\mathbf{C}$ is also symmetric

   Notice that is x is a scalar, we define its derivative with respect to a matrix $\mathbf{A}$ according to:

$$\frac{dx}{d\mathbf{A}} \equiv \begin{pmatrix} \frac{dx}{da_{11}} & \frac{dx}{da_{12}} & \cdot & \cdot & \cdot \\ \frac{dx}{da_{21}} & \frac{dx}{da_{22}} & \cdot & \cdot & \cdot \\ \cdot & & \cdot & & \\ \cdot & & \cdot & \cdot & \\ \cdot & & \cdot & & \cdot \end{pmatrix}$$

   where a_{ij} is the (i,j)-th element of matrix $\mathbf{A}$.

[2] If a random variable z is defined as the summation of two independent random variables w and v the probability of z can be obtained via the convolution integral:

$$p_z(z) = \int_{-\infty}^{\infty} p_w(z-v)p_v(v)\,dv$$

5. Show that

$$\mathbf{G}_{k+1} = \mathbf{P}_{k+1}\mathbf{H}_{k+1}^T\mathbf{O}_{k+1}^{-1}$$

is an alternative expression for the gain matrix $\mathbf{G}_{k+1}$ found in the least squares estimation method.

6. Let $\mathbf{A}$ and $\mathbf{B}$ be to $n \times m$ matrices, with $n \geq m$, and with $\mathbf{B}$ full rank (m). Show that

$$\mathbf{A}^T\mathbf{A} \geq (\mathbf{B}^T\mathbf{A})^T (\mathbf{B}^T\mathbf{B})^{-1} (\mathbf{B}^T\mathbf{A}) .$$

(Hint: Use the following inequality:

$$(\mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{y})^T(\mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{y}) \geq 0$$

valid for any two m–vectors $\mathbf{x}$ e $\mathbf{y}$.) Now, to show the inequality in (4.102), without making use of the fact that $\hat{\mathbf{w}}_{MV}$ is a minimum variance estimate for the linear case, make the following choice:

$$\mathbf{A} = \mathbf{R}^{1/2}\mathbf{M}^T , \quad \mathbf{B} = \mathbf{R}^{-1/2}\mathbf{H}$$

and complete the proof as suggested in the end of section 4.5.