Nonlinear Signal Processing
ELEG 833

Gonzalo R. Arce

Department of Electrical and Computer Engineering
University of Delaware
arce@ee.udel.edu

Fall 2008

Outline

1. Statistical Foundations of Filtering
   - Properties of Estimators
     - Unbiased Estimators
   - Maximum Likelihood Estimation
   - Robust Estimation

Statistical Foundations of Filtering

Filtering and parameter estimation are related since information is carried into one or more parameters of a signal.

- In AM and FM modulation, the information resides in the envelope and instantaneous frequency.
- Information can be carried in a number of signal parameters the mean, variance, phase, frequency.

The location estimate:

Suppose that a constant signal β is transmitted through a channel which adds Gaussian noise Z_i. Several independent observations X_i are measured giving

$$X_i = \beta + Z_i \qquad i = 1, 2, \cdots, N.$$

Given $X_1,\ X_2, \cdots,\ X_N$, the goal is to derive a “good” estimate of β. Estimates of this kind are known as *location estimates*, a *key* in the formulation of the optimal filtering problem.

Several methods of estimating β are possible.

- The sample mean:

$$\bar{\beta}_N = \bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$$

- The sample median $\tilde{\beta}_N = \tilde{X}$.
- The *trimmed-mean* (the largest and smallest samples are first discarded and the remaining $N-2$ samples are averaged.)

Which one of these estimators, if any, is correct will depend on the criterion which is selected.

Outline

1. Statistical Foundations of Filtering
   - Properties of Estimators
     - Unbiased Estimators
   - Maximum Likelihood Estimation
   - Robust Estimation

Properties of Estimators

- Estimators use observations that are random variables.
- The estimates are themselves random variables.
- We can recourse to the statistical properties of the estimates to quantify their "goodness".
- The statistical properties can be used for purposes of comparison among various estimators.

$$
\begin{aligned}
\text{Example :}\quad 1)\ \{x_i\} &= \{5,\ -1,\ 3,\ 2,\ -4\} \\
\bar{X} &= 1 \\
2)\ \{x_i\} &= \{2,\ 0,\ -3,\ 1,\ 3\} \\
\bar{X} &= \frac{3}{5}
\end{aligned}
$$

Unbiased Estimators

A typical probability density $f_{\hat{\beta}}(y/\beta)$ is shown below.

- A "good" estimator will have its density function clustered about β.
- The mean of $\hat{\beta}$ should be close to β.

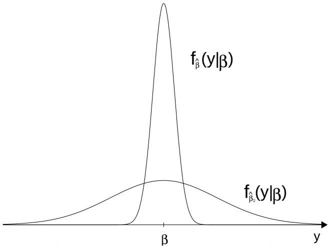


Figure: Probability density function associated with an unbiased location estimator

- In some cases, it is possible to design estimators for which the mean of $\hat{\beta}$ is always equal to the true value of β.
- If this is true for all values of β, the estimator is referred to as *unbiased*.
- The N-sample estimate of β, denoted as $\hat{\beta}_N$, is said to be *unbiased* if

$$E\{\hat{\beta}_N\} = \beta.$$

The variance of the estimate determines its precision.
If an unbiased estimate has low variance, then it is more reliable than other unbiased estimates with larger variances.

- The sample mean is unbiased since $E\{\hat{\beta}_N\} = \beta$, with variance

$$\begin{aligned} \operatorname{var}\left\{\hat{\beta}_N\right\} &= \operatorname{var}\left\{\frac{1}{N}\sum_{i=1}^{N} X_i\right\} \\ &= \frac{\sigma^2}{N}, \end{aligned} \tag{1}$$

  where σ^2 is the input variance.
  The precision of the estimate improves as N increases.

Efficient Estimators

- If we consider unbiased estimators only, the "best" estimator attains the minimum variance.
- This may seem partially useful since we would have to search among all unbiased estimators to determine which has the lowest variance.
- A lower bound on the variance of *any* unbiased estimator exists.
- If a given estimator is found to have a variance equal to that of the bound, the *best* estimator has been identified.

Let $f(\mathbf{X};\beta)$ be the density function of the observations $\mathbf{X}$ given the value of β. If $\hat{\beta}$ is an unbiased estimate of β, its variance is bounded by the Cramér-Rao bound:

$$\mathrm{var}\{\hat{\beta}\} \geq \left(E\left\{ \left[\frac{\partial}{\partial \beta} \ln f(\mathbf{X};\beta) \right]^2 \right\} \right)^{-1} \tag{2}$$

provided that the partial derivative of the log likelihood function exists and is absolutely integrable.

A second form of the Cramér-Rao bound can be written as

$$\mathrm{var}\{\hat{\beta}\} \geq \left(-E\left\{ \frac{\partial^2}{\partial \beta^2} \ln f(\mathbf{X}; \beta) \right\} \right)^{-1} \quad (3)$$

being valid if the second partial derivative of the log likelihood exists.

- There is no guarantee that an unbiased estimate exists whose variance satisfies the Cramér-Rao bound with equality.
- If one is found it is referred to as an *efficient estimator*.

Outline

Maximum Likelihood Estimation

Having a set of observation samples, a number of approaches can be taken to derive an estimate.

The method of *maximum likelihood (ML)* is the most popular approach.

Conceptually, a set of observations, $X_1,\ X_2, \cdots,\ X_N$, obey the joint distribution function $f(X_1,\ X_2, \cdots,\ X_N; \beta)$ where β is a parameter of the distributions.

- The ML estimate of β is the value $\hat{\beta}_{ML}$ for which the function $f(X_1,\ X_2, \cdots,\ X_N; \beta)$ is at its maximum:

$$\hat{\beta}_{ML} = \arg\max_{\beta} f(X_1, X_2, \cdots, X_N; \beta). \tag{4}$$

- β is variable while the samples $X_1,\ X_2, \cdots,\ X_N$ are fixed.
- The density function as a function of β, for fixed values of the observations, is known as the *likelihood function*.
- If an efficient estimate exists, the maximum-likelihood estimate is efficient.

Location Estimation in Gaussian Noise

Assume that $X_1, X_2, \cdots, X_N$, are i.i.d. Gaussian with a constant but unknown mean β. The Maximum Likelihood estimate of location is the value $\hat{\beta}$ which maximizes the likelihood function

$$\begin{aligned} f(X_1, X_2, \cdots, X_N; \beta) &= \prod_{i=1}^{N} f(X_i - \beta) \\ &= \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \, e^{-(X_i-\beta)^2/2\sigma^2} \\ &= \left(\frac{1}{2\pi\sigma^2}\right)^{N/2} e^{-\sum_{i=1}^{N}(X_i-\beta)^2/2\sigma^2}. \end{aligned} \tag{5}$$

The ML estimate of location is the value $\hat{\beta}$ which minimizes the least squares sum

$$\hat{\beta}_{ML} = \arg\min_{\beta} \sum_{i=1}^{N} (X_i - \beta)^2. \tag{6}$$

The value that minimizes the sum, results in the sample mean

$$\hat{\beta}_{ML} = \frac{1}{N} \sum_{i=1}^{N} X_i. \tag{7}$$

Note that the sample mean is unbiased since $E\{\hat{\beta}_{ML}\} = (1/N) \sum_{i=1}^{N} E\{X_i\} = \beta$. As a ML estimate, it is efficient having its variance, in (1), reach the Cramér-Rao bound.

Location Estimation in Generalized Gaussian Noise

In the generalized Gaussian distribution case, the Maximum Likelihood estimate of location is

$$\begin{aligned} f(X_1, X_2, \cdots, X_N; \beta) &= \prod_{i=1}^{N} f_\gamma(X_i - \beta) \\ &= \prod_{i=1}^{N} C\ e^{-|X_i - \beta|^\gamma / \sigma} \\ &= C^N e^{-\sum_{i=1}^{N} |X_i - \beta|^\gamma / \sigma}, \end{aligned} \tag{8}$$

where C is a normalizing constant, and γ is the dispersion parameter. Maximizing the likelihood function is equivalent to

$$\tilde{\beta}_{ML} = \arg\min_{\beta} \sum_{i=1}^{N} |X_i - \beta|^\gamma. \tag{9}$$

$$\tilde{\beta}_{ML} = \arg\min_{\beta} \sum_{i=1}^{N} |X_i - \beta|^{\gamma}.$$

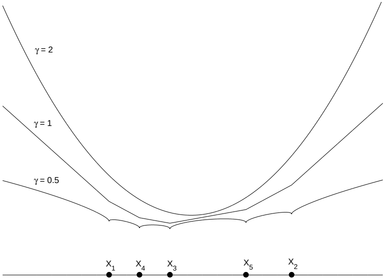


Figure: Cost functions for the observation samples $X_1 = -3, X_2 = 10, X_3 = 1, X_4 - 1, X_5 = 6$ for $\gamma = 0.5, 1,$ and 2.

When the dispersion parameter is 1, the model is Laplacian and the optimal estimator minimizes

$$\tilde{\beta}_{ML} = \arg\min_{\beta} \sum_{i=1}^{N} |X_i - \beta|. \tag{10}$$

The solution to the above is the sample median as it is shown next.
Define the cost function in (10) as $L_1(\beta)$. For values of β in the interval $-\infty < \beta \leq X_{(1)}$, $L_1(\beta)$ is simplified to

$$L_1(\beta) = \sum_{i=1}^{N} \left(X_{(i)} - \beta\right) = \sum_{i=1}^{N} X_{(i)} - N\beta. \tag{11}$$

This, as a direct consequence that in this interval, $X_{(1)} \geq \beta$.

For values of β in the range $X_{(j)} < \beta \leq X_{(j+1)}$, $L_1(\beta)$ can be written as

$$\begin{aligned} L_1(\beta) &= \sum_{i=1}^{j} \left(\beta - X_{(i)}\right) + \sum_{i=j+1}^{N} \left(X_{(i)} - \beta\right) \\ &= \left(\sum_{i=j+1}^{N} X_{(i)} - \sum_{i=1}^{j} X_{(i)} \right) - (N-2j)\,\beta, \end{aligned} \tag{12}$$

for $j = 1, 2, \cdots, N-1$.
Similarly, for $X_{(N)} < \beta < \infty$,

$$L_1(\beta) = -\sum_{i=1}^{N} X_{(i)} + N\beta. \tag{13}$$

Letting $X_{(0)} = -\infty$ and $X_{(N+1)} = \infty$, and defining $\sum_{i=m}^{n} X_{(i)} = 0$ if $m > n$, we can combine (11)-(13) into the following compactly written cost function

$$L_1(\beta) = \left(\sum_{i=j+1}^{N} X_{(i)} - \sum_{i=1}^{j} X_{(i)} \right) - (N - 2j)\,\beta, \quad j = 0, 1, \cdots, N \qquad (14)$$

for $\beta \in (X_{(j)}, X_{(j+1)}]$.

$$L_1(\beta) = \left(\sum_{i=j+1}^{N} X_{(i)} - \sum_{i=1}^{j} X_{(i)} \right) - (N - 2j)\,\beta, \quad j = 0, 1, \cdots, N$$

- $L_1(\beta)$ is piecewise linear and continuous.
- It starts with slope $-N$ for $-\infty < \beta \leq X_{(1)}$.
- As each $X_{(j)}$ is crossed, the slope is increased by 2.
- At the extreme right the slope ends at N for $X_{(N)} < \beta < \infty$.

For N odd there is an integer k, such that the slopes over the intervals $(X_{(k-1)}, X_{(k)}]$ and $(X_{(k)}, X_{(k+1)}]$, are negative and positive, respectively. From (14), these two conditions are satisfied if both

$$k < \frac{N}{2} \qquad \text{and} \qquad k > \frac{N}{2} - 1$$

hold. Both constraints are met when $k = \frac{N+1}{2}$

$$\begin{aligned} \hat{\beta}_{ML} &= \arg\min_{\beta} \sum_{i=1}^{N} |X_i - \beta| \\ &= \begin{cases} X_{(\frac{N+1}{2})} & N \text{ odd} \\ \left(X_{(\frac{N}{2})}, X_{(\frac{N}{2})}\right] & N \text{ even} \end{cases} \\ &= \mathrm{MEDIAN}(X_1, X_2, \cdots, X_N). \end{aligned} \tag{15}$$

Location Estimation in Stable Noise

Maximum likelihood estimation requires the knowledge of the density function. Among the class of symmetric stable densities, only the Gaussian ($\alpha = 2$) and Cauchy ($\alpha = 1$) distributions have closed-form expressions.

- The only non-Gaussian distribution for which we have a closed form expression is the Cauchy distribution.
- ML estimates under the Cauchy model can be made tunnable acquiring remarkable efficiency over the entire spectrum of stable distributions.

Given a set of i.i.d. samples $X_1, X_2, \cdots, X_N$ obeying the Cauchy distribution with scaling factor k,

$$f(x-\beta) = \frac{k}{\pi} \frac{1}{k^2 + (x-\beta)^2}, \tag{16}$$

the location parameter β is to be estimated from the data samples as the value $\hat{\beta}_k$ which maximizes the likelihood function

$$\hat{\beta}_k = \arg\max_{\beta} \prod_{i=1}^{N} f(X_i - \beta) = \arg\max_{\beta} \left(\frac{k}{\pi}\right)^N \prod_{i=1}^{N} \frac{1}{k^2 + (X_i - \beta)^2} \tag{17}$$

This is equivalent to minimizing

$$G_k(\beta) = \prod_{i=1}^{N}[k^2 + (X_i - \beta)^2]. \tag{18}$$

Given $k > 0$, the ML location estimate is known as the sample *myriad* and is given by

$$\begin{aligned} \hat{\beta}_k &= \arg\min_{\beta} \prod_{i=1}^{N} \left(k^2 + (X_i - \beta)^2\right) \\ &= \mathrm{MYRIAD}\{k; X_1, X_2, \cdots, X_N\}. \end{aligned} \tag{19}$$

The sample myriad involves the free parameter k (refered to as the *linearity parameter*). The behavior of the myriad is markedly dependent on the value of k.

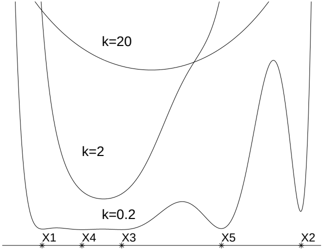


Figure: Myriad cost functions for the observation samples $X_1 = -3, X_2 = 10, X_3 = 1, X_4 - 1, X_5 = 6$ for $k = 20, 2, 0.2$.

Least Logarithmic Deviation

The sample myriad minimizes

$$G_k(\beta) = \prod_{i=1}^{N}[k^2 + (X_i - \beta)^2].$$

Since the logarithm is a strictly monotonic function, then the sample myriad will also minimize $\log G_k(\beta)$.

$$\text{MYRIAD}\{k; X_1, \cdots, X_N\} = \arg\min_{\beta} \sum_{i=1}^{N} \log\left[k^2 + (X_i - \beta)^2\right]. \tag{20}$$

If an observation in the set of input samples has a "large" magnitude such that $|X_i - \beta| \gg k$, the cost associated with this sample is approximately $\log(X_i - \beta)^2$ —the log of the square deviation.

The sample myriad (approximately) minimizes the sum of logarithmic square deviations, referred to as the LLS criterion.

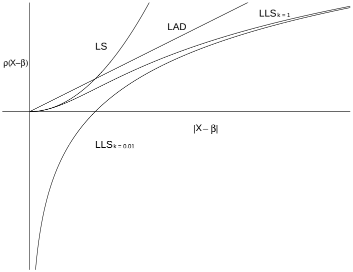


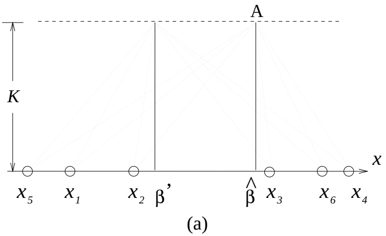
Figure: Cost functions of the mean (LS), the median (LAD) and the myriad (LLS)

Geometrical Interpretation

The observations $X_1,\ X_2, \cdots,\ X_N$ are placed along the real line. Next, a vertical bar that runs horizontally through the real line is added. The length of the vertical bar is equal to k. Each of the terms

$$\left(k^2 + (X_i - \beta)^2\right) \tag{21}$$

in (20), represents the distance from point A, at the end of the vertical bar, to the sample point X_i.



The sample myriad, $\hat{\beta}_k$, indicates the position of the bar for which the product of distances from point A to the samples $X_1, X_2, \cdots, X_N$ is minimum. Any other value, such as $x = \beta'$, produces a higher product of distances.

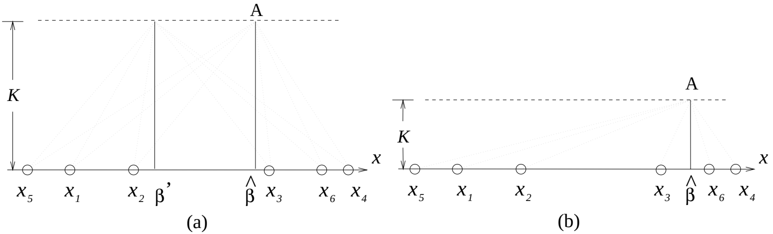


Figure: (a) The sample myriad, $\hat{\beta}$, minimizes the product of distances from point A to all samples. (b) the myriad as k is reduced.

Outline

1. Statistical Foundations of Filtering
   - Properties of Estimators
     - Unbiased Estimators
   - Maximum Likelihood Estimation
   - Robust Estimation

Robust Estimation

The maximum likelihood estimates have assumed that the form of the distribution is known. In practice, we can seldom be certain of such distributional assumptions and two types of questions arise:

1. How sensitive are optimal estimators to the precise nature of the assumed probability model?
2. Is it possible to construct *robust* estimators which perform well under deviations from the assumed model?

Sensitivity of Estimators

To answer the first question, consider an observed data set $Z_1, \ Z_2, \cdots, \ Z_N$, and consider the various location estimators previously described. We also consider two simple estimators: the *trimmed-mean* defined as

$$T_N(\alpha) = \frac{1}{N - 2\alpha} \sum_{i=\alpha+1}^{N-\alpha} Z_{(i):N} \tag{22}$$

for $\alpha = 0, 1, \cdots, \lfloor N/2 \rfloor$, and the *Windsorized mean* defined as:

$$W_N(r) = \frac{1}{N} \left[\sum_{i=r+2}^{N-r-1} Z_{(i):N} + (r+1) \left[Z_{(r+1):N} + Z_{(N-r):N} \right] \right] \tag{23}$$

The median, is a special case of trimmed mean where $\alpha = \lfloor N/2 \rfloor$.

In the first set of experiments, a sample set of size 10 including one outlier is considered. The nine i.i.d. samples are distributed as $N(\mu, 1)$ and the outlier is distributed as $N(\mu + \lambda, 1)$.

TABLE: Bias of estimators of μ for $N = 10$ when a single observation is from $N(\mu + \lambda, 1)$ and the others from $N(\mu, 1)$.

	λ							
Estimator	0	0.5	1.0	1.5	2.0	3.0	4.0	∞
$\bar{X}_{10}$	0	0.05000	0.10000	0.15000	0.20000	0.30000	0.40000	∞
$T_{10}(1)$	0	0.04912	0.09325	0.12870	0.15400	0.17871	0.18470	0.18563
$T_{10}(2)$	0	0.04869	0.09023	0.12041	0.13904	0.15311	0.15521	0.15538
Med_{10}	0	0.04932	0.08768	0.11381	0.12795	0.13642	0.13723	0.13726
$\text{W}_{10}(1)$	0	0.04938	0.09506	0.13368	0.16298	0.19407	0.20239	0.20377
$\text{W}_{10}(2)$	0	0.04889	0.09156	0.12389	0.14497	0.16217	0.16504	0.16530

TABLE: Mean squared error of various estimators of μ for $N = 10$, when a single observation is from $N(\mu + \lambda, 1)$ and the others from $N(\mu, 1)$.

	λ							
Estimator	0.0	0.5	1.0	1.5	2.0	3.0	4.0	∞
$\bar{X}_{10}$	0.10000	0.10250	0.11000	0.12250	0.14000	0.19000	0.26000	∞
$T_{10}(1)$	0.10534	0.10791	0.11471	0.12387	0.13285	0.14475	0.14865	0.14942
$T_{10}(2)$	0.11331	0.11603	0.12297	0.13132	0.13848	0.14580	0.14730	0.14745
Med_{10}	0.13833	0.14161	0.14964	0.15852	0.16524	0.17072	0.17146	0.17150
$\text{W}_{10}(1)$	0.10437	0.10693	0.11403	0.12405	0.13469	0.15039	0.15627	0.15755
$\text{W}_{10}(2)$	0.11133	0.11402	0.12106	0.12995	0.13805	0.14713	0.14926	0.14950

Table: Variance of estimators of μ for $N = 10$, where a single observation is from $N(\mu, \sigma^2)$ and the others from $N(\mu, 1)$.

	σ					
Estimator	0.5	1.0	2.0	3.0	4.0	∞
X_{10}	0.09250	0.10000	0.13000	0.18000	0.25000	∞
$T_{10}(1)$	0.09491	0.10534	0.12133	0.12955	0.13417	0.14942
$T_{10}(2)$	0.09953	0.11331	0.12773	0.13389	0.13717	0.14745
Med_{10}	0.11728	0.13833	0.15373	0.15953	0.16249	0.17150
$\text{W}_{10}(1)$	0.09571	0.10437	0.12215	0.13221	0.13801	0.15754
$\text{W}_{10}(2)$	0.09972	0.11133	0.12664	0.13365	0.13745	0.14950

The mean is better than the median as long as the variance of the outlier is not large. The trimmed mean, however, outperforms the median regardless of the variance of the outlier.