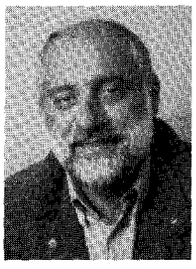
Reconnaissance supervisée de signaux aléatoires

Supervised recognition of random signals

Eric LE CARPENTIER

Laboratoire d'Automatique de Nantes URA CNRS n° 823, École Nationale Supérieure de Mécanique, 1, rue de la Noë, 44072 NANTES CEDEX 03.

Eric Le Carpentier est ingénieur ENSM et titulaire d'un DEA d'Automatique depuis 1987. Dans le cadre d'une thèse de Doctorat, il exerce des activités de recherche en Traitement du Signal et Reconnaissance de Formes, le but de son travail étant la classification automatique de signaux aléatoires.



Christian DONCARLI

Laboratoire d'Automatique de Nantes URA CNRS n° 823, École Nationale Supérieure de Mécanique, 1, rue de la Noë, 44072 NANTES CEDEX 03.

Christian Doncarli est Docteur-Ingénieur ENSM, Docteur ès Sciences, et Professeur d'Université. Ses domaines de recherche regroupent le filtrage statistique ainsi que la théorie de la détection-décision. Le champ d'application privilégié par ses recherches est l'ingénierie biomédicale.

RÉSUMÉ

On s'intéresse à la classification et à la reconnaissance d'échantillons finis de signaux aléatoires stationnaires ergodiques, modélisables par un processus ARMA d'ordre fixé. La méthode utilisée consiste à identifier les coefficients du modèle, de façon à condenser en un faible nombre de paramètres toutes les propriétés statistiques du signal. On définit ensuite une distance dans l'espace de représentation qui en résulte. Les distances les plus connues sont les distances entre spectres de puissance (distance d'Itakura-Saïto, distance cepstrale...), ou entre lois de probabilité de N-échantillons du signal (divergence de Kullback, distance de Bhattacharyya...), ces quantités étant calculées à partir des coefficients du modèle. Or, ces coefficients ne sont pas mesurés exactement, mais seulement estimés, et la répartition de ces estimateurs dépend de la méthode d'identification utilisée. On propose donc d'intégrer cette caractéristique fondamentale dans la définition d'une distance entre modèles ARMA identifiés par Maximum de Vraisemblance hors-ligne.

MOTS CLÉS

Signal aléatoire, modèle ARMA, Estimateur du Maximum de Vraisemblance, Classification, Distance.

SUMMARY

The purpose of this paper is the classification and the recognition of stationary ergodic random signals, which can be represented by a fixed-order ARMA process. The method consists in estimating the model parameters, in order to summarize the statistical properties of the signal in a small set of parameters. Then we define a distance measure in the representation space, the more classical being distances between power spectrum (Itakura-Saïto distance, cepstral distance...) or between probability laws of N-points data sequences (Kullback divergence, Bhattacharyya distance...). These quantities are evaluated from model parameters. But these parameters are not exactly measured; they are just estimated, and the probability law of these estimators depends on the chosen identification method. We propose to insert this fundamental feature in the definition of a distance between ARMA models identified by Maximum Likelihood method.

KEY WORDS

Random signal, ARMA model, Maximum Likelihood estimator, Classification, Distance.

1. Introduction

On s'intéresse au problème de la reconnaissance de phénomènes (physiques, physiologiques...) se présentant sous la forme d'enregistrements de durée finie de signaux aléatoires stationnaires ergodiques centrés. Plus précisément, chaque phénomène (encore appelé classe) correspond à un signal aléatoire ayant certaines caractéristiques (inconnues), et il s'agit, au vu d'une réalisation de durée finie d'un de ces signaux, de décider quelle est la classe la plus probable. On admettra par ailleurs qu'on dispose d'une population d'échantillons déjà affectés aux classes par un expert (tirage d'apprentissage) et le problème fondamental abordé ici consiste à élaborer une règle de décision optimale, c'est-à-dire à calculer les probabilités de chaque classe conditionnellement à l'échantillon examiné (appelé individu) et à la population d'apprentissage.

La première étape de l'approche classique de ce problème consiste à modéliser les signaux, puis à les représenter dans un espace paramétrique. La deuxième phase comporte l'élaboration d'une distance entre les modèles dans l'espace paramétrique retenu (Distances spectrales [1, 2], ou distances inspirées des problèmes de sélection de signaux en communication [3, 4, 5]). Il faut noter que ces méthodes classiques discriminent les échantillons de signaux sur leurs paramètres de représentation, en considérant qu'il s'agit d'une mesure exacte des individus. Dès lors, l'optimalité de la règle de décision, c'est-à-dire la conditionnalité à l'échantillon observé (et non à des estimations des paramètres) n'est plus garantie, quel que soit le choix du couple « espace de représentation-distance ». Pour retrouver cette optimalité, il faut utiliser une information équivalente à celle contenue dans l'échantillon de signal, c'est-à-dire la statistique des paramètres estimés conditionnellement à cet échantillon. La prise en compte de cette statistique permet de déduire une règle de décision optimale (moyennant quelques approximations justifiées), c'est-à-dire l'évaluation des probabilités des classes conditionnellement à l'individu examiné et à la population d'apprentissage, ce qui garantit l'indépendance de la procédure, vis-à-vis du choix des paramètres de représentation et de la distance utilisée.

On reviendra en détail (§2) sur cette notion d'optimalité de la décision, et sur les conséquences que le choix d'une stratégie optimale impose aux modèles et aux algorithmes d'identification à utiliser. On présentera ensuite l'apprentissage des probabilités *a posteriori* et la règle de décision qui s'en déduit (§3). On proposera enfin quelques conclusions (§4), en revenant sur la notion fondamentale de distance entre modèles, et en indiquant une extension possible vers la classification non supervisée.

2. Généralités

2.1. Choix du modèle et de la méthode d'identification

On se propose de classer des signaux aléatoires ergodiques stationnaires centrés, susceptibles d'une représentation gaussienne markovienne (densité spectrale de puissance rationnelle). La représentation la plus habituelle de cette famille de signaux est le modèle générateur (de type AR ou ARMA) permettant de synthétiser leurs propriétés statistiques par filtrage linéaire d'une séquence blanche. La question préalable à toute classification est donc le choix de la structure de modèle retenue pour représenter les individus. On peut alors avancer l'idée que si les tests habituels de structure sont aptes à discriminer efficacement les échantillons de signaux, il est inutile de poursuivre l'étude, puisque l'objectif de classification est atteint. On suppose donc que tous les individus sont susceptibles d'être modélisés par une structure identique et que seules des différences de paramètres correspondent aux diverses classes.

La règle de décision optimale (conditionnalité aux échantillons) ne peut se traduire en termes de paramètres, que si on connaît les statistiques de ces paramètres conditionnellement aux échantillons. Il est donc évident que le choix naturel de l'espace de représentation sera l'ensemble des coefficients et que la méthode d'identification utilisée sera le Maximum de Vraisemblance. En effet, ces choix donnent très facilement les statistiques cherchées [6]. On suppose en outre que tous les échantillons correspondent à une structure ARMA (plus générale que la structure AR).

2.2. Modélisation ARMA et identification par Maximum de Vraisemblance

Soit $(Y(t))_{t\in\mathbb{Z}}$ un signal aléatoire discret modélisé par une séquence blanche gaussienne $(B_0(t))_{t\in\mathbb{Z}}$ de variance λ filtrée par un filtre ARMA de transmittance

$$H(\theta, z^{-1}) = \frac{C(\theta, z^{-1})}{A(\theta, z^{-1})} \quad \text{(cf. } \textit{fig.}\ 1\text{)}.$$

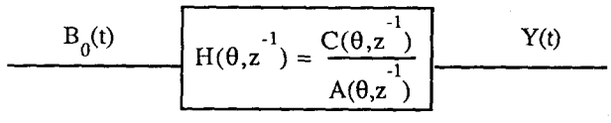


Fig. 1. – Modèle générateur du signal $(Y(t))_{t\in\mathbb{Z}}$.

En outre, on note:

$$C(\theta, z^{-1}) = 1 + \sum_{i=1}^{m} c_i z^{-i}$$

$$A(\theta, z^{-1}) = 1 + \sum_{i=1}^{n} a_i z^{-i}$$

$$d = n + m$$

$$\theta = [a_1 \ldots a_n c_1 \ldots c_m].$$

Soit le vecteur aléatoire $(Y(t))_{1\leq t\leq N}$, échantillon de N points du signal. L'estimateur de θ par maximum de vraisemblance, noté X^N, est défini par:

$$X^N = \arg\min_{x\in\mathbb{R}^d} J_N(x)$$

avec

$$J_N(x) = \frac{1}{N} \sum_{t=1}^{N} B^2(t,x)$$

et $B(t,x)$ pseudo-innovation définie par récurrence par :

$$B(t,x) = Y(t) + [Y(t-1) \ldots Y(t-n) - B(t-1,x) \ldots - B(t-m,x)] . x.$$

On rappelle la propriété fondamentale suivante : quand N tend vers l'infini, le vecteur aléatoire $N^{1/2} X^N$ est régi par la loi gaussienne de moyenne θ et de variance-covariance $V(\theta)$ définie par :

$$V(\theta) = 2\lambda . \left[E \frac{d^2 J_N}{dx^2}(x) \right]^{-1}_{x=\theta},$$

où $d^2 J_N / dx^2$ désigne le hessien de J_N [6].
Par exemple, dans le cas ARMA (1/1) :

$$V^{-1}(\theta) = V^{-1}(a_1, c_1) = \begin{bmatrix} \frac{1}{1-a_1^2} & \frac{1}{a_1 c_1 - 1} \\ \frac{1}{a_1 c_1 - 1} & \frac{1}{1-c_1^2} \end{bmatrix}$$

2.3. Hypothèses

On suppose qu'il existe un modèle unique de vecteur-paramètre $\theta_i \in \mathbb{R}^d$ représentatif de la classe $i \in \{1, \ldots, c\}$. Soit alors $(Y_{ik}(t))_{1 \leq t \leq N_{ik}}$ le k-ième échantillon affecté à la classe i dans la population d'apprentissage ($k \in \{1, \ldots, n_i\}$, en notant n_i le nombre d'échantillons affectés à la classe i). L'identification de θ_i par Maximum de Vraisemblance à partir de cet échantillon fournit un vecteur-paramètre $X_{ik}^{N_{ik}} \in \mathbb{R}^d$.
De même, soit $(Y(t))_{1 \leq t \leq N}$ l'échantillon à classer. L'identification par Maximum de Vraisemblance fournit un vecteur-paramètre $X^N \in \mathbb{R}^d$ qui est un estimateur du vecteur-paramètre $\theta \in \mathbb{R}^d$. On cherche donc $i \in \{1, \ldots, c\}$ tel que $\theta = \theta_i$.
Notons $P_i = (X_{ik}^{N_{ik}})_{1 \leq k \leq n_i}$ l'ensemble des estimateurs de θ_i, et $p_i = (x_{ik})_{1 \leq k \leq n_i}$ une réalisation de P_i.
De même, $P = (P_i)_{1 \leq i \leq c}$ est l'ensemble de tous les estimateurs, et $p = (p_i)_{1 \leq i \leq c}$ est une réalisation de P ; enfin, soit $x \in \mathbb{R}^d$ une réalisation de X^N à classer.
On se propose d'utiliser alors le classifieur à taux d'erreur minimale, *i.e.* celui qui affecte l'individu à classer à la classe i pour laquelle la probabilité *a posteriori*, notée $\text{Prob}(I = i / X^N = x, P = p)$, est maximale.

3. Règle de décision

3.1. Calcul de $\text{Prob}(I = i / X^N = x . P = p)$

En notant $f_{X/Y=y}(x)$ la densité de probabilité du vecteur aléatoire X conditionnellement à l'événement $\{Y = y\}$ calculée au point x, on peut écrire [7] :

$$\text{Prob}(I = i / X^N = x, P = p) = \frac{f_{X^N/I=i;\, P_i=p_i}(x) \, \text{Prob}(I = i / P = p)}{c}$$
$$\sum_{j=1} f_{X^N/I=j;\, P_i=p_i}(x) \, \text{Prob}(I = j / P = p)$$

En supposant que $f_{X^N/I=i;\, P_i=p_i}$ ne dépend pas de P_j pour $j \neq i$, et que les probabilités *a priori* des classes $\text{Prob}(I = i)$, $i \in \{1, \ldots, c\}$ sont parfaitement connues alors :

$$\text{Prob}(I = i / X^N = x, P = p) = \frac{f_{X^N/I=i;\, P_i=p_i}(x) \, \text{Prob}(I = i)}{c}$$
$$\sum_{j=1} f_{X^N/I=j;\, P_j=p_j}(x) \, \text{Prob}(I = j)$$

La fonction $f_{X^N/I=i;\, P_i=p_i}$ est paramétrée par θ_i. L'estimation bayesienne de θ_i consiste alors à modéliser ce vecteur inconnu par un vecteur aléatoire Θ_i de densité f_{Θ_i} constante sur le domaine S défini par $S = \{\theta \in \mathbb{R}^d : H(\theta, z^{-1})$ stable à inverse stable$\}$, et nulle hors de S. Alors :

$$f_{X^N/I=i;\, P_i=p_i}(x) = \int_{\mathbb{R}^d} f_{X^N/\Theta_i=\theta;\, I=i;\, P_i=p_i}(x) f_{\Theta_i/I=i;\, P_i=p_i}(\theta) \, d\theta$$

Il est naturel de supposer Θ_i et P_i indépendants de I, et de supposer que $X^N/I = i$ et $X_{ik}^{N_{ik}}$, $k \in \{1, \ldots, n_i\}$ sont 2 à 2 indépendants conditionnellement à l'événement $\{\Theta_i = \theta\}$. On peut donc écrire :

(1) $$f_{X^N/I=i;\, P_i=p_i}(x) = \int_{\mathbb{R}^d} f_{X^N/\Theta_i=\theta;\, I=i}(x) f_{\Theta_i/P_i=p_i}(\theta) \, d\theta$$

avec

$$f_{\Theta_i/P_i=p_i}(\theta) = \frac{f_{P_i/\Theta_i=\theta}(p_i) f_{\Theta_i}(\theta)}{\int_S f_{P_i/\Theta_i=u}(p_i) f_{\Theta_i}(u) \, du}$$

soit

(2) $$f_{\Theta_i/P_i=p_i}(\theta) = \frac{\prod_{k=1}^{n_i} f_{X_{ik}^{N_{ik}}/\Theta_i=\theta}(x_{ik})}{\int_S \prod_{k=1}^{n_i} f_{X_{ik}^{N_{ik}}/\Theta_i=u}(x_{ik}) \, du}.$$

3.2. Approximations

On rappelle que le vecteur aléatoire $W = N^{1/2}(X^N - \Theta_i)/\Theta_i = \theta$; $I = i$ est régi asymptotiquement (quand N tend vers l'infini) par la loi gaussienne centrée de variance-covariance $V(\theta)$:

$$f_{N^{1/2}(X^N - \Theta_i)/\Theta_i=\theta;\, I=i}(w) = \frac{1}{(2\pi)^{d/2} (\det V(\theta))^{1/2}} e^{-(1/2) w^T V^{-1}(\theta) w}.$$

On se propose d'approcher cette densité pour N fini par :

$$f_{N^{1/2}(X^N-\Theta_i)/\Theta_i=\theta;\, I=i}(w) \# \frac{1}{(2\pi)^{d/2}(\det V(\theta+(w/N^{1/2})))^{1/2}} e^{-(1/2)\, w^T V^{-1}(\theta+(w/N^{1/2}))\, w}.$$

En effet, cette fonction vérifie la normalité asymptotique, et converge vers la densité $f_{N^{1/2}(X^N-\Theta_i)/\Theta_i=\theta;\, I=i}$ quand N tend vers l'infini.

Le changement de variable linéaire $W = N^{1/2}(X^N-\Theta_i)/\Theta_i=\theta;\ I=i$ permet alors de déduire une approximation $g_N(x,\theta)$ de $f_{X^N/\Theta_i=\theta;\, I=i}(x)$, définie pour tout $\theta\in\mathbb{R}^d$ et tout $x\in S$ par :

(3)
$$g_N(x,\theta) = \frac{N^{d/2}}{(2\pi)^{d/2}(\det V(x))^{1/2}} e^{-(N/2)(x-\theta)^T V^{-1}(x)(x-\theta)}.$$

De même, $f_{X_{ik}^{N_{ik}}/\Theta_i=\theta}(x_{ik})$ est approché par $g_{N_{ik}}(x_{ik},\theta)$. On peut donc écrire, d'après (1), (2) et (3) :

(4) $f_{X^N/I=i;\, P_i=p_i}(x)$
$$\# \int_S g_N(x,\theta) \frac{\prod_{k=1}^{n_i} g_{N_{ik}}(x_{ik},\theta)}{\int_S \prod_{k=1}^{n_i} g_{N_{ik}}(x_{ik},u)\,du}\, d\theta.$$

D'autre part, il est aisé de vérifier (*cf.* annexe) que $\prod_{k=1}^{n_i} g_{N_{ik}}(x_{ik},\theta)$ est une fonction gaussienne de θ, de moyenne μ_i et de variance-covariance A_i définies par :

(5)
$$A_i^{-1} = \sum_{k=1}^{n_i} N_{ik} V^{-1}(x_{ik})$$
$$\mu_i = A_i \cdot \sum_{k=1}^{n_i} N_{ik} V^{-1}(x_{ik})\, x_{ik}.$$

Donc, pour un n_i ou des N_{ik} suffisamment grands, la variance-covariance A_i devient petite et par conséquent, $\prod_{k=1}^{n_i} g_{N_{ik}}(x_{ik},\theta)$ est petit pour θ loin de μ_i; il est ainsi justifié d'approcher les intégrales sur S par des intégrales sur $\mathbb{R}^d$.

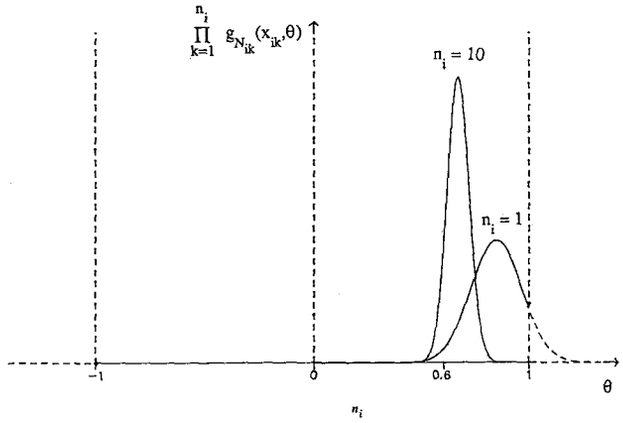


Fig. 2. – Fonction $\prod_{k=1}^{n_i} g_{N_{ik}}(x_{ik},.)$ pour différentes valeurs de n_i.

A titre d'exemple, on visualise une réalisation de cette fonction pour deux valeurs de n_i dans le cas d'un modèle AR d'ordre 1 (simulations de signaux de 20 points; modèle exact : $a_1 = 0{,}6$) (cf. fig. 2). L'approximation est nettement justifiée pour $n_i = 10$.

Avec l'approximation précédente, l'équation (4) s'écrit alors :

(6) $f_{X^N/I=i;\, P_i=p_i}(x)$
$$\neq \int_{\mathbb{R}^d} g_N(x,\theta) \frac{\prod_{k=1}^{n_i} g_{N_{ik}}(x_{ik},\theta)}{\int_{\mathbb{R}^d} \prod_{k=1}^{n_i} g_{N_{ik}}(x_{ik},u)\,du}\, d\theta$$

avec

$$g_N(x,\theta) = \frac{N^{d/2}}{(2\pi)^{d/2}(\det V(x))^{1/2}} e^{-(N/2)(x-\theta)^T V^{-1}(x)(x-\theta)}$$

$$\frac{\prod_{k=1}^{n_i} g_{N_{ik}}(x_{ik},\theta)}{\int_{\mathbb{R}^d} \prod_{k=1}^{n_i} g_{N_{ik}}(x_{ik},u)\,du} = \frac{1}{(2\pi)^{d/2}(\det A_i)^{1/2}} e^{-(1/2)(\theta-\mu_i)^T A_i^{-1}(\theta-\mu_i)}.$$

Après quelques calculs, on montre alors sans difficulté (*cf.* annexe) que :

(7) $f_{X^N/I=i;\, P_i=p_i}(x) \# \frac{1}{(2\pi)^{d/2}(\det(A_i+(1/N)V(x))^{1/2})} \times e^{-(1/2)(x-\mu_i)^T (A_i+(1/N)V(x))^{-1}(x-\mu_i)}.$

3.3. Règle de décision

L'application de la théorie bayesienne de la classification à erreur minimale permet alors d'adopter la règle de décision suivante : on affecte l'individu x à la classe i

qui maximise $\mathrm{Prob}(I=i/X^N=x, P=p)$,

i.e. qui maximise $f_{X^N/I=i;\, P_i=p_i}(x).\mathrm{Prob}(I=i)$,

i.e. qui maximise

$\mathrm{Log}[(2\pi)^{d/2} f_{X^N/I=i;\, P_i=p_i}(x).\mathrm{Prob}(I=i)]$,

i.e. qui minimise la fonction discriminante $h_i(x)$ définie d'après (7) par :

$$h_i(x) = \frac{1}{2}\mathrm{Log}\left[\det\left(A_i+\frac{1}{N}V(x)\right)\right] - \mathrm{Log}[\mathrm{Prob}(I=i)] + \frac{1}{2}(x-\mu_i)^T\left(A_i+\frac{1}{N}V(x)\right)^{-1}(x-\mu_i)$$

où A_i et μ_i sont définis par les formules (5).

On définit ainsi une distance du point x affecté par $(1/N)V(x)$ au point μ_i affecté par A_i.

En particulier, on peut remarquer que si la classe i ne dispose que d'un individu d'apprentissage x_i identifié à partir d'un échantillon de N_i points, la distance devient alors :

$$h_i(x) = \frac{1}{2}\mathrm{Log}\left[\det\left(\frac{1}{N_i}V(x_i)+\frac{1}{N}V(x)\right)\right] - \mathrm{Log}\,[\mathrm{Prob}\,(I=i)] + \frac{1}{2}(x-x_i)^T\left[\frac{1}{N_i}V(x_i)+\frac{1}{N}V(x)\right]^{-1}(x-x_i).$$

Enfin, la borne de Cramer-Rao $V(x)$ étant difficile à calculer analytiquement, on l'approche par $\overline{V(x)} = 2J_N(x).[(d^2 J_N/dx^2)(x)]^{-1}$, obtenu par exemple en minimisant J_N par une méthode de quasi-Newton.

4. Simulations

On considère des échantillons de signaux de longueurs différentes tirées aléatoirement entre 50 et 500. Leur longueur peut donc être faible, afin de tester la robustesse de la méthode dans un cadre qui est loin d'être asymptotique. Ces échantillons sont divisés en deux classes : la première correspond au filtre formeur ARMA de coefficients $\theta_1 = [-1{,}72;\ 0{,}75;\ 0{,}4;\ 0{,}15]$ (pôles : $0{,}86 \pm 0{,}102$]; zéros : $-0{,}2 \pm 0{,}332$ j), la seconde au filtre de coefficients $\theta_2 = [-1{,}75;\ 0{,}78;\ 0{,}3;\ 0{,}2]$ (pôles : $0{,}875 \pm 0{,}12$ j; zéros : $-0{,}15 \pm 0{,}421$ j).

On teste alors la règle de décision obtenue à partir d'une population d'apprentissage croissant de 1 à 50 individus par classe sur une population test de 1 000 individus par classe, disjointe de la population d'apprentissage : on obtient donc une courbe représentant le pourcentage d'échantillons mal classés dans la population-test de 2 000 individus en fonction du cardinal des populations d'apprentissage. Cette population-test importante permet d'évaluer correctement la probabilité d'erreur pour une population d'apprentissage donnée.

Nous avons répété cette opération sur 6 populations d'apprentissage différentes, en utilisant d'une part la règle de décision proposée, d'autre part la distance cepstrale au centre de gravité calculée sur les 8 premiers coefficients cepstraux.

Si, sur l'exemple représenté figure 3, il est difficile de conclure à une supériorité de la règle de décision proposée, l'exemple de la figure 4 montre clairement que cette règle de décision, en accordant une faible pondération aux coefficients identifiés à partir d'un échantillon de signal de faible longueur, permet une franche amélioration des performances de la classification. La moyenne sur les 6 essais (figure 5) permet de constater une amélioration quel que soit le cardinal de la population d'apprentissage, et en particulier lorsque celui-ci est faible. De plus, il est clair, d'après la figure 5, que dans cet exemple, une population d'apprentissage de cardinal 50 est suffisante pour caractériser une classe (à partir de 30, l'adjonction d'individus supplémentaires n'apporte que peu d'information).

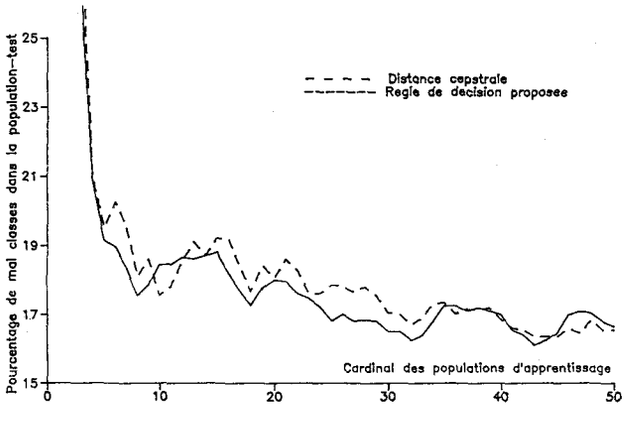


Fig. 3. – Premier essai.

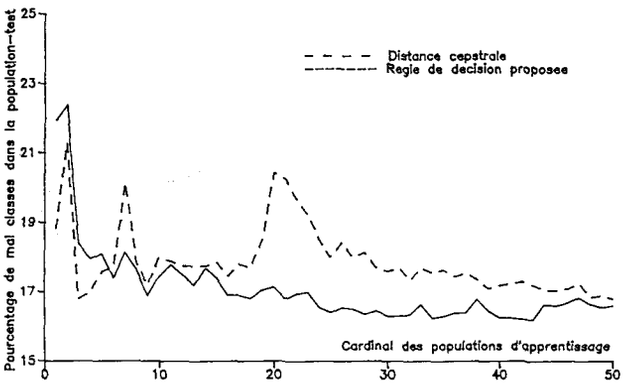


Fig. 4. – Deuxième essai.

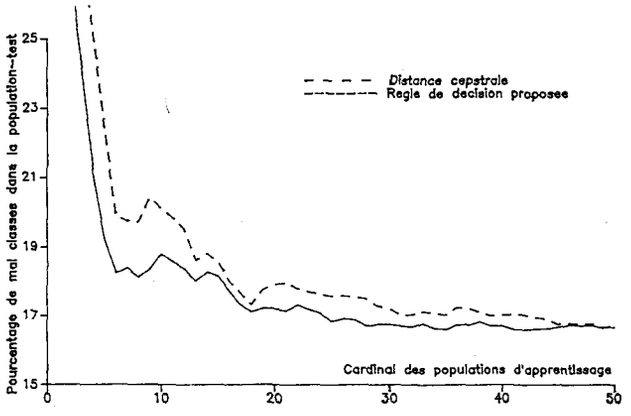


Fig. 5. – Moyenne des 6 essais.

5. Conclusions et perspectives

L'algorithme de classification proposé conduit à une règle de décision optimale au sens de la probabilité d'erreur. Les diverses approximations nécessaires au calcul analytique des probabilités *a posteriori* sont justifiées, et les fonctions discriminantes obtenues font directement apparaître les résultats de l'algorithme d'identification par Maximum de Vraisemblance. Enfin, l'expression de ces fonctions discriminantes à

la forme d'une distance, ce qui permet d'envisager une extension de ces résultats vers la classification non supervisée.

Annexe

ÉVALUATION D'UN PRODUIT DE DENSITÉS GAUSSIENNES ET DE SON INTÉGRALE

Pour tout $i \in \{1, \ldots, n\}$, soit f_i la densité de probabilité de la loi normale de moyenne $x_i \in \mathbb{R}^d$ et de variance V_i:

$$f_i(\theta) = \frac{1}{(2\pi)^{d/2} (\det V_i)^{1/2}} e^{-(1/2)(\theta - x_i)^T V_i^{-1} (\theta - x_i)}.$$

Calculons $\prod_{i=1}^{n} f_i(\theta)$:

$$\prod_{i=1}^{n} f_i(\theta) = \frac{1}{(2\pi)^{dn/2} \prod_{i=1}^{n} (\det V_i)^{1/2}} e^{-(1/2) \sum_{i=1}^{n} (\theta - x_i)^T V_i^{-1} (\theta - x_i)}.$$

Alors, en posant $\begin{cases} A^{-1} = \sum_{i=1}^{n} V_i^{-1} \\ \mu = A \sum_{i=1}^{n} V_i^{-1} x_i \end{cases}$

$$\prod_{i=1}^{n} f_i(\theta) = \frac{(\det A)^{1/2}}{(2\pi)^{d(n-1)/2} \prod_{i=1}^{n} (\det V_i)^{1/2}} \times e^{(1/2)(\mu^T A^{-1} \mu - \sum_{i=1}^{n} x_i^T V_i^{-1} X_i)} \times \frac{1}{(2\pi)^{d/2} (\det A)^{1/2}} e^{-(1/2)(\theta - \mu)^T A^{-1} (\theta - \mu)}.$$

Il est alors évident que

$$(a) \quad \frac{\prod_{i=1}^{n} f_i(\theta)}{\int_{\mathbb{R}^d} \prod_{i=1}^{n} f_i(u)\, du} = \frac{1}{(2\pi)^{d/2} (\det A)^{1/2}} \times e^{-(1/2)(\theta - \mu)^T A^{-1} (\theta - \mu)}$$

$$\int_{\mathbb{R}^d} \prod_{i=1}^{n} f_i(u)\, du = \frac{(\det A)^{1/2}}{(2\pi)^{d(n-1)/2} \prod_{i=1}^{n} (\det V_i)^{1/2}} \times e^{(1/2)(\mu^T A^{-1} \mu - \sum_{i=1}^{n} x_i^T V_i^{-1} x_i)}$$

Calculons ce dernier terme dans le cas $n = 2$:

$$\int_{\mathbb{R}^d} f_1(u) f_2(u)\, du = \frac{(\det A)^{1/2}}{(2\pi)^{d/2} \prod_{i=1}^{n} (\det V_i)^{1/2}} \times e^{(1/2)(\mu^T A^{-1} \mu - x_1^T V_1^{-1} x_1 - x_2^T V_2^{-1} x_2)}.$$

En utilisant la relation matricielle

$$(V_1^{-1} + V_2^{-1})^{-1} = V_1 (V_1 + V_2)^{-1} V_2 = V_1 (V_1 + V_2)^{-1} V_2,$$

on calcule d'une part:

$$\det A = \det (V_1^{-1} + V_2^{-1})^{-1} = \det V_1 (V_1 + V_2)^{-1} V_2 = \frac{\det V_1 \det V_2}{\det (V_1 + V_2)}.$$

D'autre part:

$$\begin{aligned} \mu^T A \mu &= (x_1^T V_1^{-1} + x_2^T V_2^{-1}) (V_1^{-1} + V_2^{-1}) (V_1^{-1} x_1 + V_2^{-1} x_2) \\ &= x_1^T V_1^{-1} V_1 (V_1 + V_2)^{-1} V_2 (V_1^{-1} x_1 + V_2^{-1} x_2) \\ &\quad + x_2^T V_2^{-1} V_2 (V_1 + V_2)^{-1} V_1 (V_1^{-1} x_1 + V_2^{-1} x_2) \\ &= x_1^T (V_1 + V_2)^{-1} V_2 V_1^{-1} x_1 \\ &\quad + x_2^T (V_1 + V_2)^{-1} V_1 V_2^{-1} x_2 + 2 x_1^T (V_1 + V_2)^{-1} x_2. \end{aligned}$$

Donc, en utilisant la relation matricielle

$$V_1^{-1} - (V_1 + V_2)^{-1} V_2 V_1^{-1} = (V_1 + V_2)^{-1}:$$

$$\mu^T A \mu - x_1^T V_1^{-1} x_1 - x_2^T V_2^{-1} x_2 = -(x_1 - x_2)^T (V_1 + V_2)^{-1} (x_1 - x_2).$$

Finalement:

$$(b) \quad \int_{\mathbb{R}^d} f_1(u) f_2(u)\, du = \frac{1}{(2\pi)^{d/2} (\det (A_1 + A_2))^{1/2}} \times e^{-(1/2)(x_1 - x_2)^T (A_1 + A_2)^{-1} (x_1 - x_2)}.$$

APPLICATION

En tirant parti du fait que les fonctions $g_N(x, .)$ sont des densités de probabilité gaussiennes de moyenne x et de variance $(1/N) V(x)$, la relation (5) est une application directe de la relation (*a*), et le passage de la relation (6) à la relation (7) est une application directe de (*b*).

Manuscrit reçu le 27 février 1989.

BIBLIOGRAPHIE

[1] R. M. GRAY, A. BUZO, A. H. GRAY et Y. MATSUYAMA, Distortion measures for speech processing, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-28, n° 4, août 1980.

[2] A. H. GRAY et J. D. MARKEL, Distance measures for speech processing *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-24, n° 5, août 1976.

[3] T. L. GRETTENBERG, Signal selection in communication and radar systems, *IEEE Transactions on Information Theory*, IT-9, october 1963.

[4] D. KAZAKOS, The Bhattacharyya distance and detection between Markov chains, *IEEE Transactions on Information Theory*, IT-24, n° 6, november 1978.

[5] T. KAILATH, The Divergence and Bhattacharyya distance measures in signal selection, *IEEE Transactions on Communication Technology*, COM-15, n° 1, février 1967.

[6] L. LJUNG, *System Identification: Theory for the User*, Prentice Hall, Inc, 1987, p. 239-248.

[7] R. O. DUDA et P. E. HART, *Pattern Classification and Scene Analysis*, Wiley Interscience, Inc, 1973, p. 44-59.