PROBABILITY 2019: PART 3
GAUSSIAN PROBABILITY MEASURE

GIOVANNI PISTONE

Contents

The present handout covers generalities on independence and conditioning, Central Limit Theorem (IID case), multivariate Gaussian distributions and the relevant matrix theory. A classical reference on Gaussian random variables is [1] (many reprints available). A modern advanced reference for positive definite matrices is [2].

1. Introduction

Exercise 1 (Gaussian distribution). The standard Gaussian distribution is the probability measure ν with density $f_\nu(x) = (2\pi)^{-1/2}\mathrm{e}^{-x^2/2}$. We have $F_\nu(x) = (2\pi)^{-1/2}\int_{-\infty}^{x}\mathrm{e}^{-u^2/2}\,du$ with no closed form expression.

(1) Check that f_ν is indeed a density.
(2) Compute the moments $\nu_n = \int x^n\,\nu(dx)$, $n \in \mathbb{N}$. [Use $p'_\nu(x) = -xp_\nu(x)$.]
(3) Compute the moment generating function $M_\nu(t) = \int \mathrm{e}^{tx}\,\nu(dx)$. Check that $M_\nu^{(n)}(0) = \nu_n$.
(4) Compute the characteristic function $\Phi_\nu(t) = \int \mathrm{e}^{\sqrt{-1}tx}\,\nu(dx)$.
(5) Compute the first two derivatives of the cumulant generating function $\kappa_\nu(t) = \log M_\nu(t)$.
(6) Compute the density of $X = aZ + b$ with $a, b \in \mathbb{R}$ and $Z \sim p_\nu$. These are the general Gaussian random variables.
(7) Compute the density of $Z + b$ with respect to the distribution of Z.
(8) Compute $\delta\psi$ such that

$$\int \phi'(x)\psi(x)\nu(dx) = \int \phi(x)\delta\psi(x)\,\nu(dx)$$

Date: DRAFT April 3, 2019.

for all $\phi, \psi \in C^1$ such that the integrals are well defined.

(9) Compute $H_n = \delta^n 1$, $n \in \mathbb{N}$.

See recap on product measures and independence the slides *Probability 2019: measure Theory* or any textbook.

Exercise 2 (Independent Gaussian random variables).
(1) Show that the Lebesgue measure on $]0,1[^2$ is the product measasure of two Lebesgue measure on $]0,1[$.
(2) Use the previous remark to construct two independent Gaussian random variables.
(3) If Y_1, Y_2 are independent standard Gaussian random variables, compute the distribution of $Y = (Y_1 + Y_2)/\sqrt{2}$.
(4) If Y_1, Y_2, Y_3 are independent standard Gaussian random variables, compute the distribution of $Y = Y_1 + Y_2 + Y_3$.

2. Central Limit Theorem

The Central Limit Theorem CLT is a weak convergence result about the distribution of standardized sums of independent random variables. It is usually stated assuming the existence of an infinite sequence of Independent Identically Distributed IID random variables.

There are many possible statement with variate assumptions. Possibly, the simplest statement is the following: *Let* $(X_n)_{n\in\mathbb{N}}$ *be an IID sequence such that* $\mathbb{E}(X_1) = 0$ *and* $\mathbb{E}(X_1^2) = 1$. *The sequence* $\left(\frac{X_1+\cdots+X_n}{\sqrt{n}}\right)_{n\in\mathbb{N}}$ *converges weakly to the standard Gaussian distribution i.e., for all* $\phi \in C_b$

$$\lim_{n\to\infty} \mathbb{E}\left(\phi\left(\frac{X_1 + \cdots + X_n}{\sqrt{n}}\right)\right) = \frac{1}{\sqrt{2\pi}} \int \phi(z) \mathrm{e}^{-z^2/2}\, dz \ .$$

Exercise 3 (Proof of the CLT).
(1) Show that $C_{\mathrm{b}}^3(\mathbb{R})$ separates points.
(2) If $\phi \in C_{\mathrm{b}}^3$, then the first Taylor approximation is

$$\phi(y) - \phi(x) - \phi'(x)(y-x) = \int_x^y \phi''(t)(y-t)\, dt$$

so that

$$R(x,y) = \phi(y) - \phi(x) - \phi'(x)(y-x) - \frac{1}{2}\phi''(x)(y-x)^2 = \int_x^y (y-t)(\phi''(t) - \phi''(x))\, dt \ .$$

We have the bound

$$|R(x,y)| \leqslant \frac{1}{2} \left\|\phi'' - \phi''(x)\right\|_\infty (y-x)^2 = C_1 |y-x|^2 \ .$$

The second Taylor approximation is

$$R(x,y) = \phi(y) - \phi(x) - \phi'(x)(y-x) - \frac{1}{2}\phi''(x)(y-x)^2 = \frac{1}{2}\int_x^y \phi'''(t)(y-t)^2\, dt$$

and we have the bound

$$|R(x,y)| \leqslant \frac{1}{3!} \left\|\phi'''\right\|_\infty |y-x|^3 = C_2 |y-x|^3 \ .$$

Putting together the two bounds, $|R(x,y)| \leqslant CL(y-x)$ with $C = C_1 \vee C_2$ and $L(z) = |z|^2 \wedge |z|^3$.

(3) From the previous computations,

$$\phi(y+z)-\phi(x+z)=(\phi(y+z)-\phi(z))-(\phi(x+z)-\phi(z))=$$
$$\left(\phi'(z)y+\frac{1}{2}\Phi''(z)y^2+R(z,y+z)\right)-\left(\phi'(z)x+\frac{1}{2}\Phi''(z)x^2+R(z,y+x)\right)=$$
$$\phi'(z)(y-x)+\frac{1}{2}\phi''(z)(y^2-x^2)+(R(z,y+z)-R(x,x+z))\ ,$$

and $|R(z,y+z)-R(x,x+z)|\leqslant C(L(x)+L(y))$.

(4) For each $n\in N$ let $Z_1,\dots,Z_n$ be a independent standard Gaussian random variables and assume $X_1,\dots,X_n,Z_1,\dots,Z_n$ are independent. Write

$$\phi\left(\frac{X_1+\cdots+X_n}{\sqrt{n}}\right)-\phi\left(\frac{Z_1+\cdots+Z_n}{\sqrt{n}}\right)=$$
$$\phi\left(\frac{X_1+\cdots+X_n}{\sqrt{n}}\right)-\phi\left(\frac{Z_1+X_2\cdots+X_n}{\sqrt{n}}\right)+$$
$$\phi\left(\frac{Z_1+X_2+\cdots+X_n}{\sqrt{n}}\right)-\phi\left(\frac{Z_1+Z_2+X_3\cdots+X_n}{\sqrt{n}}\right)+\dots$$

(5) A typical term has expected value bounded as follows:

$$\left|\mathbb{E}\left(\Phi\left(\frac{X_1}{\sqrt{n}}+\cdots+\frac{X_k}{\sqrt{n}}+\frac{Z_{k+1}}{\sqrt{n}}+\cdots+\frac{Z_n}{\sqrt{n}}\right)-\Phi\left(\cdots+\frac{Z_k}{\sqrt{n}}+\cdots\right)\right)\right|\leqslant$$
$$C\,\mathbb{E}\left(L\left(\frac{X_k}{\sqrt{n}}\right)+L\left(\frac{Z_k}{\sqrt{n}}\right)\right)=C\,\mathbb{E}\left(L\left(\frac{X_1}{\sqrt{n}}\right)+L\left(\frac{Z_1}{\sqrt{n}}\right)\right)\ .$$

The sum is bounded by

$$nC\,\mathbb{E}\left(L\left(\frac{X_1}{\sqrt{n}}\right)+L\left(\frac{Z_1}{\sqrt{n}}\right)\right)=nC\,\mathbb{E}\left(\left|\frac{X_1}{\sqrt{n}}\right|^2\wedge\left|\frac{X_1}{\sqrt{n}}\right|^3+\left|\frac{Z_1}{\sqrt{n}}\right|^2\wedge\left|\frac{Z_1}{\sqrt{n}}\right|^3\right)=$$
$$C\,\mathbb{E}\left(\left|\frac{X_1}{\sqrt{n}}\right|^2\wedge\left|\frac{X_1}{\sqrt{n}}\right|^3+\left|\frac{Z_1}{\sqrt{n}}\right|^2\wedge\left|\frac{Z_1}{\sqrt{n}}\right|^3\right)=C\,\mathbb{E}\left(|X_1|^2\wedge\frac{|X_1|^3}{\sqrt{n}}+|Z_1|^2\wedge\frac{|Z_1|^3}{\sqrt{n}}\right)$$

which converges to zero by dominated convergence.

(6) The convergence holds for all $\phi\in C^3_{\mathrm{b}}$. Show that it holds for all $\phi\in C_{\mathrm{b}}$.

3. Standard Gaussian Distribution

3.1. **Recap: Determinant and area.** Let $A=[a_1\cdots a_n]$ be a $n\times n$ generic real matrix identified with the n-tuple of its columns. Consider a mapping $\Delta\colon [a_1\cdots a_n]\mapsto \Delta A$ which is

(1) multi-linear,
(2) alternating (if two columns are equal then the value is zero),
(3) normalized ($\Delta I=1$).

The first and second condition imply for example

$$0=\Delta[(a_1+a_2)\ (a_1+a_2)\cdots]=$$
$$\Delta[a_1\ a_1\cdots]+\Delta[a_1\ a_2\cdots]+\Delta[a_2\ a_1\cdots]+\Delta[a_2\ a_2\cdots]=$$
$$\Delta[a_1\ a_2\cdots]+\Delta[a_2\ a_1\cdots]$$

that is, the exchange of two columns changes the sign of Δ. Conversely, this property inplies the nullity if equal columns.

The operator Δ is characterized by the three conditions above as it is shown by representing each column is the standard basis and $\Delta A = \det(A)$. A matrix such that $\det(A) = 0$ is said to be singular.

Let A, B be non-singular matrices. consider the mapping

$$[b_1 \cdots b_n] \mapsto (\det(A))^{-1} \det(A[b_1 \cdots b_n]) =\mapsto (\det(A))^{-1} \det([Ab_1 \cdots Ab_n])$$

All conditions above are verified hence $\det(AB) = \det(A)\det(B)$. In particular, $\det(A^{-1}) = (\det(A))^{-1}$.

Gauss-Jordan elimination An elementary matrix is a permutation matrix or, a matrix of the form $[ae_1\ e_2 \cdots e_n]$, $a \neq 0$, or the matrix $[(e_1 + e_2)\ e_2 \cdots e_n]$. *Every matrix is the product of elementary matrix.* In fact, every matrix can be reduced to the diagonal form $[e_1 \cdots e_k\ 0]$ by left and right multiplication by elementary matrices. k is the rank of the matrix.

Linear change-of-variables Let $T\colon \mathbb{R}^n \to \mathbb{R}^n$ be linear and invertible. For each Borel set A the set $T^{-1}(A)$ is Borel and the image measure of the Lebesgue measure m is $A \mapsto m(T^{-1}(A)) = T_{\#}m(A)$, so that $\int g(y)\, T_{\#}m(dx) = \int g(T(x))\, dx$. Let us show that $T_{\#}m$ is translation invariant. In fact

$$T_{\#}m(A + y) = m(T^{-1}(A + y)) = m(T^{-1}(A) + T^{-1}y) = m(T^{-1}(A)) = T_{\#}m(A)\ .$$

It follows that $T_{\#}m$ *is proportional to* m, $m(T^{-1}(A)) \propto m(A)$.

Let us show that *the proportionality constant is* $|\det(T)|^{-1}$, that is,

$$\int g(T(x))\ dx = |\det(T)|^{-1} \int g(y)\ dy\ .$$

Let us write the proportionality constant $\Delta(T)$. Note that $m((ST)^{-1}(A)) = m(T^{-1}S^{-1}(A)) = \Delta(T)m(S^{-1}(A)) = \Delta(T)\Delta(S)m(A)$ that is, $\Delta(ST) = \Delta(T)\Delta(S)$. If T is a permutation matrix, then $\Delta(T) = 1 = |\det(T)|^{-1}$; If $T = [\alpha e_1 \cdots e_n]$, then $\Delta(T) = |\alpha|^{-1} = |\det(T)|^{-1}$; If $T = [(e_1 + e_2)\ e_2 \cdots e_n]$ the same result follows. As all matrices are a product of such matrices, the result is proved.

3.2. **Change of variable formula in $\mathbb{R}^d$.** Let $\mathcal{A}, \mathcal{B} \subset \mathbb{R}^d$ be open and ϕ be a diffeomerphism from $\mathcal{A}$ onto $\mathcal{B}$. Let $J\phi\colon \mathcal{A} \to \mathrm{Mat}(d \times d)$ be the Jacobian mapping of ϕ and $J\phi^{-1}\colon \mathcal{B} \to \mathrm{Mat}(d \times d)$ the Jacobian mapping of ϕ^{-1}, so that $J\phi^{-1} = (J\phi \circ \phi^{-1})^{-1}$. For each non-negative $f\colon \mathcal{B} \to \mathbb{R}^n$,

$$\int_{\mathcal{B}} f(\boldsymbol{y})\ d\boldsymbol{y} = \int_{\mathcal{A}} f \circ \phi(\boldsymbol{x})\ \ |\det(J\phi(\boldsymbol{x}))|\ \ d\boldsymbol{x}$$

Exercise 4. $\mathcal{A} =]0, 2\pi[\times]0, +\infty[$, $\mathcal{B} = \mathbb{R}^2_* = \mathbb{R}^2 \setminus \{(x, y) \in \mathbb{R}^2 | x \geqslant 0, y = 0\}$, $\phi(\theta, \rho) = (\rho\cos\theta, \rho\sin\theta)$.

$$J\phi(\theta, \rho) = \begin{bmatrix} -\rho\sin\theta & \cos\theta \\ \rho\cos\theta & \sin\theta \end{bmatrix}, \quad \det(J\phi(\theta, \rho)) = -\rho$$

$$\iint_{\mathbb{R}^2_*} \mathrm{e}^{-(x^2+y^2)/2}\ dxdy = \iint_{]0,2\pi[\times]0,+\infty[} \mathrm{e}^{-(\rho^2\cos^2\theta + \rho^2\sin^2\theta)/2}\ \rho\ d\theta d\rho =$$

$$\iint_{]0,2\pi[\times]0,+\infty[} \mathrm{e}^{-\rho^2/2}\ \rho\ d\theta d\rho = 2\pi$$

1. *(Image of an absolutely continous measure)*Let $(S, \mathcal{F}, \mu)$ be measure space, $p\colon S \to \mathbb{R}_{>0}$ a probability density, $(\mathbb{X}, \mathcal{G})$ a measurable space, $\phi\colon S \to \mathbb{X}$ a measurable function. If ϕ has a measurable inverse, then the image measure is characterised by

$$\int f \; d\phi_{\#}(p \cdot \mu) = \int (f \circ \phi) p \; d\mu = \int (f \circ \phi)(p \circ \phi^{-1} \circ \phi) \; d\mu = \int f p \circ \phi^{-1} \; d\phi_{\#}\mu$$

hence $\phi_{\#}(p \cdot \mu) = (p \circ \phi^{-1}) \cdot \mu$. Eq. (3.2) applied to $f \circ \phi$ and the diffeomorphism ϕ^{-1} gives

$$\int_{\mathcal{B}} f \; d(\phi_{\#}\ell) = \int_{\mathcal{A}} f \circ \phi(\boldsymbol{x}) \; d\boldsymbol{x} = \int_{\mathcal{B}} f \circ \phi \circ \phi^{-1}(\boldsymbol{y}) \; \left|\det\left(J\phi^{-1}(\boldsymbol{y})\right)\right| \; d\boldsymbol{y} =$$
$$\int_{\mathcal{B}} f(\boldsymbol{y}) \; \left|\det\left(J\phi^{-1}(\boldsymbol{y})\right)\right| \; d\boldsymbol{y} = \int_{\mathcal{B}} f(\boldsymbol{y}) \; \left|\det\left(J\phi \circ \phi^{-1}(\boldsymbol{y})\right)\right|^{-1} \; d\boldsymbol{y}$$

This shows that the image of the Lebesgue measure ℓ under a diffeomorphism is

$$\phi_{\#}\ell = \left|\det\left(J\phi^{-1}\right)\right| \cdot \ell = \left|\det\left(J\phi \circ \phi^{-1}\right)\right|^{-1} \cdot \ell$$

Exercise 5. $\mathcal{A} =]0,1[\times]0,1[$, $\mathcal{B} = \mathbb{R}^2_*$, $\phi(u,v) = (\sqrt{-2\log u}\cos(2\pi v), \sqrt{-2\log u}\sin(2\pi v))$,

$$J\phi(u,v) = \begin{bmatrix} -\frac{1}{2}(-2\log u)^{-1/2}\frac{2}{u}\cos(2\pi v) & -2\pi\sqrt{-2\log u}\sin(2\pi v) \\ -\frac{1}{2}(-2\log u)^{-1/2}\frac{2}{u}\sin(2\pi v) & 2\pi\sqrt{-2\log u}\cos(2\pi v) \end{bmatrix},$$

$$\det\left(J\phi(u,v)\right) = -\frac{2\pi}{u}, \quad \det\left(J\phi \circ \phi^{-1}(x,y)\right) = \frac{2\pi}{\mathrm{e}^{(x^2+y^2)/2}} \; .$$

The image of the uniform probability measure on $]0,1[^2$ under ϕ is $(2\pi)^{-1}\mathrm{e}^{-(x^2+y^2)/2} \; dxdy$.

2 (Marginalization)**.** The previous argument does not apply when Φ is not 1-to-1. We will show in the chapter on conditioning that in such a case

$$\Phi_{\#}(p \cdot \mu) = \hat{p} \cdot \Phi_{\#}(\mu)$$

where $\hat{p}$ is the conditional expectation of p with respect to Φ.

However, there are two common and simple cases namely, the finite state space case and the marginalisation. Assume $\mu = \mu_1 \otimes \mu_2$ on $S = S_1 \times S_2$ and consider the marginal projection $\Phi\colon (x_1, x_2) \mapsto x_1$. Then $\Phi^{-1}(A_1) = A_1 \times S_2$ and $\mu(\Phi^{-1}(A_1)) = \mu(A_1 \times S_2) = \mu_1(A_1)$ hence, $\Phi_{\#}(\mu) = \mu_1$. Let p be a density on S with respect to μ. For each positive $f\colon S_1$ we have

$$\int f \; d\Phi_{\#}(p \cdot \mu) = \int f \circ \Phi \; d(p \cdot \mu) = \iint f(x_1) p(x_1, x_2) \; \mu(dx_1, dx_2) =$$
$$\int f(x_1) \left(\int p(x_1, x_2) \; \mu_2(dx_2) \right) \; \mu_1(dx_1)$$

so that

$$\Phi_{\#}(p \cdot \mu) = p_1(x_1) \cdot \mu_1, \quad p_1(x_1) = \int p(x_1, x_2) \; \mu_2(dx_2)$$

For example, if $p(x_1, x_2) = (2\pi)^{-1}\mathrm{e}^{-(x_1^2+x_2^2)/2}$, then

$$\int p(x_1, x_2) \; dx_2 = (2\pi)^{-1/2}\mathrm{e}^{-x_1^2/2} \int (2\pi)^{-1/2}\mathrm{e}^{-x_2^2/2} \; dx_2 = c(2\pi)^{-1/2}\mathrm{e}^{-x_1^2/2}$$

with $c = \int (2\pi)^{-1/2}\mathrm{e}^{-x_2^2/2} \; dx_2 = 1$ as the further integration with respect to dx_1 shows. Notice that the argument applies to all $p(x_1, x_2) = cf(x_1)f(x_2)$.

3. The real random variable Z is *standard Gaussian*, $Z \sim \mathrm{N}_1(0,1)$, if its distribution ν has density

$$\mathbb{R} \ni z \mapsto \gamma(z) = (2\pi)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}z^2\right)$$

with respect to the Lebesgue measure. It is in fact a density, see above the computation of its two-fold product.

Exercise 6. All moments $\mu(n) = \int z^n \gamma(z)\ dz$ exists. As $z\gamma(z) = -\gamma'(z)$, integration by parts produces a recurrent relation for the moments. [Hint: Write $\int z^n \gamma(z)\ dz = \int z^{n-1} z\gamma(z)\ dz = \int z^{n-1}(-\gamma'(z))\ dz$ and perform an integration by parts]

Exercise 7. If $f\colon \mathbb{R} \to \mathbb{R}$ absolutely continuous i.e., $f(z) = f(0) + \int_0^z f'(u)\ du$, with $\int |f'(u)|\, \gamma(u)\ du < +\infty$ then $\int |zf(z)|\, \gamma(z)\ dz < +\infty$. In fact,

$$\int |zf(z)|\, \gamma(z)\ dz = \int \left| z\left(f(0 + \int_0^z f'(u)\ du\right)\right| \gamma(z)\ dz \leqslant$$
$$|f(0)| \int |z|\, \gamma(z)\ dz + \int \left| z \int_0^z f'(u)\ du \right| \gamma(z)\ dz\ .$$

The first term in the RHS equals $\sqrt{2/\pi}\, |f(0)|$, while in the second term we have for $z \geqslant 0$,

$$\left| \int_0^z f'(u)\ du \right| \leqslant \int (0 \leqslant u \leqslant z)\, |f'(u)|\ du\ .$$

We have

$$\int \left| z \int_0^z f'(u)\ du \right| \gamma(z)\ dz \leqslant \int |z| \left(\int (0 \leqslant u \leqslant z)\, |f'(u)|\ du \right) \gamma(z)\ dz =$$
$$\int |f'(u)| \int_u^\infty z\gamma(z)\ dz\ du = \int |f'(u)| \int_u^\infty (-\gamma'(z))\ dz\ du =$$
$$\int |f'(u)|\, \gamma(u)\ du < \infty\ .$$

A similar argument applies to the case $z \leqslant 0$. This implies

$$\int zf(z)\gamma(z)\ dz = \int f(z)(-\gamma'(z))\ dz = \int f'(z)\ \gamma(z) dz\ .$$

Exercise 8. The *Stein operator* is $\delta f(z) = zf(z) - f'(z)$. We have

$$\int f(z)g'(z)\gamma(z)\ dz = \int \delta f(z) g(z) \gamma(z) dz$$

We define the *Hermite polynomials* to be $H_n(z) = \delta^n 1$. For example, $H_1(z) = z$, $H_2(z) = z^2 - 1$, $H_3(z) = z^3 - 3z$. Hermite polynomials are orthogonal with respect to γ,

$$\int H_n(z) H_m(z) \gamma(z)\ dz = 0 \quad \text{if } n > m\ .$$

4. Let $Z \sim \mathrm{N}_1(0,1)$, $Y = b + aZ$, $a, b \in \mathbb{R}$. Then $\mathbb{E}(X) = b$, $\mathbb{E}(X^2) = a^2 + b^2$, $\mathrm{Var}(X) = a^2$. If $a \neq 0$, then $\phi(z) = b + az$ is a diffeomorphism with inverse $\phi^{-1}(x) = a^{-1}(x - b)$, hence the density of X is

$$\gamma(a^{-1}(x-b))\, |a|^{-1} = (2\pi a^2)^{-1/2} \exp\left(\frac{1}{2a^2}(x-b)^2\right)$$

If $a = 0$ then the distribution of $X = b$ is the Dirac measure at b. We say that X is Gaussian with mean b and variance a^2, $X \sim \mathrm{N}_1(b, a^2)$. Viceversa, if $X \sim \mathrm{N}_1(\mu, \sigma^2)$ and $\sigma^2 \neq 1$, then $Z = \sigma^{-1}(X - \mu) \sim \mathrm{N}_1(0, 1)$.

5. The *characteristic function* of a probability measure μ is

$$\widehat{\mu}(t) = \int \mathrm{e}^{itx}\ \mu(dx) = \int \cos(tx)\ \mu(dx) + i \int \sin(tx)\ \mu(dx), \quad i = \sqrt{-1}$$

If two probability measure have the same characteristic function, then they are equal.

Exercise 9. For the standard Gaussian probability measure we have

$$\widehat{\gamma}(t) = \int \cos(tz)\ \gamma(z)dz = \mathrm{e}^{-\frac{t^2}{2}}\ .$$

In fact, by derivation under the integral

$$\frac{d}{dt}\widehat{\gamma}(t) = -\int z \sin(tz)\ \gamma(z)dz = \int \sin(tz)\gamma'(z)\ dz = -t\gamma(t)$$

and $\widehat{\gamma}(0) = 1$. The characteristic function of $X \sim \mathrm{N}_1(\mu, \sigma^2)$ is

$$\mathbb{E}\left(\mathrm{e}^{itX}\right) = \mathbb{E}\left(\mathrm{e}^{it(\mu+\sigma Z)}\right) = \mathrm{e}^{it\mu}\,\mathbb{E}\left(\mathrm{e}^{i(\sigma^t)Z}\right) = \mathrm{e}^{-t\mu+\frac{1}{2}\sigma^2 t^2}$$

Exercise 10. The characteristic function $\widehat{\mu}$ of the probability measure μ on $\mathbb{R}$ is *non-negative definite*. Take $t_1, \ldots, t_n$ in $\mathbb{R}$ with $n = 1, 2, \ldots$. The matrix

$$T = [\widehat{\mu}(t_i - t_j)]_{i,j=1}^n = \left[\int \mathrm{e}^{i(t_i - t_j)x}\ \mu(dx)\right]_{i,j=1}^n$$

is *Hermitian*, that is the transposed matrix is equal to the conjugate matrix equivalently, T is equal to its adjoint T^*. An Hermitian matrix T is non-negative definite if for all complex vector $\boldsymbol{\zeta} \in \mathbb{C}^n$ it holds $\boldsymbol{\zeta}^* T \boldsymbol{\zeta} \geqslant 0$. In our case

$$\boldsymbol{\zeta}^* \left[\int \mathrm{e}^{i(t_i - t_j)x}\ \mu(dx)\right] \boldsymbol{\zeta} = \sum_{i.j=1}^n \int \overline{\zeta_i}\zeta_j \mathrm{e}^{i(t_i - t_j)x}\ \mu(dx) =$$

$$\sum_{i.j=1}^n \int \overline{\zeta_i}\mathrm{e}^{it_i x}\overline{\zeta_j \mathrm{e}^{it_j x}}\ \mu(dx) = \int \left\| \sum_{i=1}^n \overline{\zeta_i}\mathrm{e}^{it_i x} \right\|^2\ \mu(dx) \geqslant 0\ .$$

Exercise 11. let $X \sim \mathrm{N}_1(b, \sigma^2)$ and $f \colon \mathbb{R} \to \mathbb{R}$ continuous and bounded. Show that $\lim_{\sigma \to 0} \mathbb{E}(f(X)) = f(b)$.

Exercise 12. Let X be a real random variable with density p with respect to the Lebesgue measure, and let $Z \sim \mathrm{N}_1(0, 1)$. Assume X and Z are independent i.e., the joint random variable (X, Z) has density $p \otimes \gamma$ with respect to the Lebesgue measure of [2]. Compute the density of $X + Z$. [Hint: make a change of variable $(x, z) \mapsto (x + z, z)$ then marginalize.]

6. The product of absolutely continuous probability measures is

$$(p_1 \cdot \mu_1) \otimes (p_2 \cdot \mu_2) = (p_1 \otimes p_2) \cdot \mu_1 \otimes \mu_2$$

The $\mathbb{R}^d$-valued random variable $Z = (Z_1, \ldots, Z_d)$ is multivariate *standard Gaussian*, $Z \sim \mathrm{N}_n(0_d, I_d)$ if its components are IID $\mathrm{N}_1(0, 1)$. We write $\nu_d = \nu^{\otimes d}$ to denote the d-fold product measure. The distribution $\nu_d = \gamma^{\otimes d}$ of $Z \sim \mathrm{N}_n(0, I)$ has the product density

$$\mathbb{R}^n \ni \boldsymbol{z} \mapsto \gamma(\boldsymbol{z}) = \prod_{j=1}^n \phi(z_j) = (2\pi)^{-\frac{n}{2}} \exp\left(-\frac{1}{2}\left\|\boldsymbol{z}\right\|^2\right)$$

Exercise 13. The *moment generating function* $t \mapsto \mathbb{E}\left(\exp\left(t \cdot Z\right)\right) \in \mathbb{R}_{>}$ is

$$\mathbb{R}^n \ni t \mapsto M_Z(t) = \prod_{j=1}^{n} \exp\left(\frac{1}{2} t_i^2\right) = \exp\left(\frac{1}{2} \left\|t\right\|^2\right)$$

M_Z is everywhere strictly convex and analytic.

Exercise 14. The *characteristic function* $\zeta \mapsto \widehat{\gamma}_n(\zeta) = \mathbb{E}\left(\exp\left(\sqrt{-1}\zeta \cdot Z\right)\right)$ is

$$\mathbb{R}^n \ni \zeta \mapsto \widehat{\gamma}_n(\zeta) = \prod_{j=1}^{2} \exp\left(-\frac{1}{2} \zeta_i^2\right) = \exp\left(-\frac{1}{2} \left\|\zeta\right\|^2\right)$$

$\widehat{\gamma}_n$ is non-negative definite.

4. Recap: Positive Definite Matrices

7. We collect here a few useful properties of matrices. * denotes transposition.

(1) Denote by $\mathrm{Mat}\left(m \times n\right)$ the *vector* space of $m \times n$ real matrices. We have $\mathrm{Mat}\left(m \times 1\right) \leftrightarrow \mathbb{R}^m$. Let $\mathrm{Mat}\left(n \times n\right)$ be the vector space of $n \times n$ real matrices, $\mathrm{GL}\left(n\right)$ the group of invertible matrices, $\mathrm{Sym}\left(n\right)$ the vector space of real symmetric matrices.

(2) Given $A \in \mathrm{Mat}\left(n \times n\right)$, a real eigen-value of A is a real number λ such that $A - \lambda I$ is singular i.e., $\det\left(A - \lambda I\right) = 0$. If λ is an eigen-value of A, $\boldsymbol{u}$ an eigen-vector of A associated to λ if $A\boldsymbol{u} = \lambda \boldsymbol{u}$.

(3) By identifying each matrix $A \in \mathrm{Mat}\left(m \times n\right)$ with its vectorized form $\mathrm{vec}(A) \in \mathbb{R}^{mn}$, the vector space $\mathrm{Mat}\left(m \times n\right)$ is an Euclidean space for the scalar product $\left\langle A, B \right\rangle = \mathrm{vec}(A)^* \,\mathrm{vec}(B) = \mathrm{Tr}\left(AB^*\right)$. The general linear group $\mathrm{GL}\left(n\right)$ is an open subset of $\mathrm{Mat}\left(n \times n\right)$.

(4) A square matrix whose columns form an orthonormal system, $S = \left[\boldsymbol{s}_1 \cdots \boldsymbol{s}_n\right]$, $\boldsymbol{s}_i^* \boldsymbol{s}_j = (i = j)$, has determinant ± 1. The property is characterised by $S^* = S^{-1}$. The set of such matrices is the orthogonal group $\mathrm{O}(n)$.

(5) *Each symmetric matrix $A \in \mathrm{Sym}\left(n\right)$ has n real eigen-values λ_i, $i = 1, \ldots, n$ and correspondingly an orthonormal basis of eigen-vectors $\boldsymbol{u}_i$, $i = 1, \ldots, n$.*

(6) Let $A \in \mathrm{Mat}\left(m \times n\right)$ and let $r > 0$ be its rank i.e., the dimension of the space generated by its columns, equivalently by its rows. There exist matrices $S \in \mathrm{Mat}\left(m \times r\right)$, $T \in \mathrm{Mat}\left(n \times r\right)$, and a positive diagonal $r \times r$ matrix Λ, such that $S^*S = T^*T = I_r$, and $A = S\Lambda^{1/2}T^*$. The matrix SS^* is the orthogonal projection onto image A. In fact image SS^* = image A, $SS^*A = A$, and SS^* is a projection. Similarly, TT^* is the ortogonal projection unto image A^*.

(7) A symmetric matrix $A \in \mathrm{Sym}\left(n\right)$ is positive definite, $A \in \mathrm{Sym}^{+}\left(n\right)$, respectively strictly positive definite, $A \in \mathrm{Sym}^{++}\left(n\right)$, if $\boldsymbol{x} \in \mathbb{R}^n \neq 0$ implies $\boldsymbol{x}' A \boldsymbol{x} \geqslant 0$, respectively > 0. $\mathrm{Sym}^{+}\left(n\right)$ is a closed pointed cone of $\mathrm{Sym}\left(n\right)$, whose interior is $\mathrm{Sym}^{++}\left(n\right)$. A positive definite matrix is strictly positive definite if it is invertible.

(8) A symmetric matrix A is positive definite, respectively strictly positive definite, if, and only if, all eigen-values are non-negative, respectively positive.

(9) A symmetric matrix B is positive definite if, and only if, $A = B'B$ for some $B \in \mathbb{M}_n$. Moreover, $A \in \mathrm{GL}_n$ if, and only if, $B \in \mathrm{GL}_n$.

(10) A symmetric matrix A is positive definite if, and only if $A = B^2$ and B is positive definite. We write $B = A^{\frac{1}{2}}$ and call B the *positive square root* of A.

Exercise 15. If you are not familiar with the previous items, try the following exercise.

Consider the matrices

$$R(\theta) = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} , \quad \theta \in \mathbb{R} .$$

Check that $R(\theta)^* R(\theta) = I$, $\det R(\theta) = 1$, and $R(\theta_1)R(\theta_2) = R(\theta_1 + \theta_2)$. Compute the matrix

$$\Sigma(\theta) = R(\theta) \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} R(\theta)^* , \quad \lambda_1, \lambda_2 \geqslant 0 .$$

Chech that $\det \Sigma(\theta) = \lambda_1 \lambda_2$, $\Sigma(\theta)^* = \Sigma(\theta)$, the eigenvalues of $\Sigma(\theta)$ are λ_1, λ_2, and $\Sigma(\theta)R(\theta) = R(\theta) \operatorname{diag}(\lambda_1, \lambda_2)$. Compute

$$A(\theta) = R(\theta) \begin{bmatrix} \lambda_1^{1/2} & 0 \\ 0 & \lambda_2^{1/2} \end{bmatrix} R(\theta)^* , \quad \lambda_1, \lambda_2 \geqslant 0 .$$

Check that $A(\theta)A(\theta)^* = A(\theta)A(\theta) = \Sigma(\theta)$.

Exercise 16. Let $A \in \mathrm{O}(n)$ and $Z \sim \mathrm{N}_n(0, I)$. Check that $AZ \sim \mathrm{N}_n(0, I)$. let $B \in \mathrm{Mat}(n \times r)$, $r < n$, and assume that the columns are orthonormal. Check that $BZ\,\mathrm{N}_r(0, I)$. [Hint: complete B to an orthogonal matrix by adding columns, $[B|C] \in \mathrm{O}(n)$ and use the marginalization.]

Exercise 17. Let $Z \sim \mathrm{N}_1(0, 1)$, $A = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \in \mathrm{Mat}(2 \times 1)$. Check that AZ has no density with respect to the Lebesgue measure.

Exercise 18. Let $Z \sim \mathrm{N}_2(0, I)$, $A = \begin{bmatrix} 1 & 1 \end{bmatrix} \in \mathrm{Mat}(1 \times 2)$. Compute the density od AZ.

5. General Gaussian Distribution

Proposition 1.

(1) ***Definition*** *Let $Z \sim \mathrm{N}_n(0, I)$, $A \in \mathrm{Mat}(m \times n)$, $b \in \mathbb{R}^m$, $\Sigma = AA^*$. Then $Y = b + AZ$ has a distribution that depends on Σ and b only. The distribution of Y is called* Gaussian with mean b and variance Σ, $\mathrm{N}_m(b, \Sigma)$.

(2) ***Statility*** *If $Y \sim \mathrm{N}_m(b, \Sigma)$, $B \in \mathrm{Mat}(r \times m)$, $c \in \mathbb{R}^r$, then $c + BY \sim \mathrm{N}_r(c + Bb, B\Sigma B^*)$.*

(3) ***Existence*** *Given any non-negative definite $\Sigma \in \mathrm{Sym}^+(n)$ and any vector $b \in \mathbb{R}^n$, the Gaussian distribution $\mathrm{N}_n(b, \Sigma)$ exists.*

(4) ***Density*** *If $\Sigma \in \mathrm{Sym}^{++}(n)$ e.g., $\Sigma \in \mathrm{Sym}^+(n)$ and moreover $\det(\Sigma) \neq 0$, then the Gaussian distribution $\mathrm{N}_m(b, \Sigma)$, has a density with respect to the Lebesgue measure on $\mathbb{R}^n$ given by given by*

$$p_Y(y) = (2\pi)^{-\frac{m}{2}} \det(\Sigma)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(y - b)^T \Sigma^{-1} (y - b) \right) .$$

(5) ***No density*** *If the rank of Σ is $r < m$, then the distribution of $\mathrm{N}_m(b, \Sigma)$ is supported by the image of Σ. In particular it has no density w.r.t. the Lebesgue measure on $\mathbb{R}^n$.*

(6) ***Characteristic function*** *$Y \sim \mathrm{N}_m(b, \Sigma)$ if, and only if, the characteristic function is*

$$\mathbb{R}^m \ni t \mapsto \exp\left(-\frac{1}{2} t^* \Sigma t + i b^* t \right)$$

Proof.

(1) Assume $b_1, b_2 \in \mathbb{R}^m$, $A_i \in \mathrm{Mat}\,(m \times n_i)$, $Y_i = b_i + A_i Z_i$, $Z_i \sim \mathrm{N}_{n_i}\,(0, I)$, $i = 1, 2$. If $b_1 \neq b_2$ then the expected values of Y_1 and Y_2 are different, hence the distribution is different. Assume $b_1 = b_2 = b$, and consider the distribution of $Y_i - b = A_i Z_i$, $i = 1, 2$. We can write $A_i = S_i \Lambda_i^{1/2} T_i^*$, which in turn implies implies $\Sigma = S_i \Lambda S_i^*$, but $\Sigma = S \Lambda S^*$, hence $S_1 = S_2 = S$ and $\Lambda_1 = \Lambda_2 = \Lambda$ (a part the order). We are reduced to the case $Y_i - b = S \Lambda T_i^* Z_i$, $T_i \in \mathrm{Mat}\,(n_i \times r)$ with both with orthonormal columns. The conclusion follows from $T_1^* Z_1 \sim T_2^* Z_2$.

(2) $Y \sim \mathrm{N}_m\,(b, \Sigma)$ means $Y = b + AZ$ with $Z\,\mathrm{N}_n\,(0, I)$ and $AA^* = \Sigma$. It follows

$$c + BY = c + B(b + AZ) = (c + Bb) + (BA)Z \ ,$$

wth $(BA)(BA)^* = BAA^*B^* = B\Sigma B^*$.

(3) Take $Y = b + \Sigma^{1/2} Z$, $Z \sim \mathrm{N}_n\,(0, I)$.

(4) Use the change of variable formula in $Y = b + AZ$ with $A = \Sigma^{1/2}$ to get

$$p_Y(y) = \left|\det\left(A^{-1}\right)\right| p_Z(A^{-1}(y - b)) \ .$$

The express each term with Σ.

(5) use the decomposition $\Sigma = S \Lambda S^*$ and note that some elements on the diagonal of Λ are zero.

(6) The "if" part is a computation, the "only if" part requires the injection property of characteristic function.

□

Exercise 19 (Linear interpolation of the Brownian motion). Let Z_n, $n = 1, 2 \ldots$ be IID $\mathrm{N}_1\,(0, 1)$. Given $0 < \sigma \ll 1$, define recursively the times $t_0 = 0$ and $t_{n+1} = t_n + \sigma^2$. Let $T = \{t_n | n = 0, 1, \ldots\}$. Define recursively $B(0) = 0$, $B(t_{n+1}) = B(t_n) + \sigma Z_n$. As $B(t_n) = \sum_{i=1}^n \sigma Z_i = \sigma \sum_{i=1}^n Z_i$, then $\mathrm{Var}\,(B(t_n)) = \sigma^2 \mathrm{Var}\,(\sum_{i=1}^n Z_i) = n\sigma^2 = t_n$. For each $t \in \mathbb{R}_{>0} \backslash T$, define $B(t)$ by linear interpolation i.e.,

$$B(t) = \frac{t_{n+1} - t}{t_{n+1} - t_n} B(t_n) + \frac{t - t_n}{t_{n+1} - t_n} B(t_{n+1}) \ , \quad t \in [t_n, t_{n+1}] \ .$$

Compute the variance of $B(t)$ and the density of $B(t)$.

6. Independence of Jointly Gaussian Random Variables

Proposition 2. *Consider a partitioned Gaussian vector*

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} \sim \mathrm{N}_{n_1+n_2}\left(\begin{bmatrix} b_1 \\ b_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}\right) \ .$$

Let $r_i = \mathrm{Rank}\,(\Sigma_{ii})$, $\Sigma_{ii} = S_i \Lambda_i S_i^*$ *with* $S_i \in \mathrm{Mat}\,(n_i \times r_i)$, $S_i^* S = I_{r_i}$, *and* $\Lambda_i \in \mathrm{diag}_{++}\,(r_i)$, $i = 1, 2$.

(1) *The blocks* Y_1, Y_2 *are independent,* $Y_1 \perp\!\!\!\perp Y_2$, *if, and only if,* $\Sigma_{12} = 0$, *hence* $\Sigma_{21} = \Sigma_{12}^* = 0$. *More precisely, if, and only if, there exist two independent standard Gaussian* $Z_i \sim \mathrm{N}_{r_i}\,(0, I)$ *and matrices* $A_i \in \mathrm{Mat}\,(n_i \times r_i)$, $i = 1, 2$, *such that*

$$\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} \sim \begin{bmatrix} A_1 & 0 \\ 0 & A_1 \end{bmatrix} \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} \ .$$

(2) *(The following property is sometimes called* Schur complement lemma.*) Write* $\Sigma_{22}^{+} = S_2 \Lambda_2^{-1} S_2^*$. *Then,*

$$\begin{bmatrix} I & -\Sigma_{12}\Sigma_{22}^{+} \\ 0 & I \end{bmatrix} \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \begin{bmatrix} I & 0 \\ -\Sigma_{22}^{+}\Sigma_{21} & I \end{bmatrix} = \begin{bmatrix} \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{+}\Sigma_{21} & 0 \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \begin{bmatrix} I & 0 \\ -\Sigma_{22}^{+}\Sigma_{21} & I \end{bmatrix} = \begin{bmatrix} \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{+}\Sigma_{21} & 0 \\ 0 & \Sigma_{22} \end{bmatrix},$$

hence the last matrix is non-negative definite. The Shur complement of the partitioned covariance matrix Σ *is*

$$\Sigma_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{+}\Sigma_{21} \in \mathrm{Sym}^{+}(n_1) \ .$$

(3) *Assume* $\det(\Sigma) \neq 0$. *Then both* $\det\left(\Sigma_{1|2}\right) \neq 0$ *and* $\det(\Sigma)_{22} \neq 0$. *If we define the partitioned* concentration *to be*

$$K = \Sigma^{-1} = \begin{bmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{bmatrix} ,$$

then $K_{11} = \Sigma_{1|2}^{-1}$ *and* $K_{11}^{-1}K_{12} = -\Sigma_{12}\Sigma_{22}^{-1}$.

Exercise 20. Let $\Sigma \in \mathrm{Sym}^{+}(n)$ and let $r = \mathrm{Rank}(\Sigma)$. We know that $\Sigma = S\Lambda S^*$ with $S \in \mathrm{Mat}(n \times r)$, $S^*S = I_r$, $\lambda \in \mathrm{diag}_{++}(r)$. Let us define $\Sigma^{+} = S\Lambda^{-1}S^*$. Then it follows by simple computation that $\Sigma^{+}\Sigma = \Sigma\Sigma^{+} = SS^*$. Also, $\Sigma\Sigma^{+}\Sigma = \Sigma$ and $\Sigma^{+}\Sigma\Sigma^{+} = \Sigma^{+}$. If $Y \sim \mathrm{N}_n(0, \Sigma)$, then $Y = SS^*Y$. In fact, $Y - SS^*Y = (I - SS^*)Y$ is a Guassian random variable with variance $(I - SS^*)S\Lambda S^*(I - SS^*) = 0$ because $(I - SS^*)S = S - SS^*S = S - S = 0$.

Proof. (1) If the blocks are independent, they are uncorrelated. Conversely, if $\Sigma_{ii} = S_i\Lambda_i S_i^*$, $i = 1, 2$, define $A_i = S_i\Lambda_i^{1/2}$ to get

$$\begin{bmatrix} A_1 & 0 \\ 0 & A_2 \end{bmatrix} \begin{bmatrix} A_1 & 0 \\ 0 & A_2 \end{bmatrix}^* = \Sigma \ .$$

(2) Computations using Ex. 20.

(3) From the computation above we see that the Schur complement is positive definite and that

$$\det\left(\begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}\right) = \det\left(\Sigma_{1|2}\right)\det(\Sigma_{22}) \ .$$

It follows that $\det(\Sigma) \neq 0$ implies both $\det\left(\Sigma_{1|2}\right) \neq 0$ and $\det(\Sigma_{22}) \neq 0$. The condition

$$\begin{bmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{bmatrix} \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix}$$

is equivalent to

$$\begin{aligned} I &= K_{11}\Sigma_{11} + K_{12}\Sigma_{21} \\ 0 &= K_{11}\Sigma_{12} + K_{12}\Sigma_{22} \\ &\vdots \end{aligned}$$

Right-multiply the second equation by Σ_{22}^{-1} and substitute in the first one, to get $K_{11}\Sigma_{1|2} = I$, hence $K_{11}^{-1} = \Sigma_{1|2}$. The other equality follows by left-multiplying the second equation by K_{11}^{-1}.

□

Exercise 21 (Whitening). Let $Y \sim \mathrm{N}_n(b, \Sigma)$. Assume Σ has rank r and decomposition $\Sigma = S\Lambda S^*$, $S^*S = I_r$, $\lambda \in \mathrm{diag}_{++}(r)$. Then $Z = \Lambda^{-1/2}S^*(Y - b)$ ia a white noise, $Z \sim \mathrm{N}_r(O, I)$. Moreover, $b + S\Lambda^{1/2}Z = Y$. In fact,

$$Y - (b + S\Lambda^{1/2}Z) = (Y - b) - S\Lambda^{1/2}\Lambda^{-1/2}S^*(Y - b) = (I - SS^*)(Y - b) = 0 \ .$$

Conditioning is one among the core concepts in reasoning about uncertainty in Probability, in Statistics, in Economics, in Machine Learning. See the textbook by D. Williams [4, Ch. 9] and E. Çınlar [3, Ch. IV].

7. Conditional expectation

Exercise 22. 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 23 (Examples). (1) If $\mathcal{G} = \{\varnothing, \Omega\}$, then $\mathrm{E}_\mu(X|\mathcal{G}) = \mathbb{E}_\mu[X]$.
(2) If $\mathcal{G} = \mathcal{F}$, then $\mathrm{E}_\mu(X|\mathcal{G}) = X$.
(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

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

Exercise 24. 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 $\mathbb{E}\left(\widehat{X}G\right) = \mathbb{E}(XG)$ 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 25. 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 26. 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.)]

8. 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 [4, ¶3.14].]

9 (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

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

10 (Existence)**.** The fact that the previous mapping is actually defined on all of $L^1(\mathcal{F},\mu)$, is discussed in [4, ¶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.

11 (Properties of the conditional expectation)**.** 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.

12 (Image of a density)**.** On the measurable space $(\Omega, \mathcal{F})$, consider the probability 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 μ-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.

13 (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 $L^1(\mathcal{F},\mu)$. In terms of Functional Analysis, one could say that it is the transposed operator of the injection operator $\mathcal{L}^\infty(\mathcal{G}) \to \mathcal{L}^\infty(\mathcal{F})$.

14 (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 be given $Y \in L^2$ and let $X_1,\dots,X_m \in L^2$ be *explanatory variables*. We want a vector $\boldsymbol{\theta} = (\theta_0,\theta_1,\dots,\theta_d) \in \mathbb{R}^{d+1}$ 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 27. Check all detail of the previous paragraph.

15 (Conditional expectation of a real function of a r.v.)**.** 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 $\mathbb{E}_\mu\left[\phi(Y)\widehat{\phi}_{\mu,X}(Y)\right] = \mathbb{E}_\mu\left[\phi(Y)X\right]$. 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 $\phi_{\mu,X}(y) = \mathrm{E}_\mu\left(X|Y=y\right)$, which is called the *expected value of X, given $Y = y$*.

16 (Special cases)**.** (1) If $X \perp\!\!\!\perp Y$ then $\mathrm{E}_\mu\left(X|\sigma(Y)\right) = \mathbb{E}_\mu\left[X\right]$. 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 $\mathrm{E}_\mu\left(f(X,Y)|\sigma(Y)\right) = \int f(x,Y)\ \mu_X(dx)$. 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

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

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

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

8. Conditional distribution

17 (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.

18 (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

$$\int f_1f_2\ d\mu = \int\left(\int f_1(x_1)\ \mu_{1|2}(dx_1|x_2)\right) f_2(x_2)\ \mu_2(dx_2)\ .$$

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

19 (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 28 (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 $Y_1^2 + Y_2^2$.

20 (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,*

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

9. Conditioning of jointly Gaussian vectors

Exercise 29. 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 3.

(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

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

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

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

(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] T. W. Anderson, *An introduction to multivariate statistical analysis*, third ed., Wiley Series in Probability and Statistics, Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, 2003. MR 1990662

[2] Rajendra Bhatia, *Positive definite matrices*, Princeton Series in Applied Mathematics, Princeton University Press, Princeton, NJ, 2007. MR 2284176 (2007k:15005)

[3] Erhan Çınlar, *Probability and stochastics*, Graduate Texts in Mathematics, vol. 261, Springer, New York, 2011.

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

Collegio Carlo Alberto Room 203a
E-mail address: giovanni.pistone@carloalberto.org
URL: https://www.giannidiorestino.it/