Multivariate Probability and Related Properties

Wei-Chang Lee, Chi-Ning Chou

December 7, 2015

Contents

Chapter 1

Statistical Inference I **Prof. Chin-Tsang Chiang**

Lecture Notes 17

December 7, 2015 **Scribe: Wei-Chang Lee, Chi-Ning Chou**

For now, we have learned how to use a single random variable to construct a model. And we have introduced several well-studied distributions and analyze their good properties. But, is that enough? Can we use only one random variable to model everything? The answer is clearly negative. But now we need to ask what are the differences between single random variable and more than one random variables? What can we benefits from that and what are the main assumptions and structures?

We start from defining random vector.

Definition 1 (random vector) *A p-dimensional random vector is a function* $X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{pmatrix}$ *from* Ω *to* $\mathbb{R}^p$ *such that each component is a random variables.*

Definition 2 (joint distribution) *We say* F_X *is a joint distribution of* X *such that* $F_X(\underset{\sim}{x}) = P[X_1 \leq x_1, ..., X_p \leq x_p]$.

Now, we might wonder, what if some components are discrete and some are continuous? How to define the joint distribution?

Example: X_1 is discrete and X_2 is continuous.

- Only X_1: $f_{X_1}(x_1) = P(X_1 = x_1)$
- Only X_2: $f_{X_2}(x_2) = \frac{\partial}{\partial x_2} F_{X_2}(x_2)$
- Both X_1 and X_2: $\lim_{\Delta_2 \to 0} \frac{P[X_1 = x_1, X_2 \in [x_2 - \frac{\Delta_2}{2}, x_2 + \frac{\Delta_2}{2}]]}{\Delta_2}$

Finally, we give an example to show the benefit of using random vector instead of using single random variable.

Example: Consider the regression problem as follow. We have a dependent variable y and some explanatory variables $X_1, ..., X_p$. Now we want to know the distribution of Y conditioned on $X_1, ..., X_p$ or the distribution of $X_1, ..., X_p$ conditioned on Y. To characterize these, we have to look into the correlation between Y and $X_1, ..., X_p$ which can be simply modeled by joint distribution.

To sum up, random vector and joint distribution provide us a structure that can describe the correlation among different sources.

Chapter 2

Statistical Inference I **Prof. Chin-Tsang Chiang**

Lecture Notes 18

December 7, 2015 **Scribe: Wei-Chang Lee, Chi-Ning Chou**

After defining random vector, how do we characterize its distribution? As a warm up, let's think about what kind of distribution about random vector we would like to know? First of all we can regard the whole random vector as a single random variable. That is, it itself has a distribution, which is called *joint probability function* formally. On the other hands, we might want to consider the distribution of certain part of the random vector, and thus the marginal distribution is introduced. Finally, we would like to know how one part of the random vector being affected by another. Hence, the *conditional distribution* shows up. In the following lectures, we are going to define the above three important aspects of random vector formally. Now, let's start with an example.
Example: Let $F(x, y)$ be the cdf of a bivariate random vector $(x, y)^T$. Then, $\forall x_l < x_r, y_l < y_r$, we have

$$\begin{aligned} P[x_l \le x \le x_r,\ y_l \le y \le y_r] &= P[x \le x_r,\ y_l \le y \le y_r] - P[x \le x_l,\ y_l \le y \le y_r] \\ &= P[x \le x_r,\ y \le y_r] - P[x \le x_r,\ y \le y_l] \\ &\quad - P[x \le x_l,\ y \le y_r] + P[x \le x_l,\ y \le y_l] \end{aligned}$$

With this example, we can have a feeling about how to quantify the probability of a certain interval type event. Now, we can define the joint probability function.

Definition 3 (joint probability mass function) *The pmf of a discrete random vector* $\underset{\sim}{X}$ *is a function* $f_{\underset{\sim}{X}}(\underset{\sim}{x})$ *from* $\mathbb{R}^p$ *to [0,1] is defined by* $f_{\underset{\sim}{X}}(\underset{\sim}{X} = \underset{\sim}{x})$

Definition 4 (joint probability density function) *The pdf of a continuous random vector* $\underset{\sim}{X}$ *is a function* $f_{\underset{\sim}{X}}(\underset{\sim}{x})$ *that satisfies*

$$F_{\underset{\sim}{X}}(\underset{\sim}{x}) = \int \cdots \int_{\{u_1 \le x_1, \ldots, u_p \le x_p\}} f_{\underset{\sim}{X}}(\underset{\sim}{u}) d\underset{\sim}{u}$$

We can define the marginal distribution as

Definition 5 (marginal distribution) *The marginal distribution of* $\underset{\sim}{X}_1$*, which is a part of random vector* $\underset{\sim}{X}$*, is a function* $F_{\underset{\sim}{X}_1}(\underset{\sim}{x}_1) = P[X_{11} \leq x_{11},\ ..., X_{1p_1} \leq x_{1p_1}]$

Chapter 3

Statistical Inference I **Prof. Chin-Tsang Chiang**

Lecture Notes 19

December 7, 2015 **Scribe: Wei-Chang Lee, Chi-Ning Chou**

3.1 Condition and independence

Here, we formally define the concept of conditional distribution in random vector.

Definition 6 (conditional probability density function) *Let $\underset{\sim}{X}_1, \underset{\sim}{X}_2$ be two random vectors with $f_{\underset{\sim}{X}_1}(\underset{\sim}{x}_1) > 0$. Then, the conditional pdf of $\underset{\sim}{X}_2$ given $\underset{\sim}{X}_1 = \underset{\sim}{x}_1$ is defined as*

$$f_{\underset{\sim}{X}_2|\underset{\sim}{X}_1}(\underset{\sim}{x}_2|\underset{\sim}{x}_1) = \frac{f_{\underset{\sim}{X}}(\underset{\sim}{x})}{f_{\underset{\sim}{X}_1}(\underset{\sim}{x}_1)}$$

Intuitively, we can think of the whole sample space has a partition $\Omega_1 \times \Omega_2$ over random vector $\underset{\sim}{X}_1$ and $\underset{\sim}{X}_2$. Then conditional sense it to restrict $\underset{\sim}{X}_1$ occurs in A_1. Thus, the sample space becomes $A_1 \times \Omega_2$.

With the concept of conditional pdf, we can know further characterize the idea of independence. In probability theory, we say two **events** are independent if the probability of an events is not affected by another. Now, what we are consider is **random variables**, whose probability is defined over the whole $\sigma-$algebra. As a results, the formation is much more complicated. However, the idea is quite the same. And actually, the condition is more simple. As $\sigma-$algebra have the good closeness property, here once we define the independence relation for some specific representative events, the independence has been characterized.Formally, we have

Definition 7 (independence) *Let $\underset{\sim}{X} = (\underset{\sim}{X}_1{}^T, ..., \underset{\sim}{X}_p{}^T)^T$ be a random vector with joint pdf $F_{\underset{\sim}{X}}(\underset{\sim}{x})$ and we denote the marginal distribution of $\underset{\sim}{X}_i$ as $F_{\underset{\sim}{X}_i}(\underset{\sim}{x}_i)$. Now, we say $\underset{\sim}{X}_1, ..., \underset{\sim}{X}_p$ are mutually independent if*

$$F_{\underset{\sim}{X}}(\underset{\sim}{x}) = \prod_{i=1}^{p} F_{\underset{\sim}{X}_i}(\underset{\sim}{x}_i)$$

The concept of independence is very important in probability theory and statistics. With help of independence, we can derive lots of great properties for the models. Take a look at the following example.

Example: Consider the example of hazard time and missing data.

- Failure time: $T \sim f_T(t)$, where the cdf is $F_T(t)$ and the hazard function $S_T(t) = 1 - F_T(t)$.
- Censure time: $C \sim f_C(t)$, where the cdf is $F_C(t)$ and the hazard function $S_C(t) - 1 - F_C(t)$.
- $X = \min\{T, C\}$
- $\delta = \mathbf{1}_{T \neq X}$

Assume that T and C are independent. Consider

$$\begin{aligned} f_{X,\delta}(x,1) &= \lim_{\Delta \to 0} \frac{P[X \in [x, x+\Delta],\ \delta = 1]}{\Delta} = \lim_{\Delta \to 0} \frac{P[X \in [x, x+\Delta],\ T \neq C]}{\Delta} \\ &= \lim_{\Delta \to 0} \frac{P[X \in [x, x+\Delta],\ C \geq x+\Delta]}{\Delta} \\ (\because T, C \text{ are independent}) &= \lim_{\Delta \to 0} P[X \in [x, x+\Delta]] \cdot P[C \geq x+\Delta] \\ &= f_T(x) \cdot S_C(x) \end{aligned}$$

Theorem 1 (necessary and sufficient condition for mutually independence) *Let $X_1, ..., X_p$ be random variables, they are mutually independent iff $\exists g$ such that $F_{\underset{\sim}{X}}(\underset{\sim}{x}) = \prod_{i=1}^{p} g_i(x_i)$.*

Proof:
($\Rightarrow$) Take g_i to be the marginal distribution of X_i is sufficient.
($\Leftarrow$) Consider the marginal distribution $F_{X_i}(x_i) = F_{\underset{\sim}{X}}(\infty, ..., x_i, ..., \infty) = g_i(x_i) \prod_{j \neq i} g_j(\infty)$. Let $c_i = g_i(\infty)$, then $\prod_{i=1}^{p} c_i = 1$. Thus, we can further write $F_{X_i}(x_i) = \frac{g_i(x_i)}{c_i}$. Finally, by the assumption, we have

$$\begin{aligned} F_{\underset{\sim}{X}}(\underset{\sim}{x}) &= \prod_{i=1}^{p} g_i(x_i) \\ (\because \prod_{i=1}^{p} \frac{1}{c_i}) &= \prod_{i=1}^{p} \frac{g_i(x_i)}{c_i} = \prod_{i=1}^{p} F_{X_i}(x_i) \end{aligned}$$

■

Intuition (necessary and sufficient condition for mutually independence)

Mutually independence of random variables is equivalent to the partition of their marginal distribution.

Example: Latent analysis. Suppose there are three random variables X, Y, and Z. Suppose only Y and Z are mutually independence, how can we analyze the network? A possible solution is to assume there are some underlying latent effect W_1 among X and Y and another latent effect W_2 among X and Z. Now, as we consider X and Y conditioned on W_1, they will be independent. That is, $X \perp\!\!\!\perp Y | W_1$ and $X \perp\!\!\!\perp Z | W_2$.

So far, we construct the concept of independence. Now, we may want to discover the good properties implied by mutually independence. In the following context, we assume random variables $X_1, ..., X_p$ are mutually independence.

Property 1 (measurable function) *If $g_1, ..., g_p$ are measurable, then*

$$\mathbb{E}[\prod_{i=1}^{p} g_i(X_i)] = \prod_{i=1}^{p} \mathbb{E}[g_i(X_i)]$$

Property 2 ($\sigma-$algebra) *Suppose $dim(X_i) = r_i$, then $\forall A_i \in \mathcal{B}^{r_i}$, we have*

$$Pr[X_1 \in A_1, ..., X_p \in A_p] = \prod_{i=1}^{p} Pr[X_i \in A_i]$$

Proof: Take $g_i = \mathbf{1}_{\{x_i \in A_i\}}$ and apply Property 1. ■

Property 3 (induced random variables) *Suppose g_i are measurable and let $U_i = g_i(X_i)$, then $\{U_i\}$ are mutually independent.*

Proof: We can see that the induced $\sigma-$algebra is a subset of the original one. By Property 2, we know that $\{U_i\}$ are mutually independent. ■

Intuition (independence and correlation)

Independence talks about the relation of random variables over the whole σ-algebra. Thus, it is actually very difficult to achieve independence. However, correlation deals with some basic relation such as linearity among random variables.

3.2 Characteristic function

For univariate random variable, we define the characteristic function as the expectation of e^itX. However, now we are in the world of multivariate random variables, how can we define the characteristic function in a similar fashion? The idea is actually quite simple, we also take t as a vector instead of a scaler, which turns the characteristic function into a multivariate function. Then using inner product to create the exponent term. Thus, result in a similar expectation form $e^{i\underset{\sim}{t}^T\underset{\sim}{X}}$. Formally, we define the characteristic function for multivariate random variables as follow:

Definition 8 (characteristic function) *The characteristic function of multivariate random variable $\underset{\sim}{X}$ is*

$$\phi_{\underset{\sim}{X}}(\underset{\sim}{t}) = \mathbb{E}[e^{i\underset{\sim}{t}^T \underset{\sim}{X}}]$$

Similarly, we can define the moment generating function as $M_{\underset{\sim}{X}}(\underset{\sim}{t}) = \mathbb{E}[e^{\underset{\sim}{t}^T \underset{\sim}{X}}]$.

Chapter 4

Statistical Inference I **Prof. Chin-Tsang Chiang**

Lecture Notes 20

December 7, 2015 **Scribe: Wei-Chang Lee, Chi-Ning Chou**

4.1 Multivariate Transformation

Let $\underset{\sim}{X} = (X_1, X_2, ..., X_p)^T$ be a continuous random vector with pdf $f_{\underset{\sim}{X}}$ on $\mathcal{A} := \{\underset{\sim}{x} : f_{\underset{\sim}{X}(\underset{\sim}{x})\geq 0}\}$. Consider another random vector $\underset{\sim}{U} = (U_1, U_2, ..., U_p)^T$ with $U = g_i(\underset{\sim}{X})$ being 1-1 transformation from $\mathcal{A}_1$ to $\mathcal{B}$ where $\mathcal{A}_0, \mathcal{A}_1, ..., \mathcal{A}_k$ being a partition of $\mathcal{A}$ and $f_{\underset{\sim}{X}}(\mathcal{A}_0) = 0$. Thus, $\exists h_i$ such that $\underset{\sim}{x} = h_i(\underset{\sim}{u})$ on $\mathcal{A}_i$. Moreover, we will have

$$f_{\underset{\sim}{U}}(\underset{\sim}{u}) = \sum_{i=1}^{k} f_{\underset{\sim}{X}}(h_i(\underset{\sim}{x}))|J_i|, \ \forall \underset{\sim}{u} \in \mathcal{B}$$

, where

$$J_i = \begin{vmatrix} \frac{\partial h_{i1}}{\partial u_1} & \cdots & \frac{\partial h_{i1}}{\partial u_p} \\ \vdots & \ddots & \vdots \\ \frac{\partial h_{ip}}{\partial u_1} & \cdots & \frac{\partial h_{ip}}{\partial u_p} \end{vmatrix}$$