Multiple Random Variables

Lecture 5

Sept. 23, 25, 30, and Oct. 7 and 14, 2025

Outline

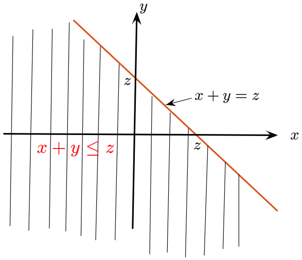
Functions of Two Random Variables

- Suppose we are given two RVs X and Y with known joint cdf $F_{XY}(x, y)$ (or pdf $f_{XY}(x, y)$) and a function $z = g(x, y)$. What is the cdf (or pdf) of the random variable $Z = g(X, Y)$?
- Use

$$\begin{aligned} F_Z(z) =& P(Z \leq z) = P((x, y) \,:\, g(x, y) \leq z) \\ =& \int\int_{\{(x,y)\,:\,g(x,y)\leq z\}} f_{XY}(x, y)\, dx\, dy \end{aligned}$$

- Then $f_Z(z) = \frac{dF_Z(z)}{dz}$.

Example: Sum of Two RVs. Let $Z = X + Y$ and $(X, Y) \sim f_{XY}(x, y)$. Find $f_Z(z)$.



$$F_Z(z) = \int \int_{\{(x,y)\,:\, x+y \leq z\}} f_{XY}(x, y)\, dx\, dy = \int_{-\infty}^{\infty} \left[\int_{-\infty}^{z-y} f_{XY}(x, y)\, dx \right] dy$$

$$\Rightarrow \; f_Z(z) = \frac{dF_Z(z)}{dz} = \int_{-\infty}^{\infty} f_{XY}(z - y, y)\, dy$$

If X and Y are independent, then $f_Z(z) = \int_{-\infty}^{\infty} f_X(z - y) f_Y(y)\, dy$. This is convolution of $f_X(.)$ with $f_Y(.)$

Example: Let $Z = X + Y$, X and Y are independent, and both $\sim U[0,1]$. Find $f_Z(z)$.

We have

$$f_X(x) = f_Y(y) = \begin{cases} 1 & x, y \in [0,1] \\ 0 & \text{otherwise} \end{cases}$$

and $f_Z(z) = \int_{-\infty}^{\infty} f_X(z-y) f_Y(y)\, dy$. Therefore,

$$f_X(z-y) = \begin{cases} 1 & z - y \in [0,1], \text{ i.e., } z-1 \leq y \leq z \\ 0 & \text{otherwise} \end{cases}$$

Hence

$$f_Z(z) = \int_{\max(0,z-1)}^{\min(z,1)} dy = \begin{cases} 0 & z < 0 \\ z & 0 \leq z \leq 1 \\ 2 - z & 1 \leq z \leq 2 \\ 0 & z > 2 \end{cases}$$

Convolution of two identical rectangles yields a triangle.

Two Functions of Two Random Variables

- Suppose we are given two RVs X and Y with known joint cdf $F_{XY}(x,y)$ (or pdf $f_{XY}(x,y)$) and two functions $z = g(x,y)$ and $w = h(x,y)$. What is the joint cdf (or pdf) of random variables $Z = g(X,Y)$ and $W = H(X,Y)$?
- Approach 1: Use

$$\begin{aligned} F_{ZW}(z,w) &= P(Z \leq z, W \leq w) \\ &= P(\{(x,y) : g(x,y) \leq z, h(x,y) \leq w\}) \\ &= \int\int_{\{(x,y) : g(x,y) \leq z, h(x,y) \leq w\}} f_{XY}(x,y)\, dx\, dy \end{aligned}$$

- Then $f_{ZW}(zw) = \frac{\partial^2 F_{ZW}(z,w)}{\partial z\, \partial w}$.

- Approach 2: Suppose $(X, Y) \sim f_{XY}(x, y)$, $Z = g(X, Y)$, $W = H(X, Y)$, and $g(x, y)$ and $h(x, y)$ are differentiable.
- For a fixed (z, w), simultaneously solve $g(x, y) = z$ and $h(x, y) = w$ for real-valued (x, y). If there exists no real-valued solution, then $f_{ZW}(z, w) = 0$.
- Else, let there be n solutions $(x_1, y_1), \cdots, (x_n, y_n)$ satisfying $g(x_i, y_i) = z$ and $h(x_i, y_i) = w$. Then

$$f_{ZW}(z, w) = \sum_{i=1}^{n} \frac{f_{XY}(x_i, y_i)}{\left|\det J(x_i, y_i)\right|}$$

if $\det J(x_i, y_i) \neq 0$ for any i, where the Jacobian of transformation

$$J(x_i, y_i) = \left[\begin{array}{cc} \frac{\partial g(x,y)}{\partial x} & \frac{\partial g(x,y)}{\partial y} \\ \frac{\partial h(x,y)}{\partial x} & \frac{\partial h(x,y)}{\partial y} \end{array} \right] \Bigg|_{(x,y)=(x_i,y_i)}$$

- This method fails if $\det J(x_i, y_i) = 0$

Example: Let $Z = X + Y$ and $(X, Y) \sim f_{XY}(x, y)$. Find $f_Z(z)$.

We have $Z = X + Y = g(X, Y)$. We create a new RV: $W = X = h(X, Y)$, and then use the "expression" given earlier. For given (z, w), we have a unique solution $(n = 1)$: $(x_1, y_1) = (w, z - w)$. The Jacobian is

$$J(x,y) = \begin{bmatrix} \frac{\partial(x+y)}{\partial x} & \frac{\partial(x+y)}{\partial y} \\ \frac{\partial x}{\partial x} & \frac{\partial x}{\partial y} \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}$$

Then $\det J(x, y) = -1$ and $|\det J(x, y)| = 1$. Thus

$$\begin{aligned} f_{ZW}(z, w) =& f_{XY}(w, z - w) \\ \Rightarrow \; f_Z(z) =& \int_{-\infty}^{\infty} f_{ZW}(zw)\, dw = \int_{-\infty}^{\infty} f_{XY}(w, z - w)\, dw \end{aligned}$$

If X and Y are independent, $f_Z(z) = \int_{-\infty}^{\infty} f_X(w) f_Y(z - w)\, dw$: convolution!

Example: Let $Z = \sqrt{X^2 + Y^2} = g(X, Y)$, $(X, Y) \sim f_X(x) f_Y(y)$ and both X and $Y \sim \mathcal{N}(0, \sigma^2)$. Find $f_Z(z)$.

Approach 2: Given

$$f_{XY}(x, y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}}$$

We create a new RV: $W = X = h(X, Y)$. For given (z, w), we have $x = w$ and therefore, $y^2 = z^2 - x^2 = z^2 - w^2 \Rightarrow y = \pm\sqrt{z^2 - w^2}$ if $z^2 > w^2$.

- If $z < |w|$, there is no real-valued solution, hence, $f_{ZW}(z, w) = 0$ for $z < |w|$.
- For $z > |w| \geq 0$, there are two solutions: $(x_1, y_1) = (w, \sqrt{z^2 - w^2})$, $(x_2, y_2) = (w, -\sqrt{z^2 - w^2})$

The Jacobian is

$$J(x,y) = \begin{bmatrix} \frac{\partial\sqrt{x^2+y^2}}{\partial x} & \frac{\partial\sqrt{x^2+y^2}}{\partial y} \\ \frac{\partial x}{\partial x} & \frac{\partial x}{\partial y} \end{bmatrix} = \begin{bmatrix} \frac{x}{\sqrt{x^2+y^2}} & \frac{y}{\sqrt{x^2+y^2}} \\ 1 & 0 \end{bmatrix}$$

$$\Rightarrow\ |\det J(x,y)| = \frac{|y|}{|\sqrt{x^2+y^2}|} = \frac{\sqrt{z^2-w^2}}{z} \text{ for } (x,y) = (x_1,y_1),\ (x_2,y_2)$$

Thus, for $z > |w| \geq 0$,

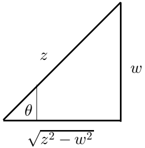
$$f_{ZW}(z,w) = \sum_{i=1}^{2} \frac{f_{XY}(x_i,y_i)}{\left|\det J(x_i,y_i)\right|} = \frac{z}{\sqrt{z^2-w^2}} \left[\frac{1}{2\pi\sigma^2} e^{-\frac{z^2}{2\sigma^2}} + \frac{1}{2\pi\sigma^2} e^{-\frac{z^2}{2\sigma^2}} \right]$$

$$\Rightarrow\ f_{ZW}(z,w) = \begin{cases} \frac{z}{\pi\sigma^2\sqrt{z^2-w^2}} e^{-\frac{z^2}{2\sigma^2}} & z > |w| \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

Noting that $z > |w| \geq 0$ is equivalent to $-z < w < z$, $z > 0$,

$$f_Z(z) = \int_{-\infty}^{\infty} f_{ZW}(z,w)\,dw = \int_{-z}^{z} \frac{z}{\pi\sigma^2\sqrt{z^2-w^2}} e^{-\frac{z^2}{2\sigma^2}}\,dw$$

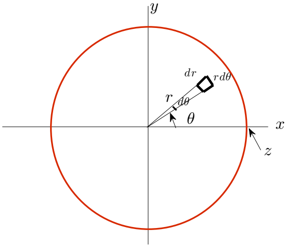
$$= \frac{ze^{-\frac{z^2}{2\sigma^2}}}{\pi\sigma^2} \int_{-z}^{z} \frac{1}{\sqrt{z^2-w^2}} dw = \frac{2ze^{-\frac{z^2}{2\sigma^2}}}{\pi\sigma^2} \int_{0}^{z} \frac{1}{\sqrt{z^2-w^2}} dw$$



Set $w = z\sin(\theta) \Rightarrow dw = z\cos(\theta)d\theta$ and $\sqrt{z^2-w^2} = z\cos(\theta)$. Then $\int_0^z \frac{1}{\sqrt{z^2-w^2}} dw = \int_0^{\pi/2} d\theta = \frac{\pi}{2}$. Thus, we have the Rayleigh pdf

$$f_Z(z) = \frac{z}{\sigma^2} e^{-\frac{z^2}{2\sigma^2}} u(z)$$

Approach 1: Clearly $F_Z(z) = 0$ for $z \leq 0$. Set $x^2 + y^2 = r^2$, $dx\,dy = r\,dr\,d\theta$:



$$F_Z(z) = \int\int_{\{(x,y):\sqrt{x^2+y^2}\leq z\}} \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}}\,dx\,dy$$

$$= \frac{1}{2\pi\sigma^2}\int_0^z r e^{-\frac{r^2}{2\sigma^2}} \left[\int_0^{2\pi} d\theta\right] dr = \frac{1}{\sigma^2}\int_0^z r e^{-\frac{r^2}{2\sigma^2}}\,dr$$

$$\Rightarrow\ f_Z(z) = \frac{dF_Z(z)}{dz} = \frac{z}{\sigma^2} e^{-\frac{z^2}{2\sigma^2}} u(z)$$

Random Vectors

- Let $X_1, X_2, \cdots, X_n$ be random variables defined on the same probability space. We define a random vector (RV) as

$$\boldsymbol{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix}$$

- $\boldsymbol{X}$ is completely specified by its joint CDF for $\boldsymbol{x} = (x_1, x_2, \cdots, x_n)$:

$$F_{\boldsymbol{X}}(\boldsymbol{x}) = P(X_1 \leq x_1, X_2 \leq x_2, \cdots, X_n \leq x_n), \quad \boldsymbol{x} \in \mathbb{R}^n$$

- If $\boldsymbol{X}$ is continuous, i.e., $F_{\boldsymbol{X}}(\boldsymbol{x})$ is a continuous function of $\boldsymbol{x}$, then $\boldsymbol{X}$ can be specified by its joint pdf

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = f_{X_1, X_2, \cdots, X_n}(x_1, x_2, \cdots, x_n) = \frac{\partial^n F_{\boldsymbol{X}}(\boldsymbol{x})}{\partial x_1 \cdots \partial x_n}$$

- If $\boldsymbol{X}$ is discrete, then it can be specified by its joint pmf

$$P_{\boldsymbol{X}}(\boldsymbol{x}) = P_{X_1, X_2, \cdots, X_n}(x_1, x_2, \cdots, x_n) = P(X_1 = x_1, \cdots, X_n = x_n)$$

- A marginal cdf (pdf, pmf) is the joint cdf (pdf, pmf) for a subset of $\{X_1, X_2, \cdots, X_n\}$, e.g., for

$$\boldsymbol{X} = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix}$$

the marginal pdf's are $f_{X_1}(x_1)$, $f_{X_2}(x_2)$, $f_{X_3}(x_3)$, $f_{X_1 X_2}(x_1, x_2)$, $f_{X_1 X_3}(x_1, x_3)$, $f_{X_3 X_2}(x_3, x_2)$.

- The marginals can be obtained from the joint in the usual way. For the previous example,

$$\begin{aligned} F_{X_1}(x_1) &= F_{X_1 X_2 X_3}(x_1, \infty, \infty) \\ f_{X_1 X_2}(x_1, x_2) &= \int_{-\infty}^{\infty} f_{X_1 X_2 X_3}(x_1, x_2, x_3)\, dx_3 \end{aligned}$$

- Conditional cdf (pdf, pmf) can also be defined in the usual way. E.g., the conditional pdf of $\boldsymbol{X}_{k+1}^n \stackrel{def}{=} (X_{k+1}, X_{k+2}, \cdots, X_n)$ given $\boldsymbol{X}^k \stackrel{def}{=} (X_1, X_2, \cdots, X_k)$ is

$$f_{\boldsymbol{X}_{k+1}^n|\boldsymbol{X}^k}(\boldsymbol{x}_{k+1}^n|\boldsymbol{x}^k) = \frac{f_{\boldsymbol{X}}(x_1, x_2, \cdots, x_n)}{f_{\boldsymbol{X}^k}(x_1, x_2, \cdots, x_k)} = \frac{f_{\boldsymbol{X}}(\boldsymbol{x})}{f_{\boldsymbol{X}^k}(\boldsymbol{x}^k)}$$

- Chain rule: We can write

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = f_{X_1}(x_1) f_{X_2|X_1}(x_2|x_1) f_{X_3|X_2,X_1}(x_3|x_2, x_1) \cdots f_{X_n|\boldsymbol{X}^{n-1}}(x_n|\boldsymbol{x}^{n-1})$$

Proof: By definition of conditional pdf, the rule holds for $n = 2$ Now use induction: suppose it holds for $n-1$. Then

$$\begin{aligned} f_{\boldsymbol{X}}(\boldsymbol{x}) = & \underbrace{f_{\boldsymbol{X}^{n-1}}(\boldsymbol{x}^{n-1})}_{iterate} f_{X_n|\boldsymbol{X}^{n-1}}(x_n|\boldsymbol{x}^{n-1}) \\ & = f_{\boldsymbol{X}^{n-2}}(\boldsymbol{x}^{n-2}) f_{X_{n-1}|\boldsymbol{X}^{n-2}}(x_{n-1}|\boldsymbol{x}^{n-2}) f_{X_n|\boldsymbol{X}^{n-1}}(x_n|\boldsymbol{x}^{n-1}) \\ & \vdots \\ & = f_{X_1}(x_1) f_{X_2|X_1}(x_2|x_1) f_{X_3|X_2,X_1}(x_3|x_2, x_1) \cdots f_{X_n|\boldsymbol{X}^{n-1}}(x_n|\boldsymbol{x}^{n-1}) \end{aligned}$$

Independence and Conditional Independence

- Independence is defined in the usual way, e.g., $X_1, X_2, \cdots, X_n$ are independent if

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \prod_{i=1}^{n} f_{X_i}(x_i) \text{ for all } (x_1, x_2, \cdots, x_n)$$

- Important special case: i.i.d. RVs: $X_1, X_2, \cdots, X_n$ are said to be independent and identically distributed (i.i.d.) if they are independent and have the same marginals.
  Example: if we flip a coin n times independently, we generate i.i.d. Bern(p) RVs $X_1, X_2, \cdots, X_n$
- RVs X_1 and X_3 are said to be conditionally independent given X_2 if

$$f_{X_1 X_3|X_2}(x_1, x_3|x_2) = f_{X_1|X_2}(x_1|x_2) f_{X_3|X_2}(x_3|x_2) \text{ for all } (x_1, x_2, x_3)$$

- Conditional independence neither implies nor is implied by independence; X_1 and X_3 independent given X_2 does not mean that X_1 and X_3 are independent (or vice versa).

- Example: Coin with random bias. Given a coin with random bias $P \sim f_P(p)$, flip it n times independently to generate the RVs $X_1, X_2, \cdots, X_n$, where $X_i = 1$ if the i-th flip is heads, $= 0$ otherwise. Example: if we flip a coin n times independently, we generate i.i.d. Bern(p) RVs $X_1, X_2, \cdots, X_n$
  - $X_1, X_2, \cdots, X_n$ are not independent.
  - However, $X_1, X_2, \cdots, X_n$ are conditionally independent given P; in fact, they are i.i.d. Bern(p) for every $P = p$.
- Example: Additive noise channel. Consider an additive noise channel with signal X, noise Z, and observation $Y = X + Z$, where X and Z are independent random variables.
  - Although X and Z are independent, they are not in general conditionally independent given Y.

Mean and Covariance Matrix

- The mean of the random vector $\boldsymbol{X}$ is defined componentwise as

$$E\{\boldsymbol{X}\} = \begin{bmatrix} E\{X_1\} & E\{X_2\} & \cdots & E\{X_n\} \end{bmatrix}^\top$$

- Denote the covariance between X_i and X_j, $\mathrm{Cov}(X_i, X_j)$, by σ_{ij} (so the variance of X_i is denoted by σ_{ii}, $\mathrm{Var}(X_i)$, or $\sigma^2_{X_i}$
- The covariance matrix of $\boldsymbol{X}$ is defined as

$$\Sigma_{\boldsymbol{X}} = E\{[\boldsymbol{X} - E\{\boldsymbol{X}\}]\,[\boldsymbol{X} - E\{\boldsymbol{X}\}]^\top\} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_{nn} \end{bmatrix}$$

- For $n = 2$, use the definition of correlation coefficient to obtain

$$\Sigma_{\boldsymbol{X}} = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix} = \begin{bmatrix} \sigma^2_{X_1} & \rho_{X_1X_2}\sigma_{X_1}\sigma_{X_2} \\ \rho_{X_1X_2}\sigma_{X_1}\sigma_{X_2} & \sigma^2_{X_2} \end{bmatrix}$$

Properties of Covariance Matrix $\Sigma_{\boldsymbol{X}}$

- $\Sigma_{\boldsymbol{X}}$ is real, and symmetric (i.e., $\Sigma_{\boldsymbol{X}} = \Sigma_{\boldsymbol{X}}^{\top}$) since $\sigma_{ij} = \sigma_{ji}$.
- $\Sigma_{\boldsymbol{X}}$ is positive semidefinite, i.e., the quadratic form

$$\boldsymbol{z}^{\top}\Sigma_{\boldsymbol{X}}\boldsymbol{z} \geq 0 \text{ for every real vector } \boldsymbol{z}$$

  Equivalently, all the eigenvalues of $\Sigma_{\boldsymbol{X}}$ are nonnegative.
- To show that $\Sigma_{\boldsymbol{X}}$ is positive semidefinite, consider

$$\begin{aligned}
\boldsymbol{z}^{\top}\Sigma_{\boldsymbol{X}}\boldsymbol{z} &= \boldsymbol{z}^{\top}E\{[\boldsymbol{X} - E\{\boldsymbol{X}\}]\,[\boldsymbol{X} - E\{\boldsymbol{X}\}]^{\top}\}\boldsymbol{z} \\
&= E\{\boldsymbol{z}^{\top}\,[\boldsymbol{X} - E\{\boldsymbol{X}\}]\,[\boldsymbol{X} - E\{\boldsymbol{X}\}]^{\top}\,\boldsymbol{z}\} \\
&= E\{\left(\boldsymbol{z}^{\top}\,[\boldsymbol{X} - E\{\boldsymbol{X}\}]\right)^{2}\} \geq 0
\end{aligned}$$

Gaussian Random Vector

- A random vector $\boldsymbol{X} = (X_1, X_2, \cdots, X_n)$ is a Gaussian random vector (or $X_1, X_2, \cdots, X_n$ are jointly Gaussian RVs) if the joint pdf is of the form

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}$$

  where $\boldsymbol{\mu} = E\{\boldsymbol{X}\}$, Σ is the covariance matrix of $\boldsymbol{X}$ and $|\Sigma| > 0$, i.e., Σ is positive definite. We use the notation $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$

- The above definition requires that $|\Sigma| > 0$. An alternative definition that does not requires $|\Sigma| > 0$ is as follows. Random vector $\boldsymbol{X}$ is Gaussian, or $X_1, X_2, \cdots, X_n$ are jointly Gaussian RVs, if

$$\boldsymbol{c}^\top \boldsymbol{X} = \sum_{i=1}^{n} c_i X_i \sim \mathcal{N}(\mu, \sigma^2)$$

  for any nonzero vector $\boldsymbol{c}$. That is, any linear combination of the components of $\boldsymbol{X}$ is a scalar Gaussian random variable.

Characteristic Function

- The characteristic function $\phi_X(\boldsymbol{\nu})$ of a random vector $\boldsymbol{X} \in \mathbb{R}^n$ is defined as (the integral is n-dimensional and $j = \sqrt{-1}$)

$$\phi_X(\boldsymbol{\nu}) = E\{e^{j\boldsymbol{\nu}^\top \boldsymbol{X}}\} = E\{e^{j\sum_{i=1}^n \nu_i X_i}\} = \int_{-\infty}^{\infty} e^{j\boldsymbol{\nu}^\top \boldsymbol{x}} f_{\boldsymbol{X}}(\boldsymbol{x})\, d\boldsymbol{x}$$

- In terms of multidimensional integration

$$\phi_X(\boldsymbol{\nu}) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} e^{j\sum_{i=1}^n \nu_i X_i} f_{\boldsymbol{X}}(x_1, \cdots, x_n)\, dx_1 \cdots dx_n$$

  It is n-dimensional Fourier transform of $f_{\boldsymbol{X}}(\boldsymbol{x})$.

- Inverse Fourier transform of $\phi_X(\boldsymbol{\nu})$ yields $f_{\boldsymbol{X}}(\boldsymbol{x})$:

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{1}{(2\pi)^n} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} e^{-j\sum_{i=1}^n \nu_i X_i} \phi_X(\nu_1, \cdots, \nu_n)\, d\nu_1 \cdots d\nu_n$$

- The characteristic function of $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ is given by

$$\phi_X(\boldsymbol{\nu}) = e^{j\boldsymbol{\nu}^\top \boldsymbol{\mu} - \frac{1}{2}\boldsymbol{\nu}^\top \boldsymbol{\Sigma} \boldsymbol{\nu}}$$

  This expression is valid whether or not $|\Sigma| > 0$. Recall that $|\Sigma| \geq 0$

Properties of Gaussian Random Vectors (GRV)

- Property 1: Uncorrelation implies independence.
  This can be verified by substituting $\sigma_{ij} = 0$ for all $i \neq j$ in the joint pdf. Then Σ becomes diagonal and so does Σ^{-1}, and the joint pdf reduces to the product of the marginals $X_i \sim \mathcal{N}(\mu_i, \sigma_{ii})$.
- Property 2: Linear transformation of a GRV yields a GRV.
  Suppose $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$, $\boldsymbol{X} \in \mathbb{R}^n$. Define $\boldsymbol{Y} = \boldsymbol{AX} + \boldsymbol{b} \in \mathbb{R}^m$. The characteristic function of $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ is given by

  $$\phi_X(\boldsymbol{\nu}) = e^{j\boldsymbol{\nu}^\top \boldsymbol{\mu} - \frac{1}{2}\boldsymbol{\nu}^\top \Sigma \boldsymbol{\nu}}$$

  The characteristic function of $\boldsymbol{Y}$ is

  $$\begin{aligned}\phi_Y(\boldsymbol{\nu}) =& E\{e^{j\boldsymbol{\nu}^\top \boldsymbol{Y}}\} = E\{e^{j\boldsymbol{\nu}^\top(\boldsymbol{AX}+\boldsymbol{b})}\} \\ =& E\{e^{j(\boldsymbol{A}^\top\boldsymbol{\nu})^\top \boldsymbol{X}}\} e^{j\boldsymbol{\nu}^\top \boldsymbol{b}} = e^{j(\boldsymbol{A}^\top\boldsymbol{\nu})^\top \boldsymbol{\mu} - \frac{1}{2}(\boldsymbol{A}^\top\boldsymbol{\nu})^\top \Sigma \boldsymbol{A}^\top \boldsymbol{\nu}} e^{j\boldsymbol{\nu}^\top \boldsymbol{b}} \\ =& e^{j\boldsymbol{\nu}^\top(\boldsymbol{A\mu}+\boldsymbol{b}) - \frac{1}{2}\boldsymbol{\nu}^\top(\boldsymbol{A}\Sigma\boldsymbol{A}^\top)\boldsymbol{\nu}}.\end{aligned}$$

  Thus, $\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{A\mu} + \boldsymbol{b}, \boldsymbol{A}\Sigma\boldsymbol{A}^\top)$.

- Property 3: Marginals of a GRV are Gaussian, i.e., if $\boldsymbol{X}$ is GRV then for any subset $\{i_1, i_2, \cdots, i_k\} \subset \{1, 2, \cdots, n\}$ of indexes, the RV

$$\boldsymbol{Y} = [X_{i_1}\ X_{i_2}\ \cdots\ X_{i_k}]^\top$$

is a GRV.
This follows from Property 2.

- The converse of Property 3 does not hold in general, i.e., Gaussian marginals do not necessarily mean that the RVs are jointly Gaussian.
- Property 4: Conditionals of a GRV are Gaussian, more specifically, if

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1 \\ \cdots \\ \boldsymbol{X}_2 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}_1 \\ \cdots \\ \boldsymbol{\mu}_2 \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \vdots & \boldsymbol{\Sigma}_{12} \\ \cdots & \vdots & \cdots \\ \boldsymbol{\Sigma}_{21} & \vdots & \boldsymbol{\Sigma}_{22} \end{bmatrix} \right)$$

where $\boldsymbol{X}_1$ is a k-dim RV and $\boldsymbol{X}_2$ is an $n-k$-dim RV, then

$$\boldsymbol{X}_2 \Big| \{\boldsymbol{X}_1 = \boldsymbol{x}_1\} \sim \mathcal{N}\left(\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\boldsymbol{x}_1 - \boldsymbol{\mu}_1) + \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\right)$$

Mean-Square Estimation

- Consider two random vectors $\boldsymbol{X} \in \mathbb{R}^n$ and $\boldsymbol{Y} \in \mathbb{R}^m$, where $\boldsymbol{Y}$ is observed but $\boldsymbol{X}$ is not. (For example, $\boldsymbol{Y}$ could be noisy measurements of some function of $\boldsymbol{X}$.) Given $\boldsymbol{Y}$, we wish to estimate $\boldsymbol{X}$ as $\hat{\boldsymbol{X}}(\boldsymbol{Y}) = \boldsymbol{g}(\boldsymbol{Y})$ to minimize the mean-square error (MSE)

$$\boldsymbol{g}(\boldsymbol{Y}) = \arg\min E\{\|\boldsymbol{X} - \boldsymbol{g}(\boldsymbol{Y})\|^2\}$$

  where

$$\|\boldsymbol{x}\|^2 = \boldsymbol{x}^\top \boldsymbol{x} = \sum_{i=1}^{n} x_i^2$$

- The solution is given by

$$\boldsymbol{g}(\boldsymbol{Y}) = E\{\boldsymbol{X} \mid \boldsymbol{Y}\} \;\Rightarrow\; \boldsymbol{g}(\boldsymbol{y}) = E\{\boldsymbol{X} \mid \boldsymbol{Y} = \boldsymbol{y}\}$$

  We have

$$E\{\|\boldsymbol{X} - \boldsymbol{g}(\boldsymbol{Y})\|^2\} = E_Y\{E_X\{\|\boldsymbol{X} - \boldsymbol{g}(\boldsymbol{Y})\|^2 \mid \boldsymbol{Y} = \boldsymbol{y}\}\}$$

  Now for each $\boldsymbol{Y} = \boldsymbol{y}$, $E_X\{\|\boldsymbol{X} - \boldsymbol{g}(\boldsymbol{y})\|^2 \mid \boldsymbol{Y} = \boldsymbol{y}\}$ is minimized if $\boldsymbol{g}(\boldsymbol{y}) = E\{\boldsymbol{X} \mid \boldsymbol{Y} = \boldsymbol{y}\}$

- To establish the claim, consider the first-order optimality condition:

$$E_X\{\|\boldsymbol{X} - \boldsymbol{g}(\boldsymbol{y})\|^2 \mid \boldsymbol{Y} = \boldsymbol{y}\} = \int \sum_{i=1}^{n} [x_i - g_i(\boldsymbol{y})]^2 f_{\boldsymbol{X}|\boldsymbol{Y}}(\boldsymbol{x}|\boldsymbol{y})\, d\boldsymbol{x}$$

$$0 = \frac{\partial()}{\partial g_\ell(\boldsymbol{y})} = -2 \int [x_\ell - g_\ell(\boldsymbol{y})] f_{\boldsymbol{X}|\boldsymbol{Y}}(\boldsymbol{x}|\boldsymbol{y})\, d\boldsymbol{x}\,, \quad \ell = 1, 2, \cdots, n$$

$$\Rightarrow g_\ell(\boldsymbol{y}) = E\{X_\ell \mid \boldsymbol{Y} = \boldsymbol{y}\} \Rightarrow \boldsymbol{g}(\boldsymbol{y}) = E\{\boldsymbol{X} \mid \boldsymbol{Y} = \boldsymbol{y}\}$$

- Thus,

$$E\{\|\boldsymbol{X} - \boldsymbol{g}(\boldsymbol{Y})\|^2\} = \int E_X\{\|\boldsymbol{X} - \boldsymbol{g}(\boldsymbol{y})\|^2 \mid \boldsymbol{Y} = \boldsymbol{y}\} f_{\boldsymbol{Y}}(\boldsymbol{y})\, d\boldsymbol{y}$$

  is minimized for $\boldsymbol{g}(\boldsymbol{Y}) = E\{\boldsymbol{X} \mid \boldsymbol{Y}\}$

- Thus $E\{\boldsymbol{X} \mid \boldsymbol{Y}\}$ minimizes the MSE conditioned on every $\boldsymbol{Y} = \boldsymbol{y}$ and not just its average over $\boldsymbol{Y}$!

Jointly Gaussian $\boldsymbol{X}$ and $\boldsymbol{Y}$

- Suppose ($\boldsymbol{X} \in \mathbb{R}^n$ and $\boldsymbol{Y} \in \mathbb{R}^m$)

$$\boldsymbol{Z} = \begin{bmatrix} \boldsymbol{Y} \\ \cdots \\ \boldsymbol{X} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}_Y \\ \cdots \\ \boldsymbol{\mu}_X \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{YY} & \vdots & \boldsymbol{\Sigma}_{YX} \\ \cdots & \vdots & \cdots \\ \boldsymbol{\Sigma}_{XY} & \vdots & \boldsymbol{\Sigma}_{XX} \end{bmatrix} \right)$$

- Recall that

$$\boldsymbol{X}\Big|\{\boldsymbol{Y} = \boldsymbol{y}\} \sim \mathcal{N}\left(\boldsymbol{\Sigma}_{XY}\boldsymbol{\Sigma}_{YY}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}_Y) + \boldsymbol{\mu}_X, \boldsymbol{\Sigma}_{XX} - \boldsymbol{\Sigma}_{XY}\boldsymbol{\Sigma}_{YY}^{-1}\boldsymbol{\Sigma}_{YX} \right)$$

- Therefore, the optimal estimator is

$$E\{\boldsymbol{X} \mid \boldsymbol{Y}\} = \boldsymbol{\Sigma}_{XY}\boldsymbol{\Sigma}_{YY}^{-1}(\boldsymbol{Y} - \boldsymbol{\mu}_Y) + \boldsymbol{\mu}_X$$

and the corresponding minimum MSE (MMSE) $E\{\|\boldsymbol{X} - E\{\boldsymbol{X} \mid \boldsymbol{Y}\}\|^2\}$ is

$$\operatorname{tr}\left(\boldsymbol{\Sigma}_{XX} - \boldsymbol{\Sigma}_{XY}\boldsymbol{\Sigma}_{YY}^{-1}\boldsymbol{\Sigma}_{YX} \right)$$

Example 8.21 (Gubner) If scalar signal $X \sim \mathcal{N}(0,1)$ and noise $W \sim \mathcal{N}(0,\sigma^2)$, find the MMSE estimator of X given noisy observation $Y = X + W$. The signal and noise are independent.

We have $\Sigma_{XY} = \Sigma_{XX} = 1$ and $\Sigma_{YY} = \Sigma_{XX} + \Sigma_{WW} = 1 + \sigma^2$ since $\Sigma_{XW} = 0$. Also $\mu_X = \mu_Y = 0$. Therefore,

$$\begin{aligned} E\{X \mid Y = y\} &= \Sigma_{XY}\Sigma_{YY}^{-1}(y - \mu_Y) + \mu_X \\ &= \frac{1}{1+\sigma^2}(y-0) + 0 = \frac{y}{1+\sigma^2} \end{aligned}$$

The corresponding MMSE is

$$\Sigma_{XX} - \Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX} = 1 - \frac{1}{1+\sigma^2} = \frac{\sigma^2}{1+\sigma^2}$$

Linear Mean-Square Estimation

- Consider two random vectors $\boldsymbol{X} \in \mathbb{R}^n$ and $\boldsymbol{Y} \in \mathbb{R}^m$, where $\boldsymbol{Y}$ is observed but $\boldsymbol{X}$ is not. Given $\boldsymbol{Y}$, we wish to estimate $\boldsymbol{X}$ as $\hat{\boldsymbol{X}}(\boldsymbol{Y}) = \boldsymbol{A}\boldsymbol{Y} + \boldsymbol{b}$ to minimize the mean-square error (MSE)

$$\{\hat{\boldsymbol{A}}, \hat{\boldsymbol{b}}\} = \arg\min_{\boldsymbol{A},\boldsymbol{b}} E\{\|\boldsymbol{X} - \hat{\boldsymbol{X}}(\boldsymbol{Y})\|^2\}$$

  where $\|\boldsymbol{x}\|^2 = \boldsymbol{x}^\top \boldsymbol{x} = \sum_{i=1}^n x_i^2$.

- Let

$$E\left\{\begin{bmatrix}\boldsymbol{X}\\ \boldsymbol{Y}\end{bmatrix}\right\} = \begin{bmatrix}\boldsymbol{m}_X\\ \boldsymbol{m}_Y\end{bmatrix}, \quad \mathrm{cov}\left(\begin{bmatrix}\boldsymbol{X}\\ \boldsymbol{Y}\end{bmatrix}\right) = \begin{bmatrix}\mathrm{cov}(\boldsymbol{X},\boldsymbol{X}) & \mathrm{cov}(\boldsymbol{X},\boldsymbol{Y})\\ \mathrm{cov}(\boldsymbol{Y},\boldsymbol{X}) & \mathrm{cov}(\boldsymbol{Y},\boldsymbol{Y})\end{bmatrix}$$

- Define $\tilde{\boldsymbol{X}} = \boldsymbol{X} - \boldsymbol{m}_X$ and $\tilde{\boldsymbol{Y}} = \boldsymbol{Y} - \boldsymbol{m}_Y$. Then $E\{\tilde{\boldsymbol{X}}\} = \boldsymbol{0} = E\{\tilde{\boldsymbol{Y}}\}$, $\mathrm{cov}(\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{X}}) = \mathrm{cov}(\boldsymbol{X}, \boldsymbol{X})$, $\mathrm{cov}(\tilde{\boldsymbol{Y}}, \tilde{\boldsymbol{Y}}) = \mathrm{cov}(\boldsymbol{Y}, \boldsymbol{Y})$, and $\mathrm{cov}(\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{Y}}) = \mathrm{cov}(\boldsymbol{X}, \boldsymbol{Y}) = E\{[\boldsymbol{X}-\boldsymbol{m}_X][\boldsymbol{Y}-\boldsymbol{m}_Y]^\top\} = (\mathrm{cov}(\boldsymbol{Y}, \boldsymbol{X}))^\top$.

- Set $\boldsymbol{X} = \tilde{\boldsymbol{X}} + \boldsymbol{m}_X$ etc, and rewrite

$$\begin{aligned}\text{MSE} &= E\{\|\boldsymbol{X} - (\boldsymbol{AY} + \boldsymbol{b})\|^2\} = E\{[\boldsymbol{X} - (\boldsymbol{AY} + \boldsymbol{b})]^\top [\boldsymbol{X} - (\boldsymbol{AY} + \boldsymbol{b})]\} \\ &= E\{\|\tilde{\boldsymbol{X}} + \boldsymbol{m}_X - \boldsymbol{A}(\tilde{\boldsymbol{Y}} + \boldsymbol{m}_Y) - \boldsymbol{b}\|^2\} \\ &= E\{\|\tilde{\boldsymbol{X}} - \boldsymbol{A}\tilde{\boldsymbol{Y}}\|^2\} + \|\boldsymbol{m}_X - \boldsymbol{A}\boldsymbol{m}_Y - \boldsymbol{b}\|^2 \\ &\qquad + 2\underbrace{E\{(\tilde{\boldsymbol{X}} - \boldsymbol{A}\tilde{\boldsymbol{Y}})^\top\}}_{=\mathbf{0}}(\boldsymbol{m}_X - \boldsymbol{A}\boldsymbol{m}_Y - \boldsymbol{b})\end{aligned}$$

- Therefore, to minimize MSE, we must estimate $\boldsymbol{b}$ as

$$\hat{\boldsymbol{b}} = \boldsymbol{m}_X - \boldsymbol{A}\boldsymbol{m}_Y = \boldsymbol{m}_X - \hat{\boldsymbol{A}}\boldsymbol{m}_Y$$

  which minimizes the second term $\|\boldsymbol{m}_X - \boldsymbol{A}\boldsymbol{m}_Y - \boldsymbol{b}\|^2$ whatever the choice of $\hat{\boldsymbol{A}}$, and it does not affect $E\{\|\tilde{\boldsymbol{X}} - \boldsymbol{A}\tilde{\boldsymbol{Y}}\|^2\}$.

- Now consider minimization of $E\{\|\tilde{\boldsymbol{X}} - \boldsymbol{A}\tilde{\boldsymbol{Y}}\|^2\}$ w.r.t. $\boldsymbol{A}$. Using $\boldsymbol{\Sigma}_{XY} = \text{cov}(\boldsymbol{X}, \boldsymbol{Y})$, rewrite (explained in next slide)

$$E\{\|\tilde{\boldsymbol{X}} - \boldsymbol{A}\tilde{\boldsymbol{Y}}\|^2\} = \text{tr}(\boldsymbol{\Sigma}_{XX}) + \text{tr}(\boldsymbol{\Sigma}_{YY}\boldsymbol{A}^\top\boldsymbol{A}) - \text{tr}(\boldsymbol{\Sigma}_{YX}\boldsymbol{A}) - \text{tr}(\boldsymbol{\Sigma}_{XY}\boldsymbol{A}^\top)$$

- Notation "tr" stands for the trace operator. This operator acts on square matrices and it equals the sum of its diagonal elements, i.e., $\mathrm{tr}(\boldsymbol{C}) = \sum_{i=1}^{n} C_{ii}$ for $\boldsymbol{C} \in \mathbb{R}^{n\times n}$, where C_{ii} is the (i,i)th elements of $\boldsymbol{C}$.
- We have $\mathrm{tr}(\boldsymbol{AB}) = \mathrm{tr}(\boldsymbol{BA})$. Also, for a scalar y, $y = \mathrm{tr}(y)$. For instance, $\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} = \mathrm{tr}(\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}) = \mathrm{tr}(\boldsymbol{A}\boldsymbol{x}\boldsymbol{x}^\top)$.
- We have

$$
\begin{aligned}
&E\{\|\tilde{\boldsymbol{X}} - \boldsymbol{A}\tilde{\boldsymbol{Y}}\|^2\} = E\{[\tilde{\boldsymbol{X}} - \boldsymbol{A}\tilde{\boldsymbol{Y}}]^\top[\tilde{\boldsymbol{X}} - \boldsymbol{A}\tilde{\boldsymbol{Y}}]\} \\
&\quad = E\{\tilde{\boldsymbol{X}}^\top\tilde{\boldsymbol{X}} + \tilde{\boldsymbol{Y}}^\top\boldsymbol{A}^\top\boldsymbol{A}\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{X}}^\top\boldsymbol{A}\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{Y}}^\top\boldsymbol{A}^\top\tilde{\boldsymbol{X}}\} \\
&\quad = E\{\tilde{\boldsymbol{X}}^\top\tilde{\boldsymbol{X}}\} + E\{\tilde{\boldsymbol{Y}}^\top\boldsymbol{A}^\top\boldsymbol{A}\tilde{\boldsymbol{Y}}\} - E\{\tilde{\boldsymbol{X}}^\top\boldsymbol{A}\tilde{\boldsymbol{Y}}\} - E\{\tilde{\boldsymbol{Y}}^\top\boldsymbol{A}^\top\tilde{\boldsymbol{X}}\} \\
&\quad = \mathrm{tr}(\boldsymbol{\Sigma}_{XX}) + \mathrm{tr}(\boldsymbol{\Sigma}_{YY}\boldsymbol{A}^\top\boldsymbol{A}) - \mathrm{tr}(\boldsymbol{\Sigma}_{YX}\boldsymbol{A}) - \mathrm{tr}(\boldsymbol{\Sigma}_{XY}\boldsymbol{A}^\top) \qquad (1)
\end{aligned}
$$

where we have used the fact that
$E\{\tilde{\boldsymbol{X}}^\top\tilde{\boldsymbol{X}}\} = E\{\mathrm{tr}(\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^\top)\} = \mathrm{tr}(E\{\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^\top\}) = \mathrm{tr}(\boldsymbol{\Sigma}_{XX})$, etc.

Matrix Calculus

- The gradient $\nabla f(\boldsymbol{x})$ of the differentiable scalar function $f : \mathbb{R}^n \to \mathbb{R}$, $f(\boldsymbol{x})$, is a column vector $\in \mathbb{R}^n$, given by

$$\nabla f(\boldsymbol{x}) = \begin{bmatrix} \frac{\partial f(\boldsymbol{x})}{\partial x_1} \\ \frac{\partial f(\boldsymbol{x})}{\partial x_2} \\ \vdots \\ \frac{\partial f(\boldsymbol{x})}{\partial x_n} \end{bmatrix} . \tag{2}$$

- $f(\boldsymbol{x}) = \boldsymbol{b}^\top \boldsymbol{x} = \sum_{i=1}^n b_i x_i$. Then $\nabla f(\boldsymbol{x}) = \boldsymbol{b}$. Since $\boldsymbol{b}^\top \boldsymbol{x} = \boldsymbol{x}^\top \boldsymbol{b}$, $\nabla \boldsymbol{x}^\top \boldsymbol{b} = \boldsymbol{b}$.
- $f(\boldsymbol{x}) = \boldsymbol{x}^\top A \boldsymbol{x} = \sum_{i=1}^n \sum_{j=1}^n x_i A_{ij} x_j$. Then $\nabla f(\boldsymbol{x}) = A\boldsymbol{x} + A^\top \boldsymbol{x}$. This follows from

$$\frac{\partial \boldsymbol{x}^\top A \boldsymbol{x}}{\partial x_\ell} = \sum_{j=1}^n A_{\ell j} x_j + \sum_{i=1}^n x_i A_{i\ell} = \sum_{j=1}^n A_{\ell j} x_j + \sum_{j=1}^n A^\top_{\ell j} x_j = [A\boldsymbol{x} + A^\top \boldsymbol{x}]_\ell .$$

- Similarly, define gradient $\nabla f(X)$ of the differentiable scalar function $f : \mathbb{R}^{n\times m} \to \mathbb{R}$, $f(X)$, as a matrix $\in \mathbb{R}^{n\times m}$, whose (i,j)th element is given by
$$[\nabla f(X)]_{ij} = \frac{\partial f(X)}{\partial X_{ij}}\,. \tag{3}$$
We will also write $\nabla f(X)$ as $\nabla_X f(X)$ and $\frac{\partial f(X)}{\partial X}$
- Some useful results: A, X and B below are all matrices.
  - $\frac{\partial \mathrm{tr}(AXB)}{\partial X} = A^\top B^\top$.
  - $\frac{\partial \mathrm{tr}(AX^\top B)}{\partial X} = BA$
  - $\frac{\partial \mathrm{tr}(X^\top AXB)}{\partial X} = 2AXB$ if $A = A^\top$ and $B = B^\top$.
  - $\frac{\partial \mathrm{tr}(XBX^\top A)}{\partial X} = 2AXB$ if $A = A^\top$ and $B = B^\top$.

Back to LMMSE Estimation

- To minimize

$$C = \mathrm{tr}(\boldsymbol{\Sigma}_{XX}) + \mathrm{tr}(\boldsymbol{\Sigma}_{YY}\boldsymbol{A}^\top\boldsymbol{A}) - \mathrm{tr}(\boldsymbol{\Sigma}_{YX}\boldsymbol{A}) - \mathrm{tr}(\boldsymbol{\Sigma}_{XY}\boldsymbol{A}^\top)$$

w.r.t. $\boldsymbol{A}$, we set

$$\begin{aligned}\mathbf{0} &= \frac{\partial C}{\partial \boldsymbol{A}} \\ &= \mathbf{0} + 2\boldsymbol{A}\boldsymbol{\Sigma}_{YY} - \boldsymbol{\Sigma}_{YX}^\top - \boldsymbol{\Sigma}_{XY} \\ &= 2\boldsymbol{A}\boldsymbol{\Sigma}_{YY} - 2\boldsymbol{\Sigma}_{XY}\end{aligned}$$

- Thus, optimal $\boldsymbol{A}$, denoted by $\hat{\boldsymbol{A}}$, satisfies

$$\hat{\boldsymbol{A}}\boldsymbol{\Sigma}_{YY} = \boldsymbol{\Sigma}_{XY} \;\Rightarrow\; \hat{\boldsymbol{A}} = \boldsymbol{\Sigma}_{XY}\boldsymbol{\Sigma}_{YY}^{-1} \text{ if the inverse exists.}$$

- Thus, the solution to $\{\hat{\boldsymbol{A}}, \hat{\boldsymbol{b}}\} = \arg\min_{\boldsymbol{A},\boldsymbol{b}} E\{\|\boldsymbol{X} - (\boldsymbol{AY} + \boldsymbol{b})\|^2\}$ is given by $\hat{\boldsymbol{A}} = \boldsymbol{\Sigma}_{XY}\boldsymbol{\Sigma}_{YY}^{-1}$ and $\hat{\boldsymbol{b}} = \boldsymbol{m}_X - \hat{\boldsymbol{A}}\boldsymbol{m}_Y$, leading to

$$\hat{\boldsymbol{X}} = \hat{\boldsymbol{A}}\boldsymbol{Y} + \hat{\boldsymbol{b}} = \hat{\boldsymbol{A}}(\boldsymbol{Y} - \boldsymbol{m}_Y) + \boldsymbol{m}_X$$

- The optimal MSE is

$$\begin{aligned}
\text{MSE}_o &= E\{\|\boldsymbol{X} - (\hat{\boldsymbol{A}}\boldsymbol{Y} + \hat{\boldsymbol{b}})\|^2\} = E\{\|\tilde{\boldsymbol{X}} - \hat{\boldsymbol{A}}\tilde{\boldsymbol{Y}}\|^2\} \\
&= \text{tr}(\boldsymbol{\Sigma}_{XX}) + \text{tr}(\boldsymbol{\Sigma}_{YY}\hat{\boldsymbol{A}}^\top\hat{\boldsymbol{A}}) - \text{tr}(\boldsymbol{\Sigma}_{YX}\hat{\boldsymbol{A}}) - \text{tr}(\boldsymbol{\Sigma}_{XY}\hat{\boldsymbol{A}}^\top) \\
&= \text{tr}(\boldsymbol{\Sigma}_{XX}) - \text{tr}(\boldsymbol{\Sigma}_{YX}\hat{\boldsymbol{A}}) = \text{tr}(\boldsymbol{\Sigma}_{XX}) - \text{tr}(\boldsymbol{\Sigma}_{XY}\boldsymbol{\Sigma}_{YY}^{-1}\boldsymbol{\Sigma}_{YX})
\end{aligned}$$

where we have used
$\text{tr}(\boldsymbol{\Sigma}_{YY}\hat{\boldsymbol{A}}^\top\hat{\boldsymbol{A}}) - \text{tr}(\boldsymbol{\Sigma}_{XY}\hat{\boldsymbol{A}}^\top) = \text{tr}((\hat{\boldsymbol{A}}\boldsymbol{\Sigma}_{YY} - \boldsymbol{\Sigma}_{XY})\hat{\boldsymbol{A}}^\top) = 0$

Example

Let X be the random variable representing a signal with mean m_X and variance P. The observations are $Y_i = X + Z_i$, for $i = 1, 2, \cdots, n$, where the Z_is are zero-mean, uncorrelated noise with variance σ^2, and X and Z_is are also uncorrelated. Find the MMSE linear estimate of X given $\boldsymbol{Y} = [Y_1 \ Y_2 \ \cdots \ Y_n]^\top$, and its MSE.

Solution: We have $\mathrm{cov}(X, Y_i) = \mathrm{cov}(X, X) = P$, and

$$\begin{aligned} \mathrm{cov}(Y_i, Y_j) &= \mathrm{cov}(X + Z_i, X + Z_j) = \mathrm{cov}(X, X) + \mathrm{cov}(Z_i, Z_j) + 0 \\ &= \begin{cases} P + \sigma^2 & \text{if } i = j \\ P & \text{if } i \neq j \end{cases} \end{aligned}$$

Thus, $\Sigma_{XX} = P$, $\boldsymbol{\Sigma}_{XY} = P\mathbf{1}_n^\top$, and $\boldsymbol{\Sigma}_{YY} = P\mathbf{1}_n\mathbf{1}_n^\top + \sigma^2\boldsymbol{I}_n$ where $\boldsymbol{I}_n$ denotes the $n \times n$ identity matrix and $\mathbf{1}_n$ denotes the n-dimensional column vector of all ones. The matrix inversion lemma states

$$(A + BCD)^{-1} = A^{-1} - A^{-1}B(C^{-1} + DA^{-1}B)^{-1}DA^{-1}$$

$$\Rightarrow \boldsymbol{\Sigma}_{YY}^{-1} = (\sigma^2\boldsymbol{I}_n + \mathbf{1}_n P \mathbf{1}_n^\top)^{-1} = \frac{1}{\sigma^2}\left(\boldsymbol{I}_n - \frac{P}{\sigma^2 + nP}\mathbf{1}_n\mathbf{1}_n^\top\right)$$

We have $\hat{\boldsymbol{A}} = \boldsymbol{\Sigma}_{XY}\boldsymbol{\Sigma}_{YY}^{-1}$ leading to

$$\begin{aligned}
\hat{\boldsymbol{A}}\boldsymbol{Y} &= \frac{P}{\sigma^2}\mathbf{1}_n^\top\left(\boldsymbol{I}_n - \frac{P}{\sigma^2+nP}\mathbf{1}_n\mathbf{1}_n^\top\right)\boldsymbol{Y} \\
&= \frac{P}{\sigma^2}\left(\mathbf{1}_n^\top - \frac{P}{\sigma^2+nP}n\mathbf{1}_n^\top\right)\boldsymbol{Y} \\
&= \frac{P}{\sigma^2}\left(1 - \frac{nP}{\sigma^2+nP}\right)\sum_{i=1}^n Y_i = \frac{P}{\sigma^2+nP}\sum_{i=1}^n Y_i
\end{aligned}$$

$$\begin{aligned}
\Rightarrow\ \hat{X} = \hat{\boldsymbol{A}}\tilde{\boldsymbol{Y}} + m_X &= \frac{P}{\sigma^2+nP}\sum_{i=1}^n (Y_i - m_X) + m_X \\
&= \frac{P}{\sigma^2+nP}\sum_{i=1}^n Y_i + \frac{\sigma^2}{\sigma^2+nP}m_X
\end{aligned}$$

Optimal MSE is (you work the details!)

$$\mathrm{MSE}_o = \frac{P\sigma^2}{\sigma^2+nP}$$