Fundamentals on Estimation Theory

Marcos Martín Fernández

May 4, 2004

Under the *estimation theory* title a doctrinal body is collected which treats to solve a problem with very simple formulation: given a set of noisy observations from reality to guess the value a magnitude has taken is under consideration departing from those observations. Obviously, we have got no direct access to that magnitude but it has some functional relationship with the obtained observations.

The importance of the *noisy* adjective in the previous paragraph must be realized. If the obtained observations taken from reality have not got any degree of uncertainty, the unknown magnitude would be determined without no error from solving, in favor of that magnitude, the functional relationship mentioned beforehand. As many times as the experiment being repeated, as many times the same value would be obtained. However, the nature do not behave like that; let repeat any measurement of any magnitude several times and surely different values will be obtained. The fluctuations shall be with higher or lower amplitude depending on the random factors which affect the problem, and thus averaging all those measurements, the fluctuation will decrease as the number of averaged measurements increase.

The estimation theory deals then with, from given observations, building a function of them which allows achieving a value for the unknown magnitude as accurate as possible.

1 Classification

We can make the following distinction within the *estimation theory*:

- **Parametric estimators.** These estimators are built from the knowledge or the assumption of the probability density function for the data and for the magnitudes to estimate. These estimators used to be very dependent on the adjustment goodness of the data to the density functions.
- **Non-parametric estimators.** In this case, none assumption on the behavior of the data is carried on. The result is the foundation of robust estimators, that is, with few dependences on the true statistics of the data, though without having any optimality properties.

For particular problems, both estimation philosophies could give rise the same estimator.

In what follows, we come round the parametric methods, as in many cases the previous hypotheses used to be very realistic[1]. Anyway, in our exposition we will interleave some non parametric method, as the minimum variance unbiased estimator and the least square estimator. Within those methods is common to make the following classification:

- **Classical Estimation.** The magnitude (vectorial in general) is modeled as a deterministic vector with unknown value. In that case, under the assumptions that the observations were

[1] The central limit theorem gives us some guarantee for that.

denoted by the random vector $\mathbf{x}$ and the unknown parameter by $\boldsymbol{\vartheta}$, the density function is *parameterized* by $\boldsymbol{\vartheta}$, so we can write

$$\mathsf{f}(\mathbf{x};\boldsymbol{\vartheta}). \tag{1}$$

This case is referred to as parameter estimation.

- **Bayesian estimation.** The magnitude to estimate is modeled as a random vector whose probability density function models the prior knowledge about having that magnitude. In this case, following the previous notation, we can write

$$\mathsf{f}(\mathbf{x}/\boldsymbol{\vartheta}), \tag{2}$$

where it can be seen that the density function of the observations is, in this case, a density function conditioned on the value taken by the non-observable variable. This second case is referred to as random variable estimation.

The difference between both statements is outstanding. In the Bayesian case we can realize that we have previous knowledge about the magnitude to estimate which can be concreted into the density function $\mathsf{f}(\boldsymbol{\vartheta})$.

2 Measurement of the Estimator Quality

As we have said previously, an estimator is a function of the observations, what gives rise that the estimator results in a random variable. For that reason, the estimator quality can only be given in probabilistic terms.

Let analyze in parallel the parameter and the random variable estimation cases:

- **Parameter estimation.** The quality criterium commonly adopted for a parameter estimator is the minimum mean squared error criterium. Considering, for simplicity purposes, a scalar parameter (for a vectorial parameter, what is said here must be understood as component by component) ϑ, and denoting the estimator with $\hat{\vartheta}$, the achieved error would be equal to $\varepsilon = \vartheta - \hat{\vartheta}$, and its mean squared error would be

$$\mathsf{E}(\varepsilon^2) = \mathsf{E}((\vartheta - \hat{\vartheta})^2). \tag{3}$$

The coupled problem to that criterium is the fact that the estimators created that way used to be unfeasible, as the estimators used to be function not only from the observations but also from the parameter to estimate itself. The reason of that is due to the mean square error can be rewritten as[2]

$$\begin{aligned}
\mathsf{E}(\varepsilon^2) &= \mathsf{E}((\vartheta - \hat{\vartheta} - \mathsf{E}(\hat{\vartheta}) + \mathsf{E}(\hat{\vartheta}))^2) \\
&= (\vartheta - \mathsf{E}(\hat{\vartheta}))^2 + \mathsf{E}((\hat{\vartheta} - \mathsf{E}(\hat{\vartheta}))^2) \\
&= \mathsf{bias}^2(\hat{\vartheta}) + \mathsf{var}(\hat{\vartheta}),
\end{aligned} \tag{4}$$

that is, the squared *bias* of the estimator plus the estimator variance. The bias used to be function of the parameter to estimate itself, and, due to that, the dependence on that

[2]The $\mathsf{bias}(\cdot)$ operator represents the bias of its argument.

parameter goes, through the minimization process for the mean squared error, straight to the estimator.

The last concern could be avoided typically with the mean squared error minimization constrained to the unbiased estimators. In particular, a estimator is referred to as *unbiased* if it satisfies the condition $\mathsf{E}(\hat{\vartheta}) = \vartheta,\ \forall\ \vartheta$. For those estimators, as it can be seen from equation (4), the mean squared error matches with the estimator variance. Thus, assuming that constraint, the estimator is built, within all unbiased estimators, choosing that one which has got the minimum variance for all values of the ϑ parameter. Such an estimator, in the case of existence (which can not exist) is referred to as *minimum variance uniform unbiased estimator*.

The previous constraint is not worthless; note that we are not allowing interchanging bias with variance, so we are impeding from building an estimator that having a small bias has lower variance that other unbiased estimators and that the mean squared error being minimum. Anyway, for the frequent practical case of Gaussian estimators, it can be proven easily that if $\mathsf{E}(\hat{\vartheta}_1) = \mathsf{E}(\hat{\vartheta}_2) = \vartheta$, and $\mathsf{var}(\hat{\vartheta}_1) < \mathsf{var}(\hat{\vartheta}_2)$, then

$$\mathsf{P}(|\hat{\vartheta}_1 - \vartheta| < r) > \mathsf{P}(|\hat{\vartheta}_2 - \vartheta| < r), \tag{5}$$

which means that the true value of the parameter is found in an interval with radius r round the estimator *with hugest probability* in the minimum variance case.

The objective will be thus to find the unbiased estimator, when it exists, with minimum variance. When it does not exist, we have to resort to any other approach.

- **Random variable estimation.** The magnitude to estimate is not a deterministic parameter but a random variable with a known probability density function. So the optimality conditions, in this case, must be adapted adequately. In particular, in the Bayesian case the magnitude to determine is modeled as a random variable $\boldsymbol{\vartheta}$, which has some density function $\mathsf{f}(\boldsymbol{\vartheta})$ associated. This time we do not treat with the error itself but with the error *cost*, that is, with a function of the difference between the variable value and the estimator (assuming the scalar case)

$$\mathsf{c}(\varepsilon) = \mathsf{c}\left(\vartheta - \hat{\vartheta}(\mathbf{x})\right), \tag{6}$$

which, as before, will be a random variable which we will calculate its expectation from. The mean cost is generally referred to as *risk* and can be written as

$$\begin{aligned} \mathsf{E}(\mathsf{c}(\varepsilon)) &= \mathsf{E}\left(\mathsf{c}\left(\vartheta - \hat{\vartheta}(\mathbf{x})\right)\right) = \int \mathsf{c}\left(\vartheta - \hat{\vartheta}(\mathbf{x})\right) \mathsf{f}(\mathbf{x}, \vartheta) d\mathbf{x} d\vartheta \\ &= \int \mathsf{c}\left(\vartheta - \hat{\vartheta}(\mathbf{x})\right) \mathsf{f}(\mathbf{x}/\vartheta) \mathsf{f}(\vartheta) d\mathbf{x} d\vartheta. \end{aligned} \tag{7}$$

So in this case we are averaging the error following the joint density function for the data and parameter, that is, the error measurement is averaged for all values the random variable ϑ can reach, adequately weighted by their relative importance given by $\mathsf{f}(\vartheta)$. Here it is non sense to talk about minimum variance uniform estimator as the error dependence on the parameter value has been eliminated.

3 Building Parameter Estimators

3.1 Cramer Rao Lower Bound and Finding the Efficient Estimator

The *Cramer Rao lower bound* or CRLB gives us the minimum variance that can be expected from an unbiased estimator. In particular, and assuming by now an scalar parameter ϑ, if we

accept that the data density function satisfies the following regularity condition

$$\mathsf{E}\left(\frac{\partial \ln \mathsf{f}(\mathbf{x};\vartheta)}{\partial \vartheta}\right) = 0, \tag{8}$$

then it can be proven that the minimum achieve able variance is

$$\mathsf{var}(\hat{\vartheta}) \geq \frac{-1}{\mathsf{E}\left(\frac{\partial^2 \ln \mathsf{f}(\mathbf{x};\vartheta)}{\partial \vartheta^2}\right)}. \tag{9}$$

That fact not only tells us which the maximum expected quality for our estimators is, but also is constructive. Specifically, if we can write the density function of the estimator as

$$\frac{\partial \ln \mathsf{f}(\mathbf{x};\vartheta)}{\partial \vartheta} = \mathsf{I}(\vartheta)\left[g(\mathbf{x}) - \vartheta\right], \tag{10}$$

then that the estimator $\hat{\vartheta} = g(\mathbf{x})$ reaches the CRLB can be proven, that is, its variance is given by the equation (9). Particularly in this case, the variance results to be equal to $1/\mathsf{I}(\vartheta)$. This estimator, in case of existence, would be referred to as *efficient*.

The vectorial case extension is simple. The regularity condition given by the equation (8) must be satisfied by the derivative with respect to each component ϑ_i for the vector $\boldsymbol{\vartheta}$, with $i \in \{1, 2, \ldots, d\}$. Defining now the *Fisher information matrix* $\mathbf{I}(\boldsymbol{\vartheta})$ as that matrix whose element at position (i, j) with $1 \leq i \leq d$, and $1 \leq j \leq d$, is

$$\mathsf{I}_{ij}(\boldsymbol{\vartheta}) = -\mathsf{E}\left(\frac{\partial^2 \ln \mathsf{f}(\mathbf{x};\boldsymbol{\vartheta})}{\partial \vartheta_i \vartheta_j}\right) \tag{11}$$

and if we can write now the data density function as

$$\frac{\boldsymbol{\partial} \ln \mathsf{f}(\mathbf{x};\boldsymbol{\vartheta})}{\boldsymbol{\partial \vartheta}} = \mathbf{I}(\boldsymbol{\vartheta})\left[\mathbf{g}(\mathbf{x}) - \boldsymbol{\vartheta}\right], \tag{12}$$

then the $\hat{\boldsymbol{\vartheta}} = \mathbf{g}(\mathbf{x})$ vector is the efficient vector estimator and the variance for each component is given for the CRLB which in this case is given by

$$\mathsf{var}(\hat{\vartheta}_i) = \left[\mathbf{I}(\boldsymbol{\vartheta})\right]^{-1}_{(i,i)}. \tag{13}$$

3.2 Minimum Variance Unbiased Estimator

For a particular problem, it may happen that the efficient estimator does not exist, that is, we can not write the data density function following the equation (12), but it may happen that a *minimum variance unbiased estimator* or MVUE still exists, though with worse performance that –if it exists– the efficient one.

Whenever that is true, the method known as *Rao-Blackwell-Lehmann-Scheffe procedure* can be built that kind of estimator. That procedure is based on the concept of *sufficient statistics*. A data function $T(\mathbf{x})$ is referred to as a sufficient statistics whenever the data density function conditioned to the sufficient statistics is not a function of the parameter to estimate. It can be seen this way that the sufficient statistics captures all data dependencies with respect to the parameter to estimate thus the knowledge of that data function alone (and then forgetting about what value each individual datum from the observed sample have taken) is *sufficient* information for the correct estimation of the parameter. What is said in this paragraph can be expressed as follows: $T(\mathbf{x})$ is the sufficient statistics to estimate the $\boldsymbol{\vartheta}$ parameter if we can factorize the density function in the following way:

$$\mathsf{f}(\mathbf{x};\boldsymbol{\vartheta}) = g\left(T(\mathbf{x}), \boldsymbol{\vartheta}\right) h(\mathbf{x}). \tag{14}$$

The procedure for building the MVUE will be the following:

1. Let build any unbiased estimator $\check{\boldsymbol{\vartheta}}$.

2. Let obtain the sufficient statistics $T(\mathbf{x})$ for estimating the $\boldsymbol{\vartheta}$ parameter.

3. Let calculate

$$\hat{\boldsymbol{\vartheta}} = \mathsf{E}\left(\check{\boldsymbol{\vartheta}}/T(\mathbf{x})\right), \tag{15}$$

   then, if the sufficient statistics is complete, the $\hat{\boldsymbol{\vartheta}}$ estimator is the MVUE.

3.3 Best Linear Unbiased Estimator

The two preceding approaches are very illustrative and define the optima to take into account as reference; nevertheless, in most of the practical cases the complexity joined to such estimators, mainly the second, can be such too much great that it can be analytically untractable.

In this cases, suboptimal schemes but with analytical tractability can be used instead. Such schemes are the linear approaches, that is, we constrain the estimator to be a linear function of the data.

The question here is what linear function is the optimum one. The answer seems to be clear: such function which makes the estimator the *best unbiased linear estimator* or BLUE, that is, such estimator which besides of being a linear function of the data and unbiased, has minimum variance.

Hence, the estimator will have the form[3]

$$\hat{\boldsymbol{\vartheta}} = \mathbf{A}\mathbf{x}. \tag{16}$$

For this estimator to be unbiased, we have to suppose that the expected value of the data is a linear function of the unknown parameter

$$\mathsf{E}(\mathbf{x}) = \mathbf{H}\boldsymbol{\vartheta}. \tag{17}$$

Then, it can be proven that

$$\hat{\boldsymbol{\vartheta}} = \left(\mathbf{H}^T\mathsf{cov}(\mathbf{x})^{-1}\mathbf{H}\right)^{-1}\mathbf{H}^T\mathsf{cov}(\mathbf{x})^{-1}\mathbf{x}. \tag{18}$$

As it can be observed, the linear estimator is function only of the first and second order data statistics, with independency on the data distribution shape.

3.4 Maximum Likelihood Estimator

The *maximum likelihood estimator* or MLE is probably the parameter estimation method used mostly in practice. This fact is given because its calculus complexity is not as great as for the non-linear estimators explained above, but however the estimator presents some asymptotic optimality properties which confer it a huge practical interest.

The MLE is based on the principle of *maximum likelihood*. Specifically, if we consider a data sample $\mathbf{x}$ (think about a vector with N independent and identically distributed components) and if we know the marginal density of each, we can write the joint density function $\mathsf{f}(\mathbf{x};\boldsymbol{\vartheta})$ for the N components as the product of the marginalities.

If we consider now that expression as a function of the $\boldsymbol{\vartheta}$ parameter (so with the data sample vector $\mathbf{x}$ being constant), this function is referred to as *likelihood function*. Basically, that function means the probability that a random variable takes values around an infinitesimal volume centered about the point $\mathbf{x}$, but *as a function of the parameter* $\boldsymbol{\vartheta}$. Hence, it may seem

[3] In that follows we assume that all vectors are column vectors.

to be reasonable thinking about, once the data have been observed, the adequate value of the unknown parameter $\boldsymbol{\vartheta}$ should be such one which makes more probable to have observed the $\mathbf{x}$ sample.

Thus, the MLE is defined by

$$\hat{\boldsymbol{\vartheta}}_{ML} = \arg\max_{\boldsymbol{\vartheta}} \mathsf{f}(\mathbf{x}; \boldsymbol{\vartheta}). \tag{19}$$

As said before, the MLE enjoys some properties which make it very attractive in practice. Particularly:

- If the efficient estimator exists, the MLE will produce it. Effectively, if the efficient estimator exists, hence we can write the derivative of the logarithm of the density function (likelihood) as it was stated by the equation (12). So, as maximizing a function is equivalent to maximize the logarithm of that function, we can infer that the estimator under search is $\hat{\boldsymbol{\vartheta}}_{ML} = \mathbf{g}(\mathbf{x})$.

- Likewise, under very low constrained conditions, the MLE is asymptotically distributed (as the data sample size N goes up) with Gaussian behavior, in particular[4]

$$\lim_{N\to\infty} \hat{\boldsymbol{\vartheta}}_{ML} \sim \mathbf{N}(\boldsymbol{\vartheta}, \mathbf{I}^{-1}(\boldsymbol{\vartheta})). \tag{20}$$

  This gives rise, whenever the efficient estimator does not exist for finite samples sizes, to that the MLE to be asymptotically efficient. In addition, this property makes possible to calculate, assuming some kind of Gaussianity, confidence ranges where the true value of the parameter is confined with a given probability.

- Finally, the MLE satisfies also the *invariance* property. If to estimate $\boldsymbol{\alpha} = \mathbf{g}(\boldsymbol{\vartheta})$ is wished, hence it can be proven that $\hat{\boldsymbol{\alpha}}_{ML} = \mathbf{g}(\hat{\boldsymbol{\vartheta}}_{ML})$ is satisfied. This property can noticeably simplify the calculations wherever the likelihood function of the transformed parameter is difficult to obtain.

3.5 Least Squares Estimation

The *least squares estimation* o LSE is used whenever the probabilistic information about the data are not given. So this is an entirely deterministic approach, in general without any optimality property.

The LSE approach supposes that the observations follow the next formulation

$$x[n] = s[n; \boldsymbol{\vartheta}] + w[n], \quad 0 \leq n \leq N-1, \tag{21}$$

that is, a signal which is function of the parameters to be estimated over which a perturbation is overlaid. The latter could be understood as such a component of the observation which can not be explained by the signal generating model. The objective is, from a consistent sample for the N observations, to calculate the parameter vector $\boldsymbol{\vartheta}$ that makes the signal model to explain, as better as it can, the collected observations, that is, such a signal model that minimizes the term which is not explained by itself. So the estimator will be

$$\hat{\boldsymbol{\vartheta}}_{LS} = \arg\min_{\boldsymbol{\vartheta}} \sum_{n=0}^{N-1} \left(x[n] - s[n; \boldsymbol{\vartheta}]\right)^2. \tag{22}$$

This kind of estimation presents an operative difficulty which depends hardly on the signal generating model. In particular, if the parameters are a linear function of the data, the LSE

[4]With $\mathbf{N}(\boldsymbol{\eta}, \mathbf{C})$ we denote a multivariate Gaussian distribution with vector means $\boldsymbol{\eta}$ and covariance matrix $\mathbf{C}$.

has a closed-form solution. On the contrary, the optimization problem stated in last equation is non linear and, in each case, the way to obtain the solution have to be found out.

Let refine the linear model: if we reorganize the values $s[n;\boldsymbol{\vartheta}]$ as a column vector with N components, and the parameter vector $\boldsymbol{\vartheta}$ is formed by d parameters, we can write

$$\mathbf{s} = \mathbf{H}\boldsymbol{\vartheta}, \tag{23}$$

with $\mathbf{H}$ a matrix with dimensions $N \times d$. In order to the problem can be solved we suppose that $N > d$, and likewise that the $\mathbf{H}$ matrix is full rank, that is, with rank d. Under such assumptions, is easy to see that

$$\hat{\boldsymbol{\vartheta}}_{LS} = (\mathbf{H}^T\mathbf{H})^{-1}\mathbf{H}^T\mathbf{x}, \tag{24}$$

solution that would be equal to the BLUE given by the equation (18), in the case the used data for the latter were uncorrelated. However, as it can be seen, in LSE case we have made no assumption to that respect.

Let note that if the overlaid perturbation (reorganized as vector) $\mathbf{w}$ has Gaussian statistics, the previous solution will be the same as the MLE solution and likewise the estimator would be efficient. But, as we say, those are coincidences as in the LSE approach none probabilistic philosophy is followed.

Finally and with the objective to emphasize more similarities, realize that the equation (22) could be written in matrix form as

$$(\mathbf{x} - \mathbf{H}\boldsymbol{\vartheta})^T (\mathbf{x} - \mathbf{H}\boldsymbol{\vartheta}) . \tag{25}$$

If, by any means, it would be interesting to give more importance to some errors than to the others we could make use of a positive definite matrix $\mathbf{W}$ in the previous expression to be minimized

$$(\mathbf{x} - \mathbf{H}\boldsymbol{\vartheta})^T \mathbf{W} (\mathbf{x} - \mathbf{H}\boldsymbol{\vartheta}) , \tag{26}$$

which would give rise to the solution

$$\hat{\boldsymbol{\vartheta}}_{LS} = (\mathbf{H}^T\mathbf{W}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{W}\mathbf{x}, \tag{27}$$

which would increase the similarity with the corresponding expression for the BLUE given by the equation (18). Nonetheless, and once more, we arrive to *formal* similarities employing wholly different methodologies.

The LSE presents interesting geometrical properties and, by means of those, a recursive formulation could be developed which will progressively increase the estimator accurateness as more data are available to build it. The details in this case can be queried in any estimation theory book.

3.6 Estimation by means of the Moments Method

The *moments method* consists of a rather simple optimization approach which in spite, of being without any optimality properties, used to work satisfactorily for large sample size, as it used to be consistent. It is used as it is or as an initial estimation for algorithms like the iterative searching of the MLE.

The method consists of the following: if the data density function depends on d parameters, we can determine d moments of the distribution, for instance, the first d non-centered moments.

If we denote those with μ_i, $i \in \{1, 2, \ldots, d)$, we can write

$$\begin{aligned} \mu_1 &= h_1(\vartheta_1, \vartheta_2, \ldots, \vartheta_d) \\ \mu_2 &= h_2(\vartheta_1, \vartheta_2, \ldots, \vartheta_d) \\ &\vdots \quad\quad \vdots \\ \mu_d &= h_d(\vartheta_1, \vartheta_2, \ldots, \vartheta_d), \end{aligned} \tag{28}$$

that is, we can build a system, non-linear in general, with d equations and d unknowns

$$\boldsymbol{\mu} = \mathbf{h}(\boldsymbol{\vartheta}). \tag{29}$$

The estimator $\hat{\boldsymbol{\vartheta}}$ will be such one which makes last system of equations true when the moments vector $\boldsymbol{\mu}$ is replaced by its sample equivalent $\hat{\boldsymbol{\mu}}$, that is,

$$\hat{\boldsymbol{\vartheta}} = \mathbf{h}^{-1}(\hat{\boldsymbol{\mu}}). \tag{30}$$

Note that not for all cases the first d non-centered moments will be chosen. For instance, if the distribution presents symmetry with respect to the origin the odd moments will be all zero, so those moments will not be a functional expression of the distribution parameters. What can surely be considered as a general rule is the convenience of using moments with orders as lower as possible, as the variance of the sample moments increases with the order.

4 Building Bayesian Estimators

As said before, the *Bayesian philosophy* departs form considering the parameter to estimate as a sample of a random variable $\boldsymbol{\vartheta}$ which some kind of prior knowledge is given for. This knowledge is given by the density function of such variable, $\mathsf{f}(\boldsymbol{\vartheta})$. The observed sample is likewise a sample of the random variable $\mathbf{x}$, which has a probabilistic behavior that depends on the parameter to estimate through the density function conditioned to such parameter, $\mathsf{f}(\mathbf{x}/\boldsymbol{\vartheta})$. Hence, the Bayesian approach makes nothing but looking at how our probabilistic knowledge about the parameter $\boldsymbol{\vartheta}$ changes after knowing which value the variable $\mathbf{x}$ has taken. As it is very known, that new knowledge about the parameter distribution is given by the *posterior* density function of the parameter, which, following the Bayes' theorem, can be written as

$$\mathsf{f}(\boldsymbol{\vartheta}/\mathbf{x}) = \frac{\mathsf{f}(x/\boldsymbol{\vartheta})\mathsf{f}(\boldsymbol{\vartheta})}{\mathsf{f}(\mathbf{x})}. \tag{31}$$

Then, the Bayesian estimators are built from the posterior density function and their expression will depend on the cost error function $\mathsf{c}(\varepsilon)$ under consideration in each case, (see equation (7) in section 2). In particular:

- If $\mathsf{c}(\varepsilon) = \varepsilon^2$ is assumed, that is, the risk would be the mean squared error, the achieved Bayesian estimator is the posterior mean given by

$$\hat{\boldsymbol{\vartheta}}_{MMSE} = \mathbf{E}\left(\boldsymbol{\vartheta}/\mathbf{x}\right) \tag{32}$$

  This estimator is referred to as *minimum mean squared error estimator* or MMSEE.

- If $\mathsf{c}(\varepsilon) = |\varepsilon|$ is assumed, that is, the risk is the expectation of the absolute error value, the obtained Bayesian estimator is the median of the posterior distribution, thus, the $\hat{\boldsymbol{\vartheta}}_{MEDIAN}$ value which satisfies

$$\int_{-\infty}^{\hat{\boldsymbol{\vartheta}}_{MEDIANA}} \mathsf{f}(\boldsymbol{\vartheta}/\mathbf{x})d\boldsymbol{\vartheta} = \int_{\hat{\boldsymbol{\vartheta}}_{MEDIANA}}^{\infty} \mathsf{f}(\boldsymbol{\vartheta}/\mathbf{x})d\boldsymbol{\vartheta}. \tag{33}$$

- If $\mathsf{c}(\varepsilon) = 1$ if $|\varepsilon| > \delta$ and $\mathsf{c}(\varepsilon) = 0$ if $|\varepsilon| \leq \delta$, for an arbitrary small threshold δ, is assumed, that is, if the cost function is non zero as soon as a small error is made for the estimation, the achieved Bayesian estimator is the posterior mode given by

$$\hat{\boldsymbol{\vartheta}}_{MAP} = \arg\max_{\boldsymbol{\vartheta}} \mathsf{f}(\boldsymbol{\vartheta}/\mathbf{x}). \tag{34}$$

  This estimator is referred to as *maximum a posteriori estimator* or MAPE.

Those times for which the previous estimators are difficult to obtain, a simplification similar to that performed for the BLUEs used to be a normal practice which is to force that estimator to be a linear function of the data sample. In the Bayesian case the obtained expressions are much like the BLUE case, however, in this case, the first and second order statistics of the random variable to estimate appear. Specifically, assuming a data model as

$$\mathbf{x} = \mathbf{H}\boldsymbol{\vartheta} + \mathbf{w}, \tag{35}$$

where that the variable $\mathbf{w}$ has zero mean and that the variables $\boldsymbol{\vartheta}$ and $\mathbf{w}$ are uncorrelated, are supposed, hence it can be proved that the *linear minimum mean squared error estimator* or LMMSSEE follows the closed-form expression

$$\hat{\boldsymbol{\vartheta}}_{LMMSE} = \mathsf{E}(\boldsymbol{\vartheta}) + \mathsf{cov}(\boldsymbol{\vartheta})\mathbf{H}^T \left(\mathbf{H}\mathsf{cov}(\boldsymbol{\vartheta})\mathbf{H}^T + \mathsf{cov}(\mathbf{w})\right)^{-1} (\mathbf{x} - \mathbf{H}\mathsf{E}(\boldsymbol{\vartheta})). \tag{36}$$

Obviously, the power of that estimation philosophy is based on the fact that the previous knowledge about the problem can be merged in order to get the parameter estimation carried on. This property is specially well suit when the information given by the data can be blur, as it usually happens in many image analysis problems.