Weekly Notes for EE2012 2014/14 – Week 12

T. J. Lim

April 14, 2014

Book sections covered this week: 5.6, 5.7.1.

1 Expected Value of a Function of X and Y

1.1 Concept

We can define a new r.v. $Z = g(X, Y)$, e.g. $Z = XY$, $Z = X + Y$, $Z = \sqrt{X^2 + Y^2}$, etc., from the random vector (X, Y). The mean or expected value of $g(X, Y)$ may be interpreted as the long-term arithmetic mean of many samples of $g(X, Y)$, i.e. if (X_i, Y_i) is the i-th sample of (X, Y), then

$$E[Z] = \lim_{N\to\infty} \frac{\sum_{i=1}^{N} g(X_i, Y_i)}{N}. \tag{1}$$

A simple extension of the arguments used in the one-variable case leads us to the following *definition*:

$$E[g(X,Y)] = \left\{ \begin{array}{ll} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y) f_{X,Y}(x,y)dxdy & X, Y \text{ jointly continuous} \\ \sum_j \sum_k g(x_j, y_k) p_{X,Y}(x_j, y_k) & X, Y \text{ jointly discrete} \end{array} \right. \tag{2}$$

(See the Appendix for definitions using conditional distributions when one of X and Y is discrete and the other continuous.)

1.2 Important Special Cases

1.2.1 Sum of Two Random Variables

For any two random variables X and Y, regardless of whether they are discrete or continuous, independent or dependent, we have

$$E[X + Y] = E[X] + E[Y]. \tag{3}$$

The sophisticated proof consists of using iterated expectations; the straightforward proof consists of verifying that for all possible combinations of types of X and Y, the result holds. As an example, we consider the jointly contjnuous case below.

Example 1: Using (2), with $g(x,y) = x + y$, we have

$$\begin{aligned} E[X+Y] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x+y)f_{X,Y}(x,y)dxdy \qquad (4)\\ &= \int_{-\infty}^{\infty} x \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy \cdot dx \qquad (5)\\ &\quad + \int_{-\infty}^{\infty} y \int_{-\infty}^{\infty} f_{X,Y}(x,y)dx \cdot dy \qquad (6)\\ &= \int_{-\infty}^{\infty} x f_X(x)dx + \int_{-\infty}^{\infty} y f_Y(y)dy \qquad (7)\\ &= E[X] + E[Y]. \qquad (8)\end{aligned}$$

Thus we have just shown that the sum of two jointly continuous RVs has a mean that is the sum of the individual means. By going through the other three possible combinations of types of X and Y, we obtain the general result. ■

1.2.2 Product of Two Random Variables

If X and Y are independent, then $E[XY] = E[X]E[Y]$. However it is not true that $E[XY] = E[X]E[Y]$ implies necessarily that X and Y are independent.

Example 2: Suppose X and Y are jointly continuous and independent. Prove that $E[XY] = E[X]E[Y]$. Conversely, give a counter-example to show that $E[XY] = E[X]E[Y]$ does not imply independence between X and Y.

Ans: Let the joint PDF of X and Y be $f_X(x)f_Y(y)$. Then from (2), we have

$$\begin{aligned} E[XY] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy f_X(x)f_Y(y)dxdy\\ &= \int_{-\infty}^{\infty} x f_X(x)\int_{-\infty}^{\infty} y f_Y(y)dydx\\ &= \int_{-\infty}^{\infty} x f_X(x)E[Y]dx\\ &= E[X]E[Y].\end{aligned}$$

For a counter-example, let $\Theta \sim \mathcal{U}[0, 2\pi)$, $X = \cos\Theta$ and $Y = \sin\Theta$. Then $E[XY] = 0.5E[\sin 2\Theta] = 0$, and $E[X] = E[\cos\Theta] = 0$, therefore, $E[XY] = E[X]E[Y]$. But since (X,Y) lie on the unit circle, X and Y are not independent. This example will be re-visited in the next section. ■

1.3 Correlation and Covariance

We may be interested in a single-parameter description of the degree to which X and Y affect each other, as opposed to merely knowing whether or not they are independent. For this purpose, we may use the covariance of X and Y, defined as

$$\text{cov}(X,Y) = E[(X-\mu_X)(Y-\mu_Y)] \qquad (9)$$

where $\mu_X = E[X]$ and $\mu_Y = E[Y]$. When X and Y tend to be large together, then their covariance will be positive. Conversely, if X and Y tend to move in opposite directions, then their covariance will be negative. If a large value of X leads to small and large values of Y in equal proportion, then their covariance will be small because the positive values of $(X - \mu_X)(Y - \mu_Y)$ cancel out the negative ones. Therefore, the covariance is a suitable candidate for the one-parameter description of correlation between X and Y.

By expanding $(X - \mu_X)(Y - \mu_Y)$, and then using the linearity property of the expectation operator, we get

$$\text{cov}(X, Y) = E[XY] - E[X]E[Y]. \tag{10}$$

This second expression is often slightly easier to compute than the original.

Unfortunately, the covariance scales with X and Y, i.e. if $Z = aX$, then $\text{cov}(X, Z) = a\text{cov}(X, Y)$. But we would not want the measure of correlation between aX and Y to be a times that of X and Y, since there is no fundamental difference in the statistical relationship between aX and Y, and between X and Y. Therefore, we define a *normalized* version of covariance, called the *correlation coefficient*, that is unaffected by linear scaling of X and/or Y:

$$\rho_{X,Y} = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} \tag{11}$$

where σ_X and σ_Y are the standard deviations of X and Y, respectively.

Example 3: Show that $\rho_{X,Y}$ is invariant to scaling of X and/or Y.

Ans: Let $Z = aX$ and $W = bY$, where $a, b \neq 0$. Then

$$\begin{aligned} \rho_{Z,W} &= \frac{E[ZW] - E[Z]E[W]}{\sigma_Z \sigma_W} && (12) \\ &= \frac{abE[XY] - abE[X]E[Y]}{ab\sigma_X \sigma_Y} && (13) \\ &= \frac{E[XY] - E[X]E[Y]}{\sigma_X \sigma_Y} && (14) \\ &= \rho_{X,Y}. && (15) \end{aligned}$$

The second line follows from the linearity of expectation, and the rule that $\text{var}(aX) = a^2\text{var}(X)$. ■

When $\rho_{X,Y} = 0$, it is said that X and Y are *uncorrelated.* As we show below, uncorrelatedness must be interpreted with some care.

1.3.1 What Is The Correlation Coefficient?

It is easily shown (see the textbook) that $\rho_{X,Y}$ is bounded in magnitude:

$$|\rho_{X,Y}| \leq 1, \tag{16}$$

with equality only when $Y = aX + b$, $a \neq 0$. This hints at the nature of the correlation coefficient – it is only a measure of *linear* correlation between two random variables. It is maximized when (X, Y) points all fall on a straight line – we say that Y is an "affine" function of X, and vice versa. But what does it mean to say that X and Y uncorrelated, i.e. $\rho_{X,Y} = 0$? The following example gives us a clue.

Example 4: Let (X, Y) fall on the unit circle with a uniform distribution for $\tan^{-1}(Y/X)$, i.e.

$$X = \cos\Theta, \quad Y = \sin\Theta, \tag{17}$$

where $\Theta \sim U(0, 2\pi)$. Both X and Y have zero mean since $E[\cos\Theta] = E[\sin\Theta] = 0$. The correlation of X and Y is

$$E[XY] = E[\sin\Theta\cos\Theta] = \frac{1}{2}E[\sin 2\Theta] = 0. \tag{18}$$

Therefore, $\rho_{X,Y} = 0$ and X and Y are uncorrelated. This result is counter-intuitive to our everyday notion of correlation, because X and Y are clearly rather tightly coupled – if X is known, then Y is known to within two values! This is a prime example of X and Y being uncorrelated, but not independent. ■

The same phenomenon of X and Y being "correlated" in the sense that knowledge of one does impact on uncertainty in the other, but which are "uncorrelated" in the sense that $\rho_{X,Y} = 0$, occurs in all cases where the correlation is of a circularly symmetric nature. More generally, when all lines of best fit are equally good (or bad) at fitting the (X, Y) points then $\rho_{X,Y} = 0$. Therefore, we can conclude that

$$\text{Uncorrelatedness} \;\not\Rightarrow\; \text{Independence}. \tag{19}$$

However,

$$\text{Independence} \;\Rightarrow\; \text{Uncorrelatedness}. \tag{20}$$

This is easy to see, since if X and Y are independent, then $E[XY] = E[X]E[Y]$, and hence $\text{cov}(X, Y) = 0$. (The only case of importance where uncorrelatedness does imply independence is when X and Y are "jointly Gaussian". Since we have not spent any time on the topic of joint Gaussianity, we will not dwell on this case.)

2 Conditioning on a Random Variable

2.1 Concept

We have encountered conditional probabilities of the form $P[A|B]$ where A and B are both events, and we have also studied conditional distributions in the form of conditional CDFs $F_X(x|A)$, conditional PDFs $f_X(x|A)$ and conditional PMFs $p_X(x|A)$, where A is an event.

In this section, we will introduce the concept of "conditioning on X", where X is a r.v. There are two cases to consider: (i) When X is discrete, the conditioning event is $\{X = x\}$, where $x \in S_X$. (ii) When X is continuous, the conditioning event has to be modified to $\{x < X \leq x + h\}$, where h is small.

2.2 Discrete X

2.2.1 Discrete Y

When both X and Y are discrete, we define the conditional PMF of Y given X as

$$p_{Y|X}(y|x) = P[Y = y|X = x], \tag{21}$$

where $x \in S_X$ and $y \in S_Y$. The conditional PMF of Y given X allows one to compute the conditional probability of $Y \in B$ given $X = x$ via

$$P[Y \in B|X = x] = \sum_{y \in B} p_{Y|X}(y|x). \tag{22}$$

From (22) and the Theorem on Total Probability,

$$P[Y \in B] = \sum_{x \in S_X} P[Y \in B|X = x]p_X(x) = \sum_{x \in S_X} \sum_{y \in B} p_{Y|X}(y|x)p_X(x). \tag{23}$$

A particularly important application of this last result is when B contains just one element, in which case we obtain the marginal PMF of Y as

$$p_Y(y) = \sum_{x \in S_X} p_{Y|X}(y|x)p_X(x). \tag{24}$$

Equation (24) allows us to derive some very interesting results, as the following example on Poisson splitting shows.

Example 5: Let the customers arriving in a store over a time interval t be $N(t)$, and assume that $N(t)$ is Poisson with PMF

$$P[N(t) = k] = \frac{(\lambda t)^k e^{-\lambda t}}{k!}, \quad k = 0, 1, 2, \ldots \tag{25}$$

With probability p, a customer buys something in the store, and with probability $1 - p$, he/she doesn't. Customers make independent decisions on whether to buy something or not. Find the distribution of the number of *paying* customers entering the store in one time unit.

Ans: Let X be the number of paying customers arriving in one time unit. Conditioned on $N(1) = n$, i.e. there are n customers in total that enter in one time unit, the number who buy something follows a binomial distribution with parameters n and p, i.e.

$$P[X = k|N(1) = n] = \binom{n}{k} p^k (1 - p)^{n-k}, \quad k = 0, 1, \ldots, n. \tag{26}$$

This is because we assume that customers make independent decisions.

Using (24), we can obtain the PMF of X as

$$\begin{aligned} p_X(k) &= \sum_{n=k}^{\infty} \binom{n}{k} p^k (1-p)^{n-k} \frac{\lambda^n e^{-\lambda}}{n!} && (27) \\ &= e^{-\lambda} (\lambda p)^k \sum_{n=k}^{\infty} \frac{n!}{k!(n-k)!} \frac{[\lambda(1-p)]^{n-k}}{n!} && (28) \\ &= \frac{e^{-\lambda}(\lambda p)^k}{k!} \sum_{n=k}^{\infty} \frac{[\lambda(1-p)]^{n-k}}{(n-k)!} && (29) \\ &= \frac{e^{-\lambda}(\lambda p)^k}{k!} e^{\lambda(1-p)} && (30) \\ &= \frac{e^{-\lambda p}(\lambda p)^k}{k!}, \quad k = 0, 1, \ldots && (31) \end{aligned}$$

Therefore, the number of customers arriving in one time unit who buy something is also Poisson distributed, but with mean λp instead of λ. ■

2.2.2 Continuous Y

When Y is continuous, its conditional distribution is best described using a PDF, derived from its conditional CDF:

$$\begin{aligned} F_{Y|X}(y|x) &= P[Y \leq y | X = x] && (32) \\ f_{Y|X}(y|x) &= \frac{d}{dy} F_{Y|X}(y|x). && (33) \end{aligned}$$

From this conditional PDF, we can find conditional probabilities

$$P[Y \in B | X = x] = \int_B f_{Y|X}(y|x) dy. \qquad (34)$$

The theorem on total probability yields

$$P[Y \in B] = \sum_{x \in S_X} P[Y \in B | X = x] p_X(x) = \sum_{x \in S_X} p_X(x) \int_B f_{Y|X}(y|x) dy. \qquad (35)$$

In particular, if $B = (y, y + dy]$ where $dy \to 0$, then

$$\begin{aligned} f_Y(y) dy &= \sum_{x \in S_X} p_X(x) f_{Y|X}(y|x) dy && (36) \\ \Rightarrow \quad f_Y(y) &= \sum_{x \in S_X} f_{Y|X}(y|x) p_X(x). && (37) \end{aligned}$$

Example 6: Let $Y = X + N$, where $p_X(-1) = p_X(1) = 0.5$, and $N \sim \mathcal{N}(0, \sigma_n^2)$. X and N are independent random variables. This is a simple model for a binary transmission system with additive Gaussian noise. Let's find the conditional PDF of Y given $X = 1$.

First, we find the conditional CDF:

$$\begin{aligned}
F_{Y|X}(y|1) &= P[Y \leq y|X = 1] && (38)\\
&= P[X + N \leq y|X = 1] && (39)\\
&= P[1 + N \leq y|X = 1] && (40)\\
&= P[N \leq y - 1|X = 1] && (41)\\
&= P[N \leq y - 1] && (42)\\
&= 1 - Q\left(\frac{y-1}{\sigma_n}\right). && (43)
\end{aligned}$$

The fifth line follows due to independence between X and N. Next, we differentiate the conditional CDF to obtain the conditional PDF:

$$f_{Y|X}(y|x) = -\frac{d}{dy}Q\left(\frac{y-1}{\sigma_n}\right) = \frac{1}{\sqrt{2\pi\sigma_n^2}}e^{-(y-1)^2/2\sigma_n^2}, \tag{44}$$

where we made use of the identity

$$\frac{d}{dx}Q(x) = -\frac{1}{\sqrt{2\pi}}e^{-x^2/2}, \tag{45}$$

and the chain rule of differentiation. The result tells us that, conditioned on $\{X = 1\}$, Y is $\mathcal{N}(1, \sigma_n^2)$. Similarly, conditioned on $\{X = -1\}$, Y is $\mathcal{N}(-1, \sigma_n^2)$.

Using (37), we have the marginal PDF of Y as

$$f_Y(y) = \frac{1}{2\sqrt{2\pi\sigma_n^2}}\left[e^{-(y-1)^2/2\sigma_n^2} + e^{-(y+1)^2/2\sigma_n^2}\right], \tag{46}$$

which is the sum of two identical Gaussian PDFs, centered over $y = -1$ and $y = +1$ respectively. ■

2.3 Continuous X

With a continuous X, the conditioning event is $\{x < X \leq x + h\}$, $h \to 0$. This is to work around the problem of conditioning on the zero probability event $\{X = x\}$, which is not a well-defined operation. We can imagine conditioning on reading an instrument with limited resolution, e.g. a weighing scale that only gives weight readings to the nearest 0.1 kg – when the scale says we weight 65 kg, it only means that our true weight is close to 65 kg. It then makes sense to talk about the probability of say overloading a lift when we step into it, given our weight X is around 65 kg.

2.3.1 Discrete Y

When Y is discrete, its conditional distribution is best discussed in terms of its conditional PMF[1]:

$$p_{Y|X}(y|x) = P[Y = y|x < X \leq x + h]. \tag{47}$$

Quite often, the above conditional PMF must be obtained from the conditional PDF of X given Y, using Bayes rule:

$$\begin{aligned} p_{Y|X}(y|x) &= \frac{P[x < X \leq x + h|Y = y]p_Y(y)}{P[x < X \leq x + h]} && (48) \\ &= \frac{f_{X|Y}(x|y) \cdot h \cdot p_Y(y)}{f_X(x)h} && (49) \\ &= \frac{f_{X|Y}(x|y)p_Y(y)}{f_X(x)} && (50) \end{aligned}$$

where we note that the arbitrary, small value h has disappeared.

Example 7: We continue with the $Y = X + N$ example earlier, but now obtain $p_{X|Y}(x|y)$ using (50), with X and Y swapped. This yields us

$$\begin{aligned} p_{X|Y}(x|y) &= \frac{f_{Y|X}(y|x)p_X(x)}{f_Y(y)} && (51) \\ &= \frac{e^{-(y-x)^2/2\sigma_n^2}}{e^{-(y-1)^2/2\sigma_n^2} + e^{-(y+1)^2/2\sigma_n^2}}, \quad x \in \{-1, +1\} && (52) \end{aligned}$$

where we substituted (44) and (46), and used the fact that $p_X(x) = 0.5$ for $x \in \{-1, +1\}$.

We are interested in knowing whether, upon measuring the received signal to be approximately $Y = y$, we should decide in favour of the transmitted X being -1 or $+1$. This is the simplest case of *probabilistic inference.* A logical thing to do would be to decide in favour of $\{X = -1\}$ if $p_{X|Y}(-1|y) > p_{X|Y}(1|y)$, and in favour of $+1$ otherwise. In other words, we decide that $X = -1$, if and only if

$$\begin{aligned} \frac{e^{-(y+1)^2/2\sigma_n^2}}{e^{-(y-1)^2/2\sigma_n^2} + e^{-(y+1)^2/2\sigma_n^2}} &> \frac{e^{-(y-1)^2/2\sigma_n^2}}{e^{-(y-1)^2/2\sigma_n^2} + e^{-(y+1)^2/2\sigma_n^2}} && (53) \\ \Rightarrow \quad e^{-(y+1)^2/2\sigma_n^2} &> e^{-(y-1)^2/2\sigma_n^2} && (54) \\ \Rightarrow \quad -\frac{(y+1)^2}{2\sigma_n^2} &> -\frac{(y-1)^2}{2\sigma_n^2} && (55) \\ \Rightarrow \quad (y+1)^2 &< (y-1)^2. && (56) \end{aligned}$$

Since $|y+1|$ is the "distance" between y and -1, and $|y-1|$ is the distance between y and $+1$, the "decision rule" is to choose the point in $\{-1, +1\}$ that is closest to the

[1] It will also have a conditional PDF, in terms of impulse functions.

measured received signal. Such a decision rule is also known as "minimum-distance decoding", and was derived in this example using the maximum *a posteriori* (MAP) principle, of finding the value that maximizes the "posterior" probability $p_{X|Y}(x|y)$ over $x \in S_X$. ■

The total probability principle can be applied even when conditioning on a continuous X. We simply partition the sample space according to the events

$$\{x_k < X \leq x_{k+1}\}, \quad k = 0, 1, \ldots$$

where $x_{k+1} - x_k = \Delta \to 0$, and x_0 is a very small value approximating $-\infty$. Then

$$\begin{aligned} p_Y(y) &= \sum_{k=0}^{\infty} p_{Y|X}(y|x_k) P[x_k < X \leq x_{k+1}] && (57)\\ &= \sum_{k=0}^{\infty} p_{Y|X}(y|x_k) f_X(x_k)\Delta && (58)\\ &\overset{\Delta \to 0}{=} \int_{-\infty}^{\infty} p_{Y|X}(y|x) f_X(x) dx. && (59) \end{aligned}$$

2.3.2 Continuous Y

When X and Y are both continuous, we use the conditional PDF of Y given X, $f_{Y|X}(y|x)$ to find conditional probabilities of events involving Y, i.e.

$$P[Y \in B | x < X \leq x + h] = \int_B f_{Y|X}(y|x) dy. \quad (60)$$

Bayes' Rule can be applied in order to find $f_{Y|X}(y|x)$ from $f_{X|Y}(x|y)$ and vice versa:

$$\begin{aligned} f_{Y|X}(y|x)dy &= \frac{P[y < Y \leq y + dy, x < X \leq x + dx]}{P[x < X \leq x + dx]} && (61)\\ &= \frac{P[x < X \leq x + dx | y < Y \leq y + dy] f_Y(y) dy}{f_X(x)dx} && (62)\\ &= \frac{f_{X|Y}(x|y) dx f_Y(y) dy}{f_X(x)dx} && (63)\\ \Rightarrow \ f_{Y|X}(y|x) &= \frac{f_{X|Y}(x|y) f_Y(y)}{f_X(x)} && (64) \end{aligned}$$

The total probability expression in this case is

$$f_Y(y) = \int_{-\infty}^{\infty} f_{Y|X}(y|x) f_X(x) dx. \quad (65)$$

2.4 Joint Distribution from Conditional Distribution

When X and Y are jointly discrete, or jointly continuous, they will have a joint PMF or a joint PDF, respectively. These can be obtained from the conditional and marginal PMF/PDFs, as follows:

$$p_{X,Y}(x,y) = p_{X|Y}(x|y)p_Y(y) = p_{Y|X}(y|x)p_X(x) \tag{66}$$
$$f_{X,Y}(x,y) = f_{X|Y}(x|y)f_Y(y) = f_{Y|X}(y|x)f_X(x). \tag{67}$$

The derivation of the PMF expression is easily obtained from the definition of joint, marginal and conditional PMFs. The derivation of the PDF expression simply requires us to consider the events $\{x < X \le x + dx\}$ and $\{y < Y \le y + dy\}$, and is left as an exercise.

3 Summary

In this week, we discussed the following topics:

- The mean value of $Z = g(X, Y)$, in particular $E[X + Y]$ and $E[XY]$.
- The idea of a correlation coefficient $\rho_{X,Y}$ to quantify the degree of influence X has on Y and vice versa. This is a rather imperfect tool, because it only measures the degree of *linear* correlation.
- We saw that uncorrelated X and Y may or may not be independent, but that independent X and Y must necessarily be uncorrelated.
- Conditioning on a random variable, say X – if X is discrete, we condition on $\{X = x\}$; if X is continuous, we condition on $\{x < X \le x + dx\}$, where dx is vanishingly small.
- Depending on whether Y is discrete or not, we use the conditional PMF $p_{Y|X}(y|x)$ or the conditional PDF $f_{Y|X}(y|x)$. Conditional CDFs can also be defined, though they are not too useful in practice.
- Bayes rule and the theorem on total probability can be applied to conditional distributions, and form the bedrock of the fields of detection and estimation, and probabilistic inference.

4 Diagnostic Questions

1. If X and Y are jointly uniform within a disc of radius R, show that they are uncorrelated but not independent.
2. If X and Y are jointly uniform in the unit square $\{(x, y) : 0 \le x \le 1, 0 \le y \le 1\}$, show that they are independent by finding $f_{Y|X}(y|x)$ and showing that it is equal to $f_Y(y)$.

3. Suppose that X is binomial with $n = 5$ and $p = 0.1$, and conditioned on $\{X = k\}$, Y is Gassian with mean k and unit variance. Find $E[XY]$.

A Mean of $g(X, Y)$ In General

Suppose X is discrete, and Y continuous. Then we can show using relative frequency and limits that for discrete X and continuous Y,

$$E[g(X,Y)] = \int_{-\infty}^{\infty} \sum_{k} g(x_k, y) p(x_k|y) f_Y(y) dy. \tag{68}$$

The inner expectation is over the conditional PMF $p_{X|Y}(x|y)$, and the outer one over Y.

Similarly, when X is continuous and Y is discrete,

$$E[g(X,Y)] = \sum_{k} \int_{-\infty}^{\infty} g(x, y_k) f(x|y_k) dx \cdot p_Y(y_k). \tag{69}$$

Therefore, it is possible to find the mean of $g(X, Y)$ for all types of X and Y, but when X is of one type and Y of the another, we have to use conditional distributions.