MACS 201 : Hilbert spaces and probability

1 Hilbert spaces

Def. Let $\mathcal{H}$ be a complex linear space. An **inner-product** on $\mathcal{H}$ is a function $\langle \cdot \mid \cdot \rangle : \mathcal{H} \times \mathcal{H} \to \mathbf{C}$ which satisfies the following properties :

(i) $\forall (x, y) \in \mathcal{H} \times \mathcal{H}, \langle x \mid y \rangle = \overline{\langle y \mid x \rangle}$,

(ii) $\forall x, y, z \in \mathcal{H}, \forall (\alpha, \beta) \in \mathbf{C} \times \mathbf{C}, \langle \alpha x + \beta y \mid z \rangle = \alpha \langle x \mid z \rangle + \beta \langle y \mid z \rangle$,

(iii) $\forall x \in \mathcal{H}, (\langle x \mid x \rangle = 0) \iff (x = 0)$

Then $\|\cdot\| : x \mapsto \sqrt{\langle x \mid x \rangle} \geqslant 0$ defines a norm on $\mathcal{H}$. Both are continuous.

Th. *For all $x, y \in \mathcal{H}$, we have :*

a) Cauchy-Schwarz inequality : $|\langle x \mid y \rangle| \leqslant \|x\| \cdot \|y\|$,

b) triangular inequality : $|\|x\| - \|y\|| \leqslant \|x - y\| \leqslant \|x\| + \|y\|$,

c) Parallelogram inequality : $\|x + y\|^2 + \|x - y\|^2 = 2\|x\|^2 + 2\|y\|^2$.

Def. An inner-product space $\mathcal{H}$ is called an **Hilbert space** if it is complete.

Prop. *For all measured space $(\Omega, \mathcal{F}, \mu)$, the space $L^2(\Omega, \mathcal{F}, \mu)$ endowed with $\langle f \mid g \rangle = \int f \bar{g} \, \mathrm{d}\mu$ is a Hilbert space.*

Def. Two vectors $x, y \in \mathcal{H}$ are **orthogonal** if $\langle x \mid y \rangle = 0$ which we denoted by $x \perp y$. If $\mathcal{S}$ is a subspace of $\mathcal{H}$, we write $x \perp \mathcal{S}$ if $\forall s \in \mathcal{S}, x \perp s$. Also we write $\mathcal{S} \perp \mathcal{T}$ if all vectors in $\mathcal{S}$ are orthogonal to $\mathcal{T}$.

Not. If $\mathcal{H} = \mathcal{A} + \mathcal{B}$ and $\mathcal{A} \perp \mathcal{B}$ we will denote $\mathcal{H} = \mathcal{A} \overset{\perp}{\oplus} \mathcal{B}$.

Def. Let $\mathcal{E}$ be a subset of an Hilbert space $\mathcal{H}$. The orthogonal set of $\mathcal{E}$ is $\mathcal{E}^\perp = \{x \in \mathcal{H} \mid \forall y \in \mathcal{E}, \langle x \mid y \rangle = 0\}$.

Th. *If $\mathcal{E}$ is a subset of an Hilbert space $\mathcal{H}$, then $\mathcal{E}^\perp$ is closed.*

Orthogonal and orthonormal bases

Def. Let E be a subset of $\mathcal{H}$. It is an orthogonal set if for all $(x, y) \in E \times E, x \neq y, x \perp y$. If moreover $\forall x \in E, \|x\| = 1$, we say that E is orthonormal.

Th. *Let $(e_i)_{i \geqslant 1}$ be an orthonormal sequence of an Hilbert space $\mathcal{H}$ and let $(\alpha_i)_{i \geqslant 1} \in \mathbf{C}^{\mathbf{N}}$. The series $\sum_{i=1}^{\infty} \alpha_i e_i$ converges in $\mathcal{H}$ if and only if $\sum_i |\alpha_i|^2 < \infty$, in which case $\|\sum_{i=1}^{\infty} \alpha_i e_i\|^2 = \sum_{i=1}^{\infty} |\alpha_i|^2$.*

Prop. *Let $x \in \mathcal{H}$ (Hilbert space) and $E = \{e_1, \ldots, e_n\}$ a finite orthonormal set of vectors. Then $\|x - \sum_{k=1}^{n} \langle x \mid e_k \rangle e_k\|^2 = \|x\|^2 - \sum_{k=1}^{n} |\langle x \mid e_k \rangle|^2 = \inf\{\|x - y\|^2, y \in \mathrm{Span}(e_1, \ldots, e_n)\}$.*

Cor (Bessel inequality). *Let $(e_i)_{i \geqslant 1}$ be an orthonormal sequence of a Hilbert space $\mathcal{H}$. Then$\forall, x \in \mathcal{H}, \sum_{i=1}^{\infty} |\langle x \mid e_i \rangle|^2 \leqslant \|x\|^2$.*

Def. A subset E of a Hilbert space $\mathcal{H}$ is said **dense** if $\overline{\mathrm{Span}}(E) = \mathcal{H}$. An orthonormal dense sequence is called a Hilbert basis.

Prop. *Consider the measured space $(\Omega, \mathcal{F}, \mu)$ and the Hilbert space $\mathcal{H} = L^2(\Omega, \mathcal{F}, \mu)$, $\overline{\mathrm{Span}}(\mathbf{1}_A, A \in \mathcal{F}) = \mathcal{H}$.*

Th. *Let $(e_i)_{i \geqslant 1}$ be a Hilbert basis of the Hilbert space $\mathcal{H}$. Then $\forall x \in \mathcal{H}, x = \sum_{i=1}^{\infty} \langle x \mid e_i \rangle e_i$.*

Th. *Let $(e_i)_{i \geqslant 1}$ be an orthonormal sequence of the Hilbert space $\mathcal{H}$. The following assertions are equivalent :*

(i) $(e_i)_{i \geqslant 1}$ is a Hilbert basis,

(ii) if some $x \in \mathcal{H}$ satisfies $\forall i \geqslant 1, \langle x \mid e_i \rangle = 0$ then $x = 0$,

(iii) $\forall x \in \mathcal{H}, \|x\|^2 = \sum_{i=1}^{\infty} |\langle x \mid e_i \rangle|^2$.

Th. *A Hilbert space $\mathcal{H}$ if separable (i.e. contains a countable dense subset) if and only if it admits a Hilbert basis.*

Fourier series

Let $\psi_n \colon x \mapsto \frac{1}{\sqrt{2\pi}} e^{inx}, n \in \mathbf{Z}$. Let $L^1(\mathbf{T})$ denote the set of 2π-periodic locally integrable functions. For $f \in L^1(\mathbf{T})$, set $\forall n \in \mathbf{N}, f_n = \sum_{k=-n}^{n} \left(\int_{\mathbf{T}} f \bar{\phi}_k \right) \phi_k$.

Th. *Let f be a continuous 2π-periodic function. Then the Cesaro sequence $\frac{1}{n} \sum_{k=0}^{n-1} f_k$ converges uniformly to f.*

Cor. *Let μ be a finite measure on the Borel sets of $\mathbf{T} = \mathbf{R}/(2\pi\mathbf{Z})$. The sequence $(\phi_n)_{n \in \mathbf{Z}}$ is dense in the Hilbert space $L^2(\mathbf{T}, \mathcal{B}(\mathbf{T}), \mu)$.*

Cor. *The sequence $(\phi_n)_{n \in \mathbf{Z}}$ is a Hilbert basis in $L^2(\mathbf{T})$. In particular, $\forall f \in L^2(\mathbf{T}), f = \sum_{k=-\infty}^{\infty} \alpha_k \phi_k$ with $\alpha_k = \frac{1}{\sqrt{2\pi}} \int_{\mathbf{T}} f(x) e^{-ikx} \, \mathrm{d}x$ when the infinite sum converges in $L^2(\mathbf{T})$. The Parseval identity then reads $\int_{\mathbf{T}} |f(x)|^2 \, \mathrm{d}x = \sum_{k=-\infty}^{\infty} |\alpha_k|^2$.*

Projection and orthogonality principle

Th (Projection theorem). *Let $\mathcal{E}$ be a closed convex subset of a Hilbert space $\mathcal{H}$ and $x \in \mathcal{H}$. Then the following holds :*

(i) There exists a unique vector $\operatorname{proj}(x \mid \mathcal{E}) \in \mathcal{E}$ such that $\|x - \operatorname{proj}(x \mid \mathcal{E})\| = \inf_{w \in \mathcal{E}} \|x - w\|$.

(ii) If moreover $\mathcal{E}$ is a linear subspace, $\operatorname{proj}(x \mid \mathcal{E})$ is the unique $\hat{x} \in \mathcal{E}$ such that $x - \hat{x} \in \mathcal{E}^\perp$. It is called the orthogonal projection of x onto $\mathcal{E}$.

Prop. *Let $\mathcal{H}$ be a Hilbert space and $\mathcal{E}, \mathcal{E}_1, \mathcal{E}_2$ closed subspaces of $\mathcal{H}$. Then the following assertions hold.*

(i) Suppose that $\mathcal{E} = \overline{\operatorname{Span}}\left((e_k)_{k \in \mathbf{N}}\right)$ with (e_k) being an orthonormal sequence. Then $\operatorname{proj}(h \mid \mathcal{E}) = \sum_{k=0}^{\infty} \langle h \mid e_k \rangle e_k$.

(ii) The function $\operatorname{proj}(\cdot \mid \mathcal{H}) \colon x \mapsto \operatorname{proj}(x \mid \mathcal{E})$ is linear and continuous on $\mathcal{H}$.

(iii) $\|x\|^2 = \|\operatorname{proj}(x \mid \mathcal{E})\|^2 + \|x - \operatorname{proj}(x \mid \mathcal{E})\|^2$

(iv) $(x \in \mathcal{E} \iff \operatorname{proj}(x \mid \mathcal{E}) = x)$ and $\left(x \in \mathcal{E}^\perp \iff \operatorname{proj}(x \mid \mathcal{E}) = 0\right)$

(v) If $\mathcal{E}_1 \subset \mathcal{E}_2$ then $\forall x \in \mathcal{H}, \operatorname{proj}(\operatorname{proj}(x \mid \mathcal{E}_\in \mid \mathcal{E}_1) = \operatorname{proj}(x \mid \mathcal{E}_1)$

(vi) If $\mathcal{E}_1 \perp \mathcal{E}_2$ then $\forall x \in \mathcal{H}, \operatorname{proj}\left(x \mid \mathcal{E}_1 \overset{\perp}{\oplus} \mathcal{E}_2\right) = \operatorname{proj}(x \mid \mathcal{E}_1) + \operatorname{proj}(x \mid \mathcal{E}_2)$

Th. *Let $(M_n)_{n \in \mathbf{Z}}$ be an increasing sequence of closed subspaces of an Hilbert space $\mathcal{H}$.*

1. *Denote $M_{-\infty} = \bigcap_n M_n$. Then $\forall h \in \mathcal{H}, \operatorname{proj}(h \mid M_{-\infty}) = \lim_{n \to -\infty} \operatorname{proj}(h \mid M_n)$.*
2. *Denote $M_\infty = \overline{\bigcup_n M_n}$. Then $\forall h \in \mathcal{H}, \operatorname{proj}(h \mid M_\infty) = \lim_{n \to \infty} \operatorname{proj}(h \mid M_n)$.*

Prop. *Let $\mathcal{E}$ and $\mathcal{F}$ be two subspaces of a Hilbert space $\mathcal{H}$. If $\mathcal{E} \overset{\perp}{\oplus} \mathcal{F} = \mathcal{H}$, then $\mathcal{F} = \mathcal{E}^\perp$.*

Th. *If $\mathcal{E}$ is a closed subspace of a Hilbert space $\mathcal{H}$ then $\mathcal{E} \overset{\perp}{\oplus} \mathcal{E}^\perp = \mathcal{H}$. Moreover $\left(E^\perp\right)^\perp = \mathcal{E}$.*

Th (Riesz representation theorem). *Let $\mathcal{H}$ be a Hilbert space. Then $F \colon \mathcal{H} \to \mathbf{C}$ is a non-zero continuous linear form if and only if $\exists x \in \mathcal{H} \setminus \{0\}, \forall y \in \mathcal{H}, F(y) = \langle y \mid x \rangle$.*

Unitary Operator

Def. Let $\mathcal{H}$ and $\mathcal{I}$ be two Hilbert spaces. An **isometric** operator $S \colon \mathcal{H} \to \mathcal{I}$ is a linear application such that $\forall (v, w) \in \mathcal{H}^2, \langle Sv \mid Sw \rangle_{\mathcal{I}} = \langle v \mid w \rangle_{\mathcal{H}}$. If it is moreover bijective, it is a **unitary** operator. In this case we also says that $\mathcal{H}$ and $\mathcal{I}$ are isomorphic.

Th. *Let $\mathcal{H}$ be a separable Hilbert space.*

(i) If $\mathcal{H}$ has infinite dimension, it is isomorphic to l^2.

(ii) If $\mathcal{H}$ has dimension n, it is isomorphic to $\mathbf{C}^n$.

Th. *Let $\mathcal{H}$ and $\mathcal{I}$ be two Hilbert spaces and $\mathcal{G}$ a subspace of $\mathcal{H}$.*

(i) Let $S \colon \mathcal{G} \to \mathcal{I}$ be isometric on $\mathcal{G}$. Then S admits a unique isometric extension $\bar{S} \colon \bar{\mathcal{G}} \to \mathcal{I}$ and $\bar{S}\left(\bar{\mathcal{G}}\right)$ is the closure of $S(\mathcal{G})$ in $\mathcal{I}$.

(ii) Let $(v_t)_{t \in T}$ and $(w_t)_{t \in T}$ be two set of vectors in $\mathcal{H}$ and $\mathcal{I}$ indexed by an arbitrary index set T. Suppose $\forall (s, t) \in T^2, \langle v_t \mid v_s \rangle_{\mathcal{H}} = \langle w_t \mid w_s \rangle_{\mathcal{I}}$. Then, there exists a unique isometric operator $S \colon \overline{\operatorname{Span}}\left((v_t)_{t \in T}\right) \to \overline{\operatorname{Span}}\left((w_t)_{t \in T}\right)$ such that $\forall t \in T, S v_t = w_t$. Moreover, $S\left(\overline{\operatorname{Span}}\left((v_t)_{t \in T}\right)\right) = \overline{\operatorname{Span}}\left((w_t)_{t \in T}\right)$.

2 Probability

Let $(\Omega, \mathcal{F}, \mathbf{P})$ be a probability space.

Th (π - λ theorem). *If $\mathcal{A} \subset \mathcal{C}$ with $\mathcal{A}$ a π-system and $\mathcal{C}$ a λ-system, then $\sigma(\mathcal{A}) = \mathcal{C}$.*

Th (Characterization of probability measures). *Let $\mathcal{C}$ be a π-system on Ω and $\mathcal{F} = \sigma(\mathcal{C})$ the smallest σ-field containing $\mathcal{C}$. Then a probability measure μ on $(\Omega, \mathcal{F})$ is uniquely characterized by $\mu(A)$ on $A \in \mathcal{C}$.*

Not. For $p > 0$, we denote by $\mathcal{L}^p(\Omega, \mathcal{F}, \mathbf{P})$ the space of random variables X such that $\mathbf{E}\left(|X|^p\right) < \infty$ and by $L^p(\Omega, \mathcal{F}, \mathbf{P})$ the one identifying random variables that are equal $\mathbf{P}$-a.s.

Conditional calculus

Lem. *Let $X \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbf{P})$ and $\mathcal{G}$ a sub-σ-field of $\mathcal{F}$. Then there exists $Y \in \mathcal{L}^1(\Omega, \mathcal{G}, \mathbf{P})$ such that*

$$\forall A \in \mathcal{G}, \mathbf{E}(X \mathbf{1}_A) = \mathbf{E}(Y \mathbf{1}_A) \tag{1}$$

Moreover the following assertions hold.

(i) If $Y' \in \mathcal{L}^1(\Omega, \mathcal{G}, \mathbf{P})$ also satisfies (1) then $Y' = Y$ $\mathbf{P}$-a.s.

(ii) If $X \in \mathcal{L}^2(\Omega, \mathcal{F}, \mathbf{P})$, then $Y = \operatorname{proj}(X \mid L^2(\Omega, \mathcal{G}, \mathbf{P}))$.

(iii) (1) continues to hold extended as $\mathbf{E}(XZ) = \mathbf{E}(YZ)$ for all $\mathcal{G}$-measurable r.v. Z such that $\mathbf{E}(|XZ|) < \infty$.

Def. Let $X \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbf{P})$ and $\mathcal{G}$ a sub-σ-field of $\mathcal{F}$. The unique $Y \in L^1(\Omega, \mathcal{G}, \mathbf{P})$ defined by (1) is called the **conditional expectation** of X given $\mathcal{G}$, and denoted by $Y = \mathbf{E}(x \mid \mathcal{G})$.

Prop. *Suppose that $X, Y, Z, (X_n)_{n\geqslant 1} \in L^1(\Omega, \mathcal{F}, \mathbf{P})$. The following hold $\mathbf{P}$-a.s.*

(i) *(linearity) $\forall a, b \in \mathbf{R}, \mathbf{E}(aX + bY \mid \mathcal{G}) = a\mathbf{E}(X \mid \mathcal{G}) + b\mathbf{E}(Y \mid \mathcal{G})$*

(ii) *If X is $\mathcal{G}$-measurable, $\mathbf{E}(X \mid \mathcal{G}) = X$*

(iii) *If $\mathcal{G} = \{\varnothing, \Omega\}$ is the trivial σ-field, then $\mathbf{E}(X \mid \mathcal{G}) = \mathbf{E}(X)$*

(iv) *If X is independent of $\mathcal{G}$ then $\mathbf{E}(X \mid \mathcal{G}) = \mathbf{E}(X)$*

(v) *(positivity) If $X \leqslant Y$ then $\mathbf{E}(X \mid \mathcal{G}) \leqslant \mathbf{E}(Y \mid \mathcal{G})$*

(vi) *$\mathbf{E}(X \mid \mathcal{G}) \vee \mathbf{E}(Y \mid \mathcal{G}) \leqslant \mathbf{E}(X \vee Y \mid \mathcal{G})$, $\mathbf{E}(X \mid \mathcal{G})_+ \leqslant \mathbf{E}(X_+ \mid \mathcal{G})$ and $|\mathbf{E}(X \mid \mathcal{G})| \leqslant \mathbf{E}(|X| \mid \mathcal{G})$*

(vii) *(tower property) If $\mathcal{H}$ is a sub-σ-field of $\mathcal{F}$ such that $\mathcal{G} \subset \mathcal{H}$ then $\mathbf{E}(\mathbf{E}(X \mid \mathcal{H}) \mid \mathcal{G}) = \mathbf{E}(X \mid \mathcal{G})$*

(viii) *The expectation is not modified by conditional expectation : $\mathbf{E}(\mathbf{E}(X \mid \mathcal{G})) = \mathbf{E}(X)$*

(ix) *If X is $\mathcal{G}$-measurable and $XY \in L^1(\Omega, \mathcal{F}, \mathbf{P})$, then $\mathbf{E}(XY \mid \mathcal{G}) = X \cdot \mathbf{E}(Y \mid \mathcal{G})$*

Def. Let Y be a r.v. and $\sigma(X)$ the sub-σ-field generated by a r.v. X. If $\mathbf{E}(Y \mid \sigma(X))$ is well-defined, it is written as $\mathbf{E}(Y \mid X)$ and is called the **conditional expectation** of Y given X.

Def. Let $\mathcal{G}$ be a sub-σ-field of $\mathcal{F}$. For any event $A \in \mathcal{F}$, we denote $\mathbf{P}(A \mid \mathcal{G}) = \mathbf{E}(\mathbf{1}_A \mid \mathcal{G})$. The mapping $A \mapsto \mathbf{P}(A \mid \mathcal{G})$ is called a **version of the conditional probability** of A given $\mathcal{G}$.

Def. Let $\mathcal{G}$ be a sub-σ-field of $\mathcal{F}$. A **regular version** of the conditional probability of $\mathbf{P}$ given $\mathcal{G}$ is a function $\mathbf{P}^{\mathcal{G}} \colon \Omega \times \mathcal{F} \to [0\,;1]$ such that

(i) For all $A \in \mathcal{F}$, $\mathbf{P}^{\mathcal{G}}(A) \colon \omega \mapsto \mathbf{P}^{\mathcal{G}}(\omega, A)$ is $\mathcal{G}$-measurable and is a version of the conditional probability of A given $\mathcal{G}$, $\mathbf{P}^{\mathcal{G}}(A) = \mathbf{P}(A \mid \mathcal{G})$.

(ii) For all $\omega \in \Omega$, the mapping $A \mapsto \mathbf{P}^{\mathcal{G}}(\omega, A)$ is a probability on $\mathcal{F}$.

Lem. *Let $\mathbf{P}^{\mathcal{G}}$ be a regular version of the conditonal probability of $\mathbf{P}$ given $\mathcal{G}$ and let $Y \in L^1(\Omega, \mathcal{F}, \mathbf{P})$. Then $\mathbf{E}(Y \mid \mathcal{G}) = \mathbf{E}^{\mathcal{G}}(Y)$ $\mathbf{P}$-a.s., with $\mathbf{E}^{\mathcal{G}}(Y) \colon \omega \mapsto \int Y(\omega')\mathbf{P}^{\mathcal{G}}(\omega, \mathrm{d}\omega')$.*

Def. Let $\mathcal{G}$ be a sub-σ-field of $\mathcal{F}$. Let $(\mathsf{Y}, \mathcal{Y})$ be a measurable space and let Y be an Y-valued random variable. A regular version of the conditional distribution of Y given $\mathcal{G}$ is a function $\mathbf{P}^{Y|\mathcal{G}} \colon \Omega \times \mathcal{Y} \to [0\,;1]$ such that

(i) For all $A \in \mathcal{Y}$, $\omega \mapsto \mathbf{P}^{Y|\mathcal{G}}(\omega, A)$ is $\mathcal{G}$ measurable and is a version of conditional distribution of Y given $\mathcal{G}$, $\mathbf{P}^{Y|\mathcal{G}}(\cdot, A) = \mathbf{P}(Y \in A \mid \mathcal{G})$ $\mathbf{P}$-a.s.

(ii) For every ω, $A \mapsto \mathbf{P}^{Y|\mathcal{G}}(\omega, A)$ is a probability on $\mathcal{Y}$.

Def. Let $(\mathsf{X}, \mathcal{X})$ and $(\mathsf{Y}, \mathcal{Y})$ be two measurable spaces. A **kernel** is a mapping $Q \colon \mathsf{X} \times \mathcal{Y} \to [0\,;\infty]$ satisfying the following conditions :

(i) for every $A \in \mathcal{Y}$, the mapping $Q(\cdot, A) \colon x \mapsto Q(x, A)$ is a measurable function,

(ii) for every $x \in \mathsf{X}$, the mapping $Q(x, \cdot) \colon A \mapsto Q(x, A)$ is a measure on $\mathcal{Y}$.

Q is said to be finite if $\forall x \in \mathsf{X}, Q(x, \mathsf{Y}) < \infty$. It is called a probability kernel if $\forall x \in \mathsf{X}, Q(x, \mathsf{Y}) = 1$. It is called a Markov kernel if it is a probability kernel on $\mathsf{X} \times \mathcal{X}$.

Def. Let X and Y be random variables with values in the measure spaces $(\mathsf{X}, \mathcal{X})$ and $(\mathsf{Y}, \mathcal{Y})$ respectively. A **regular version of the conditional distribution of Y given X** is a probability kernel $\mathbf{P}^{X|Y} \colon \mathsf{X} \times \mathcal{Y} \to [0\,;1]$ such that $\forall A \in \mathcal{Y}, \mathbf{P}^{Y|X}(X, A) = \mathbf{P}(Y \in A \mid X)$ $\mathbf{P}$-a.s.

Th. *Let $\mathcal{G}$ be sub-σ-field of $\mathcal{F}$. Let $d \geqslant 1$ and Y be an $(\mathbf{R}^D, \mathcal{B}(\mathbf{R}^d))$-valued random variable. Then, there exists a regular version of the conditional distribution of Y given $\mathcal{G}$, $\mathbf{P}^{Y|\mathcal{G}}$, and this version is unique in the sense that for any other regular version $\bar{\mathbf{P}}^{Y|\mathcal{G}}$ of this distribution, for $\mathbf{P}$-almost every ω it holds that $\forall F \in \mathcal{F}, \mathbf{P}^{Y|\mathcal{G}}(\omega, F) = \bar{\mathbf{P}}^{Y|\mathcal{G}}(\omega, F)$. Moreover, if $\mathcal{G} = \sigma(X)$ for some r.v. X with values in a measurable space $(\mathsf{X}, \mathcal{X})$, there also exists a unique regular version (hence a probability kernel) $\mathbf{P}^{Y|X}$.*

Lem. *Let $\mathbf{P}^{Y|X}$ bee a regular version of the conditional expectation of Y given X. Then, for any real-valued measurable function g on Y such that $\mathbf{E}(|g(Y|) < \infty$, we have $\mathbf{E}(g(Y) \mid X) = \int g(Y)\mathbf{P}^{Y|X}(X, \mathrm{d}y)$, $\mathbf{P}$-a.s.*

Prop. *Let $\mathbf{X}$ and $\mathbf{Y}$ be two jointly Gaussian vectors, respectively valued in $\mathbf{R}^p$ and $\mathbf{R}^q$. Then the following holds.*

(i) *If $\operatorname{Cov}(\mathbf{Y})$ is invertible, then $\hat{\mathbf{X}} := \operatorname{proj}(\mathbf{X} \mid \operatorname{Span}(1, \mathbf{Y}))$ is given by $\hat{\mathbf{X}} = \mathbf{E}(\mathbf{X}) + \operatorname{Cov}(\mathbf{X}, \mathbf{Y}) \operatorname{Cov}(\mathbf{Y})^{-1}(\mathbf{Y} - \mathbf{E}(\mathbf{Y}))$, and $\operatorname{Cov}(\mathbf{X} - \hat{\mathbf{X}}) = \operatorname{Cov}(\mathbf{X}) - \operatorname{Cov}(\mathbf{X}, \mathbf{Y}) \operatorname{Cov}(\mathbf{Y})^{-1} \operatorname{Cov}(\mathbf{Y}, \mathbf{X})$.*

(ii) *We have $\mathbf{E}(\mathbf{X} \mid \mathbf{Y}) = \operatorname{proj}(\mathbf{X} \mid \operatorname{Span}(1, \mathbf{Y}))$.*

(iii) *Let $\hat{\mathbf{X}} = \mathbf{E}(\mathbf{X} \mid \mathbf{Y})$. Then $\operatorname{Cov}(\mathbf{X} - \hat{\mathbf{X}}) = \mathbf{E}\left(\mathbf{X}(\mathbf{X} - \hat{\mathbf{X}})^\intercal\right) = \mathbf{E}\left((\mathbf{X} - \hat{\mathbf{X}})\mathbf{X}^\intercal\right)$ and $\mathbf{P}^{\mathbf{Y}|\mathbf{X}}(\mathbf{X}, \cdot) = \mathcal{N}\left(\hat{\mathbf{X}}, \operatorname{Cov}\left(\mathbf{X} - \hat{\mathbf{X}}\right)\right)$.*

Radon-Nikodym derivative

Def. If $\forall A \in \mathcal{F}, \mu(A) = \int_A \phi \, \mathrm{d}\lambda$, we say that the λ-a.e. equivalent class of ϕ is the **Radon-Nikodym derivative** of μ with respect to λ, and write $\phi = \frac{\mathrm{d}\mu}{\mathrm{d}\lambda}$.

Def. Let λ be a measure on $(\Omega, \mathcal{F})$. We say that a σ-finite measure μ is **absolutely continuous** with respect to λ or that λ dominates μ and we write $\mu \ll \lambda$ if $\forall A \in \mathcal{F}, (\lambda(A) = 0) \implies (\mu(A) = 0)$.

Th (Radon-Nikodym theorem). *Let $\lambda, \mu \in \mathbf{M}_+(\Omega, \mathcal{F})$ be σ-finite measures such that $\mu \ll \lambda$. Then, there exists a non-negative Borel function ϕ such that $\forall A \in \mathcal{F}, \mu(A) = \int_A \phi \, \mathrm{d}\lambda$.*

Def. Let (X, Y) be two random elements admitting a density f with respect to measure $\xi \otimes \xi'$ on $(\mathsf{X} \times \mathsf{Y}, \mathcal{X} \otimes \mathcal{Y})$. Then the function $(x, y) \mapsto f(y \mid x) = \frac{f(x,y)}{\int f(x,y') \, \mathrm{d}\xi'(y')}$ is called the **conditional density** of Y given X.

Th. *Let (X, Y) be two random elements admitting a density $f : \mathsf{X} \times \mathsf{Y} \to \mathbf{R}_+$ with respect to $\xi \otimes \xi'$ on $(\mathsf{X} \times \mathsf{Y}, \mathcal{X} \otimes \mathcal{Y})$. Then, $\forall x \in \mathsf{X}, \forall A \in \mathcal{Y}, \mathbf{P}^{Y|X}(x, A) = \int_A f(y \mid x) \xi'(\mathrm{d}y)$.*

Lem. *Let P and Q be two probabilities on the measurable space $(\Omega, \mathcal{F})$ and let $\nu \in \mathbf{M}_+(\Omega, \mathcal{F})$ dominate both P and Q (e.g. $\nu = P + Q$). Let f_P and f_Q denote the densities of P and Q with respect to ν. Then, $\mathrm{KL}(P \| Q) = \int \ln\left(\frac{f_P}{f_Q}\right) \mathrm{d}P$ is always well defined and takes values in $[0\,;\infty]$. Moreover we have :*

(i) If Q does not dominate P then $\mathrm{KL}(P \| Q) = \infty$.

(ii) If $P \ll Q$ then $\mathrm{KL}(P \| Q) = \int \ln\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right) \mathrm{d}P$ (may be finite or infinite).

(iii) We have $\mathrm{KL}(P \| Q) = 0 \iff P = Q$.

Def. The quantity $\mathrm{KL}(P \| Q)$ is called the **Kullback-Leibler divergence** between P and Q.

Th. *Let P and Q be two probabilities on the measurable space $(\Omega, \mathcal{F})$ and X a measurable mapping from $(\Omega, \mathcal{F})$ to $(\mathsf{X}, \mathcal{X})$. Then we have $\mathrm{KL}\left(P^X \| Q^X\right) \leqslant \mathrm{KL}(P \| Q)$.*

Rem. Recall that $\forall A \in \mathcal{X}, P^X(A) = \int_{X^{-1}(A)} \mathrm{d}P$ while $\forall F \in \mathcal{F}, P(F) = \int_F \mathrm{d}P$.

3 Mathematical statistics

Statistical modeling

Def. Let $(\Omega, \mathcal{F})$ be a measurable space and $\mathcal{P}$ a collection of probabilities on this space. Let X be a measurable function from $(\Omega, \mathcal{F})$ to the observation space $(\mathsf{X}, \mathcal{X})$. We say that $\mathcal{P}$ is a **statistical model** for the observation variable X and denote $\mathcal{P}^X = \left(P^X\right)_{P \in \mathcal{P}}$ the corresponding collection of probability distributions.

It is usual in statistics to consider $\Omega = \mathsf{X}$, $\mathcal{F} = \mathcal{X}$ and $X(\omega) = \omega$, in which case $\forall P \in \mathcal{P}, P = P^X$.

Def. Let $\nu \in \mathbf{M}_+(\mathsf{X}, \mathcal{X})$ and $\mathcal{P}$ be a statistical model for X. We say that $\mathcal{P}$ is a ν-dominated model for X, or that $\mathcal{P}^X$ is ν-dominated, if $\forall P \in \mathcal{P}, P^X \ll \nu$.

Lem (Halmos and Savage). *Let $\nu \in \mathbf{M}_+(\mathsf{X}, \mathcal{X})$. Consider a ν-dominated model $\mathcal{P}$ for the variable X. Then there exists a countable collection $(P_n)_{n \geqslant 1}$ in $\mathcal{P}$ such that $\mathcal{P}^X$ is also dominated by $\mu = \sum_{n \geqslant 1} 2^{-n} P_n^X$.*

Def. Let $\mathcal{P}$ be a statistical model for the observation variable X. We say that $\mathcal{P}$ is a **parametric model** for X if there exists a finite dimensional set Θ such that $\mathcal{P} = (P_\theta)_{\theta \in \Theta}$.

Def. Let $\mathcal{P}$ be a statistical model for X. Any finite dimensional quantity $t(P^X)$ only depending on P^X as $P \in \mathcal{P}$ is called an **identifiable parameter**.

Def. Let $\mathcal{P}$ be a statistical model for X. A **statistic** in this context is any random variable T valued in $\left(\mathbf{R}^d, \mathcal{B}\left(\mathbf{R}^d\right)\right)$ with $d \geqslant 1$, defined by $T = g(X)$ where g is a Borel function not depending on $P \in \mathcal{P}$.

If a statistic is used as a guess for a parameter $t(P) \in \mathbf{R}^d$, it is called an **estimator** of $t(P)$. In this case, the **bias** of T for estimating $t(P)$ is defined as $\mathrm{Bias}(T, P) = \int T \, \mathrm{d}P - t(P)$ whenever $\int |T| \, \mathrm{d}P < \infty$. We say that T is an *unbiased* estimator of $t(P)$ if $\forall P \in \mathcal{P}, \int T \, \mathrm{d}P = t(P)$. The **quadratic risk** or **mean squared error** (in the case $d = 1$) is defined by $\mathrm{MSE}(T, P) = \int (T - t(P))^2 \, \mathrm{d}P = \mathrm{Var}(T) + \mathrm{Bias}(T, P)^2$.

Def. Let T be a statistic valued in $(\mathbf{R}^d, \mathbf{R}^{\mathcal{D}})$ with $d \geqslant 1$. We say that T is a **sufficient statistic** for the model $\mathcal{P}$ if, for all $P \in \mathcal{P}$, the conditional distribution of X given T does not depend on P, that is, there exists a probability kernel $Q \subset \mathbf{R}^d \times \mathcal{X}$ such that, for all $P \in \mathcal{P}$, Q is a regular version of $P^{X|T}$.

Lem. *Let S be a sufficient statistic associated to the Markov kernel Q and let $T = g(X)$ be an unbiased estimator of the parameter $t(P)$ (both real valued). Define $T^R = \int g(x) Q(S, \mathrm{d}x)$. Then T^R is an unbiased estimator of the parameter t and its variance is smaller than that of T. As a consequence we have, $\forall P \in \mathcal{P}, \mathrm{MSE}\left(T^R, P\right) \leqslant \mathrm{MSE}(T, P)$.*

Th (Fisher Factorization theorem). *Let $\nu \in \mathbf{M}_+(\mathsf{X}, \mathcal{X})$. Consider a ν-dominated model $\mathcal{P}$ for X and let $S = g(X)$ be a d-dimensional statistic. Then S is a sufficient statistic for the model $\mathcal{P}$ if and only if there exists a non-negative Borel function h on X such that $\forall P \in \mathcal{P}$, there exists a Borel function $f_P : \mathbf{R}^d \to \mathbf{R}_+$ such that $\frac{\mathrm{d}P^X}{\mathrm{d}\nu} = h \cdot f_P \circ g$.*

Def. Consider a ν-dominated model $\mathcal{P}$ for X. For all $P \in \mathcal{P}$, let us denote by f_P the density of P^X with respect to ν. The **likelihood function** is defined as $P \mapsto f_P \circ X$ on $P \in \mathcal{P}$.

Then, $f_{P_1}(X) \geqslant f_{P_2}(X)$ is an indication that $\mathrm{KL}\left(P_*^X \| P_1^X\right) \leqslant \mathrm{KL}\left(P_*^X \| P_2^X\right)$ where P_* is the true distribution of X.

Rem. Interestingly, we note that if one has a sufficient statistic $S = g(X)$, by the Fisher Factorization theorem, to compare $f_{P_1}(X)$ and $f_{P_2}(X)$, we only need to observe S.

With a parametric model we define the likelihood function directly on Θ, $\theta \mapsto f_\theta \circ X$ where f_θ denotes the density of P_θ with respect to ν.

Def. A statistic $\hat{\theta}_n$ valued in Θ such that $f_{\hat{\theta}_n} \circ X = \max_{\theta\in\Theta} f_\theta \circ X$ is called a **maximum likelihood estimator (MLE)**.

Statistical testing

We define two hypothesis, respectively called the *null hypothesis* and the *alternative hypothesis*.

- $(\mathbf{H}_0)$ the observation variable X has distribution P^X with $P \in \mathcal{P}_0$,
- $(\mathbf{H}_1)$ X has distribution P^X with $P \in \mathcal{P}_1$,

with $\{\mathcal{P}_0, \mathcal{P}_1\}$ a partition of a statistical model $\mathcal{P}$. $(\mathbf{H}_i)$ is simple if $\mathcal{P}_i$ reduces to one point.

Def. A **statistical test** is a statistic δ with values in $\{0, 1\}$. If $\delta = 0$ we say that we accept $(\mathbf{H}_0)$. Otherwise we reject it.

To evaluate the performance of a test δ, two type of risks are considered :

- The *first type* risk is defined as $P \mapsto P(\delta = 1)$ as $P \in \mathcal{P}_0$.
- The *second type* risk is defined as $P \mapsto P(\delta = 0)$ as $P \in \mathcal{P}_1$.

We call *power* of δ the application $P \mapsto P(\delta = 1)$ as $P \in \mathcal{P}_1$.

Def. Let $\alpha \in [0\,;1]$. We say that a test δ is of level α if $\sup_{P\in\mathcal{P}_0} P(\delta = 1) \leqslant \alpha$. We say that δ is uniformly more powerful then δ' for level α if both are of level α and $\forall P \in \mathcal{P}_1, P(\delta = 1) \geqslant P(\delta' = 1)$.

Simple hypotheses

We consider $\mathcal{P}_0 = \{P_0\}$ and $\mathcal{P}_1 = \{P_1\}$, with f_0 and f_1 the densities of P_0^X and P_1^X with respect to a common dominating measure.

Def. The statistic $T = \frac{f_1(X)}{f_0(X)}$ is called the **likelihood ratio statistic**. Let $t \in [0\,;\infty]$. The test defined by $\delta = \begin{cases} 1 & \text{if } T \geqslant t \\ 0 & \text{otherwise} \end{cases}$ is called the **likelihood ratio test** with threshold t.

Th. *Denote by T the likelihood ratio corresponding to $\mathcal{P}_0$ and $\mathcal{P}_1$. Let $t \in [0\,;\infty]$ and set $\alpha_t = P_0(T \geqslant t)$. Then the likelihood ratio test with threshold t is uniformly more powerful than any other test δ' for the level α_t. Moreover, if δ' is of level α_t and as powerful as δ, then they coïncide on the set $\{T \neq t\}$ P_i-a.s. for $i \in \{0, 1\}$.*

Fisher information matrix

We consider a parametric ν-dominated model $\mathcal{P} = (P_\theta)_{\theta\in\Theta}$ for the observation variable X valued in $(\mathsf{X}, \mathcal{X})$, and denote by f_θ the density of P_θ with respect to ν. We assume that Θ is an open subset of $\mathbf{R}^n$ and denote by $\|f\| := \left(\int_{\mathsf{X}} |f(x)|^2 \nu(\mathrm{d}x)\right)^{\frac{1}{2}}$ the norm of the Hilbert space $L^2(\mathsf{X}, \mathcal{X}, \nu)$. Observe that $\forall\theta \in \Theta, \xi_\theta = \sqrt{f_\theta} \in L^2(\mathsf{X}, \mathcal{X}, \nu)$.

Def. We say that $\mathcal{P}$ is **Hellinger differentiable** at θ if $\theta' \mapsto \xi_\theta$ defined from $\Theta \to L^2(\mathsf{X}, \mathcal{X}, \nu)$ admits a derivative at $\theta : \exists! \dot{\xi}_\theta \in (L^2(\mathsf{X}, \mathcal{X}, \nu))^d, \lim_{\theta'\to\theta} \frac{1}{|\theta'-\theta|} \left\| \xi_{\theta'} - \xi_\theta - \dot{\xi}_\theta^T(\theta' - \theta) \right\| = 0$.

Lem. *Let $\theta \in \Theta$ and $V \subset \Theta$ be a neighborhood of θ. Suppose that for ν-a.e. x and all $\theta' \in V$, we can write $\xi_{\theta'}(x) = \xi_\theta(x) + \int_{t=0}^1 g_{t\theta'+(1-t)\theta}^T(x)(\theta' - \theta)\,\mathrm{d}t$, where, for all $x \in \mathsf{X}$, g satisfies one of the following assertions,*

(i) we have $\lim_{\epsilon\downarrow 0} \left\| \sup_{|\theta'-\theta|\leqslant\epsilon} |g_{\theta'} - g_\theta| \right\| = 0$,

(ii) for ν-a.e. x, $\theta' \mapsto g_{\theta'}(x)$ is continuous and $\exists\epsilon > 0, \left\| \sup_{|\theta'-\theta|\leqslant\epsilon} |g_{\theta'}| \right\| < \infty$.

Then $\mathcal{P}$ is Hellinger differentiable at θ with derivative g_θ.

The derivarive of $\theta \mapsto \ln f_\theta(X)$ is called the score function.

Lem. *Suppose that $A := \{f_\theta > 0\}$ does not depend on θ and $\forall x \in A, \theta \mapsto \ln f_\theta(x)$ is continuolsy differentiable on Θ with derivative $\theta \mapsto \dot{l}_\theta(x)$. Suppose moreover that $\forall\theta \in \Theta$ there exists a neighborhood V of θ such that $\int \sup_{\theta'\in V} \left(\left|\dot{l}_\theta(x)\right|^2 f_\theta(x)\right) \nu(\mathrm{d}x) < \infty$. Then $\mathcal{P}$ is Hellinger differentiable with Hellinger derivative given by $\dot{\xi}_\theta(x) = \frac{1}{2}\dot{l}_\theta(x)\xi_\theta(x)\,\mathbf{1}_A(x)$.*

Def. Let $\mathcal{P}$ be Hellinger differentiable with Hellinger derivative $\dot{\xi}_\theta$. The **Fisher information matrix** is defined as $\mathcal{I}(\theta) := 4\int_{\mathsf{X}} \dot{\xi}_\theta(x)\dot{\xi}_\theta(x)^{\mathsf{T}}\nu(\mathrm{d}x)$.

With the conditions of the previous lemma we have $\mathcal{I}(\theta) = \mathbf{E}_\theta\left[\left(\dot{l}_\theta(X)\right)^2\right]$.

Th. *Let $\mathcal{P}$ be Hellinger differentiable with Hellinger derivative $\dot{\xi}_\theta$. Let $T = g(X)$ be a scalar statistic such that, for some $\epsilon > 0$, $\sup_{|\theta'-\theta|\leqslant\epsilon} \mathbf{E}_\theta\left(T^2\right) < \infty$. Define $\psi\colon \theta \to \mathbf{E}_\theta(T)$. Then ψ is differentiable at θ and, if $\mathcal{I}(\theta)$ is positive definite, we have $\mathrm{Var}_\theta(T) \geqslant \dot{\psi}(\theta)^\mathsf{T}\mathcal{I}(\theta)^{-1}\dot{\psi}(\theta)$.*

Def. Let T be as in the previous theorem. If $\forall\theta\in\Theta, \mathrm{Var}_\theta(T) = \dot{\psi}(\theta)^\mathsf{T}\mathcal{I}(\theta)^{-1}\dot{\psi}(\theta)$, we say that T is an efficient estimator of $\psi(\theta)$.

4 Random processes

Random processes

We consider a probability space $(\Omega,\mathcal{F},\mathbf{P})$, an index T and a measurable space $(\mathsf{X},\mathcal{X})$ called the observation space.

Def. A **random process** defined on $(\Omega,\mathcal{F},\mathbf{P})$, indexed on T and valued in $(\mathsf{X},\mathcal{X})$ is a collection $(X_t)_{t\in T}$ of r.v. defined on $(\Omega,\mathcal{F},\mathbf{P})$ and taking values in $(\mathsf{X},\mathcal{X})$.

Def. For each $\omega\in\Omega$, the application $t\mapsto X_t(\omega)$ is called the **path** associated to the experiment ω.

Def. A **filtration** of a measurable space $(\Omega,\mathcal{F})$ is an increasing sequence $(\mathcal{F}_t)_{t\in T}$ of sub-σ-fields of $\mathcal{F}$. A **filtered probability space** $(\Omega,\mathcal{F},(\mathcal{F}_t)_{t\in T},\mathbf{P})$ is a probability space endowed with a filtration. A random process $(X_t)_{t\in T}$ defined on $(\Omega,\mathcal{F},\mathbf{P})$ is said to be **adapted** to the filtration if for each $t\in T$, X_t is $\mathcal{F}_t$-measurable. Then we note $((X_t,\mathcal{F}_t))_{t\in T}$.

Def. The **natural filtration** of a process $(X_t)_{t\in T}$ is the smallest filtration with respect to which $(X_t)_{t\in T}$ is adapted, i.e. $\forall t\in T, \mathcal{F}_t^X = \sigma(X_s, s\leqslant t)$.

Def. We call **finite dimensional distributions**, or **fidi distributions**, of the process X the collection of probability measures $(\mathbf{P}_I)_{I\in\mathcal{I}}$ where $\mathbf{P}_I$ denotes the probability distribution of the random vector $\{X_t, t\in I\}$.

Let $J\subset I$ two finite subsets. Let us denote bu $\Pi_{I,J}$ the canonical projection of X^I onto X^J defined by $\forall x = (x_t)_{t\in I}\in\mathsf{X}^I, \Pi_{I,J}(x) = (x_t)_{t\in J}$. Then $\mathbf{P}_I\circ\Pi_{I,J}^{-1} = \mathbf{P}_J$ (compatibility condition). We denote $\Pi_I = \Pi_{T,I}$ and $\Pi_s = \Pi_{\{s\}}$ where $s\in T$.

Th (Kolmogorov). *Let $\mathcal{I}$ be the set of all finite subsets of T. Suppose that, for all $I\in\mathcal{I}$, ν_I is a probability measure on $(\mathsf{X}^I,\mathcal{X}^{\otimes I})$ and that the collection $\{\nu_I, I\in\mathcal{I}\}$ satisfies $\forall I,J\in\mathcal{I}, I\subset J, \nu_I\circ\Pi_{I,J}^{-1} = \nu_J$. Then there exists a unique probability measure $\mathbf{P}$ on $(\mathsf{X}^T,\mathcal{X}^{\otimes T})$ such that, $\forall I\in\mathcal{I}, \nu_I = \mathbf{P}\circ\Pi_I^{-1}$.*

Def. Let $X = (X_t)_{t\in T}$ be a random process defined on $(\Omega,\mathcal{F},\mathbf{P})$. The **law in the sense of fidi distribution** is the image measure $\mathbf{P}^X$, that is, the unique probability measure defined on $(\mathsf{X}^T,\mathcal{X}^{\otimes T})$ that satisfies $\forall I\in\mathcal{I}, \mathbf{P}^X\circ\Pi_I^{-1} = \mathbf{P}_I$, i.e. $\forall (A_t)_{t\in I}\in\mathcal{X}^I, \mathbf{P}^X\left(\prod_{i\in I} A_t\times\mathsf{X}^{T\setminus I}\right) = \mathbf{P}(X_t\in A_t, t\in I)$.

Def. The canonical functions defined on $(\mathsf{X}^T,\mathcal{X}^{\otimes T})$ is the collection of measurable functions $(\xi_t)_{t\in T}$ valued in $(\mathsf{X},\mathcal{X})$ as $\forall\omega = (\omega_t)_{t\in T}\in\mathsf{X}^T, \xi_t(\omega) = \omega_t$. When $(\mathsf{X}^T,\mathcal{X}^{\otimes T})$ is endowed with the image measure $\mathbf{P}^X$ then the **canonical process** $(\xi_t)_{t\in T}$ defined on $(\mathsf{X}^T,\mathcal{X}^{\otimes T},\mathbf{P}^X)$ has the same fidi distribution as X.

Gaussian processes

Def. The real valued r.v. X is Gaussian if its characteristic function satisfies $\phi_X(u) = \mathbf{E}(e^{iuX}) = \exp(i\mu u - \sigma^2u^2/2)$ where $\mu\in\mathbf{R}$ and $\sigma\in\mathbf{R}_+$.

If $\sigma\neq 0$ then X admits a probability density function $p(x) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$.

Def. A random vector $[X_1,\ldots,X_n]^\mathsf{T}$ valued in $\mathbf{R}^n$ is a Gaussian vector if any linear combination of $X_1,\ldots,X_n$ is a Gaussian variable.

Let μ denote the mean vector of $[X_1,\ldots,X_n]^\mathsf{T}$ and Γ its covariance matrix. Then $\forall u\in\mathbf{R}^n, Y = u^\mathsf{T}X$ is Gaussian, $\mathbf{E}(Y) = u^\mathsf{T}\mu$ and $\mathrm{Var}(Y) = u^\mathsf{T}\Gamma u$. Thus $\phi_X(u) = \mathbf{E}\left[\exp\left(iu^\mathsf{T}X\right)\right] = \exp\left(iu^\mathsf{T}\mu - \frac{1}{2}u^\mathsf{T}\Gamma u\right)$.

Prop. *The probability distribution of an n-dimensional Gaussian vector X is determined by its mean vector and covariance matrix Γ. We denote $X\sim\mathcal{N}(\mu,\Gamma)$. Conversely, for all vector $\mu\in\mathbf{R}^n$ and all non-negative symmetric matrix Γ, the distribution $\mathcal{N}(\mu,\Gamma)$ is well defined.*

Lem. *Let $X\sim\mathcal{N}(\mu,\Gamma)$ with $\mu\in\mathbf{R}^n$ and Γ a $n\times n$ non-negative symmetric matrix. Then X has independent components if and only if Γ is diagonal.*

Prop. *Let $X\sim\mathcal{N}(\mu,\Gamma)$ with $\mu\in\mathbf{R}^n$ and Γ a $n\times n$ non-negative symmetric matrix. If Γ is full rank, the probability distribution of X admits a density defined in $\mathbf{R}^n$ by $p(x) = \frac{1}{(2\pi)^{n/2}\sqrt{\det(\Gamma)}}\exp\left(-\frac{1}{2}(x-\mu)^\mathsf{T}\Gamma^{-1}(x-\mu)\right)$.*

Def. A real-valued random process $X = (X_t)_{t\in T}$ is called a **Gaussian process** if, for all finite set of indices $I = \{t_1,\ldots,t_n\}$, $[X_{t_1},\ldots,X_{t_n}]^\mathsf{T}$ is a Gaussian vector.

Th. *Let T be any set of indices, $\mu\colon T\to\mathbf{R}$ and $\gamma\colon T\times T\to\mathbf{R}$ such that all restrictions Γ_I to the set $I\times I$ with $I\subset T$ finite are nonnegative symmetrice matrices. Then one can define a probability space $(\Omega,\mathcal{F},\mathbf{P})$ and a gaussien process $(X_t)_{t\in T}$ defined on this space with mean μ and covariance function γ, that is such that, $\forall s,t\in T, \mu(t) = \mathbf{E}[X_t]$ and $\gamma(s,t) = \mathbf{E}[(X_s-\mu(s))(X_t-\mu(t))]$.*

Def. Let T, μ and γ be as above. We denote by $\mathcal{N}(\mu, \gamma)$ the law of the Gaussian process with mean μ and covariance γ in the sense of fidi distribution.

Strict stationarity of a random process in discrete time

Suppose that $T = \mathbf{Z}$ or $T = \mathbf{N}$.

Def. We denote by $S\colon \mathsf{X}^T \to \mathsf{X}^T$ and call the **shift operator** the mapping defined by $\forall x = (x_t)_{t\in T}, S(x) = (x_{t+1})_{t\in T}$. For all $\tau \in T$ we define S^τ by $S^\tau(x) = (x_{t+\tau})_{t\in\tau}$. The operator $B = S^{-1}$ is called the **backshift operator**.

Def. A random process $(X_t)_{t\in T}$ is **strictly stationary** if X and $S \circ X$ have the same law, i.e. $\mathbf{P}^{S\circ X} = \mathbf{P}^X$.

Stationarity preserving transformations

In this section, we set $T = \mathbf{Z}$, $\mathsf{X} = \mathbf{C}^d$ and $\mathcal{X} = \mathcal{B}(\mathbf{C}^d)$ for some integer $d \geqslant 1$.

Def. Let ϕ be a measurable function from $\left(\mathsf{X}^T, \mathcal{X}^{\otimes T}\right)$ to $\left(\mathsf{Y}^T, \mathcal{Y}^{\otimes T}\right)$ and $X = (X_t)_{t\in T}$ be a process with values in $(\mathsf{X}, \mathcal{X})$. A ϕ-filtering process $Y = (Y_t)_{t\in T}$ is defined as $\forall t, Y = \phi \circ X$ or, equivalently, $Y_t = \Pi_t(\phi(X))$. Thus Y makes its values in $(\mathsf{Y}, \mathcal{Y})$. If ϕ is linear, we will say that Y is obtained by linear filtering of X.

Def. A ϕ-filter is **shift invariant** if ϕ commutes with S.

Rem. A shift invariant ϕ-filter preserves the strict stationarity and is entirely determined by its composition with the canonical projection Π_0

5 Weakly stationary processes

L^2 processes

Def. The process $X = (X_t)_{t\in T}$ defined on $(\Omega, \mathcal{F}, \mathbf{P})$ with values in $\mathbf{C}^d$ is an L^2 process if $\forall t \in T, X_t = L^2(\Omega, \mathcal{F}, \mathbf{P})$. Its **mean function** is defined on T by $\mu(t) = \mathbf{E}(X_t)$ and the **covariance function** is defined on $T \times T$ by $\Gamma(s,t) = \mathrm{Cov}(X_s, X_t) = \mathbf{E}\left((X_s - \mu(s))(X_t - \mu(t))^{\mathsf{H}}\right)$.

Prop. *Let Γ be the covariance function of a L^2 process $X = (X_t)_{t\in T}$ with values in $\mathbf{C}^d$. The following properties hold.*

(i) Hermitian symmetry : $\forall s,t \in T, \Gamma(s,t) = \Gamma(t,s)^{\mathsf{H}}$.

(ii) Nonnegativity : $\forall n \in \mathbf{N}^*, t_1, \ldots, t_n \in T, a_1, \ldots, a_n \in \mathbf{C}^d, \sum_{1\leqslant k,m\leqslant n} a_k^{\mathsf{H}}\Gamma(t_k, t_m)a_m \geqslant 0$.

Conversely, if Γ satisfies these two properties, there exists an L^2 process with values in $\mathbf{C}^d$ with covariance function Γ.

In the scalar case ($d = 1$), we also use he notation $\gamma(s,t)$.

Weakly sationary processes

Def. Let $\mu \in \mathbf{C}^d$ and $\Gamma\colon \mathbf{Z} \to \mathbf{C}^{d\times d}$. A process $(X_t)_{t\in\mathbf{Z}}$ with values in $\mathbf{C}^d$ is said **weakly stationary** with mean μ and autocovariance function Γ if all the following assertions hold :

(i) X is an L^2 process, i.e. $\mathbf{E}\left(|X_t|^2\right) < +\infty$,

(ii) $\forall t \in \mathbf{Z}, \mathbf{E}(X_t) = \mu$,

(iii) $\forall (s,t) \in \mathbf{Z} \times \mathbf{Z}, \mathrm{Cov}(X_s, X_t) = \Gamma(s - t)$.

A strictly stationary L^2 process is weakly stationary.

Prop. *The autocovariance function $\gamma\colon \mathbf{Z} \to \mathbf{C}$ of a complex valued weakly stationary process satisfies the following :*

(i) Hermition symmetry : $\forall s \in \mathbf{Z}, \gamma(-s) = \overline{\gamma(s)}$.

(ii) Nonnegative definiteness : $\forall i \in \mathbf{N}^*, a_1, \ldots, a_n \in \mathbf{C}, \sum_{s=1}^n \sum_{t=1}^n \overline{a_s}\gamma(s-t)a - t \geqslant 0$.

Def. Let X be a weakly stationary process with autocovariance function γ such that $\gamma(0) \neq 0$. The **autocorrelation function** of X is defined as $\forall \tau \in \mathbf{Z}, \rho(\tau) = \frac{\gamma(\tau)}{\gamma(0)}$. It is normalized in the sense that $\rho(0) = 1$ and $\forall s \in \mathbf{Z}, |\rho(s)| \leqslant 1$.

Def. A **weak white noise** is a centered weakly stationary process whose autocovariance function satisfies $\gamma(0) = \sigma^2 > 0$ and $\forall s \neq 0, \gamma(s) = 0$. We will denote $(X_t) \sim \mathrm{WN}(0, \sigma^2)$. When a white noise is an i.i.d. process it is called a **strong white noise**. We will denote $(X_t) \sim \mathrm{IID}(0, \sigma^2)$.

Spectral measure

Th. *(Heglotz) A sequence $(\gamma(h))_{h\in\mathbf{Z}}$ is a nonnegative definite hermitian sequence iff there exists a finite nonnegative measure ν on $(\mathbf{T}, \mathbf{T})$ such that*

$$\forall h \in \mathbf{Z}, \gamma(h) = \int_{\mathbf{T}} e^{ih\lambda}\nu(\mathrm{d}\lambda)\,.$$

Moreover this relation defines ν uniquely.

Rem. This applies to all autocovariance function of a weakly stationary process X. In this case ν is called the **spectral measure** of X. If ν admits a density f, it is called the **spectral density function**.

Cor. *Let $(\gamma(h))_{h\in\mathbf{Z}} \in l^2(\mathbf{Z})$. Then it is a nonnegative definite hermitian sequence iff for almost every λ, $f(\lambda) = \frac{1}{2\pi}\sum_{h\in\mathbf{Z}} \gamma(h)e^{-ih\lambda}$ is nonnegative, where the convergence holds in $L^2(\mathbf{T})$.*

Def. A weakly stationary process X is called **linearly predictable** if $\exists n \geqslant 1, \forall t \geqslant n, X_t \in \mathrm{Span}(X_1,\dots,X_n)$ (in the L^2 sense).

Prop. *Let γ be the autocovariance function of a weakly stationary process X. If $\gamma(0) \neq 0$ and $\gamma(t) \overset{t\to\infty}{\to} 0$ then X is not linearly predictable.*

Spectral representation of a weakly stationary process

Def. A **random fields with orthogonal increments** W on $(\mathsf{X},\mathcal{X})$ is a L^2 random process indexed on $\mathcal{X}$, say $W = (W(A))_{A\in\mathcal{X}}$ such that

(i) $\forall A \in \mathcal{X}, \mathbf{E}(W(A)) = 0$,

(ii) $\forall A, B \in \mathcal{X}$ such that $A \cap B = \varnothing$, $W(A)$ and $W(B)$ are uncorrelated and $W(A\cup B) = W(A) + W(B)$,

(iii) for all nonincreasing sequence $(A_n)_{n\in\mathbf{N}} \subset \mathcal{X}, \bigcap_{n=0}^{\infty} A_n = \varnothing$, we have $\mathrm{Var}(W(A_n)) \to 0$.

Lem. *Let W a random field with orthogonal increments on $(\mathsf{X},\mathcal{X})$. Let $A \in \mathcal{X}$ and set $\nu(A) = \mathrm{Var}(W(A))$. Then ν is a finite nonnegative measure on $(\mathsf{X},\mathcal{X})$. Moreover $\forall A, B \in \mathcal{X}, \mathrm{Cov}(W(A), W(B)) = \nu(A\cap B)$.*

The measure ν is called the **intensity measure** of W.

Lem. *Let W be a L^2 random process indexed by $\mathcal{X}$ such that $\forall A \in \mathcal{X}, \mathbf{E}(W(A)) = 0$. Suppose $\exists\nu, \forall A, B \in \mathcal{X}, \mathrm{Cov}(W(A), W(B)) = \nu(A\cap B)$. Then W is a random field with orthogonal increments on with intensity measure ν.*

Th. *Let W be a random field with orthgonal increments with intensity measure ν. Then there exists a unique isometric operator w from $L^2(\mathsf{X},\mathcal{X},\nu)$ to $L^2(\Omega,\mathcal{F},\mathbf{P})$ such that $\forall A \in \mathcal{X}, w(\mathbf{1}_A) = W(A)$. For all $f \in L^2(\mathsf{X},\mathcal{X},\nu)$ we further have $\mathbf{E}(w(f)) = 0$ and $w(L^2(\mathsf{X},\mathcal{X},\nu) = \overline{\mathrm{Span}}(W(A), A \in \mathcal{X})$.*

Th. *Let ν be a finite nonnegative measure on $(\mathsf{X},\mathcal{X})$ and $J\colon L^2(\mathsf{X},\mathcal{X},\nu) \to L^2(\Omega,\mathcal{F},\mathbf{P})$ an isometric operator such that $\forall f, \mathbf{E}(J(f)) = 0$. Then there exists a random field W with orthogonal increments on X with intensity measure ν such that $\forall f, J(f) = \int_{\mathsf{X}} f\,\mathrm{d}W$.*

Prop. *Let W be a r.f.o.i. on $(\mathbf{T},\mathcal{B}(\mathbf{T}))$ with intensity measure ν. Then the sequence $(X_t)_{t\in\mathbf{Z}}$ defined by $X_t = \int_{\mathbf{T}} e^{it\lambda}\,\mathrm{d}W(\lambda)$ is a centered weakly stationary process with spectral measure ν.*

Def. Let X be a L^2 process. Its **linear closure** is defined as $\mathcal{H}^X_\infty = \overline{\mathrm{Span}}(X_t, t\in\mathbf{Z})$ (closure in $L^2(\Omega,\mathcal{F},\mathbf{P})$).

Th. *Let X be a centered weakly stationary process with spectral measure ν. Then there exists a r.f.o.i. $\hat{X}$ on $(\mathbf{T},\mathcal{B}(\mathbf{T}))$ with intensity measure ν, called the **spectral field**, such that $\forall t \in \mathbf{Z}, X_t = \int e^{it\lambda}\,\mathrm{d}\hat{X}(\lambda)$. Moreover, the mapping $f \mapsto \int f\,\mathrm{d}\hat{X}$ defines the unique operator from $L^2(\mathbf{T},\mathcal{B}(\mathbf{T}),\nu)$ to $\mathcal{H}^X_\infty$ that maps each function $\lambda \mapsto e^{it\lambda}$ to X_t.*

Innovation process

Let $X = (X_t)_{t\in\mathbf{Z}}$ denote a centered weakly stationary process. Let $\mathcal{H}^X_t = \overline{\mathrm{Span}}(X_s, s \leqslant t)$ denote the **linear past** of X up to time t.

Def. We call **innovation process** the process $\epsilon = (\epsilon_t)_{t\in\mathbf{Z}}$ defined by $\epsilon_t = X_t - \mathrm{proj}(X_t \mid \mathcal{H}^X_{t-1})$.

Def. We call **predictor of order** p the random variable $\mathrm{proj}(X_t \mid \mathcal{H}^X_{t-1,p})$ and the **partial innovation process** of order p the process $\epsilon^+_p = (\epsilon^+_{t,p})_{t\in\mathbf{Z}}$ defined by $\epsilon^+_{t,p} = X_t - \mathrm{proj}(X_t \mid \mathcal{H}^X_{t-1,p})$. The **prediction coefficients** are any coefficients $\phi^+_p = (\phi^+_{k,p})_{k\in[\![1;p]\!]}$ which satisfy $\forall t \in \mathbf{Z}, \mathrm{proj}(X_t \mid \mathcal{H}^X_{t-1,p}) = \sum_{k=1}^{p} \phi^+_{k,p} X_{t-k}$.

Cor. *The innovation process of a centered weakly stationary process is a (centered) weak white noise. Its variance is called the **innovation variance** of the process.*

Def. If the variance of its innovation process is zero, we say that X is **deterministic**. Otherwise we say that X is **regular**.

We define the intersection of the whole past of X as $\mathcal{H}^X_{-\infty} = \bigcap_{t\in\mathbf{Z}} \mathcal{H}^X_t$. We denote $\psi_s = \frac{\langle X_t | \epsilon_{t-s}\rangle}{\sigma^2}$.

Th (Wold decomposition). *Let X be a regular process, ϵ its innovation process and σ^2 its innovation variance, so that $\epsilon \sim \mathrm{WN}(0,\sigma^2)$. Define the L^2 centered process U as $U_t = \sum_{k=0}^{\infty} \psi_k \epsilon_{t-k}$. Define the L^2 centered process V b $\forall t \in \mathbf{Z}, X_t = U_t + V_t$. Then the following assertions hold,*

(i) $U_t = \mathrm{proj}(X_t \mid \mathcal{H}^\epsilon_t)$ and $V_t = \mathrm{proj}(X_t \mid \mathcal{H}^X_{-\infty})$,

(ii) ϵ and V are uncorrelated : $\forall(t,s), \langle V_t \mid \epsilon_s\rangle = 0$,

(iii) U is a purely non-dterministic process and has same innovation as X. Moreover $\forall t \in \mathbf{Z}, \mathcal{H}^\epsilon_t = \mathcal{H}^U_t$.

(iv) V is a deterministic process and $\mathcal{H}^V_{-\infty} = \mathcal{H}^X_{-\infty}$.

6 Markov chains : basic definitions

Deffinition using conditioning

Def. Let $(\Omega,\mathcal{F},(\mathcal{F}_k)_{k\in\mathbf{N}},\mathbf{P})$ be a filtered probability space and $(\mathsf{X},\mathcal{X})$ be a measurable space. An adapteted stochastic process $((X_k,\mathcal{F}_k))_{k\in\mathbf{N}}$ on X is a **Markov chain** if $\forall k \in \mathbf{N}, \forall A \in \mathcal{X}, \mathbf{P}(X_{k+1} \in A \mid \mathcal{F}_k) = \mathbf{P}(X_{k+1} \in A \mid A_k)$.

Prop. *Let $((X_k, \mathcal{F}_k))_{k\in\mathbf{N}}$ be an adapted stochastic process. The following propoerties are equivalent :*

(i) $((X_k, \mathcal{F}_k))_{k\in\mathbf{N}}$ *is a Markov chain,*

(ii) $\forall k \in \mathbf{N}, \forall Y \in L^1(\Omega, \sigma(X_l, l \geqslant k), \mathbf{P}), \mathbf{E}(Y \mid \mathcal{F}_k) = \mathbf{E}(Y \mid X_k)$ $\mathbf{P}$*-a.s.*

(iii) $\forall k \in \mathbf{N}, \forall Y \in L^1(\Omega, \sigma(X_l, l \geqslant k), \mathbf{P}), \forall Z \in L^\infty(\Omega, \mathcal{F}_k, \mathbf{P}), \mathbf{E}(YZ \mid X_k) = \mathbf{E}(Y \mid X_k)\mathbf{E}(Z \mid X_k)$ $\mathbf{P}$*-a.s.*

How to use kernels

Let N be a kernel and f be a measurable function defined on Y. We denote by Nf the function defined on X by $Nf\colon x \mapsto \int_{\mathsf{Y}} N(x, \mathrm{d}y)f(y)$ whenever this integral is well-defined (for instance if f is non-negative).

Prop. *Let N be a kernel on $\mathsf{X} \times \mathcal{Y}$. Then $\forall f \in \mathbf{F}_+(\mathsf{Y}, \mathcal{Y}), Nf \in \mathbf{F}_+(\mathsf{X}, \mathcal{X})$. Moreover, if N is a probability kernel, then $\forall f \in \mathbf{F}_b(\mathsf{X}, \mathcal{X}), Nf \in \mathbf{F}_b(\mathsf{Y}, \mathcal{Y})$.*

The canonical Chain, notation $\mathbf{P}_\xi$ and $\mathbf{E}_\xi$

Th. *Let P be a Markov kernel on $\mathsf{X} \times \mathcal{X}$ and $\nu \in \mathbf{M}_1(\mathsf{X}, \mathcal{X})$. Then, there exists a unique probability $\mathbf{P}_\nu$ on $(\mathsf{X}^{\mathbf{N}}, \mathcal{X}^{\otimes\mathbf{N}})$ is a Markov chain with initial distribution ν and kernel P.*

Def. The **canonical Markov chain** with kernel P on $\mathsf{X} \times \mathcal{X}$ is the canonical process $(X_n)_{n\in\mathbf{N}}$ on the canonical filtered space $(\mathsf{X}^{\mathbf{N}}, \mathcal{X}^{\otimes\mathbf{N}}, (\mathcal{F}_k)_{k\in\mathbf{N}})$ endowed with the collection of probability measures $\{\mathbf{P}_\nu, \nu \in \mathbf{M}_1(\mathsf{X}, \mathcal{X})\}$ given by the previous theorem.

The expectation associated to $\mathbf{P}_\nu$ is denoted by $\mathbf{E}_\nu$ and for $x \in \mathsf{X}$, $\mathbf{P}_x$ and $\mathbf{E}_x$ are shorthand for $\mathbf{P}_{\delta_x}$ and $\mathbf{E}_{\delta_x}$.

Prop. *(i)* $\forall A \in \mathcal{X}^\otimes$*, the function* $x \mapsto \mathbf{P}_x(A)$ *is* $\mathcal{X}$*-measurable.*

(ii) $\forall \xi \in \mathbf{M}_1(\mathsf{X}, \mathcal{X}), \forall A \in \mathcal{X}^\otimes, \mathbf{P}_\xi(A) = \int_{\mathsf{X}} \mathbf{P}_x(A)\xi(\mathrm{d}x)$

(iii) $\forall \xi \in \mathbf{M}_1(\mathsf{X}, \forall Y \in \mathbf{F}_+(\mathsf{X}^{\mathbf{N}}, \mathcal{X}^{\otimes\mathbf{N}}), \mathbf{E}_\xi(Y) = \int_{\mathsf{X}} \mathbf{E}_x(Y)\xi(\mathrm{d}x)$

(iv) $\forall \xi \in \mathbf{M}_1(\mathsf{X}, \forall Y \in L^1(\mathsf{X}^{\mathbf{N}}, \mathcal{X}^{\otimes\mathbf{N}}, \mathbf{P}_\xi), \mathbf{E}(|Y|) < \infty$ *for ξ-a.e. x and* $\mathbf{E}_\xi(Y) = \int_{\mathsf{X}} \mathbf{E}_x(Y)\xi(\mathrm{d}x)$

(v) $\forall \xi \in \mathbf{M}_1(\mathsf{X}, \forall Y \in L^1(\mathsf{X}^{\mathbf{N}}, \mathcal{X}^{\otimes\mathbf{N}}, \mathbf{P}_\xi), \mathbf{E}_\xi(Y \mid X_0) = \mathbf{E}_{X_0}(Y)$ $\mathbf{P}_\xi$*-a.s.*

Def. An event is $\mathbf{P}_*$-a.s. if it is $\mathbf{P}_\nu$-a.s. for all initial distribution ν.

Lem. *Let $A \in \mathcal{X}^{\otimes\mathbf{N}}$. We have $X \in A$ $\mathbf{P}$-a.s. if and only if it is true $\mathbf{P}_x$-a.s. for all $x \in \mathsf{X}$.*

7 Stationary Markov chains

Invariant measures and stationarity

Let P be a Markov kernel on $(\mathsf{X}, \mathcal{X})$.

Def. A non zero σ-finite positive measure $\mu \in \mathbf{M}_+(\mathsf{X}, \mathcal{X})$ is said to be **invariant** with respect to P (or P-invariant) if $\mu = \mu P$.

Th. *A Markov chain $(X_k)_{k\in\mathbf{N}}$ defined on $(\Omega, \mathcal{F}, \mathbf{P})$ with kernel P is a stationary process iff its initial distribution is invariant with respect to P.*

Def. A non-empty set $B \in \mathcal{X}$ is called **absorbing** if $\forall x \in B, P(x, B) = 1$.

Th. *(i) The set of invariant probability measures for P is a convex subset of the convex cone $\mathbf{M}_+(\mathsf{X}, \mathcal{X})$.*

(ii) Let π be an invariant probability and $\mathsf{X}_1 \subset \mathsf{X}$ with $\pi(\mathsf{X}_1) = 1$. There exists $B \subset \mathsf{X}_1$ such that $\pi(B) = 1$ and B is absorbing for P.

Def. A random variable $\tau\colon \Omega \to \bar{\mathbf{N}} = \mathbf{N} \cup \{\infty\}$ is called a **stopping time** if $\forall k \in \mathbf{N}, \{\tau \leqslant k\} \in \mathcal{F}_k$. The family $\mathcal{F}_\tau$ of events A such that $\forall k \in \mathbf{N}, A \cap \{\tau \leqslant k\} \in \mathcal{F}_k$ is called the σ-field of events prior to times τ.