Statistical Signal Processing

Lecture 3

chapter 1: parameter estimation

stochastic parameters

- Bayes estimation: the MMSE, absolute value and uniform cost functions
- examples: Gaussian mean in Gaussian noise, Poisson process
- vector parameters
- Fischer Information Matrix
- Cramer-Rao lower bound on the MSE
- Linear and Affine MMSE estimation

Cramer-Rao Bound

- **Theorem (CRB for Stochastic Parameters)** *If the estimator $\widehat{\theta}(Y)$ of θ is unbiased, then the correlation matrix of the parameter estimation errors $\widetilde{\theta}$ is bounded below by the inverse of the information matrix:*

$$R_{\widetilde{\theta}\widetilde{\theta}} \;=\; E\,(\widehat{\theta}-\theta)(\widehat{\theta}-\theta)^T \geq J^{-1}$$

with equality ($\forall\theta\in\Theta$ or for one θ with $\frac{\partial b_{\widehat{\theta}}^T(\theta)}{\partial\theta}=0$) iff

$$\widehat{\theta}(Y)-\theta \;=\; J^{-1}\frac{\partial\,\ln\,f(\theta|Y)}{\partial\,\theta} \quad \textit{in m.s.} \quad \text{(in mean-square)}$$

An estimator that achieves the lower bound is called *efficient*.
(note: $A\geq B \;\Leftrightarrow\; A-B\geq 0$: positive semi-definite)

- *Proof:* We shall apply the lemma on Schur complements to a vector space of random vectors with inner product $<\;X,Y\;>=\;E\,XY^T$. In particular, we choose $X_1=\frac{\partial\,\ln\,f(\theta|Y)}{\partial\,\theta}$ and $X_2=\widehat{\theta}-\theta$. The Unit Cross Correlation lemma applies for an unbiased estimator and we find

$$E\begin{bmatrix}\frac{\partial\,\ln\,f(\theta|Y)}{\partial\,\theta}\\ \widehat{\theta}-\theta\end{bmatrix}\begin{bmatrix}\frac{\partial\,\ln\,f(\theta|Y)}{\partial\,\theta}\\ \widehat{\theta}-\theta\end{bmatrix}^T = \begin{bmatrix}J & I\\ I & R_{\widetilde{\theta}\widetilde{\theta}}\end{bmatrix} = \begin{bmatrix}R_{11} & R_{12}\\ R_{21} & R_{22}\end{bmatrix} \;\geq\; 0\;.$$

We have $R_{22}\geq R_{21}R_{11}^{-1}R_{12}$ with equality iff $X_2=R_{21}R_{11}^{-1}X_1$.

Cramer-Rao Bound (2)

- *Efficiency* condition: implication for the posterior distribution? We can integrate

$$\frac{\partial \ln f(\theta|Y)}{\partial \theta} = J\widehat{\theta}(Y) - J\theta$$

over θ to yield

$$\ln f(\theta|Y) = c(Y) + \widehat{\theta}^T(Y)\, J\, \theta - \frac{1}{2}\theta^T J\theta = c'(Y) - \frac{1}{2}(\theta - \widehat{\theta})^T J\, (\theta - \widehat{\theta})$$

where $c(Y)$ and $c'(Y)$ are scalar functions of Y. This implies that $f(\theta|Y)$ is Gaussian. Using the constraint $\int_\Theta f(\theta|Y)\, d\theta = 1$, we can determine the proper integration constant and we get

$$f(\theta|Y) = \sqrt{\frac{\det J}{(2\pi)^m}}\, \exp\left(-\frac{1}{2}(\theta - \widehat{\theta}(Y))^T J\, (\theta - \widehat{\theta}(Y))\right)$$

or in other words $f(\theta|Y) \leftrightarrow \mathcal{N}(\widehat{\theta}(Y), J^{-1})$. So the posterior distribution should be Gaussian *with constant covariance matrix*. In that case, the posterior mean (which may depend on the data) is an efficient estimator. Note that neither the prior distribution nor the conditional distribution $f(Y|\theta)$ need to be Gaussian for the posterior distribution to be Gaussian. Note also: $\widehat{\theta}(Y) = \widehat{\theta}_{MMSE}(Y)$.

Cramer-Rao Bound (3)

- *Additivity of the information matrix*: using Bayes' rule, we can write

$$J = -E\,\frac{\partial}{\partial\theta}\left(\frac{\partial\ \ln\ f(\theta|Y)}{\partial\theta}\right)^T = -E\,\frac{\partial}{\partial\theta}\left(\frac{\partial\ \ln\ f(\theta)}{\partial\theta}\right)^T - E\,\frac{\partial}{\partial\theta}\left(\frac{\partial\ \ln\ f(Y|\theta)}{\partial\theta}\right)^T = J_{prior} + J_Y$$

- If $f(\theta)$ is Gaussian then the corresponding information matrix J is the inverse of the covariance matrix C : $J = C^{-1}$.
- If the different data y_i are independent given θ, then

$$f(Y|\theta) = \prod_{i=1}^{n} f(y_i|\theta) \quad \Rightarrow \quad J_Y = \sum_{i=1}^{n} J_{y_i}$$

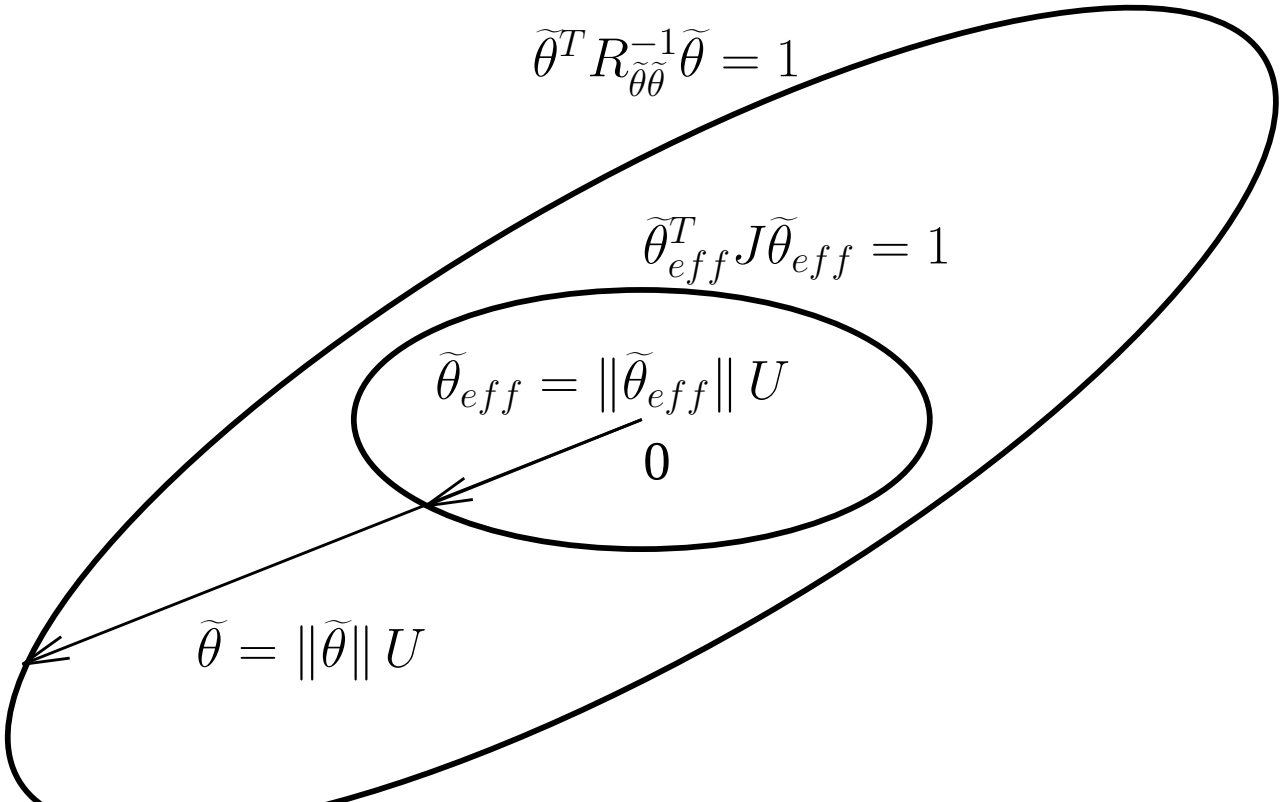
  If furthermore the data are i.i.d. given θ, then $J_Y = nJ_{y_1}$.
- $\widehat{\theta}_{MAP}$ is generally easier to determine than $\widehat{\theta}_{MMSE}$. If $\widehat{\theta}_{MAP}$ achieves efficiency (and hence is unbiased), then it equals $\widehat{\theta}_{MMSE}$ since $\widehat{\theta}_{MMSE}$ minimizes the MSE criterion but the minimum value of the MSE criterion cannot be lower than $\mathrm{tr}\left\{J^{-1}\right\}$.

  MSE = $E\,\|\widehat{\theta}(Y) - \theta\|^2 = \mathrm{tr} R_{\tilde{\theta}\tilde{\theta}}$ MMSE $\leftrightarrow$ $\min R_{\tilde{\theta}\tilde{\theta}} \geq J^{-1}$
- So in general, $R_{\tilde{\theta}\tilde{\theta}}^{MAP} \geq R_{\tilde{\theta}\tilde{\theta}}^{MMSE} \geq J^{-1}$
- $\widehat{\theta}_{MMSE}$ is unbiased: $\widehat{\theta} = E(\theta|Y) = E_{\boldsymbol{\theta}|\mathbf{Y}}\theta \ \Rightarrow\ E_{\mathbf{Y}}\widehat{\theta} = E_{\mathbf{Y}}E_{\boldsymbol{\theta}|\mathbf{Y}}\theta = E_{\boldsymbol{\theta},\mathbf{Y}}\theta$

Cramer-Rao Bound (4)

Concentration Ellipsoids

- Gaussian random vector: concentration ellipsoid = volume in which the random vector occurs which a certain probability.
- CRB $\Rightarrow$ the concentration ellipsoid for any unbiased estimator lies outside or on the concentration ellipsoid of an efficient estimator. This means that the estimation errors are the most concentrated in space (around the origin) for an efficient estimator.



Cramer-Rao Bound (5)

Example 1.6 Gaussian mean in Gaussian noise - Example 1.4 Continued

- In example 1.4, the posterior distribution was Gaussian with constant variance.
- We had $\widehat{\theta}_{MMSE} = \widehat{\theta}_{MAP} = \widehat{\theta}_{ABS}$ which is an efficient estimator.
- We found indeed for the estimation error correlation

$$\underbrace{\frac{1}{\sigma^2_{\tilde{\theta}}} = J}_{\text{efficiency}} = J_{prior} + J_Y = \frac{1}{\sigma^2_\theta} + \frac{n}{\sigma^2_v}$$

  which decomposes indeed into the prior information and n times the information in one measurement (all distributions involved are Gaussian).

Linear MMSE Estimation

- MMSE criterion is a desirable criterion but the resulting Bayes estimator $E[\theta|Y]$ may be complicated to derive.

- In practice often: suboptimal estimators, e.g. restrict the estimator to be a linear function of the data.

- Remark: the MMSE criterion for a vector parameter decomposes:

$$\min_{\widehat{\theta}} \; E\,(\widehat{\theta}-\theta)^T(\widehat{\theta}-\theta) \;=\; \min_{\widehat{\theta}} \; E\sum_{i=1}^{m}(\widehat{\theta}_i-\theta_i)^2 \;=\; \sum_{i=1}^{m}\min_{\widehat{\theta}_i} \; E\,(\widehat{\theta}_i-\theta_i)^2$$

  Hence it suffices to concentrate on the estimator $\widehat{\theta}_i$ of a scalar component θ_i of θ and then $\widehat{\theta} = \left[\widehat{\theta}_1 \cdots \widehat{\theta}_m\right]^T$.

- Linear Estimators : constrain $\widehat{\theta}_i(Y)$ to be linear:

$$\widehat{\theta}_i(Y) \;=\; F_i^T Y \;=\; [f_{i,1}\cdots f_{i,n}]\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \;=\; \sum_{k=1}^{n} f_{i,k}\, y_k$$

  where F_i is a $n \times 1$ vector of combination coefficients (of the same dimension as the data vector Y).

Linear MMSE Estimators

- MSE risk function (of F_i now !)

$$\mathcal{R}_{LMMSE}(F_i) \;=\; E\,(\theta_i - \widehat{\theta}_i)^2 \;=\; E\,(\theta_i - F_i^T Y)^2$$

and we shall obtain F_i as $F_i = \arg\min_{F_i} \mathcal{R}_{LMMSE}(F_i)$.

- Setting the gradient equal to zero leads to

$$\frac{\partial}{\partial F_i}\mathcal{R}_{LMMSE}(F_i) \;=\; -2E\,(\theta_i - F_i^T Y)Y \;=\; 0 \quad\Rightarrow\quad E\,(\theta_i - F_i^T Y)Y^T \;=\; 0$$

$$\Rightarrow\quad F_i^T = (E\,\theta_i Y^T)(E\,YY^T)^{-1} = R_{\theta_i Y} R_{YY}^{-1}$$

- The Hessian can be verified to be

$$\frac{\partial}{\partial\, F_i}\left(\frac{\partial}{\partial\, F_i}\mathcal{R}_{LMMSE}(F_i)\right)^T \;=\; 2\,R_{YY} \;>\; 0$$

Hence, the unique extremum is indeed the global minimum.

Linear MMSE Estimators (2)

- This can be done for each component θ_i of θ independently to yield the minimizer of the overall MSE criterion

$$\widehat{\theta} \;=\; \begin{bmatrix} \widehat{\theta}_1 \\ \vdots \\ \widehat{\theta}_m \end{bmatrix} \;=\; \begin{bmatrix} F_1^T Y \\ \vdots \\ F_m^T Y \end{bmatrix} \;=\; F\,Y \;=\; R_{\theta Y} R_{YY}^{-1} Y \;=\; \begin{bmatrix} R_{\theta_1 Y} R_{YY}^{-1} Y \\ \vdots \\ R_{\theta_m Y} R_{YY}^{-1} Y \end{bmatrix}$$

  where $F = [F_1 \cdots F_m]^T$.

- correlation matrix of the parameter estimation errors:

$$\begin{aligned} R_{\tilde{\theta}\tilde{\theta}}(F) \;&=\; E\,(\theta - \widehat{\theta})(\theta - \widehat{\theta})^T \;=\; E\,(\theta - FY)(\theta^T - Y^T F^T) \\ &=\; R_{\theta\theta} - R_{\theta Y} F^T - F\,R_{Y\theta} + F\,R_{YY} F^T \end{aligned}$$

- Evaluated at the minimum, this gives

$$R_{\tilde{\theta}\tilde{\theta}}^{LMMSE} \;=\; R_{\tilde{\theta}\tilde{\theta}}(R_{\theta Y} R_{YY}^{-1}) \;=\; R_{\theta\theta} - R_{\theta Y} R_{YY}^{-1} R_{Y\theta}$$

- the MSE criterion is just the trace of this correlation matrix, hence

$$\min_F E\;\underbrace{(\theta - FY)^T(\theta - FY)}_{=\|\theta - FY\|^2 = \|\tilde{\theta}\|^2 = \mathrm{tr}\{\tilde{\theta}\tilde{\theta}^T\}} \;=\; \mathrm{tr}\left\{R_{\tilde{\theta}\tilde{\theta}}^{LMMSE}\right\} \;=\; \mathrm{tr}\left\{R_{\theta\theta} - R_{\theta Y} R_{YY}^{-1} R_{Y\theta}\right\}$$

Affine MMSE Estimators

- for random variables with non-zero mean, it may be advantageous to add a constant term to the linear estimator:

$$\widehat{\theta}_i(Y) \;=\; F_i^T Y + g_i \quad \text{where } g_i \text{ is a scalar}$$

- The MSE risk function is now

$$\mathcal{R}_{AMMSE}(F_i, g_i) \;=\; E\,(\theta_i - \widehat{\theta}_i)^2 \;=\; E\,(\theta_i - F_i^T Y - g_i)^2$$

and we shall obtain F_i and g_i from the optimization problem $\min\limits_{F_i, g_i} \mathcal{R}_{AMMSE}(F_i, g_i)$.

- Setting the gradients equal to zero leads to

$$\begin{array}{lll} \frac{\partial}{\partial F_i}\mathcal{R}_{AMMSE}(F_i, g_i) \;=\; 0 \;=\; -2E\,(\theta_i - F_i^T Y - g_i)Y & \Big| & 1 \\ \frac{\partial}{\partial g_i}\mathcal{R}_{AMMSE}(F_i, g_i) \;=\; 0 \;=\; -2E\,(\theta_i - F_i^T Y - g_i) & \Big| & -m_Y \end{array}$$

- From the second equation, we get

$$g_i \;=\; m_{\theta_i} - F_i^T m_Y$$

Affine MMSE Estimators (2)

- By forming the indicated linear combination of both equations, we get

$$\begin{aligned} 0 &= E\,(\theta_i - F_i^T Y - g_i)(Y - m_Y) \;=\; E\,(\theta_i - m_{\theta_i} - F_i^T(Y - m_Y))(Y - m_Y) \\ &= E\,(Y - m_Y)(\theta_i - m_{\theta_i} - (Y - m_Y)^T F_i) \\ &= E\,\{(Y - m_Y)(\theta_i - m_{\theta_i})\} \;-\; E\,\left\{(Y - m_Y)(Y - m_Y)^T\right\}\; F_i = C_{Y\theta_i} - C_{YY} F_i \end{aligned}$$

  which leads to

$$\widehat{\theta}_i(Y) \;=\; F_i^T Y + g_i \;=\; m_{\theta_i} + C_{\theta_i Y} C_{YY}^{-1}(Y - m_Y)$$

- The Hessian can be verified to be (use $R_{XY} \;=\; C_{XY} + m_X\, m_Y^T \;=\; R_{YX}^T$, $C_{XY} = C_{YX}^T$)

$$\left[\begin{array}{cc} \frac{\partial}{\partial F_i}\left(\frac{\partial}{\partial F_i}\mathcal{R}_{AMMSE}\right)^T & \frac{\partial}{\partial F_i}\left(\frac{\partial}{\partial g_i}\mathcal{R}_{AMMSE}\right)^T \\ \frac{\partial}{\partial g_i}\left(\frac{\partial}{\partial F_i}\mathcal{R}_{AMMSE}\right)^T & \frac{\partial}{\partial g_i}\left(\frac{\partial}{\partial g_i}\mathcal{R}_{AMMSE}\right)^T \end{array}\right]\Bigg|_{\begin{cases} F_i = C_{\theta_i Y} C_{YY}^{-1} \\ g_i = m_{\theta_i} - C_{\theta_i Y} C_{YY}^{-1} m_Y \end{cases}}$$

$$= 2\begin{bmatrix} R_{YY} & m_Y \\ m_Y^T & 1 \end{bmatrix} = 2\begin{bmatrix} C_{YY} & 0 \\ 0 & 0 \end{bmatrix} + 2\begin{bmatrix} m_Y \\ 1 \end{bmatrix}\begin{bmatrix} m_Y \\ 1 \end{bmatrix}^T$$

$$= 2\begin{bmatrix} I & m_Y \\ 0 & 1 \end{bmatrix}\begin{bmatrix} C_{YY} & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} I & m_Y \\ 0 & 1 \end{bmatrix}^T \;>\; 0$$

Affine MMSE Estimators (3)

- This can again be done for each component θ_i of θ independently to yield the minimizer of the overall MSE criterion

$$\widehat{\theta} \;=\; \begin{bmatrix} \widehat{\theta}_1 \\ \vdots \\ \widehat{\theta}_m \end{bmatrix} \;=\; \begin{bmatrix} F_1^T Y + g_1 \\ \vdots \\ F_m^T Y + g_m \end{bmatrix} \;=\; F\,Y \;+\; g \;=\; m_\theta + C_{\theta Y} C_{YY}^{-1}(Y - m_Y)$$

  where $F = [F_1 \cdots F_M]^T$ and $g = [g_1 \cdots g_m]^T$.

- the affine estimator is unbiased: $E_{\mathbf{Y}}\,\widehat{\theta} \;=\; m_\theta$ or $E_{\mathbf{Y},\boldsymbol{\theta}}\widetilde{\theta} \;=\; 0$

- the correlation matrix of the parameter estimation errors evaluated for the optimal estimator is

$$\begin{aligned} R_{\tilde{\theta}\tilde{\theta}}^{AMMSE} &= E\,(\theta - \widehat{\theta})(\theta - \widehat{\theta})^T \\ &= E\,[\theta - m_\theta - C_{\theta Y} C_{YY}^{-1}(Y - m_Y)][\theta - m_\theta - C_{\theta Y} C_{YY}^{-1}(Y - m_Y)]^T \\ &= C_{\theta\theta} - C_{\theta Y} C_{YY}^{-1} C_{Y\theta} \;=\; C_{\tilde{\theta}\tilde{\theta}}^{AMMSE} \end{aligned}$$

- the MSE criterion is just the trace of this correlation matrix, hence

$$\min_{F,g} E\,(\theta - FY - g)^T(\theta - FY - g) \;=\; \mathrm{tr}\left\{C_{\tilde{\theta}\tilde{\theta}}^{AMMSE}\right\} \;=\; \mathrm{tr}\left\{C_{\theta\theta} - C_{\theta Y} C_{YY}^{-1} C_{Y\theta}\right\}$$

Linear MMSE Estimation: Remarks

- when θ and Y are jointly Gaussian, then

$$\widehat{\theta}_{MMSE} \;=\; E\,[\theta|Y] \;=\; m_\theta + C_{\theta Y}C_{YY}^{-1}(Y - m_Y) \;=\; \widehat{\theta}_{AMMSE}$$

  Hence, in the case of Gaussian variables, the Affine MMSE estimator is optimal!

- Whereas general Bayes estimators require the knowledge of the complete joint distribution $f(Y,\theta)$, the Linear and Affine MMSE estimators only require the joint first and second order moments of θ and Y.

- If θ and Y have non-zero means, it is advantageous to use an affine estimator. Indeed, using $R_{XY} = C_{XY} + m_X m_Y^T$ and the Matrix Inversion Lemma on $R_{YY}^{-1} = (C_{YY} + m_Y m_Y^T)^{-1}$, one can show that

$$\begin{aligned} R_{\tilde{\theta}\tilde{\theta}}^{LMMSE} \;&=\; R_{\tilde{\theta}\tilde{\theta}}^{AMMSE} + \underbrace{(m_\theta - C_{\theta Y}C_{YY}^{-1}m_Y)(m_Y^T C_{YY}^{-1} m_Y + 1)^{-1}(m_\theta - C_{\theta Y}C_{YY}^{-1}m_Y)^T}_{\geq 0} \\ &\geq\; R_{\tilde{\theta}\tilde{\theta}}^{AMMSE} \;=\; C_{\tilde{\theta}\tilde{\theta}}^{AMMSE}\,. \end{aligned}$$

  By taking the trace of these expressions, one finds that the affine estimator leads to lower MSE than the linear estimator when $m_Y \neq 0$ or $m_\theta \neq 0$.

  If $m_\theta = C_{\theta Y}C_{YY}^{-1}m_Y$, check that we should have $\widehat{\theta}_{LMMSE} = \widehat{\theta}_{AMMSE}$.

Linear MMSE Estimation: Remarks (2)

Linear MMSE estimator simpler than the Affine MMSE estimator
$\Rightarrow$ reduce Affine MMSE estimation to Linear MMSE estimation.

- Method 1: introduce a linear estimator for an augmented problem

$$\theta' = \theta\,,\ Y' = \begin{bmatrix} Y \\ 1 \end{bmatrix}\,,\quad \widehat{\theta}'_L \;=\; \underbrace{F'}_{1\times(n+1)}\ \underbrace{Y'}_{(n+1)\times 1} \;=\; [F\ g]\begin{bmatrix} Y \\ 1 \end{bmatrix} \;=\; F\,Y + g \;=\; \widehat{\theta}_A$$

  The Linear MMSE estimator for the augmented problem $\{\theta, Y'\}$, namely

$$\widehat{\theta}'_{LMMSE} \;=\; R_{\theta Y'}R^{-1}_{Y'Y'}Y' \;=\; m_\theta + C_{\theta Y}C^{-1}_{YY}(Y - m_Y) \;=\; \widehat{\theta}_{AMMSE}$$

  is in fact the Affine MMSE estimator for the original problem $\{\theta, Y\}$. (Exo!)

- Method 2: When $m_\theta = 0$ and $m_Y = 0$, the affine estimator reduces to the linear estimator $\Rightarrow$ centralize the variables before further treatment

$$\begin{cases} \theta' \;=\; \theta - m_\theta \\ Y' \;=\; Y - m_Y\ . \end{cases}$$

  Then the linear and the affine MMSE estimators coincide

$$\widehat{\theta}' \;=\; R_{\theta' Y'}R^{-1}_{Y'Y'}Y' \;=\; C_{\theta' Y'}C^{-1}_{Y'Y'}Y' \;=\; C_{\theta Y}C^{-1}_{YY}Y'$$

  henceforth assume zero mean.

Linear MMSE Estimation: Remarks (3)

- Except in the jointly Gaussian case, the MSE increases by imposing the constraint of linearity on the estimator. This increase can be displayed by decomposing the MSE associated with a linear estimator as follows:

$$
\begin{aligned}
E\|\theta - FY\|^2 &= E(\theta - FY)^T(\theta - FY) \\
&= E(\theta - E[\theta|Y] + E[\theta|Y] - FY)^T(\theta - E[\theta|Y] + E[\theta|Y] - FY) \\
&= E(\theta - E[\theta|Y])^T(\theta - E[\theta|Y]) + \underbrace{E(E[\theta|Y] - FY)^T(E[\theta|Y] - FY)}_{\geq 0} \\
&\quad +2\,\underbrace{E(\theta - E[\theta|Y])^T(E[\theta|Y] - FY)}_{=0} \;\geq\; E(\theta - E[\theta|Y])^T(\theta - E[\theta|Y])
\end{aligned}
$$

- the difference between the LMMSE and the MMSE is the MSE in approximating the conditional mean $E[\theta|Y]$ by a linear function FY.

- In fact, the best linear approximation of $E[\theta|Y]$ is also the best linear approximation of θ since the above implies

$$R_{\theta Y}R_{YY}^{-1} = \arg\min_F E_{Y,\theta}\|\theta - FY\|^2 = \arg\min_F E_Y\|E[\theta|Y] - FY\|^2$$

- And for any F:

$$E_{Y,\theta}\|\theta - E[\theta|Y]\|^2 = E_{Y,\theta}\|\theta - FY\|^2 - E_Y\|E[\theta|Y] - FY\|^2$$

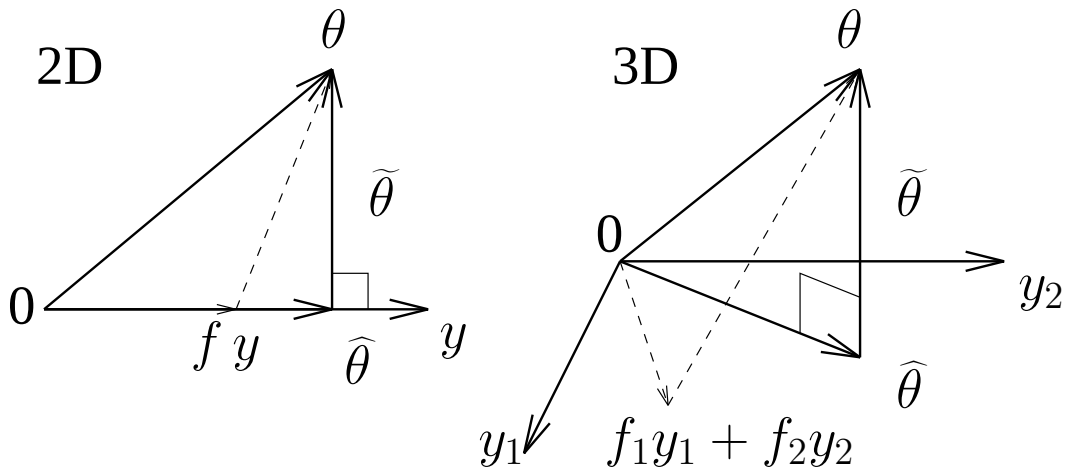
Orthogonality Principle of LMMSE

- Let θ, $y_1, \ldots, y_n$ be random variables that span a $(n+1)$-dimensional vector space with inner product $< x, y >= E\,xy$. We shall form a linear estimate (approximation) of θ in terms of $y_1, \ldots, y_n$: $\widehat{\theta} = F^T Y = \Sigma_{i=1}^n f_i\, y_i$ determine the combination coefficients f_i by minimizing the MSE

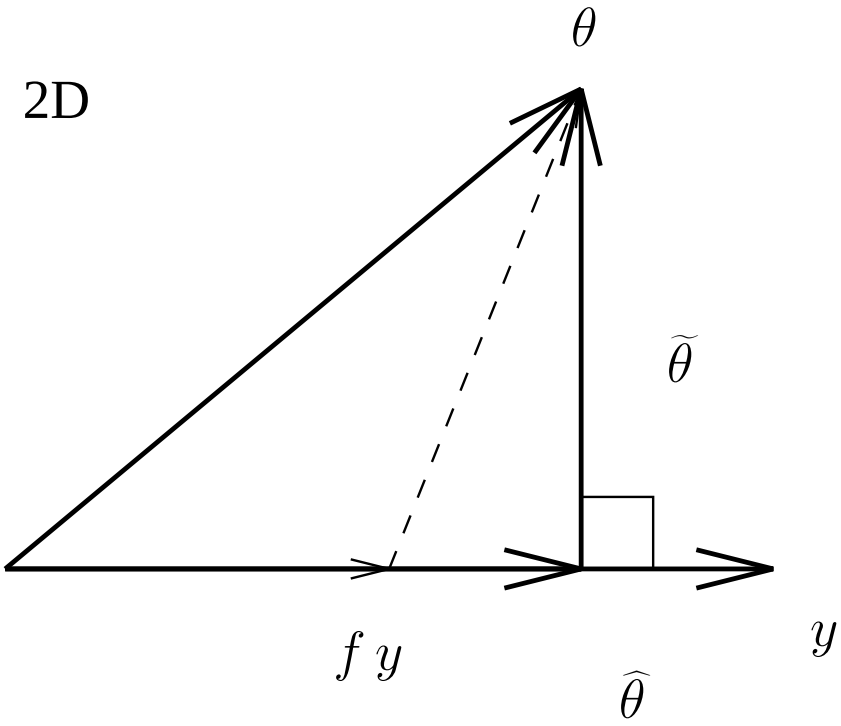
$$\min_{f_i} E\,(\theta - \widehat{\theta})^2 = \min_{f_i} E\,(\theta - \sum_{i=1}^{n} f_i\, y_i)^2 = \min_{f_i} \left\|\theta - \widehat{\theta}\right\|^2 \quad \Rightarrow$$

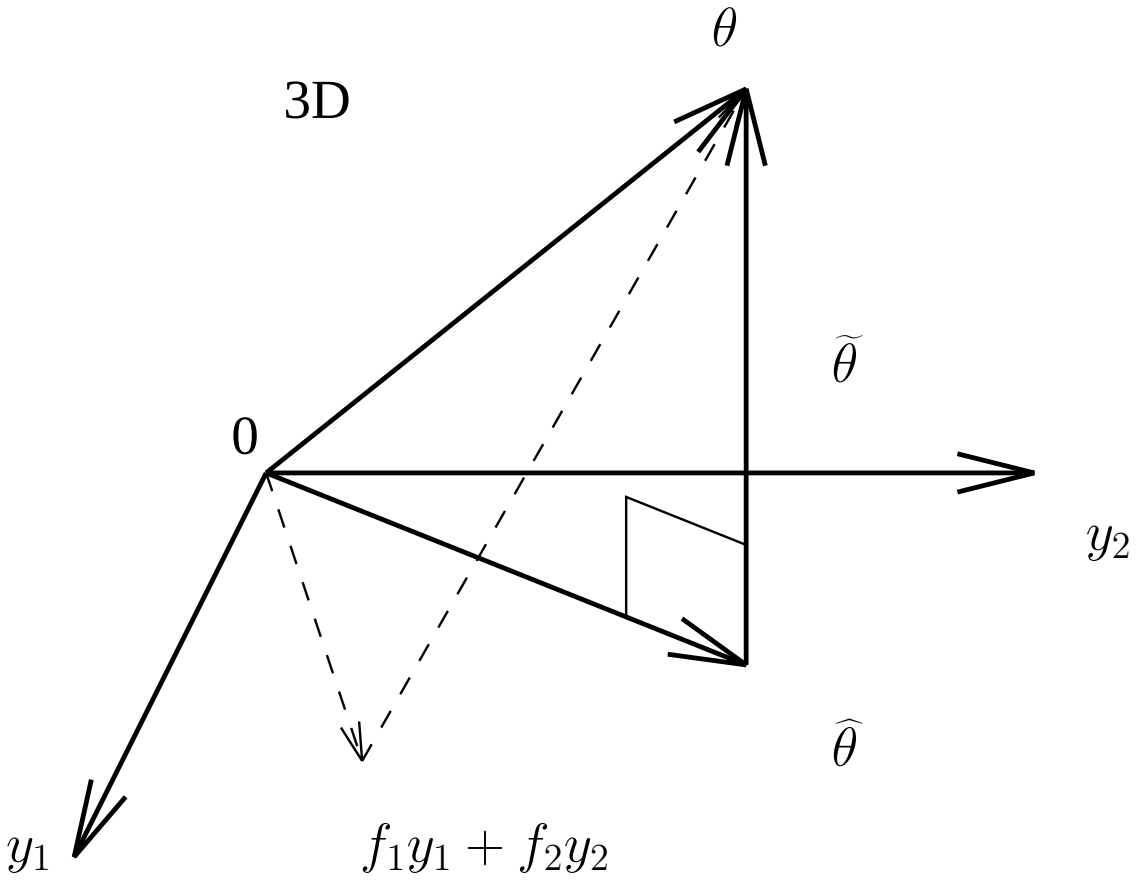
$$\frac{\partial}{\partial\, f_i} E\,(\theta - \widehat{\theta})^2 = -2\,E\,(\theta - \widehat{\theta}) y_i = 0 \quad \Rightarrow \quad < \theta - \widehat{\theta},\, y_i >= 0\,, \quad i = 1, \ldots, n$$

Hence the LMMSE estimate also satisfies the orthogonality conditions.



Orthogonality Principle of LMMSE





Orthogonality Principle of LMMSE (2)

- When $\theta = [\theta_1 \cdots \theta_m]^T$ is a random vector, then we can again consider the space spanned by the variables $\theta_1, \ldots, \theta_m, y_1, \ldots, y_n$. We shall now more generally consider vectors of random variables and the matrix inner product between them. If X and Y are random vectors, then we take their inner product to be $< X, Y >= E\, XY^T = R_{XY}$.

- A linear estimator for θ in terms of Y is now

$$\widehat{\theta} = F\, Y \quad \text{where } F \text{ is } m \times n$$

- We cannot follow directly the same path as for the case of a scalar θ since we have not seen how to take gradients w.r.t. a matrix (and even less how to consider the corresponding Hessian). Geometrical intuition: optimal F is the one that satisfies the orthogonality condition:

$$0 = < \theta - \widehat{\theta}, Y > = < \theta - FY, Y > = < \theta, Y > - F < Y, Y >$$

so we find the result we found before using a different route:

$$F = < \theta, Y >< Y, Y >^{-1} = R_{\theta Y} R_{YY}^{-1}$$

Orthogonality Principle of LMMSE (3)

- We can now show that this F which satisfies the orthogonality condition minimizes the correlation matrix $R_{\tilde{\theta}\tilde{\theta}}$ of the estimation errors. Indeed, let KY be any other linear estimator of θ. With $< X, Y >= R_{XY}$, $||X||^2 = R_{XX}$, we get

$$
\begin{aligned}
R_{\tilde{\theta}\tilde{\theta}}(K) &= \|\theta - KY\|^2 = < \theta - KY, \theta - KY > \\
&= < \theta - FY + FY - KY, \theta - FY + FY - KY > \\
&= \|\theta - FY\|^2 + \|(F-K)Y\|^2 \\
&\quad + \underbrace{< \theta - FY, Y >}_{=0}(F-K)^T + (F-K)\underbrace{< Y, \theta - FY >}_{=0} \\
&= \|\theta - FY\|^2 + \underbrace{(F-K)\,\|Y\|^2\,(F-K)^T}_{\geq 0 \ (=0 \text{ iff } F=K \ (\|Y\|^2 = R_{YY} > 0))} \\
&\geq \|\theta - FY\|^2 = R_{\tilde{\theta}\tilde{\theta}}(F) \ .
\end{aligned}
$$

Since the MSE is the trace of $R_{\tilde{\theta}\tilde{\theta}}$, this shows again that $\widehat{\theta} = R_{\theta Y} R_{YY}^{-1} Y$ is the LMMSE estimator, with a proof based on the orthogonality property. This is an example where the whole $R_{\tilde{\theta}\tilde{\theta}}$ gets minimized instead of just its trace (= MSE).

Bayesian Linear Model

- $Y = H\,\theta + V$, $\theta \sim \mathcal{N}(m_\theta, C_{\theta\theta})$ and $V \sim \mathcal{N}(0, C_{VV})$ independent
- $\begin{bmatrix} \theta \\ Y \end{bmatrix} = \begin{bmatrix} I & 0 \\ H & I \end{bmatrix} \begin{bmatrix} \theta \\ V \end{bmatrix}$ jointly Gaussian, $Y - m_Y = H\,(\theta - m_\theta) + V$
- $C_{YY} = H\,C_{\theta\theta}\,H^T + C_{VV}$, $C_{Y\theta} = H\,C_{\theta\theta}$
- Gauss-Markov theorem:

$$
\begin{aligned}
f(\theta|Y) \;&\leftrightarrow\; \mathcal{N}\left(m_{\theta|Y},\, C_{\theta|Y}\right) \\
m_{\theta|Y} \;&=\; m_\theta + C_{\theta Y} C_{YY}^{-1}\,(Y - m_Y) \\
&=\; m_\theta + C_{\theta\theta} H^T \left(H\,C_{\theta\theta}\,H^T + C_{VV}\right)^{-1} (Y - H\,m_\theta) \\
&\overset{\text{MIL}}{=}\; m_\theta + \left(C_{\theta\theta}^{-1} + H^T\,C_{VV}^{-1}\,H\right)^{-1} H^T\,C_{VV}^{-1}\,(Y - H\,m_\theta) \\
C_{\theta|Y} \;&=\; C_{\theta\theta} - C_{\theta Y} C_{YY}^{-1} C_{Y\theta} = C_{\theta\theta} - C_{\theta\theta} H^T \left(H\,C_{\theta\theta}\,H^T + C_{VV}\right)^{-1} H\,C_{\theta\theta} \\
\overset{\text{MIL}}{\Rightarrow}\; C_{\theta|Y}^{-1} \;&=\; C_{\theta\theta}^{-1} \;+\; H^T\,C_{VV}^{-1}\,H
\end{aligned}
$$

- θ, Y jointly Gaussian $\Rightarrow\; \widehat{\theta}_{MMSE} = \widehat{\theta}_{AMMSE} = m_{\theta|Y}$

Bayesian Linear Model (2)

- $f(\theta|Y)$ Gaussian $\Rightarrow$ (A)MMSE estimator = efficient:
  $R_{\tilde{\theta}\tilde{\theta}} = C_{\theta|Y} = J^{-1},\ J = J_{prior} + J_Y = C_{\theta\theta}^{-1} + H^T C_{VV}^{-1} H$

- if noise samples uncorrelated: C_{VV} diagonal: $C_{VV} = \text{diag}\{\sigma_{v_1}^2 \cdots \sigma_{v_n}^2\}$,
  then the information from the data also decomposes:
  $J_Y = H^T C_{VV}^{-1} H = \sum_{k=1}^{n} H_{k,:}^T \sigma_{v_k}^{-2} H_{k,:}$ $\qquad \text{rank}(J_Y) \leq \min\{n, m\}$

- Note: for $n < m$, J_Y has rank $\leq n$ and hence the $m \times m$ matrix J_Y is singular. However, due to $J_{prior} > 0$, $J > 0$ and hence non-singular $\Rightarrow$ importance of prior information when only little measurement data is available.

- On the other hand, as $n \to \infty$, J_{prior} becomes of negligible importance compared to J_Y. So in general, the influence of the prior information disappears asymptotically as the number of measurements becomes large.

Recap: Bayes Parameter Estimation

- obtain the estimator $\widehat{\theta}(.)$ by minimizing the risk, the average cost:

$$\min_{\widehat{\theta}(.)} \mathcal{R}(\widehat{\theta}(.)) = E_{\mathbf{Y}} \left[\min_{\widehat{\theta}(Y)} E_{\boldsymbol{\theta}|\mathbf{Y}} \mathcal{C}(\theta, \widehat{\theta}(Y)) \right] = E_{\mathbf{Y}} \left[\min_{\widehat{\theta}(Y)} \mathcal{R}(\widehat{\theta}(Y)|Y) \right]$$

  requires the posterior distribution $f_{\boldsymbol{\theta}|\mathbf{Y}}(\theta|Y) = f_{\mathbf{Y}|\boldsymbol{\theta}}(Y|\theta) f_{\boldsymbol{\theta}}(\theta) / f_{\mathbf{Y}}(Y)$

- quadratic cost function (risk=MSE): $\widehat{\theta}_{MMSE}(Y) = E(\theta|Y)$
- uniform cost function : $\widehat{\theta}_{MAP}(Y) = \arg\max_\theta f(\theta|Y) = \arg\max_\theta f(Y|\theta) f(\theta)$
- information matrix $J = E\left(\frac{\partial \ln f(\theta|Y)}{\partial \theta}\right)\left(\frac{\partial \ln f(\theta|Y)}{\partial \theta}\right)^T = -E \frac{\partial}{\partial \theta}\left(\frac{\partial \ln f(\theta|Y)}{\partial \theta}\right)^T$
- $\widehat{\theta}$ unbiased if $E_{\boldsymbol{\theta}} b_{\widehat{\theta}}(\theta) = 0$ or $\lim_{\theta \to \partial\Theta} f(\theta) b_{\widehat{\theta}}(\theta) = 0$, $b_{\widehat{\theta}}(\theta) = -E_{\mathbf{Y}|\boldsymbol{\theta}} \widetilde{\theta}$. Then
- CRB: $R_{\tilde{\theta}\tilde{\theta}} \geq J^{-1}$, = (efficiency) iff $\widehat{\theta}(Y) - \theta = J^{-1} \frac{\partial \ln f(\theta|Y)}{\partial \theta}$, $f(\theta|Y)$ Gaussian
- in general: $R_{\tilde{\theta}_{MAP}\tilde{\theta}_{MAP}} \geq R_{\tilde{\theta}_{MMSE}\tilde{\theta}_{MMSE}} \geq J^{-1}$
- linear MMSE: $\widehat{\theta}_{LMMSE} = R_{\theta Y} R_{YY}^{-1} Y$
- $R_{\tilde{\theta}_{LMMSE}\tilde{\theta}_{LMMSE}} \geq R_{\tilde{\theta}_{MMSE}\tilde{\theta}_{MMSE}}$, Gaussian case: $\widehat{\theta}_{AMMSE} = \widehat{\theta}_{MMSE} = \widehat{\theta}_{MAP}$
- special Gaussian case: linear model: $Y = H\,\theta + V$