Cours GBM6103A

Traitement d'images et de signaux biomédicaux

NOTES DE SYNTHÈSE

Yves Goussard et Benoît Hamelin

Département de génie électrique
Institut de génie biomédical

ÉCOLE POLYTECHNIQUE DE MONTRÉAL

Version provisoire

12 septembre 2012

Table des matières

1 Signaux aléatoires à temps discret **1**
- 1.1 Généralités . . . 1
  - 1.1.1 Caractéristiques à l'ordre 2 . . . 2
  - 1.1.2 Propriétés des signaux aléatoires . . . 2
- 1.2 Signaux aléatoires stationnaires d'ordre 2 . . . 3
  - 1.2.1 Position du problème . . . 3
  - 1.2.2 Représentation fréquentielle . . . 4
  - 1.2.3 Filtrage . . . 5
  - 1.2.4 Modélisation . . . 5

2 Détection et mesure d'ondes **7**
- 2.1 Introduction . . . 7
- 2.2 Cadre de travail et formulation du problème . . . 7
- 2.3 Détection d'échos « courts » . . . 8
  - 2.3.1 Résolution par maximum *a posteriori* . . . 8
  - 2.3.2 Détecteur par gabarit . . . 9
- 2.4 Amplitude et du temps d'arrivée d'une d'onde . . . 10
- 2.5 Détection et estimation de plusieurs ondes . . . 11
  - 2.5.1 Formulation comme un problème de déconvolution . . . 12

3 Traitement d'images médicales I : Échantillonnage, filtrage, segmentation **17**
- 3.1 Représentation fréquentielle . . . 17
- 3.2 Échantillonnage . . . 20
- 3.3 Convolution, corrélation, filtrage . . . 20
- 3.4 Applications . . . 21

4 Traitement d'images médicales II : restauration, reconstruction **25**
- 4.1 Restauration d'images . . . 25
- 4.2 Minimisation itérative d'un critère quadratique . . . 27
- 4.3 Reconstruction tomographique . . . 29

5 Reconnaissance de formes I : approches statistiques **35**
5.1 Présentation du problème 35
5.2 Définitions 37
5.3 Classification 37
5.4 Estimation des caractéristiques des classes 40

6 Reconnaissance de formes II : réduction du nombre des attributs ; approche par réseaux de neurones **43**
6.1 Réduction du nombre d'attributs 43
6.2 Classification par réseau de neurones 47

7 Compression de signaux biomédicaux **53**
7.1 Position du problème 53
7.2 Redondance liée à la source 55
7.3 Redondance structurelle 57

Table des figures

2.1 Filtrage par gabarit. Le signal $s(n)$ apparaît en traits pleins et le gabarit en traits pointillés. $p(\boldsymbol{z}|\mathcal{H}_1)$ prend une valeur non nulle si et seulement si le signal $z(n)$ rentre dans le gabarit. 10
2.2 Extension du détecteur par corrélation au cas d'ondes multiples sans recouvrement. 11
2.3 Représentation schématique du problème de détection-estimation de formes d'ondes. L'instant d'occurrence et l'amplitude de chaque impulsion composant le signal d'entrée x correspond à ceux des formes d'ondes présentes dans z. 12

6.1 Structure d'un réseau de neurones multicouches (d'après [1]). . . 50

Chapitre 1

Signaux aléatoires à temps discret

1.1 Généralités

L'idée intuitive est de définir un signal aléatoire à temps discret comme un signal dont chaque échantillon est une variable aléatoire (VA). On se heurte immédiatement à la difficulté de devoir définir la loi conjointe d'une *infinité* d'échantillons. Pour contourner cette difficulté, on définit la *loi temporelle* de l'ensemble infini de VA $\{X_k \ ; \ k \in \mathbb{N}\}$ comme suit :

Définition 1.1 – Loi temporelle. Pour tout entier n, pour tout ensemble d'indices $\{k_1, k_2, \ \dots \ , k_n\}$, la loi conjointe $f_{X_{k_1}, \ \dots \ , X_{k_n}}(x_{k_1}, \ \dots \ , x_{k_n})$ existe.

De manière analogue, on définit l'indépendance de deux signaux aléatoires $\{X_k \ ; \ k \in \mathbb{N}\}$ et $\{Y_l \ ; \ l \in \mathbb{N}\}$:

Définition 1.2 – Indépendance de deux signaux aléatoires. Les deux signaux aléatoires X et Y sont indépendants si et seulement si pour tout couple d'entiers (m, n), pour tous sous-ensembles d'indices $\{k_1, k_2, \ \dots \ , k_m\}$ et $\{l_1, l_2, \ \dots \ , l_n\}$, les vecteurs aléatoires $(X_{k_1}, \ \dots \ , X_{k_m})$ et $(Y_{l_1}, \ \dots \ , Y_{l_n})$ sont indépendants.

De la loi temporelle d'un signal aléatoire $\{X_k \ ; \ k \in \mathbb{N}\}$, on peut évidemment déduire ses caractéristiques instantanées, qui sont les caractéristiques statistiques de chaque échantillon pris isolément. Entre autres, on définit la *loi instantanée* du signal comme la loi de chaque échantillon $f_{X_k}(x_k)$. De ceci, on déduit les autres caractéristiques instantanées : espérance, moyenne, variance, etc.

1.1.1 Caractéristiques à l'ordre 2

L'importance des caractéristiques du second ordre apparaîtra à la section 1.2.

Définition 1.3 – Signal d'ordre 2. Un signal aléatoire est d'ordre deux si et seulement si la variance de chaque échantillon existe.

On peut montrer que, si un signal est d'ordre deux, la covariance de tout couple d'échantillons existe. On définit alors la covariance du signal aléatoire d'ordre 2 $\{X_k \ ; \ k \in \mathbb{N}\}$ comme :

$$\forall (i,j) \ ; \ r_X(i,j) \triangleq \mathrm{cov}(X_i, X_j) = E\left[(X_i - E[X_i])\,(X_j - E[X_j])\right].$$

La covariance du signal aléatoire traduit le lien statistique entre les échantillons X_i et X_j. Si :

$$\forall (i,j) \ ; \ i \neq j \Rightarrow r_X(i,j) = 0,$$

alors le signal aléatoire est *décorrélé*, ce qui indique en première approximation un lien statistique faible entre échantillons différents.

1.1.2 Propriétés des signaux aléatoires

Stationnarité

La stationnarité traduit le fait que les propriétés statistiques d'un signal aléatoire restent invariantes au cours du temps. On distingue deux types de stationnarité :

- la *stationnarité forte* traduit l'invariance de *toutes* les propriétés statistiques du signal. Elle se définit donc comme l'invariance par translation de la loi temporelle du signal aléatoire ;
- la *stationnarité faible* ou *d'ordre deux* est de portée plus limitée : elle signifie que les moments d'ordres un et deux du signal aléatoire existent et sont invariants par translation, ce qui se traduit par :

$$\forall (i,j) \in \mathbb{N}^2 \ ; \ E[X_i] = E[X_j] = m,$$
$$\forall (i,j,k) \in \mathbb{N}^3 \ ; \ r_X(i,j) = r_X(i-k, j-k).$$

  En prenant $k = j$ dans cette dernière équation, on observe que $r_X(i,j) = r_X(i-j, 0)$, ce qui indique que la covariance des deux échantillons X_i et X_j ne dépend que de leur distance $i - j$. Dans ce cas, on note :

$$r_X(i,j) \triangleq r_X(i-j) \ : \ \text{fonction d'autocorrélation.}$$

  Soulignons que, en pratique, un signal aléatoire stationnaire d'ordre deux est complètement caractérisé d'un point de vue statistique par sa moyenne m et sa fonction d'autocorrélation r_X.

Ergodisme pour une propriété

Le calcul de propriétés statistiques d'un signal aléatoire (par exemple, moments, espérance d'une fonction des échantillons, etc.) nécessite de connaître la loi temporelle du signal, de manière complète ou partielle. Or, dans la pratique, une telle connaissance peut ne pas être disponible. La propriété d'ergodisme permet de remplacer une moyenne au sens de l'espérance mathématique par une moyenne temporelle. Plus précisément, on dit que le signal $\{X_k \ ; \ k \in \mathbb{N}\}$ est ergodique pour la propriété φ si et seulement si, pour tout entier n et pour tout ensemble d'indices $\{k_1, k_2, \ \dots \ , k_n\}$:

$$E\left[\varphi(X_{k_1}, \ \dots \ , X_{k_n})\right] = \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \varphi(x_{i+k_1}, \ \dots \ , x_{i+k_n}).$$

On peut bien sûr s'interroger sur la validité d'une telle propriété lorsqu'on ne connaît pas la loi temporelle du signal aléatoire. En fait, sous des hypothèses plus faibles (bruit blanc par exemple), on peut montrer que la propriété d'ergodisme est vérifiée.

Exemples

Il existe évidemment de très nombreux types de signaux aléatoires. Mentionnons :

- les *bruits blancs*, dont tous les échantillons sont indépendants les uns des autres et distribués selon la même loi instantanée. Les bruits blancs peuvent prendre des allures très variées, puisque la distribution spatiale des échantillons est quelconque (binaire, gaussienne, etc.) ;
- les *signaux gaussiens*, caractérisés par le fait que leur loi temporelle est gaussienne. Si, de plus, un signal gaussien est décorrélé, ses échantillons sont alors indépendants et on est en présence d'un bruit blanc gaussien. Notons que l'allure d'un signal gaussien varie significativement en fonction de sa fonction de corrélation.

Ces éléments indiquent que, même en se limitant aux signaux aléatoires d'ordre 2, voire gaussiens, on peut représenter une vaste diversité de phénomènes.

1.2 Signaux aléatoires stationnaires d'ordre 2

1.2.1 Position du problème

Un de nos objectifs est d'appliquer aux signaux aléatoires les mêmes traitements (filtrage, représentation fréquentielle) qu'aux signaux déterministes. On voudrait donc pouvoir écrire :

$$\text{Représentation fréquentielle :} \quad X_k = \int u(\nu) e^{2i\pi k\nu} \, d\nu, \tag{1.1}$$

$$\text{Filtrage par un filtre de RI } h : \ Y_k = \sum_{n\in\mathbb{Z}} h(n) X_{k-n}. \tag{1.2}$$

Or nous ne disposons pas des outils mathématiques permettant de donner un sens satisfaisant à ces deux équations. Pour (1.1), il est clair que, X_k étant aléatoire et $e^{2i\pi k\nu}$ étant déterministe, la quantité $u(\nu)$ est nécessairement aléatoire. Or les intégrales usuelles ne peuvent être utilisées pour les quantités aléatoires. En ce qui concerne (1.2), on observe que Y_k est défini comme une somme *infinie* de variables aléatoires, ce qui soulève également quelques difficultés mathématiques. Dans le cadre de ce cours, plutôt que d'introduire les notions abstraites qui permettraient de donner un sens à (1.1) et (1.2), on contournera la difficulté en utilisant comme intermédiaire une quantité déterministe, la fonction de corrélation. Pour cela, on se restreint au cadre des signaux aléatoires stationnaires d'ordre deux.

1.2.2 Représentation fréquentielle

Pour introduire une représentation fréquentielle des signaux aléatoires stationnaires d'ordre deux, on se base sur le résultat suivant, qui découle du théorème de Bochner-Wiener-Khinchine :

Définition 1.4 – Représentation spectrale. Soit $\{X_k \ ; \ k \in \mathbb{N}\}$ un signal aléatoire stationnaire d'ordre 2. On peut définir $u_X(\nu)$, *représentation spectrale* de X, qui est une quantité aléatoire telle que :

$$X_k = \int u(\nu) e^{2i\pi k\nu} \, d\nu$$

où l'intégrale s'entend au sens de Lebesgue. De plus, on a :

$$E\left[|u(\nu)|^2\right] = \mathcal{F}(r_X(k)).$$

Cette dernière équation indique que l'espérance mathématique du carré du module de la représentation spectrale de X (qui est déterministe) n'est autre que la transformée de Fourier de la fonction d'autocorrélation de X (qui est également une quantité déterministe). On pose :

$$E\left[|u(\nu)|^2\right] = \Gamma_X(\nu) \ : \ \text{densité spectrale de puissance de } X,$$

et on observe que la densité spectrale de puissance est égale à la transformée de Fourier de la fonction d'autocorrélation. Ce résultat est à rapprocher du résultat équivalent pour les signaux déterministes, où la densité spectrale *d'énergie* est égale à la transformée de Fourier de la fonction d'autocorrélation. Soulignons que, bien évidemment, la fonction d'autocorrélation est définie différemment pour les signaux déterministes et les signaux aléatoires.

Remarque 1 La densité spectrale de puissance (ou plus simplement le spectre) de X est défini à partir du *module* de $u(\nu)$. On perd donc la notion de phase de la représentation spectrale des signaux aléatoires, alors que cette notion reste présente pour les signaux déterministes.

Remarque 2 Si on ne connaît ni la loi temporelle, ni la fonction d'autocorrélation du signal aléatoire, il n'est pas possible de calculer le spectre en utilisant les définitions ci-dessus. On doit donc s'appuyer sur la réalisation dont on dispose. Par définition :

$$r_X(k) = E[X_n X_{n-k}]$$

et, par ergodisme :

$$r_X(k) = \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} x_n x_{n-k}.$$

On observe que, au facteur $1/N$ près, on retrouve la définition de la fonction d'autocorrélation d'un signal déterministe. Cependant, dans la pratique, on ne dispose que d'un nombre fini d'échantillons de X et on n'obtient donc qu'une valeur approchée de $r_X(k)$. Il est alors souvent préférable de poser le problème de détermination de r_X ou Γ_X comme un problème d'estimation. Ceci fait l'objet de la séance consacrée à l'estimation spectrale.

Remarque 3 Quelle que soit la distribution de chaque échantillon, la fonction d'autocorrélation d'un bruit blanc prend la forme d'une impulsion située en zéro. Il en résulte que le spectre d'un bruit blanc est toujours plat, quelle que soit la distribution du bruit (gaussienne, binaire, etc.).

1.2.3 Filtrage

Lorsque l'on dispose des échantillons issus d'un signal aléatoire stationnaire d'ordre 2, il est toujours possible de filtrer ceux-ci par un filtre de RI h. On peut alors se demander si les échantillons y_k de la sortie ainsi obtenue peuvent être modélisés comme un autre signal aléatoire d'ordre deux et, si tel est bien le cas, quelles sont les caractéristiques de ce signal Y. Pour répondre directement à la question, on est tenté de s'appuyer sur (1.2). Comme mentionné plus haut, cette équation présente des difficultés mathématiques liées au caractère infini de la somme. On peut néanmoins donner un sens à (1.2), ce qui permet de déduire que :

$$r_Y(k) = r_X(k) * h(k) * h(-k), \tag{1.3}$$

d'où :

$$\Gamma_Y(\nu) = |H(\nu)|^2 \Gamma_X(\nu). \tag{1.4}$$

Il en découle que le résultat du filtrage de la réalisation $\{x_k \ ; \ k \in \mathbb{Z}\}$ de X peut être considéré comme la réalisation d'un signal aléatoire stationnaire d'ordre deux Y dont la fonction d'autocorrélation et le spectre sont liés à ceux de X par (1.3) et (1.4).

1.2.4 Modélisation

L'objectif est de *paramétriser* les signaux aléatoires stationnaires d'ordre deux de manière à pouvoir soit en générer des réalisations (en précisant leurs caractéristiques statistiques), soit en estimer les caractéristiques statistiques (à

partir d'une ou plusieurs réalisations). De ce point de vue, (1.4) suggère une approche simple : si on suppose que X est un bruit blanc de variance σ^2, son spectre $\Gamma_X(\nu)$ est égal à $\sigma^2 1_\nu$. Il en résulte que le spectre de Y est égal à $\sigma^2|H(\nu)|^2$. Paramétriser Y est donc équivalent à paramétriser le filtre h. Or nous avons vu précédemment que les filtres dynamiques fournissent une paramétrisation à la fois générale et simple des filtres linéaires. Leur équation d'entrée-sortie a pour forme :

$$y_n - \sum_{p=1}^{P} a_p y_{n-p} = \sum_{q=0}^{Q} b_q x_{n-q},$$

et :

$$|H(\nu)|^2 = \frac{|B(\nu)|^2}{|A(\nu)|^2} \tag{1.5}$$

où $B(\nu)$ et $A(\nu)$ désignent respectivement les transformées de Fourier à temps discret des séquences $(b_0, b_1, \ldots, b_Q)$ et $(1, -a_1, \ldots, -a_P)$. Pour pouvoir paramétriser (à une constante près) tout signal aléatoire stationnaire d'ordre deux par les paramètres $(b_0, \ldots, b_Q, a_1, \ldots, a_P)$, il suffit donc de répondre à la question suivante : est-il possible de considérer tout signal aléatoire stationnaire d'ordre deux comme le résultat du filtrage d'un bruit blanc stationnaire d'ordre deux par un filtre dynamique ?

On peut montrer que, si la condition suivante est vérifiée :

$$\int_{-1/2}^{1/2} |\ln \Gamma_X(\nu)| \, d\nu \quad \text{converge},$$

la réponse à la question posée est positive, l'ordre du filtre dynamique étant possiblement infini. De plus, sous réserve de conditions techniques supplémentaires, on peut utiliser un filtre dynamique tous pôles ($Q = 0$).

Ce résultat peut être utilisé pour résoudre les problèmes pratiques suivants :
Analyse - L'objectif est ici d'estimer les caractéristiques statistiques d'un signal, donc sa fonction d'autocorrélation ou son spectre. D'après (1.5), pour connaître le spectre, il suffit de déterminer une paramétrisation de celui-ci, donc de choisir l'ordre du filtre dynamique, puis d'estimer les paramètres $(b_0, \ldots, b_Q, a_1, \ldots, a_P)$ à partir de la réalisation dont on dispose. Cette approche sera étudiée lors de la séance portant sur l'analyse spectrale.
Synthèse - Ici, l'objectif est de générer un bruit de caractéristiques statistiques, donc spectrales, données. En procédant par filtrage d'un bruit blanc, le spectre du signal résultant est donné par (1.5). Synthétiser le signal aléatoire désiré est donc équivalent à concevoir un filtre dynamique de caractéristiques fréquentielles données, ce qui peut être accompli en utilisant les techniques classiques de synthèse des filtres.

Chapitre 2

Détection et mesure d'ondes

2.1 Introduction

Détecter la présence d'une onde de forme connues et, au besoin, estimer son amplitude sont des problèmes qui se rencontrent fréquemment en génie biomédical. À titre d'exemple, mentionnons :

- l'échographie, dans laquelle il est nécessaire de détecter une forme d'onde (écho) renvoyée par les inhomogénéités du milieu étudié. Selon la distance entre ces inhomogénéités relativement à la longueur de l'écho, le problème peut être celui de la détection d'échos « courts », c'est-à-dire ne se recouvrant pas ou d'échos « longs », pouvant se recouvrir ;
- la détection de formes d'ondes particulières, comme le complexe QRS dans les électrocardiogrammes (ECG), afin de recaler plusieurs battements les uns par rapport aux autres ou de détecter des périodes d'ECG présentant un caractère pathologique.

L'objectif de ce chapitre est de mettre en évidence la différence de nature entre divers problèmes de détection et de mesure d'ondes, de présenter des stratégies de résolution adaptées à chacun de ces problèmes et d'exposer les méthodes de détection correspondantes, le tout étant mis en perspective dans le cadre général de l'estimation. Après avoir fixé le cadre de travail et formulé le problème de manière général, on examinera dans un premier temps les cas relativement simples de la détection d'échos « courts », puis de l'estimation du temps d'arrivée d'*une* onde de forme connue. On étudiera ensuite le cas plus complexe de l'estimation du temps d'arrivée et de l'amplitude de plusieurs ondes, avec possibilité de recouvrement ; deux approches adaptées à ce dernier cas seront présentées.

2.2 Cadre de travail et formulation du problème

On se place dans le cadre des signaux réels à temps discrets. On désigne par $z(n)$ le signal observé, c'est-à-dire celui contenant l'écho ou les échos à

détecter. La forme de l'*onde de référence* ou de la *forme d'écho* est définie par $s(n)$; $0 \leq n \leq N-1$. Celle-ci est supposée connue. Le signal observé peut alors être représenté comme :

$$z(n) = \sum_{i=1}^{I} a_i s(n - \tau_i) + b(n) \tag{2.1}$$

où l'entier I représente le nombre de formes d'onde ou d'échos présents dans le signal observé. Pour chacun d'entre eux a_i et τ_i en désignent respectivement l'amplitude et temps d'occurrence. Enfin $b(n)$ représente le bruit, destiné à prendre en compte toutes les incertitudes et erreurs de modélisation du système. Sauf indication contraire, on considérera que ce bruit est blanc, gaussien et centré.

Les différents problèmes examinés dans ce chapitre s différencient par les connaissances *a priori* disponibles sur I ainsi que par les valeurs possibles de a_i et τ_i.

2.3 Détection d'échos « courts »

Dans ce cas, le signal $z(n)$ contient *au plus* une forme d'onde, et il n'y a d'incertitude ni sur son instant d'occurrence, ni sur son amplitude. On peut donc respectivement fixer ces valeurs à zéro et à un et envisager les deux hypothèses suivantes : soit $I = 0$ (pas d'écho), soit $I = 1$ (un écho est présent). De manière équivalente, on peut fixer I à un et définir les deux hypothèses par $a_1 = 0$ (pas d'écho), et par $a_1 = 1$ (un écho est présent). Dans un cas comme dans l'autre, on doit choisir entre les deux hypothèses $\mathcal{H}_0$ (pas d'écho), et par $\mathcal{H}_1$ (un écho est présent), ce choix devant être fait au sens de l'estimation.

2.3.1 Résolution par maximum *a posteriori*

Dans de cas, on suppose que l'on connaît les probabilités *a priori* d'occurrence des hypothèses $\mathcal{H}_0$ et $\mathcal{H}_1$, que l'on désigne respectivement par p_0 et p_1. On suppose de plus que le vecteur $\boldsymbol{b}$ des échantillons du bruit suit une distribution gaussienne, centrée de matrice de covariance $\boldsymbol{R}_b$ (le bruit n'est pas nécessairement blanc). On a donc :

Hypothèse	**Probabilité *a priori***	**Vraisemblance**
$\mathcal{H}_0$	p_0	$\mathcal{N}(\boldsymbol{0}, \boldsymbol{R}_b)$
$\mathcal{H}_1$	p_1	$\mathcal{N}(\boldsymbol{s}, \boldsymbol{R}_b)$

et on doit maximiser la vraisemblance *a posteriori* $p(\mathcal{H}|\boldsymbol{z}) \propto p(\boldsymbol{z}|\mathcal{H})p(\mathcal{H})$ par rapport à $\mathcal{H}$. Comme $\mathcal{H}$ ne peut prendre que les deux valeurs $\mathcal{H}_0$ et $\mathcal{H}_1$, il suffit d'évaluer la vraisemblance *a posteriori* sous chacune des deux hypothèse et d'en

comparer les valeurs. Sous nos hypothèses, on peut écrire :

$$\begin{aligned} p(\mathcal{H}_0)p(\boldsymbol{z}|\mathcal{H}_0) &= \frac{p_0}{K}\exp\left(-\frac{\boldsymbol{z}^{\mathrm{t}}\boldsymbol{R}_b^{-1}\boldsymbol{z}}{2}\right) \\ p(\mathcal{H}_0)p(\boldsymbol{z}|\mathcal{H}_0) &= \frac{p_0}{K}\exp\left(-\frac{(\boldsymbol{z}-\boldsymbol{s})^{\mathrm{t}}\boldsymbol{R}_b^{-1}(\boldsymbol{z}-\boldsymbol{s})}{2}\right) \end{aligned}$$

où la constante de normalisation K est la même sous les deux hypothèse. En comparant le rapport de ces deux quantités à 1, on aboutit à :

$$\frac{p_0}{p_1}\exp\left(-\frac{\boldsymbol{z}^{\mathrm{t}}\boldsymbol{R}_b^{-1}\boldsymbol{z}}{2}+\frac{(\boldsymbol{z}-\boldsymbol{s})^{\mathrm{t}}\boldsymbol{R}_b^{-1}(\boldsymbol{z}-\boldsymbol{s})}{2}\right) \underset{\mathcal{H}_1}{\overset{\mathcal{H}_0}{\gtrless}} 1$$

ce qui, après développement des produits et passage au logarithme, conduit à :

$$\boldsymbol{s}^{\mathrm{t}}\boldsymbol{R}_b^{-1}\boldsymbol{z} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \frac{1}{2}\boldsymbol{s}^{\mathrm{t}}\boldsymbol{R}_b^{-1}\boldsymbol{s}+\ln(p_0/p_1) \tag{2.2}$$

Le résultat ci-dessus appelle les remarques suivantes :

- le détecteur présente une structure très simple, puisqu'il est basé sur une transformation linéaire, c'est-à-dire un filtrage, des mesures ;
- le résultat de ce filtrage est comparé à un seuil, ce qui permet l'étude des caractéristiques du détecteur en termes de probabilités de détection, de fausse alarme, de courbes COR, etc.
- le détecteur se prête à une extension empirique simple aux formes d'ondes multiples *sans recouvrement*, en segmentant le signal observé z en plusieurs fenêtres ;
- dans le cas particulier où le bruit est blanc, $\boldsymbol{R}_b = \boldsymbol{I}$ et le détecteur (2.2) se simplifie en :

  $$\boldsymbol{s}^{\mathrm{t}}\boldsymbol{z} \underset{\mathcal{H}_0}{\overset{\mathcal{H}_1}{\gtrless}} \frac{1}{2}\boldsymbol{s}^{\mathrm{t}}\boldsymbol{s}+\ln(p_0/p_1) \tag{2.3}$$

  qui est la définition d'*filtre adapté*. On remarque par ailleurs que la quantité $\boldsymbol{s}^{\mathrm{t}}\boldsymbol{z} = \sum_{n=0}^{N-1} s(n)z(n)$ s'interprète comme la valeur de la fonction d'inter-corrélation γ_{sz} en zéro, sous hypothèse déterministe. Autrement dit, le filtre adapté présente des liens très forts avec le détecteur par corrélation présenté dans [2, pages 55 à 58] et dont il sera de nouveau question plus bas.

Soulignons enfin les principales limitations d'un tel détecteur : (i) les hypothèses sur lesquelles il est basées sont très fortes, de type tout ou rien, et en limitent donc la gamme d'application ; (ii) en particulier, il n'est possible de détecter qu'une seule forme d'onde par fenêtre traitée.

2.3.2 Détecteur par gabarit

On conserve ici les mêmes hypothèses que précédemment, à ceci près que le bruit est maintenant supposé blanc, centré et *uniformément distribué* sur un

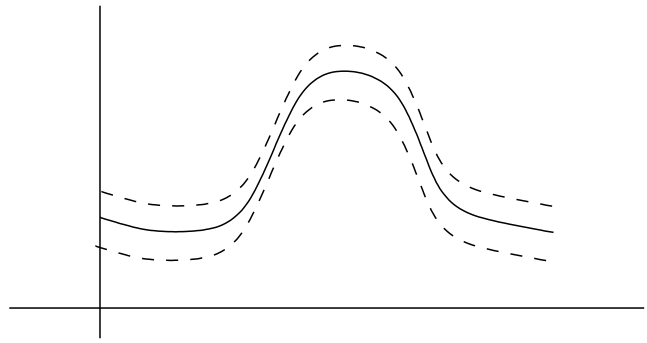


FIGURE 2.1 – Filtrage par gabarit. Le signal $s(n)$ apparaît en traits pleins et le gabarit en traits pointillés. $p(\boldsymbol{z}|\mathcal{H}_1)$ prend une valeur non nulle si et seulement si le signal $z(n)$ rentre dans le gabarit.

intervalle $[-\epsilon\ ,\ \epsilon]$. Sous l'hypothèse $\mathcal{H}_1$, $z(n)$ est uniformément distribué sur l'intervalle $[s(n)-\epsilon\ ,\ s(n)+\epsilon]$. De par l'indépendance du bruit, on peut écrire :

$$p(\boldsymbol{z}|\mathcal{H}_1) = \prod_{n=0}^{N-1} \frac{1}{2\epsilon} \mathrm{Ind}_{[-\epsilon\ ,\ \epsilon]}(z(n) - s(n))$$

où $\mathrm{Ind}_{[a\ ,\ b]}$ désigne la fonction indicatrice de l'intervalle $[a\ ,\ b]$. On observe que $p(\boldsymbol{z}|\mathcal{H}_1)$ ne peut prendre que deux valeurs : soit $1/(2\epsilon)^N$ si $z(n)$ rentre dans le gabarit $[s-\epsilon\ ,\ s+\epsilon]$ (voir figure 2.1), soit 0 dans le cas contraire. L'idée intuitive de la détection par gabarit est donc de ne détecter l'onde que si elle rentre effectivement dans le gabarit dont la taille est contrôlée par ϵ. Un tel détecteur présente des limitations similaires à celles du détecteur par MAP avec cependant l'avantage pratique d'une robustesse accrue.

2.4 Estimation de l'amplitude et du temps d'arrivée d'une forme d'onde

On suppose maintenant que le signal $z(n)\ ;\ 0 \leq n \leq L\ ,\ L > N$ contient *une* forme d'onde dont l'amplitude a et le temps d'arrivée τ sont inconnus. En se référant au cadre introduit au paragraphe 2.2, on est donc dans le cas où $I = 1$ et où a et τ doivent être estimés. On suppose de plus que le bruit b est blanc, gaussien, centré et de variance égale à σ_b^2. Ce problème se résout simplement par maximum de vraisemblance (MV), le développement de l'estimateur étant une simple extension de l'exemple donné dans [2, pages 55 à 58]. On en rappelle ici les principales étapes.

On cherche à exprimer la vraisemblance des paramètres définie par $f(\boldsymbol{z}\ ;\ a, \tau)$. Sous nos hypothèses, cette distribution de probabilité est gaussienne de moyenne $a\,\mathrm{Vect}(s(n-\tau))$ et de matrice de covariance $\sigma_b^2 \boldsymbol{I}$. L'objectif est donc de recher-

FIGURE 2.2 – Extension du détecteur par corrélation au cas d'ondes multiples sans recouvrement.

cher :

$$\arg\min_{a,\tau} -\ln(f(\boldsymbol{z}\ ;\ a,\tau)) \equiv \arg\min_{a,\tau} J(a,\tau)$$
$$\text{avec} \quad J(a,\tau) \triangleq \sum_n (z(n) - as(n-\tau))^2$$

Dans un premier temps, on observe que le minimum de J par rapport à τ ne dépend pas de a. Comme dans l'exemple traité dans [2], on aboutit à :

$$\begin{aligned} \arg\min_{\tau} J(a,\tau) &= \arg\max_{\tau} \sum_n z(n)s(n-\tau) \\ &= \arg\max_{\tau} \gamma_{zs}(\tau) \end{aligned}$$

La valeur de τ est donc donnée par le maximum (le cas échéant, en valeur absolue) de la fonction d'intercorrélation de z et s.

Dans un second temps, une fois $\hat{\tau}$ détermine on annule le gradient de J par rapport à a, ce qui conduit à :

$$-\sum_n s(n-\hat{\tau})(z(n) - as(n-\hat{\tau})) = 0$$

d'où :

$$\begin{aligned} \hat{a} &= \frac{\sum_n z(n)s(n-\hat{\tau})}{\sum_n s(n-\hat{\tau})^2} \\ &= \frac{\gamma_{zs}(\hat{\tau})}{\gamma_{ss}(0)} \end{aligned}$$

On observe que $\gamma_{zs}(\tau)$ est la sortie d'un filtre linéaire don l'entrée est le signal z et la réponse impulsionnelle (RI) le signal s retourné, que l'on désignera par s^-. On retrouve donc une structure de détecteur analogue à celle du filtre adapté présenté au paragraphe précédent. De plus, cette structure, illustrée à la figure 2.2, se prête à une extension empirique immédiate pour la détection d'ondes multiples ne se recouvrant pas. Une telle structure a cependant des limitations analogues à celles d'un simple filtre adapté, et son comportement est difficile à caractériser si les ondes se recouvrent mutuellement. On observe également qu'il n est pas possible d'obtenir une expression explicite de l'estimée de τ.

2.5 Détection et estimation de plusieurs ondes

On traite maintenant le cas général, c'est-à-dire celui dans lequel le nombre I de formes d'onde, les amplitudes a_i et les instants d'occurrence τ_i sont tous

FIGURE 2.3 – Représentation schématique du problème de détection-estimation de formes d'ondes. L'instant d'occurrence et l'amplitude de chaque impulsion composant le signal d'entrée x correspond à ceux des formes d'ondes présentes dans z.

inconnus. Après avoir formulé ce problème de détection-estimation comme une problème de déconvolution, on présente deux approches pour sa résolution.

2.5.1 Formulation comme un problème de déconvolution

On rappelle que le signal observé $z(n)$ s'exprime comme :

$$z(n) = \sum_{i=1}^{I} a_i s(n - \tau_i) + b(n) \quad ; \quad 0 \leq n \leq L - 1$$

Considérons maintenant le signal impulsionnel x défini par :

$$x(n) = \sum_{i=1}^{I} a_i \delta(n - \tau_i) \tag{2.4}$$

et exprimons le résultat y du filtrage de s par un filtre linéaire dont la RI est égale à $s(n)$; $0 \leq n \leq N - 1$. Par définition du produit de convolution discret, on a :

$$\begin{aligned} y(n) &= \sum_{k=0}^{N-1} \sum_{i=1}^{I} a_i s(k) \delta(n - k - \tau_i) \quad ; \quad 0 \leq n \leq L - 1 \\ &= \sum_{i=1}^{I} a_i \sum_{k=0}^{N-1} s(k) \delta(n - k - \tau_i) \quad ; \quad 0 \leq n \leq L - 1 \\ &= \sum_{i=1}^{I} a_i s(n - \tau_i) \quad ; \quad 0 \leq n \leq L - 1 \end{aligned}$$

où la dernière équation provient de ce que $\delta(n - k - \tau_i)$ est nul sauf pour $k = n - \tau_i$. On peut donc écrire que $z(n) = (x * s)(n) + b(n)$ avec la définition de x donnée plus haut. D'après cette définition (voir (2.4)), trouver x équivaut à trouver I, a_i et τ_i ; $1 \leq i \leq I$, donc à trouver le nombre, l'amplitude et l'instant d'occurrence des formes d'onde. Toujours d'après (2.4), x est un signal composé d'impulsions, l'amplitude et l'instant d'occurrence de chaque impulsion étant celles de la forme d'onde correspondante. Résoudre le problème de détection-estimation est donc équivalent à déconvoluer le signal impulsionnel x observé à travers le filtre de RI s, tel que schématisé à la figure 2.3.

- Restauration d'un signal Bernouilli-gaussien (B-G)

  → Restauration de signal : approche MAP (conditions difficiles).

  Modèle de signaux impulsionnels?

  Pour chaque échantillon : probabilité Ω d'occurence d'une impulsion.

  Si impulsion présente : amplitude gaussienne, de moyenne nulle et de variance σ_x^2.

  Processus blanc.
  $\Rightarrow$ Processus binaire q (présence ou absence d'impulsion)

$$p(x(n)|q(n)=1) \rightsquigarrow \mathcal{N}(0,\sigma_x^2).$$

  → Deux quantités à estimer.

  q **:** détection;
  $x|q$ **:** estimation des amplitudes.

  Estimation conjointe? $\longrightarrow f(x,q|z)$.
  Estimation marginale? $\longrightarrow f(q|z)$, puis $f(x|q,z)$.

  Possibilité de calculer, puis de maximiser les vraisemblances?

  → Estimation marginale.

  On cherche à exprimer $f(q|z) \propto f(z|q)f(q)$.
  q : séquence binaire, taille N, N_e composantes égales à 1.

  Si q est connu, x est gaussien de moyenne nulle et de matrice de covariance $\Pi = \sigma_x^2 \mathrm{diag} q$.

$$f(z|q) \rightsquigarrow \mathcal{N}(0, \underbrace{S\Pi S^T + \sigma_b^2 I}_{B})$$

$$f(z|q) = \frac{1}{(2\pi)^{N/2}|B|^{1/2}} \exp\left[-\frac{1}{2} z^T B^{-1} z\right]$$

$$f(q) = \lambda^{N_e}(1-\lambda)^{N-N_e}$$

$$\Rightarrow \log f(z|q)f(q) \propto -z^T B^{-1} z - \log|B| - 2N \log \frac{1-\lambda}{\lambda}$$

  Attention : on ne peut laisser tomber le coefficient de normalisation de la gaussienne, qui dépend de q par l'intermédiaire de B!

Maximisation explicite de la log-vraisemblance par rapport à q : impossible.

Idée : maximisation empirique par exploration de « voisinages » de la séquence courante.

Conditions :

* Pouvoir calculer facilement la vraisemblance sur le voisinage choisi.
* Pouvoir facilement mettre à jour les quantités intervenant dans le calcul de la vraisemblance lors du choix d'un nouveau q.

Conditions vérifiées (algèbre matricielle).

Compromis complexité des voisinages / efficacité de la maximisation.

→ Une fois q estimée...

$f(x|z)$: gaussienne non stationnaire.

Expression explicite : $\hat{x} = \Pi S^T z B^{-1} z$.

→ Remarques :

* Pas de solution explicite (problème de détection).
* Possibilité de tomber dans un minimum local.
* Robustesse parfois limitée de l'approche.

- Restauration d'un signal sous hypothèse $l_2 l_1$.

→ Idée générale :

* Approche « restauration d'entrée ».
* Terme de pénalisation favorisant l'allure impulsionnelle de x.

$$J(x) = (z - Sx)^T (z - Sx) + \lambda \sum_{n=1}^{N} \Phi(x_i)$$

Interprétation bayésienne

* Bruit gaussien;
* $f(x) \propto \frac{1}{K} \exp\left(-\sum_{n=1}^{N} \Phi(x_i)\right)$

Question : choix de Φ pour obtenir

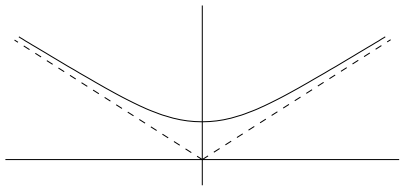
* une entrée impulsionnelle;

* un problème de minimisation possible à résoudre.

→ Si Φ convexe : J est convexe.

* Possibilité d'utiliser les algorithmes de descente (gradient, gradient conjugué, quasi-Newton, Newton...).

Exemple $\Phi(r)$ « $l_2 l_1$ »



$\Phi(r) = \sqrt{\delta^2 + r^2}$ (autres choix possibles).

Algorithmes efficaces : quasi-Newton, semi-quadratique.

→ En dernier ressort : f quadratique.

* Solution explicite (MAP sous hypothèses linéaire et gaussienne).
* Mauvais modèle pour un signal impulsionnel.
* Comparable à la détection par filtrage adapté ou corrélation! (avec un seuillage)
* Exemples de résultats.

5 Conclusion

- Attention aux hypothèses utilisées pour développer les différentes méthodes.
- Dans le cas le plus général : compromis à trouver entre
  - → robustesse,
  - → qualité des résultats,
  - → volume des calculs,
  - → caractère « vraiment impulsionnel » de l'entrée reconstruite.

Chapitre 3

Traitement d'images médicales I : Échantillonnage, filtrage, segmentation

Plan

- Définitions et cadre
- Représentation des images (domaine de Fourier)
- Échantillonnage
- Transformation (filtrage)
- Application : détection de contours, débruitage et segmentation.

3.1 Représentation fréquentielle

- Notion de fréquence : projection sur une « base fréquentielle ».

  En 1D : une variable dans le domaine de départ et une variable fréquentielle.
  En 2D : deux variables dans le domaine de départ $\Rightarrow$ deux variables fréquentielles.

Démarche intuitive par succession d'opérations.

$$\begin{aligned} f(x_1,x_2) \xrightarrow{\mathcal{F}_{\nu_1}} \overline{F}(\nu_1,x_2) &= \int f(x_1,x_2)e^{-2i\pi\nu_1 x_1}dx_1 \\ F(\nu_1,x_2) \xrightarrow{\mathcal{F}_{\nu_2}} F(\nu_1,\nu_2) &= \int F(\nu_1,x_2)e^{-2i\pi\nu_2 x_2}dx_2 \\ &= \int\int f(x_1,x_2)e^{-2i\pi(\nu_1 x_1 \nu_2 x_2)}dx_1dx_2 \end{aligned}$$

Interprétation :

→ Produit scalaire : $< f,g >= \int\int f(x_1,x_2)g^*(x_1,x_2)dx_1dx_2$.

→ « Base fréquentielle » : $\{e^{2i\pi(\nu_1 x_1 \nu_2 x_2)}\ ;\ (\nu_1,\nu_2) \in \mathbb{R}^2\}$.

Formalisation : pour $f(x_1,x_2)$ d'énergie finie (i.e. $\int\int |f(x_1,x_2)|^2 dx_1dx_2$ converge) :

TF 2D

$$F(\nu_1,\nu_2) = \int\int f(x_1,x_2)e^{-2i\pi(\nu_1 x_1+\nu_2 x_2)}dx_1dx_2$$

TF 2D inverse

$$f(x_1,x_2) = \int\int F(\nu_1,\nu_2)e^{2i\pi(\nu_1 x_1+\nu_2 x_2)}d\nu_1d\nu_2$$

Propriété importante : l'ordre d'intégration est indifférent (théorème de Fubini).

- Cas des fonctions périodiques

  → Il faut préciser la ou les variables par rapport auxquelles $f(x_1,x_2)$ est périodique.

  → La fonction f est décomposée en série de Fourier par rapport aux variables selon lesquelles elle est périodique.

  Exemple $f(x_1,x_2)$ périodique par rapport à x_1 et x_2, aux périodes T_1 et T_2.

  1. Décomposition de $f(x_1,x_2)$ en série de Fourier par rapport à x_1 :

$$f(x_1,x_2) = \sum_{k_1} a_{k_1}(x_2)e^{2i\pi k_1 x_1/T_1} \tag{3.1}$$

2. Décomposition des $a_{k_1}(x_2)$ en série de Fourier par rapport à x_2 :
$$a_{k_1}(x_2) = \sum_{k_2} a_{k_1 k_2} e^{2i\pi k_2 x_2/T_2} \tag{3.2}$$
3. En substituant (3.2) dans (3.1),
$$f(x_1, x_2) = \sum_{k_1,k_2} a_{k_1 k_2} e^{2i\pi\left(\frac{k_1 x_1}{T_1} + \frac{k_2 x_2}{T_2}\right)} \tag{3.3}$$
et on montre sans difficulté que
$$a_{k_1 k_2} = \frac{1}{T_1 T_2} \int_{T_1} \int_{T_2} f(x_1, x_2) e^{-2i\pi\left(\frac{k_1 x_1}{T_1} + \frac{k_2 x_2}{T_2}\right)} dx_1 dx_2.$$
4. En prenant la TF 2D de (3.3) et en utilisant la propriété de séparabilité, ainsi que l'expression de la TF 1D d'une exponentielle complexe, on obtient
$$F(\nu_1, \nu_2) = \sum_{k_1,k_2} a_{k_1 k_2} \delta\left(\nu_1 - \frac{k_1}{T_1}\right) \delta\left(\nu_2 - \frac{k_2}{T_2}\right).$$
La TF 2D de f est un ensemble d'impulsions disposées régulièrement sur une grille cartésienne, à des intervalles de $\frac{1}{T_1}$ dans la direction ν_1, et $\frac{1}{T_2}$ dans la direction ν_2.

→ Symétriquement, si $f(x_1, x_2)$ est constituée d'impulsions régulièrement disposées sur une grille cartésienne, $F(\nu_1, \nu_2)$ est périodique (cas f échantillonné).

→ Finalement : si f est à la fois échantillonnée et périodique, sa TF 2D l'est aussi.

- Propriété de rotation.

Image $f(x_1, x_2) \longleftrightarrow F(\nu_1, \nu_2)$
Image tournée de θ $g(x_1, x_2) \longleftrightarrow G(\nu_1, \nu_2)$
avec $g(x_1, x_2) = f[\mathcal{R}_{-\theta}(x_1, x_2)]$; $\mathcal{R}_{-\theta}$: rotation d'angle $-\theta$.

En posant $\boldsymbol{\nu} = \begin{pmatrix} \nu_1 \\ \nu_2 \end{pmatrix}$ et $\boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$,
$$G(\nu_1, \nu_2) = \int\int f(x_1, x_2) e^{-2i\pi \boldsymbol{\nu}^{\mathrm{t}} \boldsymbol{x}}.$$
Par changement de variable $\boldsymbol{u} = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \mathcal{R}_{-\theta} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$, on obtient
$$\begin{aligned} G(\boldsymbol{\nu}) &= \int\int f(u_1, u_2) e^{-2i\pi[\mathcal{R}_{-\theta}(\boldsymbol{\nu})^{\mathrm{t}} \boldsymbol{u}]} |\mathcal{R}_{-\theta}| du_1 du_2 \\ &= F[\mathcal{R}_{-\theta}(\nu_1, \nu_2)], \end{aligned}$$

puisque $|\mathcal{R}_{-\theta}| = 1$. En d'autres termes, G est égale à F à laquelle on applique une rotation d'angle θ.

- Les autres propriétés usuelles de la TF 1D se transposent à la TF 2D.

3.2 Échantillonnage

- Principe du théorème d'échantillonnage : éviter le recouvrement spectral.
- Condition suffisante pour f à bande limitée (B_1, B_2).

  Échantillonnage sans perte si les périodes d'échantillonnage f_{e_1} et f_{e_2} selon x_1 et x_2 respectivement vérifient

$$\left.\begin{array}{lcl} f_{e_1} & > & 2B_1 \\ f_{e_2} & > & 2B_2 \end{array}\right\} \Leftrightarrow \left\{\begin{array}{lcl} T_{e_1} & < & \frac{1}{2B_1} \\ T_{e_2} & < & \frac{1}{2B_2} \end{array}\right.$$

3.3 Convolution, corrélation, filtrage

- Convolution 2D : extension immédiate de la convolution 1D.

  Pour f, g deux images,

$$\begin{aligned} (f * g)(y_1, y_2) &= \int\int f(x_1, x_2) g(y_1 - x_1, y_2 - x_2) dx_1 dx_2 \\ &= \int\int f(y_1 - x_1, y_2 - x_2) g(x_1, x_2) dx_1 dx_2. \end{aligned}$$

  Si on pose $h = f * g$, alors on montre que

$$H(\nu_1, \nu_2) = F(\nu_1, \nu_2) G(\nu_1, \nu_2).$$

- Fonction de corrélation 2D

  $\rightarrow$ Dans le domaine des temps :

$$\begin{array}{ll} \text{Intercorrélation} & \gamma_{fg}(\tau_1, \tau_2) = \int\int f(x_1, x_2) g(x_1 - \tau_1, x_2 - \tau_2) dx_1 dx_2 \\ \text{Autocorrélation} & \gamma_{ff}(\tau_1, \tau_2) = \int\int f(x_1, x_2) f(x_1 - \tau_1, x_2 - \tau_2) dx_1 dx_2 \end{array}$$

→ Dans le domaine des fréquences : si on note $\Gamma_{fg}(\nu_1, \nu_2)$ (resp. $\Gamma_{ff}(\nu_1, \nu_2)$) la TF 2D de γ_{fg} (resp. γ_{ff}), on a :

$$\begin{array}{ll} \text{Intercorrélation} & \Gamma_{fg}(\nu_1, \nu_2) = F(\nu_1, \nu_2)G^*(\nu_1, \nu_2) \\ \text{Autocorrélation} & \Gamma_{ff}(\nu_1, \nu_2) = |F(\nu_1, \nu_2)|^2 \end{array}$$

- Filtre linéaire invariant

Équivalence

→ h linéaire invariant
→ Relation entrée-sortie : convolution
→ Fonctions propres : exponentielles 2D

⇒ Caractérisation du filtre par sa RI 2D $h(x_1, x_2)$.

Propriétés :

Causalité : propriété sans lien avec la physique.
Stabilité : essentielle.

Représentation :

Par RIF : pas de difficulté (notion de « masque »).
Par filtre auto-régressif : problème théorique sérieux (dans le cas le plus général, nécessité d'utiliser une structure *non causale* très difficile à mettre en oeuvre).

3.4 Applications

- Problèmes types
  → Amélioration d'images. Exemples : débruitage, rehaussement de contraste.
  → Segmentation : détection de contours, détection de régions.

- Outils
  → Outils empiriques. Exemple : égalisation d'histogramme.
  → Filtrage linéaire ou non linéaire.
  → Estimation.

- Contraintes

→ Mise en œuvre pour traitements rapides.
→ Stabilité des filtres.
→ Perception du bruit (oeil sensible aux *contrastes* fréquentiels).

- Remarque : filtrage équivalent à
  → choisir un masque ;
  → choisir l'opérateur (linéaire ou non) opérant sur le masque.

Exemple – Débruitage

- Filtrage passe-bas
  → Choix du support du masque (exemple : 3×3)
  → Choix d'une approximation discrète d'une RIF passe-bas sur le masque retenu (compromis suppression du bruit / lissage).
  → Choix du type d'opération (somme → filtrage linéaire, médiane...)

- Estimation MAP
  → A priori sur le bruit $f(z|x)$
  → A priori sur l'image $f(x)$
  → Maximisation...

Exemple – Rehaussement de contraste

- Filtrage
  → Mêmes choix que précédemment
  → RIF typique : $x \to x + \nabla^2 f$

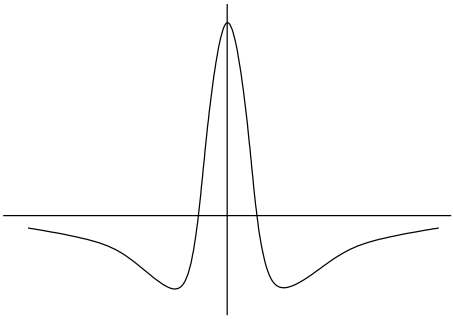
$$\text{Opérateur } \nabla^2 \ : \ \begin{bmatrix} -1 & -1 & -1 \\ -1 & 8 & -1 \\ -1 & -1 & -1 \end{bmatrix}.$$

- Estimation MAP
  → Mêmes choix que précédemment.
  → $f(x)$ doit représenter un *a priori* de « contraste élevé » (typiquement, champ de Markov).
  → Compromis complexité du modèle et de la méthode / qualité des résultats.

Exemple – Détection de contour

Contour : discontinuité dans l'image ↔ détection de composantes à haute fréquence (dans le plan de Fourier 2D...).

- Filtrage
  - → Mêmes choix que précédemment.
  - → RIF
    - ∗ Laplacien : trop sensible au bruit.
    - ∗ Laplacien de la gaussienne : filtrage passe-bas préalable.
    - ∗ Ajout d'une opération de détection par seuillage.



Laplacien de la gaussienne

$$\begin{bmatrix} 0 & 0 & -1 & 0 & 0 \\ 0 & -1 & -2 & -1 & 0 \\ -1 & -2 & 16 & -2 & -1 \\ 0 & -1 & -2 & -1 & 0 \\ 0 & 0 & -1 & 0 & 0 \end{bmatrix}$$

Masque 5×5 correspondant

  - → Possibilité de filtres directionnels.
  - → Difficulté : raccord des éléments de contours.

- Estimation
  - → Même choix que précédemment.
  - → $f(x)$ doit contenir un « processus de lignes ».
  - → Problème d'optimisation très complexe.
  - → Autre possibilité : contours actifs.
    - z : approximation du laplacien de l'image ;
    - x : contour doté d'une « énergie interne » assimilable à $f(x)$.
    - ∗ Modèle liant x à z.
    - ∗ Optimisation relativement complexe et lente.

Exemple – Segmentation

- Approches empiriques
  - → Seuillage (seuils déterminés à partir de l'histogramme). Analogue a un filtrage non linéaire avec masque 1×1.
  - → Agrégation de régions...

- Estimation
  - → Mêmes choix que précédemment.
  - → $f(x)$: modèle de l'image encourageant la formation de régions homogènes.

    Explicitement Champs de Markov pouvant être très complexes ; difficulté d'optimisation.

Implicitement Champs de Markov avec structure plus simple (pénalisation $l_2 l_1$ sur les différences premières). Optimisation plus facile.

→ Compromis difficulté de mise en oeuvre / volume de calculs / qualité des résultats à ajuster en fonction de l'objectif visé.

Chapitre 4

Traitement d'images médicales II : restauration, reconstruction

Plan

- Restauration d'images
- Minimisation itérative d'un critère quadrartique
- Reconstruction tomographique

4.1 Restauration d'images

- Justification pratique
  → Flou (sources ou détecteurs non ponctuels).
  → Imperfections de certains appareils d'imagerie.

- Cadre : signaux échantillonnés.

- Modèle adopté :
$$g(i_1, i_2) = (h * f)(i_1, i_2) + b(i_1, i_2)$$
  où
  g **:** image dégradée,
  h **:** PSF définie sur un *support fini* de taille $(2K+1, 2K+1)$,

f **:** image originale,

b **:** bruit.

Ainsi,

$$g(i_1, i_2) = \sum_{k_1=-K_1}^{K_1} \sum_{k_2=-K_2}^{K_2} h(k_1, k_2) f(i_1 - k_1, i_2 - k_2) + b(i_1, i_2). \tag{4.1}$$

- Résolution : approche de type estimation. Mise sous forme matricielle, puis application de la méthodologie générale.
  → Mise sous forme matricielle ; f, g et h rangés dans des vecteurs notés respectivement $\boldsymbol{f}$, $\boldsymbol{g}$ etr $\boldsymbol{h}$.

$$\begin{aligned} \boldsymbol{f} &= [\boldsymbol{f}_1^{\mathrm{t}} \mid \boldsymbol{f}_2^{\mathrm{t}} \mid \boldsymbol{f}_3^{\mathrm{t}} \mid \ \dots \ \mid \boldsymbol{f}_N^{\mathrm{t}}]^{\mathrm{t}} \\ \boldsymbol{g} &= [\boldsymbol{g}_1^{\mathrm{t}} \mid \boldsymbol{g}_2^{\mathrm{t}} \mid \boldsymbol{g}_3^{\mathrm{t}} \mid \ \dots \ \mid \boldsymbol{g}_N^{\mathrm{t}}]^{\mathrm{t}} \\ \boldsymbol{h} &= [\boldsymbol{h}_{-K_2}^{\mathrm{t}} \mid \boldsymbol{h}_{-K_2+1}^{\mathrm{t}} \mid \ \dots \ \mid \boldsymbol{h}_{K_2-1}^{\mathrm{t}} \mid \boldsymbol{h}_{K_2}^{\mathrm{t}}]^{\mathrm{t}} \end{aligned}$$

  où $\boldsymbol{f}_i$ (respectivement $\boldsymbol{g}_i$ et $\boldsymbol{h}_i$) désigne la colonne d'indice i de f (respectivement g et h). On peut réécrire le produit de convolution de (4.1) comme :

$$(h * f)(i_1, i_2) = \sum_{k_2=-K_2}^{K_2} (h(\cdot, k_2) * f(\cdot, i_2 - k_2))(i_1)$$

  Pour une valeur de i_2 fixée (donc pour une colonne donnée du résultat), $(h * f)(\cdot, i_2)$ apparaît donc comme la somme des produits de convolution :

$$(h * f)(\cdot, i_2) = \sum_{k_2=-K_2}^{K_2} (h(\cdot, k_2) * f(\cdot, i_2 - k_2))(.)$$

  En utilisant les quantités vectorielles définies plus haut, on a donc :

$$\boldsymbol{g}_{i_2} = \sum_{k_2=-K_2}^{K_2} \boldsymbol{H}_{k_2} \boldsymbol{f}_{i_2-k_2} + \boldsymbol{b}_{i_2} \tag{4.2}$$

  où chaque matrice $\boldsymbol{H}_{k_2}$ est une matrice de convolution 1D par la k_2^{e} colonne de h. Une telle matrice a donc une structure de Toeplitz. Par ailleurs, il est possible de réécrire l'áquation précédente sous la forme suivante :

$$\boldsymbol{g}_{i_2} = [\ \boldsymbol{0} \mid \ \cdots \ \mid \boldsymbol{0} \mid \boldsymbol{H}_{K_2} \mid \ \cdots \ \mid \boldsymbol{H}_{-K_2} \mid \boldsymbol{0} \mid \ \cdots \ \mid \boldsymbol{0}\ ]\, \boldsymbol{f} + \boldsymbol{b}$$

  Lorsque i_2 augmente de 1, donc lorsqu'on passe d'une colonne de $\boldsymbol{g}$ à la colonne suivante, (4.2) montre que les blocs non nuls de l'équation ci-dessus sont déplacés d'un bloc vers la droite. On peut donc écrire :

$$\boldsymbol{g} = \boldsymbol{H}\boldsymbol{f} + \boldsymbol{b},$$

avec

$\boldsymbol{H}$: matrice creuse, bloc-Toeplitz, chaque bloc etant lui-même Toeplitz ; taille $MN \times MN$.

$\boldsymbol{f}$ et $\boldsymbol{g}$: vecteurs de taille MN (aux effets de bord près).

→ Estimation de $\boldsymbol{f}$
- ∗ Approche MAP : $\boldsymbol{f}$ et $\boldsymbol{g}$ de taille comparable.
- ∗ Forme générale de l'estimateur :

$$\hat{\boldsymbol{f}} = (\boldsymbol{H}^{\mathrm{t}}\boldsymbol{H} + \lambda \boldsymbol{M})^{-1}\boldsymbol{H}^{\mathrm{t}}\boldsymbol{g}.$$

- ∗ Critère correspondant :

$$J(\boldsymbol{f}) = (\boldsymbol{g} - \boldsymbol{H}\boldsymbol{f})^{\mathrm{t}}(\boldsymbol{g} - \boldsymbol{H}\boldsymbol{f}) + \lambda \boldsymbol{f}^{\mathrm{t}}\boldsymbol{M}\boldsymbol{f}$$

- ∗ Problème de mise en oeuvre : taille des matrices !
- ∗ Solutions : moindres carrés récurrents asymptotiques ou approche plus générale : minimisation itérative d'un critère quadratique.

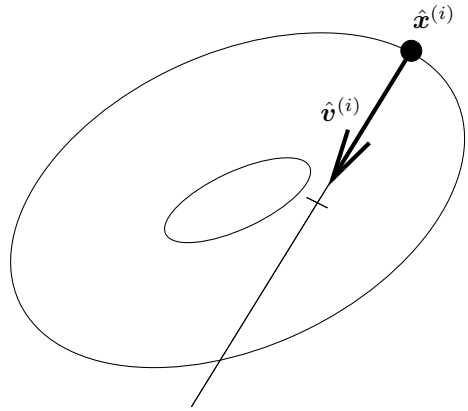
4.2 Minimisation itérative d'un critère quadratique

- Hypothèses
  → Modèle : $\boldsymbol{y} = \boldsymbol{H}\boldsymbol{x} + \boldsymbol{b}$ $\qquad$ $\boldsymbol{x}$ est l'inconnue vectorielle à estimer.
  → Critère : moindres carrés $J(\boldsymbol{x}) = (\boldsymbol{y} - \boldsymbol{H}\boldsymbol{x})^{\mathrm{t}}(\boldsymbol{y} - \boldsymbol{H}\boldsymbol{x})$.

- Approche :

$$\hat{\boldsymbol{x}}^{(i+1)} = \hat{\boldsymbol{x}}^{(i)} + \mu \boldsymbol{v}^{(i)}.$$

On peut montrer que si $\boldsymbol{v}^{(i)}$ est convenablement choisi de manière à minimiser le critère par rapport à μ, alors la procédure converge vers $\hat{\boldsymbol{x}}$.



→ Détermination de la valeur de μ.

On réécrit le critère en fonction de μ :

$$\begin{aligned} J(\mu) &= (\boldsymbol{y} - \boldsymbol{H}\hat{\boldsymbol{x}}^{(i)} - \mu \boldsymbol{H}\boldsymbol{v}^{(i)})^{\mathrm{t}}(\boldsymbol{y} - \boldsymbol{H}\hat{\boldsymbol{x}}^{(i)} - \mu H \boldsymbol{v}^{(i)}) \\ &= \mu^2 (\boldsymbol{v}^{(i)})^{\mathrm{t}} \boldsymbol{H}^{\mathrm{t}} \boldsymbol{H} \boldsymbol{v}^{(i)} - 2\mu (\boldsymbol{v}^{(i)})^{\mathrm{t}} \boldsymbol{H}^{\mathrm{t}} (\boldsymbol{y} - \boldsymbol{H}\hat{\boldsymbol{x}}^{(i)}) + (\boldsymbol{y} - \boldsymbol{H}\hat{\boldsymbol{x}}^{(i)})^{\mathrm{t}} (\boldsymbol{y} - \boldsymbol{H}\hat{\boldsymbol{x}}^{(i)}). \end{aligned}$$

Équation du deuxième degré en μ, qui atteint son minimum en

$$\mu^* = \frac{(\boldsymbol{v}^{(i)})^{\mathrm{t}} \boldsymbol{H}^{\mathrm{t}} (\boldsymbol{y} - \boldsymbol{H}\hat{\boldsymbol{x}}^{(i)})}{(\boldsymbol{v}^{(i)})^{\mathrm{t}} \boldsymbol{H}^{\mathrm{t}} \boldsymbol{H} \boldsymbol{v}^{(i)}},$$

d'où l'algorithme

$$\hat{\boldsymbol{x}}^{(i+1)} = \hat{\boldsymbol{x}}^{(i)} + \frac{(\boldsymbol{v}^{(i)})^{\mathrm{t}} \boldsymbol{H}^{\mathrm{t}} (\boldsymbol{y} - \boldsymbol{H}\hat{\boldsymbol{x}}^{(i)})}{(\boldsymbol{v}^{(i)})^{\mathrm{t}} \boldsymbol{H}^{\mathrm{t}} \boldsymbol{H} \boldsymbol{v}^{(i)}} \boldsymbol{v}^{(i)}.$$

→ Problème : choix de $\boldsymbol{v}^{(i)}$?

1. $\boldsymbol{v}^{(i)}$ nul partout sauf en la composante d'indice k, qui vaut 1.

   $$\boldsymbol{v}^{(i)} = \boldsymbol{u}_k = (0, 0 \cdots 0, 1, 0 \cdots 0)^{\mathrm{t}}$$

   C'est l'algorithme de Gauss-Seidel. Seule la composante d'indice k de $\hat{\boldsymbol{x}}^{(i)}$ est modifiée. On a

   $$\hat{x}_k^{(i+1)} = \hat{x}_k^{(i)} + \frac{\boldsymbol{c}_k^{\mathrm{t}} (\boldsymbol{y} - \boldsymbol{H}\hat{\boldsymbol{x}}^{(i)})}{\boldsymbol{c}_k^{\mathrm{t}} \boldsymbol{c}_k},$$

   avec

   $$\boldsymbol{c}_k = \boldsymbol{H}\boldsymbol{u}_k \qquad \boldsymbol{c}_k \ : \ k^{\mathrm{e}} \text{ colonne de } \boldsymbol{H}.$$

   Avantage : mise à jour de $\hat{\boldsymbol{x}}$ composante par composante.
   Inconvénient : calcul de $\boldsymbol{H}\hat{\boldsymbol{x}}^{(i)}$ à chaque itération (possibilité de simplification).

2. $\boldsymbol{v}^{(i)}$: vecteur colonne dont les composantes sont celles de la k^{e} *ligne* de $\boldsymbol{H}$; $\boldsymbol{v}^{(i)} = \boldsymbol{r}_k$.

   Algorithme :

$$\hat{\boldsymbol{x}}^{(i+1)} = \hat{\boldsymbol{x}}^{(i)} + \frac{\boldsymbol{r}_k^{\text{t}} \boldsymbol{H}^{\text{t}} (\boldsymbol{y} - \boldsymbol{H}\hat{\boldsymbol{x}}^{(i)})}{\boldsymbol{r}_k^{\text{t}} \boldsymbol{H}^{\text{t}} \boldsymbol{H} \boldsymbol{r}_k} \boldsymbol{r}_k,$$

   où

$$\boldsymbol{H}\boldsymbol{r}_k = \begin{bmatrix} \boldsymbol{r}_1^{\text{t}} \\ \boldsymbol{r}_2^{\text{t}} \\ \vdots \\ \boldsymbol{r}_k^{\text{t}} \\ \vdots \\ \boldsymbol{r}_{MN}^{\text{t}} \end{bmatrix} \boldsymbol{r}_k \approx \begin{bmatrix} 0 \\ 0 \\ \vdots \\ \|\boldsymbol{r}_k\|^2 \\ \vdots \\ 0 \end{bmatrix},$$

   pourvu que les lignes de $\boldsymbol{H}$ soient suffisamment orthogonales les unes aux autres. L'algorithme s'écrit alors

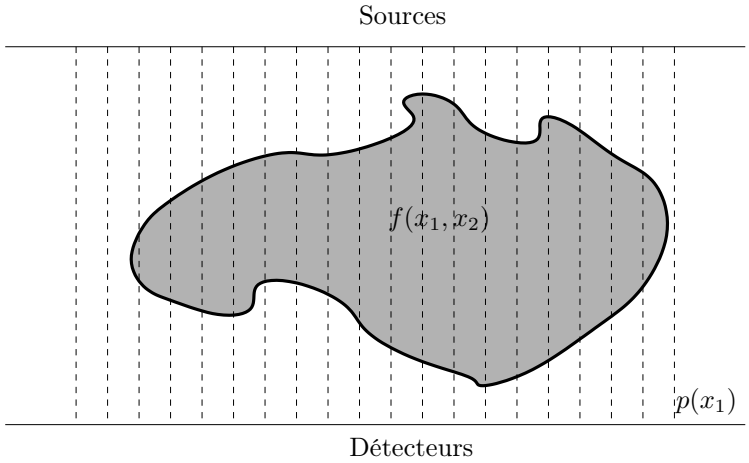
$$\begin{aligned} \hat{\boldsymbol{x}}^{(i+1)} &= \hat{\boldsymbol{x}}^{(i)} + \frac{\|\boldsymbol{r}_k\|^2 (\boldsymbol{y}_k - \boldsymbol{r}_k^{\text{t}} \hat{\boldsymbol{x}}^{(i)})}{\|\boldsymbol{r}_k\|^4} \boldsymbol{r}_k \\ &= \hat{\boldsymbol{x}}^{(i)} + \frac{y_k - \boldsymbol{r}_k^{\text{t}} \hat{\boldsymbol{x}}^{(i)}}{\|\boldsymbol{r}_k\|^2} \boldsymbol{r}_k. \end{aligned}$$

   Remarques sur l'algorithme :
   - Structure identique à celle des moindres carrés récurrents.
   - Mise à jour de $\hat{\boldsymbol{x}}$ tout entier à chaque itération.
   - Calcul d'un simple produit scalaire.

Remarques sur les deux techniques :
- Extension possible aux critères régularisés.
- Extension possible aux critères convexes.
- Possibilité d'introduire des coefficients de relaxation pour accélérer la convergence.

4.3 Reconstruction tomographique

- Formulation du problème :



Projection au sens mathématique :

$$p(x_1) = \int f(x_1, x_2) dx_2.$$

En faisant tourner le dispositif d'un angle θ,

$$p_\theta(r) = \int f(r\cos\theta - s\sin\theta, r\sin\theta - s\cos\theta) ds.$$

- Transformée de Radon

  Reconstruction tomographique : estimer f à partir de la mesure de plusieurs projections à des angles θ différents.

- Théorème de la tranche de Fourier

  En revenant à la projection d'angle θ et en prenant sa TF :

$$\begin{aligned} P(\nu_1) &= \int\int f(x_1, x_2) e^{-2i\pi\nu_1 x_1} dx_1 dx_2 \\ &= \int\int f(x_1, x_2) e^{-2i\pi\nu_1 x_1 + 0 x_2} dx_1 dx_2 \\ &= F(\nu_1, 0), \end{aligned}$$

  où $F(\nu_1, \nu_2)$ constitue la TF 2D de $f(x_1, x_2)$. En utilisant la propriété de rotation de la TF 2D,

$$P_\theta(\nu) = F(\nu\cos\theta, \nu\sin\theta).$$

  La TF 1D de la projection d'angle θ est égale à la TF 2D de f évaluée sur une droite passant par l'origine et faisant un angle θ avec l'axe $(O\nu_1)$.

- Reconstruction par interpolation dans le domaine de Fourier

  → Démarche : « remplissage » du domaine de Fourier 2D de f par TF 1D des projections à différents angles, puis « estimation » de $f(x_1, x_2)$ par TF 2D inverse du résultat.

  → Difficultés : « remplissage » inégal du domaine de Fourier 2D de f.
  - ∗ Hautes fréquences mal sensibilisées.
  - ∗ Impossibilité d'utiliser l'algorithme FFT sans interpolation.

- Reconstruction par rétro-projection filtrée.

Rappel Changement de variables en 2D

$$\begin{aligned} \begin{pmatrix} u \\ v \end{pmatrix} &= \psi \begin{pmatrix} s \\ t \end{pmatrix} \\ \int\int \phi(u,v)dudv &= \int\int \phi[\psi(s,t)]|J_\psi|dsdt, \end{aligned}$$

où $|J_\psi|$ est le déterminant du jacobien de ψ.

Démonstration Par définition de la TF 2D inverse,

$$f(x_1, x_2) = \int\int F(\nu_1, \nu_2) e^{2i\pi(\nu_1 x_1 + \nu_2 x_2)} d\nu_1 d\nu_2.$$

Changement de variables $\begin{pmatrix} \nu_1 \\ \nu_2 \end{pmatrix} = \begin{pmatrix} \nu\cos\theta \\ \nu\sin\theta \end{pmatrix}$, avec

$$J_\psi = \begin{bmatrix} \cos\theta & \nu\sin\theta \\ \sin\theta & \nu\cos\theta \end{bmatrix},$$

d'où $|J_\psi| = |\nu|$; on se ramène à

$$f(x_1, x_2) = \int_0^\pi d\theta \int_\infty P_\theta(\nu) e^{2i\pi\nu(x_1\cos\theta + x_2\sin\theta)} |\nu| d\nu.$$

Or, $\int_\infty P_\theta(\nu) e^{2i\pi\nu(x_1\cos\theta + x_2\sin\theta)} d\nu$ représente la TF inverse de $P_\theta(\nu)|\nu|$ au point $x = x_1\cos\theta + x_2\sin\theta$, ou encore $(P_\theta * h)(x)$, avec h la RI du filtre de RF $|\nu|$ et $x = x_1\cos\theta + x_2\sin\theta$.

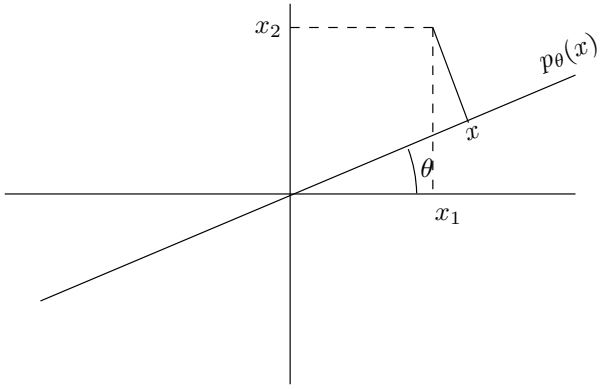
→ Obtention de $f(x_1, x_2)$ en (x_1, x_2) donné : pour chaque projection p_θ,

1. convoluer avec h,
2. déterminer la valeur en $x = x_1\cos\theta + x_2\sin\theta$, puis

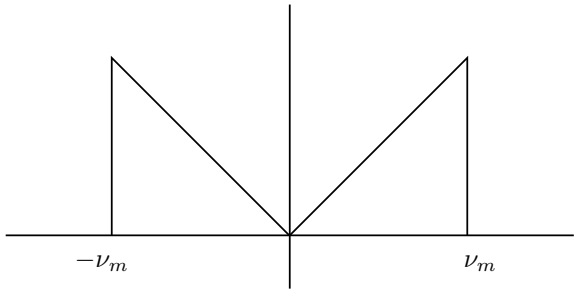
3. sommer pour toutes les valeurs de θ.

→ Reconstruction de $f(x_1, x_2)$: pour chaque projection,

1. convoluer avec h,
2. « répandre » selon les perpendiculaires.



→ Le filtre de RF $|\nu|$ n'est pas physique. On le remplace par le filtre de RF



avec $\nu_m = \frac{1}{2T_e}$, où T_e est la période d'échantillonnage des projections.

Mais : contradiction théorique (f à bande et à support limités).

- Reconstruction selon une approche « estimation » (méthodes de type ART)

→ Modèle adopté :

$$\boldsymbol{p} = \boldsymbol{H}\boldsymbol{f} + \boldsymbol{b}$$

$\boldsymbol{p}$ **:** projections concaténées en un vecteur.
$\boldsymbol{H}$ **:** matrice de projection aux différents angles.

→ Méthode ART « de base » : minimisation itérative d'un critère des moindres

carrés.

$$\hat{\boldsymbol{f}}^{(i+1)} = \hat{\boldsymbol{f}}^{(i)} + \alpha \frac{p_k - \boldsymbol{r}_k^{\mathrm{t}} \hat{\boldsymbol{f}}^i}{\|\boldsymbol{r}_k\|^2} \boldsymbol{r}_k$$

Mais : problème mal posé et pas d'*a priori*
$\Rightarrow$ Divergence lorsque $i \to \infty$.

$\to$ Méthode ART régularisée :

$$\begin{aligned} J(\boldsymbol{f}) &= (\boldsymbol{p} - \boldsymbol{H}\boldsymbol{f})^{\mathrm{t}}(\boldsymbol{p} - \boldsymbol{H}\boldsymbol{f}) + \lambda \boldsymbol{f}^{\mathrm{t}}\boldsymbol{f} \\ \hat{\boldsymbol{f}}^{(i+1)} &= \hat{\boldsymbol{f}}^{(i)} + \alpha \frac{p_k - \boldsymbol{r}_k^{\mathrm{t}} \hat{\boldsymbol{f}}^{(i)} \left(1 + \frac{\lambda}{\|\boldsymbol{r}_k\|^2}\right)}{\|\boldsymbol{r}_k\|^2 + \lambda} \boldsymbol{r}_k \end{aligned}$$

$\to$ Utilisation d'un algorithme de type Gauss-Seidel (critère régularisé) :

$$\hat{\boldsymbol{f}}^{(i+1)} = \hat{\boldsymbol{f}}^{(i)} + \alpha \frac{\boldsymbol{c}_k^{\mathrm{t}}[\boldsymbol{p} - \boldsymbol{H}\hat{\boldsymbol{f}}^{(i)} - \lambda \hat{\boldsymbol{f}}^{(i)}]}{\boldsymbol{c}_k^{\mathrm{t}}\boldsymbol{c}_k + \lambda} \boldsymbol{u}_k,$$

où

$\boldsymbol{c}_k$ **:** k^{e} colonne de $\boldsymbol{H}$;

$\boldsymbol{u}_k$ **:** k^{e} vecteur canonique.

Nécessité d'utiliser le caractère creux de $\boldsymbol{H}$!

$\to$ Extensions :

* Pénalisation par fonction $l_2 l_1$ des différences premières ou secondes.
* Autres types d'*a priori* non convexes (problèmes de minimisation...).

$\to$ Mérites des différentes approches :

Rétroprojection OK si
* Nombre de projections élevé ;
* Projections régulièrement distribuées de 0 à 180° ;
* Bruit faible.

Méthodes régularisées nécessaires dans les autres cas.

Méthodes avec pénalisation non quadratique utiles pour obtention d'images précises (pas encore utilisables de manière routinière).

Chapitre 5

Reconnaissance de formes I : approches statistiques

Plan

- Présentation du problème ; exemples.
- Définitions.
- Classification :
  → Approche empirique.
  → Approche bayésienne.
- Estimation des caractéristiques des classes.

5.1 Présentation du problème

Exemple 1 – Évaluation du risque de maladie vasculaire dans une population

- Existence de données épidémiologiques portant sur de multiples facteurs :
  → paramètres morphologiques,
  → paramètres géographiques,
  → historique médical,
  → mode de vie,
  → etc.

- Démarche pour évaluer le risque couru par un individu.

→ Fixer les objectifs (classes) : risque faible, moyen, élevé...
→ Déterminer un certain nombre d'attributs à partir desquels la classification se fera. Exemples :
  * taille,
  * âge,
  * poids,
  * lieu de résidence (ville, banlieue, milieu rural, régions nordiques...),
  * nombre d'heures d'activité physique par semaine,
  * consommation quotidienne moyenne de tabac, d'alcool...
→ Analyser les données existantes pour déterminer comment se comportent les attributs dans chaque classe (sans risque, risque léger, risque élevé).
→ En fonction de cette analyse, classer la population à évaluer.

Exemple 2 – Classification automatique de mammogrammes

- Existence de bases de données mammographiques déjà examinées par des experts (présence de tumeurs bénignes, malignes, mammogrammes normaux).

- Objectif : aider le praticien à détecter les mammogrammes suspects.

- Démarche :
  → Fixer les objectifs (deux classes, ou plus ?).
  → Déterminer des attributs dans les mammogrammes qui soient caractéristiques des classes choisies. Difficultés :
    * variabilité des mammogrammes,
    * fiabilité des traitements permettant le calcul des attributs,
    * immunité aux questions de normalisation, d'orientation...
  → Analyser le comportement des attributs dans chaque classe choisie, à partir des bases de données existantes.
  → En fonction des résultats de cette analyse, classer les mammogrammes à traiter.

Exemple 3 – Classification automatique des séquences ECG

- Objectifs possibles : surveillance de patient, aide au diagnostic.
- Démarche : analogue à celle de l'exemple 2.
- Particularité : structure forte dans un ECG (présence de plusieurs ondes, durée relativement stable de chacune d'elles).
  → Les attributs pourraient refléter cette structure (notion de chaîne, d'arbre).

Points importants
- Choix des objectifs (classes).

- Adéquation des attributs.
- Redondance / robustesse des attributs.

5.2 Définitions

Candidat Élément à classifier.

Attribut Grandeur (continue ou discrète) que l'on peut associer à un candidat.

À tout candidat est associé un vecteur d'attributs $\boldsymbol{x} = (x_1, x_2 \dots x_N)^{\mathrm{t}}$.

Classe ou hypothèse Ensemble auquel peut appartenir tout candidat et qui influe sur la valeur de ses attributs. On les désigne par $\omega_i,\ 1 \leq i \leq I$.

Ensemble d'apprentissage Ensemble de candidats pour lesquels le résultat de la classification est connu. Cet ensemble sert généralement à déterminer le comportement des attributs dans chaque classe et à valider le classificateur proposé.

5.3 Classification

- Hypothèses : on suppose que
  1. le nombre de classes est connu et que les classes sont déterminées — on les désigne par $\omega_1, \omega_2, \dots \omega_I$;
  2. la composition du vecteur d'attributs est fixée ;
  3. une certaine information sur le comportement des attributs en fonction de la classe à laquelle appartient le candidat est disponible (exemples : moyenne, distribution...) soit à partir d'une modélisation, soit à partir d'une analyse de l'ensemble d'apprentissage.

- Remarque : on n'aborde pas ici le cas de la classification « aveugle » (pas d'ensemble d'apprentissage, nombre de classes inconnu...).

- Approche empirique : distance minimale.

  On suppose que
  → le nombre I de classes est connu ;
  → la composition du vecteur d'attributs est connue ;
  → chaque classe ω_i est caractérisée par la moyenne du vecteur d'attributs lorsque le candidat appartient à cette classe, cette moyenne étant connue.

Remarque : dans l'espace des attributs,
→ à chaque classe correspond un nuage de points ;
→ la moyenne de la classe peut être estimée empiriquement par la moyenne de ce nuage de points.
Pour tout candidat de vecteur d'attributs $\boldsymbol{x}$:
→ Calcul d'une distance $d_i(\boldsymbol{x}) = d(\boldsymbol{x} - \boldsymbol{x}_i)$, avec $\boldsymbol{x}_i$ la moyenne de la classe ω_i.
→ Classification : minimum de distance.

$$\text{Classe } \omega_{i_0}; \forall i : d_{i_0}(\boldsymbol{x}) \leq d_i(\boldsymbol{x})$$

Remarques
→ Le choix de la distance est laissé à l'utilisateur.
→ Les points de l'espace des attributs tels que $d_i(\boldsymbol{x}) = d_j(\boldsymbol{x})$ sont les « frontières de décision » qui partitionnent l'espace des attributs en I régions correspondant à chacune des classes.
→ Si $d_i(\boldsymbol{x}) = \|\boldsymbol{x} - \boldsymbol{x}_i\|$ (norme euclidienne), les frontières de décision sont des hyperplans.

- Approche bayésienne.

→ Hypothèses : elles sont plus fortes que dans le cas précédent. On suppose connus
∗ le nombre de classes I et les classes $\omega_1, \omega_2 \ldots \omega_I$;
∗ la composition du vecteur d'attributs ;
∗ la distribution des vecteurs d'attributs dans chaque classe $f(\boldsymbol{x}|\omega_i)$;
∗ la probabilité a priori de chaque classe $f(\omega_i)$.

→ Cas simple : classification par MAP.

On cherche l'indice i de la classe qui maximise

$$f(\omega_i|\boldsymbol{x}) \propto f(\boldsymbol{x}|\omega_i) f(\omega_i)$$

pour chaque vecteur d'attributs $\boldsymbol{x}$ associé à un candidat.

Les quantités $f(\omega_i|\boldsymbol{x})$ jouent le rôle des distances $d_i(\boldsymbol{x})$ (fonctions de décision) dans le cas précédent.

Limitation : le « prix à payer » pour une mauvaise classification n'est pas toujours le même.
∗ Estimation à minimum de risque.

→ Cas général : estimation d'une variable θ par minimisation d'un coût (ou risque) bayésien.

θ **:** paramètre à estimer ;

z **:** observations.

On suppose connue $f(\theta|z)$.

Pour deux valeurs de θ, on définit le coût

$$C(\theta_1) = d(\theta_1, \theta),$$

où $C(\theta_1)$ constitue le coût de choisir θ_1 alors que l'inconnue vaut θ ; $d(\theta_1, \theta)$ n'est pas nécessairement symétrique.

Approche pour l'estimation de θ : minimisation du coût moyen (ou risque bayésien).

$$\mathrm{E}[C(\hat{\theta})|z] = \int d(\hat{\theta}, \theta) f(\theta|z) d\theta$$

Lorsque θ est à valeurs continues, deux choix classiques pour $C(\theta_1) = d(\theta_1, \theta)$:

1. $d(\theta_1, \theta) = 1 - \delta(\theta_1 - \theta)$
   Dans ce cas,

$$\begin{aligned} \mathrm{E}[C(\hat{\theta})|z] &= \int (1 - \delta(\theta_1 - \theta)) f(\theta|z) d\theta \\ &= \underbrace{\int f(\theta|z) d\theta}_{1} - f(\hat{\theta}|z). \end{aligned}$$

   Minimiser le coût bayésien ci-dessus revient à maximiser $f(\hat{\theta}|z)$. On retrouve l'estimateur MAP.
2. $d(\theta_1, \theta) = \|\theta_1 - \theta\|^2 = (\theta_1 - \theta)^{\mathrm{t}}(\theta_1 - \theta)$

$$\mathrm{E}[C(\hat{\theta})|z] = \int (\hat{\theta} - \theta)^{\mathrm{t}} (\hat{\theta} - \theta) f(\theta|z) d\theta$$

   En évaluant le gradient par rapport à $\hat{\theta}$ de l'expression ci-dessus (dérivation sous le signe somme) et en écrivant que ce gradient s'annule, on obtient

$$\begin{aligned} & \hat{\theta} \int f(\theta|z) d\theta = \int \theta f(\theta|z) d\theta \\ \Longleftrightarrow \quad & \hat{\theta} = E[\theta|z]. \end{aligned}$$

   On retrouve ici l'estimateur de la moyenne *a posteriori*.

→ Application à la classification.

Ici, la grandeur à estimer (indice i de la classe) est discrète. On définit empiriquement $C(\hat{\omega}_i) = d(\hat{\omega}_i, \omega_j)$ pour chacun des couples (i, j).

Pour chaque classe ω_i, le risque bayésien a pour expression

$$\mathrm{E}[C(\hat{\omega}_i)|\boldsymbol{x}] = r_i(\boldsymbol{x}) = \sum_{j=1}^{I} d(\hat{\omega}_i, \omega_j) f(\boldsymbol{x}|\omega_j) f(\omega_j),$$

où $r_i(\boldsymbol{x})$ constitue la nouvelle fonction de décision.

En général,
∗ calcul aisé ;
∗ choix de $d(\hat{\omega}_i, \omega_j)$, $f(\boldsymbol{x}|\omega_j)$, $f(\omega_j)$ plus délicat.

→ Cas particulier : $f(\boldsymbol{x}|\omega_j)$ gaussienne.

∗ Pour une fonction de coût quelconque, pas de simplification.
∗ Pour une fonction de coût $1 - \delta_{ij}$ (classification MAP),

$$r_i(\boldsymbol{x}) = \underbrace{f(\boldsymbol{x}|\omega_i)}_{\mathcal{N}(\boldsymbol{m}_i, \boldsymbol{R}_i)} \underbrace{f(\omega_i)}_{\text{constante}} .$$

En prenant le logarithme de $r_i(\boldsymbol{x})$, on montre aisément que les frontières de décision sont des quadriques.

De plus, si $\forall i : \boldsymbol{R}_i = \boldsymbol{I}$, alors

$$\log r_i(\boldsymbol{x}) \propto \boldsymbol{m}_i^{\mathrm{t}} \boldsymbol{x} - \frac{1}{2} \boldsymbol{m}_i^{\mathrm{t}} \boldsymbol{m}_i,$$

⇒ les frontières de décision sont des hyperplans.

On retrouve la classification par minimum de distance.

5.4 Estimation des caractéristiques des classes

Difficulté : comment spécifier

$$f(\boldsymbol{x}|\omega_i), \quad f(\omega_i)?$$

$f(\omega_i)$ **:** scalaire, probabilité d'occurrence de la classe ω_i. Pose en général peu de problèmes à spécifier, soit à partir de considérations générales, soit à partir de l'ensemble d'apprentissage.

$f(\boldsymbol{x}|\omega_i)$ **:** beaucoup plus difficile.

En général, utilisation de l'ensemble d'apprentissage $\Rightarrow$ estimation d'une loi de probabilité à partir d'un ensemble de réalisations.

- Approche non paramétrique (estimateurs à noyaux).
- Approche paramétrique. Souvent difficile.

Exemple – Estimation des paramètres d'une loi gaussienne par maximum de vraisemblance

Maximisation de $f(\boldsymbol{x}|\boldsymbol{m}, \boldsymbol{R})$ par rapport à $(\boldsymbol{m}, \boldsymbol{R})$. Plusieurs réalisations indépendantes $\boldsymbol{x}_i$:

$\boldsymbol{m}$ **:** toujours possible (moyenne empirique) ;

$\boldsymbol{R}$ **:** selon le cas,

dimension 1 : variance empirique ;

dimension ≥ 1 **:** ? ? ?

$\Rightarrow$ On se rabat sur des estimateurs empiriques (des moments, par exemple).

Chapitre 6

Reconnaissance de formes II : réduction du nombre des attributs ; approche par réseaux de neurones

Plan

- Réduction du nombre des attributs.
  - → Position du problème.
  - → Minimum d'entropie.
  - → Discriminant linéaire de Fisher.
  - → Décomposition en composantes principales (développement de Karhunen-Loêve).
- Classification par réseaux de neurones.
  - → Position du problème.
  - → Perceptron : définition, entraînement.
  - → Réseaux à couches multiples : définition, entraînement par rétro-propagation.

6.1 Réduction du nombre d'attributs

- Position du problème

Vecteurs d'attributs $\boldsymbol{x}$ de taille N ; I classes ω_i, $1 \leq i \leq I$.

Si N est grand,
$\rightarrow$ lourdeur des calculs ;
$\rightarrow$ possible détérioration des performances du classificateur.
Objectifs : réduction de la taille des attributs $N \rightarrow M < N$.
$\Rightarrow$ Préservation, voire amélioration de la classification.

Cadre : nouveau vecteur d'attributs $\boldsymbol{y}$ de taille M tel que $\boldsymbol{y} = \boldsymbol{T}\boldsymbol{x}$, où $\boldsymbol{T}$ désigne une transformation *linéaire*.

Notations : pour chaque classe,

$\boldsymbol{m}_i$ **:** moyenne de la classe,

$\boldsymbol{R}_i$ **:** matrice de covariance de la classe.

Ces matrices sont supposées connues (ou peuvent être estimées empiriquement sur un ensemble d'entraînement).

- Approche par maximum d'entropie

  $\rightarrow$ Idée générale de l'approche : déterminer $\boldsymbol{T}$ de manière à minimiser la dispersion de chaque classe.

  Mesure de la dispersion : entropie H. Pour une loi de probabilité f,

  $$H(f) = -E[\log f(x)] = -\int f(x) \log f(x) dx.$$

  $*$ Exemple : si f est la loi uniforme de support D.

  $$\Rightarrow H(f) = \log D,$$

  donc croît avec D.

  $*$ Propriété : si f est une loi gaussienne multivariée $\mathcal{N}(\boldsymbol{m}, \boldsymbol{R})$,

  $$H(f) = \frac{N}{2} + \frac{1}{2} \log |R| + \frac{N}{2} \log 2\pi,$$

  où N est la taille de la variable.

  $\rightarrow$ Hypothèses :

  $f(\boldsymbol{x}|\omega_i)$ gaussienne $\mathcal{N}(\boldsymbol{m}_i, \boldsymbol{R}_i)$.

  De par la propriété précédente, $H_i = H(f(\boldsymbol{x}|\omega_i))$ ne dépend que de $\boldsymbol{R}_i$. Si les $\boldsymbol{R}_i$ sont différents les uns des autres, pas de résultat simple.

  $*$ On suppose que tous les R_i sont identiques $\forall i \in \{1, 2 \ldots I\}$: $\boldsymbol{R}_i = \boldsymbol{R}$. Statistiquement, les classes ne se différencient que par leur moyenne.

→ Développement.

On a $\boldsymbol{y} = T\boldsymbol{x}$.

Dans chaque classe, $f(\boldsymbol{y}|\omega_i)$ est la densité associée à la loi $\mathcal{N}(\boldsymbol{T}\boldsymbol{m}_i, \boldsymbol{T}\boldsymbol{R}\boldsymbol{T}^{\mathrm{t}})$. Après transformation,

$$\begin{aligned} H_i &= \frac{M}{2} + \frac{1}{2}\log|\boldsymbol{T}\boldsymbol{R}\boldsymbol{T}^{\mathrm{t}}| + \frac{M}{2}\log 2\pi \\ &= \frac{M}{2}(1+\log 2\pi) + \frac{1}{2}\sum_{j=1}^{M}\log\lambda_j, \end{aligned}$$

où les λ_j sont les valeurs propres de $\boldsymbol{T}\boldsymbol{R}\boldsymbol{T}^{\mathrm{t}}$.

On peut montrer que, sous condition de normalisation de $\boldsymbol{T}$, H_i est minimisé en prenant pour $\boldsymbol{T}$ la matrice dont les M lignes sont les vecteurs propres de $\boldsymbol{R}$ associés à ses M valeurs propres les plus faibles.

- Discriminant linéaire de Fisher

→ Idée générale : déterminer $\boldsymbol{T}$ de manière à maximiser le rapport entre dispersion inter-classes et dispersion intra-classe.

→ Hypothèses : la moyenne $\boldsymbol{m}_i$ et la matrice de covariance $\boldsymbol{R}_i$ de chaque classe sont connues.

On définit

$$\begin{aligned} \boldsymbol{R} &= \sum_{i=1}^{I}\boldsymbol{R}_i, \\ \boldsymbol{m} &= \frac{1}{I}\sum_{i=1}^{I}\boldsymbol{m}_i. \end{aligned}$$

→ Mesures de dispersion :
  * Intra-classe : $\boldsymbol{R}$.
  * Inter-classes : dispersion par rapport à la moyenne

$$\boldsymbol{B} = \sum_{i=1}^{I}(\boldsymbol{m}_i - \boldsymbol{m})(\boldsymbol{m}_i - \boldsymbol{m})^{\mathrm{t}}$$

(matrice de rang $I-1$ au plus).

→ Développement.

Le critère à maximiser doit être scalaire et lié à $\boldsymbol{R}$ et $\boldsymbol{B}$ (après transformation). Plusieurs choix sont possibles. Ici, on adopte

$$J = \frac{|\boldsymbol{B}|}{|\boldsymbol{R}|},$$

qui conduit à une solution relativement simple.

Après transformation $\boldsymbol{T}$,

$$\begin{array}{rcl} \boldsymbol{R}_i \to \boldsymbol{T}\boldsymbol{R}_i\boldsymbol{T}^{\mathrm{t}} & \Rightarrow & \boldsymbol{R} \to \boldsymbol{T}\boldsymbol{R}\boldsymbol{T}^{\mathrm{t}} \\ \boldsymbol{m}_i \to \boldsymbol{T}\boldsymbol{m}_i & & \\ \boldsymbol{m} \to \boldsymbol{T}\boldsymbol{m} & \Rightarrow & \boldsymbol{B} \to \boldsymbol{T}\boldsymbol{B}\boldsymbol{T}^{\mathrm{t}}. \end{array}$$

On cherche donc $\boldsymbol{T}$ qui maximise

$$J = \frac{|\boldsymbol{T}\boldsymbol{B}\boldsymbol{T}^{\mathrm{t}}|}{|\boldsymbol{T}\boldsymbol{R}\boldsymbol{T}^{\mathrm{t}}|}.$$

$\boldsymbol{B}$ étant de rang $I-1$, M doit être au plus égal à $I-1$, sinon $J=0$.

On montre que l'on doit prendre pour $\boldsymbol{T}$ une matrice dont les lignes sont les vecteurs propres de $\boldsymbol{R}^{-1}\boldsymbol{B}$. On a donc

$$\boldsymbol{R}^{-1}\boldsymbol{B}\boldsymbol{T}^{\mathrm{t}} = \boldsymbol{T}^{\mathrm{t}}\boldsymbol{\Lambda},$$

où $\boldsymbol{\Lambda}$ est la matrice diagonale des valeurs propres associées respectivement aux colonnes de $\boldsymbol{T}^{\mathrm{t}}$ (identiques aux lignes de $\boldsymbol{T}$). Ainsi,

$$J = \frac{|\boldsymbol{T}\boldsymbol{R}\boldsymbol{T}^{\mathrm{t}}\boldsymbol{\Lambda}|}{|\boldsymbol{T}\boldsymbol{R}\boldsymbol{T}^{\mathrm{t}}} = |\boldsymbol{\Lambda}| = \prod \lambda_i.$$

Pour une valeur de M donnée, il suffit de sélectionner les vecteurs propres associés aux M plus grandes valeurs propres de $\boldsymbol{R}^{-1}\boldsymbol{B}$ et de former la matrice $\boldsymbol{T}$ correspondante.

- Analyse en composantes principales.

  → Idée générale : deux interprétations.
  - ∗ Meilleure approximation de $\boldsymbol{x}$ par $\boldsymbol{y}$ au sens de l'erreur quadratique moyenne.
  - ∗ Sélection de $\boldsymbol{y}$ de manière à maximiser la dispersion (pas seulement inter-classes).
  - ∗ On travaille sur $f(\boldsymbol{x})$, et non sur $f(\boldsymbol{x}|\omega_i)$!

- Développement :

$$\boldsymbol{y} = \boldsymbol{T}\boldsymbol{x}$$

Mesure de dispersion : $\boldsymbol{R_y} = |\boldsymbol{T}\boldsymbol{R}\boldsymbol{T}^{\mathrm{t}}|$ (identique à l'entropie sous hypothèse gaussienne).

$$|\boldsymbol{R_y}| = \prod_{m=1}^{M} \lambda_m^{\boldsymbol{y}},$$

où les $\lambda_m^{\boldsymbol{y}}$ sont les valeurs propres de $\boldsymbol{T}\boldsymbol{R}\boldsymbol{T}^{\mathrm{t}}$.

Sous condition de normalisation : $\boldsymbol{T}$ formé en prenant les vecteurs propres de $\boldsymbol{R}$ associés aux M plus grandes valeurs propres de $\boldsymbol{R}$,

$$T = \begin{bmatrix} — \boldsymbol{\phi}_1 — \\ — \boldsymbol{\phi}_2 — \\ \vdots \\ — \boldsymbol{\phi}_M — \end{bmatrix}.$$

$\boldsymbol{R}$ symétrique et définie positive $\Rightarrow$ les $\boldsymbol{\phi}_i$ sont orthogonaux (on les prend orthonormés). On a alors

$$\boldsymbol{R_y} = \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_M \end{bmatrix}.$$

Dans ce contexte, $\boldsymbol{y}$ est la projection de $\boldsymbol{x}$ sur le sous-espace engendré par ces vecteurs propres, donc la meilleure approximation de $\boldsymbol{x}$ au sens de l'EQM.

De plus, le choix de la base de vecteurs propres décorrèle les composantes de $\boldsymbol{y}$ ($\boldsymbol{R_y}$ diagonale).

- Application pratique.

  → Décomposition en valeurs propres / vecteurs propres.
  → Sélection des vecteurs propres associés aux valeurs propres supérieures à un seuil pour former T.

- *Mais*, pas d'effet garanti sur la séparation entre classes.

6.2 Classification par réseau de neurones

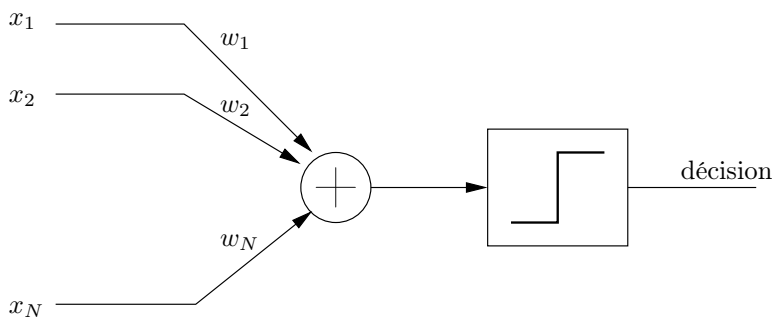
- Objectif : mise en œuvre simple d'une méthode de classification par minimum de distance quadratique.

- Perceptron.

  → Hypothèses :

∗ Deux classes.
∗ Distance quadratique ⇔ surface de décision est un hyperplan.
Classificateur $\boldsymbol{w}^{\mathrm{t}}\boldsymbol{x} \gtrless t$.

→ Mise en œuvre.



→ Détermination des paramètres $(\boldsymbol{w}, t)$?

Par convention : deux sorties possibles pour le perceptron, soit 0 ou 1 (selon la classe).

Modèle :

$$\boldsymbol{x}^{\mathrm{t}}\boldsymbol{w} - t = r \Leftrightarrow \boldsymbol{y}^{\mathrm{t}}\boldsymbol{v} = r,$$

avec $\boldsymbol{y} = \begin{pmatrix} \boldsymbol{x} \\ 1 \end{pmatrix}$ et $\boldsymbol{v} = \begin{pmatrix} \boldsymbol{w} \\ -t \end{pmatrix}$.

À partir d'un ensemble d'entraînement, on forme un ensemble de vecteurs $\boldsymbol{y}_i$ connus et le vecteur $\boldsymbol{r}$ des sorties correspondantes. On voudrait

$$\begin{aligned} \begin{bmatrix} \boldsymbol{y}_1^{\mathrm{t}} \\ \boldsymbol{y}_2^{\mathrm{t}} \\ \vdots \\ \boldsymbol{y}_E^{\mathrm{t}} \end{bmatrix} \boldsymbol{v} &= \begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_E \end{bmatrix} = \boldsymbol{r} \\ \boldsymbol{Y}\boldsymbol{v} &= \boldsymbol{r} \end{aligned}$$

∗ par moindres carrés :

$$\Rightarrow \hat{\boldsymbol{v}} = (\boldsymbol{Y}^{\mathrm{t}}\boldsymbol{Y})^{-1}\boldsymbol{Y}^{\mathrm{t}}\boldsymbol{r}.$$

Dans la pratique : mise en œuvre sous forme de gradient.

$$\hat{\boldsymbol{v}}^{(i+1)} = \hat{\boldsymbol{v}}^{(i)} + \alpha \boldsymbol{y}_{i+1}(r_{i+1} - \boldsymbol{y}_{i+1}^{\mathrm{t}}\hat{\boldsymbol{v}}^{(i)}).$$

Divers choix possibles pour α, selon la difficulté du problème.

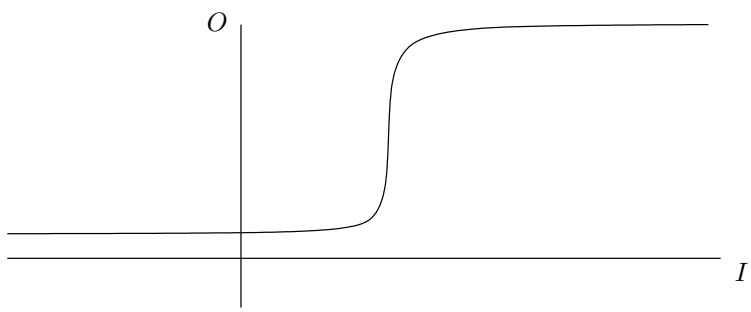
Approche par perceptron :

1. limitée en tant que telle ;
2. « brique de base » pour les réseaux de neurones.

- Réseaux à couches multiples.

→ Structure : voir figure 6.1

→ Quantités à fixer :
  * Nombre de couches ;
  * Nombre de cellules par couche ;
  * Fonction seuil $O = h(I)$.



→ Difficulté : évolution de l'ensemble des coefficients w_j.

Une solution : algorithme de rétro-propagation.

→ Remarque préliminaire : l'algorithme de mise à jour des coefficients du perceptron peut se mettre en œuvre sous la forme

$$\Delta v_n = -\alpha \frac{\partial E}{\partial v_n},$$

avec

$$E = \frac{1}{2}\left(r - \sum_{n=1}^{N} y_n v_n\right)^2$$

où l'indice i de la récurrence a été omis pour alléger les notations.

→ Idée : étendre le principe en parcourant le réseau de droite à gauche.

Dernière couche : pas de difficulté. On peut calculer toutes les entrées et on connaît les sorties désirées.

$$\Delta w_{qp} = \alpha \underbrace{(r_q - O_q)h'(I_q)}_{\delta_q} O_p,$$

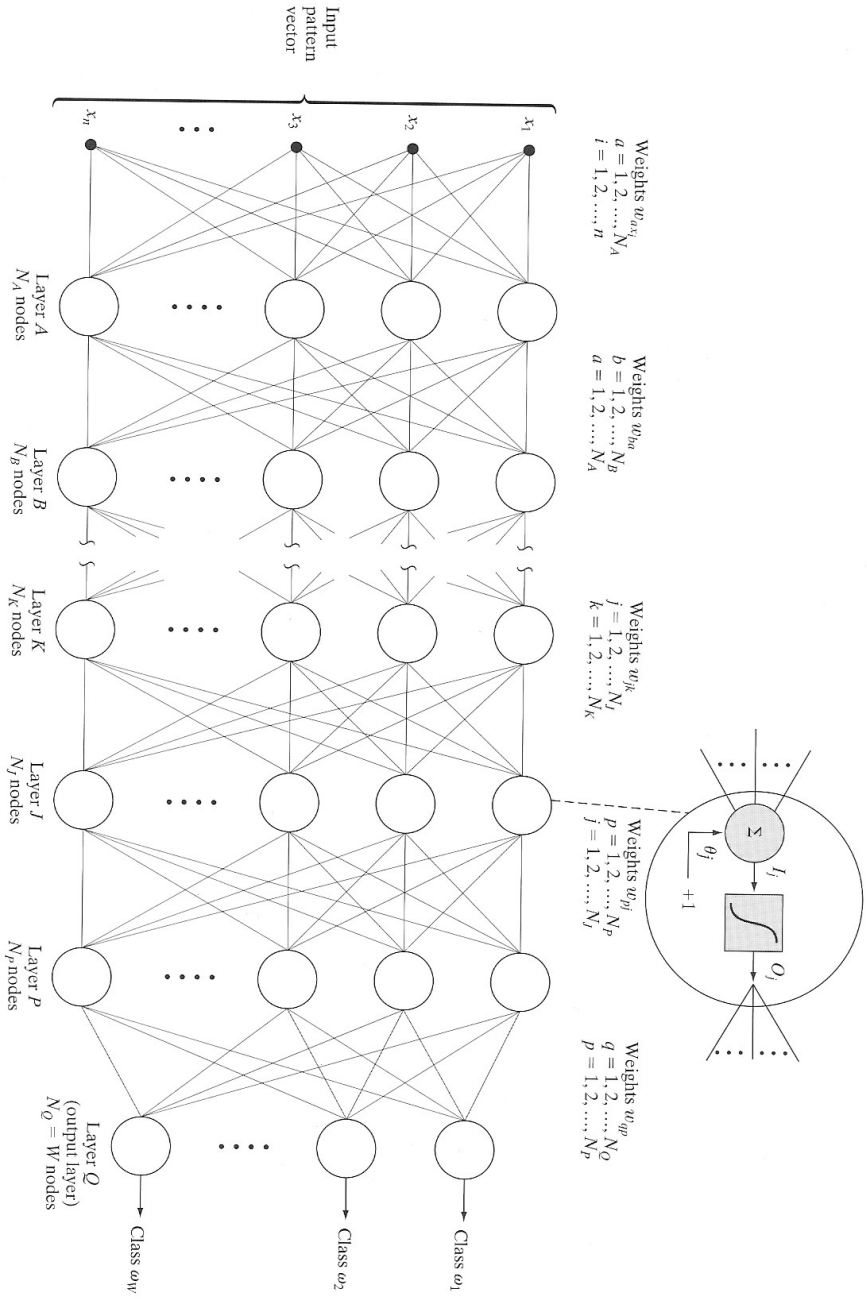


FIGURE 6.1 – Structure d'un réseau de neurones multicouches (d'après [1]).

avec

$$E_q = \frac{1}{2}\sum(r_q - O_q)^2.$$

Avant-dernière couche : on ne connaît pas la sortie désirée r_p !

$$E_p = \frac{1}{2}\sum_p(r_p - O_p)^2$$

Cependant, on peut montrer que

$$\delta_p = \frac{\partial E_p}{\partial I_p} = \sum_q \delta_q w_{qp}.$$

* $\Delta w_{pj} = \sum_q \delta_q w_{qp}$,
Cela permet de poursuivre la récurrence.

→ Avantages de l'approche :
* algorithme relativement simple ;
* souplesse de la structure.

→ Mais,
* les fonctions de décision sont des hyperplans par morceaux ;
* très lourd à entraîner (mauvaise sensibilité des paramètres internes, non-linéarités).

Chapitre 7

Compression de signaux biomédicaux

Plan

- Position du problème.
- Redondance liée à la source.
  - → Codage de Huffman.
  - → Codage RLE.
  - → Méthodes à base de dictionnaires — codage LZW.
- Redondance structurelle et liée à la perception.
  - → Codage prédictif, avec et sans perte.
  - → Décomposition sur des bases : méthodes à base de transformées, quantification.

7.1 Position du problème

Exemple Compression d'une image représentée avec 256 niveaux de gris.

Chaque pixel	1 octet = 8 bits
Taille totale	$N \times 8$ bits
	N : nombre de pixels

Idée générale : exploitation de la redondance pour réduire la taille totale (en nombre de bits) permettant de représenter l'image.

- Première approche : transformation image → suite d'octets (niveaux de gris, symboles...) et codage des plus fréquents avec un faible nombre de bits.

  Exploitation des statistiques d'apparition des symboles. Stockage
  → de la table de codage,
  → de l'image codée.
  $\boxed{\text{Redondance liée à la source}}$ (autres approches : RLE, dictionnaires...).

- Seconde approche : utilisation de la « structure » de l'image par une modélisation adéquate.

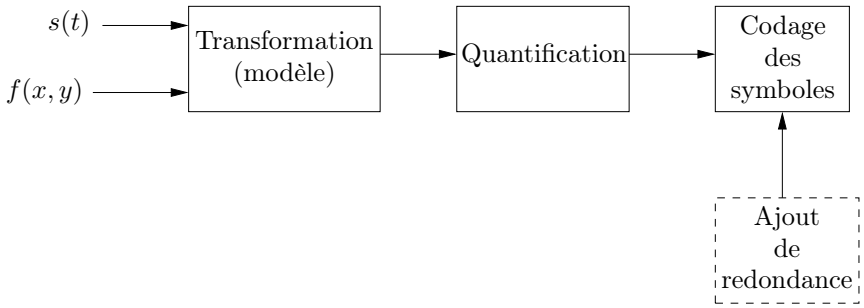
  Exemple Contenu spectral limité de l'image.

  → Décomposition de Fourier avec de nombreux coefficients nuls ou voisins de zéro.

  Codage sans perte (*lossless*) : tous les coefficients de Fourier sont conservés (gain au niveau du codage de la source).

  Codage avec perte (*lossy*) : élimination de coefficients faibles ou ayant peu d'influence sur la perception.
  → Quantification.

- Structure d'un système de codage.



  Ajout de redondance au niveau du codage des symboles : nécessaire pour protection contre erreurs de transmission et/ou de stockage.
  → Pas abordé ici.

  Facteur de compression :

$$\frac{\text{Nb. de bits de l'image ou signal original}}{\text{Nb. de bits de l'image ou signal codé}}$$

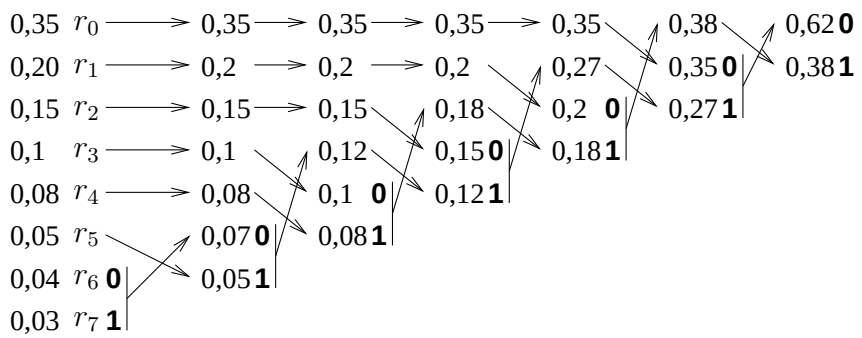
7.2 Redondance liée à la source

- Codage de Huffman : utilisation des propriétés statistiques de la suite de symboles.

  Codage des symboles les plus fréquents (probables) avec un petit nombre de bits.

  Exemple Huit niveaux $r_0 \to r_7 \Leftrightarrow 3$ bits par symbole si on utilise un codage sans compression.

  Construction du code d'Huffman pour les fréquences suivantes.



  Encodage :

r_0 : 00	r_4 : 111
r_1 : 10	r_5 : 0111
r_2 : 010	r_6 : 01100
r_3 : 110	r_7 : 01101

- Propriétés des codes d'Huffman

  → Séparabilité ⇒ pas de nécessité de coder la séparation entre symboles.

  → « Quasi-optimalité » au sens de la théorie de l'information.

  Source caractérisée par $p(r_i), 0 \leq i \leq I-1$ (distribution IID) ;

$$H = -\sum_{i=0}^{I-1} p(r_i) \log_2 p(r_i)$$

  représente l'information moyenne par symbole, en nombre de bits. Borne

inférieure du nombre de bits par symbole qu'il est possible d'atteindre *à condition que les hypothèses soient vérifiées.*

Dans un cas pratique...
→ $p(r_i)$?
→ Indépendance entre symboles ?
On montre néanmoins que le codage de Huffman permet d'approcher H d'aussi prêt que l'on veut, à condition toutefois de coder des groupes de symboles de plus en plus longs.

- Quantités à stocker.

  → Les symboles et leur probabilité (pour que le décodeur puisse construire les codes de chaque symbole) → en-tête.
  → La suite des symboles codés.

- Codage RLE (*run-length encoding*).

  Présenté pour son importance historique. Type intermédiaire entre modélisation et codage de symboles.

Exemple Signal (image) binaire.

Plutôt que de transmettre la suite de bits, on recherche les séquences de symboles identiques dont on transmet la longueur (utilité pour la transmission par télécopie).

Remarques :

→ Longueur maximale d'une séquence → possibilité de séquences de longueur nulle.

→ Pour des images sur 256 niveaux de gris ou couleur (3×256 niveaux) : chaque plan est codé.
  * Les bits de fort poids sont codés efficacement.
  * Les bits de faible poids sont codés moins efficacement.

→ À stocker : informations sur la taille de l'image, valeur du niveau du premier bit, suite des longueurs des séquences homogènes.

- Codage à base de dictionnaires.

  → Approche générale :

  * Repérer dans la suite de symboles les séquences qui se répètent fréquemment pour construire un dictionnaire.
  * Coder les symboles et les mots du dictionnaires.
  * Attention : il faut stocker le dictionnaires.
    - Compromis entre la taille du dictionnaire, l'augmentation de la taille des codes et la gain lié à l'utilisation du dictionnaire.

→ Approche plus économique : LZW (Lempel-Ziv-Welch).

Construction dynamique du dictionnaire, au codage *comme au décodage*.

Nécessité d'un tampon pour construction dynamique du dictionnaire (mémoire = 1 code).

Soit limitation de la taille du dictionnaire, soit codes de taille croissante.

À stocker : suite des codes (symboles + mots du dictionnaire), ainsi que quelques caractères de contrôle.

Efficace pour signaux images, documents complexes (utilisé notamment dans les programmes de compression ZIP).

7.3 Redondance structurelle

- Idée générale : modélisation (au sens large) de l'image ou du signal à stocker.
  → Stockage du modèle (exact ou approché).
  → Stockage de l'écart au modèle (exact ou approché).

- **Exemple 1** Codage prédictif sans perte.

  Signal $x(n)$ à compresser. Modèle :

$$x(n) = \underbrace{\sum_{p=1}^{P} a_p x(n-p)}_{\text{prédicteur}} + \varepsilon(n).$$

→ Démarche : estimation des coefficients du prédicteur (modèle) $\{a_p \mid 1 \leq p \leq P\}$.

Calcul (récurrent au besoin) des erreurs de prédiction :

$$\varepsilon(n) = x(n) - \sum_{p=1}^{P} a_p x(n-p).$$

Stockage de

$$\{a_p \mid 1 \leq p \leq P\},$$
$$\{x(n) \mid 0 \leq n \leq p-1\},$$
$$\{\varepsilon(n) \mid p \leq n \leq N\}.$$

Au décodage,

$$x(n) = \sum_{p=1}^{P} a_p x(n-p) + \varepsilon(n).$$

Le gain provient de ce que la dispersion (l'entropie) de $\varepsilon(n)$ est beaucoup plus faible que celle de $x(n)$.

→ Remarque : extension facile aux images (exemple).

→ Schéma des systèmes de codage et de décodage.

Codeur Décodeur

- **Exemple 2** Codage prédictif avec pertes.

→ Idée générale : analogue au codage prédictif sans perte, mais l'erreur de prédiction est arrondie par une étape de quantification.

$\varepsilon(n)$: erreur de prédiction ;
$\dot{\varepsilon}(n)$: erreur de prédiction quantifiée.

$$\dot{x}(n) = \sum_{p=1}^{P} a_p \dot{x}(n-p) + \dot{\varepsilon}(n)$$

⇒ Accumulation des erreurs de quantification.

Pour corriger le défaut : définir ε comme l'erreur entre x et le $\hat{x}$ reconstruit par le décodeur à partir des erreurs de prédiction quantifiées précédentes.

→ Schéma des systèmes de codage et de décodage :

Codeur
Décodeur

Boucle supplémentaire au niveau du codage : stabilisation par rétroaction (mise en boucle fermée).

→ Quantités à stocker : identiques à celles du codage prédictif sans perte.

→ Cas particulier important : modulation δ, avec $P = 1$.

Quantification de ε sur deux niveaux (en fonction de l'image).

→ Remarques :
- ∗ Nécessaire compromis entre qualité et taux de compression.
- ∗ Valeur de P ?
- ∗ Structure du prédicteur pour les images ?

• Codage par transformées : décomposition sur des bases orthogonales.

→ Idée générale. x appartient à un espace $\mathcal{E}$ muni d'une norme.

$$\{e_i, 1 \leq i \leq I\} \text{ base orthogonale de } \mathcal{E}$$
$$x = \sum_{i=1}^{I} x_i e_i$$
$$x_i = \frac{< x|e_i >}{< e_i|e_i >}$$

- ∗ Stockage des x_i plutôt que des composantes de x.
- ∗ Quantification des x_i selon l'importance de ces coefficients sur la perception (visuelle, auditive...).
- ∗ Importance du choix de la base.

→ Choix classiques :

∗ Base de Fourier (transformée en cosinus) pour les images.

Le pas de quantification est ajusté en fonction de la fréquence. Plus la fréquence est élevée, plus le pas est grand.
Exemple : $Q(u, v) = 1 + F(1 + \alpha(u^n + v^n))$.

L'image est préalablement découpée en sous-blocs de taille modeste (8×8, par exemple).
⇒ Codage compatible JPEG.

Quantités à stocker :
- Taille de la matrice.
- Taille des blocs.
- Pas de quantification.
- Coefficients de la décomposition (triplets $(u, v, x(u, v))$).

Ces grandeurs sont ensuite transformées par une étape de codage de symboles (Huffman, par exemple).

∗ Approche analogue pour les fichiers audio (codage MP3).

∗ Base d'ondelettes pour signaux et images.

Démarche analogue aux précédentes. La base est bien adaptée à la représentation de signaux comportant des « phases transitoires » (le nombre de coefficients significatifs reste faible).

Conclusion

- Dans le domaine médical : compression avec pertes ou sans perte ?
- Compromis entre le facteur de compression et la qualité.
- Influence entre la redondance structurelle et la redondance liée à la source, sans lien formel.

Bibliographie

[1] Rafael C. Gonzalez and Richard E. Woods, *Digital Image Procesing*, Perarson Education, Inc., Upper Saddle River, NJ, 3rd edition, 2008.

[2] Yves Goussard, *Traitement de Signaux Biomédicaux*, vol. 5123 of *Notes de cours*, École Polytechnique, Montréal, Canada, 1996, 118 pages.