Estimation Theory

LIN Qingfeng
March 1, 2021

1 Introduction

- There are two kinds of estimations problem. One is **Classical estimation**, which is based on PDFs. The parameters in classical estimation are assumed to be deterministic but unknown.The other approach is **Bayesian estimation** in that the parameters of interest are assumed to be a random variable θ.
- The problem of estimation theory is to find a function of the N -point data set which provides an estimate of θ, that is:

$$\hat{\theta} = g(\mathbf{x} = \{x[0], x[1], \ldots, x[N-1]\}) \tag{1}$$

  where $\hat{\theta}$ is an estimate of θ, and $g(\mathbf{x})$ is known as the estimator function.

  Once a candidate $g(\mathbf{x})$ is found, then we usually ask:
  (1). How close will $\hat{\theta}$ be to θ (i.e. how good or optimal is our estimator)?
  (2). Are there better (i.e. closer) estimators?

- A natural optimal criterion is minimisation of the mean square error:

$$\operatorname{mse}(\hat{\theta}) = E\left\{[(\hat{\theta} - E(\hat{\theta})) + (E(\hat{\theta}) - \theta)]^2\right\} = \operatorname{var}(\hat{\theta}) + [E(\hat{\theta}) - \theta]^2 \tag{2}$$

  However although $[E(\hat{\theta}) - \theta]^2$ is a function of θ the variance of the estimator, $\operatorname{var}(\hat{\theta})$, is only a function of data. Thus an alternative approach is to assume $E(\hat{\theta}) - \theta = 0$ and minimise $\operatorname{var}(\hat{\theta})$. This produces the Minimum Variance Unbiased (MVU) estimator.

- **MVU Estimator**: Estimator is unbiased, that is,

$$E(\hat{\theta}) = \theta \text{ for } a < \theta < b \tag{3}$$

  Also, Estimator should have the minimum variance,

$$\hat{\theta}_{MVU} = \arg\min_{\hat{\theta}}\{\operatorname{var}(\hat{\theta})\} = \arg\min_{\hat{\theta}}\left\{E(\hat{\theta} - E(\hat{\theta}))^2\right\} \tag{4}$$

2 The Approaches of Estimator Selection

2.1 Cramer-Rao Lower Bound (CRLB)

- **Data Model**: PDF is known.
- **Estimator**: If the equality condition for the CRLB

$$\frac{\partial \ln p(\mathbf{x};\theta)}{\partial \boldsymbol{\theta}} = \mathbf{I}(\boldsymbol{\theta})(\mathbf{g}(\mathbf{x}) - \boldsymbol{\theta}) \tag{5}$$

  is satisfied, then the estimator is

$$\hat{\boldsymbol{\theta}} = \mathbf{g}(\mathbf{x}) \tag{6}$$

- **Optimality**: $\hat{\boldsymbol{\theta}}$ achieves the CRLB, the lower bound on the variance for any unbiased estimator, and is therefore the minimum variance unbiased (MVU) estimator.
- **Comments**: An efficient estimator may not exist, and hence this approach may fail.
- **Example**: Consider the multiple observations,

$$x[n] = A + w[n] \quad n = 0, 1, \ldots, N-1 \tag{7}$$

To determine the CRLB for A,

$$p(\mathbf{x}; A) = \prod_{n=0}^{N-1} \frac{1}{\sqrt{2\pi\sigma^2}} exp\left[-\frac{1}{2\sigma^2}(x[n]-A)^2\right] \tag{8}$$

Taking the first derivative,

$$\begin{aligned} \frac{\partial \ln p(\mathbf{x}; A)}{\partial A} &= \frac{\partial}{\partial A}\left[-\ln\left[\left(2\pi\sigma^2\right)^{\frac{N}{2}}\right] - \frac{1}{2\sigma^2}\sum_{n=0}^{N-1}(x[n]-A)^2\right] \\ &= \frac{1}{\sigma^2}\sum_{n=0}^{N-1}(x[n]-A) \\ &= \frac{N}{\sigma^2}(\bar{x}-A) \end{aligned} \tag{9}$$

We can easily get,

$$\mathbf{I}(\boldsymbol{\theta}) = \frac{N}{\sigma^2} \tag{10}$$

$$\mathbf{g}(\mathbf{x}) = \bar{x} \tag{11}$$

2.2 Rao-Blackwell-Lehmann-Scheffe

- **Motivation**: The evaluation of CRLB sometimes results in a MVU estimator. However, an efficient estimator does not exist, it is still be able to find the MVU estimator. To do so requires the concept of sufficient statistics and important Rao-Blackwell-Lehmann-Scheffe theorem.
- **Data Model**: PDF is known.
- **Estimator**: **i**) Find a sufficient statistic $\mathbf{T}(\mathbf{x})$ by factoring PDF as

$$p(\mathbf{x}; \boldsymbol{\theta}) = g(\mathbf{T}(\mathbf{x}), \boldsymbol{\theta})h(\mathbf{x}) \tag{12}$$

ii) if $E[\mathbf{T}(\mathbf{x})] = \boldsymbol{\theta}$, then $\hat{\boldsymbol{\theta}} = \mathbf{T}(\mathbf{x})$.

- **Optimality**: $\hat{\boldsymbol{\theta}}$ is the MVU estimator.
- **Comments**: Completeness of sufficient statistic must be checked. If it doesn't exist, this method may fail.
- **Examples**: Same as the example in CRLB,

$$p(\mathbf{x}; A) = \underbrace{\frac{1}{\left(2\pi\sigma^2\right)^{\frac{N}{2}}} exp\left[-\frac{1}{2\sigma^2}\left(NA^2 - 2A\sum_{n=0}^{N-1}x[n]\right)\right]}_{g(T(\mathbf{x}),A)} \underbrace{exp\left[-\frac{1}{2\sigma^2}\sum_{n=0}^{N-1}x^2[n]\right]}_{h(\mathbf{x})} \tag{13}$$

$$T(\mathbf{x}) = \sum_{n=0}^{N-1} x[n] \tag{14}$$

2.3 Best Linear Unbiased Estimator (BLUE)

- **Motivation**: It frequently occurs in practice that the MVU estimator, if it exists, cannot be found. For instance, we may not know the PDF of the data or even be willing to assume a model for it. Thus, it is reasonable to resort to a suboptimal estimator. A common approach is to restrict the estimator to be linear in the data and find the linear estimator that is unbiased and has minimum variance.
- **Data Model**:

  $$\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w} \tag{15}$$

  where $E(\mathbf{w}) = \mathbf{0}$ and $\mathbf{C}_w = \mathbf{C}$.
- **Estimator**:

  $$\hat{\boldsymbol{\theta}} = \left(\mathbf{H}^T\mathbf{C}^{-1}\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{C}^{-1}\mathbf{x} \tag{16}$$

- **Optimality**: $\hat{\boldsymbol{\theta}}$ has the minimum variance of all unbiased estimators that are linear in $\mathbf{x}$.
- **Comments**: If $\mathbf{w}$ is a Gaussian random vector so that $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{C})$, then $\hat{\boldsymbol{\theta}}$ is also the MVU estimator (for all linear or nonlinear functions of $\mathbf{x}$). Unfortunately, without knowledge of the PDF there is no way to determine the loss in performance by resorting to a BLUE.
- **Examples**: Same as above, if we observe,

  $$x[n] = A + w[n] \quad n = 0, 1, \ldots, N-1 \tag{17}$$

  where $w[n]$ is white noise with variance σ^2, then the problem is to estimate A.

  $$\begin{aligned} \hat{A} &= \frac{1^T\frac{1}{\sigma^2}1\mathbf{x}}{1^T\frac{1}{\sigma^2}1} \\ &= \frac{1}{N}\sum_{n=0}^{N-1} x[n] = \bar{x} \end{aligned} \tag{18}$$

  $$\begin{aligned} \operatorname{var}(\hat{A}) &= \frac{1}{1^T\frac{1}{\sigma^2}1} \\ &= \frac{\sigma^2}{N} \end{aligned} \tag{19}$$

2.4 Maximum Likelihood Estimator (MLE)

- **Motivation**: Sometimes, the MVU estimator does not exit or can not be found even if it does exist. We need investigate an alternative to the MVU estimator, which is based on the maximum likelihood principle.For most cases of practical interest its performance is optimal for large data records. Also, it is approximately the MVU estimator due to its approximate efficiency.
- **Data Model**: PDF is known.
- **Estimator**: $\hat{\boldsymbol{\theta}}$ is the value of $\boldsymbol{\theta}$ maximizing $p(\mathbf{x}; \boldsymbol{\theta})$, where $\mathbf{x}$ is replaced by the observed data samples.
- **Optimality**: Not optimal in general. Under certain conditions on the PDF, however, the MLE is efficient for large data records. Hence, asymptotically it is the MVU estimator.

- **Comments**: If an MVU estimator exists, the maximum likelihood procedure will produce it.
  Consider the general linear model of the form:

$$\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w} \tag{20}$$

  where $\mathbf{H}$ is a known $N \times p$ matrix, $\mathbf{x}$ is the $N \times 1$ observation vector with N samples, and $\mathbf{w}$ is a noise vector of dimension $N \times 1$ with $PDF \mathbf{N}(\mathbf{0}, \mathbf{C})$. The PDF is:

$$p(\mathbf{x}; \boldsymbol{\theta}) = \frac{1}{(2\pi)^{\frac{N}{2}} det^{\frac{1}{2}}(\mathbf{C})} exp\left[-\frac{1}{2}(\mathbf{x} - \mathbf{H}\boldsymbol{\theta})^T \mathbf{C}^{-1}(\mathbf{x} - \mathbf{H}\boldsymbol{\theta})\right] \tag{21}$$

  and the MLE of $\boldsymbol{\theta}$ is found by differentiating the log-likelihood which can be shown to vield:

$$\frac{\partial \ln p(\mathbf{x}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \frac{\partial (\mathbf{H}\boldsymbol{\theta})^T}{\partial \boldsymbol{\theta}} \mathbf{C}^{-1}(\mathbf{x} - \mathbf{H}\boldsymbol{\theta}) \tag{22}$$

  and this we obtain the MLE of $\boldsymbol{\theta}$ as:

$$\hat{\boldsymbol{\theta}} = \left(\mathbf{H}^T \mathbf{C}^{-1} \mathbf{H}\right)^{-1} \mathbf{H}^T \mathbf{C}^{-1} \mathbf{x} \tag{23}$$

  which is the same as the MVU estimator.

- **Examples**: For the received data,

$$x[n] = A + w[n] \quad n = 0, 1, \ldots, N-1 \tag{24}$$

  where A is the unknown level to be estimated and $w[n]$ is WGN with known variance σ^2, the PDF is

$$p(\mathbf{x}; A) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} exp\left[-\frac{1}{2\sigma^2} \sum_{n=0}^{N-1} (x[n] - A)^2\right] \tag{25}$$

  Taking the derivative of the log-likelihood function produces,

$$\frac{\partial \ln p(\mathbf{x}; A)}{\partial A} = \frac{1}{\sigma^2} \sum_{n=0}^{N-1} (x[n] - A) \tag{26}$$

  which being set equal to zero yields the MLE,

$$\hat{A} = \frac{1}{N} \sum_{n=0}^{N-1} x[n] \tag{27}$$

2.5 Least Squares Estimator (LSE)

- **Motivation:**
  The MVU, BLUE and MLE estimators developed previously required an expression for the PDF $p(\mathbf{x}; \theta)$ in order to estimate the unknown parameter θ in some optimal fashion. An alternative approach is to assume a signal model (rather than probabilistic assumptions about the data) and achieve a design goal assuming this model.

- **Data Model**: The model is,

$$\mathbf{x} = \mathbf{s}(\boldsymbol{\theta}) + \mathbf{w} \tag{28}$$

  where $\mathbf{s}$ is a known function of $\boldsymbol{\theta}$ and the noise $\mathbf{w}$ has zero mean.

- **Estimator**: $\hat{\boldsymbol{\theta}}$ is the value of $\boldsymbol{\theta}$ that minimizes,

$$\begin{aligned} J(\boldsymbol{\theta}) &= (\mathbf{x} - \mathbf{s}(\boldsymbol{\theta}))^T (\mathbf{x} - \mathbf{s}(\boldsymbol{\theta})) \\ &= \sum_{n=0}^{N-1} (x[n] - s[n; \boldsymbol{\theta}])^2 \end{aligned} \tag{29}$$

- **Optimality**: Not optimal in general.
- **Comments**: The fact that we are minimizing a LS error criterion does not in general translate into minimizing the estimation error. Also, if $\mathbf{w}$ is a Gaussian random vector with $\mathbf{w} \sim \mathcal{N}\left(\mathbf{0}, \sigma^2 \mathbf{I}\right)$, then the LSE is equivalent to the MLE.
- **Example**: Assume that the signal model is $s[n] = A$ and we observe $x[n]$ for $n = 0, 1, \ldots, N-1$. Then, according to the LS approach, we can estimate A by minimizing,

$$J(A) = \sum_{n=0}^{N-1} (x[n] - A)^2. \tag{30}$$

Differentiating with respect to A and setting the result equal to zero produces,

$$\begin{aligned} \hat{A} &= \frac{1}{N} \sum_{n=0}^{N-1} x[n] \\ &= \bar{x} \end{aligned} \tag{31}$$

2.6 Method of Moments

- **Motivation**: The method of moments is very easy to determine and simple to implement. Although the estimator has no optimality properties, it is useful if the data record is long enough. This is because the method of moments estimator is usually consistent.
- **Data Model**: There are p moments $\mu_i = E\left(x^i[n]\right)$ for $i = 1, 2, \ldots, p$. The entire PDF need not be known.
- **Estimator**: If $\mu = h(\theta)$, where h is an invertible function of θ and $\mu = [\mu_1 \mu_2 \ldots \mu_p]^T$ then,

$$\hat{\boldsymbol{\theta}} = \mathbf{h}^{-1}(\hat{\boldsymbol{\mu}}) \tag{32}$$

- **Optimality**: Not optimal in general.
- **Comments**: It is usually very easy to implement.
- **Example**: If $x[n] = A + w[n]$ is observed for $n = 0, 1, \ldots, N-1$, where $w[n]$ is WGN with variance σ^2 and A is to be estimated, then we know that,

$$\mu_1 = E(x[n]) = A \tag{33}$$

$$\hat{A} = \frac{1}{N} \sum_{n=0}^{N-1} x[n] \tag{34}$$

2.7 Minimum Mean Square Error (MMSE) Estimation

- **Motivation:** The classic approach we have been using so far has assumed that the parameter θ is unknown but deterministic. Thus the optimal estimator $\hat{\theta}$ is optimal irrespective and independent of the actual value of θ.

  In the Bayesian philosophy the θ is treated as a random variable with a known prior $pdf, p(\theta)$. Such prior knowledge concerning the distribution of the estimator should provide better estimators than the deterministic case.

- **Data Model**:The joint PDF of $\mathbf{x}, \boldsymbol{\theta}$ or $p(\mathbf{x}, \boldsymbol{\theta})$ is known, where $\boldsymbol{\theta}$ is now considered to be a random vector. Usually, $p(\mathbf{x}|\boldsymbol{\theta})$ is specified as the data model and $p(\boldsymbol{\theta})$ as the prior PDF for $\boldsymbol{\theta}$, so that $p(\mathbf{x}, \boldsymbol{\theta}) = p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})$

- **Estimator**:

  $$\hat{\boldsymbol{\theta}} = E(\boldsymbol{\theta}|\mathbf{x}) \tag{35}$$

  where the expectation is with respect to the posterior PDF

  $$p(\boldsymbol{\theta}|\mathbf{x}) = \frac{p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}} \tag{36}$$

- **Optimality**: Not optimal in general.

- **Comments**: In the non-Gaussian case, this will be very difficult to implement.

- **Example:**
  The model is,

  $$x[n] = A + w[n] \tag{37}$$

  where as before $w[n] = N\left(0, \sigma^2\right)$ is a WGN process and the unknown parameter $\theta = A$ is to be estimated. However in the Bayesian approach we also assume the parameter A is a random variable with a prior pdf which in this case is the Gaussian pdf $p(A) = N\left(\mu_A, \sigma_A^2\right)$. We also have that $p(\mathbf{x} \mid A) = N\left(A, \sigma^2\right)$ and we can assume that $\mathbf{x}$ and A are jointly Gaussian. Thus the posterior pdf:

  $$p(A \mid \mathbf{x}) = \frac{p(\mathbf{x}|A)p(A)}{\int p(\mathbf{x} \mid A)p(A)dA} = N\left(\mu_{A|\mathbf{x}}, \sigma^2_{A|\mathbf{x}}\right) \tag{38}$$

  is also a Gaussian pdf and after the required simplification we have that:

  $$\sigma^2_{A|\mathbf{x}} = \frac{1}{\frac{N}{\sigma^2} + \frac{1}{\sigma_A^2}} \quad \text{and} \quad \mu_{A|\mathbf{x}} = \left(\frac{N}{\sigma^2}\bar{x} + \frac{\mu_A}{\sigma_A^2}\right)\sigma^2_{A|\mathbf{x}} \tag{39}$$

  and hence the MMSE is:

  $$\begin{aligned} \hat{A} = E[A|\mathbf{x}] &= \int A p(A|\mathbf{x})dA = \mu_{A|\mathbf{x}} \\ &= \alpha\bar{x} + (1-\alpha)\mu_A \end{aligned} \tag{40}$$

- **Relation with Classic Estimation**:
  In classical estimation we cannot make any assumptions on the prior, thus all possible θ have to be considered. The equivalent prior pdf would be a flat distribution, essentially $\sigma_\theta^2 = \infty$. This so-called noninformative prior pdf will yield the classic estimator where such is defined.

2.8 Maximum A Posteriori (MAP) Estimator

- **Data Model**: Same as for the MMSE estiamtor.
- **Estimator**: $\hat{\boldsymbol{\theta}}$ is the value of $\boldsymbol{\theta}$ that maximizes $p(\boldsymbol{\theta}|\mathbf{x})$ or, equivalently, the value that maximizes $p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})$.
- **Optimality**: Not optimal in general.
- **Comments**: For PDFs whose mean and mode (the location of the maximum) are the same. the MMSE and MAP estimators will be identical, i.e., the Gaussian PDF, for example.
- **Examples**:

  Assume that,

$$p(x[n]|\theta) = \begin{cases} \theta exp(-\theta x[n]) & x[n] > 0 \\ 0 & x[n] < 0 \end{cases} \tag{41}$$

  where the $x[n]$'s are conditionally IID, or

$$p(\mathbf{x} \mid \theta) = \prod_{n=0}^{N-1} p(x[n] \mid \theta) \tag{42}$$

  and the prior PDF is,

$$p(\theta) = \begin{cases} \lambda exp(-\lambda\theta) & \theta > 0 \\ 0 & \theta < 0 \end{cases} \tag{43}$$

  Then, the MAP estimator is found by maximizing

$$\begin{aligned} g(\theta) &= \ln p(\mathbf{x} \mid \theta) + \ln p(\theta) \\ &= \ln\left[\theta^N exp\left(-\theta\sum_{n=0}^{N-1} x[n]\right)\right] + \ln[\lambda exp(-\lambda\theta)] \\ &= N\ln\theta - N\theta\bar{x} + \ln\lambda - \lambda\theta \end{aligned} \tag{44}$$

  for $\theta > 0$. Differentiating with respect to θ produces

$$\frac{dg(\theta)}{d\theta} = \frac{N}{\theta} - N\bar{x} - \lambda \tag{45}$$

  and setting it equal to zero yields the MAP estimator

$$\hat{\theta} = \frac{1}{\bar{x} + \frac{\lambda}{N}} \tag{46}$$

2.9 Linear Minimum Mean Square Error (LMMSE) Estimator

- **Motivation:**
  The optimal Bayesian estimators discussed previously are difficult to determine in closed form, and in practice too computationally intensive to implement. They involve multidimensional integration for the MMSE estimator and multidimensional maximization for the MAP estimator. Although under the jointly Gaussian assumption these estimators are easily found, in general, they are not.

  When we are unable to make the Gaussian assumption, another approach must be used. To fill this gap we can choose to retain the MMSE criterion but constrain the estimator to be linear. Then, an explicit form for the estimator may be determined which depends only on the first two moments of the PDF.

- **Data Model**: The first two moments of the joint PDF $p(\mathbf{x}, \boldsymbol{\theta})$ are know or the mean and covariance,

$$\begin{bmatrix} E(\boldsymbol{\theta}) \\ E(\mathbf{x}) \end{bmatrix} \quad \begin{bmatrix} \mathbf{C}_{\theta\theta} & \mathbf{C}_{\theta x} \\ \mathbf{C}_{x\theta} & \mathbf{C}_{xx} \end{bmatrix} \tag{47}$$

- **Estimator**:

$$\hat{\boldsymbol{\theta}} = E(\boldsymbol{\theta}) + \mathbf{C}_{\theta x}\mathbf{C}_{xx}^{-1}(\mathbf{x} - E(\mathbf{x})) \tag{48}$$

- **Optimality**: $\hat{\theta}_i$ has the minimum Bayesian MSE of all estimators that are linear functions of x.
- **Comments**: If $\mathbf{x}, \boldsymbol{\theta}$ are jointly Gaussian, this is identical to the MMSE and MAP estimators.
- **Example:**
  The model is,

$$x[n] = A + w[n] \quad n = 0, 1, \dots, N-1 \tag{49}$$

where $A \sim \mathcal{U}\left[-A_0, A_0\right], w[n]$ is WGN with variance σ^2, and A and $w[n]$ are independent. to the integration required. Applying the LMMSE estimator, we first note that $E(A) = 0$, and hence $E(x[n]) = 0$. Since $E(x) = 0$, the covariances are

$$\begin{aligned} \mathbf{C}_{xx} &= E\left(\mathbf{x}\mathbf{x}^T\right) \\ &= E\left[(A\mathbf{1} + \mathbf{w})(A\mathbf{1} + \mathbf{w})^T\right] \\ &= E\left(A^2\right)\mathbf{1}\mathbf{1}^T + \sigma^2\mathbf{I} \end{aligned} \tag{50}$$

$$\begin{aligned} \mathbf{C}_{\theta x} &= E\left(A\mathbf{x}^T\right) \\ &= E\left[A(A1 + \mathbf{w})^T\right] \\ &= E\left(A^2\right)\mathbf{1}^T \end{aligned} \tag{51}$$

where 1 is an $N \times 1$ vector of all ones. Hence,

$$\begin{aligned} \hat{A} &= \mathbf{C}_{\theta x}\mathbf{C}_{xx}^{-1}\mathbf{x} \\ &= \sigma_A^2 1^T\left(\sigma_A^2\mathbf{1}\mathbf{1}^T + \sigma^2\mathbf{I}\right)^{-1}\mathbf{x} \end{aligned} \tag{52}$$

The LMMSE estimator of A is

$$\hat{A} = \frac{\frac{A_0^2}{3}}{\frac{A_0^2}{3} + \frac{\sigma^2}{N}}\bar{x} \tag{53}$$

3 Application

3.1 Expectation-Maximization Algorithm

- Motivation:
- Consider a set of data points with their classes labeled, and assume that each class is a Gaussian. Given this set of data points, finding the means of the Gaussian can be done easily by estimating the sample mean, as the class labels are known
- However, when the classes are not labeled, we could use an iterative approach: first make a guess of the class label for each data point, then compute the means and update the guess of the class labels again. We repeat until the means converge.

- **The problem of estimating parameters in the absence of labels is known as unsupervised learning.** The Expectation Maximization (EM) algorithm is a kind of unsupervised learning.

3.2 Curve Fitting Problem

- Consider fitting the data, $x(t)$, by a p^{th} order polynomial function of t

$$x(t) = \theta_0 + \theta_1 t + \theta_2 t^2 + \cdots + \theta_p t^p + w(t) \tag{54}$$

Solution: We use the Linear Model in estimation theory.

Say we have N samples of data, then:

$$\begin{aligned} \mathbf{x} &= \left[x\left(t_0\right), x\left(t_1\right), x\left(t_2\right), \ldots, x\left(t_{N-1}\right)\right]^T \\ \mathbf{w} &= \left[w\left(t_0\right), w\left(t_1\right), w\left(t_2\right), \ldots, w\left(t_{N-1}\right)\right]^T \\ \boldsymbol{\theta} &= \left[\theta_0, \theta_1, \theta_2, \ldots \theta_p\right]^T \end{aligned} \tag{55}$$

so $\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w}$, where $\mathbf{H}$ is the $N \times p$ matrix:

Hence the MVU estimate of the polynomial coefficients based on the N samples of data is:

$$\hat{\boldsymbol{\theta}} = \left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{x}$$

3.3 Least Squares

3.3.1 Linear Least Squares

- We assume the signal model is a linear function of the estimator, that is:

$$\mathbf{s} = \mathbf{H}\boldsymbol{\theta} \tag{56}$$

where $\mathbf{s} = [s[0], s[1], \ldots, s[N-1]]^T$ and $\mathbf{H}$ is a known $N \times p$ matrix with $\boldsymbol{\theta} = [\theta_1, \theta_2, \ldots, \theta_p]$.
Now:

$$\mathbf{x} = \mathbf{H}\boldsymbol{\theta} + \mathbf{w} \tag{57}$$

and with $\mathbf{x} = [x[0], x[1], \ldots, x[N-1]]^T$ we have:

$$J(\boldsymbol{\theta}) = \sum_{n=0}^{N-1}(x[n] - s[n])^2 = (\mathbf{x} - \mathbf{H}\boldsymbol{\theta})^T(\mathbf{x} - \mathbf{H}\boldsymbol{\theta}) \tag{58}$$

Differentiating and setting to zero:

$$\hat{\boldsymbol{\theta}} = \left(\mathbf{H}^T\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{x} \tag{59}$$

An interesting extension to the linear LS is the weighted LS where the contribution to the error from each component of the parameter vector can be weighted in importance by using a different from of the error criterion:

$$J(\boldsymbol{\theta}) = (\mathbf{x} - \mathbf{H}\boldsymbol{\theta})^T\mathbf{W}(\mathbf{x} - \mathbf{H}\boldsymbol{\theta}) \tag{60}$$

where $\mathbf{W}$ is an $N \times N$ postive definite (symmetric) weighting matrix.

3.3.2 Linear Least Squares

- In many cases the signal model is unknown and must be assumed. Obviously we would like to choose the model, $s(\boldsymbol{\theta})$, that minimises J_{min}, that is:

$$s_{\text{best}}\left(\boldsymbol{\theta}\right)=\arg\min_{s(\hat{\boldsymbol{\theta}})} J_{\min} \tag{61}$$

We can do this arbitrarily by simply choosing models, obtaining the LSE $\hat{\theta}$, and then selecting the model which provides the smallest $J_{\min}$. However, models are not arbitrary and some models are more ”complex” (or more precisely have a larger number of parameters or degrees of freedom) than others. The more complex a model the lower the $J_{\min}$ one can expect but also the more likely the model is to overfit the data or be overtrained (i.e. fit the noise and not generalise to other data sets.

An alternative to providing an independent LSE for each possible signal model a more efficient order-recursive LSE is possible if the models are different orders of the same base model (e.g. polynomials of different degree). In this method the LSE is updated in order (of increasing parameters). Specifially define $\hat{\boldsymbol{\theta}}_k$ as the LSE of order k (i.e. k parameters to be estimated). Then for a linear model we can derive the order-recursive LSE as:

$$\hat{\boldsymbol{\theta}}_{k+1}=\hat{\boldsymbol{\theta}}_k+\text{UPDATE}_k \tag{62}$$

3.3.3 Sequential Least Squares

- In most signal processing applications the observations samples arrive as a stream of data. All our estimation strategies have assumed a batch or block mode of processing whereby we wait for N samples to arrive and then form our estimate based on these samples.
- One problem is the delay in waiting for the N samples before we produce our estimate, another problem is that as more data arrives we have to repeat the calculations on the larger blocks of data (N increases as more data arrives). The latter not only implies a growing computational burden but also the fact that we have to buffer all the data we have seen, both will grow linearly with the number of samples we have. Since in signal processing applications samples arise from sampling a continuous process our computational and memory burden will grow linearly with time!
- One solution is to use a sequential mode of processing where the parameter estimate for n samples, $\hat{\theta}[n]$, is derived from the previous parameter estimate for $n-1$ samples, $\hat{\theta}[n-1]$. For linear models we can represent sequential LSE as

$$\hat{\theta}[n]=\hat{\theta}[n-1]+K[n](x[n]-s[n|n-1]) \tag{63}$$

where $s[n \mid n-1] \equiv s(n;\theta[n-1])$. The $K[n]$ is the correction gain and $(x[n]- s[n \mid n-1])$ is the prediction error.