

Analyse de Signaux Biomédicaux

Bruno Torrèsani



Master TSI
Aix-Marseille Université
Année 2021-22

Copyright © 2021 Bruno Torr sani

PUBLI  PAR BRUNO TORR SANI

<https://www.i2m.univ-amu.fr/perso/bruno.torresani>

Licensed under the Creative Commons Attribution-NonCommercial 3.0 Unported License (the “License”). You may not use this file except in compliance with the License. You may obtain a copy of the License at <http://creativecommons.org/licenses/by-nc/3.0>. Unless required by applicable law or agreed to in writing, software distributed under the License is distributed on an “AS IS” BASIS, WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied. See the License for the specific language governing permissions and limitations under the License.

18 novembre 2021

Table des matières

| | | |
|----------|---------------------------|----------|
| 1 | Introduction | 7 |
|----------|---------------------------|----------|

| | |
|----------|------------|
| 1 | ECG |
|----------|------------|

| | | |
|------------|---|-----------|
| 2 | Électrocardiogramme, généralités | 11 |
| 2.1 | Quelques éléments sur l'activité cardiaque et sa mesure | 11 |
| 2.1.1 | L'activité cardiaque | 11 |
| 2.1.2 | L'électrocardiogramme | 12 |
| 2.2 | Représentation d'un ECG normal | 13 |
| 2.2.1 | Rythme sinusal | 13 |
| 2.2.2 | Arythmie cardiaque | 14 |
| 2.2.3 | Traitement des signaux | 15 |
| 2.3 | Données et outils logiciels | 15 |
| 2.3.1 | La base de données Physionet | 15 |
| 2.3.2 | La boîte à outils WFDB | 16 |
| 3 | Pré-traitements : filtrage | 19 |
| 3.1 | Signaux numériques : estimation spectrale et filtrage | 19 |
| 3.1.1 | Signaux numériques | 19 |
| 3.1.2 | Filtrage numérique | 20 |
| 3.1.3 | Transformation de Fourier Discrète, transformation de Fourier Finie | 22 |
| 3.1.4 | Transformation de Fourier et filtrage | 25 |
| 3.1.5 | Transformation en z | 28 |
| 3.1.6 | Factorisation de filtres | 30 |
| 3.2 | Applications au filtrage de signaux ECG | 32 |
| 3.2.1 | Estimation spectrale | 33 |
| 3.2.2 | Suppression de la ligne de base : filtrage passe-haut | 35 |

| | | |
|------------|--|-----------|
| 3.2.3 | Suppression de lignes spectrales 60Hz : filtre à encoche | 36 |
| 3.3 | Temps-fréquence, ondelettes | 37 |
| 3.3.1 | Transformation de Fourier à court terme, spectrogramme | 37 |
| 3.3.2 | Ondelettes et variantes | 44 |
| 4 | Détection du complexe QRS | 47 |
| 4.1 | Généralités | 47 |
| 4.2 | L'algorithme de Pan & Tompkins | 48 |
| 4.2.1 | Détection de plages d'intérêt | 48 |
| 4.2.2 | Détection des ondes R | 49 |
| 4.3 | L'algorithme d'Elgendi | 51 |
| 4.3.1 | Filtrages | 51 |
| 4.3.2 | Seuillages et détection | 51 |
| 4.3.3 | Optimisation des paramètres | 52 |
| 4.4 | Filtrage adapté | 53 |
| 4.5 | Analyse de variabilité | 54 |
| 4.5.1 | Séparation de variations lentes et rapides | 54 |
| 4.5.2 | Analyse spectrale | 55 |

II

EEG

| | | |
|------------|--|-----------|
| 5 | Électroencéphalogramme, généralités | 59 |
| 5.1 | Quelques éléments sur l'activité cérébrale et sa mesure | 59 |
| 5.1.1 | Quelques éléments d'imagerie cérébrale | 60 |
| 5.1.2 | Interfaces cerveau machine | 61 |
| 5.2 | Structure de données d'EEG et BCI | 62 |
| 5.2.1 | Données EEG, potentiels évoqués | 62 |
| 5.2.2 | Le <i>P300 speller</i> | 63 |
| 6 | Probabilités, estimation, décision | 65 |
| 6.1 | Éléments de théorie des probabilités | 65 |
| 6.1.1 | Modèle probabiliste | 65 |
| 6.1.2 | Calcul des probabilités | 66 |
| 6.2 | Variables aléatoires | 71 |
| 6.2.1 | Généralités | 71 |
| 6.2.2 | Variables aléatoires discrètes | 73 |
| 6.2.3 | Variables aléatoires à densité | 76 |
| 6.2.4 | Vecteurs aléatoires | 79 |
| 6.2.5 | Loi normale univariée et multivariée | 81 |
| 6.3 | Statistique, estimation | 83 |
| 6.3.1 | Statistique | 83 |
| 6.3.2 | Estimation, exemples | 84 |
| 6.3.3 | Estimation de covariance | 86 |
| 6.3.4 | Maximum de vraisemblance | 87 |
| 6.4 | Classification, décision | 89 |
| 6.4.1 | Position du problème | 89 |
| 6.4.2 | La règle de Bayes | 91 |

| | | |
|------------|---|-----------|
| 6.5 | Analyse discriminante linéaire et quadratique | 92 |
| 6.5.1 | Analyse discriminante linéaire (LDA) | 93 |
| 6.5.2 | Analyse discriminante quadratique (QDA) | 94 |
| 6.6 | Application aux données de <i>P300-speller</i> | 95 |
| 6.6.1 | Position du problème | 95 |
| 6.6.2 | Utilisation d'un seul capteur | 96 |
| 6.6.3 | Exploitation de plusieurs capteurs | 96 |

III

Compléments

| | |
|----------------------|------------|
| Bibliographie | 99 |
| Index | 101 |

1. Introduction

Les signaux biomédicaux recouvrent un large ensemble de classes de signaux, provenant de différentes modalités. On peut par exemple citer

- Signaux bioélectriques : générés par les cellules nerveuses et musculaires : Nerf et muscle, potentiels d'action, propagation du champ électrique, ECG, EEG, EMG, EOG, EGG, GSR
- Signaux biochimiques provenant de tissus vivants ou d'échantillons analysés en laboratoire : pO_2 , pCO_2 , concentration ionique, taux de glucose, spectroscopie RMN
- Signaux biomécaniques (des mesures invasives sont souvent nécessaires) : mouvement, tension, déplacement, tension artérielle, débit
- Signaux biomagnétiques : champs générés par le cerveau, le cœur et les poumons
- Signaux bioacoustiques : sons du cœur, bruits respiratoires, articulation et contraction des muscles
- Signaux bio-optiques : propriétés fluoroscopiques du liquide amniotique, débit cardiaque mesuré par technique de dilution de colorant

L'exploitation et le traitement de ces signaux font appel à de nombreux aspects du traitement du signal, partant de problématiques classiques comme l'échantillonnage, le filtrage, le dé-bruitage ou encore la compression, jusqu'à des questions liées à des problématiques de décision ou classification.

Dans ce cours, on se focalisera sur deux modalités de signaux bioélectriques, l'ECG et l'EEG, qui nous permettront de visiter ou revisiter un certain nombre de ces problématiques. Plus précisément

- Sur les signaux ECG, on se focalisera dans un premier temps sur des pré-traitements du signal brut (filtrages, suppression de ligne de base), avant d'aborder des questions plus spécifiques comme la caractérisation du complexe QRS ou la mesure du rythme cardiaque par différentes techniques.
- Sur les signaux EEG, qui sont des signaux "multicapteurs", on se focalisera sur des problématiques de détection (détection d'événements localisés dans le temps, en l'occurrence des potentiels évoqués), par des techniques qui relèvent du filtrage adapté et/ou de la classification supervisée. On travaillera sur un jeu de données de type *P300 speller*, qui est un cas particulier de protocole d'interface cerveau machine.

| | | |
|----------|---|-----------|
| 2 | Électrocardiogramme, généralités | 11 |
| 2.1 | Quelques éléments sur l'activité cardiaque et sa mesure | |
| 2.2 | Représentation d'un ECG normal | |
| 2.3 | Données et outils logiciels | |
| 3 | Pré-traitements : filtrage | 19 |
| 3.1 | Signaux numériques : estimation spectrale et filtrage | |
| 3.2 | Applications au filtrage de signaux ECG | |
| 3.3 | Temps-fréquence, ondelettes | |
| 4 | Détection du complexe QRS | 47 |
| 4.1 | Généralités | |
| 4.2 | L'algorithme de Pan & Tompkins | |
| 4.3 | L'algorithme d'Elgendi | |
| 4.4 | Filtrage adapté | |
| 4.5 | Analyse de variabilité | |

2. Électrocardiogramme, généralités

2.1 Quelques éléments sur l'activité cardiaque et sa mesure

2.1.1 L'activité cardiaque

Le cœur est composé de deux cavités (cœur droit/gauche). Chacune de ces cavités a 2 compartiments : oreillette/ ventricule. La description qui suit est extraite du site web [6], qu'on pourra consulter pour plus de détails.

Chaque battement correspond à une même séquence

— Cœur droit : le sang pauvre en oxygène arrive au cœur par la veine cave.

- Il y entre par l'oreillette droite, et en est chassé par sa contraction appelée systole auriculaire.
- Il se déplace dans le ventricule droit.
- La systole ventriculaire (contraction des ventricules) propulse à son tour le sang du ventricule droit vers les poumons où il va se charger en oxygène.

— Cœur gauche : De retour au cœur par les veines pulmonaires,

- Le sang s'accumule dans l'oreillette gauche
- lors de la systole auriculaire, il passe dans le ventricule gauche
- lors de la systole ventriculaire le sang est envoyé vers les organes par l'artère aorte.

Ce fonctionnement est schématisé en Figure 2.1 (tirée du site web [7]). On pourra aussi consulter la vidéo sur le site https://www.youtube.com/watch?v=OtUPjnj_zC4.

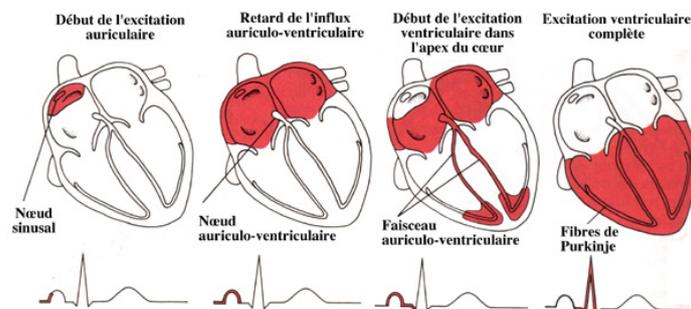


FIGURE 2.1 – Un cycle du rythme cardiaque

L'activité cardiaque (contraction/relaxation) est liée à une activité électrique se propageant dans tout l'organe. La contraction est produite par une activité électrique des fibres cardiaques. Cette activité est en fait une dépolarisation, suivie d'une repolarisation, liées à des mouvements ioniques (potassium, calcium, sodium) à travers la membrane cellulaire. Ces polarisations/dépolarisations se propagent le long des fibres cardiaques.

Au repos, la fibre cardiaque est "polarisée", chargée positivement à l'extérieur, négativement à l'intérieur. Une stimulation produit une modification de la perméabilité de la membrane cellulaire avec inversion des charges électriques, qui deviennent positives à l'intérieur et négatives à l'extérieur de la fibre. Cette dépolarisation se propage le long de la fibre, qui sera ainsi entièrement dépolarisée. Elle est transmise aux fibres voisines et les active de la même manière. La propagation de la dépolarisation au long de la fibre est représentée par un vecteur qui indique sa direction, du négatif au positif.

La dépolarisation est suivie d'une repolarisation de la fibre qui ramène les charges électriques, de part et d'autre de la membrane, à ce qu'elles étaient au repos. Comme la dépolarisation, la repolarisation se propage le long de la fibre, en sens contraire ou dans le même sens.

2.1.2 L'électrocardiogramme

L'électrocardiogramme est une mesure de cette activité électrique via des électrodes qui enregistrent le potentiel électrique en différents points. Les *dérivations cardiaques* sont des différences de potentiel électrique entre deux points, soit deux électrodes (dérivation bipolaire) ou entre un point virtuel et une électrode (dérivation unipolaire). Les dérivations classiques sont illustrées en Figure 2.2, extraite du site [1].

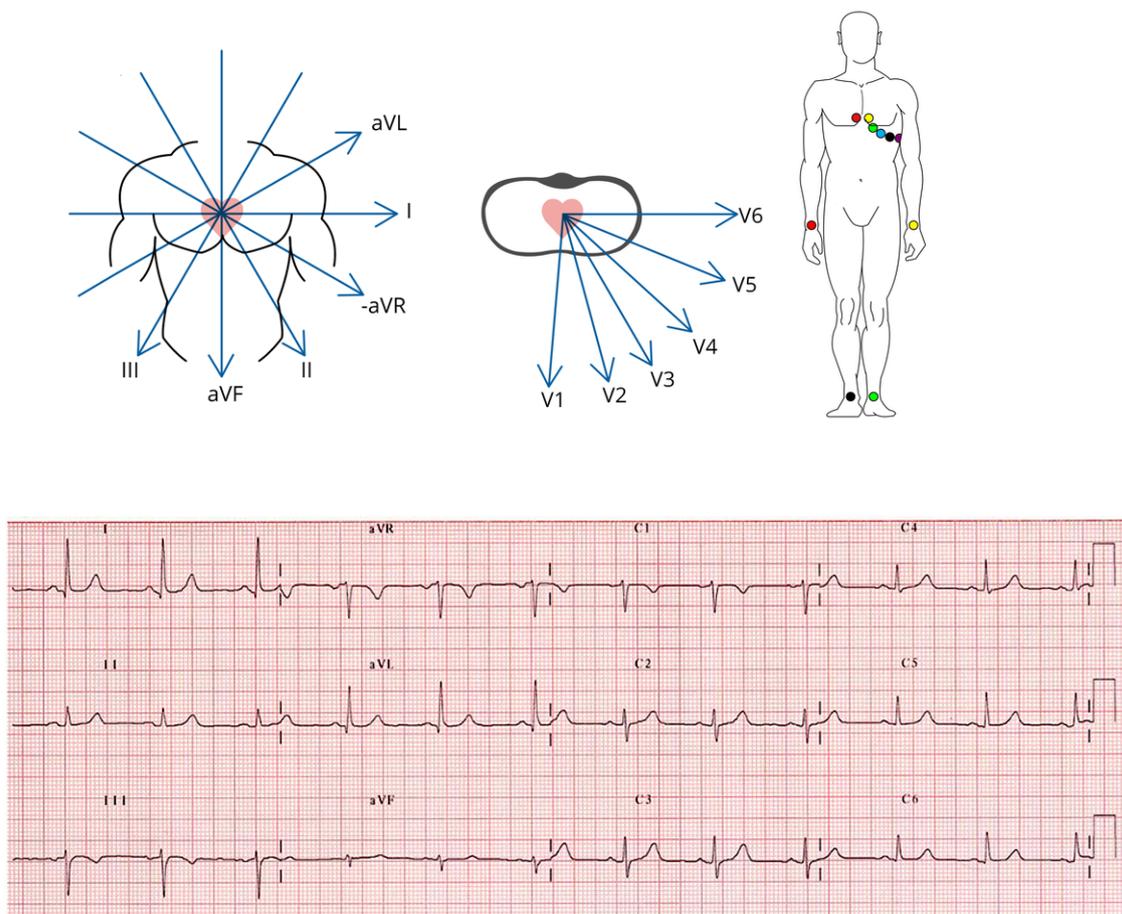


FIGURE 2.2 – Les 12 dérivations classiques (haut) et tracés ECG correspondants sur un sujet normal (bas).

Les dérivations cardiaques fournissent des points de vue différents du même stimulus électrique. On distingue

- les dérivations frontales ou standard (qui enregistrent le plan électrique frontal du cœur), notées aVL (bras gauche), aVR (bras droit) et aVF (jambe gauche). Les dérivations bipolaires standard associées sont notées $DI = aVR - aVL$, $DII = aVR - aVF$ et $DIII = aVL - aVF$ (aussi notées I , II et III). Il existe des dérivations modifiées, qu'on verra dans les exemples plus loin.
- les dérivations précordiales (qui enregistrent le plan horizontal), notées $V1, V2, \dots, V6$.

La géométrie de ces dérivations est représentée en Figure 2.2.

Pour plus de détails, on pourra par exemple consulter le site my-ekg.com.

2.2 Représentation d'un ECG normal

Le tracé électrique (voir Figure 2.3 pour un exemple) se caractérise par une répétition d'accidents appelés « ondes », et différents intervalles entre ces ondes.

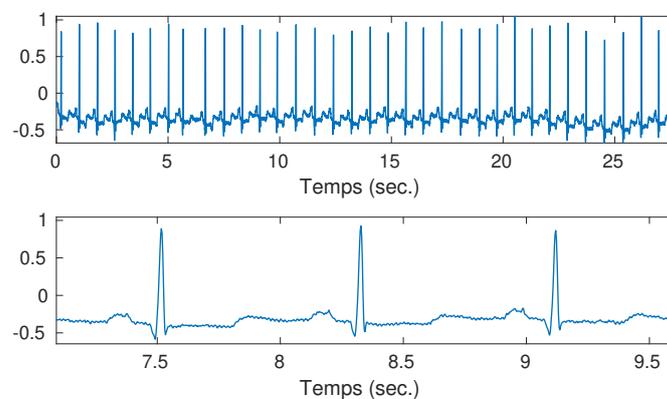


FIGURE 2.3 – Exemple de tracé ECG : segment d'approximativement 28 secondes (haut), zoom sur 3 cycles (bas).

2.2.1 Rythme sinusal

Un rythme cardiaque est défini comme la succession de trois battements présentant des formes d'onde identiques. Le fait qu'elles soient identiques indique que leur origine est identique.

Un rythme cardiaque normal est un rythme dit « sinusal » : l'activité cardiaque sous contrôle du nœud sinusal se caractérise par un certain nombre d'ondes notées P, Q, R, S et T (voir Figure 4.6 pour les morphologies des différentes « ondes »¹), et

- un rythme régulier avec un intervalle RR de longueur quasi-constante,
- la présence d'une onde P avant chaque QRS et d'un QRS après chaque onde P,
- des ondes P d'axe et de morphologie normales,
- un intervalle PR constant.

Si le rythme est régulier, on peut déterminer une fréquence cardiaque qui est égale à l'inverse de l'intervalle RR (multipliée par 60, pour être exprimée en nombre de pulsations par minute).

Lors de l'analyse d'un ECG, on se focalise surtout sur quelques caractéristiques essentielles telles que l'onde P, l'espace PR, le complexe QRS, le délai d'inscription de la déflexion intrinsécoïde (temps qui sépare le début du complexe QRS du sommet de l'onde R), du point J (point défini par l'intersection entre la fin du complexe QRS qui marque la fin de la dépolarisation, et le début du segment ST, début de la repolarisation), l'espace QT, le segment ST et l'onde T.

Les différentes ondes correspondent à différentes étapes du processus dépolarisation/polarisation, qui peuvent être interprétées comme suit.

1. Figure tirée de l'encyclopédie en ligne [Wikipedia](https://fr.wikipedia.org), comme une partie de la discussion de cette section

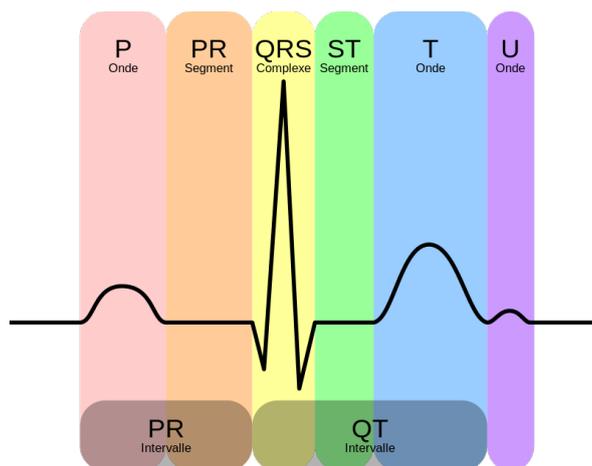


FIGURE 2.4 – Forme d'onde du complexe PQRST.

- L'onde *P* correspond à la dépolarisation (et la contraction) des oreillettes, droite et gauche. Sa durée est normalement de 0,08 à 0,1 seconde.
- L'intervalle *PR* (ou *PQ*) est le temps entre le début de *P* et le début du *QRS*. Il est le témoin du temps nécessaire à la transmission de l'influx électrique du nœud sinusal des oreillettes au tissu myocardique des ventricules. Sa durée normale, mesurée du début de l'onde *P* au début du complexe *QRS* est de 0,12 à 0,20 seconde. La durée de l'espace *PR* diminue lorsque la fréquence cardiaque augmente. Il est normalement isoélectrique.
- L'onde *QRS* (appelée aussi complexe *QRS*) qui correspond à la dépolarisation (et la contraction) des ventricules, droit et gauche. L'onde *Q* est la première onde négative du complexe. L'onde *R* est la première composante positive du complexe. L'onde *S* est la deuxième composante négative. Le complexe *QRS* a une durée normale inférieure à 0,1 seconde, le plus souvent inférieur à 0,08 s.
- Le point *J* correspond au point de transition entre le complexe *QRS* et le segment *ST*. Il est normalement isoélectrique.
- Le segment *ST* correspond au temps séparant la fin de la dépolarisation ventriculaire représentée par le complexe *QRS* et le début de l'onde *T*. Le segment *ST* normal est isoélectrique du point *J* au début de l'onde *T*.
- L'intervalle *QT* mesuré du début du *QRS* à la fin de l'onde *T* correspond à l'ensemble de la dépolarisation et de la repolarisation ventriculaire (temps de systole électrique). Sa durée varie en fonction de la fréquence cardiaque, il diminue quand la fréquence cardiaque augmente et augmente quand la fréquence cardiaque diminue. On utilise aussi le QT_c (*QT* corrigé) qui est la mesure de l'intervalle *QT* corrigé par la fréquence selon la formule $QT_c = QT / \sqrt{RR}$.
- L'onde *T* correspond à l'essentiel de la repolarisation (la relaxation) des ventricules, celle-ci commençant dès le *QRS* pour quelques cellules. Sa durée est de 0,20 à 0,25 secondes, l'analyse de sa durée est comprise dans l'analyse de la durée de l'intervalle *QT*.
- L'onde *T* atriale (repolarisation des oreillettes) est masquée par l'onde *QRS*. Elle est négative.
- L'onde *U* est une petite déflexion parfois observée après l'onde *T* dans certaines dérivations précordiales, son origine est discutée.

2.2.2 Arythmie cardiaque

On appelle arythmie cardiaque un groupe de conditions dans lesquelles le rythme cardiaque est irrégulier, trop rapide ou trop lent. La fréquence cardiaque trop rapide - supérieure à 100 battements par minute chez l'adulte - est appelée tachycardie, et une fréquence cardiaque trop lente - inférieure à 60 battements par minute - est appelée bradycardie. Certains types d'arythmies ne présentent aucun symptôme. Les symptômes, lorsqu'ils sont présents, peuvent inclure des palpitations ou une sensation de pause entre les battements cardiaques. Dans les cas plus graves, il peut y avoir des étourdissements, des évanouissements, un essoufflement ou des douleurs thoraciques. Bien que la plupart des types d'arythmie ne soient pas graves, certains prédisposent une personne à des complications telles qu'un accident vasculaire cérébral ou une

insuffisance cardiaque. D'autres peuvent entraîner une mort subite.

Les arythmies sont dues à des problèmes avec le système de conduction électrique du cœur. Il existe quatre groupes principaux d'arythmies :

- les battements supplémentaires, qui comprennent des contractions auriculaires prématurées, des contractions ventriculaires prématurées et des contractions jonctionnelles prématurées ;
- les tachycardies supraventriculaires, qui comprennent la fibrillation auriculaire, le *flutter* auriculaire et la tachycardie supraventriculaire paroxystique. Dans la tachycardie supraventriculaire paroxystique, le cœur se met tout à coup à battre très vite, 150 à 220 fois par minute. Une telle crise peut durer quelques secondes seulement, mais aussi se prolonger pendant plusieurs heures. Lors du *flutter* auriculaire, l'activité électrique du cœur est un peu plus structurée, alors que la contraction des oreillettes est en désordre total lors de la fibrillation auriculaire. Dans les deux cas, ces arythmies provoquent une diminution du débit cardiaque faisant surgir les principaux symptômes.
- les arythmies ventriculaires qui comprennent la fibrillation ventriculaire et la tachycardie ventriculaire. La tachycardie ventriculaire se produit lorsqu'un foyer électrique émet des pulsions rapides (plus de 100 battements par minute, souvent plus rapide). Suite à cela, les ventricules n'arrivent pas à se remplir correctement de sang, empêchant la circulation de s'effectuer normalement. Lors de la fibrillation ventriculaire, les ventricules du cœur n'arrivent plus à se contracter, les pulsations sont très rapides et totalement asynchrones. Le flux sanguin s'arrête alors complètement et le patient fait un arrêt cardiaque.
- les bradyarythmies, qui sont des rythmes cardiaques avec une fréquence inférieure à 60 battements/min, généralement dus à des anomalies de la conduction dans le cœur telles qu'une maladie des ganglions sinusaux ou un bloc auriculo-ventriculaire. Il existe un risque accru de mort cardiaque subite en raison du ralentissement ou de l'arrêt du cœur, et des chutes, en particulier chez les personnes âgées, dues à l'évanouissement.

2.2.3 Traitement des signaux

Le signal électrique détecté est de l'ordre du millivolt. La précision temporelle nécessaire est inférieure à 0,5 ms. Dans les systèmes d'acquisition numériques modernes, la fréquence d'échantillonnage peut atteindre 15 kHz, ce qui est généralement largement suffisant.

Le signal mesuré est généralement dégradé par de nombreuses perturbations, dans différentes bandes de fréquence. Un filtrage numérique permet d'éliminer les signaux de hautes fréquences secondaires à l'activité musculaire autre que cardiaque et aux interférences des appareils électriques. Un filtrage basse fréquence permet de diminuer les ondulations de la ligne de base secondaire à la respiration. Des filtres coupe-bande étroits permettent de supprimer l'influence de l'activité électrique domestique (50Hz ou 60Hz suivant les pays).

La qualité du signal peut être améliorée par le moyennage de plusieurs complexes, mais cette fonction entraîne des artefacts en cas d'irrégularités du rythme cardiaque ou d'extrasystoles, surtout ventriculaire. Cette technique de moyennage est particulièrement employée sur les appareils adaptés aux épreuves d'effort où le tracé est fortement perturbé par le patient en mouvement.

L'analyse et le traitement des électrocardiogrammes font appel à de multiples aspects du traitement des signaux (voir par exemple [4] pour quelques exemples, ou [12]), qu'il s'agisse de pré-traitements (filtrages, dé-bruitage,...) ou de problèmes plus spécifiques tels que par exemple l'estimation de rythme cardiaque ou la détection d'arythmies. Dans ce cours on se focalisera sur un nombre limité de techniques.

2.3 Données et outils logiciels

Il existe de nombreuses bases de données publiques fournissant des signaux ECG, et permettant donc de tester et comparer des algorithmes de traitement. Dans ce cours on se limitera à l'une d'entre elle, qui est une base de donnée généraliste de signaux physiologiques, la base Physionet.

2.3.1 La base de données Physionet

Les données sur lesquelles on travaillera dans ce cours sont issues de la plateforme en ligne Physionet [5]

<https://physionet.org/>.

PhysioNet est une ressource de recherche sur les signaux physiologiques complexes, créée en 1999 par le *National Institute of Health* (NIH). Les missions de PhysioNet sont de mener et catalyser des activités de recherche et d'éducation biomédicales, en partie en offrant un accès gratuit à de vastes collections de données physiologiques et cliniques et de logiciels libres apparentés. En coopération avec la conférence annuelle *Computing in Cardiology*, PhysioNet organise également une série annuelle de défis, axés sur la recherche sur des problèmes non résolus en sciences fondamentales et cliniques.

La plateforme PhysioNet est gérée par des membres du Laboratoire de physiologie computationnelle du MIT.

2.3.2 La boîte à outils WFDB

La *Toolbox* WFDB (disponible sous différents environnements de calcul scientifique, incluant MATLAB/Octave [10] et Python) est un ensemble de fonctions permettant de lire, écrire et traiter des signaux physiologiques et des séries chronologiques dans les formats utilisés par les bases de données PhysioBank (entre autres).

Nous utiliserons la version python, disponible sur

<https://github.com/MIT-LCP/wfdb-python>

dont la documentation peut être trouvée sur

<https://wfdb.readthedocs.io/en/latest/>

WFDB inclut principalement trois sous-paquets :

- Entrées sorties (`io`) : outils permettant de lire des données sous des formats standardisées, ainsi que différentes informations telles que ds annotations.
- Graphiques (`plot`) : outils permettant des représentations graphiques de diverses quantités
- Traitements (`processing`) : outils de traitement du signal, certains de bas niveau adaptés aux signaux physiologiques (ré-échantillonnage, filtrage,...), d'autres de plus haut niveau (détection de pics, détection du rythme cardiaque, détection du complexe QRS, et autres, incluant des outils de comparaison d'annotation)



Exemple : Jeu de données MITDB, sujet '100'.

Les quelques instructions ci-dessous chargent en mémoire (à partir du site de physionet) les 5000 premiers échantillons des signaux correspondant au sujet 100

```
rec = '100'
first_samp = 0
last_samp = 5000
full_record = wfdb.rdrecord(record_name=rec,
                             pn_dir='mitdb',
                             sampfrom=first_samp,
                             sampto=last_samp)
```

La variable de sortie `full_record` est un objet qui contient toutes les informations disponibles sur cet enregistrement, notamment la fréquence d'échantillonnage, les capteurs (en l'occurrence *MLII*, version modifiée de la dérivation *II*, et *V₅*), les unités,... Par exemple, les signaux sont stockés dans `full_record.p_signal`, leurs noms (les noms des capteurs) sont dans `full_record.sig_name`, la fréquence d'échantillonnage dans `full_record.fs` et ainsi de suite.

Les données sont également annotées par des experts, les annotations peuvent être chargées de la même façon

```
annot = wfdb.rdann(rec, 'atr',
                  pn_dir='mitdb',
                  sampfrom=first_samp,
                  sampto=last_samp,
                  return_label_elements=['symbol', 'description'])
```

La variable de sortie `annot` est aussi un objet qui contient les annotations disponibles, en particulier `annot.symbol` qui donne les symboles correspondant aux différentes annotations (N pour un battement normal, A pour une contraction prématurée de l'oreillette), et `annot.description` pour une description plus longue².

On peut tracer les signaux en utilisant des outils de la librairie classiques `matplotlib` ou les outils spécifiques de WFDB, comme les quelques lignes ci-dessous utilisées pour produire la Figure 2.5.

2. Voir <https://archive.physionet.org/physiobank/annotations.shtml> pour une description de ces annotations

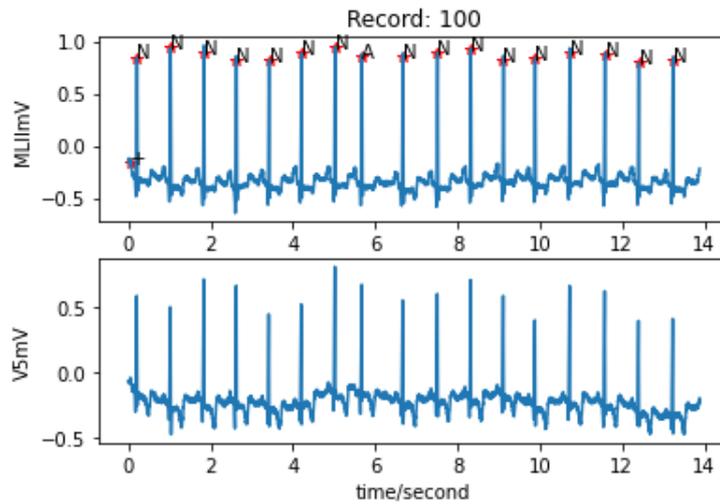


FIGURE 2.5 – Quelques secondes d'un signal EEG (sujet 100 de la base mitdb)

```
wfdb.plot.plot_wfdb(full_record,  
                    annot,  
                    time_units='seconds',  
                    plot_sym=True)
```

Ces outils permettent notamment de superposer les annotations, comme on le voit sur la Figure 2.5, où on voit par exemple que tous les battements ont été annotés comme normaux, à l'exception du huitième qui pointe une contraction anticipée des oreillettes.

3. Pré-traitements : filtrage

Comme on l'a vu, l'analyse et le traitement de signaux ECG font appel à de multiples aspects du traitement des signaux. On va se focaliser ici sur des techniques très classiques, utiles dans des phases d'acquisition et de pré-traitement des signaux. Ce sera l'occasion de revenir un peu plus en détails sur des techniques importantes telles que le filtrage ou l'estimation spectrale.

3.1 Signaux numériques : estimation spectrale et filtrage

3.1.1 Signaux numériques

On ne revient pas sur le problème d'échantillonnage, on traitera ici uniquement des signaux déjà numérisés, de la forme $x_n = x(n/\eta)$, $\eta \in \mathbb{R}_+$ étant la fréquence d'échantillonnage, et x étant un signal analogique (i.e. à temps continu), issu d'un filtrage passe-bas permettant son échantillonnage à la fréquence η . Les échantillons x_n sont donc indexés par un entier n .

- Dans un cadre théorique, on travaillera souvent dans le cadre de signaux de longueur infinie, i.e. $n \in \mathbb{Z}$, on sera amenés dans ce cas à utiliser des modèles mathématiques d'espace de signaux. Un cadre naturel est celui de l'espace $\ell^2(\mathbb{Z})$ des signaux dits d'énergie finie $\{x[n], n \in \mathbb{Z}\}$ tels que

$$\sum_{n=-\infty}^{\infty} |x[n]|^2 < \infty,$$

qui est muni du produit scalaire

$$\langle x, y \rangle_{\ell^2(\mathbb{Z})} = \sum_{n=-\infty}^{\infty} x[n] \bar{y}[n], \quad x, y \in \mathbb{C}^{\mathbb{Z}}, \quad (3.1)$$

donc on peut montrer (par l'inégalité de Cauchy-Schwarz) qu'il existe toujours, et de la norme associée $x \in \ell^2(\mathbb{Z}) \rightarrow \|x\| = \sqrt{\langle x, x \rangle_{\ell^2(\mathbb{Z})}}$.

Plus généralement, on utilise parfois les espaces $\ell^p(\mathbb{Z})$ des signaux $\{x[n], n \in \mathbb{Z}\}$ tels que $\sum_{n=-\infty}^{\infty} |x[n]|^p < \infty$, munis de la norme $x \rightarrow \|x\|_p = [\sum_n |x[n]|^p]^{1/p}$ et de l'espace $\ell^\infty(\mathbb{Z})$ des signaux x bornés, c'est à dire tels qu'il existe une constante notée $\|x\|_\infty$ telle que $|x[n]| \leq \|x\|_\infty$ pour tout n . L'application $x \in \ell^\infty(\mathbb{Z}) \rightarrow \|x\|_\infty$ est une norme sur $\ell^\infty(\mathbb{Z})$.

- Dans le cas de signaux de longueur finie, on supposera $n \in \{0, 1, \dots, N-1\}$, N étant le nombre

d'échantillons. Le cadre algébrique naturel est alors l'espace Hermitien \mathbb{C}^N , muni du produit scalaire

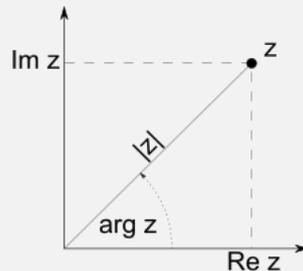
$$\langle x, y \rangle_{\mathbb{C}^N} = \sum_{n=0}^{N-1} x[n] \bar{y}[n], \quad x, y \in \mathbb{C}^N. \quad (3.2)$$

et de la norme associée $\|x\| = \sqrt{\langle x, x \rangle_{\mathbb{C}^N}}$.

Dans les calculs numériques, les signaux considérés seront toujours de longueur finie. On les considère souvent comme des restrictions à $\{0, \dots, N-1\}$ de signaux de longueur infinie périodiques de période N .

Remarque 3.1 — Bref rappel sur les nombres complexes. Les signaux d'intérêt sont généralement des signaux réels. Par contre nous verrons en étudiant la transformation de Fourier que des signaux complexes s'introduisent assez naturellement. On rappelle très brièvement ici quelques éléments sur les nombres complexes.

- Les nombres complexes sont utilisés pour représenter des couples de nombres réels : $(a, b) \in \mathbb{R}^2 \rightarrow z = a + ib$. Le nombre i est tel que $i^2 = -1$.
- La raison fondamentale pour utiliser des nombres complexes est qu'ils permettent de donner des solutions que n'en ont pas (ou pas suffisamment) si on se limite à des réels. En particulier, *Tout polynôme de degré N admet exactement N racines complexes.*
- Pour $z \in \mathbb{C}$, le nombre (réel positif ou nul) $|z| = \sqrt{a^2 + b^2}$ est le *module* de z (qui est aussi la longueur du vecteur (a, b) du plan).
- On écrit aussi sous forme polaire $z = |z|e^{i\theta}$. Le nombre $\theta \in]-\pi, \pi]$ est l'*argument* de z . $a = |z| \cos(\theta)$ est la *partie réelle* de z . $b = |z| \sin(\theta)$ est la *partie imaginaire* de z .
- Le conjugué (ou complexe conjugué) \bar{z} de $z = a + ib \in \mathbb{C}$ est le nombre $\bar{z} = a - ib = |z|e^{-i\theta}$, symétrique de z par rapport à l'axe réel dans le plan complexe.



Le nombre complexe $z = |z|(\cos(\theta) + i \sin(\theta))$ dans le plan complexe

Remarque 3.2 Nous utiliserons plus loin la transformation de Fourier discrète (TFD), et la transformation de Fourier finie (TFF). On donnera une définition prenant en compte une fréquence d'échantillonnage η (qui est prise égale à 1 dans la définition classique).

Comme on le verra, la TFD d'un signal numérique de longueur infinie est une fonction d'une variable réelle, périodique de période η , que l'on peut identifier à sa restriction à un intervalle de longueur η , l'intervalle $[-\eta/2, \eta/2]$ en l'occurrence. On sera donc amené à considérer l'espace $L^2([-\eta/2, \eta/2])$ des fonctions définies sur $[-\eta/2, \eta/2]$ de module carré intégrable. Cet espace est muni d'une structure d'espace de Hilbert par le produit scalaire

$$\langle f, g \rangle_{L^2} = \int_{-\eta/2}^{\eta/2} f(v) \bar{g}(v) dv, \quad f, g \in L^2([-\eta/2, \eta/2]), \quad (3.3)$$

duquel dérive la norme $\|f\| = \sqrt{\langle f, f \rangle_{L^2}}$.

3.1.2 Filtrage numérique

Les opérations de filtrage sont les opérations de base du traitement du signal. Comme on va le voir plus loin, le filtrage est utilisé pour modifier le "contenu fréquentiel" des signaux.

Définition 3.1 Un filtre numérique est un opérateur linéaire noté K_h , associant à un signal numérique x un autre signal $K_h x$, appelé signal filtré, de la forme

$$(K_h x)[n] = \sum_{k=-\infty}^{\infty} h[k]s[n-k], \quad (3.4)$$

pour tout n tel que la série soit convergente. La suite $h = \{h[n], n \in \mathbb{Z}\}$ est appelée *réponse impulsionnelle* du filtre.

Le filtre est dit *causal* si $h[n] = 0$ pour tout $n < 0$. Il est dit *stable* si $K_h x$ est borné pour tout x borné. Il est *réalisable* s'il est causal et stable.

La contrainte de causalité est fondamentale en traitement des signaux analogiques, car il n'est pas possible de construire de dispositif physique utilisant l'avenir. Les choses sont plus nuancées pour le traitement des signaux numériques, la causalité n'est plus une contrainte forte. Cependant on, préfère souvent utiliser quand même des filtres causaux.

La contrainte de stabilité est définie dans le cas de signaux de longueur infinie, et n'est donc théoriquement pas pertinente dans le cas de signaux de longueur finie. Cependant, dans certaines opérations, un filtre de longueur finie peut être théoriquement stable, mais présenter des instabilités numériques. C'est pour cela que cette notion conserve quand même une importance pratique.

Remarque 3.3 Il est facile de montrer qu'un filtre numérique est covariant par translations : étant donné un signal numérique x , si on note y' un translaté de x de la forme :

$$y[n] = x[n - n_0], \quad n \in \mathbb{Z}$$

pour tout entier n_0 , il vient immédiatement

$$(K_h y)[n] = (K_h x)[n - n_0].$$

Lorsque la réponse impulsionnelle h du filtre a un nombre fini de coefficients non-nuls, on dit que le filtre est à *réponse impulsionnelle finie* (filtre FIR). Dans le cas contraire, il est à *réponse impulsionnelle infinie* (filtre IIR).

■ **Exemple 3.1 — Filtrage d'un bruit blanc.** Un bruit blanc est une suite de nombres aléatoires indépendants identiquement distribués (par exemple suivant une loi Gaussienne). On peut effectuer un filtrage passe-bas par un simple lissage, en remplaçant chaque échantillon par la moyenne de cet échantillon et des $L - 1$ échantillons précédents :

$$y[n] = \frac{1}{L} \sum_{\ell=0}^{L-1} x[n-\ell].$$

Cette opération est bien un filtrage FIR au sens précédent, qui élimine (ou tout du moins réduit) les variations à une échelle inférieure à L . La réponse impulsionnelle est de longueur L , constante égale $= 1/L$. ■



Exemple : Filtrage moyenne mobile Les lignes ci-dessous illustrent le filtrage passe-bas par moyenne mobile. La fonction `lfilter` implémente (notamment) le filtrage par un filtre FIR. La séquence d'instructions ci-dessous génère une réalisation de longueur 250 d'un bruit blanc Gaussien (fonction `np.random.normal`), une réponse impulsionnelle finie constante sur son support (de longueur 10, puis 25) et effectue le filtrage correspondant. Celui-ci s'assimile à un lissage du signal initial à des échelles de 10 et 25 échantillons respectivement.

```
w = np.random.normal(loc=0.0, scale=1.0, size=250)
L1 = 10
h1 = np.ones(L1)/L1
L2 = 25
h2 = np.ones(L2)/L2
y1 = signal.lfilter(h1,1,w)
y2 = signal.lfilter(h2,1,w)
plt.plot(w)
```

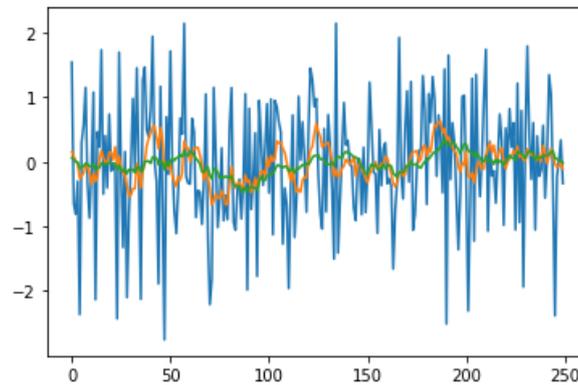


FIGURE 3.1 – Deux filtrages de type "lissage" d'un bruit blanc gaussien (bleu), à deux résolutions différentes (orange : $L = 10$; vert : $L = 25$).

```
plt.plot(y1)
plt.plot(y2)
```

Le résultat est représenté en Figure 3.1. Comme on peut le voir, le signal filtré avec $L = 10$ est une version "lissée" du bruit blanc, et le signal filtré avec $L = 25$ apparaît encore plus lisse.

■ **Exemple 3.2 — Filtrage de la composante basse-fréquence d'un signal ECG.** Le tracé ECG brut fait généralement apparaître une fluctuations lentes de ligne de base, aussi appelés instabilités (voir Figure 3.2, tracé du haut). La ligne de base, aussi appelée ligne isoélectrique doit correspondre à ce que serait la mesure en l'absence d'activité électrique. Pour ne pas fausser l'interprétation, il est fondamental de supprimer, ou tout du moins atténuer au maximum cette instabilité. Les tracés de la Figure 3.2 montrent le résultat de l'estimation de la ligne de base par lissage comme dans l'exemple 3.1 ci-dessus (superposée au signal ECG, tracé du milieu), puis le signal corrigé par soustraction de la ligne de base (tracé du bas). On peut constater que les principales fluctuations de la ligne de base ont bien été compensées, mais que le résultat n'est pas parfait. ■



Correction de ligne de base par lissage On utilise ici le même enregistrement de la base de données de MIT du site PhysioBank, de longueur 30 secondes environ, on se limite à une seule électrode (V5). On procède comme précédemment (filtrage par lissage) pour estimer cette fois une ligne de base, qu'on soustrait au signal. La longueur du lissage est fixée de façon *ad hoc* pour éliminer les variations trop rapides. La ligne de base estimée est superposée au signal (en orange dans la figure du milieu).

```
L = 250
h = np.ones(L)/L
V5_baseline = signal.lfilter(h,1,V5)
V5_corrected = V5 - V5_baseline
```

L'estimation que nous avons faite ici reste très grossière, le lissage étant lui même une forme de filtrage très rudimentaire. Pour aller plus loin, nous allons maintenant entrer davantage dans la théorie du filtrage, et voir comment construire des filtres de gabarits pré-définis.

3.1.3 Transformation de Fourier Discrète, transformation de Fourier Finie

Soit $x = \{x[n], n \in \mathbb{Z}\}$ un signal numérique. Sa transformée de Fourier Discrète (TFD) est la fonction périodique $v \rightarrow \hat{x}(v)$, de période η définie par

$$\hat{x}(v) = \sum_{n=-\infty}^{\infty} x[n]e^{-2i\pi vn/\eta}, \quad (3.5)$$

pour tout v tel que la série soit convergente. Ici η est un paramètre, qui peut être une fréquence d'échantillonnage, ou être pris égal à 1. La variable v est appelée "fréquence", et la variable $\omega = 2\pi v$ est appelée pulsation.

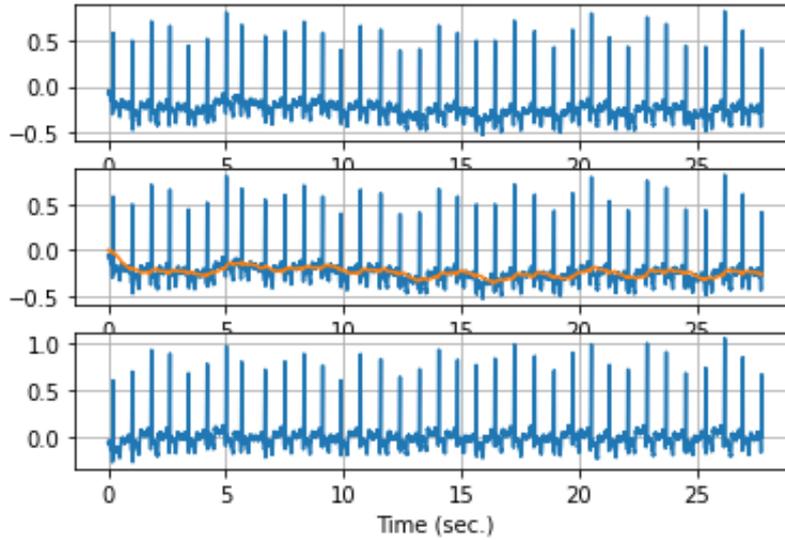


FIGURE 3.2 – Extrait de 30 secondes environ d'un signal ECG (haut), estimation de la ligne de base (milieu, courbe orange) et signal corrigé (bas).

Notons que compte tenu de la périodicité de \hat{x} , on se contente de considérer ses valeurs dans l'intervalle $[0, \eta]$, ou plus souvent $[-\eta/2, \eta/2]$.

Si on se place dans le cadre des signaux de $\ell^2(\mathbb{Z})$, La TFD est inversible. Plus précisément, on a

Théorème 3.4 La transformation de Fourier discrète est multiple d'une transformation unitaire de $\ell^2(\mathbb{Z})$ sur $L^2([-\eta/2, \eta/2])$: pour tout $x \in \ell^2(\mathbb{Z})$, $\hat{x} \in L^2([-\eta/2, \eta/2])$, et la formule de Parseval

$$\frac{1}{\eta} \int_{-\eta/2}^{\eta/2} |\hat{x}(v)|^2 dv = \sum_{-\infty}^{\infty} |x[n]|^2 \quad (3.6)$$

est vérifiée. La transformation inverse est donnée par

$$x[n] = \frac{1}{\eta} \int_{-\eta/2}^{\eta/2} \hat{x}(v) e^{2i\pi vn/\eta} dv . \quad (3.7)$$

Remarque 3.5 — Symétrie Hermitienne de la TFD. Notons que si x est un signal à valeurs réelles, alors \hat{x} possède la symétrie Hermitienne : pour tout v

$$\hat{x}(-v) = \overline{\hat{x}(v)} . \quad (3.8)$$

On montre alors que la transformation de Fourier inverse peut se simplifier comme

$$x[n] = 2 \operatorname{Re} \int_0^{\eta/2} \hat{x}(v) e^{2i\pi vn/\eta} dv , \quad (3.9)$$

où Re désigne la partie réelle.

Il existe une version de la transformation de Fourier adaptée aux signaux de longueur finie, qu'on appellera *Transformation de Fourier Finie* (TFF). En se plaçant dans l'espace \mathbb{C}^L , on définit la TFF d'un signal $x \in \mathbb{C}^L$ comme le vecteur $\hat{x} \in \mathbb{C}^L$ défini par

$$\hat{x}[k] = \frac{1}{\sqrt{L}} \sum_{\ell=0}^{L-1} x[\ell] e^{-2i\pi k\ell/L} . \quad (3.10)$$

Il est possible de démontrer que la TFF ainsi définie est inversible, avec une TFF inverse définie par

$$x[\ell] = \frac{1}{\sqrt{L}} \sum_{k=0}^{L-1} \hat{x}[k] e^{2i\pi k\ell/L}, \quad (3.11)$$

et une formule de Parseval (conservation de la norme)

$$\|\hat{x}\|^2 = \sum_{k=0}^{L-1} |\hat{x}[k]|^2 = \sum_{\ell=0}^{L-1} |x[\ell]|^2 = \|x\|^2. \quad (3.12)$$

Remarque 3.6 — Périodicité de la TFF. 1. La TFF \hat{x} de $x \in \mathbb{C}^L$ est un vecteur de \mathbb{C}^L . Ceci étant, on peut aussi la voir comme la restriction à $\{0, \dots, L-1\}$ d'un vecteur de longueur infinie, qui vérifie la propriété de périodicité $\hat{x}[k+L] = \hat{x}[k]$ pour tout $k \in \mathbb{Z}$.
2. Dans ce contexte, on peut aussi interpréter les éléments $\hat{x}[k]$ d'indice $k > L/2$ comme des éléments associés à des fréquences $k-L$ négatives.

Remarque 3.7 — Symétrie Hermitienne de la TFF de signaux réels. Comme la TFD, la TFF possède une symétrie Hermitienne : si $x \in \mathbb{C}^L$, alors

$$\hat{x}[L-k] = \hat{x}[-k] = \overline{\hat{x}[k]}. \quad (3.13)$$



TFF et représentation graphique.

Il existe différentes implémentations de la TFF sous Python. ON utilisera ici celle du package `fft` de `numpy`. La séquence d'instruction

```
fs = 8000
T = 2
L = np.round(fs*T)
times = np.linspace(0,T,L)
f0 = 440
sig = np.cos(2*np.pi*f0*t) + np.cos(4*np.pi*f0*t)/2 + np.cos(6*np.pi*f0*t)/3
fsig = np.fft.fft(sig)
```

calcule la transformée de Fourier discrète d'une somme de trois sinusoïdes de fréquences multiples de $f_0=16$. `sig` est ici un vecteur réel, et \hat{x} est un vecteur complexe, on n'en représentera que le module.

Par défaut, les fréquences calculées par la TFF sont linéairement espacées entre la fréquence nulle et la fréquence d'échantillonnage. C'est ce qui est représenté sur le premier graphique de la Figure 3.3. Il est plus courant de représenter la TFF sur un axe de fréquences symétrique par rapport à l'origine, on peut utiliser la fonction `fftshift` (qui ne calcule pas de transformée de Fourier, elle se contente de réordonner une transformée de Fourier déjà calculée). Finalement, pour des signaux réels, la transformée de Fourier possède une symétrie Hermitienne, et on se contente généralement de représenter uniquement les fréquences positives. Ces trois représentations, visibles sur la Figure 3.3 sont obtenues par les instructions suivantes

```
plt.subplot(3,1,1)
freqs = np.linspace(0,fs,L)/1000
plt.plot(freqs,np.abs(fsig))
plt.grid()
plt.subplot(3,1,2)
freqs = np.linspace(-fs/2,fs/2,L)/1000
plt.plot(freqs,np.fft.fftshift(np.abs(fsig)))
plt.grid()
plt.subplot(3,1,3)
Lover2 = np.int(np.round(L/2))
freqs = np.linspace(0,fs/2,Lover2)/1000
plt.plot(freqs,np.abs(fsig[0:Lover2]))
plt.xlabel('Freq. (kHz)')
plt.grid()
```

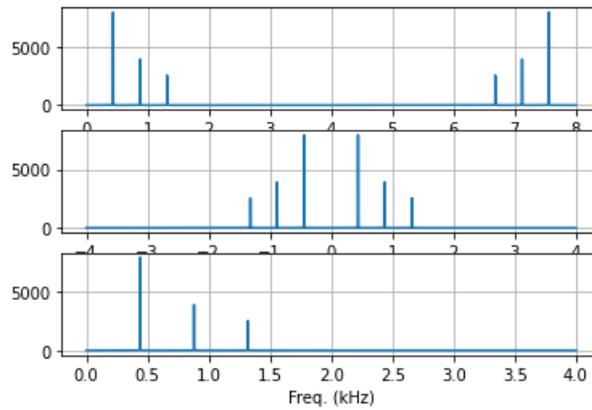


FIGURE 3.3 – Trois représentations de la transformée de Fourier d’une somme de 3 sinusoïdes : $[0, \eta]$ (haut), $[-\eta/2, \eta/2]$ (milieu) et $[0, \eta/2]$ (bas).

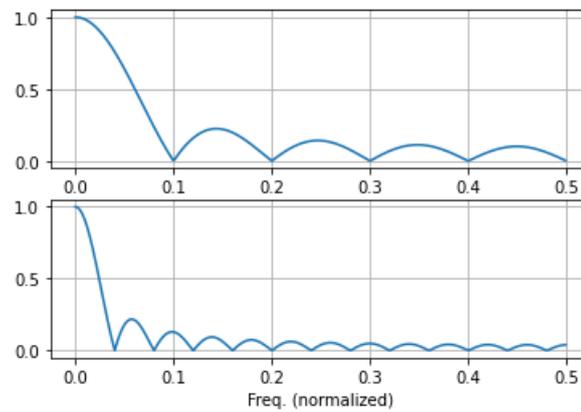


FIGURE 3.4 – Fonctions de transfert des deux filtres de lissage de l’exemple 3.1.

3.1.4 Transformation de Fourier et filtrage

Une propriété remarquable de la transformation de Fourier discrète est qu’elle transforme filtrage en multiplication point par point. En effet, étant donné un filtre de réponse impulsionnelle h , son action s’écrit dans le domaine de Fourier

$$\widehat{K_h x}(v) = \hat{h}(v)\hat{x}(v), \quad v \in [-\eta/2, \eta/2]. \quad (3.14)$$

La fonction

$$m(v) = \hat{h}(v)$$

est appelée fonction de transfert, ou impédance, ou réponse fréquentielle du filtre. On voit donc que l’effet d’un filtre est bien de modifier le contenu fréquentiel des signaux, par multiplication par cette fonction de transfert. L’action d’un filtre $T = K_h$ de fonction de transfert $m = \hat{h}$ s’écrit donc aussi sous la forme

$$Tx[n] = \frac{1}{\eta} \int_{-\eta/2}^{\eta/2} m(v)\hat{x}(v)e^{2i\pi vn/\eta} dv. \quad (3.15)$$

Les fonctions de transfert des deux filtres de lissage décrits dans l’exemple 3.1 sont donnés dans la figure 3.4. On voit bien que la fonction de transfert est proche de 1 pour des fréquences proches de 0, puis décroît, plus rapidement pour $L = 32$ que pour $L = 8$. On peut donc dire que ce filtre a tendance à préserver

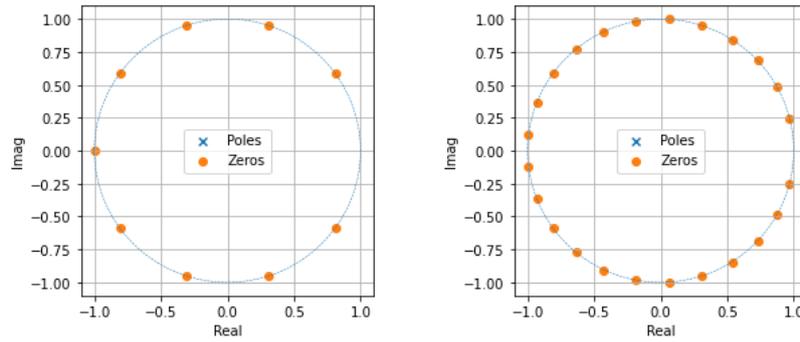


FIGURE 3.5 – Représentation des zéros des deux filtres de lissage de l'exemple 3.1.

les basses fréquences et atténuer les hautes fréquences (on parle de filtre passe-bas). Par contre on note la présence de "rebonds", qui font que ce comportement n'est pas régulier.

Plus précisément, ces rebonds sont dus à la présence de "zéros" : des valeurs v_k de la fréquence pour lesquelles la fonction de transfert s'annule. On représente souvent ces zéros dans le plan complexe, en y positionnant les nombres complexes

$$z_k = e^{2i\pi v_k/\eta} .$$

On a ainsi représenté dans la Figure 3.5 les zéros des deux filtres de lissage évoqués plus haut.

De tels filtres de lissages ne sont pas capables de supprimer (ou atténuer fortement) des domaines fréquentiels donnés, tout en en préservant d'autres. Concrètement, on aimerait pouvoir disposer de filtres (par exemple filtres passe-bas) idéaux, au sens où leur fonction de transfert soit l'indicatrice d'un intervalle symétrique par rapport à l'origine.

On voit bien d'après ce qui précède que la fonction de transfert d'un filtre FIR causal de réponse impulsionnelle $\{h[0], \dots, h[L]\}$ est en fait un polynôme trigonométrique (c'est à dire un polynôme de la variable complexe $e^{-2i\pi v/\eta}$) de degré L :

$$\hat{h}(v) = \sum_{n=0}^{L-1} h[n] e^{-2i\pi v n/\eta} .$$

Or il est connu que de tels polynômes trigonométriques peuvent difficilement approximer des fonctions discontinues (c'est le phénomène de Gibbs). On doit alors recourir à des filtres à réponse impulsionnelle infinie.

Il existe une famille de tels filtres, appelée famille de filtres récursifs, dont l'utilisation ne fait toutefois intervenir qu'un nombre fini d'opérations. Ces filtres associent à un signal d'entrée x un signal de sortie y solution de l'équation

$$\sum_{k=0}^K a[k] y[n-k] = \sum_{\ell=0}^L b[\ell] x[n-\ell] , \quad \forall n \in \mathbb{Z} , \quad (3.16)$$

appelée équation aux différences où les nombres $a[k]$ et $b[\ell]$ sont des coefficients qui caractérisent le filtre, avec en particulier $a[0] = 1$. Cette équation s'écrit aussi sous la forme

$$y[n] = \sum_{\ell=0}^L b[\ell] x[n-\ell] - \sum_{k=1}^K a[k] y[n-k] ,$$

qui fait bien apparaître la nature causale du filtre.

Remarque 3.8 — Diagramme d'un filtre d'ordre fini. Un filtre d'ordre fini est un filtre qui ne requiert qu'un nombre fini d'opérations pour le calcul de chaque échantillon du signal filtré. Un filtre FIR est par construction d'ordre fini, son ordre est égal au nombre de coefficients. Un filtre récursif tel que défini par l'équation (3.16) est aussi d'ordre fini, l'ordre est égal à $\max(K, L)$.

On peut associer à ces filtres un diagramme, qui en schématise les opérations. Par exemple, la FIGURE 3.6 représente le diagramme d'un filtre FIR. Les boîtes triangulaires représentent la multiplication d'un échantillon par un nombre, les boîtes rectangulaires (z^{-1}) représentent un décalage d'un échantillon (comme on le verra plus tard avec la transformation en z). La FIGURE 3.7 représente deux diagrammes équivalents d'un filtre FIR. Ces deux figures sont tirées de [11].

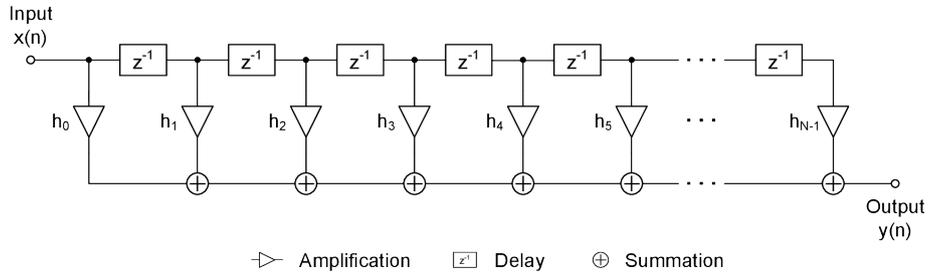


FIGURE 3.6 – Diagramme en blocs d'un filtre à réponse impulsionnelle finie.

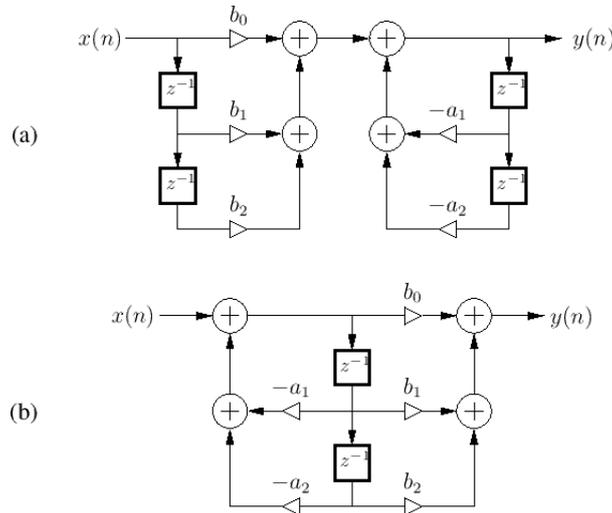


FIGURE 3.7 – Diagrammes en blocs d'un filtre à réponse impulsionnelle infinie : formes directes I (a) et II (b).

Par transformation de Fourier, l'équation (3.16) peut s'écrire

$$\hat{a}(v)\hat{y}(v) = \hat{b}(v)\hat{x}(v), \quad v \in [-\eta/2, \eta/2],$$

qui permet d'exprimer la fonction de transfert comme fonction rationnelle trigonométrique

$$m(v) = \frac{\sum_{\ell=0}^L b[\ell]e^{-2i\pi v\ell/\eta}}{\sum_{k=0}^K a[k]e^{-2i\pi vk/\eta}}, \quad (3.17)$$

c'est à dire le quotient de deux polynômes trigonométriques de degrés respectifs L et K . L'étude de telles fonctions de transfert repose généralement sur la transformation en z , que nous allons voir plus loin. On verra en particulier que sous certaines conditions sur les coefficients $a[k]$ et $b[\ell]$, il est possible de montrer que l'opérateur $T : x \rightarrow y$ solution de l'équation (3.16) est bien un filtre stable et causal.

La fonction de transfert $m(z)$ d'un filtre est à valeurs complexes, et n'est pas toujours facile à interpréter. On s'intéresse alors tout d'abord à son module ou son module au carré. On en voit un exemple ci-dessous.

■ **Exemple 3.3 — Filtre passe-bas de Butterworth d'ordre 1.** On considère le filtre dont le module carré de la fonction de transfert est donné par

$$|m(\nu)|^2 = \frac{1 + \cos(2\pi\nu/\eta)}{(1 + \cos(2\pi\nu/\eta)) + k(1 - \cos(2\pi\nu/\eta))},$$

où k est une constante. On voit en particulier que $|m(0)|^2 = 1$ et $|m(\pm\eta/2)|^2 = 0$. Ce filtre est un filtre passe-bas, qui préserve les basses fréquences (les fréquences proches de 0) et atténue les hautes fréquences (les fréquences proches de $\pm\eta/2$).

Le rôle du paramètre k est de déterminer le domaine des fréquences qui seront atténuées.

Il s'agit ici d'un filtre d'ordre 1, il existe des filtres de Butterworth d'ordres plus élevés, qui permettent une coupure plus nette dans le domaine fréquentiel, au prix de coûts en calcul plus élevés (et d'instabilités numériques qu'il faut contrôler). Les modules carrés de quelques exemples de fonctions de transfert de filtre passe-bas de Butterworth d'ordres plus élevés sont représentés en Figure 3.9. ■

3.1.5 Transformation en z

La transformation de Fourier discrète donne une représentation d'un signal dans le domaine fréquentiel. La transformation en z permet d'étendre cette représentation à un espace plus grand, dont on va voir qu'il permet une interprétation physique. Notons tout d'abord que la fonction exponentielle $\nu \rightarrow e^{2i\pi\nu/\eta}$ transforme le segment $[-\eta/2, \eta/2]$ des fréquences dans le cercle de rayon 1 dans le plan complexe, comme représenté dans la figure 3.8. Le principe de la transformation en z est de ne pas se limiter à ce cercle de rayon 1, mais de considérer le plan complexe tout entier. Comme on peut le voir sur la figure, en écrivant le nombre complexe z en coordonnées polaires $z = e^{2i\pi\nu/\eta}$, l'argument (aussi appelé phase) décrit la fréquence.

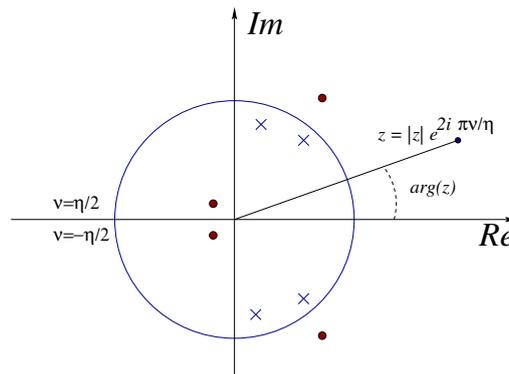


FIGURE 3.8 – Le plan complexe utilisé dans la transformation en z . Le cercle de rayon 1, donné par $z = e^{2i\pi\nu/\eta}$ reproduit le segment fréquentiel $[-\eta/2, \eta/2]$.

La transformation en z associée à un signal numérique une fonction de cette variable complexe z . Elle peut être vue comme un prolongement de la TFD dans le plan complexe, et ses propriétés en font un outil très utilisé en traitement du signal. On donne ci-dessous la définition, avant de discuter les principales propriétés mathématiques.

Séries de Laurent, transformation en z

Étant donné un signal numérique $\{x[n], n \in \mathbb{Z}\}$, il existe des cas où sa transformée de Fourier discrète n'est pas définie au sens classique. On a parfois recours à une alternative, la transformée en z , dont on décrit ci-dessous les propriétés essentielles, sans entrer dans les détails.

Définition 3.2 Soit $x = \{x[n], n \in \mathbb{Z}\}$ un signal numérique. Sa transformée en z est la série de Laurent

$$X(z) = \sum_{n=-\infty}^{\infty} x[n]z^{-n}, \quad (3.18)$$

définie dans une couronne de convergence (éventuellement vide) $r_1 < |z| < r_2$.

On sait d'après des résultats généraux sur les séries de Laurent que X est holomorphe dans sa couronne de convergence. Inversement, étant donnée une fonction X holomorphe dans une couronne $r_1 < |z| < r_2$, elle admet un unique développement en série de Laurent. De plus, on a le résultat classique suivant :

Proposition 3.9 Le rayon de convergence ρ de la série entière $z \rightarrow \sum_0^\infty a[n]z^n$ est donné par

$$\frac{1}{\rho} = \limsup_{n \rightarrow \infty} |a_n|^{1/n} = \limsup_{n \rightarrow \infty} \left| \frac{a_n}{a_{n+1}} \right| .$$

Le premier critère est le critère de Cauchy, le second est le critère de d'Alembert.

On en déduit immédiatement la couronne de convergence de la transformée en z d'un signal numérique :

Corollaire 3.10 Soit X la transformée en z de la série x . Les bornes de la couronne de convergence de X sont données par

$$r_1 = \limsup_{n \rightarrow \infty} |x[n]|^{1/n} , \quad \frac{1}{r_2} = \limsup_{n \rightarrow \infty} |x[-n]|^{1/n} . \quad (3.19)$$

■ **Exemple 3.4** On dit qu'un signal numérique x est *causal* si $x[n] = 0$ pour tout $n < 0$. Inversement, x est dit *anticausal* si $x[n] = 0$ pour tout $n \geq 0$. Supposons que x soit causal. Alors il est évident que $r_2^{-1} = 0$, de sorte que la transformée en z de x est bien définie dans le domaine $|z| > r_1$, c'est à dire à l'extérieur d'un cercle de rayon r_1 .

De même, si x est anticausal, $r_1 = 0$, et $X(z)$ est bien défini à l'intérieur du cercle de rayon r_2 . ■

Inversion de la transformation en z

Il existe plusieurs techniques permettant d'inverser une transformation en z . La plus simple consiste à expliciter un développement en série de Laurent de la fonction X considérée. Le développement en série de Laurent étant unique, ceci fournit directement une transformée inverse.

■ **Exemple 3.5** Prenons l'exemple de la fonction

$$X(z) = \frac{z}{z - z_0} , \quad |z| < |z_0| ;$$

on peut alors écrire, pour $|z| < |z_0|$,

$$X(z) = \frac{z}{z - z_0} = -\frac{z}{z_0} \frac{1}{1 - z/z_0} = \frac{z}{z_0} \sum_{n=0}^{\infty} \left(\frac{z}{z_0} \right)^n = \sum_{n=-\infty}^{-1} z_0^n z^{-n} ,$$

ce qui, conjugué à l'unicité du développement en série de Laurent, fournit

$$x[n] = \begin{cases} z_0^n & \text{pour } n < 0 \\ 0 & \text{sinon .} \end{cases}$$

■

Une alternative consiste à utiliser la TFD. Soit X la transformée en z d'un signal x , et soit r un nombre tel que $r_1 < r < r_2$. Calculons

$$\int_{-1/2}^{1/2} X(re^{2i\pi\theta}) e^{2i\pi\theta} d\theta = \sum_m x[m] r^{-m} \int_{-1/2}^{1/2} e^{2i(n-m)\pi\theta} d\theta = r^{-n} x[n] .$$

On peut donc écrire

$$x[n] = r^n \int_{-1/2}^{1/2} X(re^{2i\pi\theta}) e^{2i\pi\theta} d\theta .$$

Par un changement de variables complexes $z = re^{2i\pi\theta}$, on obtient donc

Proposition 3.11 Soit x un signal numérique, et soit X sa transformée en z , définie dans la couronne de convergence $r_1 < |z| < r_2$. Les coefficients de s sont donnés par

$$x[n] = \frac{1}{2i\pi} \oint_C X(z) z^n \frac{dz}{z}, \quad (3.20)$$

où C est un cercle centré sur l'origine du plan complexe, de rayon $r \in]r_1, r_2[$.

On a généralement recours à la méthode des résidus pour calculer de telles intégrales.

Transformation en z et filtrage numérique

L'un des intérêts de la transformation en z est son comportement vis à vis des transformations simples, et en particulier des translations. Etant donnée un signal $\{x[n], n \in \mathbb{Z}\}$, et un signal translaté $\{x'[n], n \in \mathbb{Z}\}$ donné par $x'[n] = x[n-k]$, on voit immédiatement que leurs transformées en z sont reliées par $X'(z) = z^k X(z)$. Le corollaire immédiat est le comportement de la transformation en z vis à vis du filtrage numérique.

Étant donné un signal numérique x et une filtre numérique de réponse impulsionnelle h , et $y = K_h x$ le signal filtré, alors pour tout z à l'intérieur de l'intersection des couronnes de convergence des transformées X et H de x et h respectivement, on a

$$Y(z) = \sum_n x'[n] z^{-n} = \sum_n \sum_k h[k] z^{-k} x[n-k] z^{-(n-k)},$$

de sorte que l'on a

$$Y(z) = H(z)X(z). \quad (3.21)$$

La fonction H est elle aussi appelée *fonction de transfert* du filtre.

En particulier, dans le cas d'un filtre récursif comme précédemment, on a

$$Y(z) = \frac{\sum_{\ell=0}^L b_\ell z^{-\ell}}{\sum_{k=0}^K a_k z^{-k}} X(z), \quad (3.22)$$

c'est à dire que la transformée en z de h , qui est aussi la fonction de transfert du filtre, prend la forme d'une fraction rationnelle

$$H(z) = \frac{\sum_{\ell=0}^L b_\ell z^{-\ell}}{\sum_{k=0}^K a_k z^{-k}} = \frac{B(z)}{A(z)}, \quad (3.23)$$

où A et B sont les transformées en z des suites de coefficients a et b .

Cette expression est à rapprocher de l'expression (3.17) obtenue avec la TFD. En fait (3.17) est le cas particulier de (3.23) quand on contraint la variable z à rester sur le cercle unité dans le plan complexe.

Le plan complexe peut donc être vu comme un prolongement du cercle unité, que l'on a associé au segment fréquentiel $[-\eta/2, \eta/2]$. Voir Figure 3.8.

3.1.6 Factorisation de filtres

L'expression (3.23) de la fonction de transfert d'un filtre récursif fait donc apparaître deux polynômes dans la variable complexe z^{-1} , $P(z^{-1}) = B(z)$ au numérateur et $Q(z^{-1}) = A(z)$ au dénominateur. On sait que tout polynôme d'une variable complexe est scindé, et peut donc être totalement factorisé (et caractérisé par ses racines, à une constante multiplicative près). On peut donc écrire la fonction de transfert $H(z)$ sous la forme factorisée

$$H(z) = g \frac{(1 - z_1 z^{-1})(1 - z_2 z^{-1}) \dots (1 - z_L z^{-1})}{(1 - \zeta_1 z^{-1})(1 - \zeta_2 z^{-1}) \dots (1 - \zeta_K z^{-1})} = g z^{K-L} \frac{(z - z_1)(z - z_2) \dots (z - z_L)}{(z - \zeta_1)(z - \zeta_2) \dots (z - \zeta_K)}, \quad (3.24)$$

où

- $g = H(0)$ est appelé gain du filtre.
- z_1, \dots, z_L sont les *zéros* du filtre
- ζ_1, \dots, ζ_K sont les *pôles* du filtre.

Une représentation graphique se trouve en Figure 3.8. C'est cette factorisation que nous allons maintenant exploiter.

Définition 3.3 On dit qu'un filtre causal K_h est d'ordre fini si K_h peut être réalisé comme un filtre récursif comme en (3.16).

L'équation (3.24) montre que de tels filtres d'ordre fini peuvent être caractérisés par les pôles et les zéros de leur fonction de transfert $H(z)$. On parle de **représentation pôles-zéros**.

Comme on l'a vu plus haut, la fonction de transfert elle-même est souvent difficilement interprétable, son module carré est plus facile à appréhender. En fait, la construction d'un filtre repose souvent sur la construction préalable d'un module carré de fonction de transfert, dont on extrait une "racine module carré". Le résultat qui suit montre que ceci est possible, et produit sous certaines conditions des filtres stables et causaux.

Proposition 3.12 1. Soit K_h un filtre causal stable d'ordre fini. Alors le module carré de sa fonction de transfert $M = |\hat{h}|^2$ est une fonction rationnelle non-négative de la variable réelle $w = \cos(2\pi v/\eta) \in [-1, 1]$:

$$|\hat{h}(v)|^2 = \frac{\mathcal{N}(w)}{\mathcal{D}(w)},$$

où \mathcal{N} et \mathcal{D} sont deux polynômes.

2. Inversement, étant donnée une fonction rationnelle non-négative $v \rightarrow M(v) \in \mathbb{R}^+$ de $w = \cos(2\pi v/\eta)$ comme ci-dessus, il existe un filtre causal stable d'ordre fini K_h tel que $|\hat{h}(v)|^2 = M(v)$.

Remarque 3.13 Le filtre K_h n'est pas unique, sa construction repose sur un bon choix des zéros et des pôles.

1. Les contraintes de causalité et de stabilité du filtre obligent à choisir les pôles ζ_k de la fonction de transfert à l'intérieur du disque unité dans le plan complexe. La proposition ci-dessus montre que cela est toujours possible.
2. Il reste la liberté de choisir les zéros z_k à l'intérieur ou à l'extérieur du disque unité pour former la fonction d . Choisir tous les zéros à l'intérieur du disque unité conduit aux filtres dits à *phase minimale*.

Un exemple est donné par la famille des *filtres de Butterworth* (dans leur version numérique), qui constituent des approximations de filtres passe-bas idéaux. En introduisant de nouveau la variable $w = \cos(2\pi v/\eta)$, les filtres de Butterworth sont caractérisés par une fonction de transfert telle que

$$|\hat{h}(v)|^2 = W(w) = \frac{(w+1)^L}{(w+1)^L + k(1-w)^L}, \quad (3.25)$$

où L est un entier positif fixé, et c une constante positive, qui détermine la fréquence de transition entre bande passante et bande atténuée, via la relation

$$k = 3 \left(\frac{1 + \cos(2\pi v_0/\eta)}{1 - \cos(2\pi v_0/\eta)} \right)^L$$

(par exemple, $k = 3$ correspond à une fréquence de coupure égale à $v_0 = \eta/4$).



Filtres de Butterworth.

Le package `scipy.signal` implémente de nombreux outils de filtrage et de *design* de filtres. Les filtres IIR les plus couramment employés sont probablement les filtres de Butterworth (voir aussi les filtres de Chebyshev). La fonction `butter` de `scipy.signal` génère les coefficients de filtres de Butterworth d'ordre et fréquence de coupure donnés.

Dans les lignes qui suivent on génère un filtre passe-bas de Butterworth d'ordre 4 et de fréquence de coupure égale à 880 Hz (pour une fréquence d'échantillonnage de 8kHz), et on représente sa réponse fréquentielle.

```
fs = 8000
b_ord = 4
fc = 880
b,a = signal.butter(b_ord,fc,fs=fs)
f,m = signal.freqz(b,a)
f = f/1000
```

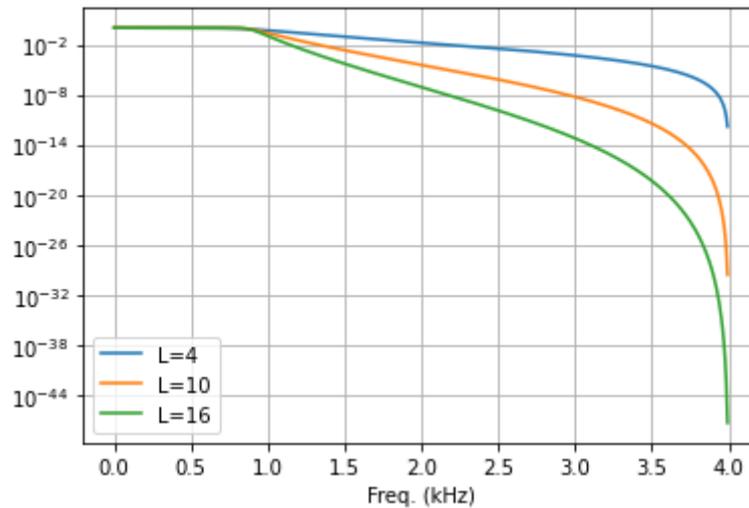


FIGURE 3.9 – Réponses fréquentielles de filtres passe-bas de Butterworth d'ordres 4, 10 et 16, de fréquence de coupure égale à 880 Hz. La réponse en fréquence est très proche de 1 et "plate" dans la bande 0-880 Hz, et décroît ensuite d'autant plus vite que l'ordre est élevé.

```
plt.semilogy(f*fs/2/np.pi,np.abs(m))
plt.grid()
```

La Figure 3.9 représente les réponses fréquentielles de filtres de Butterworth d'ordres 4, 10 et 16 respectivement. Il faut signaler le risque d'instabilités numériques liés aux filtres d'ordres élevés (présents ici pour les ordres 10 et 16). Dans ce cas, il est préférable d'utiliser une autre forme des filtres, appelée *second order section* (sos), basée sur des factorisations du filtre en filtres plus courts. C'est cette syntaxe qui a été utilisée pour les filtres d'ordres 10 et 16 :

```
b_ord = 10
sos = signal.butter(b_ord,fc,fs=fs,output='sos')
f,m = signal.sosfreqz(sos)
f = f/1000
plt.semilogy(f*fs/2/np.pi,np.abs(m))
```

On représente en Figure 3.10 les zéros et pôles des filtres ainsi générés. La syntaxe pour obtenir ces zéros et pôles est de la forme

```
z,p,k = signal.tf2zpk(b,a)
```

ou

```
z,p,k = signal.sos2zpk(sos)
```

suivant le format utilisé pour générer les filtres.

Comme on peut le voir, il y a bien un zéro pour $z = -1$, qui correspond à $v = \pm\eta/2$ (ce zéro est en fait multiple, d'ordre égal à l'ordre du filtre), et des pôles à l'intérieur du disque unité.

De même, on représente en Figure 3.11 la réponse fréquentielle et le diagramme pôle-zéro d'un filtre passe-bande de Butterworth d'ordre 10, de fréquence de bande passante [1500,2000] Hz (fréquence d'échantillonnage : $\eta = 8\text{kHz}$).

3.2 Applications au filtrage de signaux ECG

On va maintenant voir des applications des techniques de filtrage vues plus haut au pré-traitement de signaux ECG. Ces différentes étapes de filtrage permettent de supprimer un certain nombre d'artefacts qui peuvent perturber l'analyse, comme une ligne de base (composante très basse fréquence) provenant de la respiration du sujet, une composante oscillante à une fréquence bien déterminée correspondant à l'électricité domestique, ou diverses autres sources. Nous allons commencer ce chapitre par une discussion de la problématique d'estimation spectrale

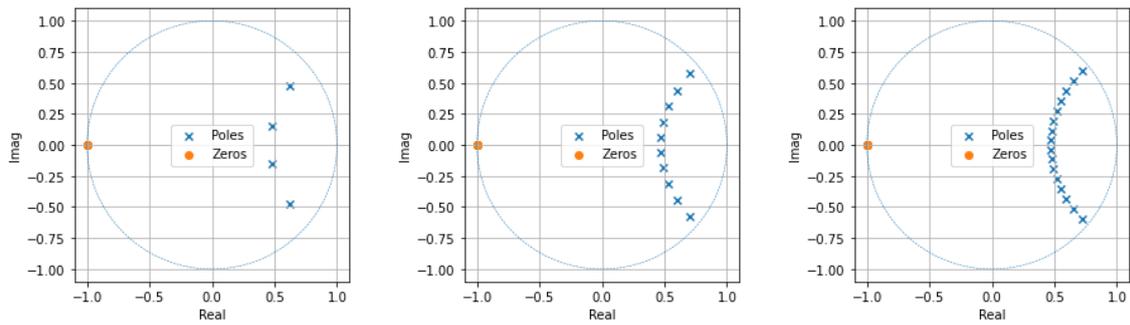


FIGURE 3.10 – Représentation des zéros et des pôles dans le plan complexe pour des filtres passe-bas de Butterworth d’ordres 4, 10 et 16, de fréquence de coupure égale à 880 Hz.

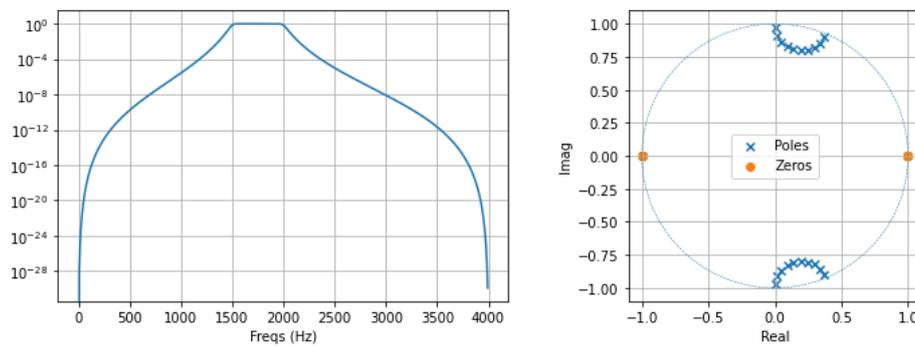


FIGURE 3.11 – Représentation fréquentielle (gauche) et représentation des zéros et des pôles dans le plan complexe (droite) pour un filtre passe-bande de Butterworth d’ordre 10, de fréquence de bande passante [1500,2000] Hz.

3.2.1 Estimation spectrale

On entend généralement par spectre d’un signal une mesure de son "contenu fréquentiel". Ceci suppose que ce dernier ne varie pas au cours du temps (au moins dans un intervalle suffisamment long). Cette hypothèse est appelée stationnarité, elle peut être définie dans le cadre de modèles de signaux aléatoires, qu’on n’abordera pas dans ce cours.

Le problème d’estimation spectrale est un problème classique en statistique et traitement du signal. Le problème est le suivant : on dispose d’une signal de longueur finie, que l’on suppose être une observation de longueur finie d’un signal stationnaire. La question est la suivante : comment estimer le spectre de celui-ci à partir de cette unique observation.

Étant donné un signal de longueur L échantillons, on dispose de plusieurs approches pour estimer un spectre.

— **Le Périodogramme** : basé sur une simple transformation de Fourier finie, élevée au module carré :

$$S\left(\frac{k\eta}{L}\right) = \frac{1}{L} \left| \sum_{\ell=0}^{L-1} x[\ell] e^{-2i\pi k\ell/L} \right|^2 = \frac{1}{L} |\hat{x}[k]|^2 \quad (3.26)$$

Le spectre estimé par le périodogramme est souvent très "irrégulier".

— **Le Périodogramme moyenné** : le signal est découpé en segments de longueur N , un spectre est estimé par calcul du périodogramme sur chacun des segments, puis et on effectue les moyennes des estimées correspondantes du spectre

$$S\left(\frac{k\eta}{N}\right) = \frac{1}{P} \sum_{p=0}^{P-1} \frac{1}{N} \left| \sum_{n=0}^{N-1} x[n+p\delta] e^{-2i\pi kn/N} \right|^2 \quad (3.27)$$

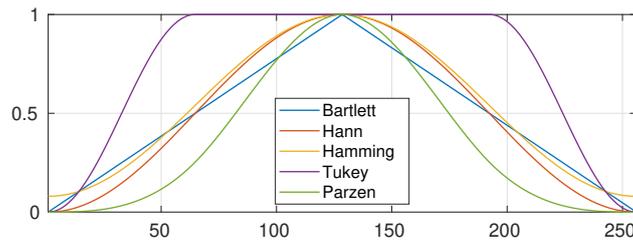


FIGURE 3.12 – Quelques exemples de fenêtres couramment utilisées pour l'estimation spectrale

Le spectre estimé correspondant est généralement plus "lisse", mais souffre de biais introduit par l'utilisation d'une fenêtre rectangulaire.

- **Le Périodogramme de Welch** : il est similaire au périodogramme moyenné, à ceci près que des fenêtres sont utilisées pour la segmentation, au lieu de la fenêtre rectangulaire. En notant $h \in \mathbb{R}^N$ la fenêtre, on définit cette fois

$$S\left(\frac{k\eta}{N}\right) = \frac{1}{P} \sum_{p=0}^{P-1} \left| \sum_{n=0}^{N-1} h[n+p\delta]x[n+p\delta]e^{-2i\pi kn/N} \right|^2. \quad (3.28)$$

Le périodogramme de Welch est plus lisse que le périodogramme classique, et le biais introduit par la fenêtre rectangulaire est réduit quand une autre fenêtre est utilisée (voir la Figure 3.12 pour quelques fenêtres classiques).

Cependant, le périodogramme de Welch (et dans une moindre mesure le périodogramme moyenné) demande de choisir avec soin quelques paramètres, notamment la fenêtre et la taille de celle-ci, ainsi que le degré de recouvrement des fenêtres. La régularité du périodogramme dépend de ces paramètres.

L'estimation spectrale est un préalable à la plupart des tâches de traitement du signal. Nous allons l'illustrer dans le cas de signaux ECG.

On représente dans les Figures 3.13 et 3.14 le périodogramme et le périodogramme de Welch d'un enregistrement EEG (deux capteurs, $MLII$ et V_5 , échantillonnage à $\eta = 360Hz$). Comme on pouvait s'y attendre, le périodogramme est peu lisible, très bruité. On peut toutefois y remarquer la présence de pics à des fréquences multiples de 60Hz (qui est la fréquence du courant électrique domestique aux USA). On peut également voir que les basses fréquences sont dominantes, ce qui peut être interprété comme le signe d'une ligne de base, qu'on cherchera à éliminer ou en tous cas atténuer.

Le périodogramme de Welch (calculé avec des valeurs par défaut des paramètres) est beaucoup plus lisse. On y retrouve la dominance des basses fréquences (avec deux fréquences particulières), et on y retrouve aussi les pics correspondant aux fréquences du courant domestique, qui ont elles aussi été lissées.



(Estimation spectrale)

Les signaux sont issus de la base de données mitdb. On dispose ici de deux capteurs, on a extrait environ 4,5 minutes d'enregistrement. Le périodogramme (et la version réduite) a été obtenu en par les instructions suivantes

```
freqs,Px = signal.periodogram(sigs,fs=fs,axis=0)
plt.semilogy(freqs,Px)
plt.xlabel('Frequency (Hz)')
plt.grid(which='both', axis='both')
plt.show()
```

```
lowfreqs_ind = np.where(freqs<10)
lowfreqs_ind = lowfreqs_ind[0]
lowfreqs = freqs[lowfreqs_ind]
plt.semilogy(lowfreqs,Px[lowfreqs_ind,:])
plt.grid(which='both', axis='both')
plt.xlabel('Frequency (Hz)')
plt.show()
```

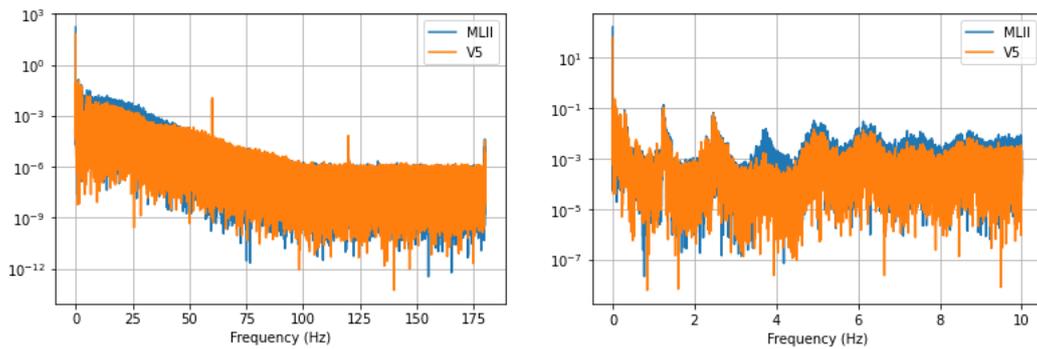


FIGURE 3.13 – Périodogramme d'un enregistrement ECG (4 minutes environ, 2 capteurs, échantillonné à 360 Hz). Gauche : spectre complet. Droite : zoom sur la bande de fréquence 0-10 Hz

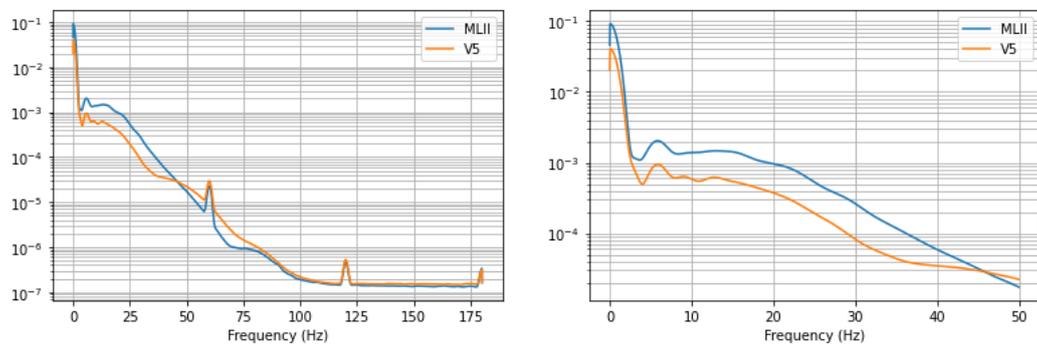


FIGURE 3.14 – Périodogramme de Welch d'un enregistrement ECG (4 minutes environ, 2 capteurs, échantillonné à 360 Hz). Gauche : spectre complet. Droite : zoom sur la bande de fréquence 0-50 Hz

Pour obtenir le périodogramme de Welch on a utilisé les instructions suivantes

```

freqs,Wx = signal.welch(sigs,fs=fs,nfft=L/10,axis=0)
plt.semilogy(freqs,Wx)
plt.grid(which='both', axis='both')
plt.xlabel('Frequency (Hz)')
plt.show()

lowfreqs_ind = np.where(freqs<50)
lowfreqs_ind = lowfreqs_ind[0]
lowfreqs = freqs[lowfreqs_ind]
plt.semilogy(lowfreqs,Wx[lowfreqs_ind])
plt.grid(which='both', axis='both')
plt.show()

```

3.2.2 Suppression de la ligne de base : filtrage passe-haut

Revenons tout d'abord sur le problème abordé dans l'exemple 3.2, c'est à dire la correction de ligne de base. Dans cet exemple, elle était abordée en deux étapes : tout d'abord un lissage fournissait une estimée de la ligne de base, qui était ensuite soustraite du signal. On va ici faire directement un filtrage passe-haut, en utilisant un filtre de Butterworth défini par

$$|\hat{h}(v)|^2 = W(w) = \frac{(1-w)^L}{(1-w)^L + k(1+w)^L}, \quad (3.29)$$

où $w = \cos(2\pi v/\eta)$, et où k contrôle la fréquence de coupure (comparer à (3.25) pour le filtre passe-bas correspondant).

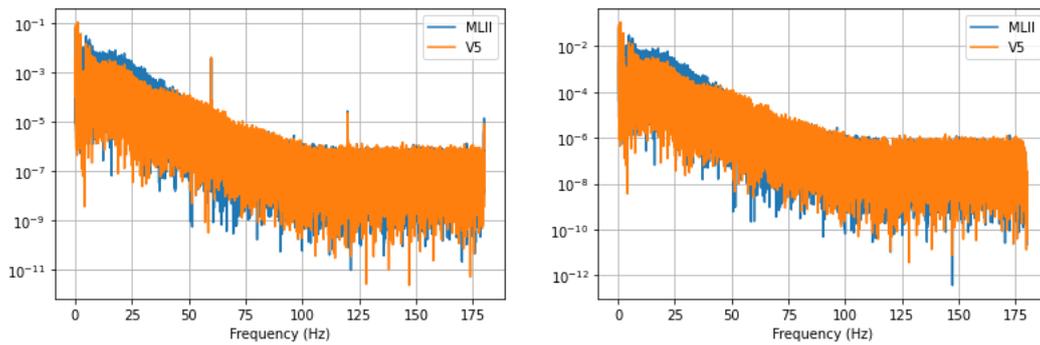


FIGURE 3.15 – Périodogramme d'un enregistrement ECG après filtrages. Gauche : filtrage passe-haut éliminant les fréquences inférieures à 0.3 Hz. Droite : Filtrage à encoche éliminant les fréquences multiples de 60 Hz (courant électrique domestique)

La Figure 3.15 (gauche) représente le périodogramme du signal filtré passe haut ; on peut y voir que les très basses fréquences ont été éliminées.



(Filtrage passe-haut)

Ces résultats ont été obtenus via les instructions suivantes. On a choisi un filtre d'ordre 10 et une fréquence de coupure de 0.3 Hz. Compte tenu de l'ordre du filtre, il a été généré sous format *sos* (*second order section*, qui est une forme factorisée) qui assure davantage de stabilité.

```
f_stop = .3
filt_ord = 10
sos_hp = signal.butter(filt_ord, f_stop, 'hp', fs=fs, output='sos')
sigs_filtered1 = signal.sosfilt(sos_hp, sub_sigs,axis=0)
```

3.2.3 Suppression de lignes spectrales 60Hz : filtre à encoche

Un filtre à encoche est un filtre dont l'objectif est de supprimer ou atténuer une bande de fréquence étroite, généralement une ligne spectrale centrée sur une fréquence donnée, et de largeur donnée (telle que celles qui apparaissent sur les ECG vers 60Hz, 120 Hz et dans une moindre mesure 180 Hz). Ils sont principalement déterminés par les zéros de la fonction de transfert, qui se trouvent sur le cercle complexe de rayon 1 et correspondent donc à des zéros de $\hat{h}(v)$.

Un exemple classique de filtre à encoche est donné par la fonction de transfert

$$H(z) = \frac{1 + az^{-1} + z^{-2}}{1 + (1 - \varepsilon)az^{-1} + (1 - \varepsilon)^2z^{-2}}, \quad \text{avec } |a| < 2. \quad (3.30)$$

Le nombre a caractérise la fréquence à filtrer, via la relation

$$v_0 = \eta \arccos\left(-\frac{a}{2}\right)$$

et ε caractérise la largeur de bande. Un exemple est donné en Figure 3.2.3. On voit que les zéros sont bien localisés sur le cercle unité, alors que les pôles sont à l'intérieur. Plus les pôles sont proches du cercle unité plus la bande de fréquence supprimée autour de la fréquence de coupure est étroite.

On représente en Figure 3.15 (graphe de droite) le périodogramme du signal ECG analysé dans cette section, après trois filtrages à encoche permettant de supprimer la ligne spectrale à 60 Hz ainsi que deux harmoniques. On voit que ces lignes ont effectivement disparu du spectre estimé.



(Filtrage à encoche)

Les fréquences à supprimer sont la fréquence fondamentale 60Hz, ainsi que les deux harmoniques à 120Hz et 180Hz.

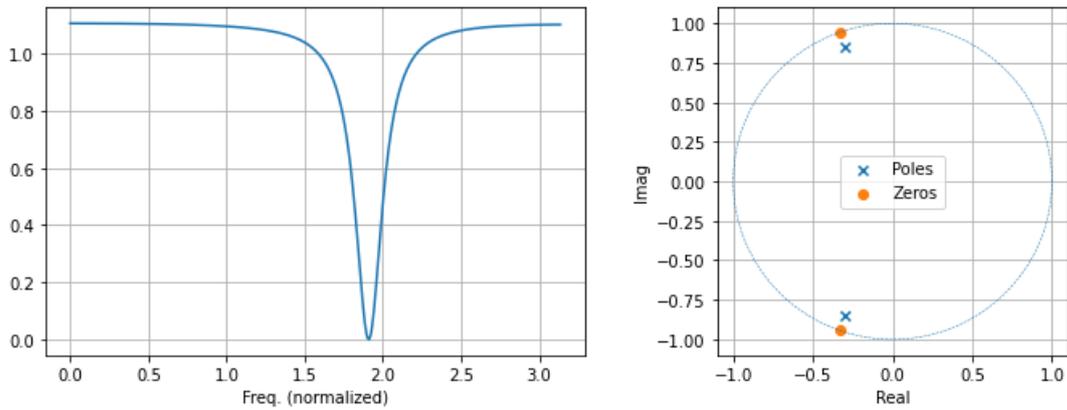


FIGURE 3.16 – Réponse fréquentielle (gauche) et zéros et pôles (droite) d'un filtre à encoche.

```

f_cut = 60
b,a = signal.iirnotch(f_cut,Q=f_cut,fs=fs)
sigs_filtered2 = signal.lfilter(b,a,sigs_filtered1,axis=0)
f_cut = 120
b,a = signal.iirnotch(f_cut,Q=f_cut,fs=fs)
sigs_filtered2 = signal.lfilter(b,a,sigs_filtered2,axis=0)
f_cut = 180
b,a = signal.iirnotch(f_cut,Q=f_cut,fs=fs)
sigs_filtered2 = signal.lfilter(b,a,sigs_filtered2,axis=0)

```

3.3 Temps-fréquence, ondelettes

Les signaux ECG sont intrinsèquement stationnaires (le rythme cardiaque est susceptible de varier au cours du temps, soit de façon lente et régulière lors de phases d'effort ou de retour au repos par exemple), soit de façon plus abrupte. Il est donc naturel d'utiliser des méthodes permettant de tracer de telles variations.

L'analyse temps-fréquence et ses variantes offrent un cadre naturel à ces approches.

3.3.1 Transformation de Fourier à court terme, spectrogramme

L'idée la plus simple est de *localiser* la transformée de Fourier. Une façon de faire cela est d'utiliser une fenêtre pour localiser le signal au voisinage d'un instant d'intérêt, avant de calculer la transformée de Fourier. En faisant glisser, ou coulisser la fenêtre le long de l'axe temporel, on obtient une analyse de Fourier en tout instant. Ce que l'on obtient dépend évidemment de l'outil d'analyse, ici la fenêtre.

Le cas des signaux de longueur infinie

On décrit tout d'abord la TFCT dans le cas de signaux de longueur infinie. Dans ce cas, le domaine temporel est \mathbb{Z} , et le domaine fréquentiel est l'intervalle $[-\eta/2, \eta/2]$.

Définition 3.4 — Transformation de Fourier à court terme (TFCT/STFT). Soit $g \in \ell^2(\mathbb{Z})$ une fenêtre (c'est à dire un signal régulier, localisé au voisinage de l'instant $t = 0$, tel que représenté dans la Figure 3.12). La transformée de Fourier à court terme de $x \in \ell^2(\mathbb{Z})$ est la fonction de deux variables $\mathcal{V}_g x$ donnée par

$$\mathcal{V}_g x[v, n] = \sum_{\ell=-\infty}^{\infty} x[\ell] \bar{g}[\ell - n] e^{-2i\pi v(\ell - n)/\eta}.$$

Le nombre (complexe) $\mathcal{V}_g x[v, n]$ décrit donc le contenu de x au voisinage de la fréquence v et de l'échantillon n . On représente généralement cette fonction de deux variables par une image, appelée **spectrogramme**, qui encode son module (ou le logarithme de celui-ci) par un code de couleurs ou un niveau de gris.

On va maintenant voir ce que signifie cette notion de voisinage, et l'importance de la fenêtre dans cette notion.

Remarque 3.14 — Atomes temps-fréquence. 1. Une autre interprétation importante fait intervenir ce que l'on appellera *atomes temps-fréquence* $g_{(v,n)}$, définis par

$$g_{(v,n)}[\ell] = g[\ell - n]e^{2i\pi v(\ell-n)/\eta}, \quad (3.31)$$

on a en effet

$$\mathcal{V}_g x[v, n] = \langle x, g_{(v,n)} \rangle, \quad (3.32)$$

cette interprétation sera importante pour la suite. Des exemples d'atomes STFT sont tracés dans la figure 3.17.

2. On a déjà vu que la transformation de Fourier discrète satisfait la formule de Parseval. On peut donc écrire (après calcul)

$$\mathcal{V}_g x[v, n] = \langle \hat{x}, \widehat{g_{(v,n)}} \rangle = \frac{1}{\eta} \int_{-\eta/2}^{\eta/2} \hat{x}(\xi) \bar{\hat{g}}(\xi - v) e^{2i\pi \xi n / \eta} d\xi \quad (3.33)$$

3. On utilise parfois la définition alternative basée sur des atomes définis par

$$g_{(v,n)}[\ell] = g[\ell - n]e^{2i\pi v(\ell-n)/\eta},$$

on écrit alors

$$\mathcal{V}_g x[v, n] = \langle x, g_{(v,n)} \rangle = \sum_{\ell=-\infty}^{\infty} x[\ell] \bar{g}[\ell - n] e^{-2i\pi v \ell / \eta} = \int_{-\eta/2}^{\eta/2} \hat{x}(\xi) \bar{\hat{g}}(\xi - v) e^{2i\pi(\xi - v)n/\eta} d\xi.$$

Ces deux formulations ne diffèrent que d'un facteur de phase. On se limitera à la première.

Remarque 3.15 — Interprétation en termes de bancs de filtres. on peut aussi, en introduisant les fenêtres modulées g_v définies par

$$g_v[\ell] = e^{2i\pi v \ell / \eta} g[\ell] \quad (3.34)$$

écrire

$$\mathcal{V}_g x[v, n] = (x * \tilde{g}_v)[n], \quad (3.35)$$

où \tilde{y} représente génériquement la suite retournée et complexe conjuguée de la suite y

$$\tilde{y}[\ell] = \bar{y}[-\ell].$$

Ainsi, $\mathcal{V}_g x$ peut être vu comme une famille (infinie continue) de signaux, obtenus par filtrage de x par des filtres de réponses impulsionnelles respectives \tilde{g}_v . Cette famille de filtres est appelé banc de filtres (continu).

La transformation de Fourier à court terme est une transformation linéaire. A l'instar de la transformation de Fourier, elle est inversible (sous des hypothèses naturelles), ce qu'exprime le résultat qui suit.

Théorème 3.16 — Inversion de la TFCT. Si g n'est pas uniformément nulle, la TFCT correspondante est inversible, et satisfait une formule de Parseval (condition de conservation de l'énergie). On a en effet, pour tout $x \in \ell^2(\mathbb{Z})$

$$x[\ell] = \frac{1}{\eta \|g\|^2} \sum_{n=-\infty}^{\infty} \int_{-\eta/2}^{\eta/2} \mathcal{V}_g x[v, n] g_{v,n}[\ell] dv,$$

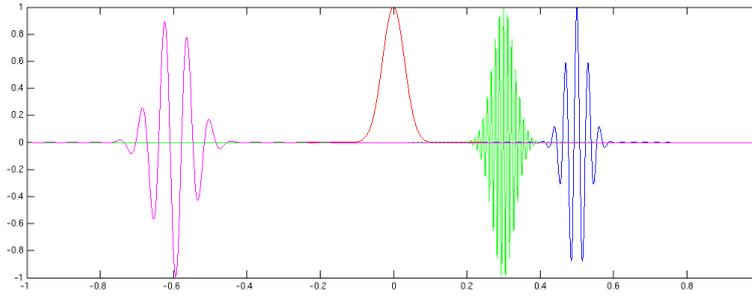


FIGURE 3.17 – Exemples d’atomes temps-fréquence (vert et bleu), copies translattées et modulées d’une fenêtre g (en rouge). L’atome temps-fréquence tracé en magenta correspond à une fenêtre différente (plus large)

et

$$\frac{1}{\eta \|g\|^2} \|\mathcal{V}_g x\|^2 := \sum_{n=-\infty}^{\infty} \int_{-\eta/2}^{\eta/2} |\mathcal{V}_g x[v, n]|^2 dv = \|x\|^2 .$$

Remarque 3.17 — Sous-échantillonnage. La variable de fréquence est dans le cas considéré une variable continue. Il s’avère qu’il est possible de discrétiser cette variable, en considérant M valeurs de la fréquence régulièrement espacées dans l’intervalle $[-\eta/2, \eta/2]$. On pose $b = \eta/M$, et on considère les fréquences de la forme $mb - \eta/2$, avec $m = 0, \dots, M-1$.

Dans le même ordre d’esprit, on peut également sous-échantillonner la variable temporelle, en introduisant un pas $a \in \mathbb{N}^*$, et ne conserver des échantillons de la forme na , pour $n \in \mathbb{Z}$.

Ceci conduit à considérer des atomes temps-fréquence (aussi appelés atomes de Gabor) de la forme

$$g_{mn}[\ell] = g[\ell - na] e^{2i\pi mb\ell/\eta} .$$

La transformation correspondante, appelé transformation de Gabor associée à tout $x \in \ell^2(\mathbb{Z})$ la matrice semi-infinie $\mathcal{V}_g x$ définie par

$$\mathcal{V}_g x[m, n] = \sum_{\ell=-\infty}^{\infty} x[\ell] \bar{g}[\ell - na] e^{-2i\pi mb\ell/\eta} , \quad n \in \mathbb{Z}, m \in \{0, \dots, M-1\} .$$

Il est possible de démontrer que sous certaines hypothèses sur a, b et g cette transformation de Gabor est inversible. Nous n’allons pas entrer dans ces détails. En effet, cette transformée est encore de taille infinie et n’a donc pas de valeur pratique. On va maintenant se focaliser sur le cas de signaux de longueur finie.

Le cas des signaux de longueur finie

Plaçons nous maintenant dans l’espace \mathbb{C}^L . La transformation de Fourier à court terme dans \mathbb{C}^L peut être définie comme suit.

Définition 3.5 — TFCT en dimension finie. Étant donné un vecteur $g \in \mathbb{C}^L$, la transformation de Fourier à court terme correspondante est la transformation linéaire $\mathcal{V}_g : \mathbb{C}^L \rightarrow \mathbb{C}^{L \times L}$ définie par

$$\mathcal{V}_g x[m, n] = \langle x, g_{mn} \rangle = \sum_{\ell=0}^{L-1} x[\ell] \bar{g}[\ell - n] e^{-2i\pi m(\ell-n)/L} ,$$

où on a encore introduit les atomes temps-fréquence g_{mn} définis par

$$g_{mn}[\ell] = g[\ell - n] e^{2i\pi m(\ell-n)/L} , \quad \ell = 0 \dots L-1 .$$

Le vecteur g sur lequel est basée l'analyse est appelé fenêtre.

A l'instar de la TFCT sur $\ell^2(\mathbb{Z})$ vue dans la section précédente, cette version de dimension finie est inversible. Calculons par exemple

$$\begin{aligned} \sum_{m,n=0}^{L-1} \mathcal{V}_g x[m,n] g_{m,n}[\ell] &= \sum_{m,n=0}^{L-1} \left[\sum_{k=0}^{L-1} x[k] \bar{g}[k-n] e^{-2i\pi m(k-n)/L} \right] g[\ell-n] e^{2i\pi m(\ell-n)/L} \\ &= \sum_{k,n=0}^{L-1} \bar{g}[k-n] g[\ell-n] x[k] \sum_{m=0}^{L-1} e^{2i\pi m(\ell-k)/L} \\ &= L \sum_{n=0}^{L-1} |g[n-\ell]|^2 x[\ell] = L \|g\|^2 x[\ell] \end{aligned}$$

où on a utilisé l'expression de la somme d'une série géométrique finie. Ceci montre que x peut être reconstruit à partir de sa TFCT. Avec la même démonstration, on peut également démontrer une formule de Parseval, et plus généralement on a le résultat suivant :

Théorème 3.18 — Inversion de la TFCT en dimension finie. Soient $g, h \in \mathbb{C}^L$ tels que $\langle g, h \rangle \neq 0$. Alors on a la formule d'inversion de la TFCT : pour tout $x \in \mathbb{C}^L$,

$$x = \frac{1}{L \langle h, g \rangle} \sum_{m,n=0}^{L-1} \mathcal{V}_g x[m,n] h_{m,n},$$

et la formule de Parseval

$$\langle \mathcal{V}_g x, \mathcal{V}_h x \rangle = L \langle h, g \rangle \|x\|^2.$$

h est appelée fenêtre de synthèse, par opposition à g qui est la fenêtre d'analyse.

La première expression montre en particulier que l'on peut utiliser une fenêtre de synthèse h différente de la fenêtre d'analyse. En particulier, en prenant $h[\ell] = \delta_{\ell,0}$, et en supposant donc que $g[0] \neq 0$, on a la formule d'inversion simple

$$x[\ell] = \frac{1}{L \bar{g}[0]} \sum_{m=0}^{L-1} \mathcal{V}_g x[m, \ell]. \quad (3.36)$$

En pratique, lorsque la longueur des signaux d'intérêt devient grande, on ne peut plus conserver toutes les valeurs possibles des indices de temps et de fréquence, et on est amenés à sous-échantillonner la TFCT. Dans ce qui suit, on notera toujours L la longueur des signaux $x \in \mathbb{C}^L$, et on notera

- a le pas d'échantillonnage temporel ; a doit être un diviseur de L , on notera $L = Na$.
- b le pas d'échantillonnage fréquentiel, qui doit aussi être un diviseur de L , on notera $L = Mb$.

Avec ces notations, on introduit les atomes temps-fréquence, qu'on appellera **atomes de Gabor**, pour les différencier de la TFCT : partant d'une fenêtre $g \in \mathbb{C}^L$

$$g_{mn}[\ell] = e^{2i\pi mb(\ell-na)/L} g[\ell-na]. \quad (3.37)$$

Définition 3.6 — Transformation de Gabor. Étant donnée une fenêtre $g \in \mathbb{C}^L$, et a, b deux diviseurs de L , la transformation de Gabor correspondante est la transformation linéaire qui associe à tout $x \in \mathbb{C}^L$ la matrice $\mathcal{V}_g x \in \mathbb{C}^{M \times N}$ définie par

$$\mathcal{V}_g x[m,n] = \langle x, g_{mn} \rangle = \sum_{\ell=0}^{L-1} x[\ell] \bar{g}[\ell-na] e^{-2i\pi mb(\ell-na)/L}.$$

- Remarque 3.19**
1. La TFCT est donc un cas particulier de transformée de Gabor, pour laquelle on a pris $a = b = 1$.
 2. La transformation de Gabor hérite des propriétés d'invariance par translation de la TFCT. Si y est une copie translatée d'un multiple de a de x , c'est à dire $y[\ell] = x[\ell - \tau a]$, alors $\mathcal{V}_g y[m, n] = \mathcal{V}_g x[m, n - \tau]$. En particulier, on retrouve une forme de conditions aux bords périodiques, en temps comme en fréquence :

$$\mathcal{V}_g x[m, n + N] = \mathcal{V}_g x[m, n], \quad \mathcal{V}_g x[m + M, n] = \mathcal{V}_g x[m, n].$$



(Transformation de Gabor Discrète)

Le package `lftfatpy` implémente des outils classiques d'analyse temps-fréquence. La fonction `dgtrealm` calcule la transformée de Gabor discrète d'un signal, en précisant la fenêtre g , ainsi que les paramètres a et M . La fenêtre peut quant à elle être générée par différentes fonctions, par exemple `pgauss` qui génère une fenêtre gaussienne, ou `gabwin` qui est plus générique.

L'exemple qui suit génère une sinusoïde de fréquence égale à 440 Hz, et calcule sa transformée de Gabor discrète. Le résultat est montré en Figure 3.18 (gauche), où est représenté le spectrogramme correspondant, c'est à dire le module carré, codé sur une échelle de couleurs. On voit que le spectrogramme prend ses plus grandes valeurs au voisinage d'une droite horizontale, qui correspond en fait à la fréquence de la sinusoïde.

```

Fs = 8000
T = 1
L = T*Fs
tt = np.linspace(0,T,L)
f0 = 440
y = np.cos(2*np.pi*f0*tt)

g = lftfatpy.pgauss(L,tfr=1)[0]
a = 8; M = np.int(L/a)
N = 2000; b = np.int(L/N)

Vgy = lftfatpy.dgtreal(y,g,a,M)
Vgy = Vgy[0]
spy = np.abs(Vgy)
plt.imshow(spy,aspect='auto',origin='lower',cmap='hot',extent=[0,T,0,Fs/2000])
plt.xlabel('Time (sec)')
plt.ylabel('Freq. (kHz)')
plt.colorbar()
plt.show()

```

Le graphe de droite dans la Figure 3.18 représente le spectrogramme d'un *chirp* quadratique (c'est à dire une sinusoïde de fréquence variable, ici dépendant du temps de façon quadratique). On voit que le spectrogramme prend ses plus grandes valeurs au voisinage d'une courbe, qui représente en fait la fréquence instantanée.

```

f_start = 440
f_end = f_start*4
z = signal.chirp(tt, f_start, T, f_end,method='quadratic')
Vgz = lftfatpy.dgtreal(z,g,a,M)
Vgz = Vgz[0]
spz = np.abs(Vgz)
plt.imshow(spz,aspect='auto',origin='lower',
           cmap='hot',extent=[0,T,0,Fs/2000])
plt.xlabel('Time (sec)')
plt.ylabel('Freq. (kHz)')
plt.colorbar()
plt.show()

```

Remarque 3.20 — Le rôle de la fenêtre. — On voit bien que pour donner à l'analyse une précision temporelle suffisante, on aimerait que la fenêtre g soit suffisamment courte. Par convention, et

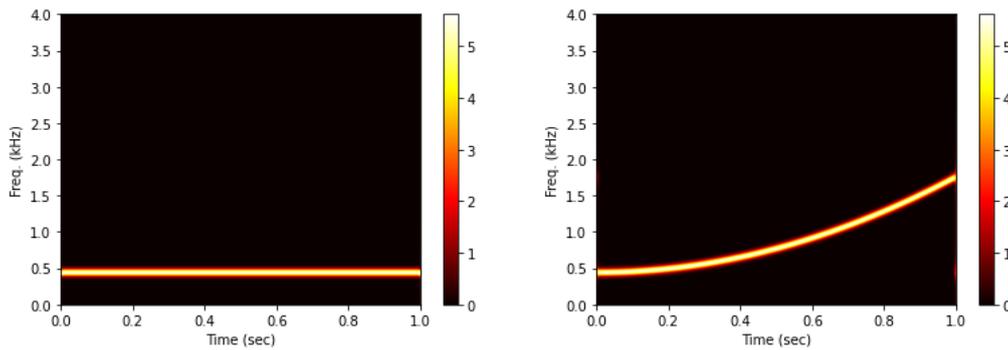


FIGURE 3.18 – Transformée de Gabor discrète d'une sinusoïde et d'un *chirp* de fréquence instantanée quadratique

sans perte de généralité on suppose g centrée sur $\ell = 0$. Par contre, l'équation (3.33) montre aussi que pour que $\mathcal{V}_g x[m, n]$ décrive le contenu fréquentiel de x au voisinage de la fréquence m , on a aussi besoin que \hat{g} soit centrée sur la fréquence nulle (ceci est fait sans perte de généralité aussi), mais aussi que \hat{g} soit bien localisée autour de $k = 0$ (c'est à dire décroissant assez vite lorsque l'on s'éloigne de l'origine des fréquences). Or ces deux conditions de localisation (en temps et en fréquence) sont relativement incompatibles. Le principe d'incertitude stipule qu'il est impossible de trouver une fonction ou une suite qui soit optimalement localisée à la fois en temps et en fréquence. On doit donc rechercher des compromis, qui conduisent à utiliser une fenêtre gaussienne, ou l'une des fenêtres classiques dont les graphes se trouvent en Figure 3.12.

- Même avec une fenêtre fixée, la taille de la fenêtre joue un rôle important. Plus la fenêtre est longue, plus elle est imprécise dans le domaine temporel et précise dans le domaine fréquentiel, et vice-versa.

Un exemple se trouve dans la Figure 3.19, sur lequel on a représenté la DGT de la somme d'une sinusoïde et d'un "pic" avec deux tailles de fenêtre. Suivant la taille de fenêtre, le pic est bien résolu au contraire de la sinusoïde (fenêtre courte) et inversement (fenêtre longue).

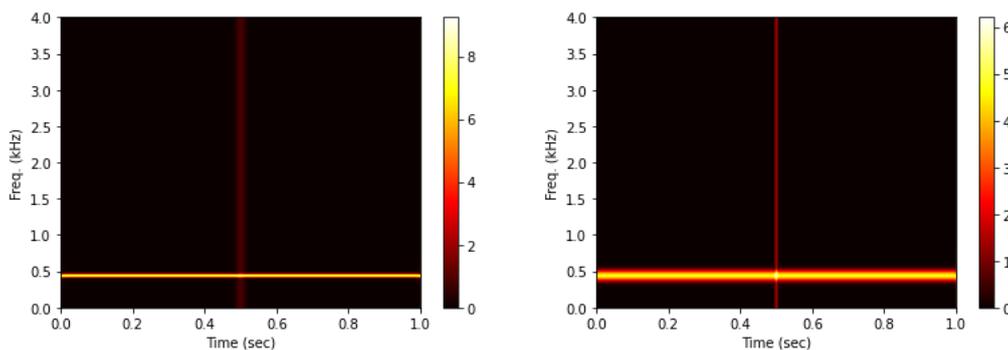


FIGURE 3.19 – Transformée de Gabor discrète de la somme d'une sinusoïde et d'un 'pic', avec deux tailles de fenêtre différentes : fenêtre courte (gauche) et longue (droite).

Compte tenu de la nature presque périodique des signaux ECG, on pourrait espérer que des représentations temps-fréquence puissent permettre une bonne analyse de la périodicité et de là permettre une estimation fiable du rythme cardiaque. La situation est plus complexe, car un ECG est loin d'être une simple sinusoïde. Une représentation temps-fréquence d'ECG est bien plus complexe à interpréter. Un exemple se trouve en Figure 3.20. On voit bien apparaître une ligne assez régulière vers les basses fréquences, ainsi que plusieurs harmoniques, mais la structure de l'ensemble est très complexe.

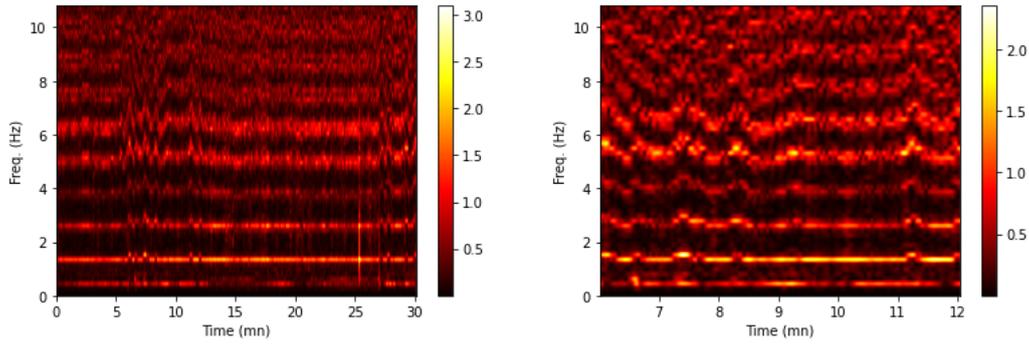


FIGURE 3.20 – Transformée de Gabor discrète (module) d’un signal ECG de la base mitdb (sujet 100). Gauche, 30mn de signal, droite : zoom sur un extrait de 6mn.

Inversion de la transformation de Gabor finie

La représentation de Gabor finie (sous-échantillonnée) est plus économique que la STFT, mais il y a un prix à payer. L’inversion de la transformée reste possible sous certaines conditions, mais elle n’est généralement pas aussi simple que dans les théorèmes 3.16 et 3.18. L’objet central pour étudier cette inversion est l’**opérateur de trame** \mathcal{R} , défini de la façon suivante : étant donnée une fenêtre (d’analyse) $g \in \mathbb{C}^L$, l’action de cet opérateur sur $x \in \mathbb{C}^L$ est

$$\mathcal{R}x = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} \langle x, g_{mn} \rangle g_{mn} . \quad (3.38)$$

On peut voir tout de suite que lorsque cet opérateur est inversible, et si son inverse peut être calculé, alors on a directement une formule d’inversion de la forme

$$x = \mathcal{R}^{-1} \left(\sum_{m,n} \langle x, g_{mn} \rangle g_{mn} \right) = \sum_{m,n} \langle x, g_{mn} \rangle \mathcal{R}^{-1} (g_{mn}) .$$

Il est par ailleurs possible de démontrer que les vecteurs $\mathcal{R}^{-1} (g_{mn})$ sont encore des atomes de Gabor

$$\mathcal{R}^{-1} (g_{mn}) = (\mathcal{R}^{-1} g)_{mn} ,$$

de sorte qu’en posant

$$\gamma = \mathcal{R}^{-1} g \quad (3.39)$$

on obtient la formule d’inversion de la transformée de Gabor

$$x = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} \langle x, g_{mn} \rangle \gamma_{mn} . \quad (3.40)$$

Ainsi, on obtient une formule d’inversion similaire à celle obtenue dans le théorème 3.18, à ceci près que la fenêtre de synthèse γ est cette fois fixée. γ est appelée **fenêtre duale canonique**. L’évaluation numérique de cette fenêtre duale est un élément important de l’analyse de Gabor.

Remarque 3.21 Il est important de signaler que la fenêtre duale, lorsqu’elle existe, dépend non seulement de la fenêtre d’analyse, mais aussi des paramètres d’échantillonnage a et b (ou M). On donne en figure 3.21 quelques exemples de fenêtres duales γ , calculées à partir de la même fenêtre g pour différentes valeurs des paramètres. Dans les figures du haut, on a pris $L = 128$, $a = 2$ et $M = 32$, on voit que la fenêtre bduale est très proche (à un facteur multiplicatif près) de la fenêtre g . Pour $a = 8$ et $M = 16$ (milieu), la fenêtre de synthèse se différencie davantage de la fenêtre d’analyse, alors que pour $a = 16$ et $M = 16$, l’opérateur de trame n’est plus inversible, et la fenêtre duale n’a plus de sens (voir les valeurs des pics).

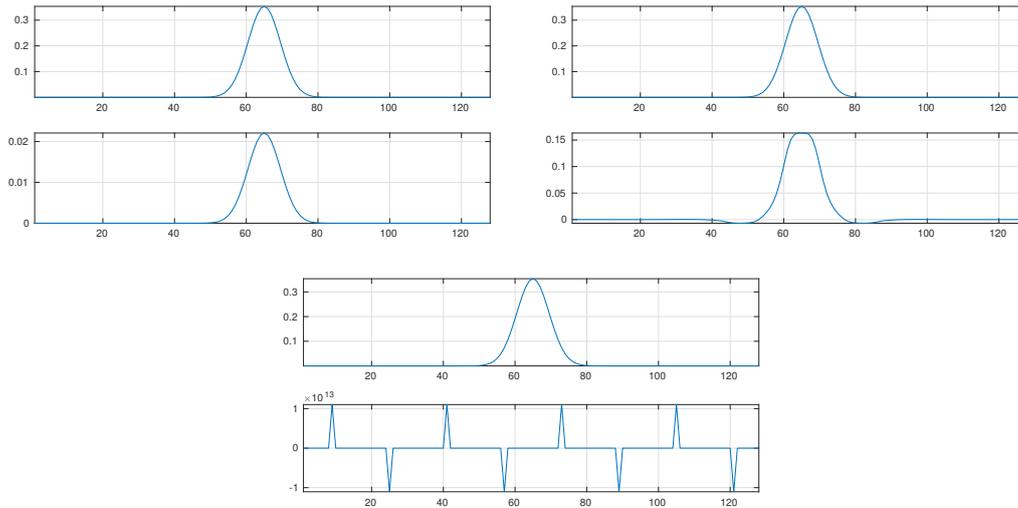


FIGURE 3.21 – Fenêtres d’analyse (haut) et duale canonique (bas), pour la même fenêtre d’analyse et différentes valeurs des paramètres a et b . Dans le dernier cas, l’opérateur de trame \mathcal{R} n’est pas inversible, la fenêtre duale canonique est singulière.

3.3.2 Ondelettes et variantes

Comme on l’a vu plus haut (voir notamment l’équation (3.35) de la remarque 3.14, et l’exercice ??), la transformation de Gabor peut s’assimiler à un banc de filtres (complété par un sous-échantillonnage), avec un filtre de réponse fréquentielle (fonction de transfert) égal à la transformée de Fourier $\hat{g}(v - mb)$ de la fenêtre g , translatée. Ainsi les réponses fréquentielles des filtres correspondants sont identiques, à une translation près, et la largeur de bande est donc constante.

De façon similaire, on peut aussi construire des transformations temps-fréquence sous forme de bancs de filtres, dont les gabarits peuvent être différents. On aborde brièvement deux exemples dans ce qui suit.

Transformation en ondelettes, transformée à Q constant

Dans le cas de la transformation en ondelettes, la largeur de bande des filtres est proportionnelle à leur fréquence centrale. Dans le cas de signaux de longueur infinie, on introduit donc un banc de filtres $\{\psi_j, j = 1, \dots, \infty\}$, où s est un paramètre que l’on appelle généralement paramètre d’échelle. La transformation correspondante est notée \mathcal{W}_ψ

$$\mathcal{W}_\psi x[j, n] = (x * \widetilde{\psi}_j)[na_j] = \sum_{\ell} x[\ell] \widetilde{\psi}_j[\ell - na_j], \quad (3.41)$$

où a_j est un pas d’échantillonnage temporel, qui peut dépendre ou pas du paramètre j , et où comme précédemment $\widetilde{\psi}_j[\ell] = \overline{\psi_j[-\ell]}$.

Dans la version la plus simple, $j = 1$ correspond à une échelle égale à 2 échantillons, $j = 2$ à une échelle de 4 échantillons, et plus généralement j est associé à une échelle de 2^j échantillons. Les réponses fréquentielles des filtres sont des copies dilatées les unes des autres, c’est à dire que l’on a essentiellement

$$\widehat{\psi}_j(v) = 2^{j/2} \widehat{\psi}(2^j v [\text{mod } 1]) \quad (3.42)$$

(le $[\text{mod } 1]$, lire "modulo 1", est là pour tenir compte de la périodicité du domaine de Fourier dans ce cas, voir plus loin pour davantage de précisions). Le filtre ψ est appelé "ondelette".

Si on oublie ce terme $[\text{mod } 1]$ pour un instant, et si on suppose que $|\widehat{\psi}|$ est maximum pour une fréquence donnée, disons $v_0 = \pm 1/2$, alors $|\widehat{\psi}_s|$ est quant à lui maximum pour une fréquence v_j telle que $2^j v_j = v_0$, soit $v_j = 2^{-j} v_0$. Comme par ailleurs on voit que $\widehat{\psi}_j$ est une copie de $\widehat{\psi}$ contractée d’un facteur 2^j , sa largeur

σ_j (quelle que soit la façon dont on la mesure) est multiple de la largeur de bande σ de ψ : $\sigma_j = 2^{-j}\sigma$, et est donc proportionnelle à la fréquence centrale du filtre. On a donc pour tout j la propriété de largeur de bande relative constante

$$Q^{-1} = \frac{\sigma_j}{v_j} = \frac{\sigma_0}{v_0}.$$

On dit que les ondelettes effectuent une analyse à Q constant.

L'échantillonnage en échelles (qui est aussi un échantillonnage fréquentiel, la fréquence étant définie comme l'inverse de l'échelle, à une constante multiplicative près) est généralement un échantillonnage suivant les puissances de 2, comme vu plus haut. Deux cas de figure sont couramment utilisés concernant l'échantillonnage temporel

- Échantillonnage régulier $a_j = a_0$ (par exemple $a_j = 1$ pour tout j). Ceci correspond à la transformation en ondelettes stationnaire, qui possède une intéressante propriété d'invariance par translation.
- Échantillonnage dépendant de l'échelle j : ceci correspond aux transformées en ondelettes décimées.

On n'entrera pas ici dans les détails de la théorie. Il est possible de ; montrer que dans ces deux cas de figure, si les paramètres (en l'occurrence l'ondelette ψ et les pas d'échantillonnage) sont bien choisis, la transformation

$$x \longmapsto \mathcal{W}_\psi x$$

est inversible, dans un cadre similaire au cadre rencontré dans l'analyse de Gabor. Les coefficients d'ondelettes $\mathcal{W}_\psi x[j, n]$ d'un signal x en donnent donc une représentation alternative, à partir de laquelle un certain nombre d'opérations peuvent être effectuées.

Il est à noter que la transformation en ondelettes constitue un cas particulier de transformées dites à Q constant. Toutes ne sont pas générées en utilisant des dilatations comme les ondelettes, mais conservent cette propriété de bancs de filtres dont la largeur de bande est proportionnelle à la fréquence centrale. Des transformées à Q constant peuvent aussi être construites par transformation de Gabor non-stationnaire.

Transformation de Gabor non-stationnaire

Les transformations de Gabor non-stationnaires sont également basées sur des bancs de filtres satisfaisant certaines propriétés génériques particulières. On sait en particulier construire maintenant des bancs de filtres dont les réponses fréquentielles reproduisent des réponses fréquentielles de filtres développés pour modéliser la perception (on parle d'échelle MEL, ERB ou encore BARK suivant le contexte). La théorie décrite dans le cadre de l'analyse de Gabor peut s'appliquer dans ce cadre, et on montre que ces transformations sont inversibles si les paramètres sont bien choisis.

4. Détection du complexe QRS

Une tâche essentielle dans l'analyse des signaux ECG est la détection des complexes QRS, et en particulier la détection des ondes R, qui sont généralement les pics les plus significatifs sur les capteurs classiquement utilisés. Une fois les ondes R détectées, on peut s'intéresser aux différences entre deux R consécutifs (l'intervalle RR), qui dépendent du temps. On note ce signal $RR[t]$, et de là il est usuel de calculer le rythme cardiaque, défini par

$$\rho[t] = \frac{60}{RR[t]}, \quad (4.1)$$

qui s'exprime en *battements par minute* (d'où le facteur 60 au numérateur). D'autres quantités d'intérêt (intervalles QR,...) peuvent également être calculées à partir de là, ainsi que diverses annotations.

4.1 Généralités

La première étape est donc une étape de détection, qui peut s'apparenter à une détection de pics, les pics étant en général les ondes R. Ce problème de détection de pics est un problème souvent complexe, on liste ci-dessous les trois grandes approches que l'on peut trouver dans la littérature.

1. Les approches les plus courantes reposent sur l'évaluation empirique de dérivées, couplées à des critères concernant la valeur du signal. Un pic est ici défini comme un emplacement où le signal est supérieur à un certain seuil absolu donné, tandis que sa dérivée (ou une version lissée de celle-ci) est égale à (ou proche de) zéro. En option, d'autres critères peuvent être inclus, par exemple en imposant que deux pics ne soient pas trop proches l'un de l'autre, ou qu'un pic doit être entouré de valeurs de signal bien en dessous de celui-ci.
2. Une autre approche à la détection de pics se base sur une modélisation explicite. En termes simples, on se base sur un signal de fond supposé obéir à un modèle particulier, par exemple un modèle auto-régressif, (AR), ou auto-régressif à moyenne mobile (ARMA), donc les paramètres sont ajustés à partir des données. Les pics sont alors détectés à partir de la qualité de l'ajustement des données : ce sont les points aberrants, c'est à dire les points qui ne sont pas bien expliqués par le modèle. Un critère quantitatif peut être donné par une interprétation probabiliste.
3. Dans un autre ordre d'idées, il est possible d'exploiter plusieurs représentations à différentes résolutions pour détecter les pics. Dans ces approches, les signaux passent d'abord par une transformation en ondelettes, choisie de manière à rehausser les pics et à lisser le signal de fond. Ensuite, les pics aux différentes échelles sont appariées et leurs amplitudes sont additionnées de manière à donner le critère final sur lequel est réalisée la détection.

Ces approches sont généralement complétées par divers pré-traitements et post-traitements.

Dans un contexte de signaux ECG, les approches que nous allons voir plus loin partagent un certain nombre d'étapes communes :

Identification de pics

Cette étape dépend de façon cruciale de la définition que l'on donne à un pic.

- De façon heuristique, un pic correspond à un point où le signal prend une valeur supérieure aux valeurs voisines. Il s'agit donc d'un maximum local. ceci dit, il faut bien définir la notion de voisinage, car elle va fortement conditionner la détection. Un voisinage trop petit va conduire à un trop grand nombre de détections, un voisinage trop grand va conduire à des détections manquées.
- Une autre question importante est celle de la significativité d'un pic détecté : il importe de s'assurer que la valeur du signal au pic est significativement plus grande que les valeurs en dehors du voisinage. Il va donc falloir éliminer les pics détectés qui sont jugés non significatifs, c'est à dire inférieurs à un certain seuil. La question est alors : comment déterminer le seuil ? d'autant plus que celui-ci peut (et va) varier au cours du temps.

Pré-traitements :

- Pour éliminer des composantes parasites (hautes et basses fréquences)

Réhaussement du contraste :

- Filtrage non-linéaire permettant de mieux faire "émerger" les pics

Mesure locale de ligne de base

- Permet de définir ce qu'on entend par "émerger"

La plupart des algorithmes de référence pour la mesure de rythme cardiaque (et la détection d'arythmie) reposent sur une succession de plusieurs étapes. On va décrire dans ce qui suit deux algorithmes différents, qui sont relativement simples (suffisamment pour pouvoir fonctionner en temps réel, sur des dispositifs portables par exemple).

4.2 L'algorithme de Pan & Tompkins

L'algorithme de Pan et Tompkins [9] est l'un des plus anciens algorithmes de référence pour la détection du complexe QRS, l'estimation de rythme cardiaque et plus généralement l'annotation. Il peut paraître à la fois simpliste (du point de vue du traitement du signal) et très heuristique, mais il faut réaliser qu'il a été proposé à une époque où le calcul numérique était beaucoup plus compliqué qu'à l'heure actuelle. Il repose sur un certain nombre d'étapes, décrites plus en détail dans ce qui suit.

4.2.1 Détection de plages d'intérêt

La première phase est une phase de pré-détection de la localisation des complexes QRS. Il s'agit essentiellement d'une succession d'opérations visant à détecter une sorte d'enveloppe. Plus précisément, les étapes sont les suivantes.

1. **Débruitage** (filtrage passe-bande) : dans l'article original, cette étape est implémentée comme la succession d'un filtrage passe-bas et d'un filtrage passe-haut. Les auteurs spécifient les filtres dans le domaine de la transformée en z , pour une fréquence d'échantillonnage égale à 200 Hz. Dans les illustrations données ici, la fréquence d'échantillonnage vaut 360 Hz, et on a choisi un filtre de Butterworth *forward-backward* (c'est à dire appliqué dans un sens puis en temps inversé, pour éviter les problèmes de délai), d'ordre faible. La bande passante choisie est généralement prise aux environs de 5-15Hz (les avis diffèrent selon les articles). Du coup, les très basses fréquences sont éliminées, ainsi que la ligne spectrale de l'électricité domestique, et ces pré-traitements ne sont plus nécessaires.
2. **Dérivation** : l'objectif des auteurs est d'être sensible aux fortes pentes du signal (en particulier le démarrage de l'onde R, très abrupt) plutôt qu'à son amplitude. Ceci justifie le calcul de la dérivée du signal. Pan & Tompkins proposent un filtre d'ordre faible pour effectuer cette opération, on dispose maintenant de filtres de gradient suffisamment stables.

3. **Filtrage quadratureur** : un nom compliqué pour dire qu'on calcule le carré du signal (échantillon par échantillon). Il s'agit d'un filtre non linéaire, qui produit un signal à valeurs positives, et augmente également la différence entre l'amplitude de l'onde R et celle de l'onde T, ce qui est intéressant lorsque cette différence est positive (et pénalisant dans le cas contraire).
4. **Moyenne mobile** : la sortie du filtre quadratureur est un signal positif qui présente de fortes oscillations au voisinage de l'onde R. Ces oscillations sont lissées par un filtre à réponse impulsionnelle finie, pour lequel Pan & Tompkins préconisent un filtre de moyenne mobile, qui présente l'intérêt de ne conserver que des valeurs positives (sa mauvaise sélectivité fréquentielle n'est pas pénalisante ici). Ce filtre fournit un signal à valeurs positives, qui met en évidence les complexes QRS.
5. **Détection** : c'est l'étape suivante... qui sera décrite plus loin.

On présente en Figure 4.1 les différentes étapes décrites plus haut, sur un extrait de 14 secondes environ d'un signal de la base mitdb , échantillonné à 360 Hz. Le filtrage passe-bande est effectué avec une bande passante 10-20 Hz, et la moyenne mobile est calculée sur une fenêtre de 150 millisecondes environ. On voit sur le dernier tracé que les localisations des complexes QRS apparaissent clairement. Ceci étant cette figure est un peu trompeuse, car la situation n'est pas toujours aussi simple, en particulier dans les cas d'intérêt où des arythmies sont présentes dans les signaux.

4.2.2 Détection des ondes R

Cette détection exploite la sortie des opérations précédentes, et se base sur l'amplitude du signal résultant, mais aussi sur les intervalles entre pics. La décision est prise à partir de seuils, qui nécessitent de connaître un *niveau de signal* et un *niveau de bruit*. La détermination de ces deux niveaux est décrite un peu plus loin.

1. **Détection des pics** : un pic correspond à un instant où le signal prend une valeur supérieure à celles des instants voisins. Ceci étant, il faut définir correctement la notion de voisin. Si on permet des voisins très proches, le bruit peut générer un très grand nombre de voisins, à partir desquels il sera difficile de prendre une décision. Inversement, si un intervalle minimal entre deux voisins potentiels est choisi trop grand, on risque de manquer des pics. Le choix est de prendre un intervalle minimal assez réduit, puis d'utiliser les amplitudes pour affiner la sélection.

Ce que l'on sait, c'est que physiologiquement, deux ondes R ne peuvent être distantes de moins de 200 millisecondes (ce qui correspondrait à un rythme cardiaque de 300 battements par minute !)

2. **Seuillage** : pour éviter des maxima locaux parasites, seuls des pics de valeur supérieure à un seuil sont conservés. La détermination du seuil est un point crucial. Les signaux étant non stationnaires, il doit évoluer au cours du temps, en fonction de mesures de
 - niveau local de bruit \mathcal{B}
 - niveau local du signal \mathcal{S}

La valeur du seuil préconisée dans l'article de Pan & Tompkins est

$$\tau = \mathcal{B} + \mu(\mathcal{S} - \mathcal{B}), \quad (4.2)$$

où μ est un paramètre, pris à $\mu = 1/4$.

On représente en Figure 4.2 des pics détectés, pour deux valeurs différentes de la distance minimale entre pics : 500 millisecondes (gauche) et 160 millisecondes (droite). Alors que le résultat de gauche paraît plus satisfaisant, il faut réaliser que ce choix interdit de détecter plus de deux pics par seconde, et sera en défaut pour des rythmes cardiaques supérieurs à 120 battements par seconde, ce qui n'est pas envisageable. Il faut donc décroître significativement la distance minimale entre pics, et utiliser leur amplitude pour éliminer les pics non pertinents.

La détermination de ces niveaux est effectuée de la façon suivante

1. **Initialisation** : les niveaux \mathcal{S} et \mathcal{B} sont initialisés sur une plage de 2 secondes en début de signal
2. **Mise à jour** : Si un nouveau pic d'amplitude A est détecté sur le signal moyenné (issu de la moyenne mobile), les niveaux de signal ou de bruit sont mis à jour, suivant la décision (pic signal ou pic bruit) via

$$\mathcal{S} = \varepsilon A + (1 - \varepsilon)\mathcal{S}, \quad \text{ou} \quad \mathcal{B} = \varepsilon A + (1 - \varepsilon)\mathcal{B}, \quad (4.3)$$

Par exemple, $\varepsilon = 1/8$.

3. **Test sur le signal filtré** : chaque pic détecté est testé aussi sur le signal filtré (avec des seuils adaptés).

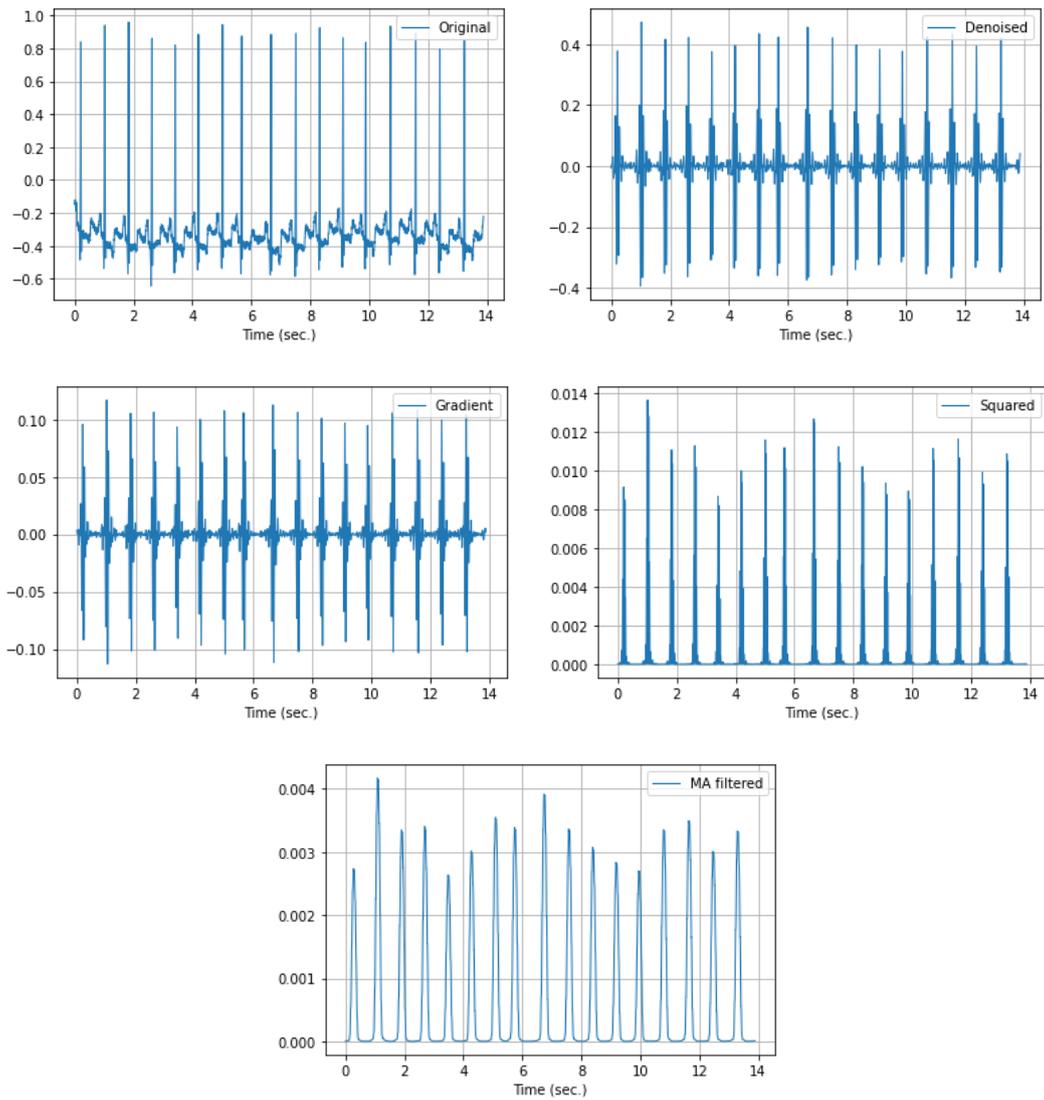


FIGURE 4.1 – Mise en évidence du complexe QRS. Haut : signal original (environ 14 secondes du sujet 100 de la base mitdb, capteur MLII) et signal filtré passe-bande. Milieu : signal dérivé (gauche), puis élevé au carré (droite). Bas : Signal filtré par moyenne mobile (moyenne sur 150 millisecondes)

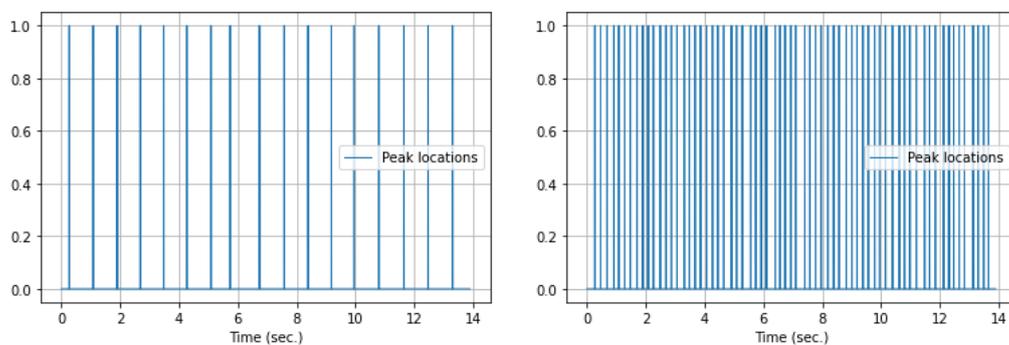


FIGURE 4.2 – Détection de pics à partir des enveloppes des complexes QRS, pour deux valeurs de la distance minimale : 500 millisecondes (gauche) et 160 millisecondes (droite).

4.3 L'algorithme d'Elgendi

L'algorithme de Pan & Tompkins est intéressant par sa simplicité et sa facilité d'implémentation, mais repose sur des choix de paramètres qui ne sont pas toujours faciles. Il était aussi particulièrement prisé à une époque où le calcul scientifique n'était pas aussi efficace que de nos jours.

L'état de l'art actuel atteint des performances sans commune mesure en termes de robustesse et de précision. Les méthodes correspondantes sont généralement basées sur des approches d'apprentissage automatique qui nécessitent des ressources de calcul très importantes.

Pendant, des méthodes simples et rapides peuvent également atteindre des taux de détection élevés. De telles approches sont particulièrement nécessaires dans des cas d'appareils ECG légers à batterie, et/ou pour l'analyse rapide d'enregistrements de très longue durée.

L'algorithme d'Elgendi [3], proposé en 2013, est une méthode de détection QRS essentiellement basée sur deux filtrages de moyenne mobile, couplés à une base de données pour optimiser les paramètres (deux paramètres seulement sont nécessaires). Là encore, il s'agit de prendre une décision à partir d'un signal non-stationnaire, à partir de seuillages où les seuils varient au cours du temps.

4.3.1 Filtrages

La première étape consiste à délimiter des segments d'intérêt dans le signal, à l'intérieur desquels est attendu le complexe QRS. Cette étape se base sur un filtrage passe-bande et un filtrage quadratureur, comme précédemment, suivi de deux lissages à moyenne mobile à deux échelles différentes.

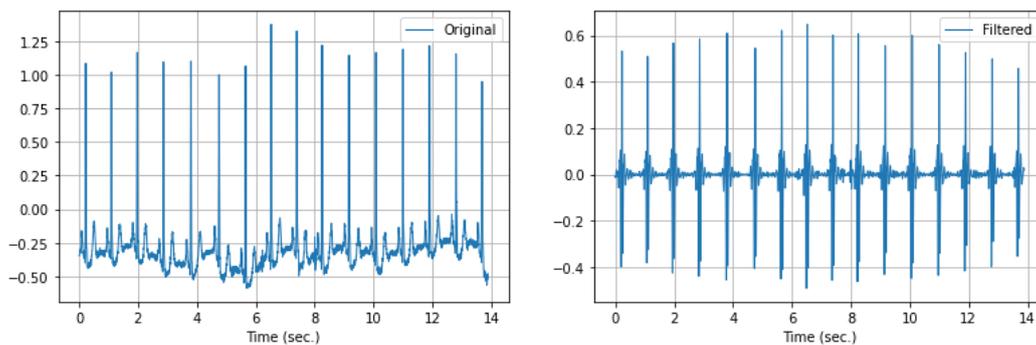


FIGURE 4.3 – Etape de débruitage sur le signal 101 de la base mitdb. Signal original (gauche) et signal filtré passe-bande (droite).

1. **Débruitage** : filtrage passe-bande, par filtre de Butterworth *forward-backward* d'ordre faible, pour une bande de fréquence $[f_1, f_2]$ donnée.
2. **Filtrage quadratureur** : la sortie du filtre de Butterworth est élevée au carré. On note x_2 le signal résultant.
3. **Filtrage moyenne mobile (QRS)** : la sortie du filtre quadratureur est filtrée par lissage par moyenne mobile, de longueur w_1 échantillons. w_1 représente la longueur approximative (en nombre d'échantillons) du complexe QRS. On note x_{QRS} le signal filtré résultant.
4. **Filtrage moyenne mobile (B)** : la sortie du filtre quadratureur est filtrée par un autre lissage par moyenne mobile, de longueur w_2 échantillons. w_2 représente la longueur approximative (en nombre d'échantillons) d'un battement cardiaque. On note x_B le signal résultant.

Il faut donc noter qu'à ce point, la méthode repose sur 5 paramètres : les fréquences f_1, f_2 , les longueurs w_1, w_2 des filtres de moyenne mobile, et le paramètre de seuillage β .

4.3.2 Seuillages et détection

Les filtrages ci-dessus sont maintenant utilisés pour détecter des segments d'intérêt.

1. **Segments d'intérêt** : les segments d'intérêt sont déterminés par un seuillage du signal x_{QRS} utilisant un seuil calculé à partir du signal x_B , donné par

$$\tau_1[n] = x_B[n] + \beta \bar{x}_2, \quad (4.4)$$

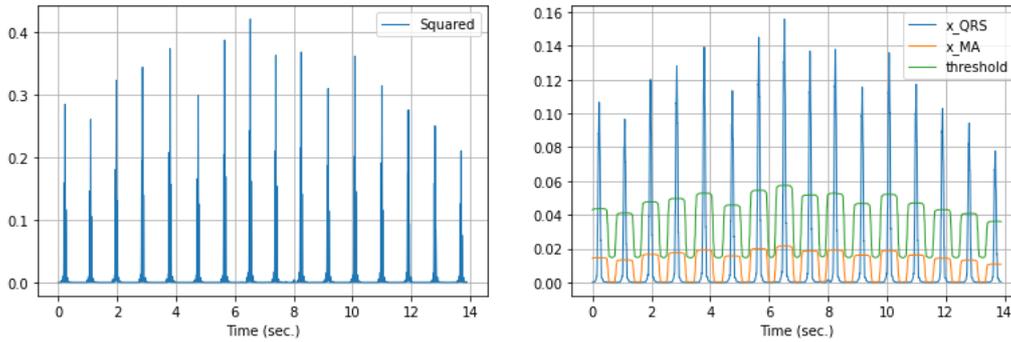


FIGURE 4.4 – Création de segments d'intérêt : sortie du filtre quadrature (gauche) et des filtres MA (droite) : signaux filtrés x_{QRS} et x_B , et seuil.

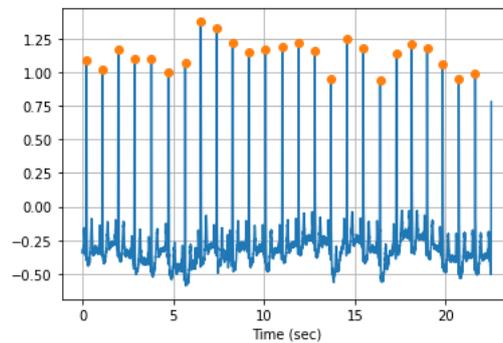


FIGURE 4.5 – Ondes R détectées par l'algorithme d'Elgendi sur les 20 premières secondes du jeu de données 101 de la base mitdb. Les paramètres utilisés ici sont $f_1 = 10\text{Hz}$, $f_2 = 25\text{Hz}$, $w_1 = 75\text{ms}$, $w_2 = 600\text{ms}$ et $\beta = 1$.

où β est un paramètre, et où \bar{x}_2 est la moyenne de la sortie x_2 du filtre quadrature.

La comparaison du signal x_{QRS} à ce seuil produit un signal "indicateur" $I[n]$

$$I[n] = 1 \quad \text{si} \quad x_{QRS}[n] \geq \tau_1[n], \quad I[n] = 0 \quad \text{sinon} . \quad (4.5)$$

Les composantes connexes de ce signal sont appelées *segments d'intérêt* S_1, S_2, \dots , qu'il reste à valider.

2. **Seuillage et rejet** : la validation se base sur la largeur de ces segments. Si celle-ci est inférieure à un certain seuil τ_2 (constant), le segment est rejeté. Elgendi suggère de prendre

$$\tau_2 = w_1 . \quad (4.6)$$

3. **Détection des ondes R** : les ondes R sont les maxima à l'intérieur des segments d'intérêt conservés.

Un exemple de résultats obtenus avec cette méthode est donné en FIGURE 4.5. On peut voir que les pics ont été correctement détectés. Ceci étant, les paramètres ont été choisis un peu au hasard, et rien n'indique que les détections seront correctes sur l'ensemble du signal, ni sur d'autres signaux de la base de données.

4.3.3 Optimisation des paramètres

L'article d'Elgendi propose de déterminer les paramètres optimaux par une recherche exhaustive à partir d'une base de données annotée, en l'occurrence la base mitdb que nous utilisons depuis le début.

Les valeurs testées sont

- Toutes les valeurs des intervalles de fréquences à l'intérieur de l'intervalle 1 – 26 Hz, par pas de 1 Hz.
- Toutes les valeurs de w_1 dans l'intervalle 55 – 111 ms, par pas de 5 ms.
- Toutes les valeurs de w_2 dans l'intervalle 555 – 694 ms, par pas de 10 ms.
- Toutes les valeurs de β dans l'intervalle 1 – 10%, par pas de 1%.

| | moyenne | incertitude |
|---------------------|---------|-------------|
| Largeur de l'onde P | 110ms | ± 20 ms |
| Largeur du QRS P | 100ms | ± 20 ms |
| Intervalle PQ/PR | 160ms | ± 40 ms |

TABLE 4.1 – Valeurs standard de certains paramètres, pour un sujet normal à un rythme de 60 battements par minute.

In fine, l'algorithme a été testé sur d'autres bases de données, atteignant des performances supérieures aux autres algorithmes de la même famille.

4.4 Filtrage adapté

Beaucoup d'approches pour la détection du complexe QRS sont basées sur des idées similaires, et débutent par un filtrage passe-bande suivi d'un filtrage quadrateur, des lissages et des seuillages avec des paramètres variés. On mentionnera ici, pour être complet, d'autres approches basées sur un filtrage adapté.

Le filtrage adapté répond au problème de détection d'un signal connu apparaissant dans une observation bruitée avec une latence (un décalage) inconnue. Le modèle est

$$y[n] = Ax[n - \tau] + \varepsilon[n], \quad (4.7)$$

où x est le signal connu (appelé signal de référence), τ est le décalage, A est une amplitude et ε représente le bruit.

Dans le cas où le bruit est supposé blanc, il est possible de démontrer que l'estimateur optimal de τ est donné par la corrélation

$$\tau_* = \underset{\tau}{\operatorname{argmax}} \sum_{\ell} y[\ell]x[\ell - \tau], \quad (4.8)$$

et on peut ensuite en déduire une estimée de l'amplitude.

Après pré-traitement (en particulier suppression de la ligne de base par filtrage passe-haut), on peut modéliser un signal ECG sous la forme

$$y[\ell] = \sum_k A_k x[\ell - \tau_k], \quad (4.9)$$

où x représente un signal PQRST de référence et les τ_k sont les positions des pics R . Le filtre adapté permet alors d'estimer ces décalages τ_k puis les amplitudes A_k .

La sortie du filtre adapté $z = K_h y$ est souvent oscillante. Une enveloppe peut être estimée par passage au signal analytique

$$E = |z_a|^2, \quad z_a = \frac{1}{2}(z + iHz), \quad \widehat{H}z(v) = 2\Theta(v)\hat{z}(v), \quad (4.10)$$

où Θ est la fonction de Heaviside. Il est possible de montrer que lorsque z est un signal oscillant, de la forme $z(t) = a(t) \cos(\phi(t))$, où $a(t) \geq 0$ pour tout t , et si l'amplitude a et la fréquence ϕ' sont lentement variables, alors $z_a(t) \approx a(t)e^{i\phi(t)}$, de sorte que $|z_a(t)|$ donne une estimée de l'enveloppe $a(t)$ du signal.

On décrit ci-dessous un algorithme récemment proposé [2] (voir aussi [8]) basé sur un filtrage adapté combiné à une estimation d'enveloppe par transformation de Hilbert.

1. Construction du filtre :

- Construction d'un signal de référence : par exemple un complexe QRS (ou PQRST), estimé sur le début du signal, ou un QRS (ou PQRST) moyen, obtenu sur une moyenne calculée sur un segment de plusieurs cycles, après détection des pics R sur ce segment (voir plus bas).
- Estimation du spectre de la ligne de base : généralement pris constant. On en déduit la réponse impulsionnelle du filtre.

2. Détection des ondes R :

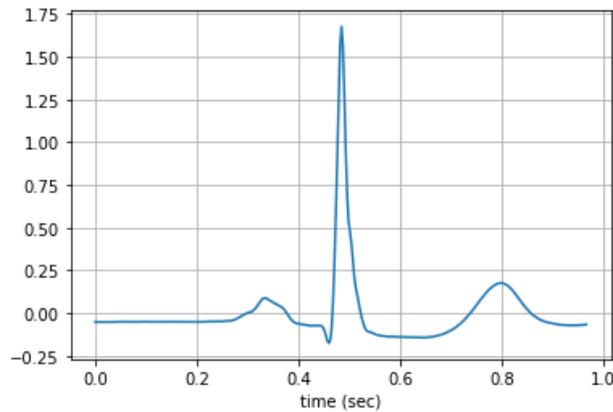


FIGURE 4.6 – Analyse du rythme cardiaque, sujet 101 de la base mitdb : PQRST moyen

- Filtrage adapté : $y \rightarrow z = K_h x$
- Calcul du signal analytique $z_a = (1 + iH)z/2$, et estimation d'enveloppe par calcul du module $|z_a|$.
- Détection des pics de $|z_a|$: seuillage, rejet des pics non physiologiques

Remarque 4.1 — Estimation d'un PQRST moyen. Il est ici nécessaire de disposer d'un signal QRS (ou PQRST) de référence. On peut par exemple extraire un exemple à partir des premières secondes du signal. Alternativement, il est possible de construire un tel signal de référence en utilisant les localisations des ondes R.

On présente en Figure 4.6 une estimation obtenue par la séquence d'opérations suivante.

1. Filtrage passe-haut (élimination de ligne de base) et à encoche (élimination des fréquences du courant électrique domestique).
2. Détection des ondes R (ici par l'algorithme d'Elgendi)
3. Construction de segments, de longueur inférieure à la moyenne de l'intervalle RR et centrés sur les pics R.
4. Moyenne des signaux supportés par ces segments.

4.5 Analyse de variabilité

Une fois détectés les QRS et les ondes R, on peut se préoccuper de l'analyse du rythme cardiaque. Ceci passe par

- Le calcul des intervalles RR
- Le calcul du rythme cardiaque, via l'expression (4.1).

4.5.1 Séparation de variations lentes et rapides

Un exemple est donné dans la Figure 4.7, où on a tracé sur la gauche le signal des différences RR, et à droite le rythme cardiaque $\rho[t]$. On peut y détecter une évolution relativement régulière, ainsi qu'un certain nombre d'accidents dont on peut se demander s'il s'agit d'artefacts ou de réelles arythmies.

Pour aller plus loin, il faut donc essayer de séparer les deux. Dans ce qui suit, on effectue un filtrage par un filtre médian pour estimer la tendance (et éliminer les variations violentes) pour on soustrait le signal filtré pour examiner les résidus.

Un filtre médian est similaire à un filtre de moyenne mobile, à ceci près qu'il remplace l'opération de moyenne par l'opération de médiane. On rappelle que la médiane d'une liste de K nombres (où K est impair), est la valeur qui a exactement $(K - 1)/2$ valeurs plus petites et $(K - 1)/2$ valeurs plus grandes.

Le filtre médian est alors défini comme suit : si N est un entier pair,

$$y[n] = \text{median}\{x[(n - N/2) : (n + N/2)]\}, \quad (4.11)$$

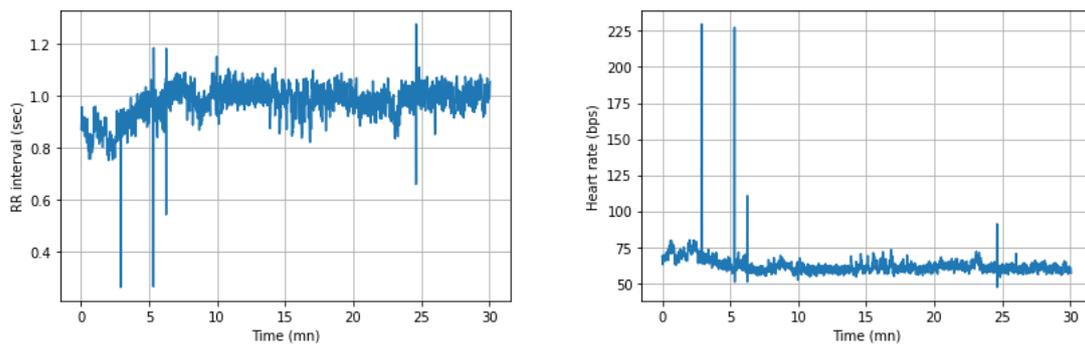


FIGURE 4.7 – Analyse du rythme cardiaque, sujet 101 de la base mitdb, après détection des complexes QRS et des ondes R par l’algorithme d’Elgendi. Gauche : intervalles RR. Droite : rythme cardiaque $\rho[t]$.

où $N + 1$ (impair) est la largeur de la fenêtre sur laquelle est calculée la médiane. Un exemple (calculé avec

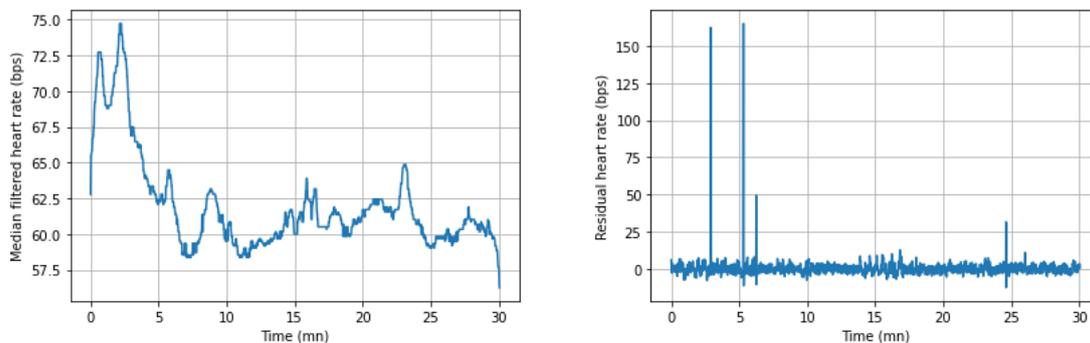


FIGURE 4.8 – Analyse du rythme cardiaque, sujet 101 de la base mitdb, après estimation du rythme cardiaque $\rho[t]$. Sortie d’un filtre médian (gauche) et résidu (droite)

un filtre médian de largeur 51) est représenté en Figure 4.8. On peut effectivement voir sur le tracé de gauche une évolution relativement lente du rythme cardiaque, qui a été séparée de variations très rapides représentées sur le tracé de droite. Ceci étant, cette séparation dépend fortement du choix de la taille de fenêtre, qui n’a pas été optimisé ici.

Une fois la ligne de base (les variations lentes) estimée, on peut essayer de faire une détection automatique de variations brusques à partir des résidus. Pour cela, on pourra comparer les résidus à une valeur seuil, par exemple calculée à partir de la distribution des résidus (voir Figure 4.9).

4.5.2 Analyse spectrale

L’analyse spectrale est difficile à mettre en œuvre sur des signaux de rythme cardiaque car ceux-ci sont échantillonnés irrégulièrement (on ne dispose de mesures de RR que sur lmes échantillons correspondant à un R). Il existe cependant des outils adaptés, par exemple le périodogramme de Lomb-Scargle. On n’ira pas plus loin ici.

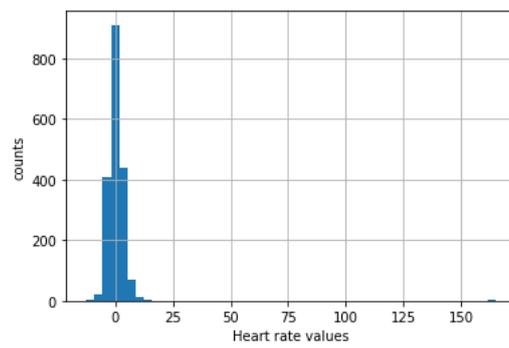


FIGURE 4.9 – Analyse du rythme cardiaque, sujet 101 de la base mi.tdb. Histogramme du résidu

| | | |
|----------|---|-----------|
| 5 | Électroencéphalogramme, généralités | 59 |
| 5.1 | Quelques éléments sur l'activité cérébrale et sa mesure | |
| 5.2 | Structure de données d'EEG et BCI | |

| | | |
|----------|--|-----------|
| 6 | Probabilités, estimation, décision | 65 |
| 6.1 | Éléments de théorie des probabilités | |
| 6.2 | Variabes aléatoires | |
| 6.3 | Statistique, estimation | |
| 6.4 | Classification, décision | |
| 6.5 | Analyse discriminante linéaire et quadratique | |
| 6.6 | Application aux données de <i>P300-speller</i> | |

5. Électroencéphalogramme, généralités

5.1 Quelques éléments sur l'activité cérébrale et sa mesure

Le cerveau¹ 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 que le réseau d'axones constitue la "matière blanche", une sorte de faisceaux de câbles.

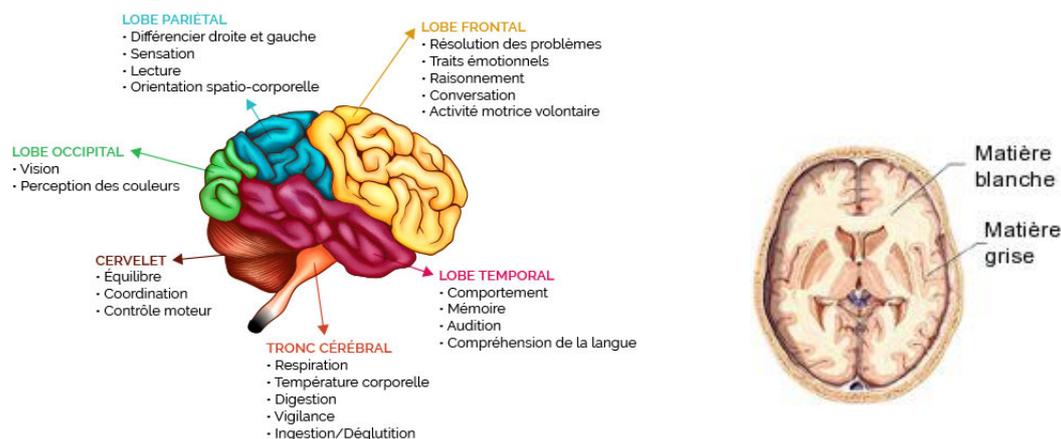


FIGURE 5.1 – Vue schématique d'un cerveau humain, avec les principales aires fonctionnelles (gauche), matière grise et matière blanche (droite)

Les informations circulent dans les neurones par influx électrique (voir FIGURE 5.2). Pour que le signal se propage, il faut que le potentiel électrique atteigne un certain seuil. Il existe des synapses excitatrices

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>

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.

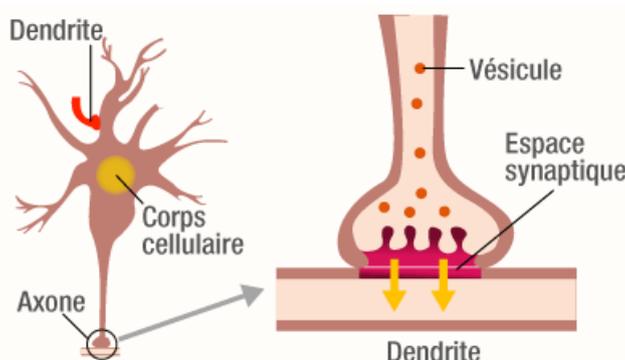


FIGURE 5.2 – Fonctionnement d'une synapse

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 5.1 gauche).

5.1.1 Quelques éléments d'imagerie cérébrale

Les lignes ci-dessous sont tirées de l'encyclopédie en ligne Wikipedia². 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 5.3). 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

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

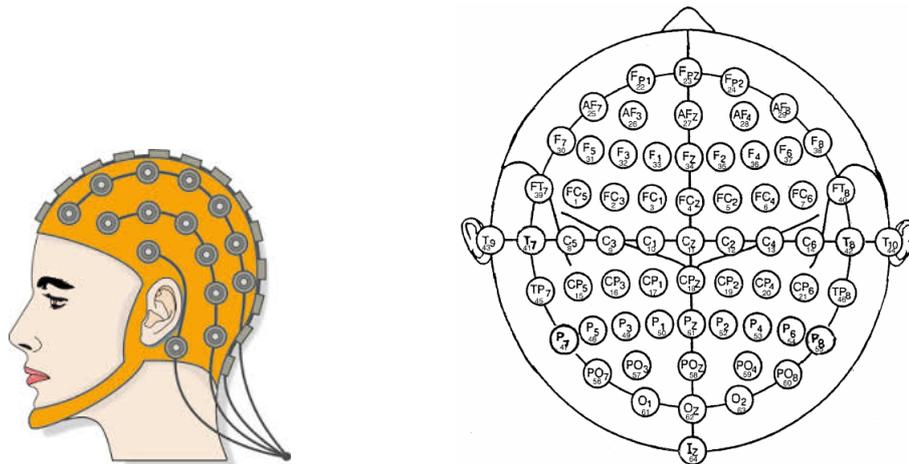


FIGURE 5.3 – Une exemple de dispositif de capteurs EEG. A gauche, représentation d'un casque EEG. A droite, nomenclature des capteurs dans un dispositif standard.

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 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.

5.1.2 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 vidéo. 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 5.4 – Un exemple de signal EEG “multicapteur”. En abscisse le temps, en ordonnée les signaux mesurés sur un certain nombre de capteurs.

5.2 Structure de données d’EEG et BCI

5.2.1 Données EEG, potentiels évoqués

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 5.4.

Chacun des signaux temporels associé 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 $v_n^c = v^c((n-1)/\eta)$, où $n = 1, \dots, L$ est la longueur du signal discrétisé, et η 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 de ce type (à deux entrées), qui peuvent être réorganisés en un tableau de plus grande dimension.

Chez l’humain, on a tendance à distinguer différents types de signaux EEG, appelés ondes ou rythmes, qui se caractérisent par leur fréquence principale ainsi que leur localisation dans l’espace des capteurs. Les principaux rythmes sont

- Le rythme alpha, dans la bande de fréquences comprises entre 8.5Hz et 12Hz. Les ondes alpha caractérisent un état de conscience apaisé, et sont principalement émises lorsque le sujet a les yeux fermés
- Le rythme bêta, dans la bande de fréquences comprises entre 12 Hz et 45 Hz. Les ondes bêta apparaissent généralement dans un premier cas d’éveil calme, encore appelé éveil interne, lorsque le sujet, les yeux fermés, est en état d’attention diffuse sans tâche perceptuelle ou mentale spécifiée, et le rythme alpha est alors souvent présent avec des activités bêta superposées ou alternées.
- Le rythme gamma, dans une bande de fréquences entre 35 Hz et 80Hz. Elles ont été récemment impliquées dans les processus de liage perceptif.
- Le rythme delta, dans les basses fréquences jusqu’à 4 Hz. Normales chez le très jeune enfant, les ondes delta peuvent ensuite caractériser certaines lésions cérébrales ou le sommeil profond
- Le rythme thêta, dans une bande de fréquences entre 4.5Hz et 8Hz. On observe les ondes thêta

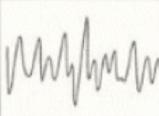
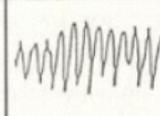
| DELTA | THETA | ALPHA | BETA | GAMMA |
|---|---|---|--|---|
| Inférieure à 4Hz | 4 à 8 Hz | 8 à 13 Hz | 13 à 35 Hz | Supérieure à 35Hz |
| Sommeil profond | Somnolence | Relaxation | Une attention à son maximum | Excitation |
|  |  |  |  |  |

FIGURE 5.5 – Illustrations des principaux rythmes cérébraux (Source : <http://tpe-batement-binauraux.webnode.fr/quest-ce-que-les-battements-binauraux-/>)

principalement chez l'enfant, l'adolescent et le jeune adulte. Elles caractérisent également certains états de somnolence, d'hypnose ou de méditation, ainsi que lors de la mémorisation d'information. Le rythme thêta est présent dans la phase de sommeil paradoxal qui s'accompagne souvent d'un rêve.

- Les fuseaux de sommeil sont des trains d'ondes comprises entre 12 et 16 Hz, caractéristiques de la phase de sommeil léger. Également reliés aux ondes SMR (rythmes sensorimoteurs) ou Mu.

En dehors des rythmes mentionnés ci-dessus, on distingue également les On se focalisera plus particulièrement sur les potentiels évoqués, qui traduisent la réponse du système nerveux central à un stimulus (visuel, auditif, moteur, cognitif ou sensitif) ou un événement interne (émotion, pensée, souvenir). Les changements de potentiels suivant un stimulus sont généralement très faibles, il est nécessaire d'enregistrer plusieurs fois le signal EEG associé à une même tâche ou à une stimulation répétée. Une moyenne sur ces répétitions permet d'augmenter le rapport signal à bruit et ainsi de dissocier les réponses neuronales associées à l'activité de fond et au bruit.

Les différentes composantes d'un potentiel évoqué sont schématisées en Figure 5.6, on elles se distinguent par leur polarité (P ou N) et le délai d'apparition après le stimulus (exprimé en millisecondes).

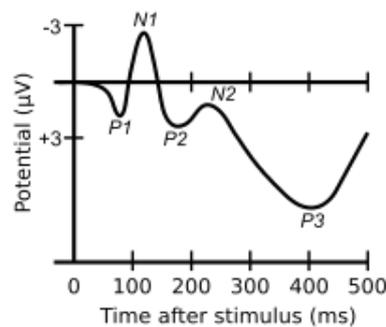


FIGURE 5.6 – Les différentes composantes d'un potentiel évoqué

5.2.2 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 5.7), 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 12 contenant le caractère



FIGURE 5.7 – Une matrice de caractères utilisée dans le dispositif *P300 speller*.

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.

La stratégie mise en place se base schématiquement sur les étapes suivantes

- A partir d'un jeu de données d'apprentissage, trouver une représentation des signaux correspondant aux caractères cible et non-cible sous forme de points dans l'espace (de dimension à déterminer)
- Construire une règle de décision
- Utiliser la règle de décision sur les données de test

Pour ce qui est de la règle de décision, elle se construit sommairement de la façon suivante

1. Estimer les moyennes des deux classes (*cible* et *non-cible*) de signaux (deux signaux moyens) : $\mu_1, \mu_0 \in \mathbb{R}^L$.
2. Estimer la matrice de covariance des signaux : $\Sigma \in \mathcal{M}_L$
3. Calculer un vecteur discriminant : $\Phi = \Sigma^{-1}(\mu_1 - \mu_0)$, et un seuil τ .
4. Pour chaque nouveau signal x , calculer le produit scalaire $s = \Phi^T x$ (score du signal).
5. Si $s \geq \tau$ on décide que la ligne ou la colonne ayant conduit à la réponse x contient le caractère cible, et ne contient pas dans le cas contraire $s < \tau$.
6. La ligne et la colonne permettent d'identifier le caractère cible.

Ces étapes constituent les étapes d'une technique de classification appelée Analyse Discriminante Linéaire (LDA). Comme on le verra le problème du P300 speller est un peu différent d'un problème de classification, on adaptera cette construction en conséquence.

6. Probabilités, estimation, décision

6.1 Éléments de théorie des probabilités

6.1.1 Modèle probabiliste

Les modèles probabilistes s'introduisent naturellement dans deux contextes différents :

- Il existe une grande quantité de systèmes qui sont trop complexes pour pouvoir être décrits de façon déterministe. Par exemple, supposons simplement que nous ayons à décrire un système de 10^{10} particules, évoluant librement dans un volume fini donné, interagissant par chocs élastiques. Clairement, les lois fondamentales de la dynamique permettent théoriquement de calculer les trajectoires individuelles de chacune des particules. Cependant, un tel calcul est irréalisable en pratique, compte tenu du grand nombre de possibilités dont il faut tenir compte. Dans le même ordre d'idée, toute mesure d'un signal est bruitée, le bruit provenant de multiples sources incontrôlées, bien qu'elles proviennent généralement de systèmes physiques bien connus, mais de très grande dimension. Là encore une description déterministe est inaccessible.
- Il existe aussi des systèmes qui ont une nature intrinsèquement probabiliste. C'est en particulier le cas des systèmes quantiques, si on s'en tient aux interprétations standard. Étant donné un système quantique, on ne peut généralement pas prédire à l'avance ce que sera le résultat d'une mesure. On doit se contenter de connaître les résultats possibles, auxquels on peut affecter une probabilité.

Dans de tels cas de figure, la modélisation est une modélisation probabiliste. Une modélisation probabiliste consiste essentiellement en un ensemble (discret ou continu) de "possibles" (appelés *événements*), auxquels on associe un nombre (la probabilité de l'évènement) qui en caractérise la vraisemblance.

■ **Exemple 6.1** Prenons l'exemple d'une pièce normale, avec laquelle on tire à pile ou face. Il est quasiment impossible de calculer avec une précision suffisante la trajectoire de la pièce pour prédire le résultat. Par contre, il est possible d'associer des probabilités aux résultats possibles. Par exemple, si on note P et F les 2 résultats possibles d'un tirage, et si on effectue un seul tirage, on a donc deux résultats possibles, avec probabilités $1/2$. Si on effectue 2 tirages, on a 4 résultats possibles équiprobables ; et ainsi de suite. On peut faire le tableau suivant :

| P | F | PP | PF | FP | FF | ... | $PP\dots P$ | ... | $FF\dots F$ |
|---------------|---------------|---------------|---------------|---------------|---------------|-----|-------------|-----|-------------|
| $\frac{1}{2}$ | $\frac{1}{2}$ | $\frac{1}{4}$ | $\frac{1}{4}$ | $\frac{1}{4}$ | $\frac{1}{4}$ | ... | 2^{-n} | ... | 2^{-n} |

On associe ainsi une probabilité à chacun des évènements considérés. On vérifie facilement que la somme des probabilités associées à un nombre de tirages n fixé vaut 1. ■

■ **Exemple 6.2** Marche au hasard : Prenons l'exemple d'un ivrogne, à la démarche hésitante. Si il a une jambe plus courte que l'autre, on peut penser qu'il a une probabilité p d'aller vers la gauche (G), et $q = 1 - p$ d'aller vers la droite (D). On peut faire le tableau suivant, représentant les configurations possibles pour un nombre n de pas donné :

| | | | | | | | | | |
|-----|-----|-------|------|------|-------|-----|-------------|-----|-------------|
| G | D | GG | GD | DG | DD | ... | $GG\dots G$ | ... | $FF\dots F$ |
| p | q | p^2 | pq | pq | q^2 | ... | p^{-n} | ... | q^{-n} |

On vérifie aussi dans ce cas que pour un nombre fixé de pas n , la somme des probabilités vaut 1. ■

■ **Exemple 6.3** On tire à pile ou face avec une pièce ordinaire. Si le résultat est pile (P), on jette un dé à 6 faces. Si le résultat est face (F), on retire à pile ou face. L'ensemble des résultats possibles est

$$\Omega = \{(P, 1), (P, 2), (P, 3), (P, 4), (P, 5), (P, 6), (F, P), (F, F)\}$$

| | | | | | | | | |
|------------------------|----------------|----------------|----------------|----------------|----------------|----------------|---------------|---------------|
| ω | $(P, 1)$ | $(P, 2)$ | $(P, 3)$ | $(P, 4)$ | $(P, 5)$ | $(P, 6)$ | (F, P) | (F, F) |
| $\mathbb{P}\{\omega\}$ | $\frac{1}{12}$ | $\frac{1}{12}$ | $\frac{1}{12}$ | $\frac{1}{12}$ | $\frac{1}{12}$ | $\frac{1}{12}$ | $\frac{1}{4}$ | $\frac{1}{4}$ |

Probabilités et statistique

La théorie des probabilités et la statistique sont deux sciences complémentaires. Plus précisément, la théorie des probabilités est une *science prédictive* : son objectif est de construire des modèles décrivant la réalité, et prédisant les fréquences d'apparition de phénomènes, ou résultats de mesure.

Comment connaît-on les probabilités de tels évènements ? La réponse à cette question est fournie par la théorie statistique, qui permet d'estimer ces probabilités à partir d'expériences. La statistique est donc une *science descriptive*.

Comment fait-on le lien entre les deux théories ? la relation est souvent donnée par ce que l'on nomme des *théorèmes limites*, comme par exemple la *loi des grands nombres*, ou encore le *théorème central limite*. L'exemple suivant permet d'illustrer la situation.

On considère un dé usuel à 6 faces. Si le dé est un dé ordinaire, les 6 faces sont équiprobables, et les 6 faces ont donc probabilités $p_1 = p_2 = \dots = p_6 = 1/6$. Supposons maintenant que l'on fasse N expériences, c'est à dire N tirages, et que l'on obtienne n_1 fois la face 1, ... et n_6 fois la face 6. Posons alors pour $i = 1, \dots, 6$: $\hat{p}_i = n_i/N$. En général, on n'a pas $\hat{p}_i = p_i$, mais le "bon sens" suggère que lorsque N est grand, on devrait avoir $\hat{p}_i \approx p_i$. La loi des grands nombres permet de donner un sens plus précis à cette assertion. On montre en fait que pour tout $\delta > 0$,

$$\mathbb{P}\{|\hat{p}_i - p_i| > \delta\} \rightarrow 0 \quad \text{quand } N \rightarrow \infty.$$

Le théorème central limite permet de préciser cette limite.

6.1.2 Calcul des probabilités

Ensembles d'évènements, union, intersection

Étant donné un ensemble Ω et une application \mathbb{P} comme ci-dessus, nous allons maintenant étendre ces notions. \mathbb{P} associe un nombre à tout élément de Ω ; on veut également être capables d'associer une probabilité à des sous-ensembles de Ω : calculer les probabilités pour que le résultat appartienne à un certain ensemble $A \subset \Omega$ de résultats possibles. Les sous-ensembles $A \subset \Omega$ auxquels on peut associer une probabilité seront appelés *évènements*.

Pour cela, examinons cas par cas un certain nombre de situations possibles.

Le cas fini

Supposons tout d'abord que Ω soit un ensemble fini, et qu'on dispose d'une application $\mathbb{P} : \Omega \rightarrow \mathbb{R}^+$, définie donc sur les éléments de Ω . Pour tout sous-ensemble $A \subset \Omega$, il est alors facile d'introduire $\mathbb{P}\{A\}$:

Définition 6.1 Soit Ω un ensemble fini, et soit $A \subset \Omega$. On définit alors $\mathbb{P}\{A\}$ par

$$\mathbb{P}\{A\} = \sum_{\omega \in A} \mathbb{P}\{\omega\} . \quad (6.1)$$

Il est immédiat qu'avec cette définition, on a nécessairement

$$\mathbb{P}\{\Omega\} = 1 . \quad (6.2)$$

Ayant posé cette définition, on se retrouve tout naturellement dans un contexte de théorie des ensembles. On a alors :

Proposition 6.1 1. Si $A, B \subset \Omega$ sont deux ensembles disjoints (c'est à dire $A \cap B = \emptyset$), alors

$$\mathbb{P}\{A \cup B\} = \mathbb{P}\{A\} + \mathbb{P}\{B\} . \quad (6.3)$$

2. Plus généralement : si $A, B \subset \Omega$,

$$\mathbb{P}\{A \cup B\} = \mathbb{P}\{A\} + \mathbb{P}\{B\} - \mathbb{P}\{A \cap B\} . \quad (6.4)$$

3. Si A^c est le complémentaire de A dans Ω :

$$\mathbb{P}\{A^c\} = 1 - \mathbb{P}\{A\} . \quad (6.5)$$

Le cas infini dénombrable

Supposons maintenant que Ω soit un ensemble infini dénombrable. Supposons aussi qu'on ait introduit une application $\mathbb{P} : \Omega \rightarrow \mathbb{R}^+$. On peut alors étendre \mathbb{P} à des sous-ensembles plus généraux de Ω , comme dans le cas fini, en définissant (comme dans la DÉFINITION 6.1) pour toute famille (finie ou infinie dénombrable) $A = \{a_1, a_2, \dots, a_i \in \Omega\}$ la probabilité de l'évènement A par

$$\mathbb{P}\{A\} = \sum_n \mathbb{P}\{a_n\} = \mathbb{P}\{a_1\} + \mathbb{P}\{a_2\} + \dots . \quad (6.6)$$

La formule (6.2) est bien évidemment toujours valable.

■ **Exemple 6.4** On considère une variante du jeu de pile ou face, dans laquelle les deux faces de la pièce ont probabilités $1/2$, mais la pièce est tirée jusqu'à ce que "face" apparaisse. L'ensemble des résultats possibles est

$$\Omega = \{F, PF, PPF, \dots\} ,$$

et les probabilités des éléments de Ω sont données par

| ω | F | PF | PPF | $PPPF$ | \dots | $P \dots PF$ (n fois P) |
|------------------------|---------------|---------------|---------------|----------------|---------|-------------------------------|
| $\mathbb{P}\{\omega\}$ | $\frac{1}{2}$ | $\frac{1}{4}$ | $\frac{1}{8}$ | $\frac{1}{16}$ | \dots | $2^{-(n+1)}$ |

A partir de là, on peut calculer les probabilités des évènements (les sous-ensembles de Ω) : par exemple, la probabilité pour que le nombre de P soit inférieur à 4 est donnée par

$$\mathbb{P}\{\{F, PF, PPF, PPPF\}\} = \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \frac{1}{16} = \frac{15}{16} .$$

On a évidemment

$$\mathbb{P}\{\Omega\} = \sum_{n=1}^{\infty} 2^{-n} = \frac{1}{1-1/2} - 1 = 1 .$$

■

Le cas continu

Le cas infini continu est un peu plus complexe. Prenons l'exemple d'un nombre tiré au hasard dans l'intervalle $[0, 1]$, toutes les valeurs ayant même probabilité. On a donc $\Omega = [0, 1]$. Mais Ω étant un ensemble continu, les probabilités des valeurs individuelles sont nulles :

$$\mathbb{P}\{t\} = 0 \quad \forall t \in [0, 1].$$

Dans ce cas, il est nécessaire d'utiliser directement des sous-ensembles de Ω pour construire le modèle. Toutes les valeurs étant équiprobables, on définit alors \mathbb{P} par

$$\mathbb{P}\{[a, b]\} = b - a$$

pour tout intervalle $[a, b[\subset \Omega$. Pour définir $\mathbb{P}\{A\}$ pour des sous-ensembles (presque) quelconques $A \subset \Omega$, on peut alors s'inspirer de la PROPOSITION 6.1, et écrire par exemple, pour toute famille d'intervalles disjoints deux à deux $I_1 = [a_1, b_1[, I_2 = [a_2, b_2[, \dots$,

$$\mathbb{P}\{I_1 \cup I_2 \cup \dots\} = \mathbb{P}\{I_1\} + \mathbb{P}\{I_2\} + \dots = (b_1 - a_1) + (b_2 - a_2) + \dots$$

Dans ce cas, nous serons naturellement amenés à considérer des intégrales, comme nous le verrons dans la section 6.1.2.

Définition fondamentale

Motivé par les discussions précédentes, on peut maintenant donner la définition suivante d'un modèle probabiliste au sens de Kolmogorov (dans la littérature mathématique, on parle d'espace probabilisé).

Définition 6.2 Un modèle probabiliste consiste en

1. Un ensemble de résultats possibles Ω (appelé *univers*, *espace de configuration* ou *référentiel*).
2. Une *tribu*, c'est à dire une famille \mathcal{F} de parties de Ω , appelées *événements*, telle que :
 - $\Omega \in \mathcal{F}$.
 - Si $A \in \mathcal{F}$, alors son complémentaire $A^c = \Omega \setminus A \in \mathcal{F}$.
 - Si $A_1, A_2, \dots \in \mathcal{F}$ est un ensemble (fini ou infini dénombrable) de parties de Ω , alors $A_1 \cup A_2 \cup \dots \in \mathcal{F}$.
 - Si $A_1, A_2, \dots \in \mathcal{F}$ est un ensemble (fini ou infini dénombrable) de parties de Ω , alors $A_1 \cap A_2 \cap \dots \in \mathcal{F}$.
3. Une application $\mathbb{P} : A \in \mathcal{F} \rightarrow \mathbb{P}\{A\} \in [0, 1]$, telle que
 - $\mathbb{P}\{\Omega\} = 1$.
 - Si $A_1, A_2, \dots \in \mathcal{F}$ est un ensemble de parties de Ω deux à deux disjointes,

$$\mathbb{P}\{A_1 \cup A_2 \cup \dots\} = \mathbb{P}\{A_1\} + \mathbb{P}\{A_2\} + \dots \quad (6.7)$$

L'ensemble \mathcal{F} possédant les propriétés ci-dessus est parfois appelé *tribu*. Une conséquence immédiate de la définition est que $\emptyset \in \mathcal{F}$, et que $\mathbb{P}\{\emptyset\} = 0$.

Calcul sur les événements

La formulation en termes de théorie des ensembles permet de formuler simplement un certain nombre de résultats. En particulier, le résultat suivant est une conséquence immédiate des définitions.

Proposition 6.2 Soit $(\Omega, \mathcal{F}, \mathbb{P})$ un modèle probabiliste.

1. Si $A, B \in \mathcal{F}$ et $A \subset B$, alors $\mathbb{P}\{A\} \leq \mathbb{P}\{B\}$.
2. Si $A \in \mathcal{F}$, $\mathbb{P}\{A^c\} = 1 - \mathbb{P}\{A\}$.
3. Si $A, B \in \mathcal{F}$ et $A \subset B$, alors $\mathbb{P}\{B \setminus A\} = \mathbb{P}\{B\} - \mathbb{P}\{A\}$.
4. Si $A, B \in \mathcal{F}$, $\mathbb{P}\{A \cup B\} = \mathbb{P}\{A\} + \mathbb{P}\{B\} - \mathbb{P}\{A \cap B\}$.

Preuve : 1) On a $B = A \cup (B \setminus A)$, et $A \cap (B \setminus A) = \emptyset$. Donc $\mathbb{P}\{B\} = \mathbb{P}\{A\} + \mathbb{P}\{B \setminus A\} \geq \mathbb{P}\{A\}$.
 2) est immédiat : $\mathbb{P}\{A\} + \mathbb{P}\{A^c\} = \mathbb{P}\{A \cup A^c\} = \mathbb{P}\{\Omega\} = 1$.
 3) Si $A \subset B : B = A \cup (B \setminus A)$ avec $A \cap (B \setminus A) = \emptyset$. Donc $\mathbb{P}\{B\} = \mathbb{P}\{A\} + \mathbb{P}\{B \setminus A\}$.
 4) $A \cup B = A \cup (B \cap A^c)$, donc $\mathbb{P}\{A \cup B\} = \mathbb{P}\{A\} + \mathbb{P}\{B \cap A^c\}$. De plus, $\mathbb{P}\{B\} = \mathbb{P}\{B \cap A\} + \mathbb{P}\{B \cap A^c\}$, d'où le résultat. ♠

Indépendance ; probabilités conditionnelles

Rappelons que l'opération d'intersection des ensembles correspond au *et* logique des évènements.

Définition 6.3 Soit $(\Omega, \mathcal{F}, \mathbb{P})$ un modèle probabiliste

1. On dit que deux évènements $A, B \in \mathcal{F}$ sont indépendants si

$$\mathbb{P}\{A \cap B\} = \mathbb{P}\{A\}\mathbb{P}\{B\}. \quad (6.8)$$

2. Une famille finie (A_1, A_2, \dots, A_n) d'éléments de \mathcal{F} est une famille d'évènements indépendants si

$$\mathbb{P}\{A_1 \cap A_2 \cap \dots \cap A_n\} = \mathbb{P}\{A_1\}\mathbb{P}\{A_2\} \dots \mathbb{P}\{A_n\}. \quad (6.9)$$

Remarque 6.3 Il est important de signaler que si (A_1, A_2, \dots, A_n) est une famille d'évènements indépendants, les évènements A_i sont indépendants deux à deux : $\mathbb{P}\{A_i \cap A_j\} = \mathbb{P}\{A_i\}\mathbb{P}\{A_j\}$ pour tous i, j . La réciproque n'est pas vraie, comme le montre l'exemple suivant.

■ **Exemple 6.5** On tire à pile ou faces 2 fois avec une pièce non biaisée. On considère les évènements suivants : $A = \{PP, PF\}$, $B = \{PF, FF\}$, $C = \{FF, PP\}$, qui ont tous probabilité 1/2. Il est facile de voir que A, B, C sont deux à deux indépendants, mais que $\mathbb{P}\{A \cap B \cap C\} = 0$, de sorte que la famille $\{A, B, C\}$ n'est pas une famille d'évènements indépendants. ■

Lorsque deux évènements A et B sont indépendants, le fait que A soit réalisé n'apporte aucune information sur le fait que B soit réalisé ou pas. Par contre, ça n'est plus le cas si A et B ne sont pas indépendants. Il est dans ce cas utile d'introduire la notion de *probabilité conditionnelle*, qui permet de calculer la probabilité $\mathbb{P}\{B|A\}$ de B , supposant A réalisé.

Définition 6.4 Soit $(\Omega, \mathcal{F}, \mathbb{P})$ un modèle probabiliste. Si A et B sont deux évènements quelconques, la probabilité de A sachant B (ou *conditionnée par B*) est le nombre

$$\mathbb{P}\{A|B\} = \frac{\mathbb{P}\{A \cap B\}}{\mathbb{P}\{B\}} \quad (6.10)$$

■ **Exemple 6.6** On considère deux dés non pipés. Calculer la probabilité pour que le résultat obtenu avec le premier soit un 4, sachant que le résultat du "bi-tirage" appartient à l'ensemble

$$B = \{(3, 6), (4, 5), (4, 6), (5, 4), (5, 5), (5, 6), (6, 3), (6, 4), (6, 5), (6, 6)\}.$$

L'espace de configurations est $\Omega = \{(i, j), i, j = 1, \dots, 6\}$ et $\text{card}(\Omega) = 36$. Pour tout $\omega \in \Omega$, $\mathbb{P}\{\omega\} = 1/36$.

$$A = \{(4, 1), (4, 2), \dots, (4, 6)\}.$$

$\mathbb{P}\{A \cap B\} = 2/36$, et $\mathbb{P}\{B\} = 10/36$. On a donc

$$\mathbb{P}\{A|B\} = \frac{1}{5}.$$

Sous forme multiplicative, on peut aussi écrire

$$\mathbb{P}\{A \cap B\} = \mathbb{P}\{A|B\}\mathbb{P}\{B\}, \quad (6.11)$$

d'où on déduit la formule de Bayes

$$\mathbb{P}\{A|B\}\mathbb{P}\{B\} = \mathbb{P}\{B|A\}\mathbb{P}\{A\}. \quad (6.12)$$

On en déduit également le résultat plus général suivant :

Proposition 6.4 Pour toute famille finie d'évènements (A_1, A_2, \dots, A_n) :

$$\mathbb{P}\{A_1 \cap A_2 \cap \dots \cap A_n\} = \mathbb{P}\{A_1\}\mathbb{P}\{A_2|A_1\}\mathbb{P}\{A_3|A_1 \cap A_2\} \dots \times \mathbb{P}\{A_n|A_1 \cap \dots \cap A_{n-1}\}. \quad (6.13)$$

Preuve : Il suffit de procéder de proche en proche :

$$\begin{aligned}
 \mathbb{P}\{A_1 \cap \dots \cap A_n\} &= \mathbb{P}\{A_n | A_1 \cap \dots \cap A_{n-1}\} \mathbb{P}\{A_1 \cap \dots \cap A_{n-1}\} \\
 &= \mathbb{P}\{A_n | A_1 \cap \dots \cap A_{n-1}\} \mathbb{P}\{A_{n-1} | A_1 \cap \dots \cap A_{n-2}\} \\
 &\quad \times \mathbb{P}\{A_1 \cap \dots \cap A_{n-2}\} \\
 &= \dots \\
 &= \mathbb{P}\{A_1\} \mathbb{P}\{A_2 | A_1\} \mathbb{P}\{A_3 | A_1 \cap A_2\} \dots \\
 &\quad \times \mathbb{P}\{A_n | A_1 \cap \dots \cap A_{n-1}\},
 \end{aligned}$$

ce qui prouve la proposition. ♠

Une autre propriété utile est la formule de probabilité totale

Proposition 6.5 Pour toute famille dénombrable d'évènements (A_1, A_2, \dots) telle que $A_i \cap A_j = \emptyset$ pour tous i, j et $\bigcup A_i \subset \Omega$, on a pour tout $B \in \mathcal{F}$

$$\mathbb{P}\{B\} = \sum_i \mathbb{P}\{B | A_i\} \mathbb{P}\{A_i\}. \quad (6.14)$$

Preuve : Ecrivons la propriété de distributivité

$$\mathbb{P}\{B\} = \mathbb{P}\{B \cap (A_1 \cup A_2 \cup \dots)\} = \mathbb{P}\{(B \cap A_1) \cup (B \cap A_2) \cup \dots\}.$$

Les $B \cap A_i$ étant disjoints deux à deux, il suffit d'appliquer (6.7) puis la formule de Bayes. ♠

Probabilités et intégrales

La façon la plus usuelle de construire un modèle probabiliste dans le cas continu est basée sur l'heuristique suivante, que nous décrivons tout d'abord en dimension 1, par exemple dans le cas $\Omega \subset \mathbb{R}$. On a déjà vu que dans ce cas, les nombres $\mathbb{P}\{t\}, t \in \Omega$ n'ont pas de signification. On considère plutôt un petit intervalle de la forme $\Delta t = [t, t + \delta t]$, et on écrit

$$\mathbb{P}\{\Delta t\} = \mathbb{P}\{[t, t + \delta t]\} \approx \rho(t) \delta t;$$

ceci revient à écrire que la probabilité se comporte, en première approximation, de façon linéaire par rapport à la taille de l'intervalle, tout du moins lorsque celui-ci est petit, les variations sur des distances plus grandes étant décrites par une fonction ρ .

Etant donné maintenant un intervalle $I = [a, b]$, on commence par le "découper" en petits intervalles de tailles $\delta t_1, \delta t_2 \dots \delta t_N$, centrés sur les valeurs t_1, t_2, \dots, t_N , et on écrit

$$\mathbb{P}\{I\} = \rho(t_1) \delta t_1 + \rho(t_2) \delta t_2 + \dots + \rho(t_N) \delta t_N.$$

Or, ceci n'est autre qu'une approximation de l'intégrale

$$\int_I \rho(t) dt = \int_a^b \rho(t) dt$$

par une somme finie (à la Riemann). Ceci suggère de définir

$$\mathbb{P}\{I\} = \int_I \rho(t) dt.$$

De plus étant donné K intervalles I_1, \dots, I_K , disjoints deux à deux, et si $A = I_1 \cup I_2 \cup \dots \cup I_K$, on peut écrire

$$\mathbb{P}\{A\} = \int_{I_1} \rho(t) dt + \int_{I_2} \rho(t) dt + \dots + \int_{I_K} \rho(t) dt = \int_A \rho(t) dt.$$

Ainsi, dans les cas où Ω est un ensemble infini continu, on peut (sauf dans certains cas pathologiques) définir l'application \mathbb{P} à partir d'une fonction ρ , par la formule

$$\mathbb{P}\{A\} = \int_A \rho(t) dt. \quad (6.15)$$

■ **Définition 6.5** La fonction ρ ci-dessus est appelée *densité de probabilités*.

■ **Exemple 6.7** Si on reprend l'exemple ci-dessus d'un nombre tiré au hasard dans l'intervalle $[0, 1]$, on a bien

$$\mathbb{P}\{[a, b]\} = b - a = \int_a^b dt ,$$

de sorte que cette probabilité est bien construite à partir d'une densité, qui est dans ce cas

$$\rho(t) = \chi_{[0,1]}(t) .$$

Remarque 6.6 Nous avons utilisé l'intégrale de Riemann pour introduire la notion de densité de probabilités. Cependant, une formulation plus rigoureuse et générale de la théorie doit être basée sur la notion de mesure et l'intégrale de Lebesgue. Ceci ne change toutefois pas grand chose tant qu'il s'agit de faire des calculs explicites dans la plupart des cas classiques.

Lorsqu'un modèle probabiliste est construit à partir d'une densité de probabilités, celle-ci doit satisfaire un certain nombre de conditions. On en donne les deux les plus importantes ci-dessous.

Proposition 6.7 Soit ρ une densité de probabilités sur $\Omega \subset \mathbb{R}$.

1. $\rho(\omega) \geq 0$ pour presque tout $\omega \in \Omega$.
2. $\int_{\Omega} \rho(t) dt = 1$

Preuve : Tout d'abord, supposons qu'il existe un ensemble $A \subset \Omega$ de mesure non nulle, tel que $\rho(\omega) < 0$ pour tout $\omega \in A$. Alors on aurait $\mathbb{P}\{A\} = \int_A \rho(\omega) d\omega < 0$, ce qui est impossible. Par ailleurs, on a évidemment $\int_{\Omega} \rho(\omega) d\omega = \mathbb{P}\{\Omega\} = 1$ ♠

Remarque 6.8 une densité n'est pas à proprement parler une probabilité. Elle est certes à valeurs positive, mais n'a aucune raison d'être toujours inférieure à 1 : il peut exister des valeurs t telles que $\rho(t) \geq 1$.

En dimension supérieure, on peut procéder de façon similaire. Par exemple, si $\Omega \subset \mathbb{R}^n$, on est conduit à introduire une densité de probabilités ρ , qui est cette fois une fonction de plusieurs variables, ou d'une variable vectorielle $x = (x_1, \dots, x_n)$:

$$\rho : \Omega \rightarrow \mathbb{R}^+ ,$$

qui est telle que

$$\int_{\Omega} \rho(x) dx = 1 .$$

et

$$\rho(x) \geq 0 , \quad \forall x \in \Omega .$$

On peut aussi s'intéresser à des Ω plus compliqués (par exemple en considérant des points tirés au hasard sur des sphères ou dans des objets géométriques complexes). On en verra des exemples par la suite.

6.2 Variables aléatoires

Les modèles probabilistes que nous avons décrits dans le chapitre précédent ne donnent pas un cadre suffisamment large pour les applications. On introduit alors les *variables aléatoires*, qui permettent d'enrichir le calcul des probabilités.

6.2.1 Généralités

Commençons par deux exemples simples.

■ **Exemple 6.8** Dans un expérience de N tirages de "pile ou face", à chaque $\omega \in \Omega$ on peut associer le nombre de fois où on a obtenu "face" dans $\omega = (\omega_{i_1}, \dots, \omega_{i_N})$. Ceci définit donc une fonction

$$N : \omega \in \Omega \rightarrow N(\omega) \in \mathbb{Z}^+ ,$$

qui est un premier exemple de variable aléatoire. ■

■ **Exemple 6.9** On tire au hasard, selon un modèle uniforme, un point dans un carré $[0, 1] \times [0, 1]$, et on s'intéresse à la distance à l'origine :

$$R : \omega \in [0, 1] \times [0, 1] \rightarrow R(\omega) \in [0, \sqrt{2}] .$$

R est encore une fonction $\Omega \rightarrow [0, \sqrt{2}]$, et est un autre exemple de variable aléatoire. ■

En termes imagés, les variables aléatoires permettent de manipuler des “produits dérivés” des modèles probabilistes.

Définition 6.6 Étant donné un modèle probabiliste $(\Omega, \mathcal{F}, \mathbb{P})$, on appelle *variable aléatoire* toute fonction mesurable

$$X : \omega \in \Omega \rightarrow X(\omega) \in \mathbb{R} \text{ (ou } \mathbb{Z} \text{) ,}$$

c'est à dire telle que pour tout intervalle $I \subset \mathbb{R}$, son image inverse $X^{-1}(I)$ par X appartienne à \mathcal{F} .

Soit maintenant un sous ensemble Borélien $A \subset \mathbb{R}$ (c'est à dire une partie de \mathbb{R} à laquelle on sait associer une mesure). On peut alors s'intéresser à la quantité

$$\mathbb{P}_X\{A\} = \mathbb{P}\{\omega \in \Omega : X(\omega) \in A\} , \quad (6.16)$$

qui est bien définie.

Définition 6.7 La collection des nombres $\mathbb{P}_X\{A\}$, où $A \subset \mathbb{R}$ est un sous-ensemble Borélien de \mathbb{R} , est appelée *distribution de probabilités* de la variable aléatoire X .

Dans la suite, on utilisera indistinctement les notations

$$\mathbb{P}_X\{A\} = \mathbb{P}\{\omega \in \Omega : X(\omega) \in A\} = \mathbb{P}\{X(\omega) \in A\} = \mathbb{P}\{X \in A\} .$$

Remarque 6.9 Il est utile de faire ici un certain nombre de remarques.

1. En pratique, on n'a pas à se donner les nombres $\mathbb{P}_X\{A\}$ pour tout $A \subset \mathbb{R}$. On se contente d'une règle permettant de calculer ces nombres. On en verra de nombreux exemples dans ce qui suit.
2. Dans le cas où $\Omega \subset \mathbb{R}$, la variable aléatoire la plus simple correspond à la fonction identité : $X(\omega) = \omega$.
3. Souvent, il n'est pas nécessaire de connaître les détails du modèle probabiliste sous-jacent pour caractériser la distribution d'une variable aléatoire. C'est d'ailleurs préférable, dans la mesure où les ensembles Ω sont souvent gigantesques et difficiles à manipuler dans des applications pratiques.

■ **Exemple 6.10** Si nous reprenons un des exemples que nous avons déjà abordés précédemment, à savoir celui qui nous a conduits au modèle binomial, on considère l'ensemble Ω_1 de toutes les configurations possibles en N tirages. A chaque $a \in \Omega_1$ on peut alors associer le nombre $X(a)$ de fois où ω_1 est apparu. Cette fonction X définit une variable aléatoire dont la distribution est donnée par

$$\mathbb{P}_X\{n\} = \mathbb{P}\{a \in \Omega_1, X(a) = n\} = C_N^n p_1^n p_2^{N-n} ,$$

et il s'agit d'une distribution binomiale. ■

La fonction de répartition constitue un intermédiaire de calcul souvent utile (dans le cas discret et plus encore dans le cas continu).

Définition 6.8 Soit X une variable aléatoire sur $(\Omega, \mathcal{F}, \mathbb{P})$. Sa fonction de répartition est la fonction $F_X : \mathbb{R} \rightarrow [0, 1]$ définie par

$$F_X(x) = \mathbb{P}_X\{-\infty, x\} = \mathbb{P}\{\omega \in \Omega : X(\omega) < x\} . \quad (6.17)$$

La fonction de répartition possède un certain nombre de propriétés intéressantes, faciles à démontrer :

- Propriété 6.10**
1. F_X est monotone : si $x \leq y$, alors $F_X(x) \leq F_X(y)$.
 2. Par conséquent, pour tout x , $0 \leq F_X(x) \leq \lim_{y \rightarrow \infty} F_X(y) = 1$.

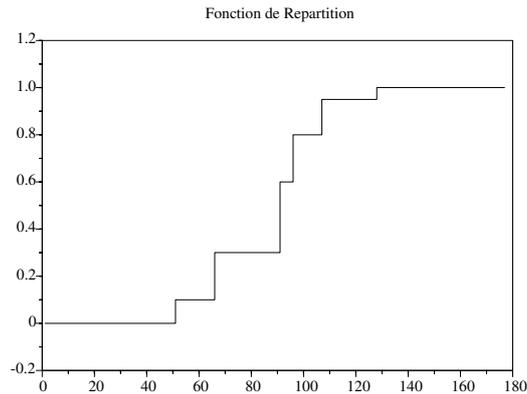


FIGURE 6.1 – Fonction de répartition d’une variable aléatoire discrète

3. La fonction de répartition est continue à gauche : pour tout x ,

$$\lim_{y \rightarrow x, y \leq x} F_X(y) = F_X(x) .$$

On distinguera trois types de variables aléatoires : les variables aléatoires discrètes (c’est à dire les fonctions X prenant un nombre fini ou infini dénombrable de valeurs), et deux catégories de variables aléatoires continues : des variables aléatoires dites “à densité”, qui sont caractérisées par une fonction appelée *densité de probabilités*, et des variables aléatoires dites “mixtes”, dont la distribution de probabilités ne peut être caractérisée par une densité.

On discutera ici les deux premiers cas (discret et continu) séparément, évitant les variables aléatoires mixtes, dont le traitement est plus complexe. Le cas discret permet d’introduire des notions dont la plupart se généralisent facilement au cas continu.

6.2.2 Variables aléatoires discrètes

Commençons donc par le cas des variables aléatoires discrètes. Étant donné un modèle probabiliste $(\Omega, \mathcal{F}, \mathbb{P})$, on considère donc la fonction mesurable $X : \Omega \rightarrow \mathbb{R}$, dont on suppose qu’elle ne prend qu’un nombre fini ou infini dénombrable de valeurs $\{x_1, x_2, \dots\}$. Il est clair que la distribution de probabilités de X est complètement caractérisée par la donnée des nombres $\mathbb{P}_X\{x_i\}$: en effet, on peut facilement en déduire la probabilité pour que X prenne une des valeurs $\{x_{k_1}, x_{k_2}, \dots, x_{k_n}\}$ grâce à la règle

$$\mathbb{P}_X\{x_{k_1}, x_{k_2}, \dots, x_{k_n}\} = \mathbb{P}_X\{x_{k_1}\} + \mathbb{P}_X\{x_{k_2}\} + \dots + \mathbb{P}_X\{x_{k_n}\} .$$

Fonction de répartition

Si X est une variable aléatoire discrète à valeurs dans $\{x_1, x_2, \dots\}$, on a alors

$$F_X(x) = \sum_{k: x_k < x} \mathbb{P}_X\{x_k\} . \quad (6.18)$$

En fait, quand X est une variable aléatoire discrète, sa fonction de répartition est une fonction constante par morceaux, comme on peut le voir dans l’exemple donné en FIG. 6.1.

Fonction d’une variable aléatoire : transfert de loi

Étant donnée la distribution d’une variable aléatoire X , et une fonction $\varphi : \mathbb{R} \rightarrow \mathbb{R}$, on peut en déduire la distribution de probabilités d’une nouvelle variable aléatoire (on parle alors de transfert de loi)

$$Y = \varphi(X) : \omega \in \Omega \rightarrow Y(\omega) = \varphi(X(\omega)) ,$$

à condition que φ vérifie des conditions de mesurabilité appropriées. On peut alors calculer

$$\mathbb{P}_Y\{y\} = \mathbb{P}\{\omega : \varphi(X(\omega)) = y\} = \mathbb{P}\{\omega : X(\omega) \in \varphi^{-1}(\{y\})\} = \mathbb{P}_X\{\varphi^{-1}(\{y\})\} , \quad (6.19)$$

où $\varphi^{-1}(\{y\})$ est l'image inverse du singleton $\{y\}$ par la fonction φ . Si φ est bijective, les choses sont plus simples car $\varphi^{-1}(\{y\})$ est le singleton $\varphi^{-1}(y)$, et on peut écrire

$$\mathbb{P}_Y\{y\} = \mathbb{P}\{\omega : X(\omega) = \varphi^{-1}(y)\} = \mathbb{P}_X\{\varphi^{-1}(y)\}.$$

■ **Exemple 6.11** Prenons l'exemple d'une variable aléatoire X à valeurs entières : $X(\omega) \in \mathbb{Z}$ pour tout $\omega \in \Omega$, et la fonction $\varphi(x) = x^2$. Alors pour $y \neq 0$ on a

$$\mathbb{P}_Y\{y\} = \mathbb{P}\{\omega : X(\omega) = \pm\sqrt{y}\} = \mathbb{P}_X\{\sqrt{y}\} + \mathbb{P}_X\{-\sqrt{y}\},$$

et $\mathbb{P}_Y\{0\} = \mathbb{P}_X\{0\}$.

Prenons le cas particulier d'une variable aléatoire X Poissonnienne : $\mathbb{P}_X\{n\} = e^{-\lambda} \lambda^n / n!$. Comme X ne prend que des valeurs positives, φ est une bijection de \mathbb{Z}^+ sur $\varphi(\mathbb{Z}^+)$, et on peut écrire

$$\mathbb{P}_Y\{y\} = \begin{cases} e^{-\lambda} \lambda^{\sqrt{y}} / (\sqrt{y}!) & \text{si } \sqrt{y} \in \mathbb{Z}^+ \\ 0 & \text{sinon.} \end{cases}$$

■

Espérance mathématique

L'espérance mathématique est la notion probabiliste qui permet de décrire la valeur moyenne d'une variable aléatoire. La définition est donnée par

■ **Définition 6.9** Soit X une variable aléatoire discrète sur $(\Omega, \mathcal{F}, \mathbb{P})$. Son espérance mathématique est donnée par

$$\mathbb{E}\{X\} = \sum_k x_k \mathbb{P}_X\{x_k\}, \quad (6.20)$$

où les x_k sont les valeurs prises par X .

L'espérance mathématique est une application linéaire : si X, Y sont deux variables aléatoires sur le même modèle probabiliste, alors $\mathbb{E}\{X + Y\} = \mathbb{E}\{X\} + \mathbb{E}\{Y\}$. De même, si $\lambda \in \mathbb{R}$, on a $\mathbb{E}\{\lambda X\} = \lambda \mathbb{E}\{X\}$.

Une autre propriété importante est la propriété suivante :

$$\mathbb{E}\{|X|\} = 0 \text{ implique } X(\omega) = 0 \text{ presque sûrement.} \quad (6.21)$$

La signification précise de cette dernière assertion est la suivante : si $\mathbb{E}\{|X|\} = 0$, alors $\mathbb{P}\{X(\omega) = 0\} = 1$.

■ **Exemple 6.12** Prenons l'exemple de la roulette : 37 résultats possibles ($\Omega = \{0, d \dots 36\}$). Si on mise une somme x_0 sur l'une des 37 cases, le gain est de $36x_0$ en cas de tirage favorable, et de 0 sinon. On peut donc modéliser le gain comme une variable aléatoire X qui vaut $36x_0$ avec probabilité $p = 1/37$, et 0 avec probabilité $1 - p = 36/37$. L'espérance de X vaut alors

$$\mathbb{E}\{X\} = \frac{36}{37} x_0.$$

On peut également miser sur rouge ou noir, ou sur pair ou impair. Dans ce cas (comme le 0 est considéré comme "ni pair ni impair", et est associé à la couleur verte), la probabilité de gagner est $p = 18/37$. Si on mise x_0 , le gain est une variable aléatoire X qui vaut $2x_0$ en cas de tirage gagnant, et 0 sinon. L'espérance du gain vaut de nouveau $\mathbb{E}\{X\} = 36x_0/37$. ■

■ **Remarque 6.11** Il existe des variables aléatoires pour lesquelles l'espérance mathématique n'est pas définie, ou plutôt telles que la somme dans l'équation (6.20) soit divergente. C'est le cas par exemple de la variable aléatoire X à valeurs dans \mathbb{Z}^+ définie par $\mathbb{P}_X\{k\} = 6/(k\pi)^2$, $k = 1, 2, \dots$. On a bien $\sum_1^\infty \mathbb{P}_X\{k\} = 1$, mais la série $\sum_1^\infty k \mathbb{P}_X\{k\}$ est divergente.

Espérance d'une fonction d'une variable aléatoire, moments, variance

Soit X une variable aléatoire discrète prenant les valeurs x_1, x_2, \dots avec probabilités p_1, p_2, \dots , et considérons la variable aléatoire $Y = X^2$. Si les x_i sont tous positifs, les nombres x_i^2 sont tous différents, et Y prendra donc les valeurs x_1^2, x_2^2, \dots avec probabilités p_1, p_2, \dots , et on pourra écrire

$$\mathbb{E}\{Y\} = \sum_i x_i^2 p_i.$$

Supposons maintenant que les x_i^2 ne soient pas tous différents, par exemple que $x_1^2 = x_2^2$, et que tous les autres x_j^2 soient deux à deux différents. Alors $\mathbb{P}_X\{x_i^2\} = p_1 + p_2$, et $\mathbb{E}\{Y\} = (p_1 + p_2)x_1^2 + \sum_{i \notin \{1,2\}} x_i^2 p_i$, qui coïncide avec l'équation précédente. On vérifie de même que cette expression reste inchangée dans les cas où plus de deux valeurs de x_i^2 coïncident. Ceci suggère la définition suivante :

Définition 6.10 Soient X et $Y = \varphi(X)$ deux variables aléatoires discrètes. On note x_1, x_2, \dots les valeurs prises par X . L'espérance de $Y = \varphi(X)$ est donnée par

$$\mathbb{E}\{Y\} = \sum_i \varphi(x_i) \mathbb{P}_X\{x_i\} . \quad (6.22)$$

On définit ainsi (lorsque les séries correspondantes sont convergentes) les moments de X

$$\mathbb{E}\{X^n\} = \sum_i x_i^n \mathbb{P}_X\{x_i\} , \quad (6.23)$$

la variance de X :

$$\text{Var}\{X\} = \mathbb{E}\{(X - \mathbb{E}\{X\})^2\} , \quad (6.24)$$

et l'écart type de X

$$\sigma_X = \sqrt{\text{Var}\{X\}} .$$

On peut en particulier remarquer que

$$\text{Var}\{X\} = \mathbb{E}\{X^2\} - \mathbb{E}\{X\}^2 .$$

La variance est une opération non-linéaire : $\text{Var}\{X + Y\} \neq \text{Var}\{X\} + \text{Var}\{Y\}$ en général (lorsque l'égalité est vérifiée, on dit que X et Y sont décorréélées). De plus, si $\lambda \in \mathbb{R}$, on a $\text{Var}\{\lambda X\} = \lambda^2 \text{Var}\{X\}$.

■ **Exemple 6.13** Considérons une variable aléatoire Poissonnienne X , de paramètre $\lambda > 0$. Les valeurs prises par X sont donc $0, 1, 2, \dots$, avec des probabilités données par une loi de Poisson :

$$\mathbb{P}_X\{k\} = \frac{\lambda^k}{k!} e^{-\lambda} .$$

Un calcul direct donne

$$\mathbb{E}\{X\} = \sum_0^\infty k \frac{\lambda^k}{k!} e^{-\lambda} = \lambda e^{-\lambda} \sum_1^\infty \frac{\lambda^{k-1}}{(k-1)!} = \lambda .$$

De même,

$$\mathbb{E}\{X^2\} = \mathbb{E}\{X\} + \sum_0^\infty k(k-1) \frac{\lambda^k}{k!} e^{-\lambda} = \lambda + \lambda^2 e^{-\lambda} \sum_2^\infty \frac{\lambda^{k-2}}{(k-2)!} = \lambda + \lambda^2 ,$$

de sorte que

$$\text{Var}\{X\} = \lambda .$$

■

Remarque 6.12 On peut construire de même l'espérance de fonctions $\varphi(X)$ de X dans le cas où φ est à valeurs complexes, comme on le verra en étudiant les fonctions caractéristiques.

La fonction caractéristique

La fonction caractéristique (parfois appelée fonction génératrice) est un intermédiaire de calcul souvent très utile lorsqu'il s'agit de calculer les moments d'une variable aléatoire.

Définition 6.11 Soit X une variable aléatoire discrète. La fonction caractéristique de X est la fonction $G_X : \mathbb{R} \rightarrow \mathbb{C}$ définie par

$$G_X(u) = \mathbb{E}\{e^{-iuX}\} = \sum_k e^{-iu x_k} \mathbb{P}_X\{x_k\} . \quad (6.25)$$

Il est facile de voir que cette série est absolument convergente, puisque

$$|G_X(u)| \leq \sum_k |e^{-iux_k} \mathbb{P}_X\{x_k\}| = \sum_k \mathbb{P}_X\{x_k\} = 1.$$

Remarquons que cette définition est réminiscente de la transformation de Fourier. Cette analogie sera plus claire dans le cas des variables aléatoires à densité.

La fonction caractéristique permet de calculer facilement les moments de X (lorsqu'ils existent) :

Proposition 6.13 Soit G_X la fonction caractéristique de la variable aléatoire X . Alors

$$G_X(0) = 1, \quad (6.26)$$

et plus généralement, pour tout $n \in \mathbb{Z}^+$ tel que $\mathbb{E}\{X^n\}$ existe

$$\left[\left(i \frac{d}{du} \right)^n G_X \right] (0) = \mathbb{E}\{X^n\}. \quad (6.27)$$

• **Exemple 6.13** (suite). Reprenons l'exemple de la variable aléatoire Poissonnienne. La fonction caractéristique vaut

$$G_X(u) = \sum_0^\infty e^{-iuk} \frac{\lambda^k}{k!} e^{-\lambda} = \sum_0^\infty \frac{(\lambda e^{-iu})^k}{k!} e^{-\lambda} = e^{-\lambda} e^{\lambda e^{-iu}}.$$

Comme attendu, $G_X(0) = 1$. En dérivant une première fois,

$$iG_X'(u) = ie^{-\lambda} (-i)\lambda e^{-iu} e^{\lambda e^{-iu}} = \lambda e^{-\lambda} e^{-iu} e^{\lambda e^{-iu}},$$

de sorte que $iG_X'(0) = \lambda$. En dérivant une seconde fois,

$$-G_X''(u) = i\lambda e^{-\lambda} \left(-ie^{-iu} e^{\lambda e^{-iu}} - i\lambda e^{-2iu} e^{\lambda e^{-iu}} \right) = (\lambda e^{-iu} + \lambda^2 e^{-2iu}) e^{-\lambda} e^{\lambda e^{-iu}},$$

ce qui redonne bien $-G_X''(0) = \lambda + \lambda^2$, et donc $\sigma_X = \lambda$.

6.2.3 Variables aléatoires à densité

On s'intéresse maintenant aux variables aléatoires prenant leurs valeurs dans un ensemble infini continu, et on s'intéresse aux variables aléatoires dites "à densité".

Définition 6.12 Une variable aléatoire X sur $(\Omega, \mathcal{F}, \mathbb{P})$ est dite à densité si il existe une fonction $\rho_X \in L^1(\mathbb{R})$ telle que pour tout ensemble mesurable $U \subset \mathbb{R}$,

$$\mathbb{P}_X\{U\} = \int_U \rho_X(x) dx. \quad (6.28)$$

Propriétés de la densité

Une densité de probabilités ρ_X est à valeurs réelles. Cependant, $\rho_X(x) \geq 0$ presque partout (par contre, rien ne contraint la densité de probabilités à prendre des valeurs inférieures à 1). De plus, on a

$$\int_{-\infty}^{\infty} \rho_X(x) dx = \mathbb{P}_X\{\mathbb{R}\} = 1. \quad (6.29)$$

Fonction de répartition

La fonction de répartition d'une variable aléatoire à densité est définie comme dans le cas discret (voir DÉFINITION 6.8), et les propriétés énoncées dans la section PROPRIÉTÉS 6.10 restent valables. La nouveauté est que la fonction de répartition est maintenant donnée par

$$F_X(x) = \mathbb{P}_X\{]-\infty, x]\} = \int_{-\infty}^x \rho_X(x) dx,$$

et est continue. De plus, on a

$$\rho_X(x) = \frac{d}{dx} F_X(x). \quad (6.30)$$

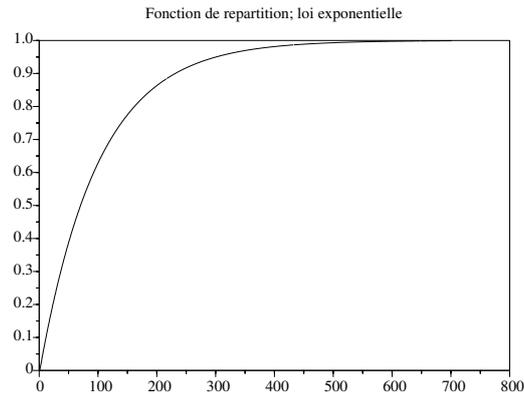


FIGURE 6.2 – Fonction de répartition d’une variable aléatoire exponentielle

■ **Exemple 6.14** Soit X une variable aléatoire exponentielle, de paramètre α : pour tout $x > 0$,

$$\mathbb{P}\{X(\omega) \geq x\} = e^{-\alpha x}.$$

On a donc la fonction de répartition

$$F_X(x) = (1 - e^{-\alpha x})\Theta(x),$$

et la densité

$$\rho_X(x) = F'_X(x) = \alpha e^{-\alpha x}\Theta(x).$$

La fonction de répartition (définie sur \mathbb{R}^+) est montrée en FIGURE 6.2. ■

Densité d’une fonction de variable aléatoire

Étant données deux variables aléatoires à densité X et $Y = \varphi(X)$ sur $(\Omega, \mathcal{F}, \mathbb{P})$, on cherche à relier la densité de Y à la densité ρ_X de X . Dans le cas général, il faut procéder en deux étapes : dans un premier temps, calculer la fonction de répartition de Y en fonction de la densité ρ_X de X ; puis en déduire la densité par dérivation.

Prenons l’exemple d’une variable aléatoire à valeurs réelles (positives ou négatives) X , de densité ρ_X , et de $Y = X^2$. La fonction de répartition de Y est donnée par

$$F_Y(y) = \mathbb{P}\{X(\omega)^2 < y\} = \mathbb{P}\{X(\omega) < \sqrt{y}\}\Theta(y) = \Theta(y) \int_{-\sqrt{y}}^{\sqrt{y}} \rho_X(x) dx.$$

On en déduit la densité de Y par dérivation :

$$\rho_Y(y) = \Theta(y) \left(\frac{1}{2\sqrt{y}} \rho_X(\sqrt{y}) - \frac{-1}{2\sqrt{y}} \rho_X(-\sqrt{y}) \right) = \Theta(y) \frac{1}{2\sqrt{y}} (\rho_X(\sqrt{y}) + \rho_X(-\sqrt{y})).$$

On voit bien sur l’exemple ci-dessus la difficulté introduite par le fait que le φ choisi ne soit pas bijectif. Lorsque φ est bijectif, les choses sont un peu plus simples, et on peut énoncer le résultat suivant :

Théorème 6.14 Si φ est une bijection différentiable, à inverse différentiable, la densité ρ_Y de $Y = \varphi(X)$ est donnée par

$$\rho_Y(y) = |(\varphi^{-1})'(y)| \rho_X(\varphi^{-1}(y)). \quad (6.31)$$

Preuve : Les hypothèses faites sont précisément celles qui nous permettent de faire un changement de variables. Supposons pour simplifier que φ soit croissante, alors φ^{-1} est croissante aussi, et

$$\rho_Y(y) = \frac{d}{dy} \int_{\varphi^{-1}(-\infty, y[} \rho_X(x) dx = \frac{d}{dy} \int_{-\infty}^y \rho_X(\varphi^{-1}(u)) (\varphi^{-1})'(u) du,$$

ce qui prouve le résultat. Si φ est décroissante le calcul est le même à un changement de signe près. ♠

Espérance, moments,...

Les définitions des espérances mathématiques des variables aléatoires à densité sont similaires à celles vues dans le cas discret.

Définition 6.13 1. L'espérance mathématique d'une variable aléatoire continue X de densité ρ_X est définie par

$$\mathbb{E}\{X\} = \int_{-\infty}^{\infty} x\rho_X(x) dx . \quad (6.32)$$

2. Etant données deux variables aléatoires continues X et $Y = \varphi(X)$, l'espérance mathématique de Y est définie par

$$\mathbb{E}\{\varphi(X)\} = \int_{-\infty}^{\infty} \varphi(x)\rho_X(x) dx . \quad (6.33)$$

En utilisant la notation introduite plus haut, on peut noter l'espérance d'une fonction de variable aléatoire quelconque (continue ou discrète)

$$\mathbb{E}\{\varphi(X)\} = \int_{-\infty}^{\infty} \varphi(x) d\mathbb{P}_X(x) . \quad (6.34)$$

On introduit par exemple les moments

$$\mathbb{E}\{X^n\} = \int_{-\infty}^{\infty} x^n \rho_X(x) dx ,$$

et la variance

$$\text{Var}\{X\} = \int_{-\infty}^{\infty} (x - \mathbb{E}\{X\})^2 \rho_X(x) dx .$$

La fonction caractéristique est introduite de façon similaire :

Définition 6.14 Soit X une variable aléatoire continue de densité ρ_X . La fonction caractéristique de X est la fonction $G_X : \mathbb{R} \rightarrow \mathbb{C}$ définie par

$$G_X(u) = \mathbb{E}\{e^{-iuX}\} = \int_{-\infty}^{\infty} e^{-iux} \rho_X(x) dx . \quad (6.35)$$

En utilisant le symbole $d\mathbb{P}_X$, on notera dans le cas général

$$G_X(u) = \int_{-\infty}^{\infty} e^{-iux} d\mathbb{P}_X(x) .$$

Dans le cas d'une variable aléatoire à densité, la fonction caractéristique est proportionnelle à la transformée de Fourier de la densité :

$$G_X(u) = \sqrt{2\pi} \hat{\rho}_X(u) .$$

On peut également en déduire (si la formule d'inversion de Fourier a un sens dans ce contexte) une expression de la densité à partir de la fonction caractéristique :

$$\rho_X(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} G_X(u) e^{iux} du .$$

Il est facile de montrer que la PROPOSITION 6.13 reste valide dans le cas des variables aléatoires à densité (la démonstration est similaire à celle donnant la dérivée d'une transformée de Fourier).

■ **Exemple 6.15** Considérons une variable aléatoire exponentielle X , de paramètre α . La densité est donc

$$\rho_X(x) = \alpha e^{-\alpha x} \Theta(x) ,$$

et la fonction caractéristique est donnée par

$$G_X(u) = \alpha \int_0^{\infty} e^{-\alpha x} e^{-iux} dx = \frac{\alpha}{\alpha + iu} .$$

On vérifie facilement que $G_X(0) = 1$. En dérivant une première fois, on a

$$iG'_X(u) = \frac{\alpha}{(\alpha + iu)^2},$$

d'où

$$\mathbb{E}\{X\} = \frac{1}{\alpha}.$$

De même,

$$-G''_X(u) = \frac{2\alpha}{(\alpha + iu)^3}$$

donne

$$\mathbb{E}\{X^2\} = \frac{2}{\alpha^2},$$

et donc $\text{Var}\{X\} = \alpha^{-2}$. ■

6.2.4 Vecteurs aléatoires

Paires de variables aléatoires

On s'intéresse maintenant à des familles de variables aléatoires. Considérons pour commencer une paire de variables aléatoires X, Y sur un même modèle probabiliste $(\Omega, \mathcal{F}, \mathbb{P})$. Les distributions de probabilités \mathbb{P}_X et \mathbb{P}_Y donnent des indications sur le comportement de X et Y , mais pas sur les relations existant entre X et Y . C'est pourquoi il est important de s'intéresser au comportement de la paire (X, Y) .

Définition 6.15 Soient X, Y deux variables aléatoires sur un même modèle probabiliste $(\Omega, \mathcal{F}, \mathbb{P})$. La distribution de probabilités jointe de X, Y est l'ensemble des nombres

$$\mathbb{P}_{X,Y}\{U\} = \mathbb{P}\{\omega \in \Omega : (X(\omega), Y(\omega)) \in U\}, \quad (6.36)$$

où $U \subset \mathbb{R}^2$ est un ensemble mesurable quelconque. Les distributions de probabilités \mathbb{P}_X et \mathbb{P}_Y sont appelées distributions marginales de $\mathbb{P}_{X,Y}$.

Il est facile d'obtenir les distributions marginales à partir de la distribution jointe. Supposons que X et Y soient des variables aléatoires discrètes, à valeurs dans $\{x_1, x_2, \dots\}$ et $\{y_1, y_2, \dots\}$ respectivement. Soit V un sous-ensemble mesurable de \mathbb{R} . On peut écrire

$$\begin{aligned} \mathbb{P}_X\{V\} &= \sum_{x_i \in V} \mathbb{P}_X\{x_i\} \\ &= \sum_{x_i \in V} \mathbb{P}\{X(\omega) = x_i, Y(\omega) \in \{y_1, y_2, \dots\}\} \\ &= \sum_{x_i \in V} \sum_j \mathbb{P}\{X(\omega) = x_i, Y(\omega) = y_j\} \\ &= \sum_j \mathbb{P}_{X,Y}\{V \times \{y_j\}\} \end{aligned}$$

Notons que la réciproque est fautive : la connaissance des distributions marginales ne permet pas de retrouver la distribution jointe en général. Il suffit pour s'en convaincre de considérer le cas extrême $Y = X$. La connaissance des distributions marginales \mathbb{P}_X et \mathbb{P}_Y (au demeurant identiques) ne permet pas de savoir que $Y = X$.

■ **Exemple 6.16** Soient deux variables aléatoires X, Y , prenant des valeurs entières $0, 1, \dots, N$, de distribution jointe multinômiale

$$\mathbb{P}_{X,Y}\{n, m\} = \frac{N!}{n!m!(N-n-m)!} p_1^n p_2^m (1-p_1-p_2)^{N-n-m},$$

où $p_1, p_2 \in [0, 1]$ sont tels que $0 \leq p_1 + p_2 \leq 1$.

On obtient de là la distribution de X :

$$\begin{aligned}\mathbb{P}_X\{n\} &= \sum_{m=0}^{N-n} \frac{N!}{n!m!(N-n-m)!} p_1^n p_2^m (1-p_1-p_2)^{N-n-m}, \\ &= \frac{N!}{n!(N-n)!} p_1^n \sum_{m=0}^{N-n} \frac{(N-n)!}{m!(N-n-m)!} p_2^m (1-p_1-p_2)^{(N-n)-m}, \\ &= \frac{N!}{n!(N-n)!} p_1^n (1-p_1)^{N-n}.\end{aligned}$$

Il s'agit là encore d'une distribution binômiale. ■

Espérance de fonction de deux variables aléatoires

Étant données deux variables aléatoires X, Y , et une fonction mesurable $\varphi : \mathbb{R}^2 \rightarrow \mathbb{R}$, on considère la variable aléatoire $Z = \varphi(X, Y)$, définie par : pour tout $\omega \in \Omega$,

$$Z(\omega) = \varphi(X(\omega), Y(\omega)).$$

Pour calculer l'espérance de Z , une première démarche consiste à calculer tout d'abord la distribution de probabilités de Z , puis de calculer l'espérance grâce à l'expression classique : dans le cas discret :

$$\mathbb{E}\{Z\} = \sum_i z_i \mathbb{P}_Z\{z_i\},$$

où les z_i sont les valeurs possibles de Z . On peut en fait montrer que l'espérance de Z peut également être obtenue de la façon suivante :

Définition 6.16 Soient $X, Y, Z = \varphi(X, Y)$ trois variables aléatoires discrètes sur $(\Omega, \mathcal{F}, \mathbb{P})$. On note $\{x_1, x_2, \dots\}$ et $\{y_1, y_2, \dots\}$ les valeurs possibles de X et Y respectivement. L'espérance mathématique de Z est donnée par

$$\mathbb{E}\{Z\} = \sum_{i,j} \varphi(x_i, y_j) \mathbb{P}_{X,Y}\{x_i, y_j\}. \quad (6.37)$$

Ces notions nous sont utiles pour introduire la covariance et la corrélation :

Définition 6.17 Soient X, Y deux variables aléatoires sur $(\Omega, \mathcal{F}, \mathbb{P})$.

1. La covariance de X et Y est le nombre

$$\gamma_{X,Y} = \mathbb{E}\{(X - \mathbb{E}\{X\})(Y - \mathbb{E}\{Y\})\}. \quad (6.38)$$

2. X et Y sont décorrélées si $\gamma_{X,Y} = 0$.

3. Le coefficient de corrélation de X et Y est donné par

$$r(X, Y) = \frac{\gamma_{X,Y}}{\sigma_X \sigma_Y}. \quad (6.39)$$

Remarquons que

$$\gamma_{X,X} = \text{Var}\{X\},$$

et que

$$\text{Var}\{X + Y\} = \text{Var}\{X\} + \text{Var}\{Y\} + 2\gamma_{X,Y}.$$

La covariance donne des indications sur une possible "association" systématique des variables aléatoires X et Y . Le coefficient de corrélation (qui est toujours un nombre sans dimension) fournit une version "normalisée" de la covariance. Il a les propriétés suivantes.

Proposition 6.15 Soient X et Y deux variables aléatoires sur $(\Omega, \mathcal{F}, \mathbb{P})$.

1. $-1 \leq r(X, Y) \leq 1$.
2. Si $r(X, Y) = 1$, alors il existe une relation du type $Y = aX + b$, où $a, b \in \mathbb{R}$ et $a > 0$.
3. Si $r(X, Y) = -1$, alors il existe une relation du type $Y = aX + b$, où $a, b \in \mathbb{R}$ et $a < 0$.

Vecteurs aléatoires

On considère maintenant des familles de plusieurs variables aléatoires, que l'on nomme vecteurs aléatoires.

Définition 6.18 Soient X_1, X_2, \dots, X_n n variables aléatoires sur un même modèle probabiliste $(\Omega, \mathcal{F}, \mathbb{P})$. La distribution de probabilités jointe de X_1, \dots, X_n est l'ensemble des nombres

$$\mathbb{P}_{X_1, \dots, X_n}\{U\} = \mathbb{P}\{(X_1(\omega), \dots, X_n(\omega)) \in U\}, \quad (6.40)$$

où $U \subset \mathbb{R}^n$ est un ensemble mesurable quelconque. Les distributions de probabilités \mathbb{P}_{X_i} sont appelées distributions marginales.

Comme dans le cas des paires de variables aléatoires, il est facile d'obtenir les distributions marginales à partir de la distribution jointe. La réciproque est toujours fautive : la connaissance des distributions marginales ne permet pas de retrouver la distribution jointe en général.

La distribution jointe permet également de calculer l'espérance de fonctions du vecteur aléatoire considéré. Si $Z = \varphi(X_1, \dots, X_n)$ est fonction du vecteur aléatoire (X_1, \dots, X_n) , on a alors

$$\mathbb{E}\{Z\} = \sum_{x_1, \dots, x_n} \varphi(x_1, \dots, x_n) \mathbb{P}_{X_1, \dots, X_n}\{x_1, \dots, x_n\}. \quad (6.41)$$

6.2.5 Loi normale univariée et multivariée

Définition 6.19 — Variable aléatoire normale. Une variable aléatoire normale, de moyenne μ et variance σ^2 est définie par la densité

$$\rho_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma^2}(x-\mu)^2\right]. \quad (6.42)$$

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

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

Par contre, sa fonction caractéristique admet une expression explicite

$$\phi_X(u) = \exp\left(i\mu u - \frac{\sigma^2 u^2}{2}\right), \quad u \in \mathbb{R}. \quad (6.43)$$

Définition 6.20 — 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 fautive, la normalité des coordonnées n'implique pas la gaussianité du vecteur.

■ **Exemple 6.17** 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 . ■

On a toutefois la propriété suivante

Proposition 6.16 Soit $X = (X_1, \dots, X_L)^T$ un vecteur aléatoire gaussien de matrice de covariance Σ_X . Les variables aléatoires X_1, \dots, X_L sont indépendantes si et seulement si la matrice Σ_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

$$\phi_X(u) = \exp\left(i\mu^T u - \frac{1}{2}u^T \Sigma u\right), \quad u \in \mathbb{R}^L. \quad (6.44)$$

■ **Exemple 6.18 — Un vecteur gaussien 2D à densité.** On considère un vecteur gaussien X de dimension 2, de moyenne μ et de matrice de covariance Σ données par

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

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

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

de sorte que la densité vaut

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

Les ensembles d'équiprobabilité sont les lignes de niveau de la densité, qui sont aussi les lignes de niveau de la fonction quadratique $(x_1, x_2) \rightarrow (x_1 - 1)^2 - 2(x_1 - 1)(x_2 - 2) + (x_2 - 2)^2$. 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 Σ . Les lignes de niveau de la densité sont représentées en FIGURE 6.3.

Quant à la fonction caractéristique, elle est donnée par

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

■

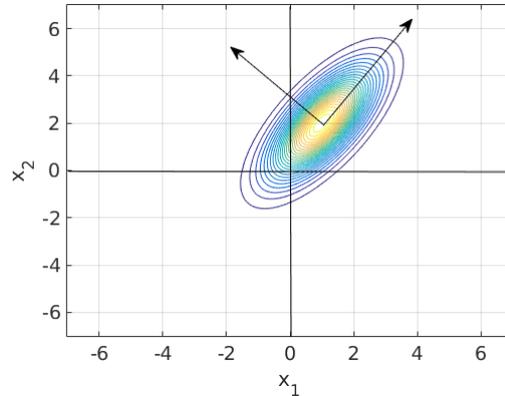


FIGURE 6.3 – Lignes de niveau de la densité ρ_X de l'exemple 6.18. Elles prennent la forme d'ellipsoïdes centrées sur la moyenne, dont les axes sont dirigés par les deux vecteurs propres de la matrice de covariance (représentés par des flèches).

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.

■ **Exemple 6.19 — Un vecteur gaussien 2D sans densité.** Soit maintenant X un vecteur gaussien de dimension 2, de moyenne μ et de matrice de covariance Σ données par

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

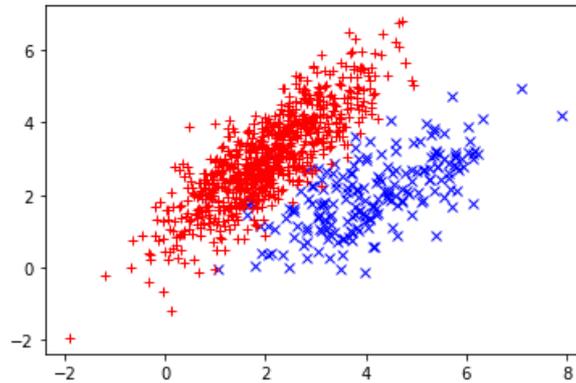


FIGURE 6.4 – Échantillon d'un mélange de lois normales bivariées. Les échantillons correspondant aux deux lois ont été colorés de deux couleurs différentes.

On voit facilement que $\det(\Sigma) = 0$, donc Σ n'est pas inversible, et X n'admet donc pas de densité.

Sa fonction caractéristique Φ_X prend la forme

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

En particulier, le noyau de Σ 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, ϕ_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 6.17 Soit $X = (X_1, \dots, X_L)^T$ un vecteur aléatoire gaussien d'espérance $\mu \in \mathbb{R}^L$ et de matrice de covariance Σ . Si Σ est définie positive (c'est à dire de déterminant $\det(\Sigma) > 0$), alors X admet pour densité la fonction ρ_X définie par

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

■ **Exemple 6.20 — Loi de mélange de deux lois normales multivariées.** Soit X un vecteur aléatoire de dimension N , défini par la loi suivante.

On se donne deux vecteurs $\mu_0, \mu_1 \in \mathbb{R}^N$, et deux matrices symétriques définies positives $\Sigma_0, \Sigma_1 \in \mathcal{M}_N(\mathbb{R})$:

- Avec probabilité p_0 , X suit une loi $\mathcal{N}(\mu_0, \Sigma_0)$.
- Avec probabilité $p_1 = 1 - p_0$, X suit une loi $\mathcal{N}(\mu_1, \Sigma_1)$.

On peut montrer que la densité de probabilités de X est de la forme

$$\rho_X(x) = p_0 \rho_0(x) + p_1 \rho_1(x)$$

où ρ_0, ρ_1 sont les densités des deux lois normales.

On représente en Figure 6.4 un échantillon d'un mélange de lois normales bivariées. On voit bien se dessiner les deux ellipses correspondant aux courbes de niveau des deux densités. ■

6.3 Statistique, estimation

6.3.1 Statistique

La théorie des probabilités est l'étude mathématique des phénomènes caractérisés par le hasard et l'incertitude, il s'agit à la base d'une discipline théorique. La statistique consiste à recueillir, traiter et

interpréter un ensemble de données, on peut la voir comme une science plus appliquée. Il existe des connexions fortes entre ces deux domaines, l'étude des procédures statistiques faisant fortement appel à la théorie des probabilités.

On distingue généralement deux approches de la statistique :

- La statistique descriptive (on parle aussi maintenant d'analyse de données), qui cherche à donner une description simplifiée d'une population de taille finie. La statistique descriptive fait appel aussi bien à des outils de visualisation qu'à des techniques mathématiques, à commencer par l'algèbre linéaire,...
- La statistique inférentielle, qui considère les données comme un échantillon de taille finie d'une population plus importante, dont on cherche à induire des informations. La statistique inférentielle fait plus fortement appel à la théorie des probabilités.

6.3.2 Estimation, exemples

Le problème de l'estimation statistique est un problème prototype de statistique inférentielle. Il peut s'énoncer comme suit. Supposons que l'on observe N valeurs $\underline{x} = (x^1, \dots, x^N)$ 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 θ , considérés comme étant déterministes. Par exemple, si la loi de X admet une densité, on note celle-ci $\rho(x; \theta)$.

Une *estimation* de θ est une fonction mesurable $T(\underline{x}) = T(x^1, \dots, x^N)$ des observations. La valeur de l'estimée dépend de la réalisation \underline{x} .

- **Exemple 6.21 — Moyenne et variance empirique.** 1. Supposons que X soit une variable aléatoire de moyenne μ inconnue. Une estimation pour μ est donné par la moyenne empirique

$$m = \frac{1}{N} \sum_{n=1}^N x^n. \quad (6.45)$$

2. Supposons que la variance σ^2 de X soit également inconnue. Une estimation pour σ^2 est donné par la variance empirique

$$s^2 = \frac{1}{N-1} \sum_{n=1}^N (x^n - m)^2, \quad (6.46)$$

qui utilise aussi la moyenne empirique définie en (6.45). ■

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

La qualité d'un estimateur se mesure souvent par deux quantités, son biais B (*bias* en anglais) et son erreur quadratique moyenne MSE (*mean square error* en anglais), définis par

$$B(T(\underline{X})) = \mathbb{E}\{T(\underline{X}) - \theta\}, \quad \text{MSE}(T(\underline{X})) = \mathbb{E}\{(T(\underline{X}) - \theta)^2\}. \quad (6.47)$$

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

$$\text{MSE}(\hat{\theta}) = \text{Var}\{\hat{\theta}\} - B(\hat{\theta})^2. \quad (6.48)$$

Exercice 6.1 Le démontrer, en utilisant la relation $\hat{\theta} - \theta = \hat{\theta} - \mathbb{E}\{\hat{\theta}\} + \mathbb{E}\{\hat{\theta}\} - \theta$ ■

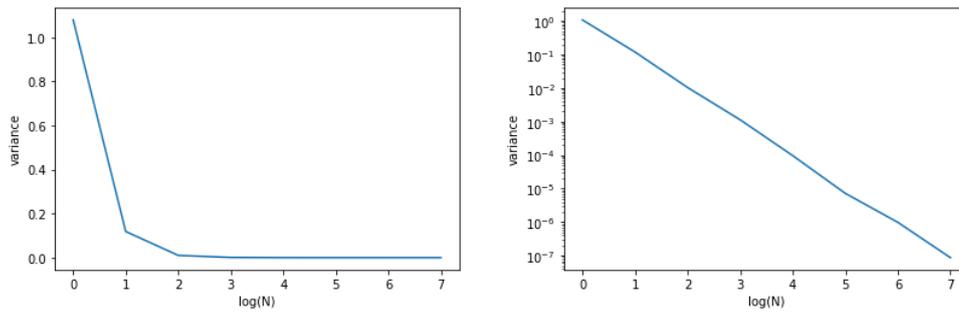


FIGURE 6.5 – Evolution de la variance de la moyenne empirique en fonction du logarithme du nombre d'échantillons. A gauche la variance, à droite son logarithme.

■ **Exemple 6.22 — Propriétés de la moyenne empirique.** On reprend les notations de l'exemple 6.21. Calculons

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

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

Concernant l'erreur quadratique moyenne, calculons

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

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

On représente en Figure 6.5 l'évolution de la variance de la moyenne empirique en fonction du nombre d'échantillons, calculée à partir d'échantillons pseudo-aléatoires d'une loi normale centrée réduite. On peut observer que le logarithme de la variance est bien proportionnel (en fait égal) au logarithme du nombre d'échantillons, ce qui est consistant avec le résultat ci-dessus. ■

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

Définition 6.21 — Consistance d'un estimateur. Un estimateur $\hat{\theta}_N$ d'un paramètre θ est consistant si $\hat{\theta}_N$ converge en probabilités vers θ lorsque la taille N de l'échantillon tend vers l'infini : pour tout $\varepsilon > 0$

$$\mathbb{P}\{|\hat{\theta}_N - \theta| \geq \varepsilon\} \rightarrow 0 \quad \text{quand } N \rightarrow \infty.$$

On a montré plus haut que

$$\lim_{N \rightarrow \infty} \text{Var}\{\hat{\mu}_N\} = 0.$$

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

$$\mathbb{P}\{|\hat{\mu}_N - \mu| \geq \varepsilon\} = \mathbb{P}\{(\hat{\mu}_N - \mu)^2 \geq \varepsilon^2\} \leq \frac{\mathbb{E}\{(\hat{\mu}_N - \mu)^2\}}{\varepsilon^2} = \frac{\text{Var}\{\hat{\mu}_N\}}{\varepsilon^2} \xrightarrow{N \rightarrow \infty} 0.$$

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

Proposition 6.18 La moyenne empirique fournit un estimateur $\hat{\mu}$ sans biais et consistant de la moyenne.

■ **Exemple 6.23 — Propriétés de la variance empirique.** Considérons maintenant la variance empirique, et calculons

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

Pour le dernier terme on a

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

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

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

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

Remarque 6.19 — Variance de la variance empirique. Avec les mêmes hypothèses, et en supposant de plus que $\mathbb{E} \{ X^4 \} < \infty$, il est possible d'obtenir une expression pour la variance de la variance empirique :

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

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

6.3.3 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 $\mu = \mathbb{E} \{ X \} \in \mathbb{R}^p$ 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

$$m = \frac{1}{N} \sum_{n=1}^N x^n, \quad S = \frac{1}{N-1} \sum_{n=1}^N (x^n - \mu)(x^n - \mu)^T. \quad (6.49)$$

En d'autres termes,

$$m_k = \frac{1}{N} \sum_{n=1}^N x_k^n, \quad S_{k\ell} = \frac{1}{N-1} \sum_{n=1}^N (x_k^n - \mu_k)(x_\ell^n - \mu_\ell).$$

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

Proposition 6.20 1. $\hat{\mu}$ est un estimateur sans biais et consistant de la moyenne μ .

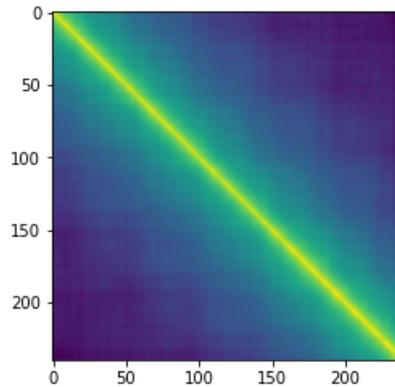


FIGURE 6.6 – Représentation de la matrice de covariance des données *non-cible* pour le *P300 speller* sous forme d’une image.

2. $\widehat{\Sigma}$ est un estimateur sans biais de la matrice de covariance.
3. La matrice de covariance empirique $\widehat{\Sigma}$ est symétrique, semi-définie positive, et de rang inférieur ou égal à N .

■ **Exemple 6.24 — 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_ℓ de x représentant la mesure à l’instant ℓ 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 6.6, 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). ■

6.3.4 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}_θ cette distribution de probabilités, pour expliciter la dépendance en θ . Supposons qu’on observe N réalisations indépendantes de X , notées x^1, \dots, x^N . On appellera vraisemblance de θ la fonction

$$\mathcal{L}(\theta) = \prod_{n=1}^N \mathbb{P}_\theta \{X = x^n\} . \quad (6.50)$$

L’idée de l’estimation par maximum de vraisemblance est de rechercher la valeur de θ 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 ρ_θ dépendant d’un paramètre θ . On appellera cette fois vraisemblance la fonction de la variable θ définie par

$$\mathcal{L}(\theta) = \prod_{n=1}^N \rho_\theta(x^n) . \quad (6.51)$$

Définition 6.22 Soient x^1, \dots, x^N , N réalisations i.i.d. d'une variable aléatoire X , dont la distribution dépend d'un paramètre θ . L'estimation de θ par maximum de vraisemblance est définie par la valeur de θ qui maximise la vraisemblance définie en (6.50) ou (6.51) selon le contexte :

$$\theta_* = \arg \max_{\theta \in \Omega} \mathcal{L}(\theta) .$$

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

$$\ell(\theta) = \ln(\mathcal{L}(\theta)) , \quad (6.52)$$

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

■ **Exemple 6.25 — Paramètre d'une loi 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 $\lambda \in \mathbb{R}^+$

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

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

Remarque 6.21 — 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 définir une vraisemblance $\Theta \in \Omega \rightarrow \mathcal{L}(\Theta)$ et une log-vraisemblance $\Theta \in \Omega \rightarrow \ell(\Theta)$, qui sont des fonctions de plusieurs variables. L'estimation par maximum de vraisemblance revient à résoudre le problème d'optimisation

$$\Theta_* = \arg \max_{\Theta \in \Omega} \mathcal{L}(\Theta) . \quad (6.53)$$

■ **Exemple 6.26 — 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 μ et variance σ^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

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

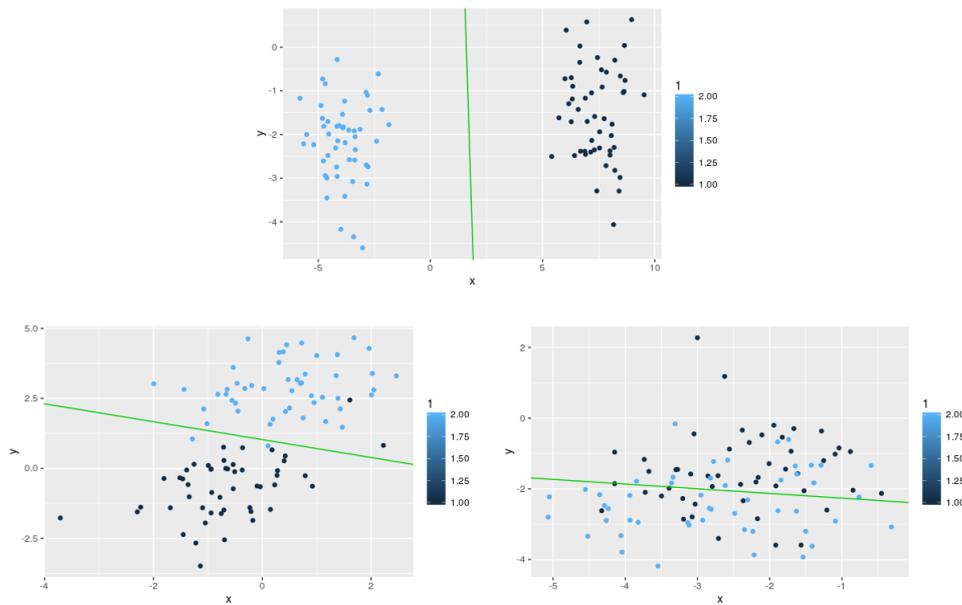


FIGURE 6.7 – Trois exemples de discrimination : on recherche une règle simple permettant de discriminer les points noirs des points bleus. La droite verte sépare l'espace en deux parties, d'après la règle fournie par l'analyse discriminante linéaire.

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

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

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

μ_* 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.

6.4 Classification, décision

6.4.1 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 6.7. Dans ce cas, les deux familles de points sont très bien séparées, et une règle 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 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 6.7 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 règle de

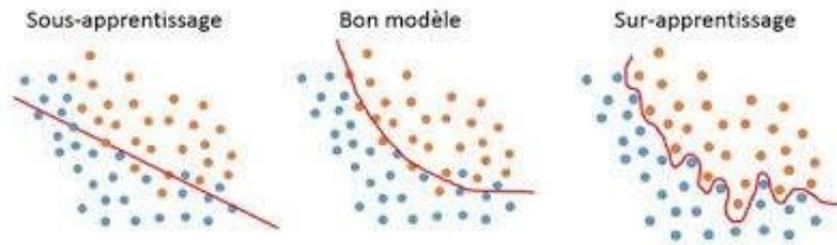


FIGURE 6.8 – Illustration du sur-apprentissage sur un problème de classification.

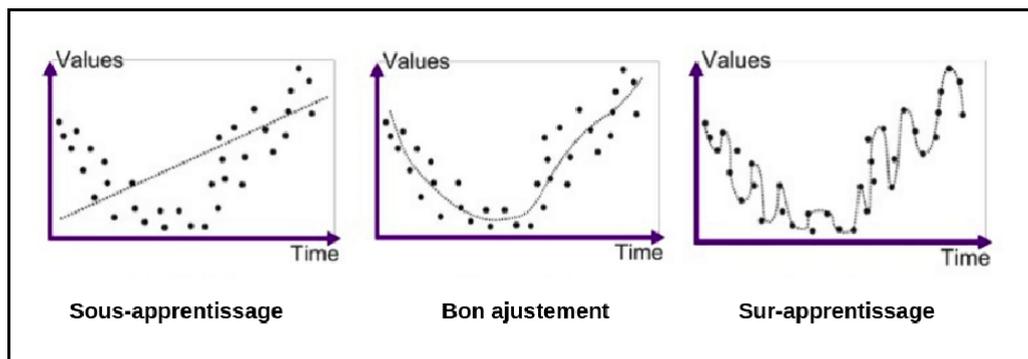


FIGURE 6.9 – Illustration du sur-apprentissage sur un problème de régression.

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. Cet aspect est important car l'objectif est de construire une règle qui sera efficace sur de nouveaux jeux de données, et pas seulement capable de reproduire le jeu de données d'apprentissage. Lorsque tel n'est pas le cas, on fait face au problème dit de *sur-apprentissage* (*overfitting* en anglais).

Le sur-apprentissage correspond à un modèle qui est trop bien adapté au jeu d'apprentissage, mais n'est pas capable de se généraliser à un autre jeu de données du même type. En d'autres termes, le modèle s'est adapté à la fois aux corrélations présentes dans les données mais aussi au bruit.

Le sur-apprentissage est illustré en FIGURE 6.8 sur un problème de classification. On cherche à répartir les données en deux classes en déterminant une courbe séparant l'espace en deux régions. La figure de gauche correspond à une séparation linéaire, qui ne décrit pas bien les données. On se tourne alors vers une régression plus complexe, la figure du centre donne un résultat assez satisfaisant, en dépit de quelques erreurs. Dans la figure de droite l'accord aux données d'apprentissage parfait, mais la séparatrice est très complexe, probablement trop, et il est vraisemblable qu'elle ne donnera pas de bons résultats sur un jeu de données de test.

Le sur-apprentissage n'est pas spécifique à la classification, il est illustré en FIGURE 6.9 sur un problème de régression. On cherche à décrire les données par une fonction décrivant l'ordonnée en fonction de l'abscisse. La figure de gauche correspond à une régression linéaire, qui à l'évidence ne reproduit pas correctement les données. On se tourne alors vers une régression plus complexe, la figure du centre donne un résultat assez satisfaisant. Tel n'est pas le cas de la figure de droite, où l'accord aux données d'apprentissage est encore meilleur, mais la fonction de régression est très complexe, probablement trop.

6.4.2 La règle de Bayes

Le problème se formule comme suit. On suppose donnés

- un ensemble de variables d'entrée (continues) : $x^1, \dots, x^N \in \mathbb{R}^L$, aussi appelées caractéristiques, prédicteurs ou variables indépendantes, et
- un ensemble correspondant de variables de sortie discrètes (variables de classe) : $C \in \mathcal{C} \triangleq \{0, 1, \dots, K-1\}$.

A partir de ces données, on cherche à construire un prédicteur, défini comme une fonction

$$f : x \in \mathbb{R}^p \rightarrow f(x) \in \mathcal{C} \quad (6.54)$$

que l'on va chercher à optimiser. Pour cela, on introduit une fonction appelée fonction de perte $\mathcal{L} : \mathcal{C} \times \mathcal{C} \rightarrow \mathbb{R}_+$, 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

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

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 appelée coût (0,1) définie par

$$\mathcal{L}(c, c') = 1 - \delta_{cc'} = \begin{cases} 0 & \text{si } c = c' \\ 1 & \text{sinon,} \end{cases} \quad (6.55)$$

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}^L . On note $p_k = \mathbb{P}\{C = k\}$ la probabilité a priori de la classe $k \in \mathcal{C}$, et $x \in \mathbb{R}^L \rightarrow \rho_k(x)$ la densité conditionnelle de x sachant que $C = k$.

Définition 6.23 — Risque. Étant donnée une règle $f : \mathbb{R}^L \rightarrow \mathcal{C}$, le risque $R(f)$ de f est l'erreur moyenne de prédiction lorsque cette règle est utilisée

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

Comme conséquence de la formule de Bayes, on introduit

Définition 6.24 — Probabilité de classe a posteriori. La probabilité a posteriori de la classe $k \in \mathcal{C}$ étant donnée une observation $x \in \mathbb{R}^L$ est donnée par

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

où ρ_X est la densité marginale de X , donnée par

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

Théorème 6.22 — Règle de Bayes. La règle de classification f^* qui minimise l'erreur moyenne de prédiction pour le coût (0,1) est la règle de Bayes : pour tout $x \in \mathbb{R}^L$,

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

Ainsi, pour classer une donnée $x \in \mathbb{R}^L$, il faut évaluer les nombres $p_k \rho_k(x)$ et déterminer la classe k qui maximise ce nombre.

Remarque 6.23 Pour effectuer une classification en utilisant la règle de Bayes, il faut que les lois de probabilités ρ_k et les probabilités p_k soient connues. Elles ne le sont pas en général, et il faut les estimer à partir d'un jeu de données d'apprentissage.

Il est toutefois très difficile d'estimer une densité de probabilités sans hypothèse supplémentaire, en particulier dans le cas multivarié. On se contente souvent de modèles paramétriques, c'est à dire de lois classiques, définies par un ou plusieurs paramètres. Ceci conduit à l'analyse discriminante, étudiée dans le chapitre qui suit.

6.5 Analyse discriminante linéaire et quadratique

On supposera dans cette section que les lois des classes sont des lois normales. Commençons par deux exemples, dans le cas unidimensionnel.

■ **Exemple 6.27** 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

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

Cet exemple est illustré dans la Figure 6.10, 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 6.28** 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

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

où la valeur critique est cette fois donnée par

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

On note que la valeur x_d de l'exemple précédent a été décalée en tenant compte des probabilités des classes. ■

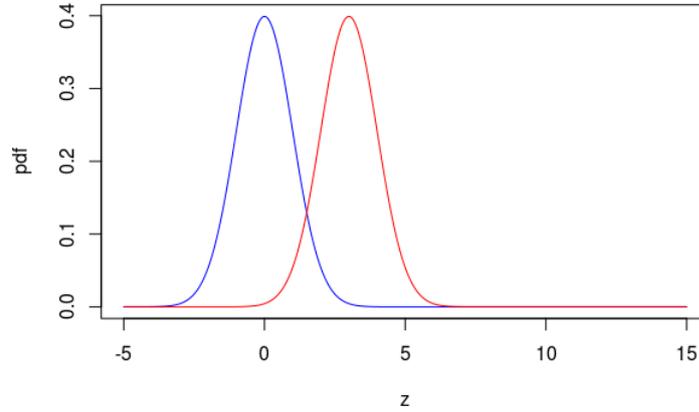


FIGURE 6.10 – Densités de deux lois gaussiennes de même variance et moyennes différentes

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.

Considérons maintenant le cas multidimensionnel et multiclassé : 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 L , distribué selon une loi normale de moyenne μ_k et covariance Σ_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

$$\rho_k(x) \triangleq \rho_{X|C=k}(x) \quad (6.56)$$

$$= \frac{1}{(2\pi)^{L/2} \sqrt{|\det(\Sigma_k)|}} \exp\left(-\frac{1}{2}(x - \mu_k)^T \Sigma_k^{-1} (x - \mu_k)\right) \quad (6.57)$$

On note cela sous la forme

$$X \sim p_0 \mathcal{N}(\mu_0, \Sigma_0) + p_1 \mathcal{N}(\mu_1, \Sigma_1) + \dots + p_{K-1} \mathcal{N}(\mu_{K-1}, \Sigma_{K-1}), \quad (6.58)$$

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

$$\Delta_{k,\ell} \triangleq \ln\left(\frac{\mathbb{P}\{C = k|X = x\}}{\mathbb{P}\{C = \ell|X = x\}}\right) \quad (6.59)$$

$$= \ln\left(\frac{p_k}{p_\ell}\right) - \frac{1}{2} \ln\left(\frac{\det(\Sigma_k)}{\det(\Sigma_\ell)}\right) - \frac{1}{2}(x - \mu_k)^T \Sigma_k^{-1} (x - \mu_k) + \frac{1}{2}(x - \mu_\ell)^T \Sigma_\ell^{-1} (x - \mu_\ell) \quad (6.60)$$

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

Définition 6.25 L'hypersurface de \mathbb{R}^L d'équation $\Delta_{k,\ell} = 0$ est appelée frontière de décision entre les classes k et ℓ , qui coupe \mathbb{R}^L en deux régions.

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

6.5.1 Analyse discriminante linéaire (LDA)

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

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

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

$$\Delta_{k,\ell}(x) = \ln\left(\frac{p_k}{p_\ell}\right) - \frac{1}{2}(x - \mu_k)^T \Sigma^{-1} (x - \mu_k) + \frac{1}{2}(x - \mu_\ell)^T \Sigma^{-1} (x - \mu_\ell). \quad (6.61)$$

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

$$\Delta_{k,\ell}(x) = \delta_k(x) - \delta_\ell(x), \quad (6.62)$$

où on a introduit la fonction de discrimination

$$\delta_k(x) = \ln p_k(x) + x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k. \quad (6.63)$$

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

$$\Phi_{k\ell} = \Sigma^{-1} (\mu_k - \mu_\ell), \quad (6.64)$$

que l'on appellera vecteur discriminant.

Proposition 6.24 Avec les notations ci-dessus, la règle de Bayes en analyse discriminante linéaire conduit à la règle de décision : pour $x \in \mathbb{R}^L$,

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

Remarque 6.25 En pratique, la matrice de covariance Σ n'est pas connue, de même que les moyennes μ_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 Σ , car la LDA demande que la matrice estimée soit inversible.

■ **Exemple 6.29 — Le cas de la classification biclassé.** Pour simplifier, supposons qu'il n'existe que deux classes, numérotées 0 et 1. La décision pour une donnée x sera alors

$$\hat{c}(x) = \begin{cases} 1 & \text{si } x^T \Phi > -\ln(p_1/p_0) + \frac{1}{2} (\mu_1 + \mu_0)^T \Phi \\ 0 & \text{sinon} \end{cases}$$

où

$$\Phi = \Sigma^{-1} (\mu_1 - \mu_0)$$

est le *vecteur discriminant*. Il s'agit donc de comparer la projection de la donnée x sur l'axe défini par le vecteur discriminant Φ et de la comparer à la projection de la moyenne des deux moyennes (tout en corrigeant le résultat pour tenir compte des probabilités d'apparition des deux classes). ■

6.5.2 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 (6.60)) 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 6.26 — 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

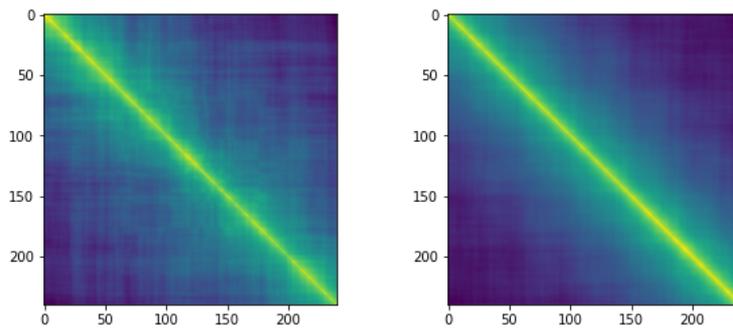


FIGURE 6.11 – Matrices de covariance empiriques pour les signaux "cible" (gauche) et "non-cible" (droite) du jeu de données P300 speller.

régulariser, en les remplaçant par une moyenne pondérée d'elles mêmes et de la matrice globale :

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

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 λ ...

6.6 Application aux données de *P300-speller*

6.6.1 Position du problème

Pour terminer, considérons de nouveau le problème de décision associé au *P300-speller*. On dispose de deux jeux de 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 μ_0 et μ_1 , les probabilités des classes p_0 et p_1 et la matrice de covariance Σ (ou les matrices de covariance des classes Σ_0 et Σ_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.

Dans un premier temps, commençons par justifier le choix d'une décision basée sur la LDA plutôt que la QDA. On a représenté en FIGURE 6.11 les matrices de covariance empiriques pour les jeux de données "cible" et "non-cible" du P300 speller. On constate que ces deux matrices sont très proches l'une de l'autre (en fait elles diffèrent de moins de 10%). On ne comment donc pas une grosse erreur en utilisant une matrice de covariance unique. Par ailleurs, les jeux de données sont déséquilibrés, le jeu "non-cible" contient 5 fois plus d'échantillons que le jeu "cible", dont la matrice de covariance est donc moins bien estimée.

Par ailleurs, il faut noter que le problème du P300 speller est différent du problème classique de classification, au sens où les données sont "structurées" : pour chaque *run* (caractère cible), 6 lignes et 6 colonnes sont illuminées (avec 15 répétitions), et il faut déterminer laquelle est la plus susceptible de contenir le caractère cible. On ne compare donc pas le score à un seuil, on recherche le meilleur score.

Plus précisément, la stratégie utilisée dans le TP est la suivante

1. Sur le jeu d'apprentissage :
 - Estimer la matrice de covariance et les moyennes des deux classes
 - Calculer le vecteur discriminant

2. Sur le jeu de test (peut être fait aussi sur le jeu d'apprentissage) : Pour chaque *run*,
 - pour chaque ligne et chaque colonne, calculer la moyenne sur les 15 essais.
 - Calculer le score de chaque ligne et chaque colonne (produit scalaire avec le vecteur discriminant)
 - Déterminer la ligne et la colonne de score maximum, en déduire le caractère prédit.

On peut ensuite comparer les caractères prédits aux vrais, et calculer le pourcentage de bonnes réponses.

On voit bien qu'il ne s'agit pas d'une analyse discriminante classique, de par la structure des données.

- Dans une analyse discriminante, chaque signal est traité indépendamment des autres, son produit scalaire avec le vecteur discriminant Φ est comparé à un seuil (voir l'exemple 6.29).
- Dans le cas d'intérêt ici, on sait qu'à chaque *run* (c'est à dire pour chaque caractère cible), une seule ligne et une seule colonne doivent être sélectionnées. On va donc comparer les scores des lignes et sélectionner la meilleure (indépendamment du seuil), même chose pour la sélection de la colonne. Le seuil de la LDA n'est donc pas utilisé.

6.6.2 Utilisation d'un seul capteur

On dispose de données enregistrées sur 64 capteurs, mais tous ne portent pas d'information également importante pour le problème considéré. Comme on s'attend à ce que les signaux P300 soient particulièrement visibles dans la région occipitale, on se limite tout d'abord au capteur Cz (voir Figure 5.3).

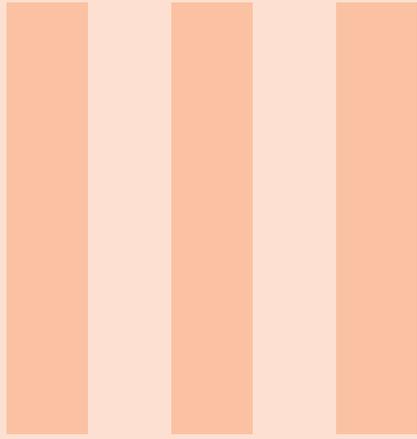
On dispose donc de signaux de longueur 240 (1 seconde échantillonnée à 240 Hz). Il y a 85 *runs*, 6 lignes et 6 colonnes, et 15 essais par ligne et par colonne, donc 15 300 signaux au total dans le jeu d'apprentissage. Faisant l'hypothèse d'égalité des matrices de covariance des deux classes, on peut considérer que ces 15 300 signaux sont suffisants pour estimer une matrice de covariance 240×240 .

Les résultats de prédiction sur le sujet A sont les suivants

- 65% de bonnes classifications sur le jeu d'apprentissage
- 27% de bonnes classifications sur le jeu de test

Ces deux pourcentages sont très supérieurs à ce que l'on obtiendrait en tirant au hasard un caractère pour chaque *run* (moins de 3%), mais ne sont pas vraiment satisfaisants. En particulier, on a près de 3/4 d'échecs sur le jeu de test, ce qui est inacceptable pour une application pratique. De plus, le taux de succès avec le jeu de test est très inférieur au taux de succès avec le jeu d'apprentissage, ce qui est un signe très net de sur-apprentissage.

6.6.3 Exploitation de plusieurs capteurs



Compléments

Bibliographie 99

Index 101

Bibliographie

- [1] CARDIOSECUR : ECG lead systems. Disponible en ligne sur <https://www.cardiosecur.com/magazine/specialist-articles-on-the-heart/lead-systems-how-an-ecg-works>.
- [2] T. CHANWIMALUEANG, W. von ROSENBERG et D. P. MANDIC : Enabling R-peak detection in wearable ECG : Combining matched filtering and Hilbert transform. In *2015 IEEE International Conference on Digital Signal Processing (DSP)*, p. 134–138, July 2015.
- [3] M. ELGENDI : Fast QRS detection with an optimized knowledge-based method : Evaluation on 11 standard ecg databases. *PLOS ONE*, 8(9):1–18, 09 2013.
- [4] A. GACEK et W. PEDRYCZ : *ECG Signal Processing, Classification and Interpretation : A Comprehensive Framework of Computational Intelligence*. Springer Publishing Company, Incorporated, 1st édn, 2011.
- [5] A. L. GOLDBERGER, L. A. N. AMARAL, L. GLASS, J. M. HAUSDORFF, P. C. IVANOV, R. G. MARK, J. E. MIETUS, G. B. MOODY, C.-K. PENG et H. E. STANLEY : Physiobank, physiotoolkit, and physionet. *Circulation*, 101(23):e215–e220, 2000.
- [6] INFIRMIERDEREA : Fonctionnement électrique du cœur et ECG, 2021. Disponible en ligne sur <http://infirmierderea.unblog.fr/2016/12/24/fonctionnement-electrique-du-coeur-et-ecg/>.
- [7] T. MUNNIER, M. BILLAUD, M. MAURY et C. GUINFARD : Le fonctionnement électrique du cœur, 2916. Disponible en ligne sur <https://trimamacle.wixsite.com/defibrillateurpe/blank-i9xgb>.
- [8] F. E. OLVERA : Electrocardiogram waveform feature extraction using the matched filter. In *ECE510; Statistical Signal Processing II*, 2001.
- [9] J. PAN et W. J. TOMPKINS : A real-time QRS detection algorithm. *IEEE Transactions on Biomedical Engineering*, BME-32(3):230–236, 1985.
- [10] I. SILVA et G. B. MOODY : An open-source toolbox for analysing and processing physionet databases in matlab and octave. *Journal of open research software*, 2(1):e27, 2014.
- [11] J. O. SMITH : *Introduction to Digital Filters, with audio applications*. Stanford University, Center for Computer Research in Music and Acoustics (CCRMA), 2007.
- [12] L. SÖRNMO et P. LAGUNA : *Electrocardiogram (ECG) Signal Processing*. American Cancer Society, 2006.

Index

- Arythmie cardiaque, 14
- Atome de Gabor, 39, 40
- Atome temps-fréquence, 38
- Biais, 84
- Bradycardie, 14
- Caractéristique, 91
- Complexe QRS, 13, 14
- Convergence en probabilités, 85
- Cortex, 59
- Coût (0,1), 91
- Détection de pics, 47
- EEG, 60
- Electroencéphalographie, 60
- Equation aux différences, 26
- Erreur quadratique moyenne, 84
- Espace $\ell^2(\mathbb{Z})$, 19
- Estimateur du maximum de vraisemblance, 88
- Fenêtre, 37
- Fenêtre duale canonique, 43
- Filtrage adapté, 53
- Filtre de Butterworth, 31
- Filtre dérivateur, 48
- Filtre médian, 54
- Filtre numérique, 21
- Filtre numérique causal, 21
- Filtre numérique d'ordre fini, 31
- Filtre numérique réalisable, 21
- Filtre numérique stable, 21
- Filtre quadrateur, 49
- Filtre récursif, 26
- Fonction erf, 81
- Fonction de perte, 91
- Fonction de transfert, 25
- Fonction de transfert d'un filtre numérique, 30
- Fréquence, 22
- Fréquence d'échantillonnage, 19, 62
- Gain du filtre, 30
- i.i.d., 84
- Imagerie fonctionnelle, 60
- Imagerie structurelle, 60
- Impédance, 25
- Interface cerveau machine, 61
- Intervalle RR, 13
- Inégalité de Markov, 85
- Largeur de bande relative, 45
- Ligne de base, 22
- Ligne isoélectrique, 22
- Log rapport de vraisemblance, 93
- Log-vraisemblance, 88
- Magnétoencéphalographie, 60
- Matière blanche, 59
- Matière grise, 59
- MEG, 60
- Modèle de mélange de gaussiennes, 93
- Moyenne mobile, 49
- Ondelette, 44
- Ondes P, Q, R, S, T, 13
- Opérateur de trame, 43
- P300-speller, 61

Prédicteur, 91
Pulsation, 22

QRS, 13

Réponse impulsionnelle, 21
Représentation pôles-zéros, 31
Risque, 91
Rythme cardiaque, 47
Rythme sinusal, 13
Règle de Bayes, 91
Réponse fréquentielle, 25

Série de Laurent, 28
Spectre, 33
Spectrogramme, 37
Statistique descriptive, 84
Statistique inférentielle, 84
Sur-apprentissage, 90, 96
Symétrie Hermitienne, 23, 24

Tachycardie, 14
TFCT inverse, 38, 40
TFD, 22
Transformée en z , 28
Transformation de Fourier Discrète, 22
Transformation de Fourier discrète inverse, 23
Transformation de Fourier Finie, 23
Transformation de Gabor, 39, 40
Transformation de Gabor inverse, 43
Transformation en z inverse, 29
Transformation en ondelettes stationnaire, 45
Transformation à Q constant, 45

Variable d'entrée, 91
Variable de sortie, 91
Variable indépendante, 91
Vecteur gaussien, 81
Vraisemblance, 87