Conditional Probability Theory

Ioanid Rosu

1 Introduction

Conditional probability theory is one of the most difficult parts of basic probability theory. The reason is that it is hard to come up with good intutitions for it. Just for fun, let me give the formal definition of conditional expectations. Start with a probability space $(\Omega, \mathcal{F}, P)$, with Ω the event space, $\mathcal{F}$ a σ-algebra on Ω, and P a probability on $\mathcal{F}$, i.e., P is a positive real measure on $(\Omega, \mathcal{F})$ with $P(\Omega) = 1$. Let $\mathcal{B}$ be a sub-σ-algebra of $\mathcal{F}$. Let f be a random variable on Ω (which is $\mathcal{F}$-measurable, but not necessarily $\mathcal{B}$-measurable). Then the conditional expectation $\mathbf{E}(f|\mathcal{B})$ of f given $\mathcal{B}$ is a $\mathcal{B}$-measurable function g on Ω which satisfies

$$g = \mathbf{E}(f|\mathcal{B}) \quad \Longleftrightarrow \quad \int_B f \, \mathrm{d}P = \int_B g \, \mathrm{d}P \quad \forall B \in \mathcal{B}. \tag{1}$$

Not very enlightening now, is it? I think that, unless you are a mathematical genius, the most reasonable intuition you can come up with by just looking at the definition is this: Let's assume f is a simple function, i.e., f is constant on some partition of Ω by $\mathcal{F}$-measurable sets. Since $\mathcal{F}$ has a finer "resolution" than $\mathcal{B}$ (there are more sets in $\mathcal{F}$), conditional expectation of f given $\mathcal{B}$ must be some way of forcing f to be identified by someone who only "sees" $\mathcal{B}$. The way to do that is by averaging out f on those smaller subsets of $\mathcal{F}$, and making it constant on the larger subsets of $\mathcal{B}$.

Directly from the definition we can deduce two properties which are used a lot in practice:

- Notice that if f is already $\mathcal{B}$-measurable, we can take $g = f$ in the above definition, hence $\mathbf{E}\{f|\mathcal{B}\} = f$. In other words, $\mathcal{B}$-measurable functions are treated as constants by the operator $\mathbf{E}\{\cdot|\mathcal{B}\}$. In particular, if X is any ($\mathcal{F}$-measurable) random variable, we deduce this important property about conditional expectations:

$$\mathbf{E}\{fX \mid \mathcal{B}\} = f\mathbf{E}\{X|\mathcal{B}\} \quad \forall f \ \ \mathcal{B}\text{-measurable}. \tag{2}$$

- Suppose $X : \Omega \to \mathbb{R}$ is a $\mathcal{F}$-measurable random variable. Define $\mathbf{E}\{f|X\} = \mathbf{E}\{f|\mathcal{B}_X\}$, where $\mathcal{B} = \mathcal{B}_X$ is the σ-algebra generated by all sets of the form $X^{-1}(B)$, with B a Borel set[1] in $\mathbb{R}$. Then we have the following property:

$$\mathbf{E}\{f|X\} = \phi(X) \quad \text{ with } \phi : \mathbb{R} \to \mathbb{R} \text{ measurable}. \tag{3}$$

Date: March 17,2003.

[1] A Borel set on $\mathbb{R}$ is simply an element in the σ-algebra $\mathcal{B}_{\mathbb{R}}$ generated by all open intervals in $\mathbb{R}$. Equivalently, a Borel set can be obtained by taking a countable number of intersections, unions, and complements of open intervals.

To explain briefly why this is the case, let $g = \mathbf{E}\{f|X\}$, which according to the definition of conditional expectation is $\mathcal{B}_X$-measurable. But for a function $g : \Omega \to \mathbb{R}$ to be $\mathcal{B}_X$-measurable is the same as g being a composition of the form $\phi \circ X$, with $\phi : \mathbb{R} \to \mathbb{R}$ *measurable*. (It is trivial that $g = \phi \circ X$ as functions; the fact that ϕ is measurable comes from the fact that X induces a one-to-one correspondence between $\mathcal{B}_X$ and $\mathcal{B}_\mathbb{R}$ restricted to $\operatorname{Im} X \subseteq \mathbb{R}$.

But enough about the definition of conditional expectations. I'm not a big fan of it, although this is what you have to use in order to prove formulas. Luckily, there is an easier way of interpreting conditional expectations, but it comes at the cost of losing some generality. Namely, if we restrict ourselves at the subclass of "square-integrable" random variables, we can come up with a nice geometric intuition. This is because the square-integrable functions form something called a Hilbert space, where one can do geometry in the usual way.

We define $\mathcal{H}$ the Hilbert space[2] of square-integrable $\mathcal{F}$-measurable functions on Ω by

$$(X, Y) = \mathbf{E}\{XY\}. \tag{4}$$

The great thing about Hilbert spaces is that the apparently innocuous inner product generates a lot of structure, to the extent that in a Hilbert space one can work smoothly with geometric concepts such as distance, projection, angles, etc. For example, the distance $d(X, Y)$ and the angle $\alpha(X, Y)$ are defined as follows:

$$d(X, Y) = \mathbf{E}\{(X - Y)^2\}^{\frac{1}{2}}, \quad \cos\alpha(X, Y) = \frac{\mathbf{E}\{XY\}}{\mathbf{E}\{X^2\}^{\frac{1}{2}}\mathbf{E}\{Y^2\}^{\frac{1}{2}}}. \tag{5}$$

Notice that if we restrict our attention at the subspace $\mathcal{H}_0 = \{X \in \mathcal{H} \mid \mathbf{E}X = 0\}$ of de-meaned random variables in $\mathcal{H}$, the formula for the cosine of the angle becomes

$$\cos\alpha(X, Y) = \operatorname{corr}(X, Y) \quad \forall X, Y \in \mathcal{H}_0, \tag{6}$$

where $\operatorname{corr}(X, Y)$ denotes correlation between X and Y. Also, we say that X and Y are perpendicular, and write $X \perp Y$, if $\cos\alpha(X, Y) = 0$. Equivalently,

$$X \perp Y \quad \iff \quad \mathbf{E}\{XY\} = 0. \tag{7}$$

What is then the projection of an element $X \in \mathcal{H}$ on the subspace $\mathcal{H}_\mathcal{B}$ of all $\mathcal{B}$-measurable functions in $\mathcal{H}$? It must be some element $Y \in \mathcal{H}_\mathcal{B}$ so that $(b - Y) \perp (X - Y)$ for all b in $\mathcal{H}_\mathcal{B}$. But Y is also in $\mathcal{H}_\mathcal{B}$, so we can replace b by $b+Y$, and deduce that $b \perp (X-Y)$ for all b in $\mathcal{H}_\mathcal{B}$. This means that $\mathbf{E}\{b(X - Y)\} = 0$ for all b in $\mathcal{H}_\mathcal{B}$. In particular, take $b = \mathbf{1}_B$, the indicator function of some subset $B \in \mathcal{B}$. We then get $E\{\mathbf{1}_B(X - Y)\} = 0$, which translates into $\int_B X \, \mathrm{d}P = \int_B Y \, \mathrm{d}P$. The subset B was arbitrary, so it follows that Y satisfies our definition of conditional expectation of X with respect to B. In other words,

$$\operatorname{Proj}_{\mathcal{H}_\mathcal{B}} X = \mathbf{E}\{X|\mathcal{B}\}, \tag{8}$$

where $\operatorname{Proj}_{\mathcal{H}_\mathcal{B}} X$ is the projection of X on $\mathcal{H}_\mathcal{B}$.

[2] A real Hilbert space is a real vector space $\mathcal{H}$, endowed with an inner product $(\cdot, \cdot) : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$, so that $(x, x) \geq 0$ with equality iff $x = 0$, and such that with the resulting metric $\mathcal{H}$ is complete. The metric is defined by $d(x, y) = (x - y, x - y)^{1/2}$. For details see for example Rudin, "Real and Complex Analysis".

One last subtle point here is about the existence of the projection of X on the closed subspace $\mathcal{L} = \mathcal{H}_{\mathcal{B}}$ (the fact that $\mathcal{L}$ is closed is proved by some easy measure theory). The existence of the projection $Y = \mathrm{Proj}_{\mathcal{L}} X$ is a standard fact in Hilbert space theory, and is proved as follows: First notice that $\mathcal{L}$ is a Hilbert space in its own right (easy). Then one can define the continuous linear functional $\Phi(\cdot) = (X, \cdot) : \mathcal{L} \to \mathbb{R}$. According to the Riesz' representation theorem, there exists $Y \in \mathcal{L}$ so that $\Phi(\cdot) = (Y, \cdot)$ on $\mathcal{L}$. But this is equivalent to $(X - Y, b) = 0$ for all $b \in \mathcal{L}$, which by the definition of projection means exactly that Y is the projection of X on $\mathcal{L}$. So the crucial step in this proof is Riesz' theorem, which only makes use of the fact that $\mathcal{H}$ is complete and $\mathcal{L}$ is closed (and of course plays around with the inner product).

By contrast, to see what we gained by reducing the generality from all measurable functions to the square-integrable ones (and thus entering the neat world of Hilbert spaces), let's see what is involved in the general existence proof for the conditional expectation $g = \mathbf{E}\{f|\mathcal{B}\}$ in (1). First notice that the measure $B \mapsto \mu(B) = \int_B f \, \mathrm{d}P$ is absolutely continuous with respect to P (that's easy). Then the hard part is proved by Radon–Nikodym, namely that there exists g a $\mathcal{B}$-measurable function such that $\mu(B) = \int_B g \, \mathrm{d}P$. But then, given our definition (1), g is exactly the conditional expectation of f given $\mathcal{B}$. The Radon–Nikodym theorem is one of the deepest theorems in measure theory, and its proof is one of the most technical proofs I know, given such an elementary statement. I dare you to read the proof and remember it after just one day.

This concludes our discussion about the geometric interpretation of the conditional expectation. Now we want to put it to use.

2 Formulas

There are two basic formulas in conditional probability theory: the law of iterated expectations (9), also called the ADAM formula, and the EVE formula (10)[3]. Let X be a $\mathcal{F}$-measurable random variable, and $\mathcal{B}' \subseteq \mathcal{B}$ two sub-σ-algebras of $\mathcal{F}$. Then:

$$\mathbf{E}\{X|\mathcal{B}'\} = \mathbf{E}\big\{\, \mathbf{E}\{X|\mathcal{B}\} \,\big|\, \mathcal{B}' \,\big\}, \tag{9}$$

$$\mathbf{V}\{X\} = \mathbf{E}\big\{\, \mathbf{V}\{X|\mathcal{B}\} \,\big\} + \mathbf{V}\big\{\, \mathbf{E}\{X|\mathcal{B}\} \,\big\}. \tag{10}$$

Both of them have a nice geometric interpretation, which we now discuss.

The law of iterated expectations (9) is quite simple. With our interpretation of conditional expectations, it says that projecting X on the smaller subspace $\mathcal{H}_{\mathcal{B}'}$ gives the same result as first projecting X on the larger subspace $\mathcal{H}_{\mathcal{B}}$, and then taking $\mathrm{Proj}_{\mathcal{H}_{\mathcal{B}}} X$ and projecting it further on $\mathcal{H}_{\mathcal{B}'}$. This has a simple geometric proof if you draw the picture (I learned it in 8'th grade as the theorem of the 3 perpendiculars). But, to be more rigorous, consider the decomposition of $\mathcal{H}$ as an orthogonal sum, $\mathcal{H} = \mathcal{H}_{\mathcal{B}} \oplus \mathcal{H}_{\mathcal{B}}^{\perp}$. Notice that via this decomposition the first component of X is exactly the projection $\mathrm{Proj}_{\mathcal{H}_{\mathcal{B}}} X$. Now $\mathcal{H}_{\mathcal{B}}$ also has an orthogonal decomposition, $\mathcal{H}_{\mathcal{B}} = \mathcal{H}_{\mathcal{B}'} \oplus \mathcal{H}_{\mathcal{B}'\mathcal{B}}^{\perp}$. Then the projection of $\mathrm{Proj}_{\mathcal{H}_{\mathcal{B}}} X$ on the first component, $\mathcal{H}_{\mathcal{B}'}$, is $\mathrm{Proj}_{\mathcal{H}_{\mathcal{B}'}} \mathrm{Proj}_{\mathcal{H}_{\mathcal{B}}} X$. We now look at the decomposition $\mathcal{H} = \mathcal{H}_{\mathcal{B}'} \oplus \mathcal{H}_{\mathcal{B}'\mathcal{B}}^{\perp} \oplus \mathcal{H}_{\mathcal{B}}^{\perp}$, which

[3] If you haven't figured it out yet, the reason these are called the ADAM and EVE formulas comes from the form of the EVE equation (10), which should perhaps be called the EVVE equation. The name ADAM comes then by association with EVE, and from the fact that in some sense it is the first formula of conditional probability theory (or it might just be a plot of the feminist movement to have ADAM be begotten by EVE, and not the other way around, as it should be :)

is the same as the decomposition $\mathcal{H} = \mathcal{H}_{\mathcal{B}'} \oplus \mathcal{H}_{\mathcal{B}'}^{\perp}$. This shows that

$$\operatorname{Proj}_{\mathcal{H}_{\mathcal{B}'}} X = \operatorname{Proj}_{\mathcal{H}_{\mathcal{B}'}} \operatorname{Proj}_{\mathcal{H}_{\mathcal{B}}} X, \tag{11}$$

which is the same as law of iterated expectations (9).

The EVE formula nicely turns out to be just the Pythagorean theorem in the Hilbert space $\mathcal{H}$. Denote by $Y = \mathsf{E}\{X|\mathcal{B}\}$. Then to prove (10) it is enough to show it for X of zero mean, i.e., for $X \in \mathcal{H}_0$ (by the law of iterated expectations, Y also has zero mean). So let $X \in \mathcal{H}_0$. Then apply the Pythagorean theorem for the triangle OXY (the right triangle is Y):

$$d(O,X)^2 = d(O,Y)^2 + d(X,Y)^2. \tag{12}$$

Now notice that $d(O,X)^2 = \mathsf{E}\{X^2\} = \mathsf{V}\{X\}$; $d(O,Y)^2 = \mathsf{E}\{Y^2\} = \mathsf{V}\big\{\ \mathsf{E}\{X|\mathcal{B}\}\ \big\}$; and $d(X,Y)^2 = \mathsf{E}\big\{\big(X - \mathsf{E}\{X|\mathcal{B}\}\big)^2\big\} = \mathsf{E}\big\{\ \mathsf{E}\big\{\big(X - \mathsf{E}\{X|\mathcal{B}\}\big)^2 \,|\, \mathcal{B}\ \big\}\big\}$, where the last equality comes from the law of iterated expectations. But $\mathsf{E}\big\{\big(X - \mathsf{E}\{X|\mathcal{B}\}\big)^2 \,|\, \mathcal{B}\ \big\} = \mathsf{V}\{X|\mathcal{B}\}$, the variance of X conditional on $\mathcal{B}$, so $d(X,Y)^2 = \mathsf{E}\big\{\ \mathsf{V}\{X|\mathcal{B}\}\ \big\}$. Putting together the three terms in the Pythagorean theorem, we get exactly formula (10).

But let's give the formal proof of (9), since it does add some value. Let's look at the sum $X = \mathsf{E}\{X|\mathcal{B}\} + \big(X - \mathsf{E}\{X|\mathcal{B}\}\big)$. This would corresponds to the orthogonal decomposition $\mathcal{H} = \mathcal{H}_{\mathcal{B}} \oplus \mathcal{H}_{\mathcal{B}}^{\perp}$, but remember that we are not in $\mathcal{H}$ anymore, so we have to try to get by without Hilbert space arguments. First, we show that the covariance of the two terms is zero:

$$\operatorname{cov}\big(\mathsf{E}\{X|\mathcal{B}\}, X - \mathsf{E}\{X|\mathcal{B}\}\big) = \mathsf{E}\Big\{\ \mathsf{E}\{X|\mathcal{B}\} \cdot \big(X - \mathsf{E}\{X|\mathcal{B}\}\big)\ \Big\} =$$
$$\mathsf{E}\Big\{\mathsf{E}\Big\{\ \mathsf{E}\{X|\mathcal{B}\} \cdot \big(X - \mathsf{E}\{X|\mathcal{B}\}\big) \,|\, \mathcal{B}\Big\}\Big\} = \mathsf{E}\Big\{\mathsf{E}\{X|\mathcal{B}\} \cdot \mathsf{E}\big\{X - \mathsf{E}\{X|\mathcal{B}\} \,|\, \mathcal{B}\big\}\Big\} = 0.$$

Here we used the law of iterated expectations in all the equalities except for the third one, where we used (2). Since covariance is zero, it follows that the variance of the sum is the sum of the variances. To get EVE, we only have to calculate $\mathsf{V}\big\{X - \mathsf{E}\{X|\mathcal{B}\}\big\}$, which, since the inside has zero mean, equals $\mathsf{E}\big\{\big(X - \mathsf{E}\{X|\mathcal{B}\}\big)^2\big\}$. But, as we saw above, this indeed equals $\mathsf{E}\big\{\ \mathsf{V}\{X|\mathcal{B}\}\ \big\}$.

3 Applications

A nice application of the above discussion is when we look at linear regression of normal variables. Suppose Y and X are normal variables (bivariate normal, to be more precise) with marginal distributions $X \sim N(\mu_X, \sigma_X^2)$ and $Y \sim N(\mu_Y, \sigma_Y^2)$. We consider a regression of the type

$$Y = a + bX + \epsilon, \quad \text{with } \operatorname{cov}(X,\epsilon) = 0, \mathsf{E}\epsilon = 0. \tag{13}$$

The regression coefficients are computed as usual (using $\operatorname{cov}(X,\epsilon) = 0$):

$$b = \beta_{YX} = \operatorname{cov}(Y,X)/\operatorname{var}(X), \quad a = \mu_Y - b\mu_X. \tag{14}$$

Equation (13) looks suspiciously like the orthogonal decomposition we had before. But we are not quite ready to say that $\mathsf{E}\{Y|X\} = a + bX$. All we know is that $\mathsf{E}\{Y|X\}$ should be of the form $\phi(X)$, with ϕ measurable (because of (3)). But why should ϕ be linear? Here's

the idea: First, denote the space $\mathcal{H}_{\mathcal{B}_X}$ by $\mathcal{H}_X$. Now $\mathcal{H}_X$ is the space of all $\mathcal{B}_X$-measurable functions, which from the discussion after equation (3) is the same as the space of all $\phi(X)$, with ϕ measurable. So to say that $\mathbf{E}\{Y|X\} = a+bX$ is the same as showing that $Y-(a+bX)$ is orthogonal on every function of the form $\phi(X)$ with ϕ measurable. Equivalently, we have to show that ϵ is uncorrelated with all $\phi(X)$. From the regression assumption, we know that ϵ is uncorrelated with X, but why would it be uncorrelated with all $\phi(X)$?

Now comes the crucial fact: for (bivariate) normal variables ϵ and X, uncorrelated implies *independent* (for proofs see Cassella and Berger). But since ϵ is independent from X, it is independent from any $\phi(X)$, hence it is also uncorrelated to all $\phi(X)$, and we are done. We just showed that $\mathbf{E}\{Y|X\} = a + bX$, so the decomposition $Y = (a + bX) + \epsilon$ is orthogonal. The EVE formula becomes $\operatorname{var}(Y) = \operatorname{var}(a + bX) + \operatorname{var}(\epsilon)$, and this implies

$$\mathbf{E}\operatorname{var}(\epsilon|X) = \operatorname{var}(\epsilon) = \operatorname{var}(\epsilon|X). \tag{15}$$

The second application uses again the geometric intuition we gave for the law of iterated expectations and the EVE formula. Let X be a $\mathcal{F}$-measurable random variable, and $\mathcal{B}' \subseteq \mathcal{B}$ two σ-algebras of $\mathcal{F}$, as in the previous section. In some sense $\mathbf{V}\{X|\mathcal{B}\}$ measures the informativeness of $\mathcal{B}$ relative to X. Denote by $Y = \mathbf{E}\{X|\mathcal{B}\}$ and by $Y' = \mathbf{E}\{X|\mathcal{B}'\}$. In the triangle XYY', the right angle is Y. As we saw in the proof of EVE, $d(X,Y)^2 = \mathbf{EV}\{X|\mathcal{B}\}$ and $d(X,Y')^2 = \mathbf{EV}\{X|\mathcal{B}'\}$. So we get

$$\mathbf{EV}\{X|\mathcal{B}'\} - \mathbf{EV}\{X|\mathcal{B}\} = \mathbf{E}\Big\{\big(\mathbf{E}\{X|\mathcal{B}\} - \mathbf{E}\{X|\mathcal{B}'\}\big)^2\Big\}. \tag{16}$$

To connect with our previous discussion about normal variables, suppose u_t is a Gaussian process (with normally distributed increments $\mathrm{d}u_t$); u_t is a publicly observed variable. Define by $\mathcal{I}_t$ the σ-algebra generated by all u_s for $s \leq t$; $\mathcal{I}_t$ is the public information process. Let v be some other normal variable, to be thought of as the "true value" of some asset; v is not public information;. Let p_t be the price process of the asset. Then, if markets are efficient, we have

$$p_t = \mathbf{E}\{v|\mathcal{I}_t\}. \tag{17}$$

Also, if we denote by ϵ_t the regression error, we have $v = p_t + \epsilon_t$. To measure the informativeness of prices at time t, we look at

$$\Sigma_t = \mathbf{E}\operatorname{var}\{v|\mathcal{I}_t\} = \mathbf{E}\operatorname{var}\{\epsilon_t|\mathcal{I}_t\} = \operatorname{var}\epsilon_t, \tag{18}$$

where the last equality comes from the independence of ϵ_t from $\mathcal{I}_t$, as we saw in the previous application. We want to study the dynamics of Σ_t. Consider $\mathcal{B}' = \mathcal{I}_t$ and $\mathcal{B} = \mathcal{I}_{t+\,\mathrm{d}t}$. Then applying formula (16), we get $\Sigma_t - \Sigma_{t+\,\mathrm{d}t} = \mathbf{E}\big\{(p_t - p_{t+\,\mathrm{d}t})^2\big\}$. But $-\,\mathrm{d}p_t = \mathrm{d}\epsilon_t$, so we get

$$-\,\mathrm{d}\Sigma_t = \mathbf{E}\big\{(\,\mathrm{d}p_t)^2\big\} = \mathbf{E}\big\{(\,\mathrm{d}\epsilon_t)^2\big\}. \tag{19}$$

These formulas, besides being pretty, are quite useful when studying continuous auctions, such as in Kyle's paper (EMA, 1985).