ESE 520 Probability and Stochastic Processes

Lecture 12

"Conditional expectation of one variable with respect to another. Application to mean least squares estimation (MLSE-estimation)"

Let $X, Y : (\Omega, \mathcal{F}, P) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ be two random variables (discrete or continuous). As it is well-known, X and Y generate on Ω the σ-algebras

$$\mathcal{F}^X := \{X^{-1}(B), B \in \mathcal{B}(\mathbb{R})\} \subset \mathcal{F}$$

and

$$\mathcal{F}^Y := \{Y^{-1}(B), B \in \mathcal{B}(\mathbb{R})\} \subset \mathcal{F}.$$

Given X and Y are integrable, that is $E|X| < \infty$ and $E|Y| < \infty$, and $\mathcal{G} \subset \mathcal{F}$ is a sub-σ-algebra of $\mathcal{F}$, one can define what are called the *conditional expectations of X and Y with respect to $\mathcal{G}$*:

$$E[X|\mathcal{G}] \text{ and } E[Y|\mathcal{G}].$$

If $\mathcal{G} = \{\emptyset, \Omega\}$ is the trivial σ-algebra on Ω, then

$$E[X|\mathcal{G}] = E[X] \text{ and } E[Y|\mathcal{G}] = E[Y].$$

The concept of conditional expectation is arguably a difficult one and we won't discuss in detail its definition and prove all related properties.

However, we notice some properties of it and will consider an application related to estimation of one signal by information available from another.

First, in general, $E[X|\mathcal{G}]$ is a *random variable* not a constant.

Notation-wise, $E[X|\mathcal{F}^Y]$ is often denoted by $E[X|Y]$.

The following result can be proven: There exists a Borel-measurable function $\phi : \mathbb{R} \to \mathbb{R}$ such that

$$E[X|Y] = \phi(Y). \tag{1}$$

To understand better the relation (1), recall the definition of conditional density of X given $\{Y = y\}$:

$$f_{X|Y=y}(x,y) := \frac{f_{XY}(x,y)}{f_Y(y)}$$

and the calculation of the curve of regression of X on Y (in continuous case):

$$E[X|Y=y] = \int_{\mathbb{R}} x f_{X|Y=y}(x,y)dx := \phi(y). \tag{2}$$

Formally, replacing $\{Y = y\}$ by just Y "we obtain" from (2) the relation (1).

Example 1. Consider a jointly continuous vector (X, Y) with the joint pdf

$$f_{XY}(x,y) = \begin{cases} xe^{-x(1+y)}, & x > 0, y > 0 \\ 0, & x \le 0, y \le 0. \end{cases}$$

Find $E[X|Y]$.

Solution: We have just to find $\phi(y)$ as in (2) and substitute Y into ϕ.

We first find $f_Y(y)$:

$$f_Y(y) = \int_0^\infty f_{XY}(x,y)dx = \int_0^\infty xe^{-x(1+y)}dx =$$

$$\frac{1}{1+y}\int_0^\infty (1+y)xe^{-x(1+y)}dx = \frac{1}{(1+y)^2}$$

since the integral $\int_0^\infty (1+y)xe^{-x(1+y)}dx$ is equal to the expectation of the exponential distribution with parameter $\lambda = 1+y$ which is equal to $\frac{1}{\lambda}$.

By the formula (2) one has then

$$E[X|Y=y] = \int_0^\infty x f_{X|Y}(x,y)dx = \int_0^\infty (1+y)^2x^2e^{-x(1+y)}dx =$$

$$(1+y)\int_0^\infty (1+y)x^2e^{-x(1+y)}dx = \frac{2}{1+y}$$

since the integral $\int_0^\infty (1+y)x^2e^{-x(1+y)}dx$ is equal to the second moment of the exponential distribution with parameter $\lambda = 1+y$ which is equal to $\frac{2}{\lambda^2}$.

The answer is then $E[X|Y] = \frac{2}{1+Y}$.

Conditional expectation has a lot of important properties but we mention only one here.

Proposition. For two integrable random variables X and Y and any Borel-measurable function $\psi : \mathbb{R} \to \mathbb{R}$ one has

$$E[\psi(Y)E[X|Y]] = E[\psi(Y)X]. \tag{3}$$

In particular, for $\psi = 1$, one has then

$$E[E[X|Y]] = E[X] = constant.$$

Proof. Formally, we have (in continuous case):

$$E\left[\psi(Y)E[X|Y]\right] = \int_{\mathbb{R}} \psi(y)\left(\int_{\mathbb{R}} x\frac{f_{XY}(x,y)}{f_Y(y)}dx\right) f_Y(y)dy =$$

$$\int_{\mathbb{R}}\int_{\mathbb{R}} x\psi(y)f_{XY}(x,y)dxdy = E[\psi(Y)X].$$

As next, we have the following important

Theorem. Assume the same conditions as in the Proposition before. Then, it holds

$$E[(X - \psi(Y))^2] \geq E[(X - E[X|Y])^2]. \tag{4}$$

Interpretation: The best estimate of the random variable X based on knowledge of Y is provided by the conditional expectation of X given Y. The estimate is understood in the sense of minimizing the difference $[X - \psi(Y)]$ in *mean square.* That's why $E[X|Y]$ is called *the mean least squares estimate* (MLSE), the best one.

Proof. The estimate (4) follows from property (3):

$$E[(X - \psi(Y))^2] = E[(X - E[X|Y]) + (E[X|Y] - \psi(Y))]^2 =$$

$$E(X - E[X|Y])^2 + 2E\Big(X - E[X|Y])(E[X|Y] - \psi(Y)\Big) + (E[X|Y] - \psi(Y))^2 =$$

$$E(X - E[X|Y])^2 + (E[X|Y] - \psi(Y))^2 + 0.$$

Since $(E[X|Y] - \psi(Y))^2 \geq 0$, the statement follows. $\square$

Remark 1. Everything remains the same when we choose X and Y to be vectors: $X = (X_1, ..., X_n)^T, Y = (Y_1, ..., Y_m)^T$, where $n, m \geq 1$.

One important case is when the vectors X and Y are Gaussian vectors. We can calculate then the best MLSE directly.

In other words, assume that X and Y have mean vectors μ_X and μ_Y, respectively and the (positive definite) covariance matrix R of the form (written as a block matrix)

$$R = \begin{bmatrix} R_{XX} & R_{XY} \\ R_{YX} & R_{YY} \end{bmatrix},$$

where R_{XX} and R_{YY} are the covariance matrices of the vectors X and Y, respectively, and

$$R_{XY} = cov(X, Y) := E[(X - \mu_X)(Y - \mu_Y)^T]$$

is the *cross-covariance matrix* of X and Y. Clearly, $R_{XY} = R_{YX}^T$. Also, R_{XX} is then $n \times n$ matrix, R_{YY} is $m \times m$ matrix, and R_{XY} is $n \times m$ matrix.

Proposition. It holds

$$E[X|Y] = \mu_X + R_{XY} R_{YY}^{-1}(Y - \mu_Y). \tag{5}$$

Proof. We need to calculate the conditional density:

$$f_{X|Y}(x, y) =$$

$$\frac{\sqrt{(2\pi)^m \det R_{YY}} \exp[-\frac{1}{2}(x - \mu_X, y - \mu_Y)^T R^{-1}(x - \mu_X, y - \mu_Y)]}{\sqrt{(2\pi)^{m+n} \det R} \exp[-\frac{1}{2}(y - \mu_Y)^T R_{YY}^{-1}(y - \mu_Y)]}. \tag{6}$$

Let S be the inverse matrix of R: $S = R^{-1}$. In block matrix form we can write it as

$$\begin{matrix} S_{XX} & S_{XY} \\ S_{YX} & S_{YY} \end{matrix} \tag{7}$$

where

$$RS = SR = I \tag{8}$$

with I being the identity matrix.
We obtain from (7) and (8) that

$$S_{XX}R_{XX} + S_{XY}R_{YX} = I,$$

or, equivalently,

$$R_{XX} + S_{XX}^{-1}S_{XY}R_{YX} = S_{XX}^{-1}. \tag{9}$$

Also, it follows from (7) and (8) that

$$S_{XX}R_{XY} + S_{XY}R_{YY} = 0,$$

or

$$R_{XY} = -S_{XX}^{-1}S_{XY}R_{YY},$$

or,

$$R_{XY}R^{-1} = S_{XX}^{-1}S_{XY}. \tag{10}$$

One obtains from (9) and (10) that

$$S_{XX}^{-1} = R_{XX} - R_{XY}R_{YY}^{-1}R_{YX}. \tag{11}$$

As next, we have to look at the exponent in the expression (6):

$$-\frac{1}{2}(x-\mu_X, y-\mu_Y)^T R^{-1}(x-\mu_X, y-\mu_Y) + \frac{1}{2}(y-\mu_Y)^T R_{YY}^{-1}(y-\mu_Y) =$$

$$-\frac{1}{2}[(x-\mu_X)^T S_{XX}(x-\mu_X) + (y-\mu_Y)^T S_{YX}(x-\mu_X) + (x-\mu_X)^T S_{XY}(y-\mu_Y) +$$
$$(y-\mu_Y)^T(S_{YY} - R_{YY}^{-1})(y-\mu_Y)] \widetilde{=}$$

Now we will what is called "complete the square" in the last relation. One has then

$$\widetilde{=} -\frac{1}{2}(x-\mu_X + S_{XX}^{-1}S_{XY}(y-\mu_Y))^T S_{XX}(x-\mu_X + S_{XX}^{-1}S_{XY}(y-\mu_Y)) +$$
$$\frac{1}{2}(y-\mu_Y)^T[R_{YY}^{-1} - S_{YY} + S_{YX}S_{XX}^{-1}S_{XY}](y-\mu_Y).$$

Since

$$R_{YY}^{-1} - S_{YY} + S_{YX}S_{XX}^{-1}S_{XY} = 0,$$

we have then that the exponent is equal to

$$-\frac{1}{2}(x-\mu_X + S_{XX}^{-1}S_{XY}(y-\mu_Y))^T S_{XX}(x-\mu_X + S_{XX}^{-1}S_{XY}(y-\mu_Y))).$$

Furthermore, we can also calculate that (working with block matrices)

$$\det R = \det(R_{XX} - R_{XY}R_{YY}^{-1}R_{YX}) \det R_{YY} = \det S_{XX}^{-1} \det R_{YY}.$$

In summary, we calculated that

$$f_{X|Y}(x,y) = \frac{1}{\sqrt{(2\pi)^n \det S_{XX}^{-1}}}$$

$$\exp -\frac{1}{2}(x - \mu_X + S_{XX}^{-1}S_{XY}(y-\mu_Y))^T S_{XX}(x - \mu_X + S_{XX}^{-1}S_{XY}(y-\mu_Y)).$$

In other words, $f_{X|Y}(x,y)$ is the conditional density of a Gaussian vector with mean

$$\mu_X - S_{XX}^{-1}S_{XY}(y-\mu_Y) = \mu_X + R_{XY}R_{YY}^{-1}(y-\mu_Y)$$

and the covariance matrix S_{XX}^{-1}. $\square$

Remark 2. The formula (5) remains true for *arbitrary vectors* X and Y if we restrict ourselves to best *linear estimators* of X in terms of Y in mean least squares sense. Then, the best linear estimator has the form (5).

Example 2. Let X and W be i.i.d. random variables with $\mathcal{N}(0,1)$ distribution. Define

$$Y := X^3 + W.$$

Find the best *linear* mean least squares estimator of X by Y.

Solution: Following Remark 2, we calculate:

$$R_{XY} = cov(X,Y) = E[(X-\mu_X)(Y-\mu_Y)] = E[XY] - (EX)(EY) =$$

$$E[XY] = [X^4 + XW] = E[X^4] + (EX)(EW) = E[X^4] = 3 \times 1 = 3$$

using the fact that $EX = 0$, the independence of X and W and the formula for the even moments of $\mathcal{N}(0,1)$ distribution we derived earlier.

In a similar way we obtain:

$$R_{YY} = E[(X^3+W)(X^3+W)] = E[X^6 + 2X^3W + W^2] =$$

$$E[X^6] + 2E[X^3](EW) + E[W^2] = 5 \times 3 \times 1 + 1 = 16$$

since EW=0.

In summary, the best estimator is then

$$E[X|Y] = \frac{3}{16}Y$$

since $\mu_X = \mu_Y = 0$.