TD7 : Gaussian vectors, conditionning

Exercice 1 — *Gaussian vectors.*
Let X be a random vector in $\mathbb{R}^n$. We say that it is a Gaussian vector if for every $t \in \mathbb{R}^n$, the random variable $\langle t, X\rangle \in \mathbb{R}$ has a Gaussian distribution (with possibly null variance).

(1) Recall the parameters, the characteristic function, and (when it exists) the p.d.f. of a Gaussian distribution on $\mathbb{R}$.
The parameters are the mean $\mu \in \mathbb{R}$ and the variance $\sigma^2 \geq 0$. When $\sigma^2 = 0$, the distribution is just the Dirac in μ, and when $\sigma^2 > 0$, it has pdf $f(t) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(t-\mu)^2/(2\sigma^2)}$. In both cases the characteristic function is $\phi(t) = e^{i\mu t - \sigma^2 t^2/2}$.

(2) Show that $t \mapsto \mathbb{E}[\langle t, X\rangle]$ is a linear form, and $(s,t) \mapsto \mathrm{Cov}[\langle s, X\rangle, \langle t, X\rangle]$ is a *positive semi-definite* bilinear form. Let them be represented by $\langle \cdot, m\rangle$ and $\langle \cdot, \Sigma \cdot\rangle$. Give an interpretation of m_i and Σ_{ij} for every $i, j \in \{1, \dots, n\}$.
It is immediate to check that $t \mapsto \mathbb{E}[\langle t, X\rangle]$ is a linear form, and $(s,t) \mapsto \mathrm{Cov}[\langle s, X\rangle, \langle t, X\rangle]$ is a *positive semi-definite* bilinear form. By decomposing on the standard Euclidean basis it turns out that $m_i = \mathbb{E}[X_i]$ and $\Sigma_{i,j} = \mathrm{Cov}(X_i, X_j)$. We call those the mean vector and the covariance matrix of X.

(3) Let X be a Gaussian vector, for every $t \in \mathbb{R}^n$, compute $\mathbb{E}[e^{i\langle t, X\rangle}]$. Briefly explain why the distribution of X is characterized by the parameters m and Σ.
We have that $\langle t, X\rangle$ is a Gaussian of mean $\langle t, m\rangle$ and variance $\langle t, \Sigma t\rangle$. So by taking the characteristic function of $\langle t, X\rangle$ at point 1 we get $\mathbb{E}[e^{i\langle t, X\rangle}] = \exp(i\langle t, m\rangle - \frac{1}{2}\langle t, \Sigma t\rangle)$. So the distribution of X is completely characterized by the parameters m and Σ.

(4) Let X be a Gaussian vector with parameters (m, Σ) and A be a $p \times n$ matrix, show that $AX \in \mathbb{R}^p$ is a Gaussian vector, and compute its parameters.
Compute $\mathbb{E}[e^{i\langle t, Ax\rangle}] = \mathbb{E}[e^{i\langle {}^\intercal At, x\rangle}] = \exp(i\langle {}^\intercal At, m\rangle - \frac{1}{2}\langle {}^\intercal At, \Sigma {}^\intercal At\rangle) = \exp(i\langle t, Am\rangle - \frac{1}{2}\langle t, A\Sigma {}^\intercal At\rangle)$. Gaussianity and identification of the parameters follows.

(5) We say that two processes A and B are uncorrelated when for every index t, s, $Cov(A_t, B_s) = 0$. Let V_1 and V_2 be two subspaces of $\mathbb{R}^n$ and X a Gaussian vector. Show that the σ-algebras $\sigma(\langle t, X\rangle, t \in V_1)$ and $\sigma(\langle t, X\rangle, t \in V_2)$ are independent if and only if $(\langle t, X\rangle)_{t \in V_1}$ and $(\langle s, X\rangle)_{s \in V_2}$ are uncorrelated.
If we have the independence condition, then for $t \in V_1$ and $s \in V_2$, we have $\mathrm{Cov}[\langle t, X\rangle, \langle s, X\rangle] = 0$ by Fubini's theorem (justified since everybody is in L^2). But the converse is also true: Suppose that for every $t \in V_1$ and $s \in V_2$, we have $\mathrm{Cov}[\langle t, X\rangle, \langle s, X\rangle] = 0$. Let $f_1, \dots f_m$ be a finite family in V_1 followed by a finite family in V_2. Set $Y = (\langle f_1, X\rangle, \dots, \langle f_m, X\rangle) = (Y_1, Y_2)$. Then, by computing covariances, we see that the covariance matrix of Y is block-diagonal. This means that we have a product decomposition $\mathbb{E}[e^{i(\langle t_1, Y1\rangle + \langle t_2, Y_2\rangle}] = \mathbb{E}[e^{i\langle t_1, Y_1\rangle}]\,\mathbb{E}[e^{i\langle t_2, Y_2\rangle}]$. By

injectivity of the characteristic distribution, we have identified the distribution of (Y_1, Y_2) as one of an independent couple of two Gaussian vectors. Now because by definition the σ-algebra spanned by a family of variables is generated by the finite subfamilies, we get the independence of the two σ-algebras.

(6) Build two standard Gaussian variables X and Y that are uncorrelated yet not independent (they obviously do not form a Gaussian vector !)

The classic example : set (X, A) to be an independent couple of a standard Gaussian and a Rademacher variable (uniform on $\{\pm 1\}$). Set $Y = AX$. Then Y is not independent of X ($\mathbb{P}(|X| \leq 1, |Y| \leq 1) = \mathbb{P}(|X| \leq 1) \neq \mathbb{P}(|X| \leq 1)^2$). Yet $\mathrm{Cov}(X, Y) = \mathbb{E}[AX^2] = \mathbb{E}[A]\,\mathbb{E}[X^2] = 0 \times 1 = 0$.

(7) Show that the vector $(X_1, \dots X_n)$ with $X_1, \dots, X_n$ independent standard Gaussian variables, is Gaussian. Use it to build a Gaussian vector with arbitrary parameters. Deduce its p.d.f. when it has one.

If $X = (X_1, \dots X_n)$ then we compute $\mathbb{E}[e^{i\langle t, X\rangle}] = e^{-\frac{1}{2}\langle t,t\rangle}$. So it's Gaussian. For m a vector and Σ a semi-definite positive matrix, and consider $Y = m + \Sigma^{1/2}X$. It should have the prescribed parameters. Furthermore, the pdf of X is given by

$$f_X(t) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-t_i^2/2} = \frac{1}{(2\pi)^{n/2}} e^{-|t|^2/2}.$$

Therefore, since

$$(Y \in A) = (X \in \Sigma^{-1/2}(A - m)) = \int_{\Sigma^{-1/2}(A-m)} f_X(t)dt$$

performing the change of variables $s = m + \Sigma^{1/2}t$ we obtain that the pdf of Y is given by,

$$f_Y(t) = \frac{1}{\sqrt{(2\pi)^n \det(\Sigma)}} e^{-\frac{1}{2}(s-m)\cdot\Sigma^{-1}(s-m)}.$$

Exercice 2 — *Conditioning and independence.*
Let $\mathcal{G}$ be a σ-algebra, $X \in \mathcal{G}$ and $Y \perp\!\!\!\perp \mathcal{G}$ be two random variables, and $f : \mathbb{R}^2 \to \mathbb{R}$ such that $f(X, Y) \in L^1$. Compute $\mathbb{E}[f(X, Y) \mid \mathcal{G}]$. Deduce the conditional distribution of $f(X, Y)$ given $\mathcal{G}$.
Set $u(x) = \mathbb{E}[f(x, Y)] = \int f(x, y)d\,\mathbb{P}_Y(y)$. According to Fubini's theorem, $u(x)$ is defined $\mathbb{P}_X$-a.e. Let us check that the almost-surely defined random variable $u(X)$ satisfies the universal property required from the conditional expectation $\mathbb{E}[f(X, Y) \mid \mathcal{G}]$. Let Z be a $\mathcal{G}$-measurable bounded random variable. Then $Zf(X, Y) \in L^1$, and since Y is independent

of (X, Z), which means $\mathbb{P}_{(X,Z,Y)} = \mathbb{P}_{(X,Z)} \otimes \mathbb{P}_Y$. We deduce

$$\begin{aligned}\mathbb{E}[Zf(X,Y)] &= \int z f(x,y) d\,\mathbb{P}_{(X,Z,Y)}(x,z,y) = \int z f(x,y) d(\mathbb{P}_{(X,Z)} \otimes \mathbb{P}_Y)(x,z,y) \\ &= \int z \left(\int f(x,y) d\,\mathbb{P}_Y(y) \right) d\,\mathbb{P}_{(X,Z)}(x,z) \text{ (Fubini)} \\ &= \mathbb{E}[Zu(X)].\end{aligned}$$

This proves the claim.
For the second part, Let $\mu(x,\cdot)$ denote the distribution of $f(x,Y)$. Then for every bounded measurable ϕ,

$$\mathbb{E}[\phi(f(X,Y))|\mathcal{G}] = \mathbb{E}[\phi(f(x,Y))]_{x=X} = \left(\int \phi(u)\mu(x,du) \right)_{x=X} = \int \phi(u)\mu(X,du).$$

This implies that the distribution of $f(X,Y)$ given $\mathcal{G}$ is $\mu(X,\cdot)$.

Exercice 3 — *Gaussian conditional distribution and Bayesian statistics 101.*
Let (X,Y) be a non-degenerate centered Gaussian vector in $\mathbb{R}^2$ with covariance matrix

$$\Sigma = \begin{pmatrix} \sigma_x^2 & \rho \\ \rho & \sigma_y^2 \end{pmatrix}.$$

(1) For every $y \in \mathbb{R}$, compute the conditional distribution of X given $Y = y$.
To do this, we project X on $\sigma(Y)$ to write

$$X = \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(Y)} Y + \left(X - \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(Y)} Y \right),$$

the two terms of this sum being uncorrelated hence independent, as they themselves form a Gaussian vector. Writing Z the second term, we end up with

$$X = \frac{\rho}{\sigma_Y^2} Y + Z,$$

where Z is independent of Y. We deduce $\mathrm{Var}(X) = \frac{\rho^2}{\sigma_Y^4}\mathrm{Var}(Y) + \mathrm{Var}(Z)$, and hence $\mathrm{Var}(Z) = \sigma_X^2 - \frac{\rho^2}{\sigma_Y^2}$. Using the previous exercise, we deduce that the conditional probability kernel of X given Y is

$$(y,\cdot) \mapsto \mathbb{P}\left(\frac{\rho}{\sigma_Y^2} y + Z \in \cdot \right) = \mathcal{N}\left(\frac{\rho}{\sigma_Y^2} y, \sigma_X^2 - \frac{\rho^2}{\sigma_Y^2} \right)(\cdot).$$

(2) Let $\theta \sim \mathcal{N}(0,\tau^2)$ and $Y_1, \dots, Y_n$ i.i.d. $\sim \mathcal{N}(0,\sigma^2)$ random variables, assume that $(\theta, Y_1, \dots, Y_n)$ is a Gaussian vector. Define $X_i = \theta + Y_i$. What is the conditional distribution of θ given $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i = \overline{x}$?
Applying the previous question, we get that the law of θ given $\overline{X} = \overline{x}$ is the normal law of mean $\frac{\overline{x}}{1+\frac{\sigma^2}{n\tau^2}}$ and variance $\frac{1}{\frac{n}{\sigma^2}+\frac{1}{\tau^2}}$.

(3) Give an interpretation of the situation discribed in the previous question.
We may interpret this as follows: a real-world parameter θ must be measured. Prior (theoretical or based on the past) knowledge gives us its *a priori* distribution $\mathcal{N}(0, \tau^2)$. We are also given noisy measurements $X_1, \dots, X_n$ of this parameter, and wonder what the distribution of θ becomes after adding this supplementary information.

(4) Compute the limit of the distribution of θ given $\overline{X} = \overline{x}$ and give an interpretation in each of the following cases.
(a) $\sigma \to +\infty$
The limit as $\sigma \to \infty$ is $\mathcal{N}(0, \tau^2)$. When the observations are very random, they give no information about θ.
(b) $\sigma \to 0$
The limit as $\sigma \to 0$ is $\mathcal{N}(\overline{x}, 0) = \delta_{\overline{x}}$. When the observations are not random, they equal θ almost surely, hence the distribution of θ given the observations is not random.
(c) $\tau \to +\infty$
The limit as $\tau \to \infty$ is $\mathcal{N}(\overline{x}, \sigma^2/n)$. The prior distribution of θ is very random hence contains no information. That is why the conditional distribution given $\overline{X}$ is not biased towards 0 anymore. Note that we recover the point of view of *inferential statistics*: when θ is unknown but deterministic, we indeed have $\theta - \overline{x} \sim \mathcal{N}(0, \sigma^2/n)$.
(d) $\tau \to 0$
The limit as $\tau \to 0$ is $\mathcal{N}(0, 0) = \delta_0$. Indeed since the prior distribution of θ becomes deterministically equal to 0, then the posterior does too.

(5) ($\star$) What about the conditional distribution of θ given $(X_1, \dots X_n)$? It turns out that the conditional distribution of θ given $(X_1, \dots X_n)$ is the same as the one given $\overline{X}$. Indeed if we replay the proof of question 1 and project θ on $\overline{X}$, we get $\theta = \frac{n\tau^2}{n\tau^2+\sigma^2}\overline{X} + Z$, and it turns out that not only $\mathrm{Cov}(\overline{X}, Z) = 0$ but also $\mathrm{Cov}(X_i, Z) = 0$, $1 \leq i \leq n$. Hence we may continue as in question 1.

Exercice 4 — *Limit in distribution of Gaussian vectors.*
Let $(X_n)_{n\geq 0}$ be a sequence of Gaussian variables $(X_n)_{n\geq 0}$. Give a necessary and sufficient condition for convergence in distribution, show that the limit is always Gaussian, and determine its parameters.
Hint: You can use tightness to show that when $(X_n)_{n\geq 0}$ converges in distribution, the sequence $(\mathbb{E}\, X_n)_{n\geq 0}$ is bounded.
We are going to show that a sequence of Gaussian random variables $(X_n)_{n\geq 0}$ converges in distribution if and only if its mean and variance converge. That is, there exist real numbers μ and $\sigma^2 \geq 0$ such that

$$\mathbb{E}[X_n] \to \mu \quad \text{and} \quad \mathrm{Var}(X_n) \to \sigma^2 \quad \text{as } n \to \infty.$$

Furthermore we are going to show that in this case the limit in distribution of (X_n) is a Gaussian random variable with mean μ and variance σ^2. Assume that $\mathbb{E}[X_n]$ and $\mathrm{Var}(X_n)$

respectively converge to μ and σ^2 then the characteristic function φ_n of X_n is given by

$$\varphi_n(t) = \exp\left(it\mu_n - \frac{1}{2}t^2\sigma_n^2\right).$$

The sequence $(\varphi_n)_n$ converges pointwise to

$$\varphi(t) = \lim_{n\to\infty} \varphi_n(t) = \exp\left(it\mu - \frac{1}{2}t^2\sigma^2\right).$$

Observe that φ is the characteristic function of the normal law with mean μ and variance σ^2. By Lévy's continuity theorem, this means that the sequence $(X_n)_{n\geq 1}$ converges in distribution toward the normal law with mean μ and variance σ^2. Conversely assume that the sequence $(X_n)_n$ converges in distribution towards some random variable X with characteristic function ϕ_X. By Lévy's continuity theorem, we have for every $t \in \mathbb{R}$

$$\lim_{n\to+\infty} \exp\left(it\mu_n - \frac{1}{2}t^2\sigma_n^2\right) = \phi_X(t).$$

Taking the modulus, we observe that $\lim_{n\to+\infty} \exp(-t^2\sigma_n^2/2) = |\phi_X(t)|$. Hence, the sequence σ_n^2 converges towards $\sigma^2 = -2\log|\phi_X(t)|$. Let $M = \sup_{n\geq 1}|\mu_n|$ and assume for now that $M < +\infty$. Let $t = \pi/(2M) > 0$, for any limit point μ of (μ_n) we have

$$\exp(it\mu) = \lim_{n\to+\infty} \exp(it\mu_n) = \phi_X(t)/|\phi_X(t)|.$$

Since $t\mu \in (-\pi, \pi)$ this uniquely characterizes μ. Hence, the bounded sequence $(\mu_n)_n$ converges. We now observe that

$$\phi_X(t) = \lim_{n\to+\infty} \exp\left(it\mu_n - \frac{1}{2}t^2\sigma_n^2\right) = \exp\left(it\mu - \frac{1}{2}t^2\sigma^2\right).$$

Thus we have proven that the mean and the variance of X_n converge toward μ and σ^2 and that X is a Gaussian random variable with mean μ and variance σ^2. It remains to show that the sequence $(\mu_n)_n$ is indeed bounded. To do so, we rely on the fact that the sequence $(X_n)_n$ is tight. Indeed recall that since for every $t \geq 0$, we have that

$$\lim_{n\to+\infty} \mathbb{P}(|X_n| \geq t) = \mathbb{P}(|X| \geq t).$$

One can prove that for every $\varepsilon > 0$ there exists $T(\varepsilon) > 0$ such that for every $n \geq 1$, $\mathbb{P}(|X_n| \geq T(\varepsilon)) \leq \varepsilon$ (this an instance of a result known as Prokhorov's theorem). If we assume by contradiction that the sequence $(\mu_n)_n$ is not bounded then for every $k \in \mathbb{N}$ there exists $n_k \geq 1$ such that $|\mu_{n_k}| \geq T(1/k)$. Then,

$$1/k \geq \mathbb{P}(|X_{n_k}| \geq T(1/k)) \geq \min\{\mathbb{P}(X_{n_k} \geq \mu_{n_k}) + \mathbb{P}(X_{n_k} \leq -\mu_{n_k})\} = 1/2,$$

a contradiction.

Exercice 5 — *Borel-Kolmogorov paradox.*
Let P denote a uniform point in the sphere $\mathbb{S}^2$, i.e. for every bounded measurable f,

$$\int f(p)\,\mathbb{P}_P(dp) = \frac{1}{\mathrm{Leb}_3(B_{\mathbb{R}^3}(0,1))}\int_{B_{\mathbb{R}^3}(0,1)} f\left(\frac{p}{|p|}\right)\mathrm{Leb}_3(dp).$$

Denote $\phi_P \in (-\pi/2, \pi/2]$ its latitude and $\theta_P \in (-\pi, \pi]$ its (almost surely defined) longitude.

(1) Compute the joint distribution of (θ_P, ϕ_P). We start by computing the joint distribution of (θ_P, ϕ_P) which we will use throughout.

$$\begin{aligned}
\mathbb{E}[h(\theta_P,\phi_P)] &= \int h(\theta_p,\phi_p)\,\mathbb{P}_P(dp)\\
&= \frac{3}{4\pi}\int_{B_{\mathbb{R}^3}(0,1)} h(\theta_{p/|p|},\phi_{p/|p|})\mathrm{Leb}_3(dp)\\
&= \frac{3}{4\pi}\int_0^1 r^2dr\int_{-\pi}^{\pi}d\theta\int_{-\pi/2}^{\pi/2}\cos(\phi)d\phi\,h(\theta,\phi)\\
&= \int_{-\pi}^{\pi}\frac{d\theta}{2\pi}\int_{-\pi/2}^{\pi/2}\frac{\cos(\phi)d\phi}{2}\,h(\theta,\phi),
\end{aligned}$$

where we applied Lebesgue's change of variable theorem in line 3, setting

$$p = (r\cos(\theta)\cos(\phi), r\sin(\theta)\cos(\phi), r\sin(\phi)),$$

which gives

$$\begin{aligned}
\mathrm{Leb}_3(dp) &= r^2\cos(\phi)drd\theta d\phi\\
\theta_{p/|p|} &= \theta\\
\phi_{p/|p|} &= \phi.
\end{aligned}$$

On the last line, we read that ϕ_P and θ_P are independent, θ has uniform distribution on $[-\pi,\pi]$, while ϕ has density $\cos(\phi)/2$ on $[-\pi/2,\pi/2]$.

(2) Let $\overline{\theta}_P \in [0,\pi)$ denote a representant of θ_P modulo π. Compute the conditional distribution of P given $\overline{\theta}_P$. With a step further in the computation above, we may deduce that $(\overline{\theta}_P, \mathrm{sign}(\theta_P), \phi_P)$ are independent random variables whose respective distributions are : uniform in $[0,\pi]$, uniform in $\{-1,1\}$, and with density $\cos(\phi)/2$. From exercise 2, we deduce that conditional on $\overline{\theta}_P = \theta$, the distribution of P is that of a point of latitude uniformly chosen in $\{\theta, \theta-\pi\}$ and longitude chosen in $[-\pi/2,\pi/2]$ with density $\cos(\phi)/2$.

(3) Compute the conditional distribution of P given ϕ_P. It comes directly from exercise 2 that conditional on $\phi_P = \phi$, the distribution of P is that of a point with latitude ϕ and uniform longitude.

(4) Justify that there is only one "right way" of specializing those answers when computing the conditional distribution of P given $\overline{\theta}_P = 0$ and the conditional distribution of P given $\phi_P = 0$). A conditional probability kernel given some variable Z is only defined up to $\mathbb{P}_Z$-almost-everywhere equality, so it does not really make sense

to specialize it at a given point z. However, if there is a continuous representative (i.e. $z \mapsto \mu(z, \cdot)$ is continuous in the space of probability measures), then it is unique. Hence specialization makes sense. This is the case for the two conditional probability kernels defined above.

(5) What is the paradox ? The paradox is that both procedures yield a probability measure on some great circle of the sphere, that are really different. In one case the measure is the image of the Lebesgue measure in $\mathbb{S}^1$, while in the other case it is not. It comes from the fact that conditioning on negligible events is not well-defined.