E U R E C O M

Statistical Signal Processing

Lecture 2

chapter 1: parameter estimation

stochastic parameters

- the parameter estimation problem

- Bayes estimation: the MMSE, absolute value and uniform cost functions

- examples: Gaussian mean in Gaussian noise, Poisson process

- vector parameters
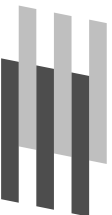
- Fischer Information Matrix

Vector Parameters

- $\boldsymbol{\theta} = \begin{bmatrix} \theta_1 \\ \cdots \\ \theta_m \end{bmatrix}$, $\widehat{\boldsymbol{\theta}}(Y) = \begin{bmatrix} \widehat{\theta}_1(Y) \\ \cdots \\ \widehat{\theta}_m(Y) \end{bmatrix}$, $\widetilde{\boldsymbol{\theta}} = \widetilde{\boldsymbol{\theta}}(\boldsymbol{\theta}, Y) = \boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}(Y)$

- problem formulation:

  – a prior distribution $f_{\boldsymbol{\theta}}(\theta)$

  – a conditional distribution $f_{Y|\boldsymbol{\theta}}(Y|\theta)$

  – Bayes' rule : joint distribution $f_{Y,\boldsymbol{\theta}}(Y,\theta) = f_{Y|\boldsymbol{\theta}}(Y|\theta) f_{\boldsymbol{\theta}}(\theta)$
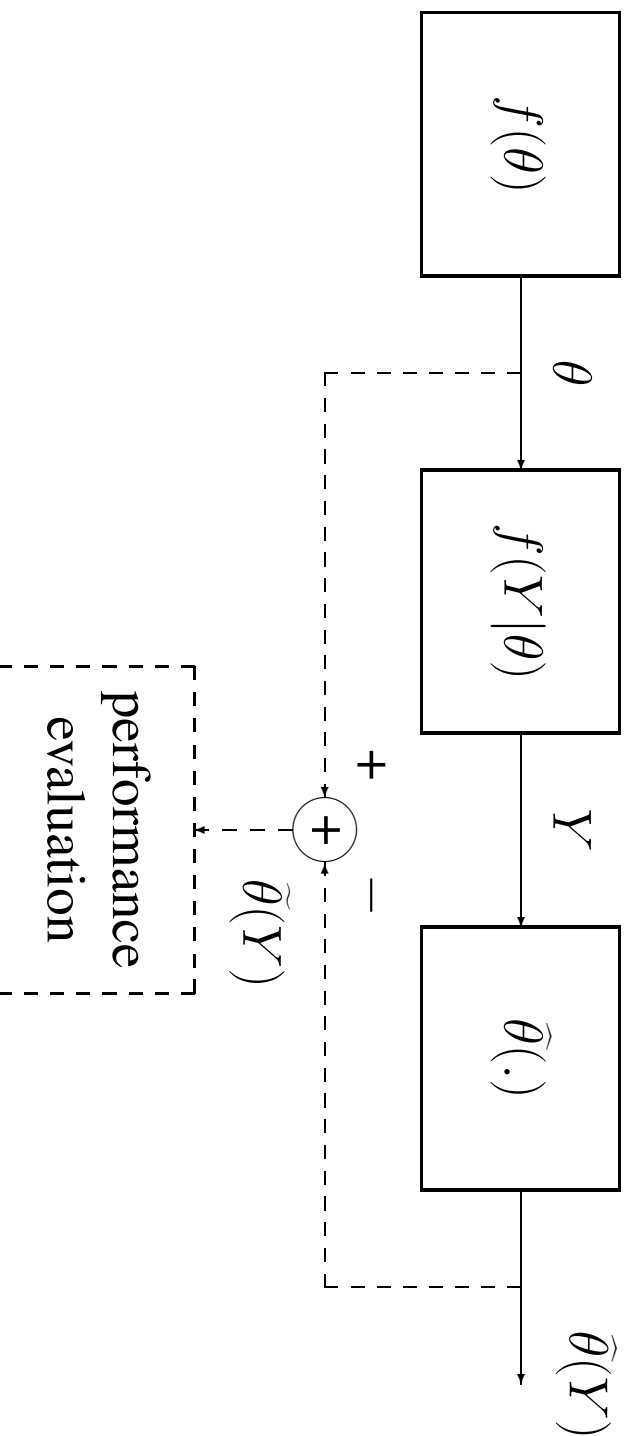
Bayes Risk Function

- cost $C(\theta, \widehat{\theta}(Y))$

- $C(\theta, \widehat{\theta}(Y))$ often directly a funtion of the estimation error $\widetilde{\theta}$ and in fact often a function of the length of the estimation error $\|\widetilde{\theta}\| = \sqrt{\widetilde{\theta}^T \widetilde{\theta}} = \sqrt{\sum_{i=1}^n \widetilde{\theta}_i^2}$

- We obtain the estimator function $\widehat{\theta}(.)$ by minimizing the risk, which is the expected value of the cost:

$$
\begin{aligned}
\min_{\widehat{\theta}(.)} \mathcal{R}(\widehat{\theta}(.)) \;=\; & \min_{\widehat{\theta}(.)} E\, C(\theta, \widehat{\theta}(Y)) \;=\; \min_{\widehat{\theta}(.)} E_{Y, \boldsymbol{\theta}} C(\theta, \widehat{\theta}(Y)) \\
=\; & \min_{\widehat{\theta}(.)} E_Y E_{\boldsymbol{\theta}|Y} C(\theta, \widehat{\theta}(Y)) \;=\; E_Y \left[\min_{\widehat{\theta}(Y)} E_{\boldsymbol{\theta}|Y} C(\theta, \widehat{\theta}(Y)) \right] \\
=\; & E_Y \left[\min_{\widehat{\theta}(Y)} \mathcal{R}(\widehat{\theta}(Y)|Y) \right] .
\end{aligned}
$$

- $\mathcal{R}(\widehat{\theta}(.))$ is a weighted average of $\mathcal{R}(\widehat{\theta}(Y)|Y)$, weighted by the nonnegative weighting function $f_Y(Y)$. The minimum of $\mathcal{R}(\widehat{\theta}(.))$ w.r.t. $\widehat{\theta}(.)$ will hence be obtained by minimizing $\mathcal{R}(\widehat{\theta}(Y)|Y)$ w.r.t. $\widehat{\theta}(Y)$ for every Y.

- again $\mathcal{R}(\widehat{\theta}(Y)|Y)$ depends on the posterior distribution $f_{\boldsymbol{\theta}|Y}(\theta|Y)$.
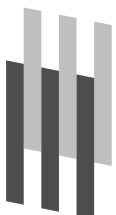
Optimization w.r.t. Vector Parameters

- $g(\theta) = [g_1(\theta) \cdots g_l(\theta)]$: $1 \times l$ row vector function, its gradient w.r.t. θ:

$$\frac{\partial\, g(\theta)}{\partial \theta} = \begin{bmatrix} \frac{\partial\, g(\theta)}{\partial \theta_1} \\ \cdots \\ \frac{\partial\, g(\theta)}{\partial \theta_m} \end{bmatrix} = \begin{bmatrix} \frac{\partial\, g_1(\theta)}{\partial \theta_1} & \cdots & \frac{\partial\, g_l(\theta)}{\partial \theta_1} \\ \cdots & \cdots & \cdots \\ \frac{\partial\, g_1(\theta)}{\partial \theta_m} & \cdots & \frac{\partial\, g_l(\theta)}{\partial \theta_m} \end{bmatrix} \quad m \times l$$
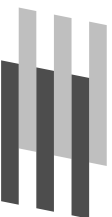
If $g(\theta)$ is a scalar ($l = 1$), then $\frac{\partial\, g(\theta)}{\partial \theta}$ is a column vector of the same dimensions as θ.

- in particular: $\quad \frac{\partial\, \theta^T}{\partial \theta} = \left[\frac{\partial\, \theta_j}{\partial \theta_i} \right] = [\delta_{i,j}] = I_m$

- The gradient operator commutes with linear operations. Let X be $m \times 1$

$$\frac{\partial}{\partial \theta}(\theta^T X) = \left(\frac{\partial\, \theta^T}{\partial \theta}\right) X = I_m\, X = X\,.$$

- Since a scalar equals its transpose, we get

$$\frac{\partial}{\partial \theta}(X^T \theta) = \frac{\partial}{\partial \theta}(\theta^T X) = X$$

Optimization w.r.t. Vector Parameters (2)

- If A is $m \times l$: $\quad \frac{\partial}{\partial \theta} (\theta^T A) = \left(\frac{\partial \theta^T}{\partial \theta} \right) A = I_m A = A$

- scalar case: $\quad (uv)' = u'v + uv'$

- vector case: let $g(\theta)$ and $h(\theta)$ be $l \times 1$. Since

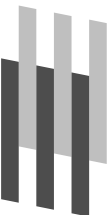$$g^T(\theta) h(\theta) = (g^T(\theta) h(\theta))^T = h^T(\theta) g(\theta)$$

we get

$$\frac{\partial}{\partial \theta} (g^T(\theta) h(\theta)) = \left(\frac{\partial g^T(\theta)}{\partial \theta} \right) h(\theta) + \left(\frac{\partial h^T(\theta)}{\partial \theta} \right) g(\theta)$$

- Particular application with $g(\theta) = \theta$ and $h(\theta) = A\theta$:

$$\frac{\partial}{\partial \theta} (\theta^T A \theta) = \left(\frac{\partial \theta^T}{\partial \theta} \right) A\theta + \left(\frac{\partial \theta^T A^T}{\partial \theta} \right) \theta = (A + A^T) \theta$$

- When A is symmetric, this gradient reduces to $2A\theta$.

MMSE Criterion: Vector Parameters

- quadratic cost function $C_{MMSE}(\theta, \widehat{\theta}) = \|\widetilde{\theta}\|_2^2 = \widetilde{\theta}^T \widetilde{\theta} = \sum_{i=1}^n \widetilde{\theta}_i^2$

- minimizing the conditional Bayes risk :

$$\min_{\widehat{\theta}(Y)} \mathcal{R}_{MMSE}(\widehat{\theta}(Y)|Y) = \min_{\widehat{\theta}(Y)} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(\theta|Y) (\theta - \widehat{\theta})^T (\theta - \widehat{\theta}) d\theta_1 \cdots d\theta_m$$
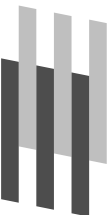
- extrema:

$$\frac{\partial}{\partial \widehat{\theta}} \mathcal{R}_{MMSE}(\widehat{\theta}|Y) = \frac{\partial}{\partial \widehat{\theta}} \int f(\theta|Y) (\theta - \widehat{\theta})^T (\theta - \widehat{\theta}) d\theta$$

$$= \int f(\theta|Y) \left(\frac{\partial}{\partial \widehat{\theta}} (\theta - \widehat{\theta})^T (\theta - \widehat{\theta}) \right) d\theta = 2 \int f(\theta|Y) (\widehat{\theta} - \theta) d\theta = 0$$

- or hence $\widehat{\theta}(Y) \underbrace{\int f(\theta|Y) d\theta}_{= 1} = \int \theta f(\theta|Y) d\theta \Rightarrow \widehat{\theta}_{MMSE}(Y) = E(\theta|Y)$

which is again the *mean* of the a posteriori distribution of θ given Y.

- extremum = minimum?

$$Hessian = \left[\frac{\partial^2}{\partial \widehat{\theta}_i \partial \widehat{\theta}_j} \mathcal{R}_{MMSE}(\widehat{\theta}|Y) \right] = \frac{\partial}{\partial \widehat{\theta}} \left(\frac{\partial}{\partial \widehat{\theta}} \mathcal{R}_{MMSE}(\widehat{\theta}|Y) \right)^T$$

$$= 2 \int f(\theta|Y) \left[\frac{\partial \widehat{\theta}^T}{\partial \widehat{\theta}} - \frac{\partial \theta^T}{\partial \widehat{\theta}} \right] d\theta = 2I \int f(\theta|Y) d\theta = 2I > 0$$

MMSE Criterion: Vector Parameters (2)

● MMSE estimation commutes over linear transformations: with $\phi = A\theta$

$$\hat{\phi}_{MMSE} = E(\phi|Y) = E(A\theta|Y) = A E(\theta|Y) = A\hat{\theta}_{MMSE}$$

● orthogonality property of MMSE estimators:

$$\hat{\theta}(Y) = E(\boldsymbol{\theta}|Y) \quad \text{iff} \quad E((\boldsymbol{\theta} - \hat{\theta}(Y))g(Y)) = 0, \quad \forall g(.)$$

where $g(.)$ is a scalar function. Equivalently :

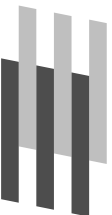$$E(\hat{\theta}(Y)g(Y)) = E(\boldsymbol{\theta} g(Y)), \quad \forall g(.)$$

which represents an alternative way of defining $E(\boldsymbol{\theta}|Y)$.

● use orthogonality to show optimality: let $\hat{\theta}(Y)$ be any function of Y,

$$E\|\boldsymbol{\theta} - \hat{\theta}(Y)\|_2^2 = E\|\boldsymbol{\theta} - E(\theta|Y) + E(\theta|Y) - \hat{\theta}(Y)\|_2^2$$

$$= E\|\boldsymbol{\theta} - E(\theta|Y)\|_2^2 + \underbrace{E\|E(\theta|Y) - \hat{\theta}(Y)\|_2^2}_{\geq 0} + \underbrace{2\,E\left((\boldsymbol{\theta} - E(\theta|Y))^T (E(\theta|Y) - \hat{\theta}(Y))\right)}_{=0}$$

$$\geq E\|\boldsymbol{\theta} - E(\theta|Y)\|_2^2$$

● correlation matrices: $\quad E\left((\boldsymbol{\theta} - E(\theta|Y))(\boldsymbol{\theta} - E(\theta|Y))^T\right) \leq E\left((\boldsymbol{\theta} - \hat{\theta})(\boldsymbol{\theta} - \hat{\theta})^T\right)$

MAP Estimators: Vector Parameters

- introduce: a ball centered around θ_o with radius δ

$$B_\delta(\theta_o) = \{\theta \in \Theta : \|\theta - \theta_0\|_2 \leq \delta\}$$

- Then the natural extension to the vector case of the uniform cost function is

$$C_{UNIF}(\theta, \widehat{\theta}) = \begin{cases} 0 & , \theta \in B_\delta(\widehat{\theta}) \\ 1 & , \theta \in \Theta \setminus B_\delta(\widehat{\theta}) \end{cases}$$
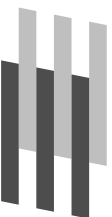
- The conditional Bayes risk becomes

$$\mathcal{R}_{UNIF}(\widehat{\theta}(Y)|Y) = \int_\Theta f(\theta|Y) C_{UNIF}(\theta, \widehat{\theta}) d\theta = \int_{\Theta \setminus B_\delta(\widehat{\theta})} f(\theta|Y) d\theta$$

$$= \int_\Theta f(\theta|Y) d\theta - \int_{B_\delta(\widehat{\theta})} f(\theta|Y) d\theta = 1 - \int_{B_\delta(\widehat{\theta})} f(\theta|Y) d\theta$$

- The optimization problem $\min_{\widehat{\theta}(Y)} \mathcal{R}_{UNIF}(\widehat{\theta}(Y)|Y)$ hence leads to

$$\max_{\widehat{\theta}(Y)} \int_{B_\delta(\widehat{\theta})} f(\theta|Y) d\theta \approx \text{Vol}(B_\delta(0)) \max_{\widehat{\theta}(Y)} f(\theta|Y)$$

the approximation becomes arbitrarily accurate as δ becomes small

MAP Estimators: Vector Parameters (2)

- This leads to the Maximum A Posteriori (likelihood) estimator

$$\widehat{\theta}_{MAP}(Y) = \arg\max_{\theta \in \Theta} f(\theta|Y) \quad \text{posterior likelihood}$$

- The same remarks as in the scalar case hold here also. In particular, one may equivalently obtain $\widehat{\theta}_{MAP}(Y)$ from the optimization problem

$$\widehat{\theta}_{MAP}(Y) = \arg\max_{\theta \in \Theta} \ln f(\theta|Y) \quad \text{posterior log likelihood}$$
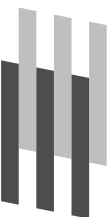
- Under certain regularity conditions, $\widehat{\theta}_{MAP}(Y)$ can be found from

$$\frac{\partial}{\partial\theta}\ln f(\theta|Y) = 0 = \frac{\partial}{\partial\theta}\ln f(Y|\theta) + \frac{\partial}{\partial\theta}\ln f(\theta)$$

- Also MAP commutes over linear transformations: $\phi = A\theta$ (A invertible)

$$\widehat{\phi}_{MAP}(Y) = \arg\max_{\phi} f_{\boldsymbol{\phi}|\mathbf{Y}}(\phi|Y) = A \arg\max_{\theta} f_{\boldsymbol{\phi}|\mathbf{Y}}(A\theta|Y)$$

$$= A \arg\max_{\theta} \frac{1}{|\det A|} f_{\boldsymbol{\theta}|\mathbf{Y}}(\theta|Y) = A\widehat{\theta}_{MAP}(Y) \,.$$

This argument can be extended to the case in which $\dim\phi \neq \dim\theta$

Fisher Information Matrix

There exists a lower bound on the correlation matrix of the estimator errors. It is independent of the Bayes estimator (cost) used; it depends only on the posterior distribution. The lower bound is specified in terms of the *information matrix*, which should express in quantitive terms the information carried by the posterior distribution about the parameters θ. For such an information measure, the following properties are desirable:

- The information should increase as the *sensitivity* of $f(\theta|Y)$ to changes in θ increases. Hence, the information should be an increasing function of $\frac{\partial f(\theta|Y)}{\partial \theta}$ or of $\frac{\partial \ln f(\theta|Y)}{\partial \theta}$.

- The information should be *additive* in the sense that it should be the sum of the informations from the prior distribution ($f(\theta)$) and from the data ($f(Y|\theta)$). Furthermore if, given θ, Y_1 and Y_2 are independent ($f(Y_1, Y_2|\theta) = f(Y_1|\theta)f(Y_2|\theta)$), then the informations in Y_1 and Y_2 should add up.

- The information should be positive and should be insensitive to a change of sign of θ.

- The information should be a *deterministic* quantity.
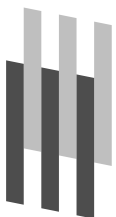
Fisher Information Matrix (2)

- The information matrix is defined as

$$J = E \left(\frac{\partial \ln f(\theta|Y)}{\partial \theta} \right) \left(\frac{\partial \ln f(\theta|Y)}{\partial \theta} \right)^T$$

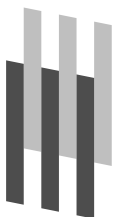It can be shown to satisfy all the properties mentioned above.

- The perturbation of Mutual Information (Information Theory) w.r.t. a parameter can be expressed in terms of its Fisher Information.

- With $\frac{\partial \ln f(\theta|Y)}{\partial \theta} = \frac{1}{f(\theta|Y)} \frac{\partial f(\theta|Y)}{\partial \theta}$ we can write the Hessian of $\ln f(\theta|Y)$ as

$$\frac{\partial}{\partial \theta} \left(\frac{\partial \ln f(\theta|Y)}{\partial \theta} \right)^T = \frac{1}{f^2(\theta|Y)} \left[f(\theta|Y) \frac{\partial}{\partial \theta} \left(\frac{\partial f(\theta|Y)}{\partial \theta} \right)^T - \left(\frac{\partial f(\theta|Y)}{\partial \theta} \right) \left(\frac{\partial f(\theta|Y)}{\partial \theta} \right)^T \right]$$

$$= \frac{1}{f(\theta|Y)} \frac{\partial}{\partial \theta} \left(\frac{\partial f(\theta|Y)}{\partial \theta} \right)^T - \left(\frac{\partial \ln f(\theta|Y)}{\partial \theta} \right) \left(\frac{\partial \ln f(\theta|Y)}{\partial \theta} \right)^T$$

- For the expectation of the first term, we get

$$E_{f(\theta|Y)} \frac{1}{f(\theta|Y)} \frac{\partial}{\partial \theta} \left(\frac{\partial f(\theta|Y)}{\partial \theta} \right)^T = \int d\theta \int dY \, f(Y) \frac{\partial}{\partial \theta} \left(\frac{\partial f(\theta|Y)}{\partial \theta} \right)^T$$

$$= \frac{\partial}{\partial \theta} \left(\frac{\partial}{\partial \theta} \int d\theta \int dY \, f(Y) f(\theta|Y) \left(|Y \right) \right)^T = \frac{\partial}{\partial \theta} \left(\frac{\partial}{\partial \theta} 1 \right)^T = 0$$

Fisher Information Matrix (3)

- It follows that we can rewrite the information matrix as

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

This expression will often allow us to obtain J more easily.

- Note also that

$$\frac{\partial \ln f(Y, \theta)}{\partial \theta} = \frac{\partial \ln f(\theta|Y)}{\partial \theta} + \underbrace{\frac{\partial \ln f(Y)}{\partial \theta}}_{=0} = \frac{\partial \ln f(\theta|Y)}{\partial \theta}$$
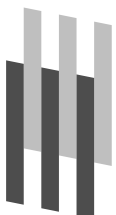
so that as long as derivatives are taken, we can interchange $f(Y, \theta)$ and $f(\theta|Y)$.
Hence

$$\frac{\partial \ln f(Y, \theta)}{\partial \theta} = \frac{\partial \ln f(\theta|Y)}{\partial \theta} = \frac{\partial \ln f(Y|\theta)}{\partial \theta} + \frac{\partial \ln f(\theta)}{\partial \theta}$$

and

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

Conditions on the Estimator Bias

- The (conditional) *bias* of an estimator $\widehat{\theta}(Y)$ of θ is defined as

$$b_{\widehat{\theta}}(\theta) \;=\; -E_{Y|\theta}\widetilde{\theta} \;=\; E_{Y|\theta}\,(\widehat{\theta}(Y)-\theta) \;=\; E_{Y|\theta}\widehat{\theta}(Y) - \theta$$

- An estimator will be called *unbiased* if either

$$E\theta\,b_{\widehat{\theta}}(\theta) \;=\; 0 \;\Leftrightarrow\; E_{\boldsymbol{\theta},Y}\,\widetilde{\theta} = 0$$

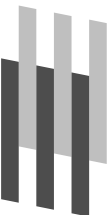which means that the unconditional or average bias is zero, or

$$\lim_{\theta\to\partial\Theta} f(\theta)b_{\widehat{\theta}}(\theta) = 0$$

where Θ is the domain for θ and $\partial\Theta$ is its boundary.

- **Lemma 0.1 (Unit Cross Correlation)** *If either condition above is satisfied, then*

$$E\,\frac{\partial\,\ln f(Y,\theta)}{\partial\theta}\,(\widehat{\theta}-\theta)^{T} \;=\; I\,.$$

In words, the cross correlation matrix between $\frac{\partial\,\ln f(Y,\theta)}{\partial\theta}$ and the estimation error of any unbiased estimator is the identity matrix.

Inner Products

- An inner product $< \cdot, \cdot >$ associates a real number $< x, y > \in \mathcal{R}$ with two vectors x and y of the vector space $\mathcal{V}$ we are considering, and it has the following properties:

1. linearity: $\forall \alpha_1, \alpha_2, \beta_1, \beta_2 \in \mathcal{R}, \quad \forall x, x_1, x_2, y, y_1, y_2 \in \mathcal{V}$:

$$
\begin{aligned}
< \alpha_1 x_1 + \alpha_2 x_2, y > &= \alpha_1 < x_1, y > + \alpha_2 < x_2, y > \\
< x, \beta_1 y_1 + \beta_2 y_2 > &= < x, y_1 > \beta_1 + < x, y_2 > \beta_2
\end{aligned}
\tag{1}
$$

2. symmetry: $< x, y > = < y, x >$

3. non-degeneracy (of the norm induced by the inner product):

$$< x, x > = \|x\|^2 \geq 0. \text{ If } \|x\| = 0, \text{ then } x = 0.$$

- One particular example is a space of random variables with the correlation as inner product: $< x, y > = E\, xy$. Non-degeneracy subtlety:

$$E\, x^2 = 0 \quad \Rightarrow \quad x = 0 \text{ in m.s.}$$

$x = 0$ "in mean square". Indeed, $E x^2 = m_x^2 + \sigma_x^2 = 0 \quad \Rightarrow \quad m_x = 0, \sigma_x^2 = 0$. This often (not always) implies $x = 0$ "almost surely" (a.s.) or "almost every-where" (a.e.) or "with probability 1" (w.p. 1): $\Pr(x = 0) = 1$.

Matrix Inner Products

- We now consider a vector space in which the vectors have multiple components such that the inner product is a real matrix. Example 1: consider a vector space of random vectors with inner product $< X, Y > = E\, X Y^T$ where X and Y are column vectors of random variables (not necessarily with the same number of rows).

- Example 2: a vector space in which the "vectors" are $* \times k$ real matrices where k is fixed and $*$ is arbitrary. Inner product: $< X, Y > = X Y^T$.
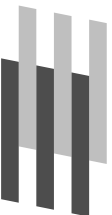
- Matrix valued inner products satisfy the following properties, which are natural generalizations of the scalar case.

1. linearity: let $X, X_1, X_2 \in \mathcal{V}$ have m rows and $Y, Y_1, Y_2 \in \mathcal{V}$ have n rows. Then $\forall \alpha_1, \alpha_2 \in \mathcal{R}^{k \times m}, \forall \beta_1, \beta_2 \in \mathcal{R}^{l \times n}$, for any k and l,

$$< \alpha_1 X_1 + \alpha_2 X_2, Y > \; = \; \alpha_1 < X_1, Y > + \alpha_2 < X_2, Y >$$
$$< X, \beta_1 Y_1 + \beta_2 Y_2 > \; = \; < X, Y_1 > \beta_1^T + < X, Y_2 > \beta_2^T$$

2. symmetry: $< X, Y > = < Y, X >^T$

3. non-degeneracy: $< X, X > = \|X\|^2 \geq 0$. If $\|X\|^2 = 0$, then $X = 0$. $\qquad (2)$

Schur Complements

- **Lemma 0.2 (Schur Complements)** *Let X_1 and X_2 be vectors in a certain vector space with a certain inner product and denote $R_{ij} = <X_i, X_j>$, $i, j = 1, 2$ so that $R_{ij} = R_{ji}^T$. Assume that R_{11} is nonsingular. Then, because of property 3 of the inner product (non-degeneracy), we have*

$$\|X_2 - R_{21}R_{11}^{-1}X_1\|^2 = <X_2 - R_{21}R_{11}^{-1}X_1, X_2 - R_{21}R_{11}^{-1}X_1>$$
$$= R_{22} - 2R_{21}R_{11}^{-1}R_{12} + R_{21}R_{11}^{-1}R_{11}R_{11}^{-1}R_{12}$$
$$= R_{22} - R_{21}R_{11}^{-1}R_{12} \geq 0$$

with equality iff $X_2 = R_{21}R_{11}^{-1}X_1$.

(matrix version of Cauchy-Schwarz inequality)

- **The name for this lemma stems from the following congruence relation**

$$< \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}, \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} > = \begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{bmatrix} = \begin{bmatrix} I & O \\ R_{21}R_{11}^{-1} & I \end{bmatrix} \begin{bmatrix} R_{11} & 0 \\ 0 & R_{22} - R_{21}R_{11}^{-1}R_{12} \end{bmatrix} \begin{bmatrix} I & R_{11}^{-1}R_{12} \\ O & I \end{bmatrix}$$

$$= \begin{bmatrix} I & \\ R_{21}R_{11}^{-1} & \end{bmatrix} R_{11} \begin{bmatrix} I \\ R_{21}R_{11}^{-1} \end{bmatrix}^T + \begin{bmatrix} 0 & 0 \\ 0 & R_{22} - R_{21}R_{11}^{-1}R_{12} \end{bmatrix} .$$

(block LDU triangular factorization). The matrix $R_{22} - R_{21}R_{11}^{-1}R_{12}$ is called the *Schur complement* of R_{11} within the big matrix on the **LHS** .