Modern Digital Communications: A Hands-On Approach

MMSE Estimation and LS Approximation

Dr. Nicolae Chiurtu

- Course material of Prof. Bixio Rimoldi -

Last revision: Nov. 27, 2023

Minimum Mean Squared Error Estimation

In this note, we study the concept of Minimum Mean Squared Error (MMSE) Estimation. We start with a general description of the problem, then focus on the special case of MMSE estimation for jointly Gaussian random vectors — which is what we need in order to estimate the OFDM channel.

Subsequently we study Least Squares (LS) approximations, which will also be used in the assignments.

Estimation

We are interested in estimating the realization $\boldsymbol{x}$ of the random variable $\boldsymbol{X} \in \mathbb{C}^m$.

For this, we have access to the realization $\boldsymbol{y}$ of $\boldsymbol{Y} \in \mathbb{C}^n$. (Here we use small letters to denote the realization of random variables). Unless $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent, knowing $\boldsymbol{y}$ should help estimating $\boldsymbol{x}$.

An $estimator$ of $\boldsymbol{X} \in \mathbb{C}^m$ based on $\boldsymbol{Y} \in \mathbb{C}^n$ can be an arbitrary function $\hat{\boldsymbol{X}} : \mathbb{C}^n \to \mathbb{C}^m$. We then let $\hat{\boldsymbol{x}} = \hat{\boldsymbol{X}}(\boldsymbol{y})$ be the estimate of $\boldsymbol{x}$ based on the observed $\boldsymbol{y}$. Clearly this estimate may or may not be a good one.

Thus we need a way to measure the performance of an estimator, and hopefully we can find an estimator which is optimal under that performance criterion.

MMSE Estimator

A popular choice for the fidelity criterion is the *mean squared error*

$$\mathrm{MSE}(\hat{\boldsymbol{X}}) \coloneqq E\left[\left\|\boldsymbol{X} - \hat{\boldsymbol{X}}(\boldsymbol{Y})\right\|^2\right]$$

where the expectation is over the joint distribution of $\boldsymbol{X}$ and $\boldsymbol{Y}$.

An estimator that minimizes the mean squared error is called a *minimum mean squared error (MMSE) estimator* of $\boldsymbol{X}$ given $\boldsymbol{Y}$.

We denote by $\hat{\boldsymbol{X}}_{\mathrm{MMSE}}(\boldsymbol{y})$ such an estimator. In other words, for any estimator $\hat{\boldsymbol{X}}(\boldsymbol{y})$,

$$E\left[\left\|\boldsymbol{X} - \hat{\boldsymbol{X}}(\boldsymbol{Y})\right\|^2\right] \geq E\left[\left\|\boldsymbol{X} - \hat{\boldsymbol{X}}_{\mathrm{MMSE}}(\boldsymbol{Y})\right\|^2\right].$$

The MMSE estimator of $\boldsymbol{X}$ given $\boldsymbol{Y}$ turns out to be $unique$ and equal to the $conditional\ expectation$ of $\boldsymbol{X}$ given $\boldsymbol{Y}$, i.e.,

$$\hat{\boldsymbol{X}}_{\mathrm{MMSE}}(\boldsymbol{y}) = E\left[\boldsymbol{X}|\boldsymbol{Y}=\boldsymbol{y}\right].$$

Derivation of the MMSE Estimator

We are looking for an estimator $\hat{\boldsymbol{X}} : \mathbb{C}^n \to \mathbb{C}^m$ such that

$$\mathrm{MSE}(\hat{\boldsymbol{X}}) = E\left[\left\|\boldsymbol{X} - \hat{\boldsymbol{X}}(\boldsymbol{Y})\right\|^2\right],$$

is minimized.

We are done if we prove that $\hat{\boldsymbol{X}}_{\mathrm{MMSE}}(\boldsymbol{y})$ minimizes

$$E_y\left[\left\|\boldsymbol{X} - \hat{\boldsymbol{X}}(\boldsymbol{Y})\right\|^2\right] \stackrel{\triangle}{=} E\left[\left\|\boldsymbol{X} - \hat{\boldsymbol{X}}(\boldsymbol{Y})\right\|^2 \middle| \boldsymbol{Y} = \boldsymbol{y}\right],$$

where E_y means that we are taking the expectation conditioning on $\boldsymbol{Y} = \boldsymbol{y}$.

We have

$$E_y\Big[\big\|\boldsymbol{X}-\hat{\boldsymbol{X}}(\boldsymbol{Y})\big\|^2\Big] = E_y\Big[\big\|\boldsymbol{X}\big\|^2\Big] + \big\|\hat{\boldsymbol{X}}(\boldsymbol{y})\big\|^2 - 2\Re\left\{E_y\Big[\boldsymbol{X}^\dagger\Big]\hat{\boldsymbol{X}}(\boldsymbol{y})\right\}.$$

The first term on the RHS of the equality does not depend on $\hat{\boldsymbol{X}}$. If we substitute it for another term that does not depend on $\hat{\boldsymbol{X}}$, then the minimizing $\hat{\boldsymbol{X}}$ remains the same.

We choose to minimize

$$\big\|E_y\big[\boldsymbol{X}\big]\big\|^2 + \big\|\hat{\boldsymbol{X}}(\boldsymbol{y})\big\|^2 - 2\Re\left\{E_y\Big[\boldsymbol{X}^\dagger\Big]\hat{\boldsymbol{X}}(\boldsymbol{y})\right\},$$

which is the quadratic form

$$\big\|E_y\big[\boldsymbol{X}\big] - \hat{\boldsymbol{X}}(\boldsymbol{y})\big\|^2.$$

The above is non-negative, and it achieves its minimum iff $\hat{\boldsymbol{X}}(\boldsymbol{y}) = E_y\big[\boldsymbol{X}\big]$.

□

MMSE Estimator for Jointly Gaussian Random Vectors

In general, finding an expression for $E[\boldsymbol{X}|\boldsymbol{Y}=\boldsymbol{y}]$ is not a trivial task.

However, when $\boldsymbol{X}$ and $\boldsymbol{Y}$ are jointly Gaussian random vectors, the MMSE estimator has a simple linear form.

In the rest of this lecture we restrict our discussion to jointly Gaussian random vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$.

First we assume that $\boldsymbol{X}$ and $\boldsymbol{Y}$ are zero-mean vectors. Later on, we will take non-zero means into account.

A Fact About Jointly Gaussian Random Vectors

If $\boldsymbol{X} \in \mathbb{C}^m$ and $\boldsymbol{Y} \in \mathbb{C}^n$ are zero-mean jointly Gaussian random vectors, we can always write

$$\boldsymbol{X} = A\boldsymbol{Y} + \boldsymbol{Z}$$

for some matrix $A \in \mathbb{C}^{m \times n}$ (that we will determine shortly) and a zero-mean Gaussian vector $\boldsymbol{Z} \sim \mathcal{N}_{\mathcal{C}}(\boldsymbol{0}, K_{\boldsymbol{Z}})$ that is independent of $\boldsymbol{Y}$.

Proof. For any matrix $A \in \mathbb{C}^{m \times n}$,

$$\boldsymbol{Z} \coloneqq \boldsymbol{X} - A\boldsymbol{Y}$$

is a (zero-mean) complex Gaussian vector. (By definition, a linear combination of jointly Gaussian vectors leads to a Gaussian vector.)

Moreover,

$$E\left[\boldsymbol{Z}\boldsymbol{Y}^\dagger\right] = E\left[\boldsymbol{X}\boldsymbol{Y}^\dagger\right] - A\,E\left[\boldsymbol{Y}\boldsymbol{Y}^\dagger\right] = K_{\boldsymbol{XY}} - A\,K_{\boldsymbol{Y}}$$

Hence, if we take

$$A = K_{\boldsymbol{XY}} K_{\boldsymbol{Y}}^{-1},$$

we will have

$$E\left[\boldsymbol{Z}\boldsymbol{Y}^{\dagger}\right] = 0,$$

i.e., with this specific choice of A , $\boldsymbol{Z}$ is independent of $\boldsymbol{Y}$. □

Note that with this choice of A,

$$\begin{aligned} K_{\boldsymbol{Z}} &= E\left[(\boldsymbol{X} - A\boldsymbol{Y})(\boldsymbol{X} - A\boldsymbol{Y})^{\dagger}\right] \\ &= K_{\boldsymbol{X}} - AK_{\boldsymbol{YX}} - (K_{\boldsymbol{XY}} - AK_{\boldsymbol{Y}})A^{\dagger} \\ &= K_{\boldsymbol{X}} - AK_{\boldsymbol{YX}} \\ &= K_{\boldsymbol{X}} - K_{\boldsymbol{XY}} K_{\boldsymbol{Y}}^{-1} K_{\boldsymbol{XY}}^{\dagger}, \end{aligned}$$

where in the third line we used the fact that $AK_{\boldsymbol{Y}} = K_{\boldsymbol{XY}}$.

MMSE Estimator for Jointly Gaussian Vectors

Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be jointly Gaussian. Then

$$\hat{\boldsymbol{X}}_{\mathrm{MMSE}}(\boldsymbol{y}) = K_{\boldsymbol{XY}} K_{\boldsymbol{Y}}^{-1} \boldsymbol{y}.$$

Moreover, the minimum mean squared error equals

$$E\left[\left\|\boldsymbol{X} - \hat{\boldsymbol{X}}_{\mathrm{MMSE}}(\boldsymbol{Y})\right\|^2\right] = \mathrm{trace}\big(K_{\boldsymbol{X}} - K_{\boldsymbol{XY}} K_{\boldsymbol{Y}}^{-1} K_{\boldsymbol{XY}}^{\dagger}\big).$$

Proof. Write

$$\boldsymbol{X} = A\boldsymbol{Y} + \boldsymbol{Z} \quad \text{with} \quad A = K_{\boldsymbol{XY}} K_{\boldsymbol{Y}}^{-1},$$

which makes $\boldsymbol{Z}$ independent of $\boldsymbol{Y}$.

Now

$$\begin{aligned} \hat{\boldsymbol{X}}_{\text{MMSE}}(\boldsymbol{y}) &= E\left[\boldsymbol{X} | \boldsymbol{Y} = \boldsymbol{y}\right] \\ &= E\left[A\boldsymbol{Y} + \boldsymbol{Z} | \boldsymbol{Y} = \boldsymbol{y}\right] \\ &= E\left[A\boldsymbol{y} + \boldsymbol{Z}\right] = A\boldsymbol{y}. \end{aligned}$$

The minimum mean squared error is

$$\begin{aligned} E\left[\left\|\boldsymbol{X} - \hat{\boldsymbol{X}}_{\text{MMSE}}(\boldsymbol{Y})\right\|^2\right] &= E\left[\left\|(A\boldsymbol{Y} + \boldsymbol{Z}) - A\boldsymbol{Y}\right\|^2\right] \\ &= E\left[\left\|\boldsymbol{Z}\right\|^2\right] = \text{trace}(K_{\boldsymbol{Z}}) = \text{trace}\big(K_{\boldsymbol{X}} - K_{\boldsymbol{XY}} K_{\boldsymbol{Y}}^{-1} K_{\boldsymbol{XY}}^{\dagger}\big). \end{aligned}$$

□

Channel Coefficients Estimation

We have seen that by using OFDM we 'forge' parallel channels described in matrix form as

$$\boldsymbol{Y}^{(m)} = D\boldsymbol{A}^{(m)} + \boldsymbol{Z}^{(m)}$$

where D is the diagonal matrix of channel coefficients. How to estimate the matrix D?

For certain values of m, we substitute $\boldsymbol{A}^{(m)}$ with an N-tuple $\boldsymbol{S}$ known to the receiver. Then, dropping the superscript (m) for notational convenience,

$$\boldsymbol{Y} = D\boldsymbol{S} + \boldsymbol{Z}$$

where D is diagonal. The same result is obtained by

$$\boldsymbol{Y} = S\boldsymbol{D} + \boldsymbol{Z}$$

where S is the diagonal matrix that has $\boldsymbol{S}$ as its diagonal elements and $\boldsymbol{D}$ is the N-tuple consisting of the diagonal elements of D.

Assume that $\boldsymbol{D}$ is zero-mean, Gaussian, and independent of $\boldsymbol{Z}$.

Then $\boldsymbol{Y}$ and $\boldsymbol{D}$ are jointly Gaussian. To see this, write

$$\begin{pmatrix} \boldsymbol{Y} \\ \boldsymbol{D} \end{pmatrix} = \begin{pmatrix} S & I \\ I & 0 \end{pmatrix} \begin{pmatrix} \boldsymbol{D} \\ \boldsymbol{Z} \end{pmatrix}.$$

This shows that the vector on the left hand side is a linear transformation of a Gaussian random vector.

Hence, the minimum mean squared error (MMSE) estimate of $\boldsymbol{D}$ based on the observable $\boldsymbol{Y}$ is

$$\hat{\boldsymbol{D}} = K_{\boldsymbol{DY}} K_{\boldsymbol{Y}}^{-1} \boldsymbol{Y},$$

where

$$\begin{aligned} K_{\boldsymbol{DY}} &\coloneqq E[\boldsymbol{D}\boldsymbol{Y}^\dagger] = K_{\boldsymbol{D}} S^\dagger \quad \text{and} \\ K_{\boldsymbol{Y}} &\coloneqq E[\boldsymbol{Y}\boldsymbol{Y}^\dagger] = S K_{\boldsymbol{D}} S^\dagger + K_{\boldsymbol{Z}} \end{aligned}$$

are covariance matrices and $\boldsymbol{D}$ and $\boldsymbol{Y}$ are zero-mean.

Both $K_{\boldsymbol{Y}}$ and $K_{\boldsymbol{DY}}$ depend on $K_{\boldsymbol{D}}$. The question is how to find $K_{\boldsymbol{D}}$.

K_D from the Channel Model

To determine K_D, we make a channel model.

A reasonable assumption for a wireless channel is the multipath channel, modeled as

$$h(t) = \sum_{l=0}^{M-1} \alpha_l \delta(t - \tau_l),$$

where α_l and τ_l are the l-th path strength and delay, respectively.

Since α_l and α_k, $l \neq k$, describe the reflection on different obstacles/surfaces, it is reasonable to assume that they are uncorrelated.

For this channel,

$$h_{\mathcal{F}}(f) = \int h(t) e^{-j2\pi ft} dt = \sum_l \alpha_l \int \delta(t - \tau_l) e^{-j2\pi ft} dt = \sum_l \alpha_l e^{-j2\pi f\tau_l}.$$

Assuming that the DFT-length N is even (usually a power of 2), $\boldsymbol{D} = (\lambda_0, \ldots, \lambda_{N-1})$ with

$$\lambda_i = \begin{cases} h_{\mathcal{F}}\left(\frac{i}{NT_s}\right), & i = 0, \ldots, \frac{N}{2} - 1 \\ h_{\mathcal{F}}\left(\frac{i-N}{NT_s}\right), & i = \frac{N}{2}, \ldots, N-1 \end{cases},$$

where

$$h_{\mathcal{F}}\left(\frac{i}{NT_s}\right) = \sum_l \alpha_l e^{-j2\pi \frac{i}{NT_s}\tau_l},$$

and T_s is both the symbol interval and the sampling interval.

If we define

$$[i] = \begin{cases} i, & i = 0, \ldots, \frac{N}{2} - 1 \\ i - N, & i = \frac{N}{2}, \ldots, N-1 \end{cases},$$

then we can write

$$\lambda_i = h_{\mathcal{F}}\left(\frac{[i]}{NT_s}\right).$$

The (i,k) entry of the covariance matrix $K_{\boldsymbol{D}}$ is

$$(K_{\boldsymbol{D}})_{i,k} = E\left[\lambda_i \lambda_k^*\right] = E\left[\sum_l \alpha_l e^{-j2\pi \frac{[i]}{NT_s}\tau_l} \sum_v \alpha_v^* e^{j2\pi \frac{[k]}{NT_s}\tau_v}\right]$$
$$= \sum_l \sum_v e^{-j2\pi \frac{[i]}{NT_s}\tau_l} E\left[\alpha_l \alpha_v^*\right] e^{j2\pi \frac{[k]}{NT_s}\tau_v} \qquad (1)$$

The above expression seems complicated but it is actually not. In fact, K_D is the product of three matrices. To see this, notice that if A, B, C are matrices such that the product ABC is well defined, then

$$(ABC)_{i,k} = \sum_l \sum_v A_{i,l} B_{l,v} C_{v,k},$$

which is exactly the right-hand side of (1) with

$$A_{i,l} = e^{-j2\pi \frac{[i]}{NT_s}\tau_l}, \qquad B_{l,v} = E\left[\alpha_l \alpha_v^*\right], \qquad C = A^\dagger.$$

From $K_{\boldsymbol{D}}$, we determine $K_{\boldsymbol{Y}}$ and $K_{\boldsymbol{DY}}$ as described earlier.

Least Squares (LS) Approximation

An MMSE estimator requires the knowledge of the statistics.

Suppose that all we know is that the observable $\boldsymbol{y} \in \mathbb{C}^n$ is obtained from $\boldsymbol{\lambda} \in \mathbb{C}^n$ according to

$$\boldsymbol{y} = S\boldsymbol{\lambda} + \boldsymbol{z},$$

where S is a diagonal matrix with non-vanishing diagonal elements, and $\boldsymbol{z} \in \mathbb{C}^n$ is a noise vector. None of the statistics are known.

We would like to find the $\hat{\boldsymbol{\lambda}} \in \mathbb{C}^n$ for which $\|\boldsymbol{y} - S\boldsymbol{\lambda}\|^2$ is minimized over all $\boldsymbol{\lambda} \in \mathbb{C}^n$.

By writing

$$\|\boldsymbol{y} - S\boldsymbol{\lambda}\|^2 = \sum_{i=0}^{n-1} |y_i - S_i\lambda_i|^2,$$

it is clear that $\hat{\lambda}_i$ needs to be chosen to minimize $|y_i - S_i\lambda_i|^2$.

Clearly the minimum is obtained when

$$\hat{\lambda}_i = \frac{y_i}{S_i},$$

in which case $\|\boldsymbol{y} - S\boldsymbol{\lambda}\|^2 = 0$.

This is a special case that we have worked out since we will need it in the assignment.

More generally, suppose that

$$\boldsymbol{y} = S\boldsymbol{\lambda} + \boldsymbol{z},$$

where $\boldsymbol{y}$ and $\boldsymbol{z}$ are in $\mathbb{C}^n$, $\boldsymbol{\lambda} \in \mathbb{C}^m$, and $S \in \mathbb{C}^{n \times m}$ is a general full-rank matrix (not necessarily diagonal).

As before, we are seeking the $\hat{\boldsymbol{\lambda}}$ that minimizes

$$\|\boldsymbol{y} - S\boldsymbol{\lambda}\|^2$$

over all $\boldsymbol{\lambda} \in \mathbb{C}^m$.

If $n \leq m$, the system $\boldsymbol{y} = S\boldsymbol{\lambda}$ can always be solved. Hence we can always find a $\boldsymbol{\lambda}$ for which $\|\boldsymbol{y} - S\boldsymbol{\lambda}\|^2 = 0$.

So we focus on the case where $n > m$ (the system is overdetermined).

We may reformulate the problem as follows. The observable $\boldsymbol{y}$ is an element of the inner-product space $\mathcal{U} = \mathbb{C}^n$ and let $\mathcal{V}$ be the subspace spanned by the columns of S. We are seeking the vector $\hat{\boldsymbol{y}} \in \mathcal{V}$ that minimizes

$$\|\boldsymbol{y} - \hat{\boldsymbol{y}}\|^2.$$

The projection theorem (see PDC) tells us that $\hat{\boldsymbol{y}}$ is the projection of $\boldsymbol{y}$ into $\mathcal{V}$. It has the property that the error vector $\boldsymbol{y} - \hat{\boldsymbol{y}}$ is orthogonal to every element of $\mathcal{V}$. In particular, it is orthogonal to the columns of S. Hence

$$\langle \boldsymbol{y} - S\hat{\boldsymbol{\lambda}}, \boldsymbol{S}_i \rangle = 0, \quad i = 1, \ldots, m$$

where $\boldsymbol{S}_i$ is the i-th column of S.

Hence

$$\langle \boldsymbol{y}, \boldsymbol{S}_i \rangle = \langle S\hat{\boldsymbol{\lambda}}, \boldsymbol{S}_i \rangle, \quad i = 1, \ldots, m. \tag{2}$$

Equivalently,

$$\boldsymbol{S}_i^\dagger \boldsymbol{y} = \boldsymbol{S}_i^\dagger S\hat{\boldsymbol{\lambda}} \quad i = 1, \ldots, m.$$

In matrix form, we obtain the so-called $normal\ equations$:

$$S^\dagger \boldsymbol{y} = S^\dagger S \hat{\boldsymbol{\lambda}}.$$

Since S has full rank, $S^\dagger S$ is nonsingular. We prove this by arguing that if $S^\dagger S$ is singular, then S does not have full rank. Indeed, if $S^\dagger S$ is singular, there exists a nonzero vector $\boldsymbol{u}$ such that $S^\dagger S\boldsymbol{u} = 0$, which implies that $\boldsymbol{u}^\dagger S^\dagger S\boldsymbol{u} = \|S\boldsymbol{u}\|^2 = 0$, so that $S\boldsymbol{u} = 0$. This means that the columns of S are linearly dependent, i.e., S is not full rank — a contradiction.

Solving for $\hat{\boldsymbol{\lambda}}$ yields the least-squares approximation of λ:

$$\hat{\boldsymbol{\lambda}} = \left(S^\dagger S\right)^{-1} S^\dagger \boldsymbol{y}.$$

To reconstruct the above formula from memory, it suffices to check the dimensions. The matrix that estimates $\boldsymbol{\lambda}$ from $\boldsymbol{y}$ must be of dimension $m \times n$. If we analyze $(S^\dagger S)^{-1} S^\dagger$ from right to left, we see that we don't have other choices (unless we make the expression more complicated). In particular, $S^\dagger$ has the right dimension to multiply $\boldsymbol{y}$ and notice that $(S^\dagger S)^{-1}$ is $m \times m$, hence it is a valid matrix to multiply the $m \times n$ matrix $S^\dagger$, whereas $(S S^\dagger)^{-1}$ would not fit as it is $n \times n$.

To summarize, recall the fundamental difference between the MMSE estimate and the LS approximation. On the one hand, the MMSE setup assumes two random vectors, the unobserved $\boldsymbol{X}$ and the observed $\boldsymbol{Y}$, and the objective is to estimate $\boldsymbol{X}$ by means of $\boldsymbol{Y}$, where the measure of performance is $E[\|\boldsymbol{X} - \hat{\boldsymbol{X}}(\boldsymbol{Y})\|^2]$.

On the other hand, the LS approximation is about adjusting the parameters of a model function to best fit the observed data. (See the example that follows.) It is also called data fitting or regression analysis. (See e.g. Least Squares in Wikipedia).

Example: Suppose that the model function is $y(x) = \alpha_1 + \alpha_2 x + \alpha_3 x^2$, with unknown parameters α_1, α_2, α_3. We seek to find the parameters so that the model fits the noisy observations

$$y_i = y(x_i) + z_i, \quad i = 1, 2, \ldots, n.$$

Letting $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^T$, $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)^T$, $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3)^T$, and $\boldsymbol{z} = (z_1, z_2, \ldots, z_n)^T$, the observations are of the form

$$\boldsymbol{y} = S\boldsymbol{\alpha} + \boldsymbol{z},$$

where

$$S = \begin{pmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{pmatrix}.$$

It is worth pointing out that even if we are estimating a nonlinear function of x, we are dealing with a data observation model where x has fixed values and the variable is the parameter vector $\boldsymbol{\alpha}$. As a function of $\boldsymbol{\alpha}$, the observation model is linear.

The $\boldsymbol{\alpha}$ that minimizes $\|\boldsymbol{y} - \hat{\boldsymbol{y}}(\boldsymbol{x})\|^2$ is the LS approximation

$$\hat{\boldsymbol{\alpha}} = (S^\dagger S)^{-1} S^\dagger \boldsymbol{y}.$$

For a MATLAB example, check out the file example.m