Lecture Notes 6
Random Vectors

- Joint, Marginal, and Conditional CDF, PDF, PMF
- Independence and Conditional Independence
- Mean and Covariance Matrix
- Mean and Variance of Sum of RVs
- Gaussian Random Vectors
- MSE Estimation: the Vector Case

Specifying Random Vectors

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

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

- $\mathbf{X}$ is completely specified by its joint cdf for $\mathbf{x} = (x_1, x_2, \ldots, x_n)$:

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

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

$$f_{\mathbf{X}}(\mathbf{x}) = f_{X_1, X_2, \ldots, X_n}(x_1, x_2, \ldots, x_n), \quad \mathbf{x} \in \mathbf{R}^n$$

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

$$p_{\mathbf{X}}(\mathbf{x}) = p_{X_1, X_2, \ldots, X_n}(x_1, x_2, \ldots, x_n), \quad \mathbf{x} \in \mathcal{X}^n$$

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

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

the marginals are

$$f_{X_1}(x_1)\,,\ f_{X_2}(x_2)\,,\ f_{X_3}(x_3)$$
$$f_{X_1,X_2}(x_1,x_2)\,,\ f_{X_1,X_3}(x_1,x_3)\,,\ f_{X_2,X_3}(x_2,x_3)$$

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

$$F_{X_1}(x_1) = \lim_{x_2,x_3\to\infty} F_{\mathbf{X}}(x_1,x_2,x_3)$$
$$f_{X_1,X_2}(x_1,x_2) = \int_{-\infty}^{\infty} f_{X_1,X_2,X_3}(x_1,x_2,x_3)\,dx_3$$

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

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

- *Chain Rule*: We can write

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

Proof: By induction. The chain rule holds for $n = 2$ by definition of conditional pdf. Now suppose it is true for $n-1$. Then

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{x}) &= f_{\mathbf{X}^{n-1}}(\mathbf{x}^{n-1}) f_{X_n|\mathbf{X}^{n-1}}(x_n|\mathbf{x}^{n-1}) \\ &= f_{X_1}(x_1) f_{X_2|X_1}(x_2|x_1)\cdots f_{X_{n-1}|\mathbf{X}^{n-2}}(x_{n-1}|\mathbf{x}^{n-2}) f_{X_n|\mathbf{X}^{n-1}}(x_n|\mathbf{x}^{n-1})\,, \end{aligned}$$

which completes the proof

Independence and Conditional Independence

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

$$f_{\mathbf{X}}(\mathbf{x}) = \prod_{i=1}^{n} f_{X_i}(x_i) \quad \text{for all } (x_1, \ldots, x_n)$$

- Important special case, *i.i.d. r.v.s*: $X_1, X_2, \ldots, X_n$ are said to be *independent, identically distributed* (i.i.d.) if they are independent and have the same marginals

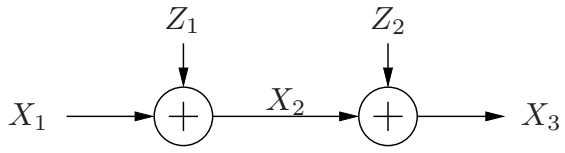
  Example: if we flip a coin n times independently, we generate i.i.d. $\mathrm{Bern}(p)$ r.v.s. $X_1, X_2, \ldots, X_n$

- R.v.s X_1 and X_3 are said to be *conditionally independent* given X_2 if

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

- 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: *Series Binary Symmetric Channels*



  Here $X_1 \sim \mathrm{Bern}(p)$, $Z_1 \sim \mathrm{Bern}(\epsilon_1)$, and $Z_2 \sim \mathrm{Bern}(\epsilon_2)$, where X_1, Z_1, Z_2 are independent and $X_3 = X_1 + Z_1 + Z_2 \bmod 2 = X_1 \oplus Z_1 \oplus Z_2$

  - In general, X_1 and X_3 are not independent
  - However, X_1 and X_3 are conditionally independent given X_2
  - Also X_1 and Z_1 are independent but not conditionally independent given X_2

- Example: *Coin with Random Bias*. Given a coin with random bias $P \sim f_P(p)$, flip it n times independently to generate the r.v.s $X_1, X_2, \ldots, X_n$, where $X_i = 1$ if i-th flip is heads, 0 otherwise

  - $X_1, X_2, \ldots, X_n$ are *not* independent
  - However, $X_1, X_2, \ldots, X_n$ are conditionally independent given P; in fact, for any $P = p$, they are i.i.d. $\mathrm{Bern}(p)$

Mean and Covariance Matrix

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

$$\mathrm{E}(\mathbf{X}) = \begin{bmatrix} \mathrm{E}(X_1) & \mathrm{E}(X_2) & \cdots & \mathrm{E}(X_n) \end{bmatrix}^T$$

- 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 $\mathbf{X}$ is defined as

$$\Sigma_{\mathbf{X}} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_{nn} \end{bmatrix}$$

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

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

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

- $\Sigma_{\mathbf{X}}$ is *real* and *symmetric* (since $\sigma_{ij} = \sigma_{ji}$)

- $\Sigma_{\mathbf{X}}$ is *nonnegative definite*, i.e., the *quadratic form*

$$\mathbf{a}^T \Sigma_{\mathbf{X}} \mathbf{a} \geq 0 \quad \text{for any real vector } \mathbf{a}$$

  Equivalently, all the *eigenvalues* of $\Sigma_{\mathbf{X}}$ are nonnegative, and also all *leading principal minors* are nonnegative

- To show that $\Sigma_{\mathbf{X}}$ is nonnegative definite we write

$$\Sigma_{\mathbf{X}} = \mathrm{E}\left[(\mathbf{X} - \mathrm{E}(\mathbf{X}))(\mathbf{X} - \mathrm{E}(\mathbf{X}))^T\right],$$

  i.e., as the expectation of an *outer product*. Thus

$$\begin{aligned} \mathbf{a}^T \Sigma_{\mathbf{X}} \mathbf{a} &= \mathbf{a}^T \mathrm{E}\left[(\mathbf{X} - \mathrm{E}(\mathbf{X}))(\mathbf{X} - \mathrm{E}(\mathbf{X}))^T\right] \mathbf{a} \\ &= \mathrm{E}\left[\mathbf{a}^T(\mathbf{X} - \mathrm{E}(\mathbf{X}))(\mathbf{X} - \mathrm{E}(\mathbf{X}))^T\mathbf{a}\right] \\ &= \mathrm{E}\left[(\mathbf{a}^T(\mathbf{X} - \mathrm{E}(\mathbf{X})))^2\right] \geq 0 \end{aligned}$$

Which of the Following Can Be a Covariance Matrix ?

1. $\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$ 2. $\begin{bmatrix} 1 & 2 & 1 \\ 2 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$ 3. $\begin{bmatrix} 1 & 0 & 1 \\ 1 & 2 & 1 \\ 0 & 1 & 3 \end{bmatrix}$

4. $\begin{bmatrix} -1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$ 5. $\begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 3 \end{bmatrix}$ 6. $\begin{bmatrix} 1 & 2 & 3 \\ 2 & 4 & 6 \\ 3 & 6 & 9 \end{bmatrix}$

Mean and Variance of Sum of RVs

- Let $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ be a RV and let Y be the sum of the X_is. In vector notation
$$Y = \mathbf{1}^T \mathbf{X},$$
where $\mathbf{1}$ is the all 1 vector

  By linearity of expectation, the expected value of Y is
$$\mathrm{E}(Y) = \mathrm{E}\left(\mathbf{1}^T \mathbf{X}\right) = \mathbf{1}^T \mathrm{E}(\mathbf{X}) = \sum_{i=1}^{n} \mathrm{E}(X_i)$$

- Example: *Mean of Binomial r.v.* One way to define a binomial r.v. is as follows: Flip a coin with bias p independently n times and define the Bernoulli r.v. X_i to be 1 if the i-th flip is a head and 0 if it is a tail. Let $Y = \sum_{i=1}^{n} X_i$. Then Y is a binomial r.v. Thus
$$\mathrm{E}(Y) = \sum_{i=1}^{n} \mathrm{E}(X_i) = np$$
Note that we did not need independence for this result to hold, i.e., the result holds even if the coin flips are not independent

Let's compute the variance of Y:

$$\begin{aligned}
\mathrm{Var}(Y) &= \mathrm{E}\left[(Y - \mathrm{E}(Y))^2\right] \\
&= \mathrm{E}\left[(\mathbf{1}^T(\mathbf{X} - \mathrm{E}(\mathbf{X}))^2\right] \\
&= \mathrm{E}\left[\mathbf{1}^T(\mathbf{X} - \mathrm{E}(\mathbf{X}))(\mathbf{X} - \mathrm{E}(\mathbf{X}))^T\mathbf{1}\right] \\
&= \mathbf{1}^T\Sigma_{\mathbf{X}}\mathbf{1} \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n}\mathrm{E}[(X_i - \mathrm{E}(X_i))(X_j - \mathrm{E}(X_j))] \\
&= \sum_{i=1}^{n}\mathrm{Var}(X_i) + \sum_{i=1}^{n}\sum_{j\neq i}^{n}\mathrm{Cov}(X_i, X_j)
\end{aligned}$$

If the r.v.s are independent, then $\mathrm{Cov}(X_i, X_j) = 0$, for all $i \neq j$, and

$$\mathrm{Var}(Y) = \sum_{i=1}^{n}\mathrm{Var}(X_i)$$

Note that this result requires only that $\mathrm{Cov}(X_i, X_j) = 0$ for all $i \neq j$, i.e., that the r.v.s are uncorrelated (which is in general weaker than independence)

- Example: *Variance of Binomial r.v.* Again express $Y = \sum_{i=1}^{n} X_i$, where the X_is are i.i.d. $\mathrm{Bern}(p)$. Since the X_is are independent, $\mathrm{Cov}(X_i, X_j) = 0$, for all $i \neq j$. Thus

$$\mathrm{Var}(Y) = \sum_{i=1}^{n}\mathrm{Var}(X_i) = np(1-p)$$

- Example: *Hats.* Suppose n people throw their hats in a box and then each picks one hat at random. Let N be the number of people that get back their own hat. Find $\mathrm{E}(N)$ and $\mathrm{Var}(N)$

  Solution: Define the r.v. $X_i = 1$ if a person selects her own hat, and $X_i = 0$, otherwise. Thus $N = \sum_{i=1}^{n} X_i$.

  To find the mean and variance of N, we first find the means, variances and covariances of the X_is

  Since $X_i \sim \mathrm{Bern}(1/n)$ we have $\mathrm{E}(X_i) = 1/n$ and $\mathrm{Var}(X_i) = (1/n)(1 - 1/n)$

  To find the covariance of X_i and X_j, $i \neq j$, note that

$$p_{X_i,X_j}(1,1) = \frac{1}{n(n-1)}$$

Thus

$$\begin{aligned}\mathrm{Cov}(X_i, X_j) &= \mathrm{E}(X_i X_j) - \mathrm{E}(X_i)\,\mathrm{E}(X_j) \\ &= \frac{1}{n(n-1)} \cdot 1 - \left(\frac{1}{n}\right)^2 = \frac{1}{n^2(n-1)}\end{aligned}$$

The mean and variance of N are given by

$$\begin{aligned}\mathrm{E}(N) &= n\,\mathrm{E}(X_1) = 1 \\ \mathrm{Var}(N) &= \sum_{i=1}^{n} \mathrm{Var}(X_i) + \sum_{i=1}^{n} \sum_{j \neq i}^{n} \mathrm{Cov}(X_i, X_j) \\ &= n\mathrm{Var}(X_1) + n(n-1)\mathrm{Cov}(X_1, X_2) \\ &= \left(1 - \frac{1}{n}\right) + n(n-1)\,\frac{1}{n^2(n-1)} = 1\end{aligned}$$

Method of Indicators

- In the last two examples we used the *method of indicators* to simplify the computation of expectation
- In general, the *indicator* of an event $A \subset \Omega$ is the r.v. defined as

$$I_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{otherwise} \end{cases}$$

Thus

$$\mathrm{E}[I_A] = 1 \cdot \mathrm{P}(A) + 0 \cdot \mathrm{P}(A^c) = \mathrm{P}(A)$$

- The method of indicators involves expressing a given r.v. Y as a sum of indicators in order to simplify the computation of its expectation (this is precisely what we did in the last two examples)
- Example: *Spaghetti.* We have a bowl with n spaghetti strands. You randomly pick two strand ends and join them. The process is continued until there are no ends left. Let X be the number of spaghetti loops formed. What is $\mathrm{E}(X)$?

Gaussian Random Vectors

- A random vector $\mathbf{X} = (X_1, \ldots, X_n)$ is a Gaussian random vector (GRV) (or $X_1, X_2, \ldots, X_n$ are jointly Gaussian r.v.s) if the joint pdf is of the form

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

  where $\boldsymbol{\mu}$ is the mean and Σ is the covariance matrix of $\mathbf{X}$, and $\left|\Sigma\right| > 0$, i.e., Σ is positive definite

- Verify that this joint pdf is the same as the case $n = 2$ from Lecture Notes 5

- Notation: $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ denotes a GRV with given mean and covariance matrix

- Since Σ is positive definite, Σ^{-1} is positive definite. Thus if $\mathbf{x} - \boldsymbol{\mu} \neq \mathbf{0}$,

$$(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}) > 0,$$

  which means that the contours of equal pdf are ellipsoids

- The GRV $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, aI)$, where I is the identity matrix and $a > 0$, is called *white*; its contours of equal joint pdf are spheres centered at the origin

Properties of GRVs

- Property 1: For a GRV, 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 $X_i \sim \mathcal{N}(\mu_i, \sigma_{ii})$

  For the white GRV $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, aI)$, the r.v.s are i.i.d. $\mathcal{N}(0, a)$

- Property 2: Linear transformation of a GRV yields a GRV, i.e., given any $m \times n$ matrix A, where $m \leq n$ and A has full rank m, then

$$\mathbf{Y} = A\mathbf{X} \sim \mathcal{N}(A\boldsymbol{\mu},\, A\Sigma A^T)$$

- Example: Let

$$\mathbf{X} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} 2 & 1 \\ 1 & 3 \end{bmatrix}\right)$$

  Find the joint pdf of

$$\mathbf{Y} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix} \mathbf{X}$$

Solution: From Property 2, we conclude that

$$\mathbf{Y} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix}1 & 1\\ 1 & 0\end{bmatrix}\begin{bmatrix}2 & 1\\ 1 & 3\end{bmatrix}\begin{bmatrix}1 & 1\\ 1 & 0\end{bmatrix}\right) = \mathcal{N}\left(\mathbf{0}, \begin{bmatrix}7 & 3\\ 3 & 2\end{bmatrix}\right)$$

Before we prove Property 2, let us show that

$$\mathrm{E}(\mathbf{Y}) = A\boldsymbol{\mu} \quad \text{and} \quad \Sigma_{\mathbf{Y}} = A\Sigma A^T$$

These results follow from linearity of expectation. First, expectation:

$$\mathrm{E}(\mathbf{Y}) = \mathrm{E}(A\mathbf{X}) = A\,\mathrm{E}(\mathbf{X}) = A\boldsymbol{\mu}$$

Next consider the covariance matrix:

$$\begin{aligned}\Sigma_{\mathbf{Y}} &= \mathrm{E}\left[(\mathbf{Y}-\mathrm{E}(\mathbf{Y}))(\mathbf{Y}-\mathrm{E}(\mathbf{Y}))^T\right]\\ &= \mathrm{E}\left[(A\mathbf{X}-A\boldsymbol{\mu})(A\mathbf{X}-A\boldsymbol{\mu})^T\right]\\ &= A\,\mathrm{E}\left[(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})^T\right]A^T = A\Sigma A^T\end{aligned}$$

Of course this is not sufficient to show that $\mathbf{Y}$ is a GRV — we must also show that the joint pdf has the right form

We do so using the *characteristic function* for a random vector

- Definition: If $\mathbf{X} \sim f_{\mathbf{X}}(\mathbf{x})$, the characteristic function of $\mathbf{X}$ is

$$\Phi_{\mathbf{X}}(\boldsymbol{\omega}) = \mathrm{E}\left(e^{i\boldsymbol{\omega}^T\mathbf{X}}\right),$$

where $\boldsymbol{\omega}$ is an n-dimensional real valued vector and $i = \sqrt{-1}$

Thus

$$\Phi_{\mathbf{X}}(\boldsymbol{\omega}) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} f_{\mathbf{X}}(\mathbf{x})e^{i\boldsymbol{\omega}^T\mathbf{x}}\,d\mathbf{x}$$

This is the inverse of the multi-dimensional Fourier transform of $f_{\mathbf{X}}(\mathbf{x})$, which implies that there is a one-to-one correspondence between $\Phi_{\mathbf{X}}(\boldsymbol{\omega})$ and $f_{\mathbf{X}}(\mathbf{x})$, which can be found by taking the Fourier transform

$$f_{\mathbf{X}}(\mathbf{x}) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\frac{1}{(2\pi)^n}\Phi_{\mathbf{X}}(\boldsymbol{\omega})e^{-i\boldsymbol{\omega}^T\mathbf{x}}\,d\boldsymbol{\omega}$$

- Example: The characteristic function for $X \sim \mathcal{N}(\mu, \sigma^2)$ is given by

$$\Phi_X(\omega) = e^{-\frac{1}{2}\omega^2\sigma^2 + i\mu\omega},$$

and for a GRV $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$,

$$\Phi_{\mathbf{X}}(\boldsymbol{\omega}) = e^{-\frac{1}{2}\boldsymbol{\omega}^T\Sigma\boldsymbol{\omega} + i\boldsymbol{\omega}^T\boldsymbol{\mu}}$$

- Now let's go back to proving Property 2

  Since A is an $m \times n$ matrix, $\mathbf{Y} = A\mathbf{X}$ and $\boldsymbol{\omega}$ are m-dimensional. Therefore the characteristic function of $\mathbf{Y}$ is

$$\begin{aligned}\Phi_{\mathbf{Y}}(\boldsymbol{\omega}) &= \mathrm{E}\left(e^{i\boldsymbol{\omega}^T\mathbf{Y}}\right)\\ &= \mathrm{E}\left(e^{i\boldsymbol{\omega}^T A\mathbf{X}}\right)\\ &= \Phi_{\mathbf{X}}(A^T\boldsymbol{\omega})\\ &= e^{-\frac{1}{2}(A^T\boldsymbol{\omega})^T\Sigma(A^T\boldsymbol{\omega}) + i\boldsymbol{\omega}^T A\boldsymbol{\mu}}\\ &= e^{-\frac{1}{2}\boldsymbol{\omega}^T(A\Sigma A^T)\boldsymbol{\omega} + i\boldsymbol{\omega}^T A\boldsymbol{\mu}}\end{aligned}$$

  Thus $\mathbf{Y} = A\mathbf{X} \sim \mathcal{N}(A\boldsymbol{\mu}, A\Sigma A^T)$

- An equivalent definition of GRV: $\mathbf{X}$ is a GRV iff for any real vector $\mathbf{a} \neq 0$, the r.v. $Y = \mathbf{a}^T\mathbf{X}$ is Gaussian (see HW for proof)

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

$$\mathbf{Y} = \begin{bmatrix} X_{i_1} \\ X_{i_2} \\ \vdots \\ X_{i_k} \end{bmatrix}$$

  is a GRV

- To show this we use Property 2. For example, let $n = 3$ and $\mathbf{Y} = \begin{bmatrix} X_1 \\ X_3 \end{bmatrix}$

  We can express $\mathbf{Y}$ as a linear transformation of $\mathbf{X}$:

$$\mathbf{Y} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} = \begin{bmatrix} X_1 \\ X_3 \end{bmatrix}$$

  Therefore

$$\mathbf{Y} \sim \mathcal{N}\left(\begin{bmatrix} \mu_3 \\ \mu_1 \end{bmatrix}, \begin{bmatrix} \sigma_{33} & \sigma_{31} \\ \sigma_{13} & \sigma_{11} \end{bmatrix}\right)$$

- The converse of Property 3 does not hold in general (as demonstrated by the example in Lecture Notes 5)

- Property 4: Conditionals of a GRV are Gaussian, more specifically, if

$$\mathbf{X} = \left[\begin{array}{c} \mathbf{X}_1 \\ -- \\ \mathbf{X}_2 \end{array}\right] \sim \mathcal{N}\left(\left[\begin{array}{c} \boldsymbol{\mu}_1 \\ -- \\ \boldsymbol{\mu}_2 \end{array}\right], \left[\begin{array}{c|c} \Sigma_{11} & \Sigma_{12} \\ -- & -- \\ \Sigma_{21} & \Sigma_{22} \end{array}\right]\right),$$

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

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

  Compare this to the case of $n = 2$ and $k = 1$:

$$X_2 \,|\, \{X_1 = x\} \sim \mathcal{N}\left(\frac{\sigma_{21}}{\sigma_{11}}(x - \mu_1) + \mu_2\,,\; \sigma_{22} - \frac{\sigma_{12}^2}{\sigma_{11}}\right)$$

- Example:

$$\left[\begin{array}{c} X_1 \\ -- \\ X_2 \\ X_3 \end{array}\right] \sim \mathcal{N}\left(\left[\begin{array}{c} 1 \\ -- \\ 2 \\ 2 \end{array}\right], \left[\begin{array}{c|cc} 1 & 2 & 1 \\ -- & -- & -- \\ 2 & 5 & 2 \\ 1 & 2 & 9 \end{array}\right]\right)$$

From Property 4, it follows that

$$\mathrm{E}(\mathbf{X}_2 \,|\, X_1 = x) = \begin{bmatrix} 2 \\ 1 \end{bmatrix}(x - 1) + \begin{bmatrix} 2 \\ 2 \end{bmatrix} = \begin{bmatrix} 2x \\ x + 1 \end{bmatrix}$$

$$\begin{aligned} \Sigma_{\{\mathbf{X}_2 | X_1 = x\}} &= \begin{bmatrix} 5 & 2 \\ 2 & 9 \end{bmatrix} - \begin{bmatrix} 2 \\ 1 \end{bmatrix}\begin{bmatrix} 2 & 1 \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 \\ 0 & 8 \end{bmatrix} \end{aligned}$$

- The proof of Property 4 follows from properties 1 and 2 and the orthogonality principle (HW exercise)
- A consequence of Property 4 is that if $\left[\mathbf{Y}^T\; X\right]^T$ is a GRV, then the best MSE estimate of X given $\mathbf{Y}$ is linear, i.e., the linear MMSE estimate is the MMSE estimate

MSE Estimation: the Vector Case

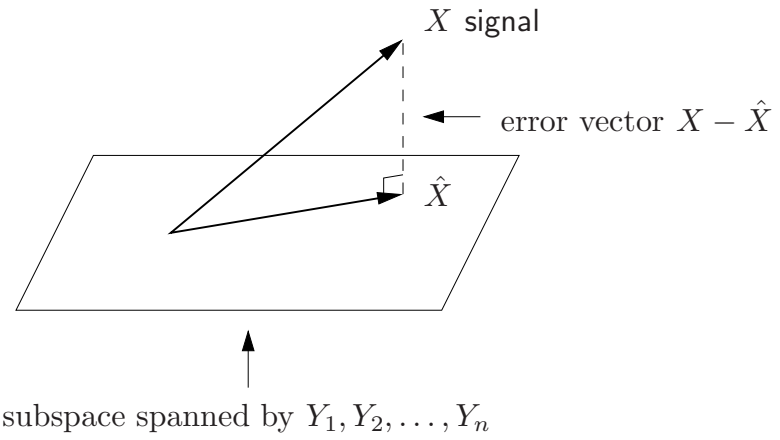
- Let $X \sim f_X(x)$ be a r.v. representing the signal and let $\mathbf{Y}$ be an n-dimensional RV representing the observations
- The minimum MSE estimate of X given $\mathbf{Y}$ is the conditional expectation $\mathrm{E}(X \,|\, \mathbf{Y})$. This is often not practical to compute either because the conditional pdf of X given $\mathbf{Y}$ is not known or because of high computational cost
- The MMSE linear (or affine) estimate is easier to find since it depends only on the means, variances, and covariances of the r.v.s involved
- To find the MMSE linear estimate, first assume that $\mathrm{E}(X) = 0$ and $\mathrm{E}(\mathbf{Y}) = \mathbf{0}$. The problem reduces to finding a real n-vector $\mathbf{h}$ such that

$$\hat{X} \;=\; \mathbf{h}^T\mathbf{Y} \;=\; \sum_{i=1}^{n} h_i Y_i$$

  minimizes the $\mathrm{MSE} = \mathrm{E}\left[(X - \hat{X})^2\right]$

MMSE Linear Estimate via Orthogonality Principle

- To find $\hat{X}$ we use the orthogonality principle: we view the r.v.s $X, Y_1, Y_2, \ldots, Y_n$ as vectors in the inner product space consisting of all zero mean r.v.s defined over the underlying probability space
- The linear estimation problem reduces to a geometry problem



- To minimize $\mathrm{MSE} = \|X - \hat{X}\|^2$, we choose $\hat{X}$ so that the error vector $X - \hat{X}$ is orthogonal to the subspace spanned by the observations $Y_1, Y_2, \ldots, Y_n$, i.e.,

$$\mathrm{E}\left[(X - \hat{X})Y_i\right] = 0\,, \quad i = 1, 2, \ldots, n\,,$$

hence

$$\mathrm{E}(Y_i X) = \mathrm{E}(Y_i \hat{X}) = \sum_{j=1}^{n} h_j \,\mathrm{E}(Y_i Y_j)\,, \quad i = 1, 2, \ldots, n$$

- Define the *cross covariance* of $\mathbf{Y}$ and X as the n-vector

$$\Sigma_{\mathbf{Y}X} = \mathrm{E}\left[(\mathbf{Y} - \mathrm{E}(\mathbf{Y}))(X - \mathrm{E}(X))\right] = \begin{bmatrix} \sigma_{Y_1 X} \\ \sigma_{Y_2 X} \\ \vdots \\ \sigma_{Y_n X} \end{bmatrix}$$

For $n = 1$ this is simply the covariance

- The above equations can be written in vector form as $\Sigma_{\mathbf{Y}}\mathbf{h} = \Sigma_{\mathbf{Y}X}$
- If $\Sigma_{\mathbf{Y}}$ is nonsingular, we can solve the equations to obtain $\mathbf{h} = \Sigma_{\mathbf{Y}}^{-1}\Sigma_{\mathbf{Y}X}$

- Thus, if $\Sigma_{\mathbf{Y}}$ is nonsingular then the best linear MSE estimate is: $\hat{X} = \mathbf{h}^T\mathbf{Y} = \Sigma_{\mathbf{Y}X}^T\Sigma_{\mathbf{Y}}^{-1}\mathbf{Y}$
- Compare this to the scalar case, where $\hat{X} = \dfrac{\mathrm{Cov}(X,Y)}{\sigma_Y^2}Y$
- Now to find the minimum MSE, consider

$$\begin{aligned}
\mathrm{MSE} &= \mathrm{E}\left[(X - \hat{X})^2\right] \\
&= \mathrm{E}\left[(X - \hat{X})X\right] - \mathrm{E}\left[(X - \hat{X})\hat{X}\right] \\
&= \mathrm{E}\left[(X - \hat{X})X\right], \text{ since by orthogonality } (X - \hat{X}) \perp \hat{X} \\
&= \mathrm{E}(X^2) - \mathrm{E}(\hat{X}X) \\
&= \sigma_X^2 - \mathrm{E}\left(\Sigma_{\mathbf{Y}X}^T\Sigma_{\mathbf{Y}}^{-1}\mathbf{Y}X\right) = \sigma_X^2 - \Sigma_{\mathbf{Y}X}^T\Sigma_{\mathbf{Y}}^{-1}\Sigma_{\mathbf{Y}X}
\end{aligned}$$

- Compare this to the scalar case, where minimum MSE is $\sigma_X^2 - \dfrac{\mathrm{Cov}(X,Y)^2}{\sigma_Y^2}$
- If X or $\mathbf{Y}$ have nonzero mean, the MMSE affine estimate $\hat{X} = h_0 + \mathbf{h}^T\mathbf{Y}$ is determined by first finding the MMSE linear estimate of $X - \mathrm{E}(X)$ given $\mathbf{Y} - \mathrm{E}(\mathbf{Y})$ (minimum MSE for $\hat{X}'$ and $\hat{X}$ are the same), which is $\hat{X}' = \Sigma_{\mathbf{Y}X}^T\Sigma_{\mathbf{Y}}^{-1}(\mathbf{Y} - \mathrm{E}(\mathbf{Y}))$, and then setting $\hat{X} = \hat{X}' + \mathrm{E}(X)$ (since $\mathrm{E}(\hat{X}) = \mathrm{E}(X)$ is necessary)

Example

- Let X be the r.v. representing a signal with mean μ and variance P. The observations are $Y_i = X + Z_i$, for $i = 1, 2, \ldots, n$, where the Z_i are zero mean uncorrelated noise with variance N, and X and Z_i are also uncorrelated

  Find the MMSE linear estimate of X given $\mathbf{Y}$ and its MSE

- For $n = 1$, we already know that $\hat{X}_1 = \dfrac{P}{P+N} Y_1 + \dfrac{N}{P+N}\mu$

- To find the MMSE linear estimate for general n, first let $X' = X - \mu$ and $Y_i' = Y_i - \mu$. Thus X' and $\mathbf{Y}'$ are zero mean

- The MMSE linear estimate of X' given $\mathbf{Y}'$ is given by $\hat{X}_n' = \mathbf{h}^T \mathbf{Y}'$, where

$$\Sigma_{\mathbf{Y}} \mathbf{h} = \Sigma_{\mathbf{Y}X}, \quad \text{thus}$$

$$\begin{bmatrix} P+N & P & \cdots & P \\ P & P+N & \cdots & P \\ \vdots & \vdots & \ddots & \vdots \\ P & P & \cdots & P+N \end{bmatrix} \begin{bmatrix} h_1 \\ h_2 \\ \vdots \\ h_n \end{bmatrix} = \begin{bmatrix} P \\ P \\ \vdots \\ P \end{bmatrix}$$

- By symmetry, $h_1 = h_2 = \cdots = h_n = \dfrac{P}{nP+N}$. Thus

$$\hat{X}_n' = \frac{P}{nP+N} \sum_{i=1}^{n} Y_i'$$

  Therefore

$$\hat{X}_n = \frac{P}{nP+N} \left(\sum_{i=1}^{n} (Y_i - \mu) \right) + \mu = \frac{P}{nP+N} \left(\sum_{i=1}^{n} Y_i \right) + \frac{N}{nP+N}\mu$$

- The mean square error of the estimate:

$$\mathrm{MSE}_n = P - \mathrm{E}(\hat{X}_n' X') = \frac{PN}{nP+N}$$

  Thus as $n \to \infty$, $\mathrm{MSE}_n \to 0$, i.e., the linear estimate becomes perfect (even though we don't know the complete statistics of X and Y)