Key Ideas in Parameter Estimation

Pavel Loskot
ZJU-UIUC Institute
Haining, China
e-mail: *pavelloskot@intl.zju.edu.cn*

***Abstract*—Parameter estimation plays a crucial role in many applications of statistical signal processing. Estimation theory is a well-established and rigorous framework for making the statistical inferences from noisy observations. It yields the best possible (in a precisely defined sense), interpretable and numerically effective procedures, provided that the models of measurements and of signals are known up to unknown parameters. Understanding the fundamental principles of parameter estimation is nowadays also important in designing the interpretable machine learning architectures, which usually exchange the computational complexity for the superior performance, while alleviating the need for knowing the models of signals and measurements. This paper comprehensively outlines the key principles of estimating the time-invariant random and non-random parameters, which are accompanied by several illustrative examples. The presentation focuses on the key ideas, and does not cover many other relevant topics; for example, neither the estimation of time-varying signals nor the survey of the research literature are considered.**

Keywords—Inference; linear estimator; noise; optimum estimator; parameter estimation; risk function; uncertainty.

I. INTRODUCTION

The common task in many applications of statistical signal processing is to learn the values of hidden parameters, and to extract other useful information from noisy measurements. This must be done statistically, i.e., the good-quality parameter estimates must be obtained with a high probability, i.e., most of the time. As illustrated in Figure 1, the unknown values of parameters are mapped to measured signals, which are distorted by the measurement noise. The goal is to find the optimum mapping for the measurements in order to recover the parameter values of interest with as small error as possible, despite the presence of other nuisance parameters and the measurement noise. Since the mapping of parameters to measurements is assumed to be known, one might be tempted to simply undo the mapping by using the corresponding inverse mapping. This may be a simple strategy for estimating the parameter values, provided that the inverse mapping can be obtained. However, the caveat is that the inverse mapping usually amplifies the measurement noise, so it is only effective if the noise is sufficiently small, and thus, can be neglected. In practice, this is often not the case, and more sophisticated methods of the parameter estimation are required.

The best possible target mapping representing the estimator is usually formulated as the solution of a constrained or unconstrained optimization problem. The optimization problem is defined, so that the estimation error is minimized in some sense. This is also dependent on how the estimated values are used in a given signal processing application. However, the resulting optimization problem may not have any solution, for example, since some important knowledge is missing, or it may be too complex to be solved effectively. In such a case, the estimator can be constrained to be a linear filter. This greatly limits the implementation complexity, although the optimality can now be only considered within the class of linear filters. The solution of the optimization problem can be sometimes found analytically in a closed-form, otherwise it must be obtained numerically.

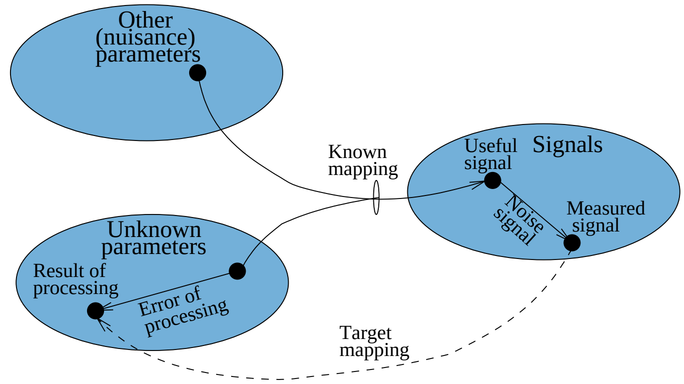

Figure 1. The mapping of parameters to measured signals, and the inverse mapping of measurements to the estimated parameter values.

The solution of the optimization problem answers one of the following two questions, depending on the application, i.e.:

1) "*Which from several defined values the parameter has?*"
2) "*How big is the parameter value?*"

The first question is central in detection theory, and it is also closely related to the optimum decision and the hypothesis testing problems. The second question is the subject of this tutorial, i.e., it defines the point estimates of the parameter within estimation theory. The parameter estimates can be also obtained as ranges of values, however, the interval estimates are not considered in this tutorial.

If the parameters vary in time, they are referred to as signals in engineering applications of statistical signal processing. In mathematics, the term, process, instead of signal is usually preferred, and the parameter estimation is studied as one of the tasks of statistical inference. In data science, the longitudinal data in discrete time are referred to as sequences or time-series, and the estimation problems are referred to as data mining.

In general, there are three levels of statistics that can be obtained for the measurements as indicated in Figure 2. In par-

 ISBN: 978-1-68558-142-8

ticular, the parametric and non-parametric descriptive statistics are used in characterizing and summarizing the measurements in observational studies. In order to estimate the values of hidden parameters, which cannot be observed directly, the inferential statistics require using more sophisticated methods as discussed above. Finally, the causal inferences are used to determine, for example, the counterfactual outcomes and other cause-effect relationships, which is, however, beyond the scope of this tutorial.

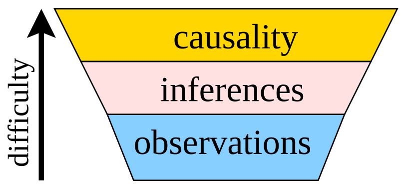

Figure 2. The three levels of statistically processing the measurements.

A random signal at any particular time instant is a random variable. The processing of continuous time signals can exploit derivatives, whereas differences are used for signals in discrete time (the time discretization does not automatically replace derivatives with differences). The important consideration in estimating the parameter values is whether their prior probability distribution is known; in such a case, the parameters can be considered to be random, and the Bayesian inference methods are used. Otherwise, without any prior knowledge, the unknown parameters must be treated as being non-random (i.e., deterministic). There are, however, many situations when some prior parameter statistics are known (e.g., the mean and the variance), or their probability distribution is known partially; these cases must be considered individually as they do not constitute the case of random nor non-random parameters.

There are generally four basic types of the parameter estimation problems depending whether the prior probability distribution of the parameter to be estimated is known or not, and whether the estimator is general (i.e., unconstrained), or it is a linear filter or a transformation.

The rest of this tutorial is organized as follows. The problem of finding the best possible estimator for a random parameter minimizing so-called risk function is outlined in Section II. It includes the Minimum Mean-Square Error (MMSE) and the Maximum a Posteriori (MAP) estimators. If the prior distribution of the parameter is not known, it must be treated as being non-random as explained in Section III. This case includes the Minimum Variance Unbiased (MVUB), the Maximum Likelihood (ML), the Least Squares (LS) and the moment-based estimators. For these estimators, the Cramer-Rao Lower Bound (CRLB) of their performance has been defined. Linear estimators of random and non-random parameters are considered in Section IV and Section V, respectively. Additional solved problems are provided in Section VI. The relevant textbooks and the topics and problems, which are not covered in this tutorial are summarized and discussed in Section VII. Finally, Section VIII concludes the paper.

II. General Estimation of Random Parameters

Consider first the case of a general estimation of a continuous or discrete random parameter, P, having the known prior Probability Density Function (PDF), $f_P(p)$, or the Probability Mass Function (PMF), $\Pr_P(p)$, respectively. The parameter P is observed as the value (or multiple values), $X(P)$, representing the system being considered, i.e., it is crucial that the dependence of X on P is known. It is then possible to derive the statistical dependence of X on P represented by the conditional density, $f_{X|P}(x|p)$, or the conditional probability, $\Pr_{X|P}(x|p)$, respectively. The estimator converts the measurements, $X(P)$, to the parameter estimates, $\hat{P}(X)$, which are used in a given application. The application dictates how to define the estimation errors, $\mu(\hat{P}(X),P)$. The overall process is summarized in Figure 3.

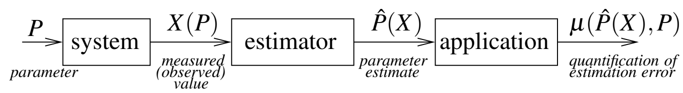

Figure 3. A formulation of the parameter estimation problem.

Since the parameter, P, is assumed to be random, the function, $\mu(\hat{P},P)$, quantifying the estimation error, $(\hat{P}-P)$, is a random variable. In order to minimize the estimation error for any value of P (which is unknown), the optimum estimator minimizes the mean value, or so-called the risk, $\mathrm{E}\left[\mu(\hat{P},P)\right]$, where $\mathrm{E}[\cdot]$ denotes expectation. For instance, the MMSE estimator minimizes the risk, $\mathrm{E}_{X,P}\left[\mu(\hat{P}(X),P)\right] = \mathrm{E}_{X,P}\left[(\hat{P}(X)-P)^2\right]$, whereas the probability that the error is greater than a threshold, Δ, assumes the risk function, $\mathrm{E}_{X,P}\left[\mu(\hat{P}(X),P)\right] = \Pr\left(|\hat{P}(X)-P|>\Delta\right)$, as illustrated in Figure 4.

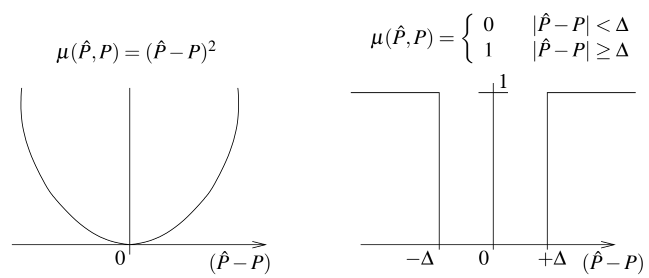

Figure 4. The two examples of risk function for designing estimator, $\hat{P}(X)$.

In particular, the optimum estimator of a random parameter, P, minimizes the general risk function,

$$\mathrm{E}_{X,P}\left[\mu(\hat{P}(X),P)\right] = \int_{\{X\}}\int_{\{P\}} \mu(\hat{P}(x),p)\, f_{X,P}(x,p)\,\mathrm{d}p\,\mathrm{d}x. \quad (1)$$

The risk (1) is minimized, if and only if,

$$\hat{P}_{\mathrm{opt}} = \mathrm{argmin}_{\hat{P}(x)}\, \mathrm{E}_{X,P}\left[\mu(\hat{P}(x),P)|X=x\right]. \quad (2)$$

Remark 1. *The expressions presented here assume that P and X are scalar random variables; the extension to random vectors is (usually) straightforward.*

A. The MMSE Estimator

Substituting the MMSE function, $\mu(\hat{P},P) = (\hat{P}-P)^2$, into (1) above, the MMSE estimator is defined as,

$$\begin{aligned}\hat{P}_{\mathrm{MMSE}}(x) &= \mathrm{E}_P[P|X=x] = \int_{\{P\}} p f_{P|X}(p|x)\,\mathrm{d}p \\ &= \frac{\int_{\{P\}} p f_{X,P}(x,p)\,\mathrm{d}p}{f_X(x)} = \frac{\int_{\{P\}} p f_{X|P}(x|p) f_P(p)\,\mathrm{d}p}{\int_{\{P\}} f_{X|P}(x|p) f_P(p)\,\mathrm{d}p}.\end{aligned} \tag{3}$$

Hence, if no measurements are available at all, the optimum MMSE estimator is, $\hat{P}_{\mathrm{MMSE}} = \mathrm{E}_P[P] = \bar{P}$ (the mean of P).

Example 1. *The signal samples, $x(i)$, $i = 1,2,\cdots,n$ represent the sum of a random, but otherwise constant parameter, P, having the uniform PDF over the interval, $(0,d)$, and a zero-mean stationary white Gaussian noise, $w(i)$, having the known variance, σ_w^2. The noise samples, $w(i)$, and the parameter, P, are independent. Find the MMSE estimate of P.*

Solution: 1. For $P = p$, the measured signal, $x(i) = p + w(i)$, has the conditional PDF,

$$\begin{aligned}f_{X|P}(x|p) &= \prod_{i=1}^{n} f_W(x(i)-p) \\ &= \frac{1}{\sqrt{(2\pi\sigma_w^2)^n}} \exp\left(-\frac{1}{2\sigma_w^2}\sum_{i=1}^{n}(x(i)-p)^2\right).\end{aligned} \tag{4}$$

The prior PDF of the parameter, $f_P(p) = \frac{1}{d}$, for $p \in (0,d)$, and 0, otherwise. Thus, the PDF of X is,

$$\begin{aligned}f_X(x) &= \int_{-\infty}^{\infty} f_{X|P}(x|p) f_P(p)\,\mathrm{d}p = \frac{1}{d}\int_0^d \prod_{i=1}^{n} f_W(x(i)-p)\,\mathrm{d}p \\ &= C_n \exp\left(-\frac{1}{2\sigma_w^2}\left(n\tilde{x}^2 - \sum_{i=1}^{n} x^2(i)\right)\right) \times \\ &\quad \times \left(Q\left(-\frac{\tilde{x}}{\sigma_w}\sqrt{n}\right) - Q\left(\frac{d-\tilde{x}}{\sigma_w}\sqrt{n}\right)\right)\end{aligned} \tag{5}$$

where $C_n = (d\sqrt{n})^{-1}(2\pi\sigma_w^2)^{-(n-1)/2}$, $Q(u) = \int_u^\infty \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-t^2/2}\,\mathrm{d}t$ (the Q-function), and, $\tilde{x} = \frac{1}{n}\sum_{i=1}^{n} x(i)$, so the MMSE estimator,

$$\begin{aligned}\hat{P}_{\mathrm{MMSE}}(n) &= \int_{-\infty}^{\infty} p\, f_{P|X}(p|x)\,\mathrm{d}p \\ &= \tilde{x} - \frac{\sigma_w}{\sqrt{n}} \frac{\dot{Q}\left(-\frac{\tilde{x}}{\sigma_w}\sqrt{n}\right) - \dot{Q}\left(\frac{d-\tilde{x}}{\sigma_w}\sqrt{n}\right)}{Q\left(-\frac{\tilde{x}}{\sigma_w}\sqrt{n}\right) - Q\left(\frac{d-\tilde{x}}{\sigma_w}\sqrt{n}\right)} \triangleq g(\tilde{x},n).\end{aligned} \tag{6}$$

For streaming data, the online MMSE estimator is obtained assuming a recurrent evaluation of the arithmetic average, i.e.,

$$\tilde{x} = \frac{1}{n}\sum_{i=1}^{n} x(i) = \frac{x(n)}{n} + \frac{n-1}{n}\underbrace{\frac{1}{n-1}\sum_{i=1}^{n-1} x(i)}_{\tilde{x}(n-1)}. \tag{7}$$

The corresponding circuit is shown in Figure 5. ■

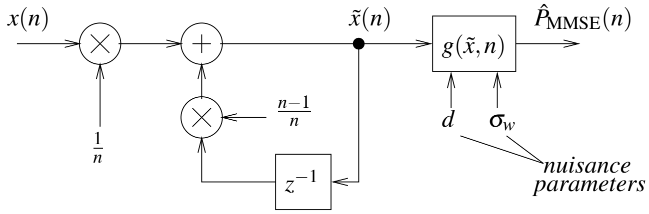

Figure 5. The MMSE estimator for streaming data from Example 1.

The nuisance parameters that appeared in Figure 1 and Figure 5 are the parameters that affect the measured signal, however, they are otherwise irrelevant in the application considered. Consequently, the nuisance parameters can be estimated, and these estimates simply ignored. Alternatively, the estimated nuisance parameters can be substituted into the estimator of the parameters of interest; this method is referred to as an adaptive estimation. The third strategy often used in practice is to eliminate the random nuisance parameters, N, from the distribution of the measurements, $X(P,N)$, i.e.,

$$\hat{P}_{\mathrm{MMSE}} = \int_{\{P\}} p \underbrace{\int_{\{N\}} f_{P,N|X}(p,n|X=x)\,\mathrm{d}n}_{f_{P|X}(p|X=x)}\,\mathrm{d}p = \mathrm{E}_P[P|X=x]. \tag{8}$$

The average MMSE estimator (8) is then considered to be good enough, for any specific values of N.

Omitting the derivations and proofs, the MMSE estimates have the following properties.

- The estimation error of the MMSE estimator has zero-mean, i.e., the MMSE estimates are unbiased:
$$\mathrm{E}\left[\hat{P}_{\mathrm{MMSE}}(X) - P\right] = 0 \;\Rightarrow\; \mathrm{E}\left[\hat{P}_{\mathrm{MMSE}}(X)\right] = \mathrm{E}[P]. \tag{9}$$
- The estimator variance can be expressed as,
$$\begin{aligned}\mathrm{var}\left[\hat{P}_{\mathrm{MMSE}} - P\right] &= \mathrm{E}\left[\left((\hat{P}_{\mathrm{MMSE}} - P) - \mathrm{E}\left[(\hat{P}_{\mathrm{MMSE}} - P)\right]\right)^2\right] \\ &= \mathrm{var}[P] - \mathrm{var}\left[\hat{P}_{\mathrm{MMSE}}\right].\end{aligned} \tag{10}$$
- The estimation error is uncorrelated with (i.e., orthogonal to) an arbitrary function of X, i.e.,
$$\begin{aligned}&\mathrm{cov}\left[\hat{P}_{\mathrm{MMSE}}(X), g(X)\right] = \mathrm{cov}[P, g(X)] \\ &= \mathrm{E}\left[\hat{P}_{\mathrm{MMSE}}(X) g(X)\right] - \underbrace{\mathrm{E}\left[\hat{P}_{\mathrm{MMSE}}(X)\right]}_{\mathrm{E}[P]} \mathrm{E}[g(X)].\end{aligned} \tag{11}$$

  Consequently, $\mathrm{cov}\left[\hat{P}_{\mathrm{MMSE}} - P, g(X)\right] = 0$, and for $g(X) = \hat{P}_{\mathrm{MMSE}}(X)$,
$$\mathrm{cov}\left[\hat{P}_{\mathrm{MMSE}} - P, \hat{P}_{\mathrm{MMSE}}\right] = 0. \tag{12}$$

Remark 2. *The unbiased estimator does not suffer from a systematic error. Moreover, the estimator quality is generally quantified as the variance of its estimation error.*

Theorem 1 (Gauss-Markov theorem)**.** *Let* $\mathbf{P}$ *and* $\mathbf{X}$ *be the vector of parameters to be estimated, and the vector of measurements, respectively. If* $\mathbf{P}$ *and* $\mathbf{X}$ *are jointly Gaussian,*

 ISBN: 978-1-68558-142-8

i.e., $\mathbf{P}$ *and* $\mathbf{X}$ *are Gaussian with the means,* $\bar{\mathbf{P}}$*, and,* $\bar{\mathbf{X}}$*, and the covariance matrices,* $\mathrm{var}[\mathbf{P}]$*, and,* $\mathrm{var}[\mathbf{X}]$*, respectively, then,*

$$\hat{\mathbf{P}}_{\mathrm{MMSE}}(\mathbf{x}) = \bar{\mathbf{P}} + \mathbf{H}(\mathbf{x}-\bar{\mathbf{X}}), \quad \mathbf{H} = \mathrm{cov}[\mathbf{P},\mathbf{X}]\,\mathrm{var}^{-1}[\mathbf{X}]. \quad (13)$$

The covariance matrix of the estimation error is,

$$\mathrm{var}\left[\hat{\mathbf{P}}_{\mathrm{MMSE}} - \mathbf{P}\right] = \mathrm{var}[\mathbf{P}] - \mathrm{cov}[\mathbf{P},\mathbf{X}]\,\mathrm{var}^{-1}[\mathbf{X}]\,\mathrm{cov}[\mathbf{X},\mathbf{P}]. \quad (14)$$

Example 2. *Let,* $x(t) = A + w(t)$*, be an observed signal over the time interval,* $t \in (0,T)$*, where* A *is a normally distributed random variable with known mean,* $\bar{A}$*, and known variance,* σ_A^2*, and* $w(t)$ *represents a zero-mean Additive White Gaussian Noise (AWGN) having the known variance,* C_0*. Estimate the value of* A *from the measured signal,* $x(t)$*.*

Solution: 2. Assuming Gauss-Markov theorem, after some straightforward derivations, the estimator is obtained as,

$$\hat{A}_{\mathrm{MMSE}} = \frac{C_0}{T\sigma_A^2 + C_0}\bar{A} + \frac{\sigma_A^2}{T\sigma_A^2 + C_0}\int_0^T x(t)\,\mathrm{d}t. \quad (15)$$

In addition, if $T\sigma_A^2 \gg C_0$, then the estimator can be simplified as,

$$\hat{A}_{\mathrm{MMSE}} = \frac{C_0\bar{A}}{T\sigma_A^2} + \frac{1}{T}\int_0^T x(t)\,\mathrm{d}t. \quad (16)$$

■

Example 3. *The samples received in a data packet of* N *symbols are expressed as,* $x(i) = p\,s(i) + w(i)$*,* $i = 1,2,\dots,N$*, where* p *represents the channel attenuation,* $s(i)$ *is transmitted modulation symbol, and* $w(i)$ *is the sample of an AWGN. Discuss how to estimate the channel attenuation,* p*.*

Solution: 3. The attenuation, p, is usually a complex-valued zero-mean Gaussian random variable. Since estimating unknown p while also detecting unknown $s(i)$ at the same time is not possible, the first n out of N data symbols are usually reserved for so-called pilot symbols, which are known at the receiver. For example, let $s(i) = s_0$, for $i = 1,\dots,n$. Then, the received symbols, ps_0, have zero mean, and the variance, $\mathrm{E}\left[|ps_0|^2\right] = \mathrm{E}\left[|p|^2\right]|s_0|^2$, and Gauss-Markov theorem can be used to estimate, ps_0, i.e., to estimate, p. ■

B. The MAP Estimator

If P is a discrete random variable, then it is meaningful to define its estimation error as,

$$\mu(\hat{P},P) = \begin{cases} 0 & \hat{P} = P \\ 1 & \hat{P} \neq P. \end{cases} \quad (17)$$

The corresponding risk is equal to the probability of error, i.e.,

$$\mathrm{E}\left[\mu(\hat{P}(X),P)\right] = \Pr\big(\hat{P}(X) \neq P\big) = 1 - \Pr\big(\hat{P}(X) = P\big). \quad (18)$$

This yields the MAP estimator,

$$\begin{aligned}\hat{P}_{\mathrm{MAP}}(X) &= \mathrm{argmin}_{p_i}\,\mathrm{E}[\mu(p_i,P)|X=x] \\ &= \mathrm{argmax}_{p_i}\,\Pr(P = p_i|X=x).\end{aligned} \quad (19)$$

If P is a continuous random variable, then the estimation error can be defined as,

$$\mu(\hat{P},P) = \begin{cases} 0 & |\hat{P}(X) - P| < \Delta \\ 1 & |\hat{P}(X) - P| \geq \Delta. \end{cases} \quad (20)$$

If Δ in (20) is small, i.e., the probability that $\hat{P}$ is close to P is large, then,

$$\begin{aligned}\hat{P}_{\mathrm{MAP}}(X) &= \mathrm{argmax}_{\hat{P}}\, f_{P|X}(\hat{P}|X=x) \\ &= \mathrm{argmax}_{\hat{P}}\, f_{X|P}(x|\hat{P})f_P(\hat{P})\end{aligned} \quad (21)$$

where $f_{X|P}(x|\hat{P})$ represents the likelihood function. In practice, the maximum can be obtained by assuming the derivatives, i.e., $\frac{\mathrm{d}}{\mathrm{d}P} f_{X|P}(x,\hat{P})f_P(\hat{P}) = 0$, or, $\frac{\mathrm{d}}{\mathrm{d}P}\log(f_{X|P}(x,\hat{P})f_P(\hat{P})) = 0$.

III. General Estimation of Non-Random Parameters

The estimators introduced in the previous subsection require that the prior statistical description of the parameters to be estimated is completely known. If this is not the case, then the parameters can treated as being non-random. The caveat is that the estimator of a non-random parameter, P, is often much more difficult to find, and it may not even exist. For example, assuming the MMSE criterion, minimizing the risk, $\mathrm{E}\left[\mu(\hat{P}(X),P)\right] = \int_{\{X\}}(\hat{P}(X)-P)^2 f_X(x,P)\,\mathrm{d}x$, yields correct, but otherwise useless solution, $\hat{P} = P$. Figure 6 shows the examples when the optimum estimator (in the sense of minimizing the average risk) exists, and when it does not exist.

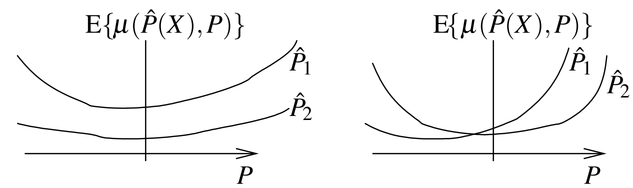

Figure 6. The examples when the optimum estimator exists (left), and when it does not exist (right).

The estimation of a non-random parameter, P, generally relies on knowledge of the statistical dependence of the measured values, X, i.e., on the parameterized PDF or PMF, $f_X(x,P)$, or, $\Pr_X(x,P)$, respectively, which must satisfy,

$$\int_{\{X\}} f_X(x,P)\,\mathrm{d}x = 1, \text{ or, } \sum_{x\in\{X\}} \Pr(X=x,P) = 1, \ \forall P. \quad (22)$$

A. The MVUB Estimator

The minimum variance unbiased (MVUB) estimator of a non-random parameter, P, is unbiased, i.e., $\mathrm{E}\left[\hat{P}\right] = P$, for $\forall P$. The variance of the estimation error of an unbiased estimator is,

$$\mathrm{var}\left[\hat{P}(X) - P\right] = \underbrace{\mathrm{E}\left[(\hat{P}(X)-P)^2\right]}_{\mathrm{MSE}} = \mathrm{var}\left[\hat{P}(X)\right]. \quad (23)$$

Consequently, the MVUB estimator minimizes the MSE equal to the variance of $\hat{P}$. However, the MVUB estimator may not exist, i.e., there may be no such function of X having the smallest variance for any value of P, as shown in Figure 6.

There are several important notions to describe the asymptotic accuracy of the estimators of non-random parameters. In particular, the CRLB defines the minimum achievable variance

 ISBN: 978-1-68558-142-8

of any unbiased estimator of a non-random parameter, P. It is mathematically formulated as,

$$\mathrm{var}\left[\hat{P}\right] \geq \frac{1}{J(P)} \tag{24}$$

where Fisher information, $J(P)$, is computed as,

$$J(P) = \mathrm{E}\left[\left(\frac{\partial \ln f_X(x,P)}{\partial P}\right)^2\right] = \mathrm{E}\left[-\frac{\partial^2 \ln f_X(x,P)}{\partial P^2}\right]. \tag{25}$$

Note that, $\ln f_X(x,P)$, represents the log-likelihood function, and showing the equality between the two expectations in (25) requires a derivation.

The estimator is said to be *efficient*, provided that it is unbiased, and it attains the CRLB defined in (24). Furthermore, the estimator is said to be *consistent*, provided that the variance, $\mathrm{var}\left[\hat{P}\right]$, is decreasing with the number of measurements, X.

Remark 3. *There may be estimators that are slightly biased, but which have the variance smaller than the CRLB. For example, the estimator design can be constrained as,* $\mathrm{E}\left[|\hat{P}-P|\right] < \Delta$, *to allow that it may possibly be unbiased. The trade-off between the bias and the variance frequently appears in training the machine learning models. Moreover, the consistency guarantees that collecting more data samples improves the estimator accuracy, which is also useful for machine learning.*

B. The ML Estimator

The MVUB estimator may not exist, or it is difficult to find. However, given the measurement, $X = x$, unless we are very unlucky, it is meaningful to choose the estimate of P to be the value with the largest likelihood, i.e.,

$$\begin{aligned} \hat{P}_{\mathrm{ML}}(X) &= \mathrm{argmax}_{\hat{P}}\, f_X(x,\hat{P}), \text{ or} \\ \hat{P}_{\mathrm{ML}}(X) &= \mathrm{argmax}_{\hat{P}}\, \mathrm{Pr}_X\left(X = x,\hat{P}\right). \end{aligned} \tag{26}$$

The estimator (26) is referred to as the ML estimator.

The ML estimator has the following properties.

- If the efficient estimate exists, then it is the ML estimate.
- If the efficient estimate does not exist, then the ML estimate is neither guaranteed to have the minimum variance, nor to be unbiased.
- The ML estimator is asymptotically unbiased as well as asymptotically efficient.
- The ML estimator is invariant to any function, $g(P)$, i.e., the ML estimate of $g(P)$ can be obtained as,

$$\hat{g}_{\mathrm{ML}}(P) = g(\hat{P}_{\mathrm{ML}}). \tag{27}$$

Remark 4. *Even though the ML estimator may not be unbiased, it is generally very attractive for its simplicity to obtain it. Moreover, if all the values of P are a priori equally likely, then the ML estimator and the MAP estimator are identical.*

Example 4. *There were x errors detected in a binary sequence of n bits. Assuming that the errors are independent, and they occur with the probability, P, decide whether the estimator* $\hat{P} = x/n$ *of P is the MVUB estimator.*

Solution: 4. The variance of the estimator, $\hat{P} = x/n$, is, $\mathrm{var}\left[\hat{P}\right] = P(1-P)/n$. The probability of x errors occurring among n bits is, $\Pr(X = x,P) = \binom{n}{x}P^x(1-P)^{n-x}$, so that $\frac{\partial}{\partial P}\ln \Pr(X = x,P) = \frac{n}{P(1-P)}\left(\frac{x}{n} - P\right)$. Since the estimator, $\hat{P} = x/n$, can be shown to be unbiased, and $\mathrm{E}\left[\left(\frac{\partial \ln \Pr(X=x,P)}{\partial P}\right)^2\right] = \mathrm{var}\left[\hat{P}\right]^{-1}$, the estimator is indeed the MVUB estimator. ■

Example 5. *The N samples of an unknown constant a were sampled in a zero-mean AWGN with an unknown variance,* σ^2. *Find the ML estimate of a.*

Solution: 5. Assuming, $x(i) = a + w(i)e$, the joint PDF of N observed samples is,

$$f_X(x,a,\sigma) = \frac{1}{\sqrt{(2\pi\sigma)^N}} \exp\left(-\frac{1}{2\sigma}\sum_{i=1}^{N}(x(i)-a)^2\right). \tag{28}$$

The ML estimates of the unknown parameters, a and σ, respectively, must satisfy,

$$\begin{aligned} \frac{\partial \ln f_X(x,\hat{a}_{\mathrm{ML}},\hat{\sigma}_{\mathrm{ML}})}{\partial \hat{a}_{\mathrm{ML}}} &= \frac{1}{\hat{\sigma}_{\mathrm{ML}}}\sum_{i=1}^{N}(x(i)-\hat{a}_{\mathrm{ML}}) \stackrel{!}{=} 0 \\ \frac{\partial \ln f_X(x,\hat{a}_{\mathrm{ML}},\hat{\sigma}_{\mathrm{ML}})}{\partial \hat{\sigma}_{\mathrm{ML}}} &= -\frac{N}{2}\frac{1}{\hat{\sigma}_{\mathrm{ML}}} + \frac{1}{2\hat{\sigma}^2_{\mathrm{ML}}}\sum_{i=1}^{N}(x(i)-\hat{a}_{\mathrm{ML}})^2 \stackrel{!}{=} 0. \end{aligned} \tag{29}$$

Solving these two equations for $\hat{a}_{\mathrm{ML}}$ and $\hat{\sigma}_{\mathrm{ML}}$, their estimators become,

$$\begin{aligned} \hat{a}_{\mathrm{ML}} &= \frac{1}{N}\sum_{i=1}^{N}x(i), \text{ and} \\ \hat{\sigma}_{\mathrm{ML}} &= \frac{1}{N}\sum_{i=1}^{N}(x(i)-\hat{a}_{\mathrm{ML}})^2. \end{aligned} \tag{30}$$

The estimate, $\hat{a}_{\mathrm{ML}}$, has the mean, $\mathrm{E}[\hat{a}_{\mathrm{ML}}] = \frac{1}{N}\sum_{i=1}^{N}\mathrm{E}[x(i)] = \frac{1}{N}\sum_{i=1}^{N}a = a$, and the variance of estimation error, $\mathrm{E}\left[(\hat{a}_{\mathrm{ML}}-a)^2\right] = \frac{1}{N^2}\sum_{i=1}^{N}\mathrm{var}[x(i)] = \frac{1}{N^2}\sum_{i=1}^{N}\sigma = \sigma/N$. On the other hand, for the estimate, $\hat{\sigma}_{\mathrm{ML}}$, $\mathrm{E}[\hat{\sigma}_{\mathrm{ML}}] = \sigma\frac{N-1}{N}$, i.e., $\mathrm{E}[\hat{\sigma}_{\mathrm{ML}}] \neq \sigma$. Hence, the ML estimate of σ is only asymptotically unbiased, and the variance of estimation error, $\mathrm{E}\left[(\hat{\sigma}_{\mathrm{ML}}-\sigma)^2\right] = \sigma^2\frac{2N-1}{N^2}$ becomes asymptotically (for large N) equal to the CRLB, $2\sigma^2/N$, proving that the ML estimator of σ is asymptotically efficient. ■

Many practical scenarios involve measurements in an AWGN. Specifically, in discrete time, assume the measured samples, $x(i) = g(p) + w(i)$, where $g(\cdot)$ denotes a non-linear function, and $w(i)$ is a zero-mean AWGN with unknown variance. The ML estimate of P is then,

$$\hat{P}_{\mathrm{ML}} = \mathrm{argmin}_{\hat{P}}\sum_{i=1}^{N}|x(i)-g(\hat{P})|^2. \tag{31}$$

Similarly, in continuous time, assume the observed signal, $x(t) = g(t,p) + w(t)$, over the time, $t \in (0,T)$, where $g(t,p)$ is a signal dependent on $P = p$, and $w(t)$ denotes a stationary zero-mean AWGN with unknown variance. The ML estimate of P is then,

$$\hat{P}_{\mathrm{ML}} = \mathrm{argmin}_{\hat{P}}\int_0^T |x(t)-g(t,\hat{P})|^2\,\mathrm{d}t. \tag{32}$$

ISBN: 978-1-68558-142-8

Example 6. *Find the ML estimate of a constant phase shift,* θ, *of the unmodulated carrier signal,* $x(t) = A\cos(\omega_c t + \theta) + w(t)$, *received over the time,* $t \in (0, T)$, *and assuming that,* $\omega_c T \gg 1$.

Solution: 6. The ML estimate of θ is,

$$\begin{aligned} \hat{\theta}_{\mathrm{ML}} &= \mathrm{argmin}_{\hat{\theta}} \int_0^T (x(t) - A\cos(\omega_c t + \hat{\theta}))^2 \,\mathrm{d}t \\ &\approx \mathrm{argmax}_{\hat{\theta}} \int_0^T x(t)\cos(\omega_c t + \hat{\theta})\,\mathrm{d}t \\ &= \angle \underbrace{\int_0^T x(t)\,\mathrm{e}^{-\mathrm{j}\omega_c t}\,\mathrm{d}t}_{X(\omega_c)}. \end{aligned} \tag{33}$$

Alternatively,

$$\begin{aligned} &\frac{\partial}{\partial \hat{\theta}_{\mathrm{ML}}} \ln f_X(x, A, \omega_c, \hat{\theta}_{\mathrm{ML}}) \stackrel{!}{=} 0 \\ \Rightarrow &\int_0^T x(t)\sin(\omega_c t + \hat{\theta}_{\mathrm{ML}})\,\mathrm{d}t = 0. \end{aligned} \tag{34}$$

The corresponding implementations are shown in Figure 7. The bottom circuit in Figure 7 is a Phase-Locked Loop (PLL). It uses the output signal of a Voltage-Controlled Oscillator (VCO) to recover the carrier signal in order to enable the coherent detection of transmitted data symbols. ∎

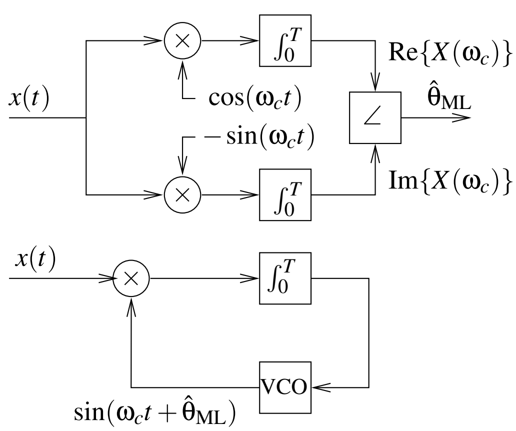

Figure 7. The two ML estimators of phase shift, θ, of the unmodulated noisy carrier signal, $x(t)$.

C. The LS Estimator

It may be sometimes impractical or impossible to obtain the distribution of the measurements, X. However, if a reasonably good model, $g(P)$, of $X(P)$ can be obtained, so that, $X(P) \approx g(P)$, then, the optimum estimator of the non-random parameter, P, can be defined as,

$$\hat{P}_{\mathrm{opt}}(X) = \mathrm{argmin}_{\hat{P}}\, \mu\big(X, g(\hat{P})\big). \tag{35}$$

The corresponding LS estimator is obtained by assuming N measurements, X_i, and the error function,

$$\mu\big(X, g(\hat{P})\big) = \sum_{i=1}^{N} v_i \left(X_i - g(\hat{P})\right)^2 \tag{36}$$

where v_i are the weights to (de-)emphasize the measurements.

For time-continuous measurements, $x(t)$, the LS estimator is,

$$\hat{P}_{\mathrm{LS}} = \mathrm{argmin}_{\hat{P}} \int_0^T \left(x(t) - g(t, \hat{P})\right)^2 \mathrm{d}t. \tag{37}$$

If the parameter, P, is continuous, the minimization of (35) or (37) can be performed by differentiation, and then numerically findings the root of a non-linear function.

Remark 5. *The LS estimator corresponds to the ML estimator, provided that the measurement noise is AWGN.*

D. The Moments Based Estimator

Both the ML and the LS estimators may be too complex to implement, since they require finding the extremum of a generally non-linear function. An alternative approach for estimating the non-random parameter is to match its statistical moments. In particular, the n-th general moment of a random variable, P, is defined as,

$$\begin{aligned} g_n(P) &= \mathrm{E}_P[P^n] = \int_{\{P\}} p^n f_P(p)\,\mathrm{d}p, \text{ or} \\ g_n(P) &= \mathrm{E}_P[P^n] = \sum_{p \in \{P\}} p^n \Pr(P = p). \end{aligned} \tag{38}$$

If the measurements are stationary, i.e, they have the same moments, the n-th moment, $g_n(P)$, can be estimated from N measurements, x_i, as,

$$\hat{g}_n(P) = \frac{1}{N}\sum_{i=1}^{N} x^n(i). \tag{39}$$

Subsequently, the estimate of P is obtained by using the inverse function, g_n^{-1}, i.e.,

$$\hat{P} = g_n^{-1}\left(\frac{1}{N}\sum_{i=1}^{N} x^n(i)\right). \tag{40}$$

Remark 6. *In practice, the moments order, n, is typically assumed to be small, since it is more difficult to reliably estimate higher-order moments, and the estimation error increases with n. The estimator* (40) *is unbiased, and consistent, i.e.,* $\lim_{N\to\infty} \hat{g}_n(P) = g_n(P)$, *however, there are otherwise no guarantees about its optimality. Note also that although the parameter, P, may be a random variable, its distribution is unknown, and thus, for the purpose of its estimation, it is considered to be non-random.*

IV. Linear Estimation of Random Parameters

The general estimation strategy is to seek an optimum function to minimize the mean estimation error, i.e., the risk. This requires a full statistical description of measurements, X, as a function of the unknown parameter, P. Such a dependence is normally expressed as the conditional distribution, $f_{X|P}(x|p)$, or the conditional probability, $\Pr(X|P)$, when P is considered to be a random variable, and the parameterized distribution, $f_X(x;P)$, or the parameterized probability, $\Pr(X;P)$, when P is non-random.

Provided that only some statistics of the parameter, P, are known, such as its mean and the variance, they are sufficient

 ISBN: 978-1-68558-142-8

to define an optimum linear estimator. Such an estimator is simply a linear filter, which can be written as,

$$\hat{\mathbf{P}}(\mathbf{X}) = \mathbf{a} + \mathbf{H}\mathbf{X} \tag{41}$$

where $\mathbf{X}$ is the vector of measurements, $\mathbf{P}$ is the vector of parameters to be estimated, and $\hat{\mathbf{P}}(\mathbf{X})$ is the vector of the estimates. The vector $\mathbf{a}$ and the matrix $\mathbf{H}$ represent the filter coefficients, which are independent of the actual values, $\mathbf{P}$, and, $\mathbf{X}$. Consequently, given the statistics of $\mathbf{X}$ and $\mathbf{P}$, such as, $\mathrm{E}[\mathbf{X}]$, $\mathrm{var}[\mathbf{X}]$, $\mathrm{E}[\mathbf{P}]$, $\mathrm{var}[\mathbf{P}]$, and $\mathrm{cov}[\mathbf{P},\mathbf{X}]$), the task is to find the optimum filter coefficients, $\mathbf{a}$, and, $\mathbf{H}$.

A. The LMMSE Estimator

The Linear MMSE (LMMSE) estimator is unbiased, i.e., $\mathrm{E}\left[\hat{\mathbf{P}}(\mathbf{X})\right] = \mathrm{E}[\mathbf{P}]$, or, $\mathrm{E}\left[\hat{\mathbf{P}}(\mathbf{X}) - \mathbf{P}\right] = \mathbf{0}$, where $\mathbf{0}$ denotes the all-zeros vector. For the scalar, P, the LMMSE estimator minimizes the variance, $\mathrm{var}\left[\hat{P}(X) - P\right] = \mathrm{E}\left[(\hat{P}(X) - P)^2\right]$. In case of the vector, $\mathbf{P}$, the LMMSE estimator minimizes the correlation matrix,

$$\mathrm{var}\left[\hat{\mathbf{P}}(\mathbf{X}) - \mathbf{P}\right] = \mathrm{E}\left[(\hat{\mathbf{P}}(\mathbf{X}) - \mathbf{P})(\hat{\mathbf{P}}(\mathbf{X}) - \mathbf{P})^T\right]. \tag{42}$$

Remark 7. *The square matrix,* $\mathbf{A}$*, is minimized, provided that it is positively semi-definite, i.e.,* $\mathbf{u}^T\mathbf{A}\mathbf{u} \geq 0$*, as well as* $\mathbf{u}^T\mathbf{A}\mathbf{u}$ *has the minimum value for some non-zero vector,* $\mathbf{u}$*.*

The LMMSE estimation minimizes the variance matrix, $\mathrm{var}\left[\hat{\mathbf{P}}(\mathbf{X}) - \mathbf{P}\right]$. It is straightforward to show that the minimum occurs, when,

$$\mathbf{a} = \bar{\mathbf{P}} - \mathbf{H}\bar{\mathbf{X}}, \text{ and, } \mathbf{H} = \mathrm{cov}[\mathbf{P},\mathbf{X}]\,\mathrm{var}^{-1}[\mathbf{X}] \tag{43}$$

requiring only knowledge of $\bar{\mathbf{P}} = \mathrm{E}[\mathbf{P}]$, $\bar{\mathbf{X}} = \mathrm{E}[\mathbf{X}]$, $\mathrm{var}[\mathbf{X}]$, and $\mathrm{cov}[\mathbf{P},\mathbf{X}]$. The LMMSE estimate is then computed as,

$$\hat{\mathbf{P}}_{\mathrm{LMMSE}}(\mathbf{X}) = \bar{\mathbf{P}} + \mathbf{H}(\mathbf{X} - \bar{\mathbf{X}}). \tag{44}$$

The LMMSE estimator has the following properties.

- The estimates are unbiased, i.e., $\mathrm{E}\left[\hat{\mathbf{P}}_{\mathrm{LMMSE}}(\mathbf{X})\right] = \mathrm{E}[\mathbf{P}]$.
- The estimation error and the measurements are uncorrelated (orthogonal), i.e., $\mathrm{cov}\left[\hat{\mathbf{P}}_{\mathrm{LMMSE}} - \mathbf{P}, \mathbf{X}\right] = \mathbf{0}$.
- The estimation error and the estimate are uncorrelated, i.e., $\mathrm{cov}\left[\hat{\mathbf{P}}_{\mathrm{LMMSE}} - \mathbf{P}, \hat{\mathbf{P}}_{\mathrm{LMMSE}}\right] = \mathbf{0}$.
- The variance matrix of the estimate, $\mathrm{var}\left[\hat{\mathbf{P}}_{\mathrm{LMMSE}}\right] = \mathbf{H}\,\mathrm{var}[\mathbf{X}]\,\mathbf{H}^T = \mathbf{H}\,\mathrm{cov}[\mathbf{X},\mathbf{P}]$.
- The covariance matrix of the estimation errors is equal to,

$$\begin{aligned} \mathbf{S} &= \mathrm{E}\left[(\hat{\mathbf{P}}_{\mathrm{LMMSE}} - \mathbf{P})(\hat{\mathbf{P}}_{\mathrm{LMMSE}} - \mathbf{P})^T\right] \\ &= \mathrm{var}\left[\hat{\mathbf{P}}_{\mathrm{LMMSE}} - \mathbf{P}\right] = \mathrm{var}[\mathbf{P}] - \mathrm{var}\left[\hat{\mathbf{P}}_{\mathrm{LMMSE}}\right]. \end{aligned} \tag{45}$$

- The estimator, which is linear, unbiased and orthogonal, is the LMMSE estimator.

Example 7. *The stationary signal,* $x(t)$*, with a known auto-covariance,* $K_x(\tau) = \mathrm{E}[(x(t+\tau) - \bar{x})(x(t) - \bar{x})]$*, is sampled at three time instances,* $t \in \{t_0 - \Delta t, t_0, t_0 + \Delta t\}$*. Find the LMMSE estimate,* $\hat{\dot{x}}(t_0)$*, of the derivative,* $\frac{\mathrm{d}}{\mathrm{d}t}x(t)|_{t=t_0} = \dot{x}(t_0)$*.*

Solution: 7. Define the vectors, $P = \dot{x}(t_0)$, and, $\mathbf{X} = [x(t_0 - \Delta t), x(t_0), x(t_0 + \Delta t)]^T$. Then, the covariance functions are written as, $K_x(\tau) = K_x(-\tau)$, $K_{\dot{x},x}(\tau) = \dot{K}_x(\tau)$, $K_{x,\dot{x}}(\tau) = -\dot{K}_x(\tau)$, $K_{\dot{x}}(\tau) = -\ddot{K}_x(\tau)$, and $\mathrm{var}[\dot{x}(t_0)] = K_{\dot{x}}(0)$. The corresponding cross-covariance vector,

$$\mathrm{cov}[P,\mathbf{X}] = [\dot{K}_x(\Delta t), \underbrace{\dot{K}_x(0)}_{0}, \underbrace{\dot{K}_x(-\Delta t)}_{-\dot{K}_x(\Delta t)}] \tag{46}$$

and the variance matrix,

$$\mathrm{var}[\mathbf{X}] = \begin{bmatrix} K_x(0) & K_x(\Delta t) & K_x(2\Delta t) \\ K_x(\Delta t) & K_x(0) & K_x(\Delta t) \\ K_x(2\Delta t) & K_x(\Delta t) & K_x(0) \end{bmatrix}. \tag{47}$$

These expressions can be substituted into, $\hat{\mathbf{P}}_{\mathrm{LMMSE}}(\mathbf{X}) = \bar{\mathbf{P}} + \mathbf{H}(\mathbf{X} - \bar{\mathbf{X}})$, and, $\mathbf{H} = \mathrm{cov}[\mathbf{P},\mathbf{X}]\,\mathrm{var}^{-1}[\mathbf{X}]$, to get the estimator,

$$\begin{aligned} \hat{\dot{x}}(t_0) = & \bar{\dot{x}}(t_0) + \underbrace{\frac{\dot{K}_x(\Delta t)}{K_x(0) - K_x(2\Delta t)}}_{\text{const}} \times \\ & (x(t_0 - \Delta t) - x(t_0 + \Delta t) - \bar{x}(t_0 - \Delta t) + \bar{x}(t_0 + \Delta t)). \end{aligned} \tag{48}$$

■

Example 8. *The signal samples,* $x(i)$*,* $i = 1, 2, \cdots, n$*, represent the sum of a random, but otherwise constant parameter,* P*, having the uniform probability distribution over the interval,* $(0,d)$*, and a stationary AWGN,* $w(i)$*, with zero-mean, and a known variance,* σ_w^2*. The noise,* $w(i)$*, and the parameter,* P*, are independent. Find the LMMSE estimate of* P*.*

Solution: 8. Define the vector, $\mathbf{X} = [x(1), \ldots, x(n)]^T$, having the elements, $x(i) = P + w(i)$, so that, $\mathbf{X} = [1, \cdots, 1]^T P + \mathbf{W}$. The parameter has the mean value, $\mathrm{E}[P] = d/2$, and the variance, $\mathrm{var}[P] = d^2/12$, while the noise, $\mathrm{E}[w(i)] = 0$, and, $\mathrm{var}[w(i)] = \sigma_w^2$. This yields the LMMSE estimator,

$$\hat{P}(n) = \frac{nd^2}{nd^2 + 12\sigma_w^2}\left(\frac{6\sigma_w^2}{nd} + \frac{1}{n}\sum_{i=1}^{n} x(i)\right). \tag{49}$$

■

V. Linear Estimation of Non-Random Parameters

If the vector of parameters, $\mathbf{P}$, is non-random, then, $\mathrm{E}[\mathbf{P}] = \mathbf{P}$, and, $\mathrm{cov}[\mathbf{P},\mathbf{X}] = \mathrm{E}\left[(\mathbf{P} - \mathbf{P})(\mathbf{X} - \bar{\mathbf{X}})^T\right] = \mathbf{0}$. Substituting these expressions into the LMMSE estimator, the solution, $\hat{\mathbf{P}}_{\mathrm{LMMSE}} = \mathbf{P}$, is correct, but useless. As for general estimators of non-random parameters, a different strategy is required.

Assuming a class of linear unbiased estimators, i.e., the estimators of the form, $\hat{\mathbf{P}}(\mathbf{X}) = \mathbf{a} + \mathbf{H}\mathbf{X}$, then,

$$\mathrm{E}\left[\hat{\mathbf{P}}(\mathbf{X})\right] = \mathbf{a} + \mathbf{H}\bar{\mathbf{X}} \stackrel{!}{=} \mathbf{P},\ \forall\mathbf{P} \tag{50}$$

and thus, $\bar{\mathbf{X}} = \mathbf{D}\mathbf{P} + \mathbf{r}$, such that, $\mathbf{H}\mathbf{D} = \mathbf{I}$, and, $\mathbf{a} = -\mathbf{H}\mathbf{r}$. The matrix, $\mathbf{D}$, and the vector, $\mathbf{r}$, are assumed to be known and independent of $\mathbf{P}$, where $\mathbf{I}$ is the identity matrix. Consequently, the linear unbiased estimator of $\mathbf{P}$ is written as,

$$\hat{\mathbf{P}} = \mathbf{H}(\mathbf{X} - \mathbf{r}) \tag{51}$$

under the constraints,

$$\bar{\mathbf{X}} = \mathrm{E}[\mathbf{X}] = \mathbf{DP} + \mathbf{r}, \text{ and, } \mathbf{HD} = \mathbf{I}. \tag{52}$$

A. The BLUE Estimator

The Best Linear Unbiased Estimator (BLUE) is a linear estimation with the smallest variance. If the number of measurements (the length of vector, $\mathbf{X}$), is equal to the number of parameters to be estimated (the length of vector, $\mathbf{P}$), then the BLUE estimator is defined as,

$$\hat{\mathbf{P}} = \mathbf{H}(\mathbf{X} - \mathbf{r}), \ \mathbf{H} = \mathbf{D}^{-1}. \tag{53}$$

On the other hand, if the length of vector, $\mathbf{X}$, is greater than the length of vector, $\mathbf{P}$, which is often the case, then the BLUE estimator is defined as the one that minimizes the correlation matrix of the estimation error, i.e.,

$$\begin{aligned}\mathbf{S} &= \underbrace{\mathrm{E}\left[(\hat{\mathbf{P}}(\mathbf{X}) - \mathbf{P})(\hat{\mathbf{P}}(\mathbf{X}) - \mathbf{P})^T\right]}_{\text{correlation matrix}} = \underbrace{\mathrm{var}\left[\hat{\mathbf{P}}(\mathbf{X}) - \mathbf{P}\right]}_{\text{variance matrix}} \\ &= \mathrm{var}\left[\mathbf{H}(\mathbf{X} - \bar{\mathbf{X}})\right] = \mathbf{H}\,\mathrm{var}[\mathbf{X}]\,\mathbf{H}^T.\end{aligned} \tag{54}$$

Thus, the matrix, $\mathbf{S}$, is equal to the variance matrix of the estimation error, since the estimator is unbiased, and it also equal to the variance matrix of the estimate, since $\mathbf{P}$ is considered to be non-random.

Furthermore, if $\mathrm{var}[\mathbf{X}]$ is independent of $\mathbf{P}$, which is not always guaranteed, the BLUE estimator, $\hat{\mathbf{P}} = \mathbf{H}(\mathbf{X} - \mathbf{r})$, is defined by the matrix,

$$\mathbf{H} = \left(\mathbf{D}^T \mathrm{var}^{-1}[\mathbf{X}]\,\mathbf{D}\right)^{-1} \mathbf{D}^T \mathrm{var}^{-1}[\mathbf{X}]. \tag{55}$$

The corresponding correlation matrix is then,

$$\mathbf{S} = \left(\mathbf{D}^T \mathrm{var}^{-1}[\mathbf{X}]\,\mathbf{D}\right)^{-1}. \tag{56}$$

Example 9. *The radar determines the distance to a target using three measurements, x_1, x_2 and x_3. The measurements are distorted by a zero-mean additive errors, w_i, $i = 1,2,3$. The correlations between the measurements are dependent on their separation in time, i.e., (a) if the time separation is T_1, then, $r_{12} = r_{23} = 0.9$, and, $r_{13} = 0.7$; (b) if the time separation is T_2, then $r_{12} = r_{23} = 0.7$, and, $r_{13} = 0.4$; and (c) if the time separation is T_3, then $r_{12} = r_{23} = r_{13} = 0$. Obtain the BLUE estimator of the distance from the three measurements, and also calculate the mean square error of the estimate. Assume that the variance of the measurement errors is,* $\mathrm{var}[w_i] = 30\ \mathrm{m}^2$.

Solution: 9. In this case, the vector, $\mathbf{D}$, and the variance matrix of $\mathbf{X}$ are defined, respectively, as,

$$\mathbf{D} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \quad \mathrm{var}[\mathbf{X}] = 30 \begin{bmatrix} 1 & r_{12} & r_{13} \\ r_{12} & 1 & r_{23} \\ r_{13} & r_{23} & 1 \end{bmatrix}. \tag{57}$$

After substituting the specific values for the correlation coefficients, r_{12}, r_{23}, and r_{13}, the matrices, $\mathrm{var}^{-1}[\mathbf{X}]$, and, the vector, $\mathbf{D}^T \mathrm{var}^{-1}[\mathbf{X}]$, can be computed including the scalar value, $\mathbf{D}^T \mathrm{var}^{-1}[\mathbf{X}]\,\mathbf{D}$. Subsequently, the estimators and their mean square errors are obtained for each case of the time-separation, i.e.,

(a) T_1: $\hat{p} = x_1 - x_2 + x_3$, $S = 24\ \mathrm{m}^2$;
(b) T_2: $\hat{p} = (x_1 + x_3)/2$, $S = 21\ \mathrm{m}^2$;
(c) T_3: $\hat{p} = (x_1 + x_2 + x_3)/3$, $S = 10\ \mathrm{m}^2$.

■

Remark 8. *In Example 9, if the measurements are uncorrelated, the estimate is a simple arithmetic average, and the estimator variance is the smallest. In other two cases, the unequal combining weights account for the non-zero correlations between the measurements.*

VI. Additional Solved Problems

Example 10. *An unmodulated carrier is measured in discrete time in the presence of a zero-mean AWGN with an unknown variance, σ_w^2, i.e.,*

$$x(i) = A\cos(\beta i + \Phi) + w(i), \quad i = 1,2,\cdots,N \tag{58}$$

where β is a known angular frequency, which can be assumed to be, $-\pi < \beta < \pi$, A is a Rayleigh distributed random amplitude, and Φ is a uniformly distributed random phase over the interval, $(-\pi,\pi)$. The amplitude, A, and the phase, Φ, are independent, and let the measured mean received power be also known. Find the MAP estimate of the phase Φ.

Solution: 10. The amplitude A is a nuisance parameter, which can be averaged out from the likelihood function, $f_{X|A,\Phi}(x|a,\phi)$, of the received signal, $x(i)$. Provided that,

$$\left|\frac{\sin(N\beta)}{N\sin(\beta)}\right| \ll 1, \tag{59}$$

the MAP estimate of Φ is obtained as (after some derivations),

$$\hat{\Phi}_{\mathrm{MAP}} = -\angle(I + \mathrm{j}Q) \tag{60}$$

where

$$\begin{aligned} I &= \sum_{i=1}^{N} x(i)\cos(\beta i) \\ Q &= \sum_{i=1}^{N} x(i)\sin(\beta i). \end{aligned} \tag{61}$$

■

Example 11. *An unmodulated carrier with unknown amplitude, phase and angular frequency is measured in a zero-mean AWGN with an unknown variance, σ_w^2, i.e.,*

$$x(i) = A\cos(\beta i + \phi) + w(i), \quad i = 0,1,\cdots,N-1 \tag{62}$$

where $\beta \in (0,\pi)$. Find the ML estimate of all the unknown parameters, provided that $|\sin(N\beta)/(N\sin(\beta))| \ll 1$.

Solution: 11. The unknown parameters are: ϕ, A, β, and σ_w^2. Define the quantities,

$$\begin{aligned} k^* &= \mathrm{argmax}_{k=0,1,\cdots,N/2} \left|\sum_{i=0}^{N-1} x(i)\,\mathrm{e}^{-\mathrm{j}\frac{2\pi}{N}k\,i}\right| \\ D_{k^*} &= \sum_{i=0}^{N-1} x(i)\,\mathrm{e}^{-\mathrm{j}\frac{2\pi}{N}k^*\,i}. \end{aligned} \tag{63}$$

The corresponding ML estimates are then computed as,

$$\hat{\beta}_{\mathrm{ML}} = \frac{2\pi}{N} k^*, \quad \hat{A}_{\mathrm{ML}} = \frac{2}{N} |D_{k^*}|,$$
$$\hat{\phi}_{\mathrm{ML}} = \angle D_{k^*}, \quad \hat{\sigma^2_w}_{\mathrm{ML}} = \frac{1}{N} \sum_{i=1}^{N} x^2(i) - \frac{2}{N^2} |D_{k^*}|^2. \tag{64}$$

■

Example 12. *As shown in Figure 8, the distance from an object at unknown locations, P_1, and, P_2, is measured at multiple spatial locations, $x_i = (i-1)\Delta$, $i = 1,2,\cdots,N$. Find the LS estimate of the object location.*

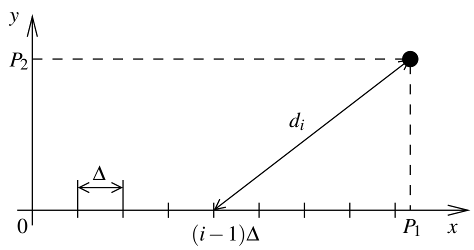

Figure 8. Determining the object location from multiple distance measurements.

Solution: 12. Denote the distances,

$$d_i = g_i(P_1, P_2) = \sqrt{(P_1 - x_i)^2 + P_2^2}. \tag{65}$$

The estimate, $[\hat{P}_1, \hat{P}_2]$, of the object location, $[P_1, P_2]$, is given by numerically solving the following set of non-linear equations:

$$\frac{1}{N} \sum_{i=1}^{N} \frac{d_i}{g_i(\hat{P}_1, \hat{P}_2)} = 1$$
$$\frac{2}{N(N-1)} \sum_{i=1}^{N} \frac{(i-1) d_i}{g_i(\hat{P}_1, \hat{P}_2)} = 1. \tag{66}$$

The initial location estimate can be computed as,

$$\hat{P}_1 = \frac{d_1^2 - d_N^2 + (N-1)^2 \Delta^2}{2(N-1)\Delta}$$
$$\hat{P}_2 = \pm \sqrt{d_1^2 - \hat{P}_1^2}. \tag{67}$$

■

Example 13. *A non-random discrete time signal, $p(i) = A\sin(\beta i + \phi)$, $i = 0,1,2,\ldots$, is measured in a zero-mean AWGN with unknown variance, σ_w^2. The frequency, β, is assumed to be known, whereas the amplitude, A, and the phase, ϕ, are unknown deterministic quantities. Find the LMMSE estimator (filter) to suppress the measurement noise at the current time instant, $i = n$. Then, simplify the estimator, provided that, $|\sin(n\beta)/\sin(\beta)| \ll 1$.*

Solution: 13. The exact signal estimate and its variance, respectively, can be derived to be,

$$\hat{p}(n) = \frac{1}{2g(n)\sin(\beta)} \times$$
$$\sum_{i=0}^{n} x(i) \left(n\sin(\beta)\cos((n-i)\beta) - \sin(n\beta)\cos((i+1)\beta) \right)$$
$$S(n) = \frac{\sigma_w^2}{2g(n)} \left(n + \frac{1}{2} - \frac{\sin((2n+1)\beta)}{2\sin(\beta)} \right) \tag{68}$$

where

$$g(n) = \frac{1}{4} \left((n+1)^2 - \left(\frac{\sin(n\beta)}{\sin(\beta)} \right)^2 - \frac{\sin((2n+1)\beta)}{\sin(\beta)} \right). \tag{69}$$

If the condition, $|\sin(n\beta)/\sin(\beta)| \ll 1$, is valid, the estimator and its variance can be approximated as,

$$\hat{p}(n) \approx \frac{2}{n+1} \sum_{i=0}^{n} x(i) \cos((n-i)\beta)$$
$$S(n) \approx \frac{2}{n+1} \sigma_w^2. \tag{70}$$

■

Example 14. *A zero-mean discrete-time Gaussian random signal, $v(n)$, has the auto-covariance, $K_v(n) = \sigma_v^2 a^{|n|}$, $a > 0$. The signal, $v(n)$, is observed through a non-linear memoryless circuit having the output signal, $x(n) = \exp(kv(n))$, $k > 0$. Find the MMSE estimate, $\hat{x}(n)$, from the past samples, $x(i)$, $i = 1,2,\cdots,n-1$. Note that the random signal, $v(n)$, can be generated as,*

$$v(n) = a\, v(n-1) + \sigma_v \sqrt{1-a^2}\, w(n) \tag{71}$$

where $v(1) = \sigma_v w(1)$, and $w(i)$ is the sample of a zero-mean, unit-variance AWGN.

Solution: 14. The extrapolated value, $x(n)$, is estimated as:

$$\hat{x}(n) = (x(n-1))^a \exp\left(\frac{1}{2} k^2 \sigma_v^2 (1-a^2) \right). \tag{72}$$

The variance of this estimator (predictor) can be found to be,

$$S = \exp(2k^2\sigma_v^2) - \exp(k^2\sigma_v^2(1+a^2)). \tag{73}$$

■

VII. Discussion

Estimation theory has been established decades ago. It is now the standard part of the undergraduate and graduate curricula in most engineering schools. It is then not surprising that many textbooks are available [1]–[17]. For example, good explanations of various topics in estimation theory at the intermediate level are provided in [7]. It should be noted that only textbooks are provided in the list of references. The survey of research papers and the state-of-the art are beyond the scope of this tutorial, which solely focuses on the fundamental principles of the parameter estimation.

Moreover, this tutorial could not cover many other important topics in parameter estimation. For instance, adaptive estimators estimate the values of multiple parameters successively rather than jointly in order to reduce the complexity. Bayesian inference relies on simple Bayes theorem; however, in practice, the underlying distributions cannot be obtained in closed-form, have many dimensions, or are only known up to a scaling constant. This requires to use sophisticated numerical algorithms involving sequential sampling, or approximations. Furthermore, estimating and predicting the values of time-dependent parameters is the subject of statistical filtering. It involves designing, e.g., Kalman filters and its variants, particle filters, and others. Importantly, signal estimation follows the same fundamental ideas of parameter estimation.

Computer simulations often perform implicit parameter estimations. It would be useful to consider the underlying parameter estimators explicitly as the components of simulations. This may be straightforward for point estimators, and it is more challenging for estimating, e.g., posterior distributions.

Unlike statistical inferences, the causal inferences are still a relatively new topic, which is not always included in the engineering curricula. Therefore, the textbooks on causal inferences are also a few and more recent [18]–[22]. A common strategy for performing causal inferences is to exploit the parameter inferences. In turn, the causal inference could enhance the parameter estimation methods.

Mathematical derivations rather than intuitive designs are often necessary to obtain the estimators, especially when the measurements are very noisy. Mathematical tractability for non-linear models can be achieved by assuming function linearization and other types of approximation. The estimator derivation translates the estimation problem into the corresponding optimization problem, and a procedure how to solve the optimization problem. This appears to be in a sharp contrast with nowadays nearly ubiquitous use of machine learning algorithms. These algorithms offer the solutions that are more flexible, but their design is based on intuition and extensive computer-based experimentation while avoiding complicated mathematical derivations altogether. Moreover, these algorithms can be implemented with a few lines of the Python code. The caveat is that the universal models used in supervised machine learning require large amounts of training data, they ignore excessive computational complexity, and their intuitive and experimental design often completely obscure their interpretability. This may explain why there are many computing libraries for machine learning, but only a few for parameter estimation.

Thus, having specialized, interpretable, but computationally efficient model-based estimators on one hand, and the universal, but inefficient model-free machine learning algorithms lacking the interpretability on the other hand indicates that there is a need to bring the principles of estimation theory into machine learning practice. For example, suppressing the measurement noise and adopting the model constraints by the means of estimation theory should greatly aid the machine learning to be either faster, or requiring less training data.

VIII. CONCLUSION

The choice of the appropriate estimator in a given signal processing scenario is completely dependent on what information is available. In particular, the model of measurements and of signals must be known, so that the statistical description of measurements and parameters to be estimated can be obtained in full, or at least partially. For instance, if the prior distribution of parameters is not known, these parameters are considered to be non-random, and their estimator may be much more difficult to find, or may not even exist. The lack of model knowledge can be replaced by the input-output samples as in the supervised machine learning. Another important consideration is how noisy the measurements are.

ACKNOWLEDGMENT

This tutorial is freely based on the undergraduate textbook [23] by Prof. Zdeněk Hrdina. E-mail: hrd@email.cz

REFERENCES

[1] R. C. Aster, B. Borchers, and C. H. Thurber, *Parameter Estimation And Inverse Problems*, 3rd ed. Elsevier, Amsterdam, Netherlands, 2019.
[2] M. Barkat, *Signal Detection and Estimation*, 2nd ed. Artech House, Norwood, MA, USA, 2005.
[3] D. D. Boos and L. Stefanski, *Essential Statistical Inference: Theory and Methods*. Spring, New York, USA, 2013.
[4] G. Casella and R. L. Berger, *Statistical Inference*, 2nd ed. Duxbury, Thomson Learning, Pacific Grove, CA, USA, 2002.
[5] C. W. Helstrom, *Elements of Signal Detection and Estimation*. Prentice Hall, Inc, Englewood Cliffs, NJ, USA, 1995.
[6] R. V. Hogg, E. A. Tanis, and D. L. Zimmerman, *Probability Statistical Inference*, 9th ed. Pearson, Upper Saddle River, NJ, USA, 2015.
[7] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Prentice Hall, Upper Saddle River, NJ, USA, 1993, vol. I.
[8] B. L. et al., *Classification, Parameter Estimation and State Estimation, An Engineering Approach Using MATLAB*, 2nd ed. John Wiley & Sons, Ltd, Hoboken, NJ, USA, 2017.
[9] H. Liero and S. Zwanzig, *Introduction to the Theory of Statistical Inference*. CRC Press, Boca Raton, FL, USA, 2012.
[10] H. S. Migon, D. Gamerman, and F. Louzada, *Statistical Inference, An Integrated Approach*, 2nd ed. CRC Press, Boca Raton, FL, USA, 2015.
[11] D. J. Olive, *Statistical Theory and Inference*. Springer Int. Publishing, Switzerland, 2014.
[12] M. J. Panik, *Statistical Inference, A Short Course*. John Wiley & Sons, Inc., Hoboken, NJ, USA, 2012.
[13] E. Pitman, *Some Basic Theory for Statistical Inference*. CRC Press, Boca Raton, FL, USA, 2018.
[14] H. V. Poor, *An Introduction to Signal Detection and Estimation*, 2nd ed. Springer-Verlag, New York, USA, 1994.
[15] C. A. Rohde, *Introductory Statistical Inference with the Likelihood Function*. Springer Int. Publishing, Switzerland, 2014.
[16] G. G. Roussas, *An Introduction to Probability and Statistical Inference*, 2nd ed. Elsevier, Amsterdam, Netherlands, 2015.
[17] J. Thijssen, *A Concise Introduction to Statistical Inference*. CRC Press, Boca Raton, FL, USA, 2014.
[18] G. W. Imbens and D. B. Rubin, *Causal Inference for Statistics, Social, and Biomedical Sciences, An Introduction*. Cambridge University Press, NY, USA, 2015.
[19] J. Pearl, *Causality, Models, Reasoning, and Inference*, 2nd ed. Cambridge University Press, New York, NY, USA, 2009.
[20] J. Pearl, M. Glymour, and N. P. Jewell, *Causal Inference In Statistics, A Primer*. John Wiley & Sons, Ltd, Hoboken, NJ, USA, 2016.
[21] J. Peters, D. Janzing, and B. Schölkopf, *Elements of Causal Inference, Foundations and Learning Algorithms*. The MIT Press, Cambridge, MA, USA, 2017.
[22] P. R. Rosenbaum, *Observation and Experiment, An Introduction to Causal Inference*. Harward Univ. Press, Cambridge, MA, USA, 2017.
[23] Z. Hrdina, *Statistical Radio-Engineering*. Publishing Company of the Czech Technical University of Prague, 1996, in Czech.