# Modélisation

——

## M1 Mathématiques et Applications
## Aix-Marseille Université
## Site Luminy

B. Torrésani

Année 2017-18

# Table des matières


**1 Préliminaires** **7**

1.1 Eléments d'imagerie cérébrale, EEG, interfaces cerveau-machine . . . 7
1.1.1 Le cerveau . . . 7
1.1.2 Quelques éléments d'imagerie cérébrale . . . 8
1.1.3 Interfaces cerveau machine . . . 9
1.2 Structure de données d'EEG, prise en main sous le logiciel R . . . 10
1.2.1 Données EEG . . . 10
1.2.2 Le *P300 speller* . . . 10
1.2.3 Structure des données étudiées sous R . . . 11
1.2.4 Analyse préliminaire . . . 12
1.3 Eléments de théorie du signal . . . 13
1.3.1 Généralités . . . 13
1.3.2 Représentation des signaux, aspects algébriques . . . 14
1.3.3 Quelques exemples d'espaces de Hilbert utiles en traitement du signal . . . 16
1.3.4 Approximation des signaux . . . 17
1.4 Rappels sur les vecteurs aléatoires . . . 20
1.4.1 Quelques définitions . . . 20
1.4.2 Vecteurs aléatoires à densité . . . 21
1.4.3 Covariance . . . 23
1.4.4 Vecteurs gaussiens . . . 23

**2 Une approche par classification supervisée** **27**

2.1 Eléments sur l'estimation . . . 27
2.1.1 Généralités, exemples . . . 27
2.1.2 Estimation de covariance . . . 29
2.1.3 Maximum de vraisemblance . . . 31
2.2 Décision . . . 32
2.2.1 Position du problème . . . 32
2.2.2 La règle de Bayes . . . 34
2.3 Analyse discriminante . . . 36
2.3.1 Analyse discriminante linéaire (LDA) . . . 37
2.3.2 Analyse discriminante quadratique (QDA) . . . 37
2.3.3 Application aux données de *P300-speller* . . . 38

**3 Solutions des exercices** **39**
3.1 Exercices du chapitre 1 . . . . . 39
3.2 Exercices du chapitre 2 . . . . . 39

exo

Chapitre

# 1 Préliminaires

## 1.1 Eléments d'imagerie cérébrale, EEG, interfaces cerveau-machine

### 1.1.1 Le cerveau

Le cerveau [1] peut être imaginé comme un immense réseau de fils électriques, ces fils étant les "queues" des neurones. Le neurone est l'élément de base du cerveau, notre cerveau en contient plus de 100 milliards, dont chacun communique avec 10 000 autres de ses voisins.
Les neurones sont des cellules de forme assez originale : à partir d'un corps central (10 à 50 millièmes de millimètres) partent des "bras", les dendrites, et une "queue", l'axone. Les corps cellulaires et les axones ne sont pas mélangés : les neurones constituent la "matière grise", ou cortex, alors quel le réseau d'axones constitue la "matière blanche", une sorte de faisceaux de câbles.

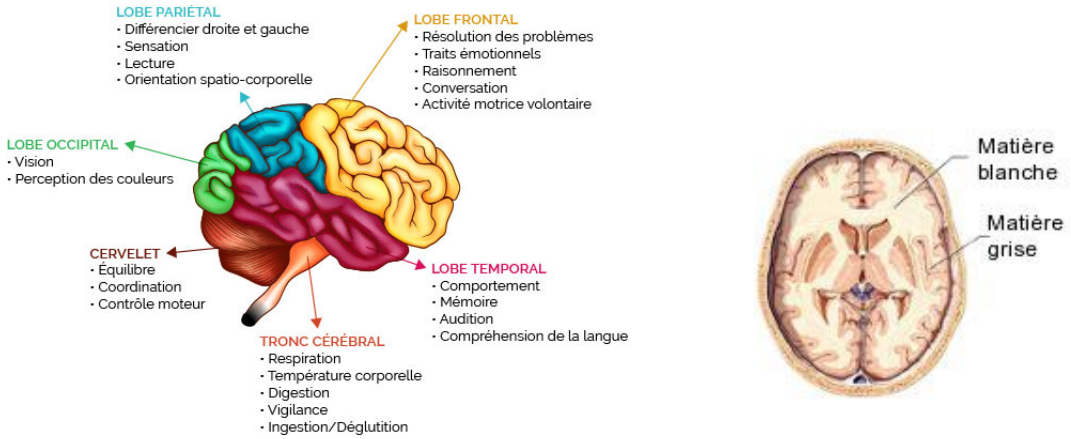


FIGURE 1.1 – Vue schématique d'un cerveau humain, avec les principales aires fonctionnelles (gauche), matière grise et matière blanche (droite)

Les informations circulent dans les neurones par influx électrique (voir FIGURE 1.2). Pour que le signal se propage, il faut que le potentiel électrique atteigne un certain seuil. Il existe des synapses excitatrices et inhibitrices : les synapses excitatrices accentuent l'amplitude électrique, et les synapses inhibitrices le modèrent. Pour atteindre le seuil de déclenchement du signal, il faut au moins 40 synapses excitatrices actives en même temps.
La science de l'observation du cerveau est appelée imagerie cérébrale. Il s'agit de reconstituer l'activité électrique cérébrale à partir de mesures, généralement des mesures externes, de courant électrique ou champ magnétique par exemple. L'imagerie cérébrale a notamment permis d'identifier les régions du cerveau impliquées dans différentes fonctions (voir la FIGURE 1.1 gauche).

1. Cette section est inspirée du dossier sur le cerveau disponible sur linternaute.fr, voir http://www.linternaute.com/science/biologie/dossiers/06/0602-cerveau/1.shtml

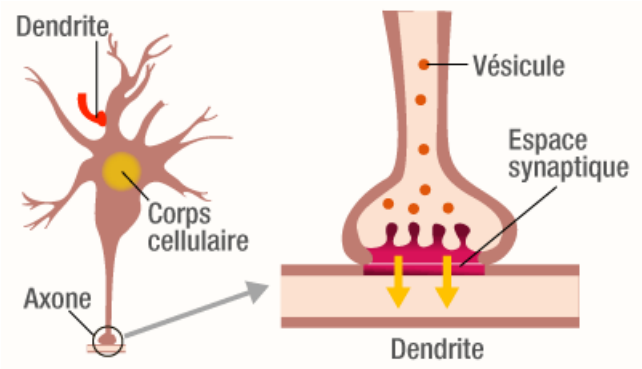


FIGURE 1.2 – Fonctionnement d'une synapse

### 1.1.2 Quelques éléments d'imagerie cérébrale

Les lignes ci-dessous sont tirées de l'encyclopédie en ligne Wikipedia [2]. L'imagerie cérébrale (aussi appelée neuro-imagerie) désigne l'ensemble des techniques issues de l'imagerie médicale qui permettent d'observer le cerveau, en particulier lorsqu'un individu exécute une tâche cognitive.
L'**imagerie structurelle** (dite aussi anatomique) cherche à identifier, localiser et mesurer les différentes parties de l'anatomie du système nerveux central. Dans la pratique médicale clinique, elle permet d'identifier la localisation et l'extension d'une lésion cérébrale dans une visée diagnostique et/ou d'intervention chirurgicale.
L'**imagerie fonctionnelle** cherche à caractériser le cerveau en action. L'usage traditionnel de ces méthodes consiste à faire effectuer une tâche cognitive à un individu et à mesurer le signal produit par l'activité cérébrale. Suivant les techniques et les outils mathématiques employés, il est possible de retrouver, avec plus ou moins de précision, quelle région du cerveau était particulièrement active et à quel moment de la tâche cognitive.
Il existe différentes techniques permettant l'imagerie fonctionnelle du cerveau. On mentionnera en particulier les suivantes :

— L'électroencéphalographie (EEG), première méthode de neuroimagerie non invasive mise au point en 1929 par le neurologue Hans Berger est une mesure directe de l'activité électrique. L'EEG mesure le potentiel électrique généré par l'activité neuronale, en un certain nombre de points sur la surface du scalp (voir Figure 1.3). L'EEG est relativement peu précise spatialement mais elle offre une résolution temporelle limitée seulement par la vitesse de l'électronique de mesure.
Une première approche consiste à mesurer des potentiels évoqués : en répétant une même stimulation un grand nombre de fois, il est possible de mettre en évidence des ondes positives et négatives caractéristiques des différentes étapes du processus traitement de l'information (e.g., N100, P300, N400). Une autre approche consiste à mesurer par Électroencéphalographie quantitative les modifications des activités rythmiques qui semblent jouer un rôle fonctionnel important dans la cognition.

— La magnétoencéphalographie (MEG) offre une information relativement similaire à l'EEG, mais elle mesure quant à elle les champs magnétiques induits par l'activité cérébrale. L'intérêt de la MEG réside dans le fait que, contrairement aux champs électriques, les champs magnétiques ne sont quasiment pas déformés par leur passage au travers des tissus organiques (notamment l'interface entre le liquide céphalo-rachidien et le crâne). Tout comme avec l'EEG, il est possible, via une analyse mathématique du signal de reconstruire les sources du signal électromagnétique. Cela permet d'identifier avec une plus ou moins grande précision les régions d'où sont émis les potentiels évoqués. Cependant, ces techniques de localisation spatiale allongent considérablement

2. https://fr.wikipedia.org/wiki/Imagerie_c%C3%A9r%C3%A9brale

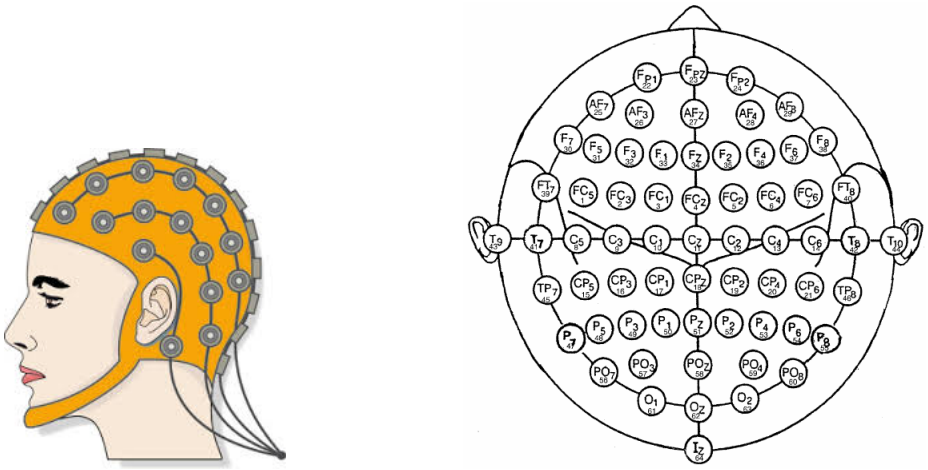

FIGURE 1.3 – Une exemple de dispositif de capteurs EEG. A gauche, représentation d'un casque EEG. A droite, nomenclature des capteurs dans un dispositif standard.

le temps de traitement des données et restent encore marginales.

— L'imagerie par résonance magnétique fonctionnelle (IRMf) consiste à mesurer un signal spécifique appelé *signal BOLD*, qui reflète le taux d'oxygénation du sang dans le cerveau. Par un mécanisme encore mal expliqué, appelé réponse hémodynamique, l'afflux de sang oxygéné augmente dans les régions qui consomment de l'énergie. Ainsi, il est possible, par cette méthode, de connaître avec une grande précision quelles régions du cerveau sont spécialement actives lors d'une tâche donnée.
— La tomographie par émission de positrons (TEP) consiste à mesurer les modifications du débit sanguin au moyen d'un traceur radioactif qu'il faut préalablement injecter par voie intraveineuse. La diffusion du traceur et la modulation du débit sanguin étant des phénomènes relativement lents, cette technique ne donne pas accès à la dynamique des mécanismes neuronaux. Ceci en fait une technique aujourd'hui de moins en moins utilisée pour l'imagerie fonctionnelle.
— De nombreuses autres modalités, telles que par exemple l'imagerie optique.

### 1.1.3 Interfaces cerveau machine

Une interface cerveau-machine (ICM), ou interface neuronale direct, désigne un système de liaison directe entre un cerveau et un ordinateur, permettant à un individu de communiquer avec son environnement sans passer par l'action des nerfs périphériques et des muscles. Les ICM permettent d'améliorer notamment le confort de vie de certains patients atteints de troubles graves de la motricité (*locked-in syndrom* consécutif à un accident vasculaire cérébral, sclérose latérale amyotrophique, lésions de la moelle épinière). Elles sont maintenant utilisées dans d'autres domaines, comme par exemple le pilotage d'objets volants, le jeu video. Les ICM constituent un domaine en évolution extrêmement rapide.
Dans ce cours on se focalisera principalement sur un exemple d'ICM, appelée *P300-speller*, développé pour permettre à des patients de communiquer lettre à lettre, en identifiant des lettres présentées sur un écran. Lorsqu'une lettre est reconnue par le sujet, la réponse d'une certaine région du cortex prend (à des fluctuations près, qui peuvent être très importantes) une forme assez spécifique appelée P300 (parce qu'elle arrive approximativement 300 millisecondes après la présentation de la lettre, et prend la forme d'une "bosse" positive). La P300 est détectable, à condition d'ajuster soigneusement les caractéristiques de l'algorithme de détection. L'objectif du projet associé à ce cours est de développer une méthode de détection aussi performante que possible.


FIGURE 1.4 – Un exemple de signal EEG "multicapteur". En abscisse le temps, en ordonnée les signaux mesurés sur un certain nombre de capteurs.

## 1.2 Structure de données d'EEG, prise en main sous le logiciel R

### 1.2.1 Données EEG

Pour un sujet et une expérience donnés, les données EEG prennent la forme d'une série de signaux temporels, comme on peut les visualiser en Figure 1.4.
Chacun des signaux temporels associe à chaque instant considéré $t$ la valeur $v^c(t)$ du potentiel électrique mesuré à cet instant là sur un capteur donné $c$. La variable temps naturellement continue, est discrétisée (on note $v^c_n = v^c((n-1)/\eta$, où $n = 1, \dots L$ est la longueur du signal discrétisé, et $\eta$ est une constante, appelée fréquence d'échantillonnage , représentant le nombre de mesures par seconde. $c = 1, \dots N_c$, le nombre de capteurs qui est fixé (par exemple $N_c = 64$). Un jeu de données EEG prend donc la forme d'un tableau à deux entrées.
En général, lors d'une expérience plusieurs enregistrements sont effectués, donc un jeu de données comprend plusieurs tableaux.

### 1.2.2 Le *P300 speller*

Le paradigme du *P300 speller* est le suivant :

1. On présente à l'utilisateur une matrice de caractères 6 par 6 (voir la figure 1.5), sa tâche est de se focaliser sur les caractères d'un mot prescrit (en phase d'apprentissage), ou des caractères qu'il désire épeler.
2. Toutes les lignes et les colonnes de cette matrice sont intensifiées de façon successive et aléatoire à un rythme de 5,71 Hz (une ligne ou colonne est intensifiée durant 100ms, puis aucune intensification durant 75ms). Par construction du dispositif, deux intensifications sur 12 de lignes ou de colonnes contiennent le caractère souhaité (c'est-à-dire une ligne particulière et une colonne particulière). Les réponses évoquées par ces stimuli peu fréquents (c'est-à-dire les 2 stimuli sur


FIGURE 1.5 – Une matrice de caractères utilisée dans le dispositif *P300 speller*.

12 contenant le caractère désiré) sont censées être différentes de celles évoquées par les stimuli ne contenant pas le caractère désiré et être similaires au P300.

3. Les signaux sont enregistrés par une série de capteurs (électrodes) mesurant le potentiel électrique sur un ensemble de points spécifiquement choisis sur le scalp. Chacun de ces signaux prend la forme d'une série temporelle de $L = 240$ valeurs correspondant à une durée d'une seconde (on dit que la fréquence d'échantillonnage vaut 240 Hz).
4. Une session consiste en un certain nombre de *runs*, chacun d'entre eux correspondant à une lettre « cible » à épeler. Pour chaque *run*, les 12 intensifications sont répétées 15 fois.

### 1.2.3 Structure des données étudiées sous R

Les jeux de données qui seront étudiées dans le cadre des TP et du projet sont issues d'une compétition internationale [3], dans le cadre de laquelle les équipes de recherche concurrentes étaient invitées à tester leurs méthodes sur un même jeu de données. Les données en question sont issues d'expériences suivant le protocole P300-speller brièvement décrit plus haut, et concernent deux sujets différents, nommés sujet `A` et sujet `B`. Ces jeux de données seront étudiés avec le logiciel `R`.

Les données d'étude concernent deux sujets, appelés A et B. Pour chacun d'entre eux, deux séries de *runs* ont été enregistrées, appelées *Train* et *Test* (le premier destiné à être utilisé pour l'apprentissage des algorithmes, le second aux tests).
On notera $L = 240$ la longueur de chaque signal, $N_{\ell c} = 12$ le nombre de lignes et colonnes intensifiées et $N_r = 15$ le nombre de répétitions.

#### *Données d'apprentissage*

Pour chaque *run*, les résultats se trouvent dans un fichier, créé sous le logiciel R, donc le nom contient les éléments nécessaires à l'identification (par exemple, `SubA_Train_run1.Rdata` contient les données relatives au premier *run* du sujet A, dans la phase d'apprentissage). Ce fichier contient les données suivantes

1. `sigs` : tableau numérique $N_{\ell c} \times N_r \times L$ contenant les signaux
2. `targ_letter` : caractère, la lettre cible
3. `targ_letter_coords` : vecteur numérique de longueur 2, contenant la position de la lettre cible dans la matrice (ligne et colonne)

3. http://www.bbci.de/competition/ii

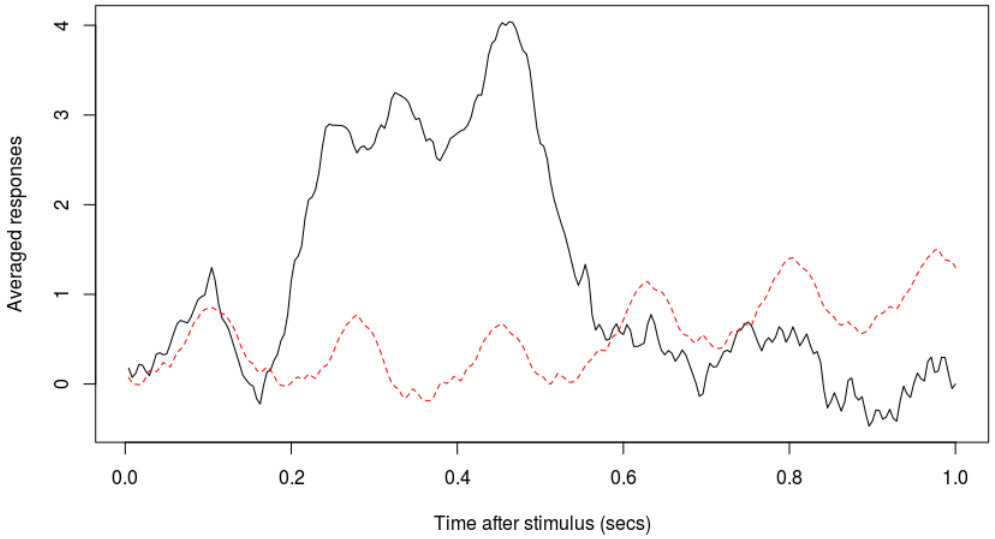


FIGURE 1.6 – Réponses moyennées sur les essais et les lettres (capteur Cz), dans une fenêtre d'une seconde après le stimulus dans le dispositif *P300 speller*. Noir/plein : caractère correct, rouge/tirets, caractère incorrect.

4. `targ_letter_ind` : vecteur numérique de longueur 2, contenant les indices de la lettre cible entre 1 et $N_{\ell c}$
5. `nontarg_letter_ind` : vecteur numérique de longueur $N_{\ell c} - 2$, contenant les indices de la lettre cible entre 1 et $N_{\ell c}$.

#### *Données de test*

Dans le cas des données de test, seuls les signaux sont fournis, le caractère cible n'étant pas censé être connu. Les caractères cible seront fournis ultérieurement.

### 1.2.4 Analyse préliminaire

Une première analyse préliminaire peut être effectuée, en calculant et traçant les signaux moyens :

— La moyenne sur tous les essais de tous les *runs* pour lesquels le caractère affiché coïncide avec le caractère « cible ».

— La moyenne sur tous les essais de tous les *runs* pour lesquels le caractère affiché ne coïncide pas avec le caractère « cible ».

Ces deux moyennes sont tracées en Figure 1.6 (en noir/trait plein le cas où le caractère coïncide avec la cible, en rouge/tirets le cas où la lettre ne coïncide pas avec la cible). On peut effectivement remarquer que lorsque le caractère correct a été présenté au sujet, le signal mesuré présente effectivement une "bosse" significative démarrant environ 300ms (en fait un peu avant) après le stimulus (qui correspond au temps $t = 0$), bosse qui n'est pas visible sur le tracé rouge pointillé. Ce dernier présente une espèce de périodicité, de période égale à 175ms, correspondant à la fréquence des *flashs*.

Le phénomène est bien moins spectaculaire lorsqu'au lieu de calculer les moyennes sur tous les essais et tous les caractères cible, on calcule la moyenne pour chaque caractère cible. Des exemples se trouvent dans la Figure 1.7, qui illustrent la diversité des situations.

— Les deux exemples sur la ligne du haut correspondent à des situations où la ligne ou la colonne illuminée contiennent bien le caractère cible. La bosse reste clairement visible sur le tracé de gauche, bien moins nettement sur le tracé de droite.

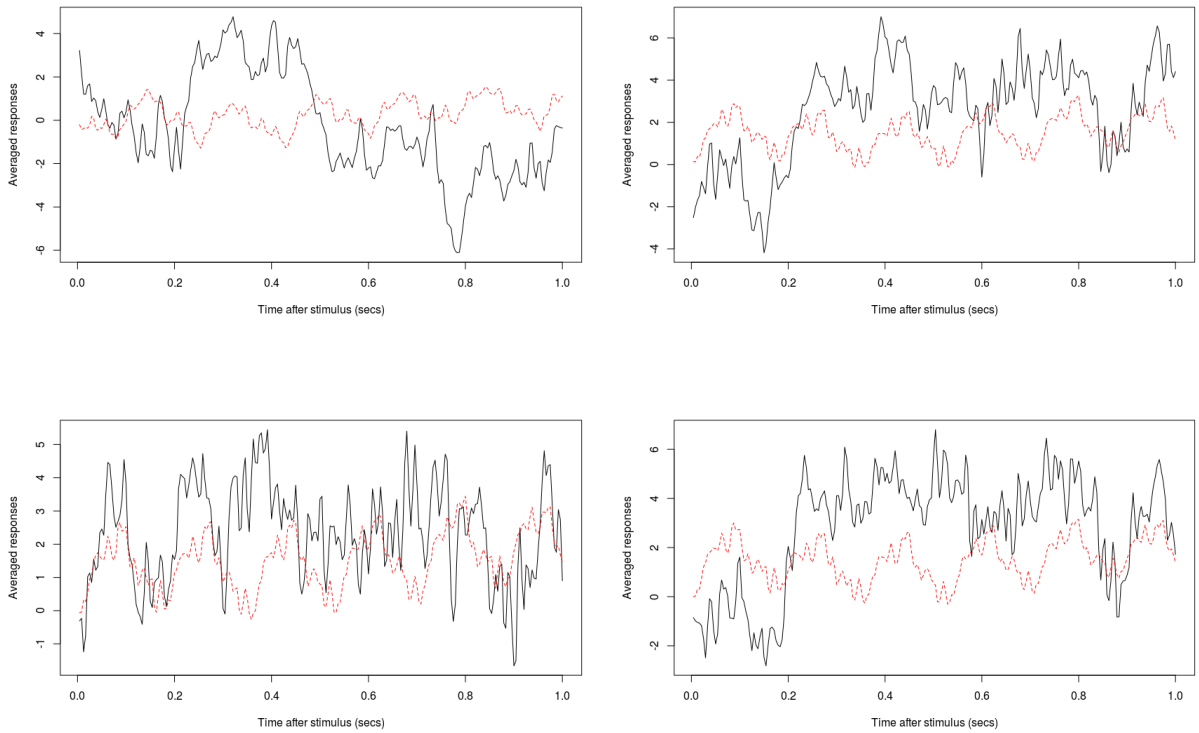


FIGURE 1.7 – Réponses moyennées sur les essais et les lettres (capteur Cz), dans une fenêtre d'une seconde après le stimulus dans le dispositif *P300 speller*. Noir/plein : caractère correct, rouge/tirets, caractère incorrect.

— Les deux exemples sur la ligne du bas correspondent à des situations où la ligne et la colonne illuminées ne contiennent pas le caractère cible. Il est difficile de trancher, même si le tracé de droite pourrait laisser penser (faussement) qu'on est dans un cas favorable...

On voit bien sur ces exemples que la tâche de décision est complexe, même sur des signaux moyennés. L'exploitation des signaux originaux, sans opération de moyenne est bien plus complexe encore.

## 1.3 Eléments de théorie du signal

### 1.3.1 Généralités

#### *Qu'est ce qu'un signal ?*

En théorie du signal, théorie de l'information et plus généralement en ingénierie, on appelle signal une grandeur dont la variation transporte une information, d'une source à une destination. La variation en question peut être une variation au cours du temps, une variation d'un point à un autre de l'espace, une variation entre les mesures de différents capteurs,... Le signal lui même peut être un courant électrique, une onde électromagnétique, une image, les bases d'une séquence d'ADN,... Concernant la façon dont l'information est encodée dans le signal, il peut s'agir de de son intensité, ses modulations d'amplitude, de fréquence,... l'extraction d'information à partir d'un signal est l'un des buts principaux du traitement du signal.

On distingue généralement deux catégories de signaux :

1. Les signaux à temps continu (si on s'intéresse aux variations au cours du temps, dans le cas contraire on pourra parler de signaux continus, ou signaux analogiques). La variable temps peut être prise soit sur l'axe réel tout entier $\mathbb{R}$, soit dans un intervalle fixé (pour lequel on prendra généralement $[0, 1]$). Les signaux analogiques sont généralement modélisés comme des fonctions

d'une (ou plusieurs dans certains cas) variable(s) réelle(s), ou fonctions aléatoires si l'on désire modéliser la variabilité.

2. Les signaux à temps discret (si on s'intéresse aux variations au cours du temps, dans le cas contraire on pourra parler de signaux discrets, ou signaux numériques). La variable temps est généralement prise dans $\mathbb{Z}$, ou dans un ensemble fini (ici on prendra généralement $\{1, 2, \dots L\}$). Les signaux numériques sont donc généralement modélisés par des suites.

**Remarque 1.1** Les signaux réels, ou physiques, sont généralement des signaux analogiques. Cependant les seuls signaux qu'un ordinateur soit capable de traiter sont des signaux discrets, qui plus est de longueur finie. L'opération qui transforme un signal analogique en signal numérique porte le nom d'échantillonnage. L'échantillonnage est une opération qui se traduit généralement par des pertes d'informations, qui peuvent toutefois être contrôlées sous des hypothèses appropriées. Dans ce cours, on ne traitera que des signaux numériques, de longueur finie.

Dans la vie réelle, la répétition d'une même expérience conduit très rarement à un résultat constant, il est donc nécessaire de pouvoir modéliser la variabilité sous-jacente. C'est pour cela que l'on est amené à travailler sur des modèles de signaux aléatoires, et pas seulement déterministes.

### *Qu'est ce que le traitement des signaux ?*

Le traitement du signal peut avoir de nombreuses finalités, parmi lesquelles on peut notamment citer
- L'analyse du signal, qui commence avec les problématiques de représentation et de visualisation, mais inclut par exemple aussi l'estimation de grandeurs à mesurer sur un signal.
- La synthèse de signaux à partir de modèles paramétriques ou non-paramétriques (par exemple, synthèse de la parole, imagerie 3D,...)
- Le codage, la compression du signal pour son stockage et sa transmission : les standards *mp3*, *aac* pour le son, *jpeg* pour les images, *mp4* pour la video en sont quelques exemples, mais on peut aussi mentionner tous les aspects liés à la sécurité (cryptage,...) et bien d'autres applications.
- L'amélioration de la qualité (restauration, par exemple défloutage, reconstitution de parties manquantes,...) selon des critères divers et variés (mathématiques, physiologiques,...), et toutes les variantes (modifications,...).
- Les problématiques de détection et décision ; c'est dans ce cadre qu'on se placera ici.

Le traitement des signaux fait appel à de nombreux domaines des mathématiques, notamment l'algèbre linéaire et multilinéaire, l'analyse (analyse harmonique, analyse complexe, analyse fonctionnelle), l'analyse numérique, l'arithmétique, les probabilités, les statistiques,...). Dans ce cours on se focalisera principalement sur l'algèbre linéaire, l'analyse Hilbertienne et les probabilités et statistiques.

### 1.3.2 Représentation des signaux, aspects algébriques

Il est utile de se rafraîchir la mémoire avec quelques notions de base. Le cadre général dans lequel on travaillera est celui des espaces vectoriels sur $\mathbb{R}$ ou $\mathbb{C}$, généralement de dimension finie, plus rarement de dimension infinie. On rappelle ci-dessous les principales notions dans le cas complexe, d'où le cas réel peut facilement être déduit.
On rappelle qu'une forme sesquilinéaire sur un espace vectoriel complexe $E$ est une application $\varphi : E \times E \longmapsto \mathbb{C}$ , $(x, y) \longmapsto \varphi(x, y)$ telle que pour tout $x \in E$ l'application $y \to \varphi(x, y)$ est anti-linéaire (c'est à dire $\varphi(x, \lambda y + \mu z) = \overline{\lambda}\varphi(x, y) + \overline{\mu}\varphi(x, z)$) et pour tout $y \in E$ l'application $x \to \varphi(x, y)$ est linéaire (c'est à dire $\varphi(\lambda x + \mu z, y) = \lambda\varphi(x, y) + \mu\varphi(z, y)$).
On dit que $\varphi$ est :
- Hermitienne si $\varphi(x, y) = \overline{\varphi}(y, x)$ pour tous $x, y \in E$ ;
- Positive si $\varphi(x, x) \geq 0$ pour tout $x \in E$ ;
- Définie si pour $x \in E$ l'égalité $\varphi(x, x) = 0$ équivaut à $x = 0$.

**Définition 1.1** *Un produit Hermitien sur un espace vectoriel complexe $E$ est une forme sesquilinéaire hermitienne définie positive.*

On note en général $(x, y) \to \langle x, y\rangle$ un tel produit Hermitien (on parle aussi de produit scalaire).

**Définition 1.2** *Un espace préhilbertien est un espace vectoriel muni d'un produit Hermitien. Un espace de Hilbert est un espace pré-hilbertien complet pour la norme $x \to \|x\| = \sqrt{\langle x, x\rangle}$.*

Dans un espace de Hilbert $E$, la norme suffit à définir le produit Hermitien via l'identité de polarisation : $\forall x, y \in E$,

$$\langle x, y\rangle = \frac{1}{4}\Big(\|x+y\|^2 - \|x-y\|^2 + i\|x+iy\|^2 - i\|x-iy\|^2\Big) \ . \tag{1.1}$$

Rappelons aussi l'inégalité de Cauchy-Schwarz : pour tous $x, y$,

$$|\langle x, y\rangle| \leq \|x\|\,\|y\| \ , \tag{1.2}$$

de sorte qu'il existe un réel $\theta \in [-\pi, \pi]$ tel que

$$|\cos(\theta)| = \frac{|\langle x, y\rangle|}{\|x\|\,\|y\|} \ ,$$

c'est à dire $\theta$ est une mesure de l'angle entre $x$ et $y$. Si $\theta = \pm\pi/2$, $x$ et $y$ sont orthogonaux.

**Définition 1.3** *On dit que deux vecteurs $x, y \in E$ sont orthogonaux si $\langle x, y\rangle = 0$. On dit qu'ils sont perpendiculaires si $\mathfrak{Re}(\langle x, y\rangle) = 0$.*

Dans un espace réel les notions d'orthogonalité et perpendicularité coïncident, ça n'est pas le cas dans un espace complexe. Notons que dans ce cas, les vecteurs $x$ et $ix$ sont perpendiculaires (mais pas orthogonaux).

**Théorème 1.1 (Pythagore)** *Les vecteurs $x$ et $y$ sont orthogonaux dans $E$ si et seulement si $\|x + y\|^2 = \|x\|^2 + \|y\|^2$.*

Les notions importantes dont nous aurons besoin sont les notions d'orthogonal d'un sous-espace, et de projection orthogonale.

**Définition 1.4** *L'orthogonal d'une partie non vide $X \subset E$ est l'ensemble :*

$$X^\perp = \{y \in E : \forall x \in X, \langle x, y\rangle = 0\} \ . \tag{1.3}$$

Il est facile de vérifier que $X^\perp$ est un sous espace vectoriel de $E$.

**Définition 1.5** *On appelle famille orthogonale dans $E$ toute famille $\{e_n,\ n = 0, 1, 2, \ldots\}$ de vecteurs de $E$ telle que $\langle e_m, e_n\rangle = 0$ pour tous $m \neq n$. Si de plus $\|e_n\| = 1$ pour tout $n$, cette famille est orthonormée ou orthonormale.*

**Définition 1.6 (Base Hilbertienne)** *Soit $E$ un espace pré-Hilbertien. Une famille orthonormale $\{e^n,\ n = 1, 2, \ldots\}$ est une base Hilbertienne de $E$ si elle est complète, dans le sens suivant : pour tout $x \in E$, il existe une famille de scalaires $\{a_n, n = 0, \ldots\}$ telle que*

$$\sum_n a_n e^n = x \ .$$

*où la sommabilité de la série dans le membre de gauche est associée à la norme. Dans le cas où la base est infinie dénombrable, l'égalité ci-dessus signifie $\lim_{N\to\infty} \left\|x - \sum_{n=1}^N a_n e^n\right\| = 0$*

Notons que dans cette définition, l'unicité de la famille de coefficients $a_n$ est assurée par l'orthonormalité de la famille.

**Théorème 1.2** *Une famille orthogonale de vecteurs non nuls de $E$ est libre. Elle est une base orthogonale du sous-espace $F$ de $E$ qu'elle engendre.*

Dans ce cas, en notant comme plus haut $\{e_n,\ n = 0, 1, 2, \ldots\}$ la famille orthogonale, et en la supposant normée, on a aussi la formule de Parseval (généralisation du théorème de Pythagore) : pour tout $x \in F$

$$\|x\|^2 = \sum_n |\langle x, e_n\rangle|^2 \ . \tag{1.4}$$

Une notion centrale dont nous aurons besoin est la notion de projection orthogonale sur un sous-espace.

**Théorème 1.3 (projection orthogonale)** *Soit $F$ un sous espace vectoriel de dimension finie de $E$. Pour tout $x \in E$, il existe un unique $y \in F$ à distance (induite par la norme) minimale de $F$, c'est à dire tel que*

$$\|x - y\| = d(x, F) = \inf_{z \in F} \|x - z\|$$

*$y$ est également l'unique vecteur de $F$ tel que $x - y \in F^\perp$ : $\langle x - y, f\rangle = 0$ pour tout $f \in F$.*

Nous verrons plus loin l'expression du projeté orthogonal lorsqu'une base de $F$ est connue.

### 1.3.3 Quelques exemples d'espaces de Hilbert utiles en traitement du signal

On va considérer ci-dessous quelques exemples d'espaces modèles, tous de même dimension, qui peuvent être utiles dans des situations de codage de signaux.
Les exemples que nous manipulerons principalement sont $\mathbb{R}^L$ et $\mathbb{C}^L$, on pourra aussi être amenés à utiliser des espaces de Lebesgue, notamment $L^2([a, b])$ ou $\ell^2(\mathbb{Z})$.

— $E = \mathbb{C}^L$ est muni d'une structure d'espace de Hilbert par le produit Hermitien

$$\langle x, y\rangle = \sum_{\ell=1}^{L} x_\ell \overline{y}_\ell \ , \qquad x, y \in \mathbb{C}^L \ . \tag{1.5}$$

Ici on a noté $(x_1, \ldots x_L)$ les coordonnées de $x \in \mathbb{C}^L$. On utilisera préférentiellement la notation matricielle, en notant génériquement $x = (x_1, \ldots x_L)^T$ la matrice colonne constituée des coordonnées de $x$. Avec ces notations on montre facilement que le produit scalaire prend la forme

$$\langle x, y\rangle = y^* x \ , \tag{1.6}$$

où on note $y^* = \overline{y}^T$ le conjugué Hermitien ou adjoint de $y$ (c'est à dire le transposé de son complexe conjugué, qui est donc une matrice "ligne").

— $E = \mathbb{R}^L$ est lui aussi un espace de Hilbert, avec le même produit scalaire que $\mathbb{C}^L$, où les complexes conjugués peuvent être ignorés car toutes les coordonnées sont réelles. Dans $\mathbb{R}^L$, la forme matricielle du produit scalaire devient $\langle x, y\rangle = y^T x$.

— L'espace $\mathcal{M}_N(\mathbb{C})$ des matrices $N \times N$ à coefficients complexes est lui aussi muni d'une structure d'espace de Hilbert grâce au produit Hermitien

$$\langle A, B\rangle = \mathsf{Tr}(B^* A) = \mathsf{Tr}(A B^*) \ , \qquad A, B \in \mathcal{M}_N(\mathbb{C}) \ , \tag{1.7}$$

où $\mathsf{Tr}$ représente la trace (somme des éléments diagonaux), et où $B^* = \overline{B}^T$ est la conjuguée Hermitienne (oui matrice adjointe) de $B$.

— Les espaces $L^2([a,b])$ définis par

$$L^2([a,b]) = \left\{ x : [a,b] \to \mathbb{C} : \ \|x\|^2 := \int_a^b |x(t)|^2 \, dt < \infty \right\} , \tag{1.8}$$

où $a < b$ sont deux réels, sont munis d'une structure d'espace de Hilbert grâce au produit Hermitien

$$\langle x, y \rangle = \int_a^b x(t)\overline{y}(t) \, dt \ . \tag{1.9}$$

— L'espace $\ell^2(\mathbb{Z})$ des suite de module carré sommable

$$\ell^2(\mathbb{Z}) = \left\{ u : \mathbb{Z} \to \mathbb{C} : \quad \sum_{n=-\infty}^{\infty} |u_n|^2 < \infty \right\} \tag{1.10}$$

est muni d'une structure d'espace de Hilbert par le produit Hermitien

$$\langle u, v \rangle = \sum_{n=-\infty}^{\infty} u_n \overline{v}_n \ , \qquad u, v \in \ell^2(\mathbb{Z}) \ . \tag{1.11}$$

On définit de même $\ell^2(\mathbb{N})$.

Plus généralement, tout sous-espace vectoriel fermé d'un espace de Hilbert, muni de la restriction du produit hermitien, est lui même un espace de Hilbert.

### 1.3.4 Approximation des signaux

Les signaux auxquels l'on s'intéresse n'ont aucune raison d'appartenir à ces espaces particuliers, on va donc les approximer par des éléments de ces espaces. Pour ce faire, le plus simple est de considérer la projection orthogonale sur ces sous-espaces modèle.

Le résultat qui suit est un corollaire du théorème de Gram-Schmidt, qui stipule qu'à toute famille libre de vecteurs dans un espace pré-Hilbertien, on peut associer une famille orthonormée qui engendre le même sous-espace, et en constitue donc une base orthonormée. La construction de cette base orthonormée est la procédure d'orthonormalisation de Gram-Schmidt.

**Théorème 1.4** *Soit $E$ un espace pré-Hilbertien, soit $F$ un sous espace de $E$. Si $F$ est de dimension finie, ou infinie-dénombrable, alors $F$ admet une base orthonormée.*

Dans ces conditions, notons $\mathcal{B} = \{e^1, e^2, e^3, \ldots\}$ une telle base orthonormée. La meilleure approximation de $x \in E$ par un élément de $F$, au sens de la distance induite par la norme de $E$ est donnée par

$$y = \Pi_F(x) = \sum_n \langle x, e^n \rangle e^n \ . \tag{1.12}$$

et l'erreur d'approximation est donnée par :

$$\|x - y\|^2 = \|x\|^2 - \|y\|^2 = \|x\|^2 - \sum_n |\langle x, e^n \rangle|^2 \ . \tag{1.13}$$

**Exemple 1.1 (Approximation constante par morceaux)** *On considère $E = \mathbb{C}^L$, et on suppose que $L$ est divisible par un entier $N > 1$ : on note $L = KN$. Pour $n = 1, \dots N$ introduit les intervalles $I_n = [\![K(n-1), Kn]\!]$ et les vecteurs $h^n \in \mathbb{C}^L$, définis par*

$$h^n_\ell = \begin{cases} \frac{1}{\sqrt{K}} & si \qquad \ell \in I_n \\ 0 & sinon \end{cases}$$

*On vérifie facilement que la famille $\{h^1, \dots h^N\}$ est orthonormée. Le projeté orthogonal $\Pi_F(x)$ d'un vecteur $x \in \mathbb{C}^L$ quelconque sur le sous-espace $F \subset \mathbb{C}^L$ engendré par cette famille est donc de la forme*

$$\Pi_F(x) = \sum_{n=1}^{N} \langle x, h^n \rangle h^n \ .$$

*Or on peut remarquer que $h^n$ est, à un facteur $1/\sqrt{K}$ près, l'indicatrice $1_{I_n}$ de l'intervalle $I_n$, et que $\langle x, h^n \rangle$ est à un facteur $\sqrt{K}$ près la moyenne de $x$ sur l'intervalle $I_n$ :*

$$\langle x, h^n \rangle = \sqrt{K}\, \overline{x}_n \ , \quad avec \quad \overline{x}_n = \frac{1}{K} \sum_{\ell \in I_n} x_\ell \ .$$

*Ainsi cette approximation prend la forme d'une approximation constante par morceaux*

$$\Pi_F(x) = \sum_{n=1}^{N} \overline{x}_n 1_{I_n} \ .$$

*Un exemple d'approximation constante par morceaux d'un signal "compliqué" se trouve en* Figure *1.8 (tracés de gauche).*

**Exercice 1.1 (Approximation constante par morceaux)** *Démontrer les résultats de l'exemple 1.1.*

**Exemple 1.2 (Approximation à bande limitée)** *$k = 1, \dots L$ on introduit le vecteur $\epsilon^k \in \mathbb{C}^L$ défini par ses coordonnées dans la base canonique*

$$\epsilon^k_\ell = \frac{1}{\sqrt{L}} e^{2i\pi k\ell/L} \ , \qquad \ell = 1, \dots L \ . \tag{1.14}$$

*Il est possible de démontrer que la famille $\{\epsilon^k,\ k = 1, \dots L\}$ est une base orthonormée de $\mathbb{C}^L$. Le paramètre $k$, appelé fréquence, décrit la vitesse de variations de ces vecteurs (on pourra en tracer quelques uns pour se convaincre). Notons que $\epsilon_L = 1/\sqrt{L}$ est un vecteur constant.*

*Etant donné un entier $N < L/2$, on considère le sous-espace $F_N \subset \mathbb{C}^L$ engendré par la famille $\mathcal{F}_N = \{\epsilon^L, \epsilon^1, \epsilon^{L-1}, \epsilon^2, \epsilon^{L-1} \dots \epsilon^N, \epsilon^{L-N}\}$. $F_N$ est de dimension $2N + 1$.*

*Il est possible de montrer que $\mathcal{F}_N$ est une famille orthonormée, ainsi le projeté orthogonal de $x \in \mathbb{C}^N$ est donné par*

$$\Pi_{F_N}(x) = \langle x, f^L \rangle f^L + \sum_{n=1}^{N} \left( \langle x, f^n \rangle f^n + \langle x, f^{L-n} \rangle f^{L-n} \right) \tag{1.15}$$

*Un exemple d'approximation "passe-bas" d'un signal "compliqué" se trouve en* Figure *1.8 (droite).*

**Exercice 1.2 (Approximation à bande limitée)**

1. *Montrer que la famille $\{\epsilon^\ell, \ell = 1 \dots L\}$ est une base orthonormée de $\mathbb{C}^L$ (commencer par montrer l'orthogonalité, on pourra utiliser l'expression de la somme partielle d'une série géométrique).*
2. *Dans l'expression de $\Pi_{F_N}(x)$ ci-dessus, comment interprète-t-on le premier terme $\langle x, f^L \rangle$ ?*

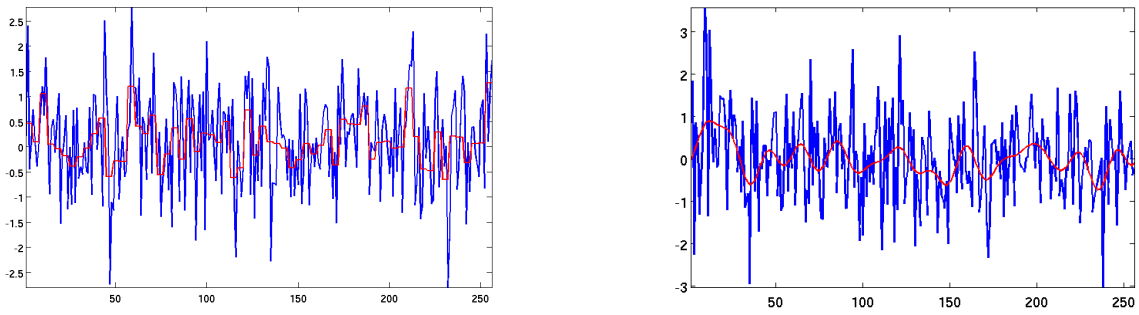

FIGURE 1.8 – Deux types d'approximations (rouge) d'un signal (bleu) : constante par morceaux (à gauche), et à bande limitée (droit).

3. *On note* $\hat{x}_k = \frac{1}{\sqrt{L}}\langle x, \epsilon^k\rangle$. *Expliciter* $\hat{x}_k$ *en fonction des* $x_\ell$. *La transformation linéaire* $x \in \mathbb{C}^L \to \hat{x} \in \mathbb{C}^L$ *est appelée transformation de Fourier finie. Dire pourquoi elle est inversible, et expliciter la transformée inverse.*

Plus généralement, il est possible de relaxer les hypothèses. On se limitera ici au cas de sous-espaces de dimension finie. Étant donnée une famille de vecteurs $f^1, \dots f^N$, la *matrice de Gram* de cette famille est définie par

$$G = \{G_{mn},\ m, n = 1, \dots N\}\ , \qquad G_{mn} = \langle f^n, f^m\rangle\ . \tag{1.16}$$

La matrice de Gram est un outil important d'algèbre linéaire, et a des propriétés importantes, par exemple

**Lemme 1.1** *Soit* $\mathcal{F} = \{f^1, \dots f^N\}$ *une famille d'éléments de* $E$.

1. *La matrice de Gram est auto-adjointe (c'est à dire telle que* $G^* = G$*), et semi-définie positive : pour tous* $\alpha_1, \dots \alpha_n \in \mathbb{C}$ *et* $m_1, \dots m_n$,

$$\sum_{k,\ell=1}^{n} \alpha_k \overline{\alpha}_\ell G_{m_k m_\ell} \geq 0\ .$$

2. *La matrice de Gram est inversible si et seulement si la famille* $\{f^1, \dots f^N\}$ *est linéairement indépendante. Dans ce cas la matrice inverse* $H = G^{-1}$ *est elle aussi auto-adjointe.*

Donc en particulier, une famille finie de vecteurs est libre si et seulement le déterminant de la matrice de Gram (appelé déterminant de Gram) est non-nul. Notons aussi qu'en conséquence de ces propriétés, la matrice de Gram est diagonalisable, ses valeurs propres sont réelles, et positives ou nulles.
Plus généralement, on montre le résultat suivant

**Proposition 1.1** *Soit* $F$ *un sous-espace de* $E$ *engendré par la famille de vecteurs* $f^1, \dots f^N$. *Supposons que* $G$ *soit inversible. Alors la projection orthogonale de* $E$ *sur* $F$ *s'écrit sous la forme*

$$E \ni x \to y = \Pi_F(x) = \sum_{n=1}^{N} a_n f^n\ , \tag{1.17}$$

*où les coefficients* $a_n$ *sont solutions du problème linéaire*

$$\sum_{n=1}^{N} G_{mn} a_n = \langle x, f^m\rangle\ , \qquad m = 1, \dots N\ . \tag{1.18}$$

*Ces coefficients s'écrivent également*

$$a_n = \langle x, \tilde{f}^n \rangle \ , \tag{1.19}$$

*où* $\{\tilde{f}^1, \dots \tilde{f}^N\}$ *est la* famille duale *de la famille des* $f^n$*, définie par*

$$\tilde{f}^n = \sum_{m=1}^{N} \overline{H}_{nm} f^m \ , \quad m = 1, \dots N \ , \tag{1.20}$$

*et les* $H_{nm}$ *sont les coefficients de la matrice* $H = G^{-1}$*. On a également*

$$\Pi_F(x) = \sum_{m=1}^{N} \langle x, f^m \rangle \tilde{f}^m \ . \tag{1.21}$$

Comme on l'a déjà vu, si la famille $\{f^1, \dots f^N\}$ est orthonormée, la matrice de Gram est égale à la matrice identité, on a $\tilde{f}^m = f^m$ pour tout $m$, et la projection orthogonale se simplifie en

$$\Pi_F(x) = \sum_{m=1}^{N} \langle x, f^m \rangle f^m \ .$$

## 1.4 Rappels sur les vecteurs aléatoires

En traitement des signaux, il est souvent fondamental de pouvoir décrire la variabilité dans une classe de signaux. L'outil adapté pour cela est la modélisation probabiliste, on modélise alors les signaux de dimension finie comme des vecteurs aléatoires, les signaux numériques infinis comme des suites aléatoires, et les signaux analogiques comme des fonctions aléatoires (ou processus aléatoires à temps continu). On ne traitera pas ces deux derniers cas ici.

### 1.4.1 Quelques définitions

Soit $(\Omega, \mathcal{A}, \mathbb{P})$ un espace probabilisé. Un vecteur aléatoire (de dimension $L$) est une application mesurable

$$X : (\Omega, \mathcal{A}, \mathbb{P}) \longmapsto (\mathbb{R}^L, \mathcal{B}_{\mathbb{R}^L}) \ .$$

La loi $\mathbb{P}_X$ du vecteur aléatoire $X$ est une mesure de probabilité sur $\mathbb{R}^L$ muni de sa tribu borélienne $\mathcal{B}_{\mathbb{R}^L}$, définie par

$$\mathbb{P}_X \{A\} = \mathbb{P}\{X \in A\} \ , \quad \forall A \in \mathcal{B}_{\mathbb{R}^L} \ . \tag{1.22}$$

Avec ces notations, la fonction de répartition de $X$ est la fonction $F = F_X$ de $L$ variables

$$F_X : (x_1, \dots x_L) \in \mathbb{R}^L \to F_X(x_1, \dots x_L) = \mathbb{P}\{X_1 < x_1, X_2 < x_2, \dots X_L < x_L\} \in [0, 1] \ . \tag{1.23}$$

On notera $\mathbb{P}_{X_\ell}$, $\ell = 1, \dots L$ les lois marginales des coordonnées de $X$.

**Définition 1.7** *Les coordonnées de* $X$ *sont indépendantes si pour tout* $k \leq L$*, pour tous* $\ell_1, \ell_2, \dots \ell_k \in [\![1, L]\!]$ *différents deux à deux, et* $\forall A = (A_1 \times A_2 \times \cdots \times A_k) \in \mathcal{B}_{\mathbb{R}^k}$*, où* $A_i \in \mathcal{B}_{\mathbb{R}}$ *pour tout* $i$*,*

$$\mathbb{P}\{X \in A_1 \times A_2 \times \cdots \times A_k\} = \mathbb{P}\{X_{\ell_1} \in A_1\} \mathbb{P}\{X_2 \in A_{\ell_2}\} \dots \mathbb{P}\{X_{\ell_k} \in A_k\}$$

Notons que si les coordonnées de $X$ sont indépendantes la fonction $F_X$ est égale au produit des fonctions de répartition $F_{X_n}$ des coordonnées : $F_X(x_1, \dots x_L) = F_{X_1}(x_1) \dots F_{X_L}(x_L)$.

**Définition 1.8 (Espérance d'un vecteur aléatoire)** *Avec les notations ci-dessus, supposons que l'espérance* $\mathbb{E}\{X_\ell\}$ *des coordonnées de* $X$ *existent. Alors l'espérance de* $X = (X_1, X_2, \dots X_L)^T$ *est définie par*

$$\mathbb{E}\{X\} = (\mathbb{E}\{X_1\}, \mathbb{E}\{X_2\}, \dots, \mathbb{E}\{X_L\})^T \ .$$

**Remarque 1.2 (Linéarité de l'espérance)** *La linéarité de l'espérance d'une variable aléatoire se transporte sur les vecteurs aléatoires. Avec les notations ci-dessus, soit $M$ un entier positif, pour tout $B \in \mathbb{R}^M$ et pour toute matrice $A \in \mathcal{M}_{M,N}(\mathbb{R})$, on montre facilement que*

$$\mathbb{E}\{AX + B\} = A\mathbb{E}\{X\} + B \ . \tag{1.24}$$

**Exercice 1.3 (Linéarité de l'espérance)** *Le démontrer.*

**Remarque 1.3** *On définit de même l'espérance d'une matrice aléatoire $A = \{A_{mn}\}$, qui est elle même une matrice de même taille, définie élément par élément par $(\mathbb{E}\{A\})_{mn} = \mathbb{E}\{A_{mn}\}$. Ceci nous sera utile pour obtenir certaines propriétés des matrices de covariance. Par exemple, notons qu'avec les notations précédentes, $XX^T$ est une matrice aléatoire carrée de taille $L \times L$, et la matrice $\mathbb{E}\{XX^T\}$ est en fait la matrice des moments d'ordre deux $\mathbb{E}\{X_k X_\ell\}$, lorsqu'ils existent.*

**Définition 1.9 (Fonction caractéristique)** *Soit $X$ un vecteur aléatoire de dimension $L$. Sa fonction caractéristique est la fonction $\phi_X : \mathbb{R}^L \to \mathbb{C}$ définie par*

$$\phi_X(u) = \mathbb{E}\left\{e^{iu^T X}\right\} = \mathbb{E}\left\{e^{i\sum_{\ell=1}^L u_\ell X_\ell}\right\} \ .$$

La fonction caractéristique est bornée, de module inférieur ou égal à 1, continue, et $\Phi_X(0) = 1$. Elle peut être utilisée pour calculer les moments de $X$ lorsqu'ils existent. Par exemple, si $\mathbb{E}\{X_\ell\}$ existe pour tout $\ell$, alors $\phi_X$ est de classe $C^1$, et

$$\mathbb{E}\{X_\ell\} = -i\frac{\partial \phi_X}{\partial x_\ell}(0, \dots 0) \ , \qquad \mathbb{E}\{X\} = -i\left[\nabla \phi_X(u)\right]_{u=0} \ ,$$

où $\nabla = (\partial/\partial u_1, \dots \partial/\partial u_L)^T$ est l'opérateur de gradient.
De même, si les moments d'ordre deux existent, alors $\Phi_X$ est de classe $C^2$ et on a

$$\mathbb{E}\{X_k X_\ell\} = (-i)^2 \frac{\partial^2 \Phi_X}{\partial u_k \partial u_\ell}(0, \dots 0) \ .$$

Le résultat suivant est la conséquence de la définition de l'indépendance et des propriétés de l'exponentielle.

**Théorème 1.5** *Si $X$ et $Y$ sont deux vecteurs aléatoires de même dimension, indépendants, alors $\Phi_{X+Y} = \Phi_X \Phi_Y$.*

### 1.4.2 Vecteurs aléatoires à densité

Ici la mesure de référence est la mesure de Lebesgue sur $\mathbb{R}^L$.

**Définition 1.10** *La loi du vecteur aléatoire $X$ de dimension $L$ est absolument continue s'il existe une fonction mesurable $\rho = \rho_X : (\mathbb{R}^L, \mathcal{B}_{\mathbb{R}^L}) \to (\mathbb{R}, \mathcal{B}_{\mathbb{R}})$, appelée densité de $X$, telle que*

1. *$\rho(x) \geq 0$ pour tout $x \in \mathbb{R}^L$.*
2. *$\int_{\mathbb{R}^L} \rho(x)dx$ existe et vaut 1,*
3. *Pour tout $A \in \mathcal{B}_{\mathbb{R}^L}$, on a*

$$\mathbb{P}\{X \in A\} = \int_A \rho(x)\, dx \ .$$

*On dira que $X$ est un vecteur aléatoire à densité.*

Dans ce cas, la densité peut s'obtenir à partir de la fonction de répartition :

$$\rho_X(x_1, x_2, \dots x_L) = \frac{\partial^L F_X(x_1, x_2, \dots x_L)}{\partial x_1 \partial x_2 \dots \partial x_L} \ . \tag{1.25}$$

La fonction caractéristique est quant à elle obtenue à partir de la densité par une transformation similaire à une transformation de Fourier

$$\phi_X(u) = \int_{\mathbb{R}^N} \rho_X(x) e^{iu^T x} \, dx \ , \qquad u \in \mathbb{R}^L \ . \tag{1.26}$$

**Exemple 1.3 (Variable aléatoire normale)** *Une variable aléatoire normale, de moyenne $\mu$ et variance $\sigma^2$ est définie par la densité*

$$\rho_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma^2}(x-\mu)^2\right] \ . \tag{1.27}$$

*Il n'existe pas de forme explicite pour sa fonction de répartition, qui s'exprime en fonction de la fonction d'erreur* erf*, définie par*

$$\mathsf{erf}(x) = \frac{1}{\sqrt{\pi}} \int_{-x}^{x} e^{-t^2} \, dt \ .$$

*Par contre, sa fonction caractéristique admet une expression explicite*

$$\phi_X(u) = \exp\left(i\mu u - \frac{\sigma^2 u^2}{2}\right) \ , \qquad u \in \mathbb{R} \ . \tag{1.28}$$

**Exercice 1.4 (Variable aléatoire gaussienne)**
1. *Exprimer la fonction de répartition d'une variable aléatoire gaussienne $\mathcal{N}(\mu, \sigma^2)$ en fonction de la fonction* erf*.*
2. *Démontrer que la fonction caractéristique est bien donnée par l'expression* (1.28)*.*

L'existence d'une densité peut se déduire de la fonction caractéristique, grâce au résultat suivant

**Théorème 1.6** *Soit $X$ un vecteur aléatoire de dimension $L$, de fonction caractéristique $\Phi_X$. Si $\phi_X \in L^1(\mathbb{R}^L)$, alors $X$ admet une densité $\rho_X$ donnée par*

$$\rho_X(x) = \frac{1}{(2\pi)^L} \int_{\mathbb{R}^L} \phi_X(u) e^{-iu^T x} \, du \ .$$

On montre que si $X$ est un vecteur aléatoire de dimension $L$ de densité $\rho_X$, alors ses coordonnées $X_\ell$ sont des variables aléatoires réelles, qui admettent des densités appelées densités marginales $\rho_{X_\ell}$ définies par

$$\rho_{X_\ell}(x_\ell) = \int_{\mathbb{R}^{L-1}} \rho_X(x_1, x_2, \dots x_L) \, dx_1 \dots dx_{\ell-1} dx_{\ell+1} \dots dx_L \ . \tag{1.29}$$

On montre également que dans ce cas, les coordonnées de $X$ sont indépendante si et seulement si

$$\rho_X(x_1, \dots x_L) = \prod_{\ell=1}^{L} \rho_{X_\ell}(x_\ell) \ . \tag{1.30}$$

On peut évidemment marginaliser par rapport à tout sous-ensemble des coordonnées, par exemple $\rho_{(X_1,X_2)}(x_1, x_2) = \int_{\mathbb{R}^{L-2}} \rho_X(x_1, x_2, \dots x_L) \, dx_3 \dots dx_L$.

**Exercice 1.5 (Marginales)** *Avec les notations ci-dessus, exprimer la fonction caractéristique $\phi_{X_\ell}$ de la coordonnée $X_\ell$ en fonction de $\phi_X$.*

### 1.4.3 Covariance

Dans le cas où les coordonnées ne sont pas indépendantes, la matrice de covariance (parfois appelée matrice de variance-covariance, sa diagonale contenant les variances des coordonnées) fournit un outil précieux. A partir de maintenant, on supposera que les vecteurs aléatoires considérés sont du second ordre, c'est à dire tels que $\mathbb{E}\left\{X_\ell^2\right\} < \infty$ pour tout $\ell$. Notons que ceci assure aussi l'existence de l'espérance $\mathbb{E}\{X\}$.

**Définition 1.11 (Matrice de covariance)** *Soit $X$ un vecteur aléatoire du second ordre, de dimension $L$. Sa matrice de covariance est la matrice $\Sigma = \Sigma_X \in \mathcal{M}_L(\mathbb{R})$ définie par ses éléments*

$$\Sigma_{mn} = \mathbb{E}\left\{(X_m - \mathbb{E}\{X_m\})(X_n - \mathbb{E}\{X_n\})\right\} .$$

Notons que la matrice de covariance peut s'écrire simplement sous forme matricielle

$$\Sigma_X = \mathbb{E}\left\{(X - \mathbb{E}\{X\})\,(X - \mathbb{E}\{X\})^T\right\} , \tag{1.31}$$

où l'espérance d'une matrice est définie de façon similaire à l'espérance d'un vecteur.
On voit facilement que $\Sigma$ est réelle et symétrique (et donc diagonalisable, admettant des valeurs propres réelles et une base orthonormée constituée de vecteurs propres). Notons que si les coordonnées de $X$ sont indépendantes, alors $\Sigma$ est diagonale. Notons aussi que la réciproque n'est pas vraie (sauf dans le cas de vecteurs gaussiens, comme on va le voir plus loin).
Par ailleurs, on a le résultat suivant

**Proposition 1.2** *La matrice de covariance d'un vecteur aléatoire du second ordre est semi-définie positive : pour tous $\alpha_1, \dots \alpha_L \in \mathbb{R}$, on a*

$$\sum_{k,\ell=1}^{L} \alpha_k \alpha_\ell \Sigma_{k\ell} \geq 0 .$$

On sait déjà que la matrice de covariance est réelle symétrique. Elle est donc diagonalisable, ses valeurs propres sont réelles et $\mathbb{R}^L$ admet une base orthonormée constituée de vecteurs propres de $\Sigma$. On peut aussi déduire de la proposition 1.2 que les valeurs propres de $\Sigma$ sont positives ou nulles.

**Remarque 1.4 (Matrice de covariance et transformation linéaire)** *Soit $X$ un vecteur aléatoire de dimension $L$, de matrice de covariance $\Sigma_X$. Soient $M$ un entier positif, $B \in \mathbb{R}^M$ et $A \in \mathcal{M}_{M,L}(\mathbb{R})$. Alors le vecteur aléatoire $Y = AX + B$ admet pour matrice de covariance*

$$\Sigma_Y = A\Sigma_X A^T . \tag{1.32}$$

### 1.4.4 Vecteurs gaussiens

Rappelons qu'une variable aléatoire $X$ est distribuée suivant une loi normale, de moyenne $\mu$ et variance $\sigma^2$ si sa fonction caractéristique est donnée par l'expression (1.28) (voir exemple 1.3).

**Définition 1.12 (Vecteur gaussien)** *Un vecteur aléatoire $X = (X_1, \dots, X_L)^T$ est un vecteur aléatoire gaussien si toute combinaison linéaire à coefficients réels de ses coordonnées $X_1, \dots, X_L$ suit une loi normale (dont la variance peut éventuellement être nulle).*

En particulier, on en déduit que les coordonnées d'un vecteur gaussien sont des variables aléatoires gaussiennes. Notons que la réciproque est fausse, la normalité des coordonnées n'implique pas la gaussianité du vecteur.

**Exemple 1.4** *Soit $Y$ une variable aléatoire normale centrée réduite (c'est à dire $\mu = 0$ et $\sigma = 1$). Soit $Z = BY$, où $B \sim \mathcal{B}(1/2)$ est une variable aléatoire de Bernoulli, valant 1 et -1 avec probabilités égales à 1/2. Alors on montre que $Y$ suit une loi normale centrée réduite. Par contre le couple $X = (Y, Z)^T$ n'est pas un vecteur gaussien. Qui plus est, $Y$ est décorrélée (mais pas indépendante) de $X$.*

**Exercice 1.6 (Bernoulli fois gaussien)** *Démontrer les propriétés énoncées dans l'exemple 1.4.*

On a toutefois la propriété suivante

**Proposition 1.3** *Soit $X = (X_1, \ldots, X_L)^T$ un vecteur aléatoire gaussien de matrice de covariance $\Sigma_X$. Les variables aléatoires $X_1, \ldots, X_L$ sont indépendantes si et seulement si la matrice $\Sigma_X$ est diagonale.*

La fonction caractéristique d'un vecteur aléatoire $X$ gaussien, de moyenne $\mu \in \mathbb{R}^L$ et covariance $\Sigma \in \mathcal{M}_L(\mathbb{R})$ est donnée par

$$\phi_X(u) = \exp\left(i\mu^T u - \frac{1}{2}u^T \Sigma u\right) \quad , \quad u \in \mathbb{R}^L \ . \tag{1.33}$$

**Exercice 1.7 (Vecteur aléatoire gaussien)** *Retrouver à partir de la fonction caractéristique la moyenne et la covariance de $X$.*

**Exemple 1.5 (Un vecteur gaussien 2D à densité)** *On considère un vecteur gaussien $X$ de dimension 2, de moyenne $\mu$ et de matrice de covariance $\Sigma$ données par*

$$\mu = \begin{pmatrix} 1 \\ 2 \end{pmatrix} \ , \qquad \Sigma = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix} \ .$$

*On voit facilement que $\det(\Sigma) = 1$, donc $\Sigma$ est inversible, et $X$ admet une densité $\rho_X$. Pour l'obtenir on doit calculer*

$$\Sigma^{-1} = \begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix}$$

*de sorte que la densité vaut*

$$\begin{aligned} \rho_X(x_1, x_2) &= \frac{1}{2\pi} \exp\left[-\frac{1}{2}(x_1 - 1, x_2 - 2) \begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix} \begin{pmatrix} x_1 - 1 \\ x_2 - 2 \end{pmatrix}\right] \\ &= \frac{1}{2\pi} \exp\left[-\frac{1}{2}\left((x_1 - 1)^2 - 2(x_1 - 1)(x_2 - 2) + (x_2 - 2)^2\right)\right] \end{aligned}$$

*Les ensembles d'équiprobabilité sont les lignes de niveau de la densité, qui sont aussi les lignes de niveau de la fonction quadratique $(x_1, x_2) \to (x_1 - 1)^2 - 2(x_1 - 1)(x_2 - 2) + (x_2 - 2)^2$. Ce sont des ellipses, de centre $\mu = (1, 2)^T$, et dont on peut montrer que les axes sont dirigés par les vecteurs propres de $\Sigma$. Les lignes de niveau de la densité sont représentées en* Figure *1.9.*

*Quant à la fonction caractéristique, elle est donnée par*

$$\phi_X(u_1, u_2) = \exp\left[i(u_1 + 2u_2) - \frac{1}{2}\left(2u_1^2 + 2u_1u_2 + u_2^2\right)\right] \ .$$

Plus généralement, étant donné un vecteur gaussien de dimension $L$ dont la matrice de covariance n'est pas dégénérée, on montre que ses ensembles d'équiprobabilité (des hyper-surfaces de niveau) sont des (hyper) ellipsoïdes dans $\mathbb{R}^L$, centrées sur la moyenne du vecteur gaussien, et dont les axes sont dirigés par les vecteurs propres de la matrice de covariance. Les valeurs propres correspondantes donnent une mesure de l'épaisseur de l'ellipsoïde dans cette direction.

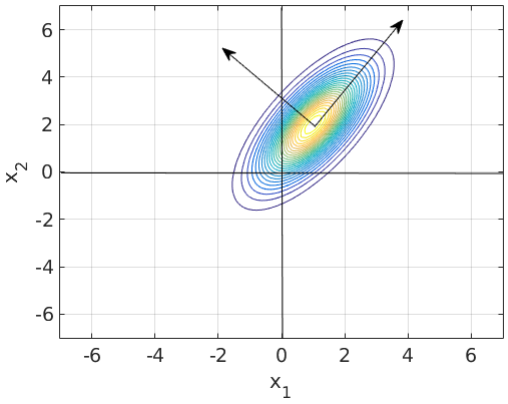


FIGURE 1.9 – Lignes de niveau de la densité $\rho_X$ de l'exemple 1.5. Elles prennent la forme d'ellipsoïdes centrées sur la moyenne, dont les axes sont dirigés par les deux vecteurs propres de la matrice de covariance (représentés par des flèches).

**Exemple 1.6 (Un vecteur gaussien 2D sans densité)** *Soit maintenant $X$ un vecteur gaussien de dimension 2, de moyenne $\mu$ et de matrice de covariance $\Sigma$ données par*

$$\mu = \begin{pmatrix} 1 \\ 2 \end{pmatrix} , \qquad \Sigma = \begin{pmatrix} 4 & 2 \\ 2 & 1 \end{pmatrix} .$$

*On voit facilement que $\mathsf{det}(\Sigma) = 0$, donc $\Sigma$ n'est pas inversible, et $X$ n'admet donc pas de densité.*
*Sa fonction caractéristique $\Phi_X$ prend la forme*

$$\phi_X(u_1, u_2) = \exp\left[ i(u_1 + 2u_2) - \frac{1}{2}\left(4u_1^2 + 4u_1u_2 + u_2^2\right)\right] .$$

*En particulier, le noyau de $\Sigma$ est constitué des vecteurs de la forme $(a, -2a)$ avec $a \in \mathbb{R}$, et on a*

$$\phi_X(a, -2a) = \exp(-3ia) .$$

*Dans cette direction, $\phi_X$ prend des valeurs de module égal à 1, et n'est donc pas absolument sommable, il ne peut donc pas exister de densité.*

Le résultat ci-dessous donne la forme générale de la densité d'un vecteur gaussien, lorsqu'elle existe.

**Proposition 1.4** *Soit $X = (X_1, \ldots, X_L)^T$ un vecteur aléatoire gaussien d'espérance $\mu \in \mathbb{R}^L$ et de matrice de covariance $\Sigma$. Si $\Sigma$ est définie positive (c'est à dire de déterminant $\mathsf{det}(\Sigma) > 0$), alors $X$ admet pour densité la fonction $\rho_X$ définie par*

$$\rho_X(x) = \frac{1}{(2\pi)^{L/2}} \frac{1}{\sqrt{\mathsf{det}(\Sigma)}} \exp\left\{ -\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu) \right\} , \qquad x \in \mathbb{R}^L .$$

**Exercice 1.8 (Vecteur gaussien sans densité)** *Le démontrer. On pourra remarquer que $\Phi_X \in L^1(\mathbb{R}^L)$ et utiliser le théorème 1.6. On pourra aussi utiliser le fait que $\int_{-\infty}^{\infty} e^{-ax^2}\, dx = \sqrt{\pi/a}$.*

Chapitre

# 2 Une approche par classification supervisée

On a vu au chapitre précédent le problème de détection de P300 qui nous intéresse, ainsi que quelques compléments mathématiques qui nous ont permis d'attaquer le problème par une approche de type "traitement du signal" (voir TD1 et TP1, où les observations sont modélisées sous la forme "bruit seul" ou "signal + bruit").
On va maintenant s'intéresser à une approche dans laquelle les aspects "traitement du signal" sont essentiellement absents, et où le problème est formulé comme un problème de décision statistique. Il s'agit d'*apprendre* un certain nombre de caractéristiques statistiques des signaux *P300* et *non-P300* à partir d'un jeu de données d'apprentissage, et de les exploiter pour la construction d'un détecteur, qui est finalement testé et validé sur un autre jeu de données de test.
Pour débuter ce chapitre, on commencera par s'intéresser à l'apprentissage des caractéristiques statistiques en question. La théorie statistique à laquelle fait appel cette étape est appelée théorie de l'estimation, et on en donne tout d'abord quelques éléments.

## 2.1 Eléments sur l'estimation

### 2.1.1 Généralités, exemples

Le problème d'estimation peut s'énoncer comme suit. Supposons que l'on observe $N$ valeurs $\underline{x} = (x^1, \ldots, x^N)$ résultant de tirages indépendants d'une variable aléatoire $X$. Celle-ci a une loi de probabilité qui dépend d'un paramètre ou un ensemble de paramètres noté génériquement $\theta$, considérés comme étant déterministes. Par exemple, si la loi de $X$ admet une densité, on note celle-ci $\rho(x;\theta)$.
Une *estimation* de $\theta$ est une fonction mesurable $T(\underline{x}) = T(x^1, \ldots x^N)$ des observations. La valeur de l'estimée dépend de la réalisation $\underline{x}$.

**Exemple 2.1 (Moyenne et variance empirique)** *1. Supposons que $X$ soit une variable aléatoire de moyenne $\mu$ inconnue. Une estimation pour $\mu$ est donné par la moyenne empirique*

$$m = \frac{1}{N}\sum_{n=1}^{N} x^n \ . \tag{2.1}$$

*2. Supposons que la variance $\sigma^2$ de $X$ soit également inconnue. Une estimation pour $\sigma^2$ est donné par la variance empirique*

$$s^2 = \frac{1}{N-1}\sum_{n=1}^{N}(x^n - m)^2 \ , \tag{2.2}$$

*qui utilise aussi la moyenne empirique définie en* (2.1).

Pour étudier les performances d'une méthode d'estimation, on modélise les valeurs observées $x^1, \ldots x^N$ comme des variables aléatoires indépendantes identiquement distribuées (on note cela i.i.d. pour simplifier) $X^1, \ldots X^N$, et on note $\underline{X} = (X^1, \ldots, X^N)$ le vecteur aléatoire dont une réalisation est $\underline{x}$. On

appellera estimateur du paramètre $\theta$ la variable aléatoire

$$\hat{\theta} = T(\underline{X}) = T(X^1, X^2, \dots X^N) \ .$$

On note parfois l'estimateur $\hat{\theta}_N$ pour souligner la dépendance dans la taille de l'échantillon. Un estimateur ne doit évidemment jamais dépendre du paramètre à estimer $\theta$ , il ne dépend que des observations. La qualité d'un estimateur se mesure souvent par deux quantités, son biais $\mathsf{B}$ (*bias* en anglais) et son erreur quadratique moyenne $\mathsf{MSE}$ (*mean square error* en anglais), définis par

$$\mathsf{B}(T(\underline{X})) = \mathbb{E}\left\{T(\underline{X}) - \theta\right\} \ , \qquad \mathsf{MSE}(T(\underline{X})) = \mathbb{E}\left\{(T(\underline{X}) - \theta)^2\right\} \ . \tag{2.3}$$

L'erreur en moyenne quadratique d'un estimateur est reliée au biais et à la variance de cet estimateur par la relation

$$\mathsf{MSE}(\hat{\theta}) = \mathsf{Var}\left\{\hat{\theta}\right\} - \mathsf{B}(\hat{\theta})^2 \ . \tag{2.4}$$

**Exercice 2.1** *Le démontrer, en utilisant la relation* $\hat{\theta} - \theta = \hat{\theta} - \mathbb{E}\left\{\hat{\theta}\right\} + \mathbb{E}\left\{\hat{\theta}\right\} - \theta$

**Exemple 2.2 (Propriétés de la moyenne empirique)** *On reprend les notations de l'exemple 2.1. Calculons*

$$\mathbb{E}\left\{\hat{\mu}\right\} = \frac{1}{N}\sum_{n=1}^{N} \mathbb{E}\left\{X^N\right\} = \mu \ ,$$

*où on a utilisé la linéarité de l'espérance et le fait que les $X^n$ sont i.i.d., de moyenne égale à $\mu$. Donc la moyenne empirique est un estimateur sans biais :* $\mathsf{B}(\hat{\mu}) = 0$.

*Concernant l'erreur quadratique moyenne, calculons*

$$\begin{aligned}
\mathbb{E}\left\{(\hat{\mu} - \mu)^2\right\} &= \mathbb{E}\left\{\frac{1}{N}\sum_{n=1}^{N}(X^n - \mu)^2\right\} \\
&= \frac{1}{N^2}\sum_{n=1}^{N}\left(\mathbb{E}\left\{(X^n)^2\right\} - 2\mu\mathbb{E}\left\{X^n\right\} + \mu^2\right) \\
&= \frac{\sigma^2}{N}
\end{aligned}$$

*qui tend vers 0 quand $N \to \infty$. Comme $\hat{\mu}$ est sans biais, on en déduit que* $\mathsf{Var}\left\{\hat{\mu}_N\right\} = \sigma^2/N$.

Une propriété importante d'un estimateur est la propriété de consistance :

**Définition 2.1 (Consistance d'un estimateur)** *Un estimateur $\hat{\theta}_N$ d'un paramètre $\theta$ est consistant si $\hat{\theta}_N$ converge en probabilités vers $\theta$ lorsque la taille $N$ de l'échantillon tend vers l'infini : pour tout $\epsilon > 0$*

$$\mathbb{P}\left\{|\hat{\theta}_N - \theta| \geq \epsilon\right\} \to 0 \quad \textit{quand} \quad N \to \infty \ .$$

On a montré plus haut que

$$\lim_{N\to\infty} \mathsf{Var}\left\{\hat{\mu}_N\right\} = 0 \ .$$

On utilise ici l'inégalité de Markov : étant donnée une variable aléatoire $X$ dont l'espérace $\mathbb{E}\left\{X\right\}$ existe, on a $\mathbb{P}\{|X| \geq \epsilon\} \leq \mathbb{E}\left\{X\right\}/\epsilon$, pour tout $\epsilon > 0$. L'inégalité de Markov et le fait que $\hat{\mu}_N$ soit non biaisé donne

$$\mathbb{P}\{|\hat{\mu}_N - \mu| \geq \epsilon\} = \mathbb{P}\left\{(\hat{\mu}_N - \mu)^2 \geq \epsilon^2\right\} \leq \frac{\mathbb{E}\left\{(\hat{\mu}_n - \mu)^2\right\}}{\epsilon^2} = \frac{\mathsf{Var}\left\{\hat{\mu}_N\right\}}{\epsilon^2} \underset{N\to\infty}{\longmapsto} 0 \ .$$

La moyenne empirique est donc un estimateur consistant de la moyenne. On peut résumer la discussion dans le résultat suivant :

**Proposition 2.1** *La moyenne empirique fournit un estimateur $\hat{\mu}$ sans biais et consistant de la moyenne.*

**Exemple 2.3 (Propriétés de la variance empirique)** *Considérons maintenant la variance empirique, et calculons*

$$\begin{aligned}\mathbb{E}\left\{\sum_{n=1}^{N}(X^n-\hat{\mu})^2\right\} &= \sum_{n=1}^{N}\mathbb{E}\left\{[(X^n-\mu)-(\hat{\mu}-\mu)]^2\right\}\\ &= \sum_{n=1}^{N}\left[\mathbb{E}\left\{(X^n-\mu)^2\right\}-2\mathbb{E}\left\{(X^n-\mu)(\hat{\mu}-\mu)\right\}\right]+N\mathbb{E}\left\{(\hat{\mu}-\mu)^2\right\}\\ &= (N+1)\sigma^2-2\sum_{n=1}^{N}\mathbb{E}\left\{(X^n-\mu)(\hat{\mu}-\mu)\right\}\ .\end{aligned}$$

*Pour le dernier terme on a*

$$\begin{aligned}\mathbb{E}\left\{(X^n-\mu)(\hat{\mu}-\mu)\right\} &= \mathbb{E}\left\{(X^n-\mu)\left(\frac{1}{N}\sum_{m=1}^{N}X^m-\mu\right)\right\}\\ &= \frac{1}{N}\sum_{m=1}^{N}\mathbb{E}\left\{X^nX^m\right\}-\mu^2\\ &= \frac{1}{N}\left[(N-1)\mu^2+(\sigma^2+\mu^2)\right]-\mu^2=\frac{\sigma^2}{N}\ ,\end{aligned}$$

*par conséquent, en notant $\hat{\sigma}^2$ l'estimateur de la variance empirique, on obtient*

$$\mathbb{E}\left\{\hat{\sigma}^2\right\}=\sigma^2$$

*l'estimateur est donc sans biais. Ceci illustre le rôle du dénominateur $N-1$ dans la définition de $\hat{\sigma}^2$.*

**Remarque 2.1 (Variance de la variance empirique)** *Avec les mêmes hypothèses, et en supposant de plus que $\mathbb{E}\left\{X^4\right\}<\infty$, il est possible d'obtenir une expression pour la variance de la variance empirique :*

$$\mathsf{Var}\left\{\hat{\sigma}^2\right\}=\frac{1}{N}\left(\mu_4-\frac{N-3}{N-1}\sigma^4\right)\ ,$$

*où $\mu_4=\mathbb{E}\left\{(X-\mu)^4\right\}$ est le quatrième moment centré de $X$. La chose importante à retenir ici est que cette variance décroît comme $1/N$ quand $N\to\infty$, c'est à dire lentement. Il faut donc que $N$ soit assez grand pour que l'estimation soit asez fiable.*

### 2.1.2 Estimation de covariance

On considère maintenant la situation où on dispose de $N$ observations i.i.d. $x^1,\dots x^N\in\mathbb{R}^p$ d'un vecteur aléatoire $X\in\mathbb{R}^p$, de moyenne $\mu=\mathbb{E}\left\{X\right\}\in\mathbb{R}^p$ et matrice de covariance $\Sigma\in\mathcal{M}_p(\mathbb{R})$ inconnus, et on cherche à estimer ces quantités à partir des observations. On forme alors la moyenne empirique et la matrice de covariance empirique des observations, définies par

$$m=\frac{1}{N}\sum_{n=1}^{N}x^n\ ,\qquad S=\frac{1}{N-1}\sum_{n=1}^{N}x^n(x^n)^T\ . \tag{2.5}$$

En d'autres termes,

$$m_k=\frac{1}{N}\sum_{n=1}^{N}x_k^n\ ,\qquad S_{k\ell}=\frac{1}{N-1}\sum_{n=1}^{N}x_k^nx_\ell^n\ .$$

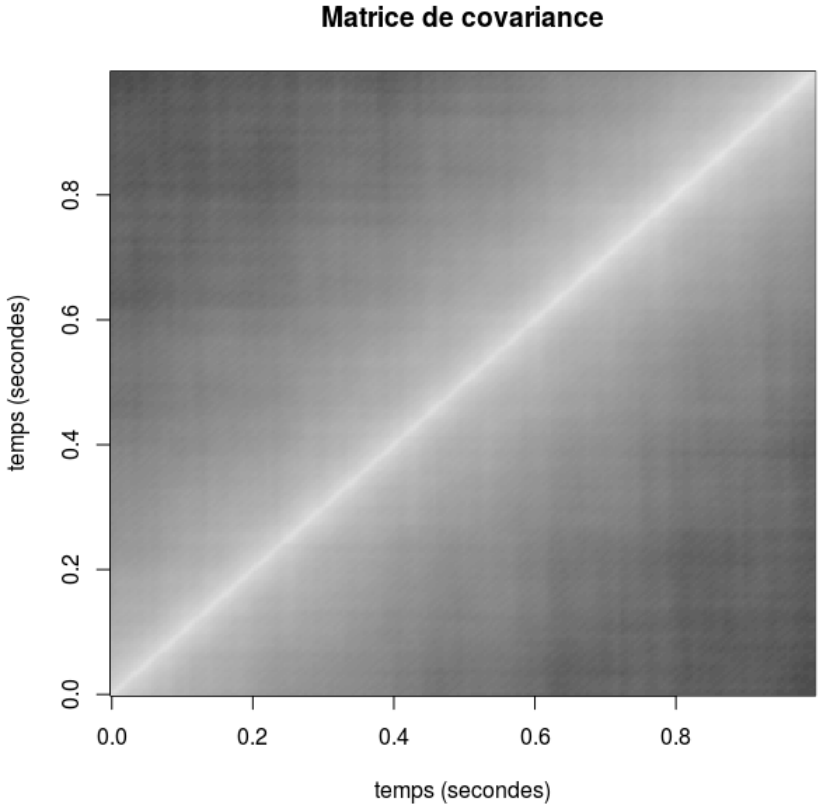


FIGURE 2.1 – Représentation de la matrice de covariance des données *non-cible* pour le *P300 speller* sour forme d'une image (les lignes sont à considérer de bas en haut)

On note $\hat{\mu}$ et $\widehat{\Sigma}$ les estimateurs correspondants. Avec les mêmes arguments que plus haut il est possible de montrer :

**Proposition 2.2** *1. $\hat{\mu}$ est un estimateur sans biais et consistant de la moyenne $\mu$.*

*2. $\widehat{\Sigma}$ est un estimateur sans biais de la matrice de covariance.*

*3. La matrice de covariance empirique $S$ est symétrique, semi-définie positive, et de rang inférieur ou égal à $N$.*

**Exemple 2.4 (Covariance de données *P300 speller*)** *Dans les données de P300 speller, les observations prennent la forme de vecteurs $x \in \mathbb{R}^L$, $L$ étant la longueur des signaux et chaque composante $x_\ell$ de $x$ représentant la mesure à l'instant $\ell$ du signal. La matrice de covariance de ces observations va jouer un rôle important dans la problématique de détection, et doit être estimée avec le plus grand soin. La matrice de covariance empirique des données non-cible (c'est à dire lorsque la ligne ou la colonne illuminée ne contient pas le caractère cible) est représentée en Figure 2.1, représentée en niveaux de gris, du blanc (représentant les grandes valeurs) au noir (faibles valeurs). On voit bien que cette matrice est "diagonale dominante", les éléments de matrice décroissent lorsque l'on s'éloigne de la diagonale.*

*Il est clair que la qualité de l'estimée dépend du nombre de réalisations (donc du nombre d'essais dans notre exemple).*

### 2.1.3 Maximum de vraisemblance

On a vu dans les sections précédentes des exemples d'estimateurs, et quelques notions permettant d'étudier les qualités d'un estimateur. On va maintenant voir une approche générique permettant de construire des estimateurs.
Prenons l'exemple d'une variable aléatoire discrète $X$ prenant ses valeurs dans l'ensemble $\{a_1, \dots a_K\}$, et supposons que la distribution de probabilités correspondante dépend d'un paramètre $\theta \in \Omega \subset \mathbb{R}$. On note $\mathbb{P}_\theta$ cette distribution de probabilités, pour expliciter la dépendance en $\theta$. Supposons qu'on observe $N$ réalisations indépendantes de $X$, notées $x^1, \dots x^N$. On appellera vraisemblance de $\theta$ la fonction

$$\mathcal{L}(\theta) = \prod_{n=1}^{N} \mathbb{P}_\theta \{X = x^n\} \ . \tag{2.6}$$

L'idée de l'estimation par maximum de vraisemblance est de rechercher la valeur de $\theta$ pour laquelle la fonction $\mathcal{L}$ atteint son maximum.
De façon similaire, supposons maintenant que $x^1, \dots x^N$ soient $N$ réalisations i.i.d. d'une variable aléatoire continue $X$, admettant une densité de probabilités $\rho_\theta$ dépendant d'un paramètre $\theta$. On appellera cette fois vraisemblance la fonction de la variable $\theta$ définie par

$$\mathcal{L}(\theta) = \prod_{n=1}^{N} \rho_\theta (x^n) \ . \tag{2.7}$$

**Définition 2.2** *Soient $x^1, \dots x^N$, $N$ réalisations i.i.d. d'une variable aléatoire $X$, dont la distribution dépend d'un paramètre $\theta$. L'estimation de $\theta$ par maximum de vraisemblance est définie par la valeur de $\theta$ qui maximise la vraisemblance définie en (2.6) ou (2.7) selon le contexte :*

$$\theta_* = \underset{\theta \in \Omega}{\mathsf{argmax}}\, \mathcal{L}(\theta) \ .$$

*On appelle l'estimateur correspondant, noté génériquement $\hat{\theta}$, estimateur du maximum de vraisemblance.*

Notons qu'il est souvent plus facile de rechercher un maximum de la fonction

$$\ell(\theta) = \ln (\mathcal{L}(\theta)) \ , \tag{2.8}$$

appelée log-vraisemblance. En effet, le logarithme transforme le produit en somme, qui est généralement plus facile à manipuler.

**Exemple 2.5 (Paramètre d'une loin de Poisson)** *Supposons que $x^1, \dots x^N$ soient des réalisations i.i.d. d'une variable aléatoire $X \sim \mathcal{P}(\lambda)$ suivant une distribution de Poisson de paramètre $\lambda \in \mathbb{R}^+$*

$$\mathbb{P}_\lambda\{x\} = e^{-\lambda} \frac{\lambda^x}{x!} \ , \qquad x \in \mathbb{N} \ .$$

*La log-vraisemblance s'écrit*

$$\ell(\lambda) = -N\lambda + \ln(\lambda) \sum_{n=1}^{N} x^n - \sum_{n=1}^{N} \ln(x^n!) \ .$$

*on voit facilement que c'est une fonction différentiable de $\lambda$, et que la valeur $\lambda_*$ qui maximise la vraisemblance est donnée par la moyenne empirique*

$$\lambda_* = \frac{1}{N} \sum_{n=1}^{N} x^n = \bar{x} \ .$$

*Comme on a aussi $\mathbb{E}\{X\} = \lambda$, on voit que l'estimateur du maximum de vraisemblance est sans biais dans ce cas.*

Plus généralement, supposons que $\theta$ soit une variable continue, et que $\ell$ soit une fonction dérivable. Si $\ell'(\theta)$ s'annule en $\theta_* \in \Omega$ et si $\ell''(\theta_*) < 0$, alors $\theta_*$ est un maximum local de la vraisemblance.

**Remarque 2.2 (Paramètres vectoriels)** *Il arrive souvent qu'on ait à considérer des lois dépendant de plusieurs paramètres $\theta_1, \dots \theta_J$, que l'on cherche à estimer. On forme alors un vecteur de paramètres $\Theta = (\theta_1, \dots \theta_J) \in \Omega \subset \mathbb{R}^J$, et on peut ainsi définie une vraisemblance $\Theta \in \Omega \to \mathcal{L}(\Theta)$ et une log-vraisemblance $\Theta \in \Omega \to \ell(\Theta)$, qui sont des fonctions de plusieurs variables. L'estimation par maximum de vraisemblance revient à résoudre le problème d'optimisation*

$$\Theta_* = \underset{\Theta \in \Omega}{\operatorname{argmax}} \mathcal{L}(\Theta) \ . \tag{2.9}$$

**Exemple 2.6 (Estimation de paramètres d'une loi normale)** *Supposons que $x^1, \dots x^N$ soient des réalisations i.i.d. d'une variable aléatoire normale de moyenne $\mu$ et variance $\sigma^2$ : $X \sim \mathcal{N}(\mu, \sigma^2)$. La densité de $X$ est*

$$\rho_\Theta(x) = \frac{1}{\sqrt{2\pi v}} \exp\left[-\frac{(x-\mu)^2}{2v}\right] \ ,$$

*où on a posé $v = \sigma^2$, et dépend donc du paramètre vectoriel $\Theta = (\mu, v)$ .*

*La log-vraisemblance s'écrit*

$$\ell(\mu, v) = -\frac{N}{2}\ln(2\pi v) - \frac{1}{2v}\sum_{n=1}^{n}(x^n - \mu)^2 \ .$$

*Le calcul donne les expressions suivantes pour les estimations du maximum de vraisemblance :*

$$\begin{aligned} \mu_* &= \frac{1}{N}\sum_{n=1}^{N} x^n = \bar{x} \\ v_* &= \frac{1}{N}\sum_{n=1}^{N}(x^n - \bar{x})^2 \ . \end{aligned}$$

*$\mu_*$ est la moyenne empirique, mais $v_*$ correspond à un estimateur biaisé de la variance (on a vu plus haut l'expression de l'estimateur non biaisé).*

Il est possible de démontrer que l'estimateur du maximum de vraisemblance possède des propriétés intéressantes (il est consistant, asymptotiquement sans biais, asymptotiquement normal,...), sous des hypothèses assez peu restrictives.

## 2.2 Décision

### 2.2.1 Position du problème

Revenons à notre problème de décision du *P300 speller*. On cherche à déterminer à partir d'observations (les signaux mesurés sur un capteur ou plusieurs capteurs) si la ligne ou la colonne qui a été illuminée contient ou pas le caractère cible. On dispose d'un jeu de données d'apprentissage, à partir duquel on cherche à construire un détecteur le plus efficace possible.
Pour illustrer le problème, prenons un exemple plus simple, en dimension 2. Les observations sont des vecteurs à 2 composantes (représentés pas des points du plan), appartenant à deux familles, les bleus et les noirs. Un premier exemple est donné sur le graphe du haut de la Figure 2.2. Dans ce cas, les deux familles de points sont très bien séparées, et une régle de décision simple peut être trouvée. Par exemple, il suffit de décider qu'on affectera à la classe bleue les points dont la première coordonnée est inférieure à 2, et à la classe noire les points dont la première coordonnée est supérieure à 2. La

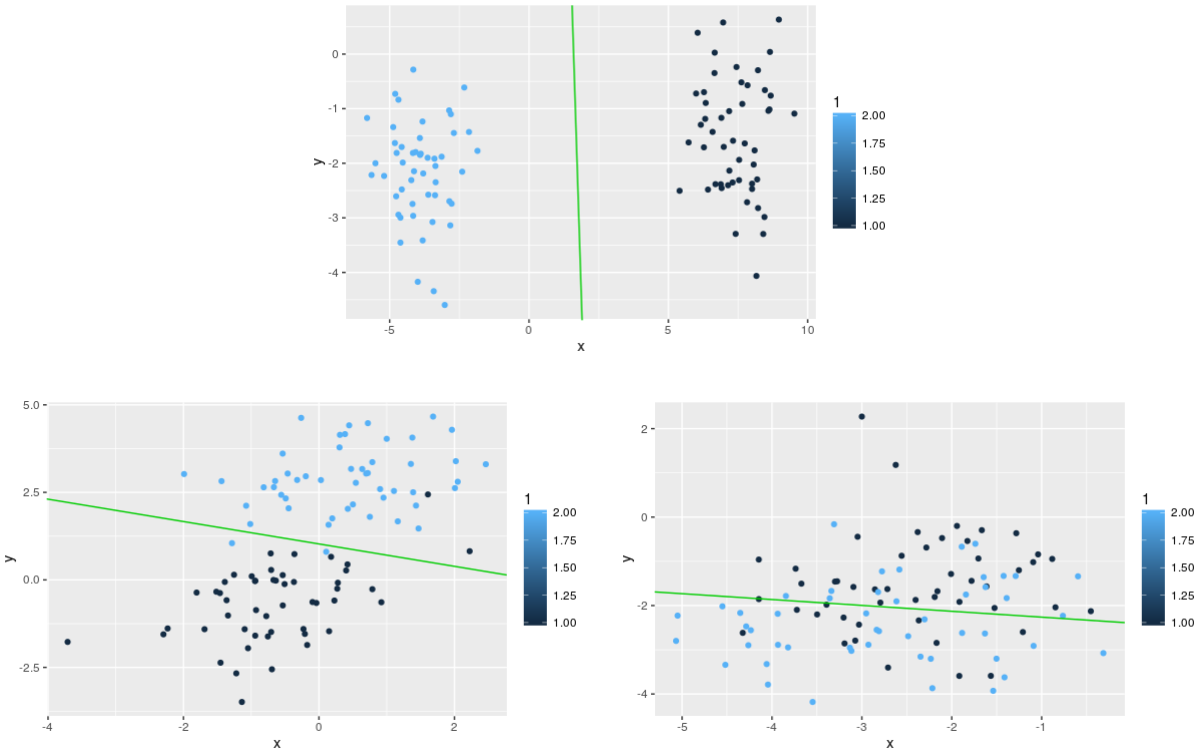


FIGURE 2.2 – Trois exemples de discrimination : on recherche une règle simple permettant de discriminer les points noirs des points bleus. La droite verte sépare l'espace en deux parties, d'après la règle fournie par l'analyse discriminante linéaire.

séparation verte apparaissant sur le graphe est la règle optimale donnée par l'analyse discriminante linéaire que nous allons voir plus loin.

Les deux graphes du bas de la FIGURE 2.2 montrent des situations dans lesquelles la séparation n'est pas si claire, surtout concernant le graphe de droite. Là encore la droite verte correspond à la régle de classification optimale donnée par l'analyse discriminante linéaire, mais on voit que cette règle conduit à des erreurs de classification.
Pour résumer, les trois principales questions qui se posent ici sont les suivantes :

1. Comment évaluer la qualité d'une règle de classification ?
2. Comment construire une règle de classification optimale ?
3. Comment construire une règle de classification à partir d'une base de données d'apprentissage ?

Dans des situations pratiques, on dispose en effet d'un jeu de données d'apprentissage, qui est utilisé pour construire la règle de classification. Cette règle est alors testée sur un autre jeu de données, appelé jeu de test, sur lequel on évalue les performances de la règle de classification.

### 2.2.2 La règle de Bayes

Le problème se formule comme suit. On suppose donnés
— un ensemble de variables d'entrée (continues) : $x^1, \dots x^N \in \mathbb{R}^p$, aussi appelées caractéristiques, prédicteurs ou variables indépendantes, et
— un ensemble correspondant de variables de sortie discrètes (variables de classe) : $C \in \mathcal{C} \stackrel{\Delta}{=} \{0, 1, ...K-1\}$.
A partir de ces données, on cherche à construire un prédicteur, défini comme une fonction

$$f : x \in \mathbb{R}^p \to f(x) \in \mathcal{C} \tag{2.10}$$

que l'on va chercher à optimiser. Pour cela, on introduit une fonction appelée fonction de perte $\mathcal{L} : \mathcal{C} \times \mathcal{C} \to \mathbb{R}_+$, qui fournit une mesure $\mathcal{L}(c, c')$ de la différence entre deux classes $c, c' \in \mathcal{C}$, et on s'intéresse à l'application

$$(c, x) \in \mathcal{C} \times \mathbb{R}^p \to \mathcal{L}(c, f(x)) \in \mathbb{R}_+$$

qui mesure une différence entre une classe $c \in \mathcal{C}$ et la classe $f(x)$ prédite à partir du signal $x$ (le choix du caractère $\mathcal{L}$ vient de l'anglais où la fonction de perte est appelée *loss function*). Le meilleur prédicteur, par rapport à cette fonction de perte, sera celui qui rend la perte moyenne la plus petite possible.
L'exemple le plus simple de fonction de perte, que nous utiliserons par la suite, est la fonction définie par

$$\mathcal{L}(c, c') = 1 - \delta_{cc'} = \begin{cases} 0 & \text{si } c = c' \\ 1 & \text{sinon ,} \end{cases} \tag{2.11}$$

mais d'autres choix sont évidemment possibles. Une fonction de perte est choisie en fonction de l'application visée.

Pour avancer dans la modélisation, on introduit un modèle aléatoire pour le couple $(c, x)$, modélisé comme une réalisation d'un couple aléatoire $(C, X)$ constitué d'une variable aléatoire discrète $C$ à valeurs dans $\mathcal{C}$ et un vecteur $X$ dans $\mathbb{R}^p$. On note $p_k = \mathbb{P}\{C = k\}$ la probabilité a priori de la classe $k \in \mathcal{C}$, et $x \in \mathbb{R}^p \to \rho_k(x)$ la densité conditionnelle de $x$ sachant que $C = k$.

**Définition 2.3 (Risque)** *Etant donnée une règle $f : \mathbb{R}^p \to \mathcal{C}$, le risque $R(f)$ de $f$ est l'erreur moyenne de prédiction lorsque cette règle est utilisée*

$$R(f) = \mathsf{EMP} = \mathbb{E}_{X,C}\left\{\mathcal{L}(C, f(X))\right\} .$$

Comme conséquence de la formule de Bayes, on introduit

**Définition 2.4 (Probabilité de classe *a posteriori*)** *La probabilité a posteriori de la classe $k \in \mathcal{C}$ étant donnée une observation $x \in \mathbb{R}^p$ est donnée par*

$$\mathbb{P}\{C = k | X = x\} = \frac{p_k \rho_k(x)}{\rho_X(x)} \ ,$$

*où $\rho_X$ est la densité marginale de $X$, donnée par*

$$\rho_X(x) = \sum_{k \in \mathcal{C}} p_k \rho_k(x) \ .$$

**Théorème 2.1 (Règle de Bayes)** *La règle de classification $f^*$ qui minimise l'erreur moyenne de prédiction est la règle de Bayes : pour tout $x \in \mathbb{R}^p$,*

$$f^*(x) = \operatorname*{argmax}_{k \in \mathcal{C}} \mathbb{P}\{C = k | X = x\} = \operatorname*{argmax}_{k \in \mathcal{C}} p_k \rho_k(x) \ .$$

**Exemple 2.7** *Prenons le cas unidimensionnel $p = 1$, et supposons que la variable aléatoire $X$ soit distribuée suivant une loi normale $\mathcal{N}(\mu_0, \sigma^2)$ avec probabilité $p_0 = 1/2$, et suivant une loi normale $\mathcal{N}(\mu_1, \sigma^2)$ avec probabilité $p_1 = 1 - p_0 = 1/2$. On suppose $\mu_0 < \mu_1$. Alors pour tout $x \in \mathbb{R}$,*

$$\rho_k(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x - \mu_k)^2}{2\sigma^2}\right) \ .$$

*On peut alors calculer*

$$\frac{p_1 \rho_1(x)}{p_0 \rho_0(x)} = \exp\left(-\frac{(x - \mu_1)^2 - (x - \mu_0)^2}{2\sigma^2}\right) \ ,$$

*et on voit que $p_1\rho_1(x) \geq p_0\rho_0(x)$ si et seulement si*

$$\frac{(\mu_1 - \mu_0)}{2\sigma^2} [2x - \mu_1 - \mu_0] \geq 0 \ ,$$

*ce qui équivaut à $x \geq x_d$, où la valeur critique $x_d$ est donnée par*

$$x_d \triangleq \frac{\mu_0 + \mu_1}{2} \ .$$

*La règle optimale est donc*

$$f^*(x) = \begin{cases} 1 & si\ x \geq x_d \\ 0 & sinon \end{cases} \ .$$

Cet exemple est illustré dans la Figure 2.3, qui représente les densités de probabilités de deux lois gaussiennes de même variance et moyennes différentes. Dans le cas où les probabilités a priori des classes sont égales, la frontière de décision sera donnée par la valeur $x_d$ pour laquelle les deux courbes s'intersectent, qui est ici égale à la moyenne des deux moyennes.

**Exemple 2.8** *Toujours dans le cas unidimensionnel, supposons maintenant que la variable aléatoire $X$ soit distribuée suivant une loi normale $\mathcal{N}(\mu_0, \sigma^2)$ avec probabilité $p_0$, et suivant une loi normale $\mathcal{N}(\mu_1, \sigma^2)$ avec probabilité $p_1 = 1 - p_0$. On peut montrer que dans ce cas la règle optimale prend encore la forme*

$$f^*(x) = \begin{cases} 1 & si\ x \geq x_d \\ 0 & sinon \end{cases} \ ,$$

*où la valeur critique est cette fois donnée par*

$$x_d \triangleq \frac{\mu_0 + \mu_1}{2} - \frac{\sigma^2}{\mu_1 - \mu_0} \ln\left(\frac{p_1}{p_0}\right) \ .$$

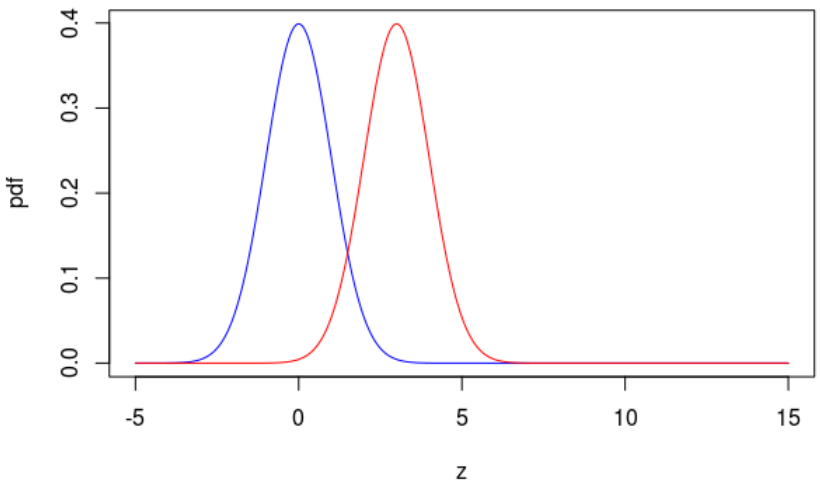


FIGURE 2.3 – Densités de deux lois gaussiennes de même variance et moyennes différentes

Dans ces deux exemples, le domaine de valeurs de $x$ (en l'occurrence $\mathbb{R}$) est partitionné en deux régions, séparées par la valeur critique $x_d$. On va voir plus loin ce qu'il en est dans le cas multidimensionnel.

## 2.3 Analyse discriminante

Considérons maintenant le cas multidimensionnel et multiclasse : les observations sont des vecteurs de dimension $p$, et peuvent appartenir à $K$ classes possibles. Le point de départ de l'analyse discriminante est un modèle appelé modèle de mélange de gaussiennes : on considère un vecteur aléatoire $X$ de dimension $p$, distribué selon une loi normale de moyenne $\mu_k$ et covariance $\Sigma_k$, avec probabilité $p_k = \mathbb{P}\{C = k\}$, pour $k = 0, \dots K-1$. La densité conditionnelle de $X$ dans la classe $k$ est donnée par

$$\rho_k(x) \triangleq \rho_{X|C=k}(x) \tag{2.12}$$

$$= \frac{1}{(2\pi)^{p/2}\sqrt{|\det(\Sigma_k)|}} \exp\left(-\frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k)\right) \tag{2.13}$$

On note cela sous la forme

$$X \sim p_0\mathcal{N}(\mu_0, \Sigma_0) + p_1\mathcal{N}(\mu_1, \Sigma_1) + \cdots p_{K-1}\mathcal{N}(\mu_{k-1}, \Sigma_{K-1}) \ , \tag{2.14}$$

Comme précédemment, on introduit le log rapport de vraisemblance : étant données deux classes $k, \ell \in \mathcal{C}$, on définit

$$\Delta_{k,\ell} \triangleq \ln\left(\frac{\mathbb{P}\{C=k|X=x\}}{\mathbb{P}\{C=\ell|X=x\}}\right) \tag{2.15}$$

$$= \ln\left(\frac{p_k}{p_\ell}\right) - \frac{1}{2}\ln\left(\frac{\det(\Sigma_k)}{\det(\Sigma_\ell)}\right) - \frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_c) + \frac{1}{2}(x-\mu_\ell)^T\Sigma_\ell^{-1}(x-\mu_\ell) \tag{2.16}$$

Si $\Delta_{k,\ell} > 0$, alors la classe $k$ sera considérée plus vraisemblable que la classe $\ell$.

**Définition 2.5** *L'hypersurface de $\mathbb{R}^p$ d'équation $\Delta_{k,\ell} = 0$ est appelée frontière de décision entre les classes $k$ et $\ell$, qui coupe $\mathbb{R}^p$ en deux régions.*

L'équation $\Delta_{k,\ell} = 0$ définit généralement une hypersurface quadratique dans $\mathbb{R}^p$. L'ensemble des frontières de décision fournit une partition de $\mathbb{R}^p$ en régions. Il y a au plus $K$ régions distinctes. Etant donnée une nouvelle donnée $x$ à classer, elle sera affectée à la classe $k$ telle que $\Delta_{k,\ell} > 0$ pour tout $\ell \neq k$.

### 2.3.1 Analyse discriminante linéaire (LDA)

On considère le cas particulier où toutes les matrices de covariances sont égales

$$\Sigma_0 = \Sigma_1 = \cdots = \Sigma_{K-1} = \Sigma \ .$$

Dans ce cas, le log rapport de vraisemblance prend une forme plus simple

$$\Delta_{k,\ell}(x) = \ln\left(\frac{p_k}{p_\ell}\right) - \frac{1}{2}(x-\mu_k)^T\Sigma^{-1}(x-\mu_c) + \frac{1}{2}(x-\mu_\ell)^T\Sigma^{-1}(x-\mu_\ell) \ . \tag{2.17}$$

On voit que les termes quadratiques s'éliminent, donc $\Delta_{k,\ell}(x)$ devient une fonction linéaire de $x$. Plus précisément, on voit que

$$\Delta_{k,\ell}(x) = \delta_k(x) - \delta_\ell(x) \ , \tag{2.18}$$

où on a introduit la fonction de discrimination

$$\delta_k(x) = \ln p_k(x) + x^T\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k\Sigma^{-1}\mu_k \ . \tag{2.19}$$

On a donc $\mathbb{P}\{C = k|X = x\} > \mathbb{P}\{C = \ell|X = x\}$ dès que $\delta_k(x) > \delta_\ell(x)$, soit

$$x^T\Sigma^{-1}(\mu_k - \mu_\ell) > -\ln(p_k/p_\ell) + \frac{1}{2}(\mu_k + \mu_\ell)^T\Sigma^{-1}(\mu_k - \mu_\ell)$$

Ceci définit un demi-espace, délimité par un hyperplan, perpendiculaire au vecteur $\Sigma^{-1}(\mu_k - \mu_\ell)$.

**Proposition 2.3** *Avec les notations ci-dessus, la règle de Bayes en analyse discriminante linéaire conduit à la règle de décision : pour $x \in \mathbb{R}^p$,*

$$k_* = \underset{k\in\{0,\dots K-1\}}{\mathsf{argmax}} \ \delta_k(x) \ .$$

**Remarque 2.3** *En pratique, la matrice de covariance $\Sigma$ n'est pas connue, de même que les moyennes $\mu_k$ des classes et leurs probabilités $p_k$. Ces quantités doivent être estimées à partir d'un jeu de données d'apprentissage. Pour que les estimées soient de précision suffisante, il importe que le jeu d'apprentissage soit assez grand. C'est surtout le cas pour la matrice $\Sigma$, car la LDA demande que la matrice estimée soit inversible.*

### 2.3.2 Analyse discriminante quadratique (QDA)

En analyse discriminante quadratique, on ne fait plus l'hypothèse d'égalité des matrices de covariance des classes. Les frontières de décision $\Delta_{k,\ell}(x) = 0$ (où $\Delta_{k,\ell}$ est défini en (2.16)) sont maintenant des hypersurfaces quadratiques dans $\mathbb{R}^p$.
Ceci étant, l'utilisation pratique de la QDA nécessite l'estimation de toutes les matrices de covariance. Il faut donc que les effectifs de toutes les classes soient suffisants pour permettre cette estimation. Pour cette raison, la LDA est souvent préférée, même dans des situations où on sait que les matrices de covariance des classes sont différentes.

**Remarque 2.4 (Estimation régularisée de matrices de covariance)** *Dans des situations où les effectifs des classes sont trop faibles pour que les matrices de covariance des classes puissent être estimées correctement, on peut soit utiliser la LDA comme mentionné ci-dessus, soit modifier l'estimation des matrices de covariance des classes pour les rendre plus robustes. Une solution simple consiste à les régulariser, en les remplaçant par une moyenne pondérée d'elles mêmes et de la matrice globale :*

$$\widehat{\Sigma}_k^{(\mathsf{reg})} = \lambda \widehat{\Sigma}_k + (1-\lambda)\widehat{\Sigma} \ ,$$

*où $\lambda \in [0,1]$ est un paramètre permettant de régler le poids donné à la matrice globale.*
*Notons que ceci pose le problème du choix de $\lambda$...*

### 2.3.3 Application aux données de *P300-speller*

Pour terminer, considérons de nouveau le problème de décision associé au *P300-speller*. On dispose de deux jeux données par sujet : données d'apprentissage et données de test. On peut penser que les sujets sont assez différents, et traiter le problème indépendamment pour chaque sujet (ceci peut évidemment être mis en question...)

1. Pour chaque sujet, le jeu d'apprentissage peut être utilisé pour "apprendre" la règle de décision : les moyennes $\mu_0$ et $\mu_1$, les probabilités des classes $p_0$ et $p_1$ et la matrice de covariance $\Sigma$ (ou les les matrices de covariance des classes $\Sigma_0$ et $\Sigma_1$ si on décide d'utiliser la QDA) sont estimées. Ceci permet de construire la règle de décision (linéaire ou quadratique).
2. Le jeu de test est alors utilisé pour quantifier les performances du détecteur. Les classes obtenues sont comparées aux classes connues, et un pourcentage de bonne ou mauvaise classification est calculé.
3. De façon pratique, il faut savoir que ce protocole peut être fatiguant pour le sujet, surtout s'il s'agit d'un sujet handicapé (ce qui est l'objectif de ce dispositif). Il est donc important de faire en sorte que le détecteur soit le plus performant possible, et demande des données (apprentissage et test) les plus petites possibles,... avec tous les inconvénients que cela comporte en terme d'estimation.

Chapitre

# 3 Solutions des exercices

## 3.1 Exercices du chapitre 1

*Solution de l'exercice 1.1 (Approximation constante par morceaux)*

*Solution de l'exercice 1.2 (Approximation à bande limitée)*

*Solution de l'exercice 1.3 (Linéarité de l'espérance)*

*Solution de l'exercice 1.4 (Variable aléatoire gaussienne)*

*Solution de l'exercice 1.5 (Marginales)*

*Solution de l'exercice 1.6 (Bernoulli fois gaussien)*

*Solution de l'exercice 1.7 (Vecteur aléatoire gaussien)*

*Solution de l'exercice 1.8 (Vecteur gaussien sans densité)*

## 3.2 Exercices du chapitre 2

# Index


Adjoint, 16
Approximation constante par morceaux, 18
Approximation à bande limitée, 18

Base Hilbertienne, 15
Biais, 28

Caractéristique, 34
Conjugué Hermitien, 16
Convergence en probabilités, 28
Cortex, 7

Densité de probabilités, 21
Données d'apprentissage, 27
Données de test, 27
Décision, 27
Déterminant de Gram, 19

Echantillonnage, 14
EEG, 8
Electroencéphalographie, 8
Erreur quadratique moyenne, 28
Espace de Hilbert, 15
Espace pré-Hilbertien, 15
Estimateur du maximum de vraisemblance, 31

Famille duale, 20
Fonction erf, 22
Fonction caractéristique, 21
Fonction de perte, 34
Fonction de répartition, 20
Forme sesquilinéaire, 14
Formule de Parseval, 16
Fréquence, 18
Fréquence d'échantillonnage, 10

i.i.d., 27

Imagerie fonctionnelle, 8
Imagerie structurelle, 8
Interface cerveau machine, 9
Inégalité de Cauchy-Schwarz, 15
Inégalité de Markov, 28

Log rapport de vraisemblance, 36
Log-vraisemblance, 31

Magnétoencéphalographie, 8
Matière blanche, 7
Matière grise, 7
Matrice de covariance, 23
Matrice de Gram, 19
MEG, 8
Modèle de mélange de gaussiennes, 35

Orthogonalité, 15

P300-speller, 9
Perpendicularité, 15
Produit Hermitien, 15
Produit scalaire, 15
Projection orthogonale, 16
Prédicteur, 34

Risque, 34
Règle de Bayes, 34

Signal Analogique, 13
Signal Numérique, 14
Signal à temps continu, 13
Signal à temps discret, 14

Variable d'entrée, 34
Variable de sortie, 34
Variable indépendante, 34

Vecteur aléatoire, 20
Vecteur gaussien, 23
Vraisemblance, 30, 31