

## ECE 6010 Lecture 3 – Random Vectors

Grimmet & Stirzaker: Section 4.9

### Random Vectors

Random vectors are an extension of the bivariate random variables.

$n$  r.v.s  $X_1, X_2, \dots, X_n$  define a measurable mapping from an underlying sample space  $(\Omega, \mathcal{F})$  to  $(\mathbb{R}^n, \mathcal{B}^n)$ , where  $\mathcal{B}^n$  is the smallest  $\sigma$ -field containing all sets of the form

$$\{(x_1, x_2, \dots, x_n) : a_1 < x_1 \leq b_1, a_2 < x_2 \leq b_2, \dots, a_n < x_n \leq b_n\}.$$

**Definition 1** The **joint distribution** of  $X_1, \dots, X_n$  is

$$P_{X_1 X_2 \dots X_n}(B) = P(\{\omega \in \Omega : (X_1(\omega), X_2(\omega), \dots, X_n(\omega)) \in B\}) \text{ for all } B \in \mathcal{B}$$

This probability is denoted as  $P_{\mathbf{X}}(B)$ . □

**Definition 2** The **joint cumulative distribution function** c.d.f. is

$$P_{X_1 X_2 \dots X_n}(a_1, a_2, \dots, a_n) = P(X_1 \leq a_1, \dots, X_n \leq a_n) = F_{\mathbf{X}}(\mathbf{a}), \quad \mathbf{a} \in \mathbb{R}^n.$$

□

**Definition 3** The **joint probability mass function** (p.m.f.) is

$$p_{\mathbf{X}}(\mathbf{a}) = P(X_1 = a_1, \dots, X_n = a_n)$$

□

**Definition 4** The **joint probability density function** (p.d.f.)  $f_{\mathbf{X}}(\mathbf{a})$  is the function that satisfies

$$F_{\mathbf{X}}(\mathbf{a}) = \int_{-\infty}^{a_n} \int_{-\infty}^{a_{n-1}} \dots \int_{-\infty}^{a_1} f_{\mathbf{X}}(\mathbf{x}) dx_1 dx_2 \dots dx_n$$

for a continuous random vector. □

Fact:  $X_1, X_2, \dots, X_n$  are independent if  $F_{\mathbf{X}}$  or  $p_{\mathbf{X}}$  or  $f_{\mathbf{X}}$  factor into products of marginals.

Suppose  $g : (\mathbb{R}^n, \mathcal{B}^n) \rightarrow (\mathbb{R}^n, \mathcal{B}^n)$  is measurable. Then  $g(X_1, \dots, X_n)$  is a random variable. Law of unconscious statistician:

$$E[g(X_1, \dots, X_n)] = \begin{cases} \int \dots \int g(x_1, \dots, x_n) f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x} & \text{continuous} \\ \sum \dots \sum g(x_{i_1}, x_{i_2}, \dots, x_{i_n}) p_{\mathbf{X}}(x_{i_1}, x_{i_2}, \dots, x_{i_n}) & \text{discrete.} \end{cases}$$

### Covariance

Suppose  $\mathbf{X} : (\Omega, \mathcal{F}) \rightarrow (\mathbb{R}^n, \mathcal{B}^n)$  and  $\mathbf{Y} : (\Omega, \mathcal{F}) \rightarrow (\mathbb{R}^m, \mathcal{B}^m)$  (that is,  $X$  and  $Y$  are random vectors of dimension  $n$  and  $m$ , respectively).

**Definition 5**

$$\text{cov}(\mathbf{X}, \mathbf{Y}) = E[(\mathbf{X} - E[\mathbf{X}])(\mathbf{Y} - E[\mathbf{Y}])^T] \in \mathbb{R}^{n \times m} = \Sigma_{XY},$$

where

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

□

Note:  $\Sigma$  is frequently used as a symbol to denote covariance. It should not be confused with a summation sign, and is usually clear from context.

Property:  $\text{cov}(\mathbf{X}, \mathbf{Y}) = [\text{cov}(\mathbf{Y}, \mathbf{X})]^T$ .

If  $A$  is  $k \times n$  and  $B$  is  $l \times m$  and  $\mathbf{a} \in \mathbb{R}^k$  and  $\mathbf{b} \in \mathbb{R}^l$  then

$$\text{cov}(A\mathbf{X} + \mathbf{a}, B\mathbf{Y} + \mathbf{b}) = A\Sigma_{XY}B^T.$$

$\text{cov}(\mathbf{X}, \mathbf{X}) = \Sigma_X$  is called the “covariance of  $\mathbf{X}$ .” It is a symmetric matrix, non-negative definite (or positive semidefinite), and thus has all non-negative eigenvalues.

If  $X_1, X_2, \dots, X_n$  are mutually uncorrelated then

$$\Sigma_X = \text{diag}(\sigma_1^2, \sigma_2^2, \dots, \sigma_n^2),$$

where  $\sigma_k^2 = \text{var}(X_k)$ .

Suppose we partition  $\mathbf{X}$  of  $n$  dimensions as

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}^{(1)} \\ \mathbf{X}^{(2)} \end{bmatrix}$$

of  $k$  and  $n - k$  elements, respectively. let

$$\boldsymbol{\mu} = E[\mathbf{X}] = \begin{bmatrix} \boldsymbol{\mu}^{(1)} \\ \boldsymbol{\mu}^{(2)} \end{bmatrix}$$

where  $\boldsymbol{\mu}^{(1)} = E[\mathbf{X}^{(1)}]$  and  $\boldsymbol{\mu}^{(2)} = E[\mathbf{X}^{(2)}]$ . Similarly,

$$\Sigma_X = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

where

$$\Sigma_{11} = \text{cov}(\mathbf{X}^{(1)}, \mathbf{X}^{(1)}) \quad \Sigma_{12} = \text{cov}(\mathbf{X}^{(1)}, \mathbf{X}^{(2)}) \quad \Sigma_{22} = \text{cov}(\mathbf{X}^{(2)}, \mathbf{X}^{(2)}),$$

or, in general,

$$\Sigma_{ij} = \text{cov}(\mathbf{X}^{(i)}, \mathbf{X}^{(j)})$$

## Characteristic functions

**Definition 6** The characteristic function of an  $n$ -dimensional random vector  $\mathbf{X}$  is defined as

$$\phi_{\mathbf{X}}(\mathbf{u}) = E[\exp(i\mathbf{u}^T \mathbf{X})]$$

where  $\mathbf{u} \in \mathbb{R}^n$ . □

As before, this is just an  $n$ -dimensional Fourier transform.

**Definition 7**  $\mathbf{X}$  is a Gaussian random vector with parameters  $\boldsymbol{\mu}$  and  $\Sigma$  if

$$\phi_X(\mathbf{u}) = \exp[i\mathbf{u}^T \boldsymbol{\mu} - \frac{1}{2}\mathbf{u}^T \Sigma \mathbf{u}]$$

We write  $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ . □

Properties of Gaussian random vectors:

1.  $E[\mathbf{X}] = \boldsymbol{\mu}$ .
2.  $X_1, X_2, \dots, X_n$  independent if and only if  $\Sigma$  is a diagonal matrix.

3. If  $\mathbf{Y} = A\mathbf{X} + \mathbf{b}$ , then  $\mathbf{Y}$  is also Gaussian,  $\mathbf{Y} \sim \mathcal{N}(A\boldsymbol{\mu} + \mathbf{b}, A\Sigma A^T)$ .

**Linear functions of Gaussians are Gaussians.**

Said another way: **Family of Gaussians closed under affine transformations.**

Suppose  $\Sigma$  is *positive definite*. Then it can be factored as

$$\Sigma = CC^T$$

where  $C$  is an  $n \times n$  invertible, lower-triangular matrix. This factorization is called the **Cholesky factorization**. This is essentially a “matrix square root.”

Suppose  $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ , with  $\Sigma$  p.d. Let  $\mathbf{Y} = C^{-1}(\mathbf{X} - \boldsymbol{\mu})$ . Then  $\mathbf{Y}$  is normal with  $\boldsymbol{\mu} = \mathbf{0}$  and  $\Sigma = I$ .

This process of diagonalizing the covariance matrix is called **whitening**. We say that uncorrelated i.i.d. components are *white*.

4. If  $\Sigma > 0$  (i.e., p.d.) then  $X$  is a continuous r.v. with

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} \exp \left[ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right].$$

where  $|\Sigma| = \det(\Sigma) =$  product of eigenvalues.

5. **Important:** Suppose  $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$  with  $\Sigma > 0$ . Partition  $\mathbf{X}$ ,

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}^{(1)} \\ \mathbf{X}^{(2)} \end{bmatrix}$$

where  $\mathbf{X}^{(1)}$  has  $k$  elements. It turns out that  $\mathbf{X}^{(1)}$  is also Gaussian. (How could we easily show this?) Let us partition

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}^{(1)} \\ \boldsymbol{\mu}^{(2)} \end{bmatrix} \quad \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

Then

$$\mathbf{X}^{(1)} \sim \mathcal{N}(\boldsymbol{\mu}^{(1)}, \Sigma_{11}) \quad \mathbf{X}^{(2)} \sim \mathcal{N}(\boldsymbol{\mu}^{(2)}, \Sigma_{22})$$

Consider  $\mathbf{X}^{(2)}$  conditioned on  $(\mathbf{X}^{(1)} = \mathbf{x}^{(1)})$ :

$$f_{\mathbf{X}^{(2)}|\mathbf{X}^{(1)}}(\mathbf{x}^{(2)}|\mathbf{x}^{(1)}) = \frac{f_{\mathbf{X}}(\mathbf{x})}{f_{\mathbf{X}^{(1)}}(\mathbf{x}^{(1)})}$$

Then it can be shown that

$$\mathbf{X}^{(2)} | (\mathbf{X}^{(1)} = \mathbf{x}^{(1)}) \sim \mathcal{N}(\boldsymbol{\mu}', \Sigma')$$

where

$$\begin{aligned} \boldsymbol{\mu}' &= \boldsymbol{\mu}^{(2)} + \Sigma_{21} \Sigma_{11}^{-1} (\mathbf{x}^{(1)} - \boldsymbol{\mu}^{(1)}) \\ \Sigma' &= \Sigma_{22} - \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12} \end{aligned}$$

This is “smaller” than  $\Sigma_{22}$ .

Discuss implications. Draw pictures.

Note: For a Gaussian vector, the conditional density is Gaussian.

## An Application: MMSE Prediction

Suppose we have a random sequence  $X_1, X_2, \dots, X_n$ , and we observe the first  $n - 1$  of them:

$$X_1 = x_1, X_2 = x_2, \dots, X_{n-1} = x_{n-1}.$$

Given this data, we want to *predict* the value of  $X_n$ .

Our estimate of  $x_n$  will be denoted as  $\hat{x}_n$ . Clearly, it could be a function of all the observed data:

$$\hat{x}_n = h(x_1, x_2, \dots, x_{n-1})$$

for some function  $h : \mathbb{R}^{n-1} \rightarrow \mathbb{R}$ . One thing we could try is to minimize the average of  $(x_n - h(x_1, \dots, x_{n-1}))^2$ . That is we would like to solve

$$\min_h E[(X_n - h(x_1, x_2, \dots, x_{n-1}))^2]$$

It is easy to see (HW!) that the best such  $h$  is

$$h(x_1, x_2, \dots, x_{n-1}) = E[X_n | X_1 = x_1, X_2 = x_2, \dots, X_{n-1} = x_{n-1}].$$

That is, the best estimator (in a **minimum mean-squared error** sense) is the conditional expectation!

Now let us take a specific distribution. Suppose  $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ , and partition according to

$$\begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_{n-1} \\ X_n \end{bmatrix}$$

Given  $X_1 = x_1, \dots, X_{n-1} = x_{n-1}$ , the variable  $X_n$  is  $\mathcal{N}(\mu', \sigma^2)$  where

$$\mu' = \mu_n + \Sigma_{n,n-1} \Sigma_{n-1}^{-1} \begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \\ \vdots \\ x_{n-1} - \mu_{n-1} \end{bmatrix}$$

$$\sigma^2 = \sigma_n^2 - \Sigma_{n,n-1} \Sigma_{n-1}^{-1} \Sigma_{n,n-1}^T$$

where

$$\Sigma = \begin{bmatrix} \Sigma_{n-1} & \Sigma_{n-1,n} \\ \Sigma_{n,n-1} & \sigma_n^2 \end{bmatrix}$$

and

$$\Sigma_{n-1} = \text{cov}([X_1, \dots, X_{n-1}], [ ])$$

$$\Sigma_{n,n-1} = (\text{cov}(X_n, X_1), \text{cov}(X_n, X_2), \dots, \text{cov}(X_n, X_{n-1})).$$

So  $\mu'$  is the conditional mean that we want and  $\sigma^2$  is the variance of the conditional distribution,

$$\sigma^2 = \text{var}(X_n | X_1, \dots, X_{n-1}) = E[(X_n - \mu')^2 | X_1 = x_1, \dots, X_{n-1} = x_{n-1}].$$

This is the minimum mean-squared error (MMSE).

Notationally, write

$$\mathbf{a}^T = \Sigma_{n,n-1} \Sigma_{n-1}^{-1}.$$

Then

$$\hat{x}_n = \mu_n + \mathbf{a}^T \begin{bmatrix} x_1 - \mu_1 \\ \vdots \\ x_{n-1} - \mu_{n-1} \end{bmatrix}$$

This is just a digital filter!

We can also show that  $\sigma^2 \leq \sigma_n^2$ , so that incorporating information from measurements decreases our uncertainty.

Note: For a Gaussian r.v., the MMSE estimator is **linear**.

## Estimation in a Markov model

Suppose that  $P(X_n | X_{n-1}, X_{n-2}, \dots, X_2, X_1) = P(X_n | X_{n-1})$ . That is, given  $X_{n-1}$ ,  $X_n$  is independent of  $X_1, X_2, \dots, X_{n-2}$ . This is actually quite common: it doesn't matter how you got to where you came from, only where you came from. Such a model is called a **Markov model**.

Under the assumption of a Markov model,

$$E[X_n | X_1, \dots, X_{n-1}] = E[X_n | X_{n-1}]$$

and

$$\hat{x}_n = \mu_n + \frac{\text{cov}(X_n, X_{n-1})}{\text{var}(X_{n-1})}(x_{n-1} - \mu_{n-1}).$$

This can be written as

$$\hat{x}_n = \mu_n + \rho(X_n, X_{n-1})\sqrt{\text{var}(X_n)} \left[ (x_{n-1} - \mu_{n-1}) / \sqrt{\text{var}(X_{n-1})} \right].$$

and

$$\sigma^2 = \text{var}(X_n)(1 - \rho^2(X_n, X_{n-1})).$$