Advanced Signal Processing Seminar - Estimation Theory 2

Vedran Dizdarević

Institute for Signal Processing and Speech Communication
Technical University Graz, Inffeldgasse 12, A-8010 Graz
email: v.dizdarevic@TUGraz.at

Abstract*—This report is a complement to an oral presentation on the topic of estimation theory held at the Advanced Signal Processing Seminar. First part of the presentation reviewed fundamental theory of signal estimation, definition of minimum variance unbiased estimator and the Cramer-Rao Lower Bound. This part deals with three classes of estimator structures. Linear estimators, ML (maximum likelihood) estimators and LS (least squares) estimators. For these three classes firstly theoretical background will be presented along with some fundamental properties. This will include the discription of ways to incorporate prior knowledge about the problem, optimality and performance criteria and calculation of estimated parameter values. Problems arising from the use of specific estimators will be addressed as well. Finally examples should help understand the theory. The main reference for this work is [1]. All theoretical derivation, if not otherwise stated, refer to this work.

1 Introduction

Problem of parameter estimation is determining an unknown value which in some way drives a process being observed. In a general case we will not be able to observe true values of the signal. Modeling of system and measurement imperfections will lead to the assumption that we have noise embedded in the signal. Two ways to incorporate a priori knowledge exist. On a statistical side we might assume that the observation is generated according to a known pdf with unknown parameters. In a case of a normal distribution this value might be the mean, the variance or both. This concept applies to arbitrary distributions. We will denote a class of known pdfs with unknown parameter θ as $p(\mathbf{x};\theta)$. If we have no statistical assumptions of the signal, we might turn to a deterministic characterization in terms of a known data model. Again one or more parameters driving the model are to be estimated. The signal observed has a functional dependence $s(\theta)$. Generally a quantity to be estimated will be a parameter vector θ. This of course includes the case, where only one-dimensional value θ is to be estimated.

Generally we will try to formulate an optimality and performance criterion for all estimator classes. However this will not always be possible. Definition of estimation optimality is closely related to the concept of CRLB. If the estimator is shown to attain the bound we define it to be optimal. Related but somewhat weaker is the performance criterion. It will include a definition of a variance of parameter vector that an estimator is producing. Optimally this will under some conditions converge towards the CRLB.

Mostly the noise process $\mathbf{w}$ will be assumed to be a zero-mean white noise. In some cases this will be a necessary part of our derivation. Sometimes however it is possible to generalise the concept and define a noise covariance structure $\mathbf{C_w}$. For some estimator classes this will allow modeling of arbitrary noise processes.

2 Linear Estimators

The class of linear estimators is based on the assumption that we known a data model that is linear in a vector parameter θ. Starting point is a simple example where we observe two noised data points at sample instants 0 and 1. If we assume a first order polynomial, a straight line, to be our signal model, the parameters to be estimated include the intercept a and slope b (eq.1 and 2).

$$x_0 = a + 0b + w_0 \tag{1}$$

$$x_1 = a + 1b + w_1 \tag{2}$$

We can formulate this in a matrix notation as

$$\begin{pmatrix} x_0 \\ x_1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} + \begin{pmatrix} w_0 \\ w_1 \end{pmatrix} \tag{3}$$

or more generally

$$\mathbf{x} = \mathbf{H}\theta + \mathbf{w} \tag{4}$$

Eq. 4 is the most general case of a linear signal model. If prior knowledge exists that allows use of eq.4 then linear estimators can be employed to estimate the value of θ. In eq.4 $\mathbf{x}$ is the vector of observed values. $\mathbf{H}$ is called observation matrix. It is known and needs to be constructed from the input values of the system. θ is the parameter vector, values θ_i need to be estimated. In the above example this would be $\theta = [a\ b]^T$. $\mathbf{w}$ is a vector of realisations of a noise process.

The theory behind the CRLB states that a MVU estimator can be found if and only if we can find a p-dimensional function $g(\mathbf{x})$ that satisfies

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

The estimator is then given as $\hat{\theta} = \mathbf{g}(\mathbf{x})$ and the variance of the estimation is the inverse of the Fisher Information matrix $\mathbf{I}(\theta)$.

Assuming the signal model from eq.4 with a given noise covariance structure $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{C})$ it can be shown [1] that

$$\frac{\partial \ln p(\mathbf{x};\theta)}{\partial \theta} = \mathbf{H}^T\mathbf{C}^{-1}\mathbf{H}\left[(\mathbf{H}^T\mathbf{C}^{-1}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{C}^{-1}\mathbf{x} - \theta\right] \tag{6}$$

The estimator for a linear model with arbitrary but known noise is then given as

$$\hat{\theta} = (\mathbf{H}^T\mathbf{C}^{-1}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{C}^{-1}\mathbf{x} \tag{7}$$

with covariance matrix of $\hat{\theta}$ being

$$\mathbf{C}_{\hat{\theta}} = (\mathbf{H}^T\mathbf{C}^{-1}\mathbf{H})^{-1} \tag{8}$$

If $\mathbf{w}$ is white gaussian noise with $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \sigma\mathbf{I})$ eq.7 and 8 simplify to

$$\hat{\theta} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{x} \tag{9}$$

$$\mathbf{C}_{\hat{\theta}} = \sigma^2(\mathbf{H}^T\mathbf{H})^{-1} \tag{10}$$

Given that eq.5 is fulfilled important implication for this class of estimators is that they yield unbiased parameter estimates and attain the CRLB. Assuming a linear model of the data leads to an optimal estimator.

It is important to note what kind of applications can be covered using linear estimators. Linear data model states that the unknown parameter has to be in linear dependence within our data model. This definition covers far more than fitting a straight line. Approximation of polynomials of arbitrary order or estimating the weights of the tapped delay line both can be covered using linear estimation.

Problems arising in this context primarily deal with the definition of the observation matrix. Since $\mathbf{H}^T\mathbf{H}$ has to be invertable in order to calculate parameter estimates $\mathbf{H}$ has to be a full rank matrix. This is closely related to the choice of input signal and the construction of $\mathbf{H}$.

2.1 Example - Curve Fitting

Noised realisations of a second order polynomial process are observed. The goal is to estimate polynomial coefficients θ. The linear signal model is defined as

$$x(t_n) = \theta_0 + \theta_1 t_n + \theta_2 t_n^2 + w(t_n) \tag{11}$$

with $\mathbf{x} = [x(t_0)x(t_1)\cdots x(t_{99})]^T$ being the observed values and the parameter vector $\theta = [\theta_0\theta_1\theta_2]^T$. Noise variance is known to be $\sigma = 10$. Observation matrix $\mathbf{H}$ is constructed as described in eq.12. Fig. 1 shows one realization of the process and the fitted signal model.

$$\mathbf{H} = \begin{bmatrix} 1 & t_0 & t_0^2 \\ 1 & t_1 & t_1^2 \\ \vdots & \vdots & \vdots \\ 1 & t_{99} & t_{99}^2 \end{bmatrix} \tag{12}$$

After observing 5000 realisations of the noise corrupted signal, the polynomial coefficients are estimated according to eq.9, averaged over all realisations resulting in $\hat{\theta} = [1.1939\ 15.107\ -2.8003]^T$. True value of the parameter vector is $\theta = [1.2\ 15.1\ -2.8]^T$.

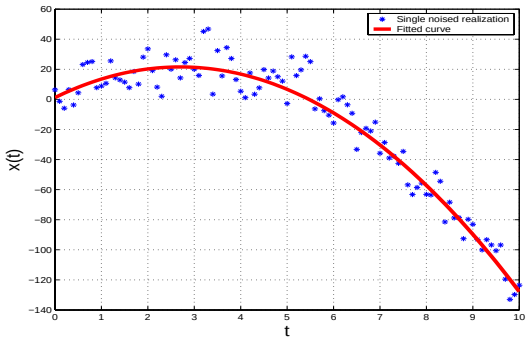

Figure 1: Second order polynomial estimation

3 Maximum Likelihood Estimators

The approach for ML estimator differs from linear modeling, since we make no deterministic assumptions of a signal model. The observed values are considered to be realisatiuons of a random process with a known probability density function (pdf). Although the analytical form of the pdf is known, its parameters are not and need to be estimated. The major principle of ML estimation is that pdf parameters are calculated which define a random process which will most likely produce the observed data. The calculation of the parameters will require finding the maximum of the likelihood function.[2] defines the likelihood function as

> The joint pdf of m sample random variables evaluated at sample points $x_1, ..., x_m$ is

$$l(\theta, x_1, ..., x_m) = l(\theta, \mathbf{x}) = f_{\mathbf{x}}(\mathbf{x} \mid \theta) = \prod_{i=1}^{m} f_{\mathbf{x}}(x_i \mid \theta) \qquad (13)$$

ML estimation yields a parameter θ for which $l(\theta, \mathbf{x})$ has a global maximum.

$$\hat{\theta} = \arg\max_{\theta} l(\theta, \mathbf{x}) \qquad (14)$$

The concepts of the probability $p(\mathbf{x}, \theta)$ and likelihood $l(\theta, \mathbf{x})$ are related in terms of that the maximization procedure for p is proportional to maximizing l. The difference lies in the way the prior knowledge is expressed. Knowing a pdf with known parameters will enable prediction of an outcome of a random process. The concept of a likelihood aims at estimating the parameters when data is already observed.

In many application it will be more convenient to maximize the logarithm of the likelihood function (eq.15). The first example in this section will show some useful properties of Λ.

$$\Lambda(\theta, \mathbf{x}) = \ln(l(\theta, \mathbf{x})) \tag{15}$$

Generally there is no optimality criterion in terms of variance of estimates attaining the CRLB. However in many cases a ML procedure will actually yield a MVU estimator if it exists. The ML estimation improves with larger data sets. If the observation interval is infinitely long $N \to \infty$, ML will yield an unbiased (eq.16) and efficient (eq.17) estimator. In practice the length of N for which ML estimator becomes asymptotically efficient will not be known and might be determined using computer simulations [1].

$$\lim_{N\to\infty} E(\hat{\theta}) = \theta \tag{16}$$

$$\lim_{N\to\infty} var(\hat{\theta}) = CRLB \tag{17}$$

A unique advantage of ML estimators is that a numerical value of θ can always be found. This is due to fact that a maximum of a known function Λ is determined within a bounded parameter space. Numerically this value can be found using grid search techniques. This property also makes ML more flexible. Discrete parameter problems can be handled as well as estimation of a parameters from a continuous range of possibilities. Two examples show this.

3.1 Example - Binomial Distribution

The binomial distribution in eq.18 gives the discrete probability distribution of obtaining exactly h successes out of N trials. Two discrete outcomes H1 and H2 of a random process exist. The question is how to calculate probabilities of H1 and H2 from a set of N observations.

$$p(h \mid N) = \frac{N!}{h!(N-h)!} p^h (1-p)^{N-h} \tag{18}$$

Out of $N = 1000$ trials H1 occurs $h = 658$ times. In order to find probabilities of H1 and H2 we need to find a value of p for which eq.18 has its maximum. Assuming $p = 0.5$ eq.19 needs to be calculated.

$$\frac{1000!}{658!342!} 0.5^{658} (1-0.5)^{342} \tag{19}$$

At this point it is clear why we would prefer to calculate the logarithm of the likelihood function instead of the values of the distribution itself. Whereas eq.19 contains both very large and very small numerical values, the introduction of eq.15 mitigates the numerical problems by calculating

$$\ln(0.5^{658}0.5^{342}) \approx -693.15 \tag{20}$$

Evaluating the log-likelihood function for different values of p we find the maximum at $p1 = 0.66$ as shown in fig.2.

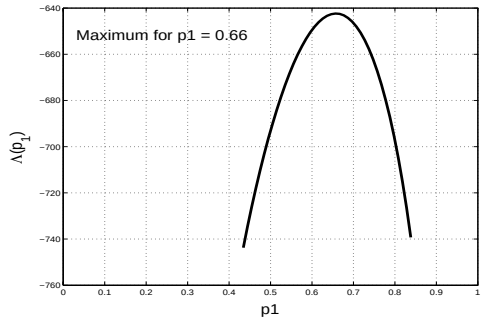

Figure 2: ML estimation of a binomial distribution

3.2 Example - Normal Distribution

In the previous example a ML estimation was performed on binary data. More general case includes estimating parameters for from a continuous set of possibilities. In this example a procedure for estimating mean μ and standard deviation σ of a normal distribution will be introduced. Although the results from eq.24 and eq.25 are well known this example should show how these results are obtained.

We assume a number of realisations of a random gaussian process. According to eq.15 the likelihood function is given as

$$L(\theta; \mathbf{x}) = L(\mu, \sigma; x_1, ..., x_n) = \prod_i \frac{1}{\sigma\sqrt{2\pi}} exp\Big(\frac{-(x_i - \mu)^2}{2\sigma^2}\Big) \tag{21}$$

or equivalently

$$L(\theta; \mathbf{x}) = \frac{(2\pi)^{-n/2}}{\sigma^n} exp\Big(-\frac{\sum_i (x_i - \mu)^2}{2\sigma^2}\Big) \tag{22}$$

Taking the natural logarithm of eq.22 yields

$$\Lambda(\mu, \sigma; \mathbf{x}) = \ln(L(\mu, \sigma; \mathbf{x})) = \frac{-1}{2} n \ln(2\pi) - n \ln(\sigma) - \frac{\sum_i (x_i - \mu)^2}{2\sigma^2} \tag{23}$$

In order to find the maximum of the function eq.23 is partial derivatives with regard to μ and σ are calculated and set to zero. This yields the

final results of the analysis, the mean and standard deviation of a normal distribution.

$$\frac{\partial \Lambda(\mu, \sigma; \mathbf{x})}{\partial \mu} = \frac{\sum_i (x_i - \mu)}{\sigma^2} = 0 \rightarrow \hat{\mu} = \frac{\sum_i x_i}{n} \tag{24}$$

$$\frac{\partial \Lambda(\mu, \sigma; \mathbf{x})}{\partial \sigma} = \frac{-n}{\sigma} + \frac{\sum_i (x_i - \mu)^2}{\sigma^3} = 0 \rightarrow \hat{\sigma} = \sqrt{\frac{\sum_i (x_i - \hat{\mu})^2}{n}} \tag{25}$$

4 Least Squares Estimators

In this section an overview over different LS approaches will be provided. A variety of estimation algorithms based on the minimization of the least square error are available. They all share some basic properties. Here again, no statistical knowledge of the data is required however a valid signal model s as a function of a parameter vector θ is essential. The advantage is a broader range of applications that can be cover using LS estimation. Lack of statistical characterization in most general case of both signal and noise leads to a lack of any optimality criterion in terms of CRLB. However an error function will be defined. Its minimization will be a quality criterion for LS estimation. If it is desirable to have any statistical measure for estimator performance, statistical descriptions of signal, noise or both will be necessary.

The goal of an LS estimator is to minimize an error function, which is the sum of squared errors (eq.26). After observing $\mathbf{x}$ optimal parameters θ driving the signal model are calculated which minimize $J(\theta)$.

$$J(\theta) = \sum_{n=0}^{N-1} (x[n] - s[n])^2 = (\mathbf{x} - \mathbf{s}(\theta))^T (\mathbf{x} - \mathbf{s}(\theta)) \tag{26}$$

Although in general case LS algorithms require no statistical description of the noise process, in most applications it will be reasonable to ensure the noise process has a zero mean, otherwise LS estimators tend to produce biased estimates.

Many classes of LS estimators exist. Depending on the functional relationship of the signal model s and its parameters θ following classes are distinguished.

- Linear estimators
- Nonlinear estimators
- Separable estimators

Linear modeling implies liner dependence of θ in s. The function itself doesn't necessary has to be linear, but only the parameter to be estimated. An example for such a function is estimation of the amplitude of sinusoidal carrier as given in eq.27.

$$J(A) = \sum_{n=0}^{N-1} (x[n] - A\cos(2\pi f_0 n))^2 \tag{27}$$

A nonlinear LS problem emerges if we estimate a parameter which is nonlinear in the signal model. In the same situation as above, estimating the carrier frequency instead of the amplitude yields a nonlinear LS problem (eq.33).

$$J(f_0) = \sum_{n=0}^{N-1} (x[n] - A\cos(2\pi f_0 n))^2 \tag{28}$$

If the parameter vector estimated using LS contains both linear and nonlinear parameters separable LS problem emerges. This situation is given in eq.29.

$$J(A, f_0) = \sum_{n=0}^{N-1} (x[n] - A\cos(2\pi f_0 n))^2 \tag{29}$$

Depending on the application several possible linear estimation approaches exist. These are summarized as follows:

- Linear estimation (batch approach)
- Weighted linear estimation
- Order-recursive estimation
- Sequntial estimation

As in the section on linear estimators, the signal model assumed is $\mathbf{x} = \mathbf{H}\theta$. However no statistical characterization of the noise process is given. The estimator is then calculated as

$$\hat{\theta} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{x} \tag{30}$$

Note that opposite to the theory of linear estimators in section 1 no performance criterion is given here. A useful generalization of eq.30 is introduction of a weighting matrix $\mathbf{W}$. Some prior knowledge about the observed samples could be applied using $\mathbf{W}$. If for example some data values are known to be more reliable than the others these could be emphasized using $\mathbf{W}$. The estimator is then given in eq.31.

$$\hat{\theta} = (\mathbf{H}^T\mathbf{W}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{W}\mathbf{x} \tag{31}$$

Order recursive LS estimation is applied if the order of the data model is unknown. Increasing order will yield a a better performance in terms of minimizing J_{min} at the cost of higher computational complexity. This trade-off needs to be addressed depending on the application, available resources and required accuracy.

Sequential LS estimation is an iterative approach which allows the incoming data to be processed sequentially as they come in. Unlike te batched LS approach as described in eq.30 where estimation is calculated when the whole observation interval is available, here after obtaining an initial estimates the value of $\hat{\theta}$is updated according to eq.32.

$$\hat{\theta}[n] = \hat{\theta}[n-1] + \mathbf{K}[n](x[n] - \mathbf{h}^T\hat{\theta}[n-1]) \tag{32}$$

After every iteration the gain vector $K[n]$ needs to be updated. This operation requires no matrix inversion. Sometimes the choice of initial values might be a problem. This is usually done by observing the process and calculating the initial estimates using batch approach, or making a reasonable guess on the values.

Nonlinear LS regression is given in its most general form in eq.33. $s(\theta)$ is an arbitrary nonlinear function in θ. Solving eq.34 requires a solution of N simultaneous nonlinear equations. This is addressed as the problem of nonlinear regression.

$$J(\theta) = \sum_{n=0}^{N-1}(\mathbf{x} - s(\theta))^2 = (\mathbf{x} - s(\theta))^T(\mathbf{x} - s(\theta)) \tag{33}$$

$$\frac{\partial J}{\partial \theta} = \frac{\partial \mathbf{s}(\theta)^{\mathbf{T}}}{\partial \theta}(\mathbf{x} - \mathbf{s}(\theta)) = 0 \tag{34}$$

Two different approaches to deal with this problem exist. One is the Newton-Raphson method. Here a solution is found by linearizing the derivative of $J(\theta)$ and iterating until the solution is converging. The other possible approach is the Gauss-Newton method which uses a linearized signal model in order to find the solution [1].

References

[1] S.M. Kay. *Fundamentals of Statistical Signal Processing*, volume Vol.1 - Estimation Theory. Prentice Hall SP Series, 1993.

[2] T.K. Moon and W.C.Stirling. *Mathematical Methods and Algorithms*. Prentice Hall, 1999.