ECE 251A: Digital Signal Processing I
Linear Minimum Mean Squared Estimation (LMMSE)

Florian Meyer

Electrical and Computer Engineering Department
University of California San Diego

UC San Diego
JACOBS SCHOOL OF ENGINEERING

Bayesian Options

Problem: Estimate random vector $\mathbf{Y}$ given measurements of random vector $\mathbf{X}$

- Posterior Density Estimation ($p(\mathbf{y}|\mathbf{x})$)
- Maximum Aposteriori Estimation (MAP) (find peak of $p(\mathbf{y}|\mathbf{x})$)
- Minimum Mean Squared Estimation (MMSE) ($E(\mathbf{Y}|\mathbf{x})$)
- Linear Minimum Mean Squared Estimation (LMMSE)

Minimum Mean Squared Estimation (MMSE)

Objective: Compute an estimate of $\mathbf{Y}$ as $\hat{\mathbf{Y}} = g(\mathbf{X})$ to minimize the mean squared error $E_{X,Y}(\|\mathbf{Y} - \hat{\mathbf{Y}}\|^2)$

Optimum minimum mean squared estimate is given by the conditional mean

$$\hat{\mathbf{y}}_{mmse} = E(\mathbf{Y}|\mathbf{X} = \mathbf{x}) = \int \mathbf{y} p(\mathbf{y}|\mathbf{x}) d\mathbf{y}$$

Challenge: Need knowledge of the conditional density $p(\mathbf{y}|\mathbf{x})$.

LMMSE Estimation

Linear Minimum Mean Squared Estimation:Estimate is constrained to be linear estimate $\hat{\mathbf{Y}} = C^H\mathbf{X}$ and C is chosen to the minimize the MSE.

Quite general: No restriction on the dimension of the random vectors.

Will consider $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, .., \mathbf{X}_M]^T$, a $M \times 1$ vector.

Will consider $\mathbf{Y} = [\mathbf{Y}_1, \mathbf{Y}_2, .., \mathbf{Y}_L]^T$, a $L \times 1$ vector.

Assumption $\mathbf{X}$ has a $M \times M$ correlation matrix $\mathbf{R}_{xx} = E(\mathbf{X}\mathbf{X}^H) = \mathbf{R}_{xx}^H$

$\mathbf{Y}$ has $L \times L$ correlation matrix $\mathbf{R}_{yy} = E(\mathbf{Y}\mathbf{Y}^H) = \mathbf{R}_{yy}^H$

The cross correlation matrix is denoted by $\mathbf{R}_{yx} = E(\mathbf{Y}\mathbf{X}^H) = \mathbf{R}_{xy}^H$. $\mathbf{R}_{yx}$ is a $L \times M$ matrix and $\mathbf{R}_{xy}$ is a $M \times L$ matrix.

Optimum Linear Estimate

$\hat{\mathbf{Y}} = C^H\mathbf{X}$. C^H is $L \times M$ matrix or C is $M \times L$ matrix.

Objective: Choose C to minimize the MSE:
$P(C) = E(\|\mathbf{Y} - \hat{\mathbf{Y}}\|^2) = E(\|\mathbf{Y} - C^H\mathbf{X}\|^2)$.

Solution: $C_o^H = \mathbf{R}_{yx}\mathbf{R}_{xx}^{-1}$, or $C_o = \mathbf{R}_{xx}^{-1}\mathbf{R}_{xy}$, where $\mathbf{R}_{yx} = E(\mathbf{Y}\mathbf{X}^H)$ and $\mathbf{R}_{xx} = E(\mathbf{X}\mathbf{X}^H)$

Note that $\mathbf{R}_{xy} = \mathbf{R}_{yx}^H$

Alternate forms

$$\mathbf{R}_{xx}C_o = \mathbf{R}_{xy} \ \ \text{(Normal Equations)}$$

$$\mathbf{R}_{xx}C_o - \mathbf{R}_{xy} = \mathbf{0} \ \text{ or } \ \mathbf{R}_{yx} - C_o^H\mathbf{R}_{xx} = \mathbf{0}$$

LMMSE Estimate of $\mathbf{Y}$

$$\hat{\mathbf{Y}}_o \stackrel{\text{def}}{=} \hat{\mathbf{Y}}_{lmmse} = C_o^H\mathbf{X} = \mathbf{R}_{yx}\mathbf{R}_{xx}^{-1}\mathbf{X}$$

Optimum Mean Squared Error

Error Covariance: $\tilde{\mathbf{Y}}_o = \mathbf{Y} - \hat{\mathbf{Y}}_o$ has Correlation matrix matrix given by

$$\mathbf{P}_o = E(\tilde{\mathbf{Y}}_o \tilde{\mathbf{Y}}_o^H) = \mathbf{R}_{yy} - \mathbf{R}_{yx}\mathbf{R}_{xx}^{-1}\mathbf{R}_{xy}$$

Mean Squared Error

$$P(C_o) = \mathrm{Tr}\left(\mathbf{R}_{yy} - \mathbf{R}_{yx}\mathbf{R}_{xx}^{-1}\mathbf{R}_{xy}\right)$$

where Tr denotes the trace operator.

Augmented Form of the Equation

If $\mathbf{Y}$ is a scalar, then C_o and $\mathbf{R}_{xy}$ are $M \times 1$ vectors and the optimal solution $C_o^H = \mathbf{R}_{yx}\mathbf{R}_{xx}^{-1}$ involves solving the following system of M equations

$$\mathbf{R}_{xx} C_o = \mathbf{R}_{xy} \text{ or } \mathbf{R}_{xy} - \mathbf{R}_{xx} C_o \text{ or } [\mathbf{R}_{xy}, \mathbf{R}_{xx}] \begin{bmatrix} 1 \\ -C_0 \end{bmatrix} = \mathbf{0} \tag{1}$$

Similarly the error covariance matrix $P_o \stackrel{\text{def}}{=} E(\tilde{\mathbf{Y}}_o \tilde{\mathbf{Y}}_o^H)$ is a scalar and $P_o = \mathbf{R}_{yy} - \mathbf{R}_{yx} C_o$ can be written in the form

$$[\mathbf{R}_{yy}, \mathbf{R}_{yx}] \begin{bmatrix} 1 \\ -C_0 \end{bmatrix} = P_o \tag{2}$$

Combining (2) and (1), we get the augmented form which consists of $(M+1)$ equations

$$\begin{bmatrix} \mathbf{R}_{yy} & \mathbf{R}_{yx} \\ \mathbf{R}_{xy} & \mathbf{R}_{xx} \end{bmatrix} \begin{bmatrix} 1 \\ -C_0 \end{bmatrix} = \begin{bmatrix} P_o \\ \mathbf{0} \end{bmatrix}$$

Formulation as L Scalar Estimation Problems

If we denote $C = [C_1, C_2, \ldots, C_L]$ where C_l is the lth column of C. The MSE can be written as

$$P(C) = E(\|\mathbf{Y} - C^H \mathbf{X}\|^2) = \sum_{l=1}^{L} E(|\mathbf{Y}_l - C_l^H \mathbf{X}|^2)$$

This can be viewed as L separate problems and $\hat{\mathbf{Y}}_{l,o} = C_{l,o}^H \mathbf{X}$, where $C_{l,o}$ is the lth column of C_o.

How to find the optimal value for C? Find the columns C_l and then arrange them in matrix form to obtain C.

Can use Wirtinger calculus to deal with this situation.

Deriving the Optimal estimator coefficients

Scalar $\mathbf{Y}_l$ and vector $\mathbf{X}$. Find C_l to minimize MSE

$$P(C_l) = E(|\mathbf{Y}_l - C_l^H \mathbf{X}|^2) = E(|\mathbf{Y}_l|^2) - C_l^H \mathbf{R}_{xy_l} - \mathbf{R}_{y_l x} C_l + C_l^H \mathbf{R}_{xx} C_l$$

Using Wirtinger calculus

$$\nabla_{C_l^*}(P(C_l)) = 0 - \mathbf{R}_{xy_l} - 0 + \mathbf{R}_{xx} C_l$$

Setting derivative to zero we get the normal equations

$$\mathbf{R}_{xx} C_{l,o} = \mathbf{R}_{xy_l}$$

With proper arrangement, we have

$$\mathbf{R}_{xx} [C_{1,o}, C_{2,o}, \ldots, C_{L,o}] = [\mathbf{R}_{xy_1}, \mathbf{R}_{xy_2}, \ldots, \mathbf{R}_{xy_L}]$$

This leads to

$$\mathbf{R}_{xx} C_o = \mathbf{R}_{xy}$$

Derivation of the Minimum Mean Squared Error

Useful Observation:
$E\|\mathbf{Y}\|^2 = E(\mathbf{Y}^H\mathbf{Y}) = E(\mathrm{Tr}(\mathbf{Y}^H\mathbf{Y})) = E(\mathrm{Tr}\mathbf{Y}\mathbf{Y}^H) = \mathrm{Tr}E(\mathbf{Y}\mathbf{Y}^H)$
The diagonal entries of $E(\mathbf{Y}\mathbf{Y}^H)$ are $E(|\mathbf{Y}_l|^2)$

Using the above relationship

$$
\begin{aligned}
P(C_o) = E\|\tilde{\mathbf{Y}}_o\|^2 &= \mathrm{Tr}E(\tilde{\mathbf{Y}}_o\tilde{\mathbf{Y}}_o^H) \\
&= \mathrm{Tr}E(\mathbf{Y} - C_o^H\mathbf{X})(\mathbf{Y} - C_o^H\mathbf{X})^H \\
&= \mathrm{Tr}E(\mathbf{Y} - C_o^H\mathbf{X})(\mathbf{Y}^H - \mathbf{X}^H C_o) \\
&= \mathrm{Tr}[\mathbf{R}_{yy} + C_o^H\mathbf{R}_{xx}C_o - \mathbf{R}_{yx}C_o - C_o^H\mathbf{R}_{xy}] \\
&= \mathrm{Tr}\left[(\mathbf{R}_{yy} - \mathbf{R}_{yx}\mathbf{R}_{xx}^{-1}\mathbf{R}_{xy}\right)
\end{aligned}
$$

The last equation was obtained by substituting $C_o = \mathbf{R}_{xx}^{-1}\mathbf{R}_{xy}$.
The error covariance matrix is given by

$$
\mathbf{P}_o \stackrel{\mathrm{def}}{=} E(\tilde{\mathbf{Y}}_o\tilde{\mathbf{Y}}_o^H) = \mathbf{R}_{yy} - \mathbf{R}_{yx}\mathbf{R}_{xx}^{-1}\mathbf{R}_{xy} = \mathbf{R}_{yy} - \mathbf{R}_{yx}C_o = \mathbf{R}_{yy} - C_o^H\mathbf{R}_{xy}
$$

Computing C_o

Consider Scalar $\mathbf{Y}$, and then C_o is a $M \times 1$ vector.

Decompose $\mathbf{R}_{xx}$ using a LU (Cholesky) factorization, i.e. $\mathbf{R}_{xx} = \mathbf{LDL}^H$, where $\mathbf{L}$ is a lower triangular matrix with ones along the diagonal and $\mathbf{D}$ is a diagonal matrix.

$$\mathbf{R}_{xx} C_o = \mathbf{R}_{xy} \rightarrow \mathbf{LDL}^H C_o = \mathbf{R}_{xy} \rightarrow \mathbf{L} \overbrace{\mathbf{D} \underbrace{\mathbf{L}^H C_o}_{B_o}}^{A_o} = \mathbf{R}_{xy}$$

Compute A_o using the system of equations $\mathbf{L} A_o = \mathbf{R}_{xy}$, where $A_o = \mathbf{DL}^H C_o$. Procedure involves forward-substitution

Having computed A_o, compute $B_o = \mathbf{D}^{-1} A_o$, where $B_o = \mathbf{L}^H C_o$.

Having computed B_o, compute C_o from $\mathbf{L}^H C_o = B_0$ by back-substitution.

Example

Signal corrupted by noise

$$x[n] = s[n] + v[n]$$

$s[n] \sim \mathcal{N}(s; 0, \sigma_s^2)$ and $v[n] \sim \mathcal{N}(v; 0, \sigma_v^2)$
$s[n]$ and $v[n]$ are independent

Having observed $x[n]$ what is the estimate of $y[n] = s[n]$?

All mean zero. $\mathbf{R}_{yy} = \sigma_s^2$, $\mathbf{R}_{xx} = \sigma_s^2 + \sigma_v^2$, and $\mathbf{R}_{yx} = \sigma_s^2$.

Using the LMMSE formula $\hat{s}[n] = \mathbf{R}_{yx}\mathbf{R}_{xx}^{-1}x[n] = \frac{\sigma_s^2}{\sigma_s^2+\sigma_v^2}x[n]$.

Variance in the estimate: $\mathbf{R}_{yy} - \mathbf{R}_{yx}\mathbf{R}_{xx}^{-1}\mathbf{R}_{xy} = \sigma_s^2 - \frac{\sigma_s^4}{\sigma_s^2+\sigma_v^2} = \frac{\sigma_s^2\sigma_v^2}{\sigma_s^2+\sigma_v^2}$

$$p(s|x) \sim \mathcal{N}\left(s; \frac{\sigma_s^2}{\sigma_s^2 + \sigma_v^2}x, \frac{\sigma_s^2\sigma_v^2}{\sigma_s^2 + \sigma_v^2}\right)$$

We have a complete characterization of the estimate.

Wiener Filtering used in speech enhancement and other applications.

Properties of $E^*(\mathbf{Y}|\mathbf{X}) \stackrel{\text{def}}{=} \hat{\mathbf{Y}}_o = \mathbf{R}_{yx}\mathbf{R}_{xx}^{-1}\mathbf{X}$

$E^*(\mathbf{Y}|\mathbf{X})$ is convenient notation to indicate linear estimate of $\mathbf{Y}$ given $\mathbf{X}$.

Will behave like the expectation operator.

- $\hat{\mathbf{Y}}_o$ is a biased estimate, i.e. $E(\hat{\mathbf{Y}}_o) = C_o^H E(\mathbf{X}) = C_o^H \mu_x \neq E(\mathbf{Y})$. Unbiased if all random vectors are zero mean.
- Orthogonality Property: $\tilde{\mathbf{Y}}_o \perp B\mathbf{X}, \forall B$, i.e $E(\tilde{\mathbf{Y}}_o(B\mathbf{X})^H) = \mathbf{0}$. Note: $E(\tilde{\mathbf{Y}}_o(B\mathbf{X})^H) = E(\tilde{\mathbf{Y}}_o\mathbf{X}^H)B^H$. Sufficient to show $E(\tilde{\mathbf{Y}}_o\mathbf{X}^H) = \mathbf{0}$.
- Estimation of linear transformed $\mathbf{Y}$: If $\mathbf{Q} = B\mathbf{Y}$, then $\hat{\mathbf{Q}}_o = B\hat{\mathbf{Y}}_o$. $(E^*(\mathbf{Q}|\mathbf{X}) = E^*(B\mathbf{Y}|\mathbf{X}) = BE^*(\mathbf{Y}|\mathbf{X}))$
- Estimation under linear transformation of the Observations $\mathbf{X}$ to $\mathbf{Z} = B\mathbf{X}$. If B is invertible, then $E^*(\mathbf{Y}|\mathbf{X}) = E^*(\mathbf{Y}|\mathbf{Z})$. (Suggests whitening or de-correlating)
- If $\mathbf{Y}$ and $\mathbf{X}$ are jointly Gaussian and zero mean, MMSE estimate = LMMSE estimate. $(E(\mathbf{Y}|\mathbf{X}) = E^*(\mathbf{Y}|\mathbf{X}))$

Orthogonality Property or Principle

Two random variables Z_1 and Z_2 are orthogonal if $E(Z_1 Z_2^*) = 0$. Two random vector $\mathbf{X}$ and $\mathbf{Y}$ are orthogonal ($\mathbf{X} \perp \mathbf{Y}$) if

$$E(\mathbf{X}_i \mathbf{Y}_j^*) = 0,\ \forall\ i, j \ \text{ or } \ E(\mathbf{X}\mathbf{Y}^H) = \mathbf{0}$$

Orthogonality Property: $\tilde{\mathbf{Y}}_o \perp B\mathbf{X},\ \ \forall B$.
Proof: $E(\tilde{\mathbf{Y}}_o (B\mathbf{X})^H) = E((\mathbf{Y} - C_o^H \mathbf{X})\mathbf{X}^H)B^H = (\mathbf{R}_{yx} - C_0^H \mathbf{R}_{xx})B^H = \mathbf{0}B^H = \mathbf{0}$

Orthogonality Principle: Instead of starting with the LMMSE criteria, choose C such that $\tilde{\mathbf{Y}} = \mathbf{Y} - C^H \mathbf{X} \perp B\mathbf{X},\ \ \forall B$

Solution: Same as the LMMSE estimate, i.e. $\mathbf{R}_{yx} - C_0^H \mathbf{R}_{xx} = \mathbf{0}$

Application of the Orthogonality Property

$\tilde{\mathbf{Y}}_o \perp B\mathbf{X}$

Special Cases of B

1. $B = \mathbf{I}$. This implies $\tilde{\mathbf{Y}}_o \perp \mathbf{X}$ (Error orthogonal to the data)
2. $B = C_o^H$. This implies $\tilde{\mathbf{Y}}_o \perp \hat{\mathbf{Y}}_o$ (Error orthogonal to the optimal linear estimate)

Consider a first order real AR process $x[n] = .9x[n-1] + w[n]$.
(know model parameters, i.e. $P = 1$, $a_1 = -0.9$, and σ_w^2.)

Problem of interest: Predict $x[n]$ given $x[n-1]$

Solution: $\mathbf{Y} = x[n]$ and $\mathbf{X} = x[n-1]$. Then $\mathbf{R}_{xx} = r[0]$ and $\mathbf{R}_{yx} = r[1]$. The optimal estimate is $\hat{x}[n] = c_o x[n-1]$ where the weight c_o is given by $r[0]c_o = r[-1]$.

Given that the process is AR, we can compute $r[0]$ and $r[1]$.

Example Continued

Same AR process, $x[n] = .9x[n-1] + w[n]$

Problem of interest: Predict $x[n]$ given $x[n-1], x[n-2], .., x[n-P]$
Solution: $\mathbf{Y} = x[n]$ and $\mathbf{X} = \mathbf{x}_P[n-1] = [x[n-1], x[n-2], \ldots, x[n-P]]^T$.
Then $\mathbf{R}_{xx}$ is a $P \times P$ Toeplitz correlation matrix and
$\mathbf{R}_{xy} = [r[-1], r[-2], .., r[-P]]^T$.
The optimal estimate is $\hat{x}[n] = C_o^H \mathbf{x}_P[n-1]$where the weight C_o is given by
$\mathbf{R}_{xx} C_o = \mathbf{R}_{xy}$.

Given that the process is AR, we can compute all the required correlations

Is this matrix inversion based approach necessary? Can we find a simpler approach?

Othogonality principle: $\hat{x}_o[n] = .9x[n-1]$ and $\tilde{x}_o[n] = w[n]$.

Since $w[n]$ is a zero mean white noise process, $\tilde{x}_o[n] = w[n] \perp B\mathbf{x}_P[n-1]$. The estimate is the same for all $P \geqslant 1$!

Estimating $\mathbf{Q}$ where $\mathbf{Q} = B\mathbf{Y}$

Want to show $\hat{\mathbf{Q}}_o = B\hat{\mathbf{Y}}_o$.

Note: $\mathbf{R}_{qx} = E(\mathbf{Q}\mathbf{X}^H) = E(B\mathbf{Y}\mathbf{X}^H) = B\mathbf{R}_{yx}$

$$\hat{\mathbf{Q}}_o = E^*(\mathbf{Q}|\mathbf{X}) = \mathbf{R}_{qx}\mathbf{R}_{xx}^{-1}\mathbf{X} = B\mathbf{R}_{yx}\mathbf{R}_{xx}^{-1}\mathbf{X} = B\hat{\mathbf{Y}}_o$$

Estimating $\mathbf{Y}$ based on $\mathbf{Z} = B\mathbf{X}$

Assumption B is invertible.

Note: $\mathbf{R}_{zz} = B\mathbf{R}_{xx}B^H$ and $\mathbf{R}_{zz}^{-1} = B^{-H}\mathbf{R}_{xx}^{-1}B^{-1}$

Also $\mathbf{R}yz = E(\mathbf{Y}\mathbf{Z}^H) = E(\mathbf{Y}\mathbf{X}^H)B^H = \mathbf{R}_{yx}B^H$.

$$
\begin{aligned}
\hat{\mathbf{Y}}_o &= E^*(\mathbf{Y}|\mathbf{Z}) = \mathbf{R}_{yz}\mathbf{R}_{zz}^{-1}\mathbf{Z} = \mathbf{R}_{yx}B^H B^{-H}\mathbf{R}_{xx}^{-1}B^{-1}B\mathbf{X} \\
&= \mathbf{R}_{yx}\mathbf{R}_{xx}^{-1}\mathbf{X} = E^*(\mathbf{Y}|\mathbf{X})
\end{aligned}
$$

Conclusion: $E^*(\mathbf{Y}|\mathbf{Z}) = E^*(\mathbf{Y}|\mathbf{X})$

Given $\mathbf{X}$, decorrelate $\mathbf{X}$ to get $\mathbf{Z}$, i.e. $\mathbf{R}_{zz}$ is a diagonal matrix, and use $\mathbf{Z}$ for estimation purposes.

Best Affine Estimate

Optimum estimate is constrained to be an affine estimate $\hat{\mathbf{Y}} = C^H\mathbf{X} + D$.

Assumption $\mathbf{X}$ has mean $E(\mathbf{X}) = \mu_x$ and Covariance
$\Sigma_{xx} = E(\mathbf{X} - \mu_x)(\mathbf{X} - \mu_x)^H = \Sigma_{xx}^H$

$\mathbf{Y}$ has mean $E(\mathbf{Y}) = \mu_y$ and Covariance $\Sigma_{yy} = E(\mathbf{Y} - \mu_y)(\mathbf{Y} - \mu_y)^H = \Sigma_{yy}^H$

The cross covariance is denoted by $\Sigma_{yx} = E(\mathbf{Y} - \mu_y)(\mathbf{X} - \mu_x)^H = \Sigma_{xy}^H$

Objective: Choose C and D to minimize

$$P(C, D) = E(\|\mathbf{Y} - C^H\mathbf{X} - D\|^2).$$

Optimal Affine Estimate

Solution: $C_o^H = \Sigma_{yx}\Sigma_{xx}^{-1}$, or $C_o = \Sigma_{xx}^{-1}\Sigma_{xy}$ and $D_o = \mu_y - C_o^H \mu_x$.
Best Affine Estimate

$$\hat{\mathbf{Y}}_{af} = \mu_y + C_o^H(\mathbf{X} - \mu_x) = \mu_y + \Sigma_{yx}\Sigma_{xx}^{-1}(\mathbf{X} - \mu_x)$$

Error $\tilde{\mathbf{Y}}_{af} = \mathbf{Y} - \hat{\mathbf{Y}}_{af}$ has Covariance matrix given by
$E(\tilde{\mathbf{Y}}_{af}\tilde{\mathbf{Y}}_{af}^H) = \Sigma_{yy} - \Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy}$

Properties

- For zero mean random variables $\hat{\mathbf{Y}}_{af} = \mathbf{R}_{yx}\mathbf{R}_{xx}^{-1}\mathbf{X} = \hat{\mathbf{Y}}_o$
- $\hat{\mathbf{Y}}_{af}$ is an unbiased estimate, i.e. $E(\tilde{\mathbf{Y}}_{af}) = \mathbf{0}$.
- $\tilde{\mathbf{Y}}_{af} \perp B\mathbf{X}$, i.e $E(\tilde{\mathbf{Y}}_{af}(B\mathbf{X})^H) = \mathbf{0}$
- If $\mathbf{Q} = B\mathbf{Y}$, then $\hat{\mathbf{Q}}_{af} = B\hat{\mathbf{Y}}_{af}$
- If $\mathbf{Y}$ and $\mathbf{X}$ are jointly Gaussian, MMSE estimate = Optimal Affine estimate