

## Lecture 5: Independence and statistical applications

Lecturer: Yiqiao Zhong

Last update: 13:36:44, 2024-09-23

*Disclaimer: materials in this pdf are used for STAT 709 “Mathematical Statistics” at the University of Wisconsin–Madison, but they are not guaranteed to be accurate. Consult textbooks whenever possible.*

Our goal is to introduce key notions in probability theory—independence and conditional independence, and various statistical applications.

## 1 Independence

In the simplest form of independence, we have two discrete random variables  $X, Y$  satisfying  $P(X = x, Y = y) = P(X = x)P(Y = y)$ . We will generalize such a discrete case.

We will assume that random variables are defined on the same probability space  $(\Omega, \mathcal{F}, P)$ .

**Definition 1.** A sequence of random variables  $X_1, \dots, X_n$  are (jointly) independent if and only if for any Borel sets  $B_1, \dots, B_n \in \mathcal{B}$ ,

$$P(X_1 \in B_1, X_2 \in B_2, \dots, X_n \in B_n) = P(X_1 \in B_1) \cdot P(X_2 \in B_2) \cdot \dots \cdot P(X_n \in B_n) \quad (1)$$

In an equivalent definition of independence: we only need the sets  $B_1, \dots, B_n$  to be of the form  $(-\infty, a]$ .

**Theorem 1.** Random variables  $X_1, \dots, X_n$  are jointly independent if and only if

$$P(X_1 \leq a_1, X_2 \leq a_2, \dots, X_n \leq a_n) = P(X_1 \leq a_1) \cdot P(X_2 \leq a_2) \cdot \dots \cdot P(X_n \leq a_n) \quad (2)$$

holds for all  $x_1, \dots, x_n \in \mathbb{R}$ .

The proof of this theorem is based on a careful study of the properties of Borel sets ( $\pi$ - $\lambda$  system), which we omit in this course. Note that, using the notation of CDF, (2) is equivalent to

$$F(a_1, \dots, a_n) = F(a_1) \dots F(a_n).$$

There is another equivalent definition: the joint measure is equal to the product measure of the individual random variables.

$$\mu_{X_1, \dots, X_n} = \mu_{X_1} \times \dots \times \mu_{X_n}. \quad (3)$$

The Lebesgue measure or counting measure in  $\mathbb{R}^n$  satisfies  $\nu_n = \nu_1 \times \dots \times \nu_1$ , so if the joint pdf or pmf (denoted by  $p_{X_1, \dots, X_n}$ ) of  $X_1, \dots, X_n$  exists, then (3) is equivalent to

$$p_{X_1, \dots, X_n}(x_1, \dots, x_n) = p_{X_1}(x_1) \dots p_{X_n}(x_n)$$

Still, another equivalent definition is stated in terms of functions: if  $f_1, \dots, f_n : \mathbb{R} \rightarrow \mathbb{R}$  are bounded Borel functions, then

$$\mathbb{E}[f_1(X_1) \dots f_n(X_n)] = \mathbb{E}[f_1(X_1)] \dots \mathbb{E}[f_n(X_n)]. \quad (4)$$

If the above identity holds, then choosing  $f_k(x) = \mathbf{1}\{x \in B_k\}$  reduces to (1); conversely, we can approximate a bounded Borel function and apply DCT to show (4) ([exercise](#)).

**Sum of independent random variables.** A basic question in probability and statistics is to understand the properties of sums of independent random variables. Consider two independent random variables  $X$  and  $Y$ , with respective CDF being  $F_X$  and  $F_Y$ , and induced measure being  $\mu_X$  and  $\mu_Y$ .

**Proposition 1.** *We have*<sup>1</sup>

$$P(X + Y \leq z) = \int_{\mathbb{R}} F(z - y) \mu(dy)$$

*Proof.*

$$\begin{aligned} P(X + Y \leq z) &= \mathbb{E}[\mathbf{1}\{X + Y \leq z\}] = \int \mathbf{1}\{x + y \leq z\} \mu_{X,Y}(dx, dy) \stackrel{(i)}{=} \int \mathbf{1}\{x + y \leq z\} \mu_X(dx) \mu_Y(dy) \\ &\stackrel{(ii)}{=} \int_{\mathbb{R}} \left( \int_{\mathbb{R}} \mathbf{1}\{x + y \leq z\} \mu(dx) \right) \mu(dy) \\ &\stackrel{(iii)}{=} \int_{\mathbb{R}} F(z - y) \mu(dy) \end{aligned}$$

where in (i) we used independence  $\mu_{X,Y} = \mu_X \times \mu_Y$ , in (ii) we used Fubini's theorem, and in (iii) we used the definition of CDF.  $\square$

This is a simple but illustrative result. If you have not learned measure theory or MATH/STAT 733 (or equivalent) before, please check each step carefully and understand how results from previous lectures are used.

With some calculation, we can apply Proposition 1 in the following examples (see [2][Sect. 2.1.3]).

**Example 1.** Suppose that  $X \sim N(\mu_1, \sigma_1^2)$  and  $Y \sim N(\mu_2, \sigma_2^2)$  are independent, then  $X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$ .

**Example 2.** If  $X_1, \dots, X_n \sim N(0, 1)$  are independent, then  $X_1^2 + \dots + X_n^2$  are given by the chi-square distribution  $\xi_n^2$  (or Gamma distribution  $\Gamma(n/2, 2)$ ) with pdf

$$f(x) = \frac{1}{2^{n/2} \Gamma(n/2)} x^{n/2-1} e^{-x/2}.$$

**Characteristic function.** In general, it is difficult to calculate the distribution of sums of independent random variables.

For a random variable  $X$ , then the characteristic function (ch.f.) is defined by

$$\varphi_X(t) = \mathbb{E}[e^{itX}] = \mathbb{E}[\cos(tX) + i \sin(tX)].$$

One important property about ch.f. is that for independent random variables  $X, Y$ ,  $\varphi_{X+Y}(t) = \varphi_X(t) \varphi_Y(t)$  (due to Eqn. 4). Also, the distribution of a random variable is uniquely determined by the ch.f. For normal variable  $X \sim N(0, \sigma^2)$ , the ch.f. is (Example 3.3.5 of [2])

$$\varphi_X(t) = e^{-\sigma^2 t^2 / 2},$$

and we can use this to derive Example 1.

---

<sup>1</sup>Here our notation follows the convention: we use  $\mu(dy)$  to emphasize the integral is over variable  $y$  instead of  $d\mu$  used in previous lectures. Both notations are legitimate, although  $\mu(dy)$  is preferred here to avoid confusion.

## 2 Some examples

**Uniform distribution.** Let  $G \sim N(0, I_d)$  be the standard normal random vector. The random variable

$$U = \frac{G}{\|G\|_2}$$

is on the unit sphere  $S^{d-1} = \{x \in \mathbb{R}^d : \|x\|_2 = 1\}$ . For any orthogonal matrix  $R \in \mathbb{R}^d$ , we know  $RG \sim N(0, I_d)$ , so the distribution of  $U$  is also invariant to orthogonal transformations.  $U$  is the uniform measure on  $S^{d-1}$ , and it is a unique measure that is invariant under the orthogonal group.

**exercise:** show that the two random variables/vectors  $\frac{G}{\|G\|_2}$  and  $\|G\|_2$  are independent.

**Mutual information.** Suppose that we have random variables  $X, Y$ . Mutual information is defined by

$$I(X; Y) = D_{\text{KL}}(\mu_{X,Y} \| \mu_X \times \mu_Y).$$

Mutual information quantifies the amount of information of a random variable upon observation another random variable.

The basic properties of mutual information: (i)  $I(X; Y) = I(Y; X)$  (ii)  $I(X, Y) \geq 0$ , and  $I(X, Y) = 0$  iff  $X$  and  $Y$  are independent.

**Differential privacy.** Consider the scenario where a database holds a data  $D \subset \Omega^n$  containing the entries of  $n$  users (e.g., Apple or Uber or a hospital) where  $\Omega$  is a domain (e.g., a subset of  $\mathbb{R}^p$ ). For a query to the database such as “computing the average annual income of the users”, it is a good practice to add noise to the answer to protect user’s privacy. Otherwise, the information of individual users may be leaked—for example, if  $n = 3$  and an adversary knows the income of two users, then the income of the remaining user can be inferred based on the exact average.

To rigorously define differential privacy, we view a **randomized algorithm**  $\mathcal{A}$  as a random map  $\mathcal{A} : \Omega^n \rightarrow \mathbb{R}$ : which is a collection of random variables  $\mathcal{A}(D)$  indexed by each  $D$ . Practically  $\mathcal{A}$  takes the protected (deterministic) data as the input and calculates a noisy answer to a given query. Fix  $\varepsilon > 0$ . We say  $\mathcal{A}$  provides  $\varepsilon$ -differential privacy (DP) if for any data  $D' \subset \Omega^n$  that differs from  $D$  at only one element, and for any measurable set  $S \subset \Omega$ ,

$$P(\mathcal{A}(D) \in S) \leq e^\varepsilon P(\mathcal{A}(D') \in S)$$

When  $\varepsilon$  is small, the definition says that modifying the entries of any user does not change the distribution much in the sense of ratios. In other words, the output of  $\mathcal{A}$  is sufficiently noisy such that it is difficult to distinguish if a user with any entries is present in the database or not.

An important property of DP is **sequential composition**: if  $\mathcal{A}_i$  is  $\varepsilon_i$ -differential private for  $i = 1, \dots, m$  and they are independent, then  $D \rightarrow (\mathcal{A}_1(D), \dots, \mathcal{A}_m(D))$  is  $\sum_{i=1}^m \varepsilon_i$ -differential private.

The proof follows directly from the definition of independence.

## 3 Conditional expectation

In elementary probability theory and statistics, if the joint pmf or Lesbegue pdf of random variables  $X$  and  $Y$  exists, we define the conditional density as

$$p_{X|Y}(x|y) = \frac{p_{X,Y}(x,y)}{p_Y(y)}. \quad (5)$$

An interpretation is that given a realization of  $Y$  at value  $y$ , the conditional density function is given by  $p_{X|Y}(x|y)$ . In particular, if we have two sets (aka events)  $A$  and  $B$ , then the conditional probability is  $P(B|A) = P(A \cap B)/P(A)$  (i.e.,  $X = \mathbf{1}_A$  and  $Y = \mathbf{1}_B$ ).

But what if we are in the general scenario where pmf or pdf may not exist (e.g., the induced probability measure is supported on a low-dim manifold such as spheres), or the random variable is complicated? (e.g.  $\mathbb{E}[X|Y_1, Y_2, Y_3, \dots]$ , which is common in stochastic processes) Can we define  $\mathbb{E}[X|Y]$  rigorously in the general scenario?

### 3.1 A simple definition

Suppose random variables (vectors)  $X, Y \in \mathbb{R}^p$  (namely taking values in  $\mathbb{R}^p$ ) are defined in the probability space  $(\Omega, \mathcal{F}, P)$ . Then there exists a collection of measures  $(\mu(y, \cdot))_{y \in \mathbb{R}^p}$  defined on  $(\Omega, \mathcal{F}, P)$  such that for Borel sets  $A, B \subset \mathbb{R}^p$ ,

$$P(X \in A, Y \in B) = \int_B \mu(y, A) \mu_Y(dy) \quad (6)$$

These measures are known as **regular conditional probabilities**. We may interpret  $\mu(y, \cdot)$  as a “pointwise” measure denoted by  $P(X \in \cdot | Y = y)$  or  $\mu_{X|Y=y}$ . The general definition is consistent with the elementary definition. Indeed, if we further assume that joint pdf  $p_{X,Y}$  (the case of joint pmf is similar), then we can related  $\mu(y, \cdot)$  to Eqn. 5:

$$\mu(y, A) = P(X \in A | Y = y) = \int_A p_{X|Y}(x|y) dx.$$

**Example 3.** Recall the measure from lecture 4:

$$\mu(B) = \frac{1}{2}\mu_0(B) + \frac{1}{2} \int_B h(x) dx.$$

defined for every Borel set  $B$ , where  $\mu_0$  is a discrete measure and  $h \geq 0$  and integrates to 1.

Consider a simple **latent variable model** where we have a latent Bernoulli variable  $Y$ , i.e.,  $P(Y = 0) = P(Y = 1) = 1/2$ , and  $X$  is determined by conditional distributions:  $X|Y = 0 \sim \mu_0$ ,  $X|Y = 1 \sim h$ . Show that the distribution of  $X$  is given by  $\mu$  (*exercise*).

**Tower property of conditional expectation.** The above example is a special case of the tower property. Given two random variable/vectors  $X, Y$ , we have

$$\mathbb{E} X = \mathbb{E}_{y \sim \mu_Y} (\mathbb{E}[X|Y = y]) = \int_{\mathbb{R}} \mathbb{E}[X|Y = y] \mu_Y(dy)$$

This is a parallel to Eqn 6, where we replace conditional probability to conditional expectation.

### 3.2 A general definition\*

More generally, we define conditional expectation in the following way; it is proof-friendly but not quite intuitive.<sup>2</sup>

**Definition 2.** Let  $X$  be a random variable on  $(\Omega, \mathcal{F}, P)$  with  $\mathbb{E}[|X|] < \infty$ . Let  $\mathcal{A}$  be a sub- $\sigma$ -algebra of  $\mathcal{F}$ . The conditional expectation of  $X$  given  $\mathcal{A}$ , denoted by  $\mathbb{E}(X|\mathcal{A})$ , is a random variable satisfying the following two conditions:

(a)  $\mathbb{E}(X|\mathcal{A})$  is measurable from  $(\Omega, \mathcal{A})$  to  $(\mathbb{R}, \mathcal{B})$ ,

(b)  $\int_A \mathbb{E}(X|\mathcal{A}) dP = \int_A X dP$  for any  $A \in \mathcal{A}$ .

The conditional expectation  $\mathbb{E}(X|\mathcal{A})$  exists and is unique a.s., that is, if  $Z, Z'$  both satisfy (i) and (ii), then  $Z = Z'$  a.s.). We may drop a.s. for simplicity unless there is possible confusion.

<sup>2</sup>It is uncommon in many areas of statistics, but if you are interested in probability theory or stochastic processes, you are encouraged to master this or check MATH 733, 734.

If we have another random variable  $Y$  on  $(\Omega, \mathcal{F}, P)$  taking values in  $(\Lambda, \mathcal{G})$ . The  $\sigma$ -algebra generated by  $Y$  is defined by

$$\sigma(Y) = \{Y^{-1}(B) : B \in \mathcal{G}\}$$

By definition,  $\sigma(Y)$  is a sub- $\sigma$ -algebra of  $\mathcal{F}$ . Intuitively,  $\sigma(Y)$  contains all sets in  $\mathcal{F}$  that are relevant to measuring observations of  $Y$ . This is useful when  $\mathcal{F}$  is potentially very large or rather abstract.

**Definition 3.** (i) Let  $A \in \mathcal{F}$ . The conditional probability of  $A$  given  $\mathcal{A}$  is defined to be  $P(A|\mathcal{A}) = E(I_A|\mathcal{A})$ .  
(ii) Let  $Y$  be measurable from  $(\Omega, \mathcal{F}, P)$  to  $(\Lambda, \mathcal{G})$ . The conditional expectation of  $X$  given  $Y$  is defined to be  $E(X|Y) = E[X|\sigma(Y)]$ .

**Remark 1.** We may consider the conditional expectation as “mean” on a coarser “mesh”, where the mesh is determined by  $\mathcal{A}$ . Consider the extreme cases.

- (i) When  $\mathcal{A}$  is a minimal (or trivial)  $\sigma$ -algebra, namely  $\mathcal{A} = \{\emptyset, \Omega\}$ , then  $E[X|\mathcal{A}] = E[X]$  is a constant.
- (ii) When  $\mathcal{A}$  contains  $\sigma(X)$ , then  $E[X|\mathcal{A}] = X$ .

**Connection to elementary definition.** There is a useful about  $\sigma$ -algebra generated by a random variable: there exists a Borel function  $h$  such that  $E(X|Y) = h \circ Y$ . Let  $Y \in \Lambda$ . We define

$$E(X|Y = y) = h(y)$$

to be the conditional expectation of  $X$  given  $Y = y$ . Note that  $h(y)$  is a function on  $\Lambda$ , whereas  $h \circ Y = E(X|Y)$  is a function on  $\Omega$ .

This definition coincides with the elementary definition.

- (i) If  $\Lambda = \mathbb{R}^p$ , then  $E(\mathbf{1}\{X \in A\}|Y = y)$  is the same as  $\mu(y, A)$  in Section 3.1.
- (ii) if  $(X, Y)$  has a joint pdf/pmf, then  $h(y) = \int_{\mathbb{R}^p} p_{X|Y}(x|y) dx$ .

**Properties.** The following results list some properties of conditional expectations.

**Proposition 2.** Let  $X, Y, X_1, X_2, \dots$  be integrable random variables on  $(\Omega, \mathcal{F}, P)$  and  $\mathcal{A}$  be a sub- $\sigma$ -field of  $\mathcal{F}$ .

- (i) (Linearity) If  $a \in \mathbb{R}$  and  $b \in \mathbb{R}$ , then  $E(aX + bY|\mathcal{A}) = aE(X|\mathcal{A}) + bE(Y|\mathcal{A})$  a.s.
- (ii) (Fatou's lemma) If  $X_n \geq 0$  for any  $n$ , then  $E(\liminf_n X_n|\mathcal{A}) \leq \liminf_n E(X_n|\mathcal{A})$  a.s.
- (iii) (DCT) Suppose that  $|X_n| \leq Y$  for any  $n$  and  $X_n \rightarrow_{a.s.} X$ . Then  $E(X_n|\mathcal{A}) \rightarrow_{a.s.} E(X|\mathcal{A})$ .
- (iv) (MCT) If  $0 \leq X_1 \leq X_2 \leq \dots$  and  $X_n \rightarrow_{a.s.} X$ , then  $E(X_n|\mathcal{A}) \rightarrow_{a.s.} E(X|\mathcal{A})$ .
- (v) (Tower property)  $E[E(X|\mathcal{A})|\mathcal{A}_0] = E(X|\mathcal{A}_0) = E[E(X|\mathcal{A}_0)|\mathcal{A}]$  a.s., where  $\mathcal{A}_0$  is a sub- $\sigma$ -field of  $\mathcal{A}$ .

You will need some exercises to familiarize the properties of conditional expectation.

## 4 Examples

### 4.1 Mutual information (continued)

We can rewrite the mutual information using conditional distributions. For random variables/vectors  $X$  and  $Y$ , assume for simplicity that  $X, Y$  has joint pdf or pmf. The conditional distribution  $\mu_{Y|X=x}$  has definition for every  $X$ . Then

$$I(X; Y) = \mathbb{E}_{x \sim \mu_X} [D_{\text{KL}}(\mu_{Y|X=x} \parallel \mu_Y)]. \quad (7)$$

We use  $p_{X,Y}(x,y) = p_{Y|X}(y|x)p_X(x)$  and Fubini's theorem to calculate

$$\begin{aligned} I(X;Y) &= \int \int p_{X,Y}(x,y) \log \left( \frac{p_{X,Y}(x,y)}{p_X(x)p_Y(y)} \right) dx dy \\ &= \int p_X(x) \left( \int p_{Y|X}(y|x) \log \left( \frac{p_{Y|X}(y|x)}{p_Y(y)} \right) dy \right) dx \\ &= \mathbb{E}_{x \sim \mu_X} \left( \int p_{Y|X}(y|x) \log \left( \frac{p_{Y|X}(y|x)}{p_Y(y)} \right) dy \right) dx \\ &= \mathbb{E}_{x \sim \mu_X} [D_{\text{KL}}(\mu_{Y|X=x} \parallel \mu_Y)]. \end{aligned}$$

When  $X$  and  $Y$  are independent, we have  $\mu_{Y|X=x} = \mu_Y$  (i.e., conditional distribution does not depend on  $x$ , [exercise](#)). This, combined with Eqn. 7, also shows that  $I(X;Y) = 0$ .

## 4.2 Regression function

Consider random variables  $Y \in \mathbb{R}$  and  $Z \in \mathbb{R}^p$ . The variable  $Y$  is the *response variable* that we want to predict using  $Z$  called covariates or predictors. We consider estimators of the form  $g(Z)$  and the goal is to minimize the MSE

$$\text{MSE} = \mathbb{E}[(g(Z) - Y)^2]$$

First, we state a simple result which is very easy to verify.

**Lemma 1** (Lemma 1.4.1 of [1]). *Suppose that  $\mathbb{E}[Y^2] < \infty$ , and we denote  $\mu = \mathbb{E}Y$ . Then*

$$\mathbb{E}[(Y - c)^2] = \text{Var}(Y) + (c - \mu)^2.$$

*Therefore, the left-hand side is minimized at  $c = \mu$ .*

Now let us define the **regression function** using conditional expectation.

$$\mu(z) = \mathbb{E}[Y|Z = z] \tag{8}$$

(Rigorously speaking,  $\mathbb{E}[Y|Z]$  is a measurable function in  $Z$ , so we can simply identify that measurable function with  $\mu$ ).

**Theorem 2** (Theorem 1.4.1 of [1]). *Suppose that  $\mathbb{E}[Y^2] < \infty$ . Then we have*

$$\mathbb{E}[(Y - g(Z))^2] = \mathbb{E}[(Y - \mu(Z))^2] + \mathbb{E}[(g(Z) - \mu(Z))^2]. \tag{9}$$

*Thus, minimizing the left-hand side over  $g$  leads to a unique minimizer  $g = \mu$  (up to a zero-measure set)*

The first term  $\mathbb{E}[(Y - \mu(Z))^2]$  is the irreducible error for predicting  $Y$  using  $Z$  even if we have infinite samples and a perfect statistical method. For that reason,  $\mu(Z)$  is usually viewed as the target (gold standard) function. A standard practice is to formulate a statistical model

$$Y = \mu(Z) + \varepsilon$$

where  $\varepsilon := Y - \mu(Z)$  is the “error” or the “noise” term. By definition,  $\mu(Z)$  is uncorrelated with  $\varepsilon$ . The formal statement is the following.

**Proposition 3** (Proposition 1.4.1 of [1]). *Suppose that  $\mathbb{E}[Y^2] < \infty$ .*

- (i)  $\mu(Z)$  and  $\varepsilon$  are uncorrelated.
- (ii)  $\text{Var}(Y) = \text{Var}(\mu(Z)) + \text{Var}(\varepsilon)$ .

Without further modeling, we can estimate the function  $\mu$  using **nonparametric** statistical methods such as kernel smoothing. A drawback is that the space of functions in a nonparametric model is usually large. A simple and convenient way is to use a **parametric** model such as a linear model where we assume  $\mu(Z) = \beta^\top f(Z) + \beta_0$  for a pre-determined function  $f$  (the simplest is identity map).

### 4.3 Conditional multivariate normal

What happens if  $Y$  and  $Z$  are jointly Gaussian? Since the regression function is simply  $\mathbb{E}[Y|Z]$  in Eqn. 8, we will study the conditional multivariate distribution of  $Y|Z$  assuming jointly Gaussian in this example.

Consider a fresh setting: let  $X \sim N(\mu, \Sigma)$  where  $\Sigma \in \mathbb{R}^{d \times d}$  is invertible, and denote  $X = (X^{(1)}, X^{(2)})$  where  $X^{(1)} \in \mathbb{R}^m$  and  $X^{(2)} \in \mathbb{R}^{d-m}$ . Accordingly, we express  $\mu$  and  $\Sigma$  in the block form

$$\mu = \begin{pmatrix} \mu^{(1)} \\ \mu^{(2)} \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma^{(11)} & \Sigma^{(12)} \\ \Sigma^{(21)} & \Sigma^{(22)} \end{pmatrix}.$$

We also denote  $\Theta = \Sigma^{-1}$  which is known as the **precision matrix**, and a well-known result about block matrix inversion gives the expression for each block of  $\Theta$ :

$$\Theta = \begin{pmatrix} \Theta^{(11)} & \Theta^{(12)} \\ \Theta^{(21)} & \Theta^{(22)} \end{pmatrix} = \begin{pmatrix} (\Sigma/\Sigma^{(22)})^{-1} & -(\Sigma/\Sigma^{(22)})^{-1}\Sigma^{(12)}(\Sigma^{(22)})^{-1} \\ -(\Sigma^{(22)})^{-1}\Sigma^{(21)}(\Sigma/\Sigma^{(22)})^{-1} & (\Sigma^{(22)})^{-1} + (\Sigma^{(22)})^{-1}\Sigma^{(21)}(\Sigma/\Sigma^{(22)})^{-1}\Sigma^{(12)}(\Sigma^{(22)})^{-1} \end{pmatrix}$$

where  $\Sigma/\Sigma^{(22)}$  denotes the **Schur complement** (link):

$$\Sigma/\Sigma^{(22)} = \Sigma^{(11)} - \Sigma^{(12)}(\Sigma^{(22)})^{-1}\Sigma^{(21)}$$

Let the condition probability density function of  $X^{(1)}|X^{(2)}$  be  $p_{1|2}(x_1|x_2)$ . From some calculations of conditional density based on Eqn. 5, we can show that  $\log[p_{1|2}(x_1|x_2)]$  is quadratic in  $x_1$ , and more specifically,  $p_{1|2}(x_1|x_2)$  is a normal density with conditional mean and variance ([exercise](#)):

$$\bar{\mu} = \mu^{(1)} + \Sigma^{(12)}(\Sigma^{(22)})^{-1}(x_2 - \mu^{(2)}) \quad \bar{\Sigma} = \Sigma^{(11)} - \Sigma^{(12)}(\Sigma^{(22)})^{-1}\Sigma^{(21)}.$$

The first equation shows a useful fact: the regression function is **linear** under the joint Gaussian assumption. In other words, under the joint Gaussian assumption, the best prediction for  $X^{(1)}$  using  $X^{(2)}$  is given by a linear function. In the special case where  $\mu = 0$ , this above result says that the (multivariate) least square regression of  $X^{(1)}$  against  $X^{(2)}$  is given by

$$\arg \min_B \mathbb{E} [\|X^{(1)} - BX^{(2)}\|_2^2] = \Sigma^{(12)}(\Sigma^{(22)})^{-1} \quad (10)$$

**Remark 2.** In linear regression, a basic fact is that minimizing the quadratic loss  $\sum_{i=1}^n \|y_i - z_i^\top \theta\|_2^2$  over  $\theta$  gives the least square estimate (to be discussed more in Lecture 7)

$$\hat{\theta} = (n^{-1} \sum_{i=1}^n z_i z_i^\top)^{-1} n^{-1} \sum_{i=1}^n z_i y_i$$

if the matrix in the parenthesis is invertible. Expressing  $\hat{\theta}$  with empirical covariance matrices,

$$\hat{\theta} = (\hat{\Sigma}_{zz})^{-1} \hat{\Sigma}_{zy} \quad (11)$$

which is similar to Eqn. 10. If we view  $(y_i, z_i)$  as a concatenated vector following jointly normal distribution, then we may interpret Eqn. 10 as the **population regression coefficients** and Eqn. 11 as the **sample regression coefficients**.

**Gaussian graphical model.** In some statistical applications, we are interested in constructing a graph/network based on observed data. As a useful modeling assumption, we suppose data are drawn from a Gaussian distribution, i.e.,  $X = (X_1, \dots, X_d) \sim N(0, \Sigma)$  where each coordinate of  $X$  represents a vertex. A naive way to construct a graph is assigning an edge to vertex pair  $i, j$  when (estimated)  $\Sigma_{ij}$  is sufficiently large.

It turns out that we have a meaningful way of constructing the graph based on  $\Theta = \Sigma^{-1}$  instead. As a homework problem, you can show that  $\Theta_{ij} = 0$  iff  $X_i$  and  $X_j$  are independent conditioning on other coordinates of  $X$ . Therefore, the absence of an edge indicates **conditional independence** between two vertices, so we assign an edge to pair  $i, j$  when (estimated)  $\Theta_{ij}$  is sufficiently large.

## 5 Further reading

Most of the materials are adapted from Chap 1.4 of [3] and Chap 2.1 of [2]. For regression functions and conditional normal distribution, a reference is [1].

**Haar measure.** Haar measures are the general definition of measures invariant under group actions; see the link for example.

**Conditional independence.**

**Definition 4.** We say that given  $Z$ , random variables  $X$  and  $Y$  are conditionally independent if and only if the following holds

$$P(A|X, Z) = P(A|Z) \quad a.s. \text{ for any } A \in \sigma(Y).$$

In the case where  $X, Y, Z$  has a joint pdf/pmf, then conditional independence is equivalent to

$$p_{X|Z}(x|z)p_{Y|Z}(y|z) = p_{X,Y|Z}(x, y|z)$$

One important use of conditional independence is the Markov property:  $X \rightarrow Z \rightarrow Y$  forms a Markov chain iff given  $Z$ , the random variables  $X$  and  $Y$  are conditionally independent.

## References

- [1] Peter J Bickel and Kjell A Doksum. *Mathematical statistics: basic ideas and selected topics, volumes I-II package*. Chapman and Hall/CRC, 2015.
- [2] Rick Durrett. *Probability: theory and examples*, volume 49. Cambridge university press, 5th edition, 2019.
- [3] Jun Shao. *Mathematical Statistics*. Springer Science & Business Media, 2nd edition, 2013.