PROBABILITY 2018
HANDOUT 3: CONDITIONING

GIOVANNI PISTONE

Contents

Conditioning is one among the core concepts in reasoning about uncertainty in Probability, in Statistics, in Economics, in Machine Learning. In this notes we refer mainly to the textbook by D. Williams [2, Ch. 9]. A concise and fully rigorous review of the basic mathematics is in the monograph by C. Dellacherie and P.-A. Meyer [1, Ch. I-III].

1. Conditional expectation

Exercise 1. Let X be a measurable function from $(\Omega, \mathcal{F})$ to $(S, \mathcal{S})$. Let $\mathcal{G}$ be the σ-algebra generated by X i.e., $\mathcal{G} = X^{-1}\mathcal{S}$. Every $\mathcal{G}$-measurable real random variable Y is of the form $Y = f \circ X$, where f is a real random variablle on $(S, \mathcal{S})$. [Hint: If Y is simple, $Y = \sum_{j=1}^n y_j \mathbf{1}_{B_j}$, with $B_j \in \mathcal{G}$, then $B_j = X^{-1}(A_j)$, $A_j \in \mathcal{S}$. It follows that $Y = \sum_{j=1}^n y_j \mathbf{1}_{X^{-1}(A_j)} = \sum_{j=1}^n y_j \mathbf{1}_{A_j} \circ X$, hence $f = \sum_{j=1}^n y_j \mathbf{1}_{A_j}$. If X is non-negative, take an increasing sequence of simple random variable ...]

Definition 1. Let $(\Omega, \mathcal{F}, \mu)$ be a probability space, X a real random variable with finite expectation, $\mathbb{E}_\mu[|X|] < +\infty$, $\mathcal{G}$ a sub-σ-algebra of $\mathcal{F}$. A random variable $\widehat{X}$ is a *version of the conditional expectation of X given $\mathcal{G}$* if, and only if,

(1) $\widehat{X}$ is integrable and $\mathcal{G}$-measurable;
(2) for all bounded and $\mathcal{G}$-measurable random variable it holds

$$\mathbb{E}_\mu\left[G\widehat{X}\right] = \mathbb{E}_\mu[GX] \ .$$

The sub-μ in the notation is there to remember that the conditional expectation *depends on the probability*. The conditions (1) and (2) in the definition provide actual equations to compute the conditional expectation, as the following examples show.

Exercise 2. If X is a real random variable with a positive density p, let $\mathcal{G}$ be the σ-algebra generated by $|X|$. That is, the absolute value only, not the sign, is observed. In this case the conditional expectation of X given $\mathcal{G} = \sigma(|X|)$, breafly, given $|X|$, is a random variable of the form $\widehat{X} = \hat{f}(|X|)$ (condition (1)) such that $\mathbb{E}\left(\widehat{X}G\right) = \mathbb{E}(XG)$

Date: April 4, 2018.

for all $G = g(|X|)$, g bounded (condition (2)). As a density is given, we write the defining equation

$$\int \hat{f}(|x|)g(|x|)p(x)\ dx = \int xg(|x|)p(x)\ dx\ .$$

[Hint: To compute $\hat{f}$, split $\int = \int_{-\infty}^{0} + \int_{0}^{+\infty}$ and change the variable $x \to -x$ in the first integral to get

$$\int_0^{+\infty} \hat{f}(|x|)g(|x|)(p(x)+p(-x))\ dx = \int_0^{+\infty} g(|x|)(xp(x)-xp(-x))\ dx\ ,$$

hence

$$\hat{f}(|x|)(p(x)+p(-x)) = xp(x)-xp(-x)\ .$$

Finally, notice that $\frac{xp(x)-xp(-x)}{p(x)+p(-x)}$ is symmetric.]

Exercise 3. Let S_1, S_2 be independent and exponential with mean 1. The joint density is $p_{S_1,S_2}(x_1,x_2) = \mathrm{e}^{-(x_1+x_2)}(x_1,x_2>0)$. We want to compute the conditional expectation of S_1 given S_1+S_2. We need to find $\hat{f}$ such that for all bounded g we have

$$\iint_0^{\infty} \hat{f}(x_1+x_2)g(x_1+x_2)\mathrm{e}^{-(x_1+x_2)}\ dx_1dx_2 = \iint_0^{\infty} x_1 g(x_1+x_2)\mathrm{e}^{-(x_1+x_2)}\ dx_1dx_2\ .$$

[Hint. Let us make the transformation $y = x_1+x_2$, $z = x_1$. The inverse transformation is $x_1 = z$, $x_2 = y-z$ with determinant -1. We have

$$(x_1,x_2>0) = (z>0)(y-z>0) = (0<z<y)$$

then the equation becomes

$$\iint_{\{0<z<y\}} \hat{f}(y)g(y)\mathrm{e}^{-y}\ dydz = \iint_{\{0<z<y\}} zg(y)\mathrm{e}^{-y}\ dzdy\ .$$

Computing the dz integrals on both sides we get

$$\int_0^{\infty} \hat{f}(y)g(y)y\mathrm{e}^{-y}\ dy = \int_0^{\infty} g(y)\frac{y^2}{2}\mathrm{e}^{-y}\ dy\ ,$$

hence $\hat{f}(y) = \frac{y}{2}$.]

Exercise 4. Let $Z = (Z_1, Z_2) \sim \mathrm{N}_2(0, I)$ and define $X = Z_1$, $Y = Z_1+Z_2$, $\mathcal{G} = \sigma(Y)$. To compute a version of $\mathrm{E}(X|\mathcal{G})$ we look for a function $\hat{f}$ such that $\hat{f}(Y)$ satisfies

$$\mathbb{E}(Xg(Y)) = \mathbb{E}\left(\hat{f}(Y)g(Y)\right) \quad \text{for all bounded } g\ .$$

[Hint: As

$$\begin{bmatrix} X \\ Y \end{bmatrix} = \begin{bmatrix} Z_1 \\ Z_1+Z_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}\begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix}$$

we have $(X,Y) \sim \mathrm{N}_2\left(0, \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}\right)$ and $Y \sim \mathrm{N}_1(0,2)$. We have $\det\left(\begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}\right) = 1$ and $\begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}^{-1} = \begin{bmatrix} 2 & -1 \\ -1 & 1 \end{bmatrix}$ so that the density of (X,Y)

$$p_{X,Y}(x,y) = (2\pi)^{-1}\exp\left(-\frac{1}{2}(2x^2-2xy+y^2)\right)\ .$$

We want

$$\iint xg(y)\ (2\pi)^{-1}\exp\left(-\frac{1}{2}(2x^2-2xy+y^2)\right)\ dx\ dy = \int \hat{f}(y)g(y)\ (2\pi\cdot 2)^{-1/2}\exp\left(-\frac{1}{2\cdot 2}y^2\right)\ dy$$

Let us perform first the dx integration in the RHS:

$$\int x\ \exp\left(-\frac{1}{2}(2x^2-2xy+y^2)\right)\ dx = \int x\ \exp\left(-\left(x^2-xy+\frac{1}{2}y^2\right)\right)\ dx = \int x\ \exp\left(-\left(x-\frac{1}{2}y\right)^2-\frac{1}{4}y^2\right)\ dx = \exp\left(-\frac{1}{4}y^2\right)\int \pi^{1/2}x\ \pi^{-1/2}\exp\left(-\left(x-\frac{1}{2}y\right)^2\right)\ dx = \frac{\pi^{1/2}}{2}y\exp\left(-\frac{1}{4}y^2\right)\ .$$

The defining equality becomes

$$\int g(y)\ (2\pi)^{-1}\frac{\pi^{1/2}}{2}y\exp\left(-\frac{1}{4}y^2\right)\ dy = \int f(y)g(y)\ (2\pi\cdot 2)^{-1/2}\exp\left(-\frac{1}{2\cdot 2}y^2\right)\ dy$$

so that, g being generic, $\hat{f}(y) = y/2$. (We are going to see below a simpler and more principled way to do this computation.)]

1. As the equation $\mathbb{E}_\mu\left[G(\widehat{X}-X)\right]=0$, $G\in\mathcal{L}^\infty(\mathcal{G})$, is linear in G and continuous under bounded pointwise convergence, it is enough to check it for random variables of the for $\mathbf{1}_C$, $C\in\mathcal{C}$, $\mathcal{C}$ π-system generating $\mathcal{G}$. [Monotone-Class Theorem [2, ¶3.14].]

2 (Almost sure equivalence)**.** If $\widehat{X}_1$, $\widehat{X}_2$, are two versions of the conditional expectation of X, then $\mathbb{E}_\mu\left[G(\widehat{X}_1-\widehat{X}_2)\right]=0$ i.e. $\widehat{X}_1=\widehat{X}_2$ μ-almost-surely. [Take $G=\operatorname{sign}\left(\widehat{X}_1-\widehat{X}_2\right)$ to get $\mathbb{E}_\mu\left[\left|\widehat{X}_1-\widehat{X}_2\right|\right]=0$.] More generally, if $X_1=X_2$ μ-almost-surely, then $\widehat{X}_1=\widehat{X}_2$ μ-almost-surely. We write $\mathrm{E}_\mu(X|\mathcal{G})$ to denote the μ-class of versions and, with abuse of notation, $\widehat{X}=\mathrm{E}_\mu(X|\mathcal{G})$. If $L^1(\mathcal{F},\mu)$ is the vector space of classes μ-equivalent real random variables, there exists a mapping

$$L^1(\mathcal{F},\mu)\ni X\mapsto \mathrm{E}_\mu(X|\mathcal{G})\in L^1(\mathcal{G},\mu)\ .$$

3 (Existence)**.** The fact that the previous mapping is actually defined on all of $L^1(\mathcal{F},\mu)$, is discussed in [2, ¶9.5]. We skip this discussion, together with a related issue namely, the notion of μ-complete σ-algebra. Many proofs of existence are actually available, either based on some result of Functional Analysis (existence of orthogonal projection), or based on results from advanced Measure Theory such as the Radon-Nikodým Theorem (see below). Here, we are mainly focused on either *computing* a version of the conditional expectation of a given random variable, or *checking* that a random variable is a version of the conditional expectation of some random variable. We have defined the conditional

expectation for integrable random variables. It is possible to define the conditional expectation for positive random variables, see the comments below about properties of the conditional expectation.

4 (Image of a density)**.** On the measurable space $(\Omega, \mathcal{F})$, consider the probility measure μ and the probability density P. If Φ is measurable from $(\Omega, \mathcal{F})$ to $(S, \mathcal{S})$, consider the image of the probability measure $p \cdot \mu$ under Φ. The image $\nu = \Phi_{\#}(p \cdot \mu)$ is characterized by

$$\int_S g(y)\ \nu(dy) = \int_\Omega g \circ \Phi(x)\ p(x)\mu(dx), \quad g \in \mathcal{L}^\infty(S, \mathcal{S})\ .$$

Now, $g \circ \Phi$ is the generic bounded $\sigma(\Phi)$-measurable random variable, then

$$\int_\Omega g \circ \Phi(x)\ p(x)\mu(dx) = \int_\Omega g \circ \Phi(x)\ \widehat{p} \circ \Phi(x)\mu(dx)\ ,$$

where $\widehat{p} \circ \Phi$ is a version of the conditional expectation of p given $\sigma(\Phi)$. Now apply again the definition of image to the RHS to get

$$\int_S g(y)\ \Phi_{\#}(p \cdot \mu)(dy) = \int_S g(y)\widehat{p}(y)\Phi_{\#}(\mu)(dy)\ .$$

We have found the density of the image measure.

5 (Projection property)**.** Let $\mathcal{H}$ be a sub-σ-field of $\mathcal{G}$. It is easy to check that

$$\mathrm{E}_\mu\left(\mathrm{E}_\mu\left(X|\mathcal{G}\right)|\mathcal{H}\right) = \mathrm{E}_\mu\left(X|\mathcal{H}\right)\ .$$

In particular, the conditional expectation operator $X \mapsto \mathrm{E}_\mu\left(X|\mathcal{F}\right)$ is a projection operator on $L^1(\mathcal{F}, \mu)$. [One could say that it is the transposed operator of the injection operator $\mathcal{L}^\infty(\mathcal{G}) \to \mathcal{L}^\infty(\mathcal{F})$.

6 (Orthogonal projection)**.** The conditioning operator is an *orthogonal projection*. Assume Y in $L^2(\Omega, \mathcal{F}, \mu)$ that is, $\mathbb{E}\left(Y^2\right) < \infty$. If $\widehat{Y} = \mathrm{E}\left(Y|\mathcal{G}\right)$, then $\widehat{Y} \in L^2(\Omega, \mathcal{G}, \mu)$ and

$$\mathbb{E}\left((Y - \widehat{Y})Z\right) = 0\ , \quad z \in L^2(\Omega, \mathcal{G}, \mu)\ .$$

This property should not be confused with *linear regression*. Let de given $Y \in L^2$ and let $X_1, \ldots, X_m \in L^2$ be *explanatory variables*. We want a vector $\boldsymbol{\theta} = (\theta_0, \theta_1, \ldots, \theta_d) \in \mathbb{R}^{d+1}$ such that

$$\text{quadratic error} = \mathbb{E}\left(\left(Y - \theta_0 - \sum_{j=1}^{d} \theta_j X_j\right)^2\right)$$

be minimum. As a function of $\boldsymbol{\theta}$ the quadratic error is a convex function then the minimum is obtained by imposing the gradient to be zero.

Exercise 5. Check all detail of the previous paragraph.

Exercise 6 (Examples).
(1) If $\mathcal{G} = \{\varnothing, \Omega\}$, then $\mathrm{E}_\mu\left(X|\mathcal{G}\right) = \mathbb{E}_\mu\left[X\right]$.
(2) If $\mathcal{G} = \mathcal{F}$, then $\mathrm{E}_\mu\left(X|\mathcal{G}\right) = X$.
(3) Let $\{A_1, \ldots, A_n\}$ be a measurable partition of Ω and let $\mathcal{G} = \sigma(A_1, \ldots, A_n)$. Assume $\mu(A_j) \neq 0$, $j = 1, \ldots, n$. It holds

$$\mathrm{E}_\mu\left(X|\mathcal{G}\right) = \sum_{j=1}^{n} \frac{\int_{A_j} X\ d\mu}{\mu(A_j)} \mathbf{1}_{A_j} = \sum_{j=1}^{n} \mathrm{E}_\mu\left(X|A_j\right) \mathbf{1}_{A_j}\ .$$

7 (Conditioning to a random variable)**.** Let $(S, \mathcal{S})$ be a measurable space, $Y\colon \Omega \to S$ a measurable mapping, and $\mathcal{Y} = \sigma(Y) = Y^{-1}(\mathcal{S})$. A real random variable is $\mathcal{Y}$-measurable if, and only if, it is of the form $\phi \circ Y$, where ϕ is a real random variable on $(S, \mathcal{S})$. In this situation, the definition of conditional expectaion is rephrased as follows. A version of the conditional expectation of X given $\sigma(Y)$ is a μ-integrable real random variable of the form $\widehat{\phi}_{\mu,X} \circ Y$ such that for all bounded measurable $\phi\colon S \to \mathbb{R}$ it holds $\mathbb{E}_\mu\left[\phi(Y)\widehat{\phi}_{\mu,X}(Y)\right] = \mathbb{E}_\mu[\phi(Y)X]$. Notice that we could write this in terms of the joint distribution of the random variables X and Y as $\int \phi(y)\widehat{\phi}_{\mu,X}(y)\ \mu_Y(dy) = \int \phi(y)x\ \mu_{X,Y}(dxdy)$. An imprecise, but widely used, notation is $\phi_{\mu,X}(y) = \mathrm{E}_\mu(X|Y=y)$, which is called the *expected value of X, given $Y = y$.*

8 (Special cases)**.** (1) If $X \perp\!\!\!\perp Y$ then $\mathrm{E}_\mu(X|\sigma(Y)) = \mathbb{E}_\mu[X]$. in fact,

$$\int \phi(y)x\ \mu_{X,Y}(dxdy) = \int \phi(y)\left(\int x\ \mu_X(dx)\right)\ \mu_Y(dy)\ .$$

(2) If $X \perp\!\!\!\perp Y$ then $\mathrm{E}_\mu(f(X,Y)|\sigma(Y)) = \int f(x,Y)\ \mu_X(dx)$. In this case we have

$$\int \phi(y)f(x,y)\ \mu_X \otimes \mu_Y(dxdy) = \int \phi(y)\left(\int f(x,y)\ \mu_X(dx)\right)\ \mu_Y(dy)\ .$$

(3) Let X, Y, be random variables in $\mathbb{R}^m$ such that $(X-Y) \perp\!\!\!\perp Y$. Then

$$\mathrm{E}_\mu(f(Y)|\sigma(Y)) = \mathrm{E}_\mu(f((X-Y)+Y)|\sigma(Y)) = \int f(s,Y)\ \mu_{(X-Y)}(ds)\ .$$

Cf. the Gaussian case below.

(4) If $\mu_{X,Y}(dx,dy) = p_{X,Y}\cdot \nu_X \otimes \nu_Y$, then $\mu_Y = \left(\int p(x,y)\ \nu_X(dx)\right)\cdot \nu_Y(dy)$ and the characteristic equality becomes

$$\int \phi(y)\phi_X(y)\ \left(\int p(x,y)\ \nu_X(dx)\right)\cdot \nu_Y(dy) = \int \phi(y)\left(\int x\ p_{X,Y}\ \nu_X(dx)\right)\ \nu_Y(dy)\ ,$$

hence we can take

$$\widehat{\phi}_X(y) = \int x\ p_{X|Y}(x|y)\ \nu_X(dx), \quad p_{X|Y}(x|y) = \frac{p_{X,Y}(x,y)}{p_X(x)}\ .$$

9 (Properties)**.** All random variables are defined on the probability space $(\Omega, \mathcal{F}, \mu)$ and $\mathcal{G}$ is a sub-σ-algebra of $\mathcal{F}$

(1) *Normalization.* $\mathrm{E}_\mu(\mathbf{1}|\mathcal{G}) = \mathbf{1}$.

(2) *$\mathcal{G}$-Linearity.* If $\mathrm{E}_\mu(X|\mathcal{G}) = \widehat{X}$ and $\mathrm{E}_\mu(Y|\mathcal{G}) = \widehat{Y}$, then $\mathrm{E}_\mu(AX+BY|\mathcal{G}) = A\widehat{X} + B\widehat{Y}$ μ-almost-surely if $A, B \in \mathcal{L}^\infty(\mathcal{G})$.

(3) *Positivity.* If $X \geqslant 0$ and $\mathrm{E}_\mu(X|\mathcal{G}) = \widehat{X}$, then $\widehat{X} \geqslant 0$. Linearity and positivity together imply monotonicity. [Hint: take $G = \mathbf{1}_{\{\widehat{X}\leqslant 0\}}$ in the characteristic property]

(4) Normalization, linearity and monotonicity together imply *Jensen inequality.* Assume $\Phi\colon \mathbb{R} \to \mathbb{R}$ and assume both X and $\Phi(X)$ are integrable. Let $x \mapsto a + bx \leqslant \Phi(x)$. Then $a + b\,\mathrm{E}_\mu(X|\mathcal{G}) \leqslant \mathrm{E}_\mu(\Phi(X)|\mathcal{G})$. Chose a version $\widehat{X} = \mathrm{E}_\mu(X|\mathcal{G})$ Because of the convexity, for each $\omega \in \Omega$, there exists coefficients $a(\omega), b(\omega)$ such that $a(\omega) + b(\omega)\widehat{X}(\omega) = \Phi(\widehat{X}(\omega))$. We have shown that $\Phi(\mathrm{E}_\mu(X|\mathcal{G})) \leqslant \mathrm{E}_\mu(\Phi(X)|\mathcal{G})$. In particular, $\mathrm{E}_\mu(|X||\mathcal{G})^\alpha \leqslant \mathrm{E}_\mu(|X|^\alpha|\mathcal{G})$ if $\alpha \geqslant 1$.

(5) *Monotone convergence.* If $0 \leqslant X_n \uparrow X$ and $\widehat{X}_n = \mathrm{E}_\mu(X_n|\mathcal{G})$, $n \in \mathbb{N}$, then random variable $\widehat{X}$ defined by $\widehat{X}_n \uparrow \widehat{X}$ is such that $\mathbb{E}_\mu\left[G\widehat{X}\right] = \mathbb{E}_\mu[GX]$ if $0 \leqslant G \in \mathcal{L}^\infty(\mathcal{G})$. It follows immediatly from the monotone convergence for the expectation [Notice that here we are assuming each X_n to be 'integrable so that the conditional expectation is defined. This is not necessary if we define conditional expectation for non-negative random variable as it was for che expectation. We do not consider this generalization in this notes.] If moreover X happens to be integrable, then $\widehat{X} = \mathrm{E}_\mu(X|\mathcal{G})$.

(6) *Fatou lemma.* If $0 \leqslant X_n$ and $\widehat{X}_n = \mathrm{E}_\mu(X_n|\mathcal{G})$, $n \in \mathbb{N}$, then $\wedge_{m\geqslant n} X_m \leqslant X_m$ if $m \geqslant n$, so that $\mathrm{E}_\mu(\wedge_{m\geqslant n} X_m|\mathcal{G}) \leqslant \wedge_{m\geqslant n} \mathrm{E}_\mu(X_m|\mathcal{G})$. From the monotone convergence it follows $\mathbb{E}_\mu[G(\liminf_{n\to\infty} X_n)] \leqslant \mathbb{E}_\mu[G(\liminf_{n\to\infty} \mathrm{E}_\mu(X_n|\mathcal{G}))]$ if $G \in \mathcal{L}^\infty(\mathcal{G})$ and $G \geqslant 0$. If $\liminf_{n\to\infty} X_n$ is integrable, then we can write $\mathrm{E}_\mu(\liminf_{n\to\infty} X_n|\mathcal{G}) \leqslant \liminf_{n\to\infty} \mathrm{E}_\mu(X_n|\mathcal{G})$.

(7) *Dominated convergence.* If in the fatou lemma we assume that the sequence $(X_n)_{n\in\mathbb{N}}$ is dominated by the integrable random variable Y, by considering the non-negative sequence $(Y - X_n)_{n\in\mathbb{N}}$ we can obtain the inequality

$$\mathrm{E}_\mu\left(\liminf_{n\to\infty} X_n\middle|\mathcal{G}\right) \leqslant \liminf_{n\to\infty} \mathrm{E}_\mu(X_n|\mathcal{G}) \leqslant \limsup_{n\to\infty} \mathrm{E}_\mu(X_n|\mathcal{G}) \leqslant \mathrm{E}_\mu\left(\limsup_{n\to\infty} X_n\middle|\mathcal{G}\right) .$$

If the sequence is convergent, then $\liminf_{n\to\infty} X_n = \lim_{n\to\infty} X_n = \limsup_{n\to\infty} X_n$ hence $\liminf_{n\to\infty} \mathrm{E}_\mu(X_n|\mathcal{G}) = \limsup_{n\to\infty} \mathrm{E}_\mu(X_n|\mathcal{G})$ and the sequence of conditional expectations is convergent to the expectation of the limit. The condition of positivity can be dropped by decomposing the positive and negative part of the sequence and the limit.

2. Conditional distribution

10 (Transition probability measure)**.** Given a product measurable space $(\Omega_1 \times \Omega_2, \mathcal{F}_1 \otimes \mathcal{F}_2)$ a *transition* is a mapping $\mu_{1|2} \colon \mathcal{F}_1 \times \Omega_2$ such that

(1) for each $x_2 \in \Omega_2$ tha mapping $\mathcal{F}_1 \ni A_1 \mapsto \mu_{1|2}(A_1|x_2)$ is a probability measure on $(\Omega_1, \mathcal{F}_1)$ and
(2) for each $A_1 \in \mathcal{F}_1$ the mapping $\Omega_2 \ni x_2 \mapsto \mu_{1|2}(A_1|x_2)$ is $\mathcal{F}_2$-measurable.

11 (Integration of probability measures)**.** Given a transition $\mu_{1|2}$ on $(\Omega_1 \times \Omega_2, \mathcal{F}_1 \otimes \mathcal{F}_2)$ and a probability measure μ_2 on $(\Omega_2, \mathcal{F}_2)$, there exists a unique probability measure $\mu = \int \mu_{1|2}\, d\mu_2$ on the product measurable space such that for each positive or μ-integrable function $f \colon \Omega_2 \times \Omega_2 \ni (x_1, x_2) \mapsto f(x_1, x_2)$ it holds

$$\int f\, d\mu = \int \left(\int f(x_1, x_2)\, \mu_{1|2}(dx_1|x_2) \right) \mu_2(dx_2) .$$

The measure μ is characterised on functions of the form $f(x_1, x_2) = f_1(x_1) f_2(x_2)$ by

$$\int f_1 f_2\, d\mu = \int \left(\int f_1(x_1)\, \mu_{1|2}(dx_1|x_2) \right) f_2(x_2)\, \mu_2(dx_2) .$$

[The proof is a simple variation of the argument for Fubini theorem.]

12 (Transition densities)**.** A simple case occurs when the transition has the form

$$\mu_{1|2}(A_1|x_2) = \int_{A_1} p_{1|2}(x_1|x_2)\, \nu_1(dx), \quad A_1 \in \mathcal{F}_1, x_2 \in \Omega_2$$

where $(x_1, x_2) \mapsto p_{1|2}(x_1|x_2)$ is measurable on the product space $(\Omega_1, \Omega_2, \mathcal{F}_1 \otimes \mathcal{F}_2)$ and $x_1 \mapsto p_{1|2}(x_!|x_2)$ is a ν_1-probability density for each $x_2 \in \Omega_2$. In such a case,

$$\int \left(\int f_1(x_1)\ \mu_{1|2}(dx_1|x_2) \right) f_2(x_2)\ \mu_2(dx_2) = \int \left(\int f_1(x_1) p_{1|2}(x_1|x_2) \nu_1(dx_1) \right) f_2(x_2)\ \mu_2(dx_2) = \iint f_1(x_1) f_2(x_2) p_{1|2}(x_1|x_2)\ \nu_1(dx_1)\mu_2(dx_2)\ ,$$

that is, $\mu = p_{1|2} \cdot \nu_1 \otimes \mu_2$. If moreover the second measure has itself a density, $\mu_2 = p_2 \cdot \nu_2$, then $\mu = (p_{1|2} \otimes p_2) \cdot \nu_1 \otimes \nu_2$

Exercise 7 (Examples).

(1) Let X be a real random variable with positive density p. The conditional distribution of X given $|X|$ is
(2) Let T_1, T_2 be independent and Exp(1). Then the distribution of T_1 given $T_1 + T_2 = t$ is uniform on $]0, t[$.
(3) If $(Y_1, Y_2) \sim \mathrm{N}_{n_1+n_2}(0, \Sigma)$, $\det \Sigma \neq 0$, find the conditional distribution of Y_1 given Y_2.
(4) If Y_1, Y_2 are independent and $\mathrm{N}_1(0, 1)$, find the distribution of (Y_1, Y_2) given $Y_1^2 + Y_2^2$.

13 (Regular version of the conditional expectation)**.** With the notations above, denoting with X_1, X_2 the coordinate projection, the random variable $\widehat{f}(X_2) = \int f(x_1, X_2)\ \mu_{1|2}(dx_1|X_2)$ is a version of the conditional expectation $\mathrm{E}_\mu(f(X_1, X_2)|\sigma(X_2))$, namely a *regular version*. In fact,

$$\mathbb{E}_\mu[f(X_1, X_2)g(X_2)] = \int \left(\int f(x_1, x_2)\ \mu_{1|2}(dx_1|x_2) \right) g(x_2)\ \mu_2(dx_2) = \mathbb{E}_\mu\left[\widehat{f}(X_2) g(X_2) \right]\ .$$

3. Conditioning of jointly Gaussian vectors

Exercise 8. Recall that for each $\Sigma \in \mathrm{Sym}_+(n)$ there exists an orthogonal $U \in \mathrm{O}(n)$ and a non-negative diagonal $\Lambda = \mathrm{diag}(\lambda_1, \dots, \lambda_n)$ such that $\Sigma = U\Lambda U^*$. By discarding the zero eigen-values, we can write $\Sigma = SDS^*$ with $S \in \mathrm{Mat}(n \times r)$, $S^*S = I_r$, and D positive diagonal, where r is the rank of Σ. If $D = \mathrm{diag}(\lambda_1, \dots, \lambda_r)$, we define $D^{-1} = \mathrm{diag}\left(\lambda_1^{-1}, \dots, \lambda_r^{-1}\right)$ and $\Sigma^+ = SD^{-1}S^*$. It follows that

$$\Sigma^+\Sigma = SD^{-1}S^*SDS^* = SS^* \quad \text{and} \quad \Sigma\Sigma^* = SDS^*SD^{-1}S^* = SS^*\ .$$

We have $\Pi = SS^* \in \mathrm{Sym}_+(n)$ and $\Pi^2 = \Pi$. The matrix Π is the orthogonal projector onto the image of Σ. In fact, for all $x \in \mathbb{R}^n$,

$$\Pi x = SS^*x = SDS^*SD^{-1}S^*x = \Sigma SD^{-1}S^*x\ .$$

Moreover, for each $x, y \in \mathbb{R}^n$

$$(x - \Pi x) \cdot (\Sigma y) = (x - \Pi x)^*(\Sigma y) = [(I - SS^*)x]^*(SDS^*y) = x^*(I - SS^*)SDS^*y = x^*(SDS^* - SS^*SDS^*) = 0$$

Proposition 1.

(1) *The Gaussian random vector with components*

$$\widetilde{Y}_1 = Y_1 - (b_1 + L_{12}(Y_2 - b_2)), \quad L_{12} = \Sigma_{12}\Sigma_{22}^{+}$$
$$\widetilde{Y}_2 = Y_2 - b_2$$

is such that $\mathbb{E}\left(\widetilde{Y}_1\right) = 0$, $\mathrm{Var}\left(\widetilde{Y}_1\right) = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{+}\Sigma_{21}$, *and* $\widetilde{Y}_1 \perp\!\!\!\perp \widetilde{Y}_2$. *It follows*

$$\mathrm{E}\left(Y_1|Y_2\right) = b_1 + L_{12}(Y_2 - b_2)$$

(2) *The conditional distribution of* Y_1 *given* $Y_2 = y_2$ *is Gaussian with*

$$Y_1|(Y_2 = y_2) \sim \mathrm{N}_{n_1}\left(b_1 + L_{12}(y_2 - b_2), \Sigma_{11} - L_{12}\Sigma_{21}\right)$$

(3) *The conditional density of* Y_1 *given* $Y_2 = y_2$ *in terms of the partitioned concentration is*

$$p_{Y_1|Y_2}(y_1|y_2) = (2\pi)^{-\frac{n_1}{2}} \det\left(K_{1|2}\right)^{\frac{1}{2}} \times$$
$$\exp\left(-\frac{1}{2}(y_1 - b_1 - K_{11}^{-1}K_{12}(y_2 - b_2))^T K_{11}(y_1 - b_1 - K_{11}^{-1}K_{12}(y_2 - b_2))\right)$$

Proof. (1) We have

$$\begin{bmatrix}\widetilde{Y}_1\\ \widetilde{Y}_2\end{bmatrix} = \begin{bmatrix}I & -\Sigma_{12}\Sigma_{22}^{+}\\ 0 & I\end{bmatrix}\begin{bmatrix}Y_1 - b_1\\ Y_2 - b_2\end{bmatrix} \sim \mathrm{N}_{n_1+n_2}\left(0, \begin{bmatrix}\Sigma_{1|2} & 0\\ 0 & \Sigma_{22}\end{bmatrix}\right)$$

It follows

$$\mathrm{E}\left(Y_1|Y_2\right) = \mathrm{E}\left(\widetilde{Y}_1 + b_1 + L_{12}(Y_2 - b_2)\middle|Y_2\right) = \mathbb{E}\left(\widetilde{Y}_1\right) + b_1 + L_{12}(Y_2 - b_2)$$

(2) The conditional distribution of Y_1 given Y_2 is a transition probability $\mu_{Y_1|Y_2} \colon \mathcal{B}(\mathbb{R}^{n_1}) \times \mathbb{R}^{n_2}$ such that for all bounded $f \colon \mathbb{R}^{n_1}$

$$\mathrm{E}\left(f(Y_1)|Y_2\right) = \int f(y_1)\ \mu_{Y_1|Y_2}(dy_1|Y_2).$$

We have

$$\mathrm{E}\left(f(Y_1)|Y_2\right) = \mathrm{E}\left(f(\widetilde{Y}_1 + \mathrm{E}\left(Y_1|Y_2\right))\middle|Y_2\right) = \int f(x + \mathrm{E}\left(Y_1|Y_2\right))\ \gamma(dx; 0, \Sigma_{1|2})$$

where $\gamma(dx; 0, \Sigma_{1|2})$ is the measure of $\mathrm{N}_{n_1}\left(0, \Sigma_{1|2}\right)$. We obtain the statement by considering the effect on the distribution $\mathrm{N}_{n_1}\left(0, \Sigma_{1|2}\right)$ of the translation $x \mapsto x + (b_1 + L_{12}(y_2 - b_2))$.

(3) A further application of the Schur complement gives

$$\begin{bmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{bmatrix} = \begin{bmatrix}I & \Sigma_{12}\Sigma_{22}^{-1}\\ 0 & I\end{bmatrix}\begin{bmatrix}\Sigma_{1|2} & 0\\ 0 & \Sigma_{22}\end{bmatrix}\begin{bmatrix}I & 0\\ \Sigma_{22}^{-1}\Sigma_{21} & I\end{bmatrix}$$

whose inverse is

$$\begin{aligned}\begin{bmatrix}K_{11} & K_{12}\\ K_{21} & K_{22}\end{bmatrix} &= \begin{bmatrix}I & 0\\ -\Sigma_{22}^{-1}\Sigma_{21} & I\end{bmatrix}\begin{bmatrix}\Sigma_{1|2}^{-1} & 0\\ 0 & \Sigma_{22}^{-1}\end{bmatrix}\begin{bmatrix}I & -\Sigma_{12}\Sigma_{22}^{-1}\\ 0 & I\end{bmatrix}\\ &= \begin{bmatrix}\Sigma_{1|2}^{-1} & 0\\ -\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{1|2}^{-1} & \Sigma_{22}^{-1}\end{bmatrix}\begin{bmatrix}I & -\Sigma_{12}\Sigma_{22}^{-1}\\ 0 & I\end{bmatrix}\\ &= \begin{bmatrix}\Sigma_{1|2}^{-1} & -\Sigma_{1|2}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\\ -\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{1|2}^{-1} & \Sigma_{22}^{-1}\Sigma_{21}\Sigma_{1|2}^{-1}\Sigma_{12}\Sigma_{22}^{-1} + \Sigma_{22}^{-1}\end{bmatrix}\end{aligned}$$

In particular, we have $K_{11} = \Sigma_{1|2}^{-1}$ and $K_{11}^{-1}K_{12} = -\Sigma_{12}\Sigma_{22}^{-1}$, hence

$$Y_1|(Y_2 = y_2) \sim \mathrm{N}_{n_1}\left(b_1 - K^{-1}K_{12}(y_2 - b_2), K_{11}^{-1}\right)$$

so that the exponent of the Gaussian density has the factor

$$(y_1 - b_1 + K_{11}^{-1}K_{12}(y_2 - b_2))^T K_{11}(y_1 - b_1 + K_{11}^{-1}K_{12}(y_2 - b_2))$$

□

References

[1] Claude Dellacherie and Paul-André Meyer, *Probabilities and potential*, North-Holland Mathematics Studies, vol. 29, North-Holland Publishing Co., Amsterdam-New York; North-Holland Publishing Co., Amsterdam-New York, 1978. MR 521810

[2] David Williams, *Probability with martingales*, Cambridge Mathematical Textbooks, Cambridge University Press, Cambridge, 1991.

Collegio Carlo Alberto

E-mail address: giovanni.pistone@carloalberto.org