THESE

présentée à

L'INSTITUT NATIONAL POLYTECHNIQUE DE LORRAINE

pour obtenir le titre de

DOCTEUR DE TROISIEME CYCLE DE MATHEMATIQUE APPLIQUEE

par

PASCAL WILD

Sujet

ESTIMATION AUTOMATIQUE DE LA DENSITE MULTIVARIEE ET APPLICATIONS A L' ANALYSE DES DONNEES

Soutenue publiquement le 20 juin 1984 devant la Commission d'Examen :

URY

MM.	M. DEPAIX	Président
	J.L. MALLET	Rapporteur
	C. CHAMBON	Examinateurs
	G. DER-MEGREDITCHIAN	
	G. SAPORTA	

THESE

présentée à

L'INSTITUT NATIONAL POLYTECHNIQUE DE LORRAINE

pour obtenir le titre de

DOCTEUR DE TROISIEME CYCLE DE MATHEMATIQUE APPLIQUEE

par

PASCAL WILD

Sujet

ESTIMATION AUTOMATIQUE DE LA DENSITE MULTIVARIEE ET APPLICATIONS A L'ANALYSE DES DONNEES

Soutenue publiquement le 20 juin 1984 devant la Commission d'Examen :

JURY

MM.	M. DEPAIX	Président
	J.L. MALLET	Rapporteur
	C. CHAMBON	Examinateurs
	G. DER-MEGREDITCHIAN	
	G. SAPORTA	

à Patrick

REMERCIEMENTS

Je remercie en premier lieu Monsieur MALLET qui a dirigé ma thèse. Me proposant un sujet intéressant, me guidant dans les moments difficiles, ne ménageant ni ses critiques, ni ses encouragements, il m'a permis de travailler dans d'excellentes conditions et m'a communiqué sa passion pour la recherche.

Je remercie Monsieur DEPAIX de présider la jury de cette thèse, mais plus encore de m'avoir en grande partie initié aux statistiques et surtout de m'en avoir donné le goût.

Je remercie Messieurs DER-MEGREDITCHIAN et SAPORTA de s'être déplacé depuis Paris pour participer au jury de ma thèse.

Je remercie Monsieur CHAMBON de participer au jury de ma thèse.

Je remercie d'autre part le centre universitaire d'études statistiques de l'université Louis Pasteur de Strasbourg de m'avoir mis à disposition ses locaux.

Je remercie Messieurs TURLOT et HEDELIN. Les discussions que j'ai pu avoir avec eux ont contribué à faire avancer mon travail.

Enfin je remercie Melle MARCHAL et Mme JULIEN qui ont eu la lourde tâche de taper ce texte.

INTRODUCTION

Le travail présenté dans cette thèse se divise en deux parties distinctes.

Dans une première partie, nous examinons une technique d'estimation de la densité multivariée. Cette méthode est d'une efficacité assez faible. Nous tachons de remédier à ce désavantage en rendant la méthode plus souple. En particulier nous adaptons cette technique de façon à ce qu'elle tienne compte des paramètres empiriques.

Dans une deuxième partie, nous examinons des applications possibles de la technique précédente en analyse des donnée. Nous traitons en particulier les problèmes de la classification et de la régression sous cet angle. Enfin à travers la notion d'indicatrice floue induite naturellement par l'estimation de la densité, nous proposons un nouveau formalisme et des applications heuristiques.

P L A N

Chapitre I : ESTIMATION DE LA DENSITE MULTIVARIEE PAR LES NOYAUX DE PARZEN

1. Introduction

a) En 1956 Murray Rosenblatt proposa un estimateur non paramétrique d'une densité réelle continue [9] sous la forme d'une généralisation du concept d'histogramme appelé histogramme décalé (shifted histogram).

$$\hat{f}_n(x) = \frac{\#\text{ points de données dans }]x - h_n, x + h_n[}{2\, n\, h_n}$$

où h_n est une constante dépendant uniquement de n, nombre de points de donnée.
Rosenblatt montra entre autre que la convergence vers la densité sous-jacente était plus rapide que celle de l'histogramme classique.

Une autre écriture de l'histogramme décalé de Rosenblatt est la suivante :

$$\hat{f}_n(x) = \frac{1}{n} \sum_{j=1}^{n} \frac{1}{h_n} w\left(\frac{x - X_j}{h_n}\right)$$

$$\text{où } w(j) = \frac{1}{2} \quad \text{si } |u| < 1$$

$$= 0 \quad \text{sinon}$$

Cette écriture amena Parzen [8] à proposer l'estimateur suivant :

$$\hat{f}_n(x) = \frac{1}{n} \sum_{j=1}^{n} \frac{1}{h_n} K\left(\frac{x - X_j}{h_n}\right)$$

où K est une densité de probabilité encore appelée noyau h_n est alors appelée fenêtre (ou largeur de la fenêtre) $f_n(x)$ est l'estimation de la densité par les noyaux de Parzen.
Depuis plusieurs générations ont été proposées, en particulier celle de Deheuvels [2] dans le cas multivarié R^p.

$$\hat{f}_n(x) = \frac{|\det A_n|}{n} \sum_{i=1}^{n} K\left(A_n (Xi - x)\right)$$

où A_n est une matrice pxp inversible.

b) Nous étudierons et utiliserons dans ce qui suit essentiellement cette dernière forme que nous transformerons légèrement.

En effet, posons $h_n^p = \frac{1}{|\det A_n|}$

et $T_n = \frac{A_n}{h_n}$

Alors $\hat{f}_n(x) = \frac{1}{nh_n^p} \sum_{i=1}^{n} K\left(T_n \left(\frac{Xi-x}{h_n}\right)\right)$

Avec $|\det (T_n)| = 1$

Cette décomposition de la matrice A_n a une signification intuitive : h_n est un paramètre de "taille" du noyau : pour les noyaux à support compact, h_n^p est proportionnel à la surface du support de chaque noyau.

T_n est alors un paramètre de forme.

c) Dans le cours de cette étude nous serons fréquemment amené à mettre en évidence les moments d'ordre 1 et 2 des noyaux, qui pour tous les types usuels de noyaux les déterminent entièrement.

Nous noterons alors :

$$f_n(x) = \frac{1}{n} \sum_{i=1}^{n} \Delta (x \,|xi, C_n)$$

où $\Delta(x \,|xi, C_n) = |\det A_n| \; K \; [A_n(xi - x)]$

et $\int x \; \Delta(x/xi, C_n)\, dx = xi$

$\int (x - xi)(x - xi) \; \Delta(x/xi, C_n)\, dx = C_n$

Un cas particulier important est le cas du noyau gaussien

$$K(x) = \frac{1}{(2\pi)^{p/2}} \exp\left\{-\frac{1}{2} x^t x\right\}$$

ou $\Delta(x/xi, C_n) = \frac{1}{(2\pi)^{p/2} \sqrt{\det C_n}} \exp\left\{-\frac{1}{2} (x - xi)^t C_n^{-1} (x - xi)\right\}$

Nous pouvons alors donner une relation entre A_n (ou hn, T_n) et C_n

$$C_n^{-1} = A_n . A_n^t = \frac{1}{h_n^2} \quad T_n \quad T_n^t$$

$$h_n^p = \det C_n$$

Ce choix d'un moyen multinormal n'est certes pas optimal au point de vue de l'efficacité asymptotique (voir [2] p 14) néanmoins ses propriétés de régularité (continuité uniforme, dérivabilité) et ses propriétés de calculabilité pratique et théorique (moments, moments conditionnels) font qu'il vérifie les hypothèses de la majorité des théorèmes de convergence et rendent son maniement agréable.

d) Nous allons dans un premier temps proposer un théorème de convergence uniforme avec une matrice forme dépendant des données puis en donner les généralisations heuristiques.

Dans un deuxième temps nous proposerons un algorithme du choix de la fenêtre qui optimise asymptotiquement l'erreur moyenne quadratique intégrée (M I S E) dans le cadre d'une estimation avec des noyaux gaussiens.

2. Un théorème de convergence

a) Dans [14] Wagner souligne un défaut de l'estimation de densité usuelle : le coefficient de lissage h_n est choisi indépendamment des points de donnée observés. Ce point de vue a été généralisé dans [4] où Devroye et Wagner obtiennent le résultat suivant.

Théorème 1

Soit f uniformément continue. Soit K une densité de probabilité sur R^p bornée, Riemann intégrable telle que (1)

$$t^{p-1} L(t)\, dt < \infty \qquad \text{où} \quad L(t) = \sup_{\|x\| > t} K(x)$$

Soit $H_n (X_1, \ldots, X_n)$ une variable aléatoire fonction des variables aléatoires mutuellement indépendantes $X_1, \ldots, X_n$ issus de f.

alors si $H_n \xrightarrow[n \to \infty]{} 0$ presque surement (2) (respectivement en probabilité)

et $\dfrac{n H_n^{2p}}{\log n} \xrightarrow{n} \infty$ ps (respectivement $nH_n^{2p} \to \infty$ en probabilité)

on a $\sup_x \left| \hat{f}_n(x) - f(x) \right| \to 0$ ps (respectivement en probabilité)

où $$\hat{f}_n(x) = \frac{1}{nH_n^p} \sum_{i=1}^{n} K\left(\frac{x - X_i}{H_n}\right)$$

Remarques

i) L'hypothèse d'uniforme continuité de la densité f(x) peut paraître très lourde, cependant dans le cas unidimensionnel Schuster [12] montre que cette hypothèse est nécessaire pour une convergence uniforme.

ii) Le noyau normal vérifie l'hypothèse $\int_0^\infty t^{p-1} L(t)\, dt < \infty$ ains que tout noyau à support compact.

iii) Pour un choix raisonnable de H_n, $H_n(x_1, \ldots, x_n)$ est invariant par toute permutation de ses variables.

iv) Un choix de H_n proposé par Wagner et vérifiant les hypothèses du théorème est le suivant :

Soit $k(n) = [n^\alpha]$ (le crochet désigne la partie entière)

$0 < \alpha < 1$, soit D_{jn} la distance de X_j à son k(n)-ième plus proche voisin.

H_n est alors choisi au hasard parmi $D_{1n}, \ldots, D_{nn}$

(1) *Dans tout le I 2. x désignera* $\sup_{i=1,p} \{|x(i)|\}$

(2) *Par la suite nous utiliserons l'abréviation usuelle ps*

Dans l'optique du 1.b le théorème 1 permet donc un choix de la "taille" de la fenêtre en fonction des données.

L'idée du résultat suivant est de permettre à la "forme" de le fenêtre de s'adapter aux données, en introduisant la matrice de covariance empirique globale de l'échantillon.

b) Théorème 2

Soit $X_1, \ldots, X_n$ une suite de vecteurs aléatoires, indépendants, identiquement distribués à valeurs dans R^p avec une densité de probabilité f que l'on supposera uniformément continue.

Soit $T_n\ (X_1, \ldots, X_n)$ une suite de matrices aléatoires p x p telles que $\det(T_n) = 1$ qui converge presque sûrement vers une matrice T constante.

Soit K une densité de probabilité uniformément continue vérifiant les conditions suivantes :

$$L(t) = \sup_{\|x\| > t} K(T.x) \quad \text{vérifie} \quad \int t^{p-1} L(t)\, dt < \infty$$

$$L_n(t) = \sup_{\|x\| > t} K(T_n.x) \quad \text{vérifie} \quad \int t^{p-1} L_n(t)\, dt < \infty$$

Soit (h_n) une suite de réels positifs telle que

$$h_n \longrightarrow 0 \qquad \frac{n h_n^{2p}}{\log n} \longrightarrow \infty$$

$$n \longrightarrow \infty \qquad n \longrightarrow \infty$$

Alors $\hat{f}_n(x) = \frac{1}{n h_n^p} \sum_{i=1}^{n} K\left(T_n\left(\frac{x - X_i}{h_n}\right)\right)$ converge uniformément presque sûrement vers f(x) c'est-à-dire

$$\sup_{x \in R^p} |f(x) - \hat{f}_n(x)| \longrightarrow 0 \quad (n \to \infty)$$

Application

Avant de démontrer ce théorème, donnons en l'application annoncée :

Soit C_n la matrice de covariance empirique du nuage des points de donnée. Nous supposerons que C_n est inversible, elle le sera à partir d'un certain rang si f admet une matrice des moments d'ordre 2 inversible. (Si cela n'était pas le cas, il existerait un sous espace propre de R^p contenant tous les points de donnée que l'on pourrait déterminer par une analyse en composantes principales, on chercherait alors une densité dans ce sous-espace)

Soit $A_n = C_n^{-1} = P_n^{-1} D_n P_n$ avec $D_n = \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_p \end{bmatrix}$

Posons $T_n = \frac{1}{\left(\prod_{i=1}^{p} \lambda_i\right)^{1/2p}} P_n^{-1} \begin{bmatrix} \sqrt{\lambda_1} & & 0 \\ & \ddots & \\ 0 & & \sqrt{\lambda_p} \end{bmatrix} P_n$

où P_n est la matrice orthogonale des vecteurs propres.

Donc d'après la loi forte des grands nombres, si f admet une matrice inversible des moments d'ordre 2

$C_n \xrightarrow[n \to \infty]{ps} C$ de même que $P_n \xrightarrow[n \to \infty]{ps} P$ $D_n \xrightarrow[n \to \infty]{ps} D$

Donc $T_n \xrightarrow[n \to \infty]{ps} T$ donc T_n ainsi définie vérifie les hypothèses du théorème.

Notons que si K admet des moments du second ordre tels que $\int x\, x^t K(x)\, dx = I_p$ matrice identité d'ordre p alors

$$\int x\, x^t\, K(Tx)\, dx = \int (T_y^{-1})(T_y^{-1})^t K(y)\, dy = T^{-1} I_p (T^t)^{-1} = (T^t T)^{-1} = \frac{C}{(\det C)^{1/p}}$$

Nous avons donc intégré la covariance globale dans la covariance locale de chaque noyau.

Démonstration du théorème 2

Pour la démonstration du théorème nous utiliserons 2 lemmes fondamentaux et un lemme de calcul,

Lemme 1

Soit f une densité de probabilité uniformément continue sur $\mathbb{R}^p$

Soit K une densité de probabilité sur $\mathbb{R}^p$

Soit (h_n) une suite de réels positifs tendant vers 0

Soit

$$g_n(x) = \frac{1}{h_n^p} \int K\left(\frac{x - y}{h_n}\right) f(y)\, dy$$

Alors $\sup_{x \in \mathbb{R}^p} |g_n(x) - f(x)| \xrightarrow[n \to \infty]{} 0$

Démonstration

Posons $t = x - y$

$$g_n(x) = \frac{1}{h_n^p} \int K\left(\frac{t}{h_n}\right) f(x - t)\, d^t$$

or $f(x) = f(x) \cdot \frac{1}{h_n^p} \int K\left(\frac{t}{h_n}\right) dt$

Donc $|f(x) - g_n(x)| \leqslant \int |f(x - t) - f(x)| \frac{1}{h_n^p} K(\frac{t}{h_n})\, dt$

Soit $\delta > 0$

$$\sup_{x \in \mathbb{R}^p} |f(x) - g_n(x)| \leqslant \sup_{x \in \mathbb{R}^p} \int_{\|t\| < \delta} |f(x) - t) - f(x)| \frac{1}{h_n^p} K(\frac{t}{h_n})\, dt$$

$$+ \sup_{x \in \mathbb{R}^p} \int_{\|t\| \geqslant \delta} |f(x - t) - f(x)| \frac{1}{h_n^p} K(\frac{t}{h_n}) dt$$

Dans le premier terme nous pouvons majorer $|f(x - t) - f(x)|$ par

$$\sup_x \sup_{\|t\| < \delta} |f(x - t) - f(x)|$$

Dans le deuxième terme $|f(x - t) - f(x)|$ est majoré par M où M est un majorant de $f(x)$.

$$\sup_{x \in \mathbb{R}^p} |f(x) - g_n(x)| \leqslant \sup_{x \in \mathbb{R}^p} \sup_{\|t\| < \delta} |f(x - t) - f(x)| + M \int_{\|t\| > \frac{\delta}{h_n}} K(y)\, dy$$

Soit $\varepsilon > 0$, comme f est uniformément continue en choisissant δ assez petit, le premier terme est inférieur à $\frac{\varepsilon}{2}$

Comme $h_n \to 0$, pour δ fixée on peut choisir h_n assez petit pour que le deuxième terme soit aussi inférieur à $\frac{\varepsilon}{2}$

donc $\forall \varepsilon > 0, \exists N > 0, \forall n > N, \sup_{x \in \mathbb{R}^p} |f(x) - g_n(x)| < \varepsilon$

c'est-à-dire $\sup_{x \in \mathbb{R}^p} |f(x) - g_n(x)| \longrightarrow 0$ $\quad n \to \infty$

Lemme 2

Soit K une fonction uniformément continue intégrable sur $\mathbb{R}^p$, positive ou nulle.

Soit (T_n) une suite de matrices inversibles p x p convergeant vers une matrice T inversible.

Alors $\forall \delta, \rho > 0$

$\exists N_0 \in \mathbb{N}$ et $\exists K^{*}(x) = \sum_{i=1}^{N} i.I_{Ai}(x)$, $\exists D \subset \mathbb{R}^p$

tel que $\forall n \quad N$

i) $\alpha 1, \ldots, \alpha_N \in \mathbb{R}^+$

ii) $A_1, \ldots, A_N$ pavés disjoints dans $[-\rho, \rho]^p$

iii) $|K^{*}(x) \leftarrow k(T_n.x)| <$ sur $([-\rho, +\rho]^p - D)$

iv) $K^{*}(x) \leqslant \sup_{x \in \mathbb{R}^p} K(x) = \sup_x K(T_n.x)$ $\quad \forall n$

v) $D \subset B = \bigcup_{i=1}^{M} Bi$ $\quad B_1, \ldots, B_M$ pavés disjoints dans $[-\rho, \rho]^p$ et

$\lambda(B) < \delta$, où λ désigne la mesure de Lebesgue

Démontrons d'abord un résultat partiel de topologie.

Résultat : Sous les hypothèses du lemme 2 $(K(T_n.x))$ est une suite de fonctions uniformément convergente vers $K(T.x)$

La démonstration de ce résultat se fait en trois étapes mais fixons d'abord quelques notations.

Si $A \subset \mathbb{R}^p$, A^c désignera le complémentaire de A dans $\mathbb{R}^p$

Soit $\varepsilon > 0$ fixé.

1ère étape : Montrons qu'il existe un compact C_ε tel que $\forall x \in C_\varepsilon^c$ on ait $K(x) < \varepsilon$

or K est uniformément continue donc $\exists \quad > 0$ tel que $\forall x,x'$ vérifiant $\|x - x'\| < n$ on ait $|K(x) - K(x')| < \frac{\varepsilon}{2}$

or on peut construire un recouvrement de $\mathbb{R}^p$ de pavés P_n de la forme

$\prod_{i=1}^{p} [x(i) - \frac{1}{4}, x(i) + \frac{1}{4}]$ tels que $\lambda(P_n \cap P_{n'}) = 0$ pour $n \neq n'$

Soit P_n un de ces pavés, s'il existe $x_0 \in P_n$ tel que $K(x_0) \geq \varepsilon$ on a $\forall x \in P_n$, $K(x) \geq \frac{\varepsilon}{2}$

comme K(x) est une fonction positive ou nulle intégrable, ces pavés sont en nombre fini, leur réunion est donc impacte d'où l'existence de C_ε

Considérons maintenant les ensembles $C_n(\varepsilon) = T_n^{-1}(C_\varepsilon) = \{x, T_n.x \in C_\varepsilon\}$

2ème étape : Montrons que $\bigcup_{n \in \mathbb{N}} C_n(\varepsilon)$ est contenue dans un compact.

T_n est une application linéaire en dimension finie, elle est donc bicontinue et par conséquent $C_n(\varepsilon)$ est compact.

Comme $T_n \longrightarrow T$ on a $\lim_{n \to \infty} C_n(\varepsilon) = T^{-1}(C_\varepsilon) = C_0(\varepsilon)$ compact

donc $\exists M > 0$ tel que $\forall x \in C_0(\varepsilon)$, $\|x\| < M$

donc $\forall \varepsilon' > 0, \exists N \in \mathbb{N}, \forall n > N \quad x \in C_n(\varepsilon) \Longrightarrow \|x\| < M + \varepsilon'$

Posons $C'_1 = \{x, \|x\| < M + \varepsilon'\}$

d'après la définition de N et de C_1 on a donc

$C'_1 \supset C_n(\varepsilon), \forall n > N$

donc la réunion de tous les $C_n(\varepsilon)$ est contenue dans un compact, son adhérence est donc compacte.

posons $C'_2(\varepsilon) = \overline{\bigcup_{n=1}^{\infty} C_n(\varepsilon)}$

3ème étape : convergence uniforme

Soit $x \in C'_2(\frac{\varepsilon}{2})$

$$\|T_n.x - T.x\| \leq \|T_n - T\| \cdot \|x\| \leq \|T_n - T\| \sup_{x \in C'_2(\frac{\varepsilon}{2})} \|x\|$$

donc $T_n.x$ converge uniformément vers $T.x$ sur $C'_2(\frac{\varepsilon}{2})$

comme K est uniformément continue on a de même $K(T_n.x)$ converge uniformément

en x vers $K(T.x)$ sur $C_2'(\frac{\varepsilon}{2})$ or si $x \in C_2'(\frac{\varepsilon}{2})^c$ par construction $T_n.x \in C^c_{\frac{\varepsilon}{2}}$ et $T.x \; C^c_{\frac{\varepsilon}{2}}$

c'est-à-dire $K(T_n.x) \leqslant \frac{\varepsilon}{2}$ et $K(T.x) \leqslant \frac{\varepsilon}{2}$ donc $|K(T_n.x) - K(T.x)| \leqslant \varepsilon$

donc $K(T_n.x)$ converge uniformément en x vers $K(T.x)$ sur $C_2'(\frac{\varepsilon}{2})^c$
donc $K(T_n.x)$ converge uniformément vers $K(T.x)$ sur $\mathbb{R}^p$ tout entier.

Démonstration du lemme 2

Soient η, δ, $\rho > 0$ fixés

$K(Tx)$ est continue intégrable, elle est donc Riemann intégrable en particulier sur $[-\rho, +\rho]^p$, il existe donc un partionnement de $[-\rho, +\rho]^p$ en pavés disjoints tel que les sommes de Riemann supérieures définies par une fonction en escalier majorante K1 et une fonction en escalier minorante K_2, différent d'au plus $\frac{\delta\eta}{3}$

c'est-à-dire $\int (K_1(x) - K_2(x))dx \leqslant \frac{\delta\eta}{3}$ (*)

avec $\forall x \in [-\rho,\rho]^p \quad K_1(x) \geqslant K(T.x) \geqslant K_2(x)$

Posons maintenant

$$K'_1(x) = K_1(x) + \frac{\eta}{4}$$

$$K'_2(x) = \sup(0, K_2(x) - \frac{\eta}{4})$$

Comme $K(T_n.x)$ converge uniformément vers $K(T.x)$, $\exists No$ tel que $\forall n > No$, $|K(T_n.x) - K(T.x)| < \frac{\eta}{4} \quad \forall x \in \mathbb{R}^p$ or $K(T_n.x) \geqslant 0$, nous avons donc les inégalités suivantes :

$$K'_1(x) > K(T_n.x) \geqslant K'_2(x)$$

Montrons que $K^* = K'_2$ vérifie les propriétés i) à v) du lemme 2, en effet i) et ii) sont vérifiées par construction.

$\{x, K(T_n.x) - K'_2(x) \geqslant \eta\} \subset \{x, K'_1(x) - K'_2(x) \geqslant \eta\}$

$\subset \{x, K'_1(x) - K_1(x) \geqslant \frac{\eta}{3}\} \cup \{x, K_1(x) - K_2(x) \geqslant \frac{\eta}{3}\} \cup \{x, K_2(x) - K'_2(x) \geqslant \frac{\eta}{3}\}$

or $K'_1(x) - K_1(x) = \frac{\eta}{4} < \frac{\eta}{3}$ et $K_2(x) - K'_2(x) \leqslant \frac{\eta}{4} < \frac{\eta}{3}$

donc $\{x, K(T_n.x) - K'_2(x) \geqslant \eta\} \subset \{x, K_1(x) - K_2(x) \geqslant \frac{\eta}{3}\}$

$\lambda(\{x, K_1(x) - K_2(x) \geqslant \frac{\eta}{3}\}) \leqslant \frac{3}{\eta} \int (K_1(x) - K_2(x))\, dx \leqslant \delta$ (d'après *)

Les Bi sont donc les pavés disjoints du partitionnement tels que

$$K_1(x) - K_2(x) \geqslant \frac{\eta}{3}$$

La mesure de leur réunion est donc inférieure à δ d'où les propriétés (iii) et (v). (iv) enfin est vérifié car $K(T.x) \geqslant K_2(x) \geqslant K'_2(x)$

Lemme 3

Notons $S(x,\rho h)$ la sphère centrée en x, de rayon ρh pour la norme du sup, $S(x,\rho h)^c$ son complémentaire dans $\mathbb{R}^p$.

Soit M_1 un majorant de $f(x)$, $L(t)$ une fonction positive de la variable réelle t.

Alors

$$\sup_{x,h} \int_{S(x,\rho h)^c} h^{-p} L(\|\frac{x-y}{h}\|_\infty) f(y)\, dy \leqslant M_1 \int_{+\rho}^{+\infty} 2p\,(2t)^{p-1} L(t)\, dt$$

si cette dernière intégrale existe

Démonstration du lemme 3

$f(x) \leqslant M_1$

donc $$\int_{S(x,\rho h)^c} h^{-p} L(\|\frac{y-x}{h}\|) f(y)\, dy \leqslant M_1 \int_{S(x,\rho h)^c} h^{-p} L(\|\frac{y-x}{h}\|)\, dy$$

posons $\zeta = \frac{y-x}{h}$

$$\int_{S(x,\rho h)^c} h^{-p} L(\|\frac{y-x}{h}\|)\, dy = \int_{S(0,\rho)^c} L(\|\zeta\|)\, d\zeta$$

$$= 2 \sum_{i=1}^{p} \int_{\rho}^{+\infty} \left[\int_{-x(i)}^{x(i)} dx(1) \ldots \int_{-x(i)}^{x(i)} dx(p)\right] L(x(i))\, dx(i)$$

$$= 2p \int_{\rho}^{+\infty} (tt)^{p-1} L(t)\, dt$$

d'où le résultat

Démonstration du théorème

Pour la démonstration du théorème fixons quelques notations..Notons F la fonction de répartition associée à $f(x)$, et μ la mesure correspondante. Notons F_n la fonction de répartition empirique et μ_n la mesure correspondante. Si $A \subset \mathbb{R}^p$, notons $A(x,h) = \{z, z = x + th,\ t \in A\}$

Notons M_1 un majorant de $f(x)$ et M_2 un majorant de $K^*(x)$ (donné par le lemme 2). Notons encore $\mathcal{C}$ la classe d'ensembles de tous les pavés de $\mathbb{R}^p$.

Avec ces notations nous pouvons écrire $f_n(x)$ sous la forme suivante

$$f_n(x) = \int h_n^{-p} K(T_n(\frac{y-x}{hn}))\, d F_n(y)$$

Enfin posons

$$g_n(x) = h_n^{-p} \int K\left(T\left(\frac{y-x}{h_n}\right)\right) dF(y)$$

$$\sup_x |f_n(x) - f(x)| < \sup_x |f_n(x) - g_n(x)| + \sup_x |g_n(x) - f(x)|$$

or d'après le lemme 1

$$\sup_x |g_n(x) - f(x)| \to 0 \quad n \to \infty$$

il suffit de montrer que $\sup_x |f_n(x) - g_n(x)| \overset{ps}{\to} 0 \quad n \to \infty$

Pour cela majorons

$$\sup_x |g_n(x) - f_n(x)| = h_n^{-p} \sup_x \left| \int K\left(T\left(\frac{y-x}{h_n}\right)\right) dF(y) - \int K\left(T_n\left(\frac{y-x}{h_n}\right)\right) dF_n(y) \right.$$

En applicant le lemme 2 nous avons l'inégalité suivante

$$\sup_x |g_n(x) - f_n(x)| < \sum_{i=1}^{3} \sup_x Ui(x)$$

avec $$U_1(x) = h_n^{-p} \int \left| K\left(T_n\left(\frac{y-x}{h_n}\right)\right) - K^*\left(\frac{y-x}{h_n}\right) \right| dF(y)$$

$$U_2(x) = h_n^{-p} \left| \int K^*\left(\frac{y-x}{h_n}\right) dF(y) - \int K^*\left(\frac{y-x}{h_n}\right) dF_n(y) \right|$$

$$U_3(x) = h_n^{-p} \int \left| K\left(T\left(\frac{y-x}{h_n}\right)\right) - K^*\left(\frac{y-x}{h_n}\right) \right| dF_n(y)$$

Majorons successivement U_1, U_2, U_3, pour cela posons :

$$C_1 = S(x, \rho h_n)^c$$

$$C_2 = S(x, \rho h_n) \cap D(x, h_n)^c$$

$$C_3 = D(x, h_n)$$

où l'ensemble D est donné par le lemme 2

1) $$U_{11}(x) = \frac{1}{h_n^p} \int_{C_1} \left| K\left(T_n\left(\frac{y-x}{h_n}\right)\right) - K^*\left(\frac{y-x}{h_n}\right) \right| dF(y)$$

si $y \in C_1$, $\frac{y-x}{h_n} \in [-\rho, \rho]^c$ donc $K^*\left(\frac{y-x}{h_n}\right) = 0$

$$U_{11}(x) < \frac{1}{h_n^p} \sup_x \int K\left(T_n\left(\frac{y-x}{h_n}\right)\right) dF(y)$$

$$U_{12}(x) = h_n^{-p} \int_{C_2} \left| K\left(T_n\left(\frac{y-x}{h_n}\right)\right) - K^*\left(\frac{y-x}{h_n}\right) \right| dF(y)$$

$\frac{y-x}{h_n} \in [-\rho, +\rho]^p$ donc d'après iii) du lemme 2

$$\left| K\left(T_n\left(\frac{y-x}{h_n}\right)\right) - K^*\left(\frac{y-x}{h_n}\right) \right| < \eta$$

donc $U_{12}(x) < \frac{\eta M_1}{h_n^p} \sup_x \lambda(C_2) < \frac{\eta M_1}{h_n^p} \sup_x \lambda(S(x,\rho h))$

$$U_{12}(x) < \frac{\eta M_1}{h_n^p} (\rho h_n)^p = \eta M_1 \rho^p$$

$$U_{13}(x) = \frac{1}{h_n^p} \int_{C_3} \left| K(T_n . (\frac{y-x}{h_n})) - K^*(\frac{y-x}{h_n}) \right| dF(y)$$

or d'après iv du lemme 2 $K(x) \leqslant M_2$ et $K^*(x) \leqslant M_2$

donc $U_{13}(x) \leqslant \frac{2M_1M_2}{h_n^p} \lambda(D(x,h_n))$

$$\leqslant 2M_1M_2\delta$$

2) $U_2(x) = h_n^{-p} \left| \sum_{i=1}^{N} \int_{Ai(x,h_n)} \alpha i \, dF(y) - \sum_{i=1}^{N} \int_{Ai(x,h_n)} \alpha i \, dF_n(y) \right|$

or $\alpha i \leqslant M_2$

$U_2(x) = MM_2 h_n^{-p} \sup_{A \in \mathcal{A}} |\mu n(A) - \mu(A)|$

3) $U_{31}(x) = h_n^{-p} \int_{C_1} \left| K(T.(\frac{y-x}{h_n})) - K^*(\frac{y-x}{h_n}) \right| dF_n(y)$

$$\leqslant h_n^{-p} \int_{C_1} K(T.(\frac{y-x}{h_n})) \, dF_n(y) \quad \text{car} \quad K^*(\frac{y-x}{h_n}) = 0$$

$$U_{32}(x) = h_n^{-p} \int_{C_2} \left| K(T.(\frac{y-x}{h_n})) - K^*(\frac{y-x}{h_n}) \right| dF_n(y)$$

sur $D^c \cap [-\rho,\rho]^p \quad \left| K(T.\frac{y-x}{h_n})) - K^*(\frac{y-x}{h_n}) \right| < \eta$

$$U_{32}(x) \leqslant \frac{\eta}{h_n^p} \mu n\,(S(x,\rho h)) < \frac{\eta}{h_n^p} \left| \mu n(S(x,\rho h)) - \mu(S(x,\rho h)) \right| + \frac{\eta}{h_n^p} \mu(S(x,\rho h))$$

$$\leqslant \frac{\eta}{h_n^p} \sup_A |\mu n(A) - \mu(A)| + \frac{\eta}{h_n^p} M_1 (2\rho h_n)^p$$

$$\leqslant \frac{\eta}{h_n^p} \sup_A |\mu n(A) - \mu(A)| + \eta M_1 (2\rho)^p$$

$$U_{33}(x) = \frac{1}{h_n^p} \int_{C_3} \left| K(T.(\frac{y-x}{h_n})) - K^*(\frac{y-x}{h_n}) \right| dF_n(y)$$

$$\leqslant \frac{1}{h_n^p} 2M_2 \sup_x (\mu n(D(x,h_n)))$$

or $D = \bigcup_{i=1}^{M} Bi$ où Bi sont des pavés disjoints

$$U_{33}(x) < \frac{2M_2}{h_n^p} \sup_x \left| \mu n(D(x,h_n) - \mu(D(x,h_n)) \right| + \frac{2M_2}{h_n^p} \sup_x \mu(D(x,h))$$

comme les pavés Bi sont disjoints

$$U_{33}(x) < \frac{2M_2M}{h_n^p} \sup_A |\mu_n(A) - \mu(A)| + 2M_1M_2\,\delta$$

Majorons maintenant la quantité H où

$$H = \frac{1}{h_n^p} \sup_x \int_{S(x,\rho h_n)^c} K\left(T_n.\left(\frac{y-x}{h_n}\right)\right) dF(y) + \frac{1}{h_n^p} \sup_x \int_{S(x,\rho h)^c} K\left(T_c.\frac{y-x}{h_n}\right) dF_n(y)$$

d'après leur définition

$$L(t) = \sup_{\|x\| > t} K(T.x)$$

$$L_n(t) = \sup_{\|x\| > t} K(T_n.x)$$

L(t) et $L_n(t)$ sont des fonctions décroissantes de t, en effet

si $t > t'$ $\quad \{x, \|x\| > t\} \subset \{x, |x| > t'\}$

donc $K(T.x) \leqslant L(\|x\|)$

et $\quad K(T_n.x) \leqslant L_n(\|x\|)$

$$H \leqslant \frac{1}{h_n^p} \sup_x \int_{S(x,\rho h_n)^c} L_n\left(\left\|\frac{y-x}{h_n}\right\|\right) dF(y) + \frac{1}{h_n^p} \sup_x \int_{S(x,\rho h_n)} L\left(\left\|\frac{y-x}{h_n}\right\|\right) dF(y)$$

$$+ \frac{1}{h_n^p} \sup_x \left| \int_{S(x,\rho h_n)^c} L\left(\left\|\frac{y-x}{h_n}\right\|\right) dF_n(y) - \int_{S(x,\rho h_n)^c} L\left(\left\|\frac{y-x}{h_n}\right\|\right) dF(y) \right|$$

En applicant le lemme 3

$$H \leqslant M_1 \int_\rho^{+\infty} 2p\,(2t)^{p-1} L(t)\,dt + M_1 \int_\rho^{+\infty} 2p\,(2t)^{p-1} L_n(t)\,dt + H_1$$

$$H_1 = \frac{1}{h_n^p} \sup_x \left| \int_{S(x,\rho h_n)^c} L\left(\left\|\frac{y-x}{h_n}\right\|\right) dF_n(y) - \int_{S(x,\rho h_n)^c} L\left(\left\|\frac{y-x}{h_n}\right\|\right) dF(y) \right.$$

posons $L'(t) = L(t)\, I_{\{t \geqslant \rho\}}$

$$H_1 = \frac{1}{h_n^p} \sup_x \left| L'\left(\left\|\frac{y-x}{h_n}\right\|\right) dF_n(y) - L'\left(\left\|\frac{y-x}{h_n}\right\|\right) dF(y) \right|$$

Soit $l \in \mathbb{N}$ arbitraire

Notons alors pour $j \leqslant l$ $\quad S_j = \{x, (j-1)\frac{L(\rho)}{l} < L(|x|) \leqslant \frac{j}{l} L(\rho)\}$

et $\quad T_j = \{x, \frac{(j-1)}{l} L(\rho) < L(\|x\|) < L(\rho)\} = \bigcup_{j=1}^{l} S_j$

Si on pose $\quad L''(x) = \sum_{j=1}^{l} (j-1)\,\frac{L(\rho)}{l}\, I_{Sj}(x)$

on a $\quad \left|L'(\|x\|) - L''(x)\right| < \frac{L(\rho)}{l}$

Compte tenu de ces notations

$$H_1 \leq h_n^{-p} \Big(\sup_x \int \Big| L'\Big(\frac{y-x}{h_n}\Big) - L'\Big(\Big\|\frac{y-x}{h_n}\Big\|\Big)\Big| \, dF_n(y)$$

$$+ \sup_x \Big| \int L''\Big(\frac{y-x}{h_n}\Big) \, dF_n(y) - \int L''\Big(\frac{y-x}{h_n}\Big) \, dF(y)\Big|$$

$$+ \sup_x \int \Big| L''\Big(\frac{y-x}{h_n}\Big) - L'\Big(\Big\|\frac{y-x}{h_n}\Big\|\Big)\Big| \, dF(y)\Big)$$

$$\leq h_n^{-p} \Big(\frac{2L(\rho)}{l} + \frac{L(\rho)}{l} \sup_x \Big| \sum_{j=1}^{l} (j-1)(\mu_n(S_j(x,h_n)) - \mu(S_j(x,h_n)))\Big|\Big)$$

$$\leq h_n^{-p} \Big(\frac{2L(\rho)}{l} + \frac{L(\rho)}{l} \sup_x \Big| \sum_{j=1}^{l} [\mu_n(Tj(x,h_n)) - \mu(Tj(x,h_n))]\Big|\Big)$$

D'après la définition des ensembles T_j, ces ensembles sont la différence de 2 pavés centrés en 0.

Donc $H_1 \leq \dfrac{2L(\rho)}{l h_n^p} + \dfrac{2L(\rho)}{h_n^p} \sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)|$

finalement

$$\sup_x |f_n(x) - g_n(x)| \leq 2pM_1 \int_\rho^{+\infty} (L(t) + L_n(t))(2t)^{p-1} dt + 2\eta M_1 (2\rho)^p$$

$$+ 4M_1M_2\delta + \frac{2L(\rho)}{h_n^p l} + (NM_2 + \eta + 2MM_2 + 2L(\rho))\frac{1}{h_n^p} \sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)|$$

l étant arbitraire on peut choisir l comme étant la partie entière de h_n^{-p}

Donc comme $\lim_{\rho \to +\infty} L(\rho) = 0$ on peut choisir ρ assez grand pour que $\dfrac{L(\rho)}{h_n^p . l}$ soit petit

de même que $\displaystyle\int_{+\rho}^{+\infty} (L(t) + L_n(t)) (2t)^{p-1} dt$

En choisissant η et δ suffisamment petit on a pour $\varepsilon > 0$ donné et en posant

$$C = NM_2 + \eta + 2MM_2 + 2L(\rho)$$

$$\sup_x |f_n(x) - g_n(x)| \leq \frac{\varepsilon}{2} + \frac{C}{h_n^p} \sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)| \qquad (1)$$

or $C . \sup_A |\mu_n(A) - \mu(A)| \leq C' . \sup_x |F_n(x) - F(x)|$

où C' est une constante inférieure ou égale à $2^p C$, donc

$$P\Big[\frac{C}{h_n^p} \sup_A |\mu_n(A) - \mu(A)| \geq \frac{\varepsilon}{2}\Big] \leq P\Big[\sup_x |F_n(x) - F(x)| \geq \frac{\varepsilon h_n^p}{C'}\Big] \qquad (2)$$

D'après une inégalité de Kiefer et Wolfowitz (voir [6]) qui s'écrit :

$\forall p, \exists c_0, c_1 > 0, \forall r > 0, \forall F, \forall n$

$$P\Big(\sup_x |F_n(x) - F(x)| > \frac{r}{\sqrt{n}}\Big) < c \exp(-c_1 r^2)$$

Nous pouvons écrire

$$P(\sup_x |F_n(x) - F(x)| \geqslant \frac{\varepsilon h_n^p}{c'}) \leqslant c_0 \exp\{-\varepsilon c_1 n h_n^{2p}\}$$

or $\lambda n = \dfrac{n h_n^{2p}}{\log_n}$ tend vers l'infini par hypothèse

donc en tenant compte de (1) et de (2), $c'_0, c'_1 \in R_+^*$

$$P[\sup_x |f_n(x) - g_n(x)| \geqslant \varepsilon] \leqslant c'_0 \exp\{-\varepsilon c'_1 \lambda n \log n\}$$

$$\leqslant c'_0 \left(\frac{1}{n}\right)^{\varepsilon c'_1 \lambda n}$$

Or cette dernière expression est le terme général d'une série convergente, en applicant le lemme de Borel-Cantelli nous pouvons conclure :

$$P\left[\limsup_n \sup_x |f_n(x) - g_n(x)| \geqslant \varepsilon\right] = 0$$

c'est-à-dire

$$P\left[\sup_n |f_n(x) - g_n(x)| \to 0\right] = 1$$

Ce qui démontre le théorème 2

c) Vitesse de convergence pour une fonction lipschitzienne

Nous nous proposons de majorer asympotiquement la quantité $\sup_x |f_n(x) - f(x)|$ et d'en déduire un ordre de convergence de h_n optimal dans ce sens.

$$\sup_x |f_n(x) - f(x)| \leqslant \sup_x |f(x) - g_n(x)| + \sup_x |g_n(x) - f(x)|$$

en appliquant le lemme 1 du théorème 1 - 2, nous obtenons

$$\sup_x |f(x) - g_n(x)| \leqslant \sup_x \sup_{\|t\| < \delta} |f(x-t) - f(x)| + 2M \int_{\|t\| \geqslant \frac{\delta}{h_n}} K(y)dy$$

Considérons un δ dépendant de n et notons δ_n; pour que le 2ème terme tende vers 0, il faut la condition $\dfrac{\delta_n}{h_n} \to 0$ $\quad n \to \infty$

Si f est lipschitzienne de rapport λ (cette condition n'est guère plus forte que l'uniforme continuité) alors

$$\sup_x \sup_{\|t\| < \delta_n} |f(x - t) - f(x)| < \lambda . \delta_n$$

ceci majore le premier terme, majorons le deuxième terme dans deux cas particuliers

- Si K est de support compact alors comme $\frac{\delta_n}{h_n} \to \infty$

le deuxième terme est nul à partir d'un certain rang.

- Si K est le noyau gaussien

$$\int_{\|t\| > \frac{\delta_n}{h_n}} K(y)\, dy \leqslant \frac{1}{\sqrt{2\pi}} \int_{\frac{\delta_n}{h_n}}^{+\infty} \exp\left\{-\frac{t^2}{2}\right\} dt$$

$$\leqslant \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{\delta_n^2}{h_n^2}\right\}$$

Or, dans la démonstration du théorème 1 - 2, nous obtenons le résultat suivant :

si $\epsilon_n \cdot \lambda_n \to \infty$

c'est à dire $\dfrac{n}{\dfrac{\log n}{n\, h_n^{2p}}} \to \infty$

alors $P\left\{ \sup_x |f_n(x) - g_n(x)| < \epsilon_n \right\} \to 1$

donc avec une probabilité tendant vers 1

$$\sup_x |f_n(x) - f(x)| < 0(\delta_n) + 0(\epsilon_n)$$

supposons $h_n = 0(n^{-\alpha})$

$\delta_n = 0(n^{-\beta})$

$\epsilon_n = 0(n^{-\gamma})$

Les différentes conditions s'expriment alors en inégalités sur , ,

$$0 < \beta < \alpha < \frac{1}{2p}$$

$$0 < \gamma < 1 - 2p\alpha$$

Avec inf (γ , β) le plus grand possible. Au vu des inégalités précédentes, ceci est réalisé quand $\alpha = 1 - 2p\alpha$, c'est à dire

$$\alpha = \frac{1}{2p + 1}$$

la meilleure majoration est donc réalisé pour r

$$\sup_x |f_n(x) - f(x)| < O(n^{-r}) \quad , \quad r < \frac{1}{2p+1}$$

D) Justification du choix précédent

Dans [2] Deheuvels examine les propriétés du MISE (voir 3) et obtient en particulier la propriété que si f (x) admet les hyperplans de coordonnées comme plan de symétrie alors la matrice Tn optimale est diagonale. Cette propriété est conservée si l'on considère le noyau proposé dans l'application du théorème ; en effet, sous la condition précédente, la matrice de covariance est diagonale.

E) Généralisations heuristiques

α) Considérons maintenant deux généralisations des méthodes précédentes avec des noyaux variant avec les points de donnée.

α) Si f est un mélande de deux lois symétriques, alors le résultat précédent cité en d) conduit à utiliser deux noyaux différents qui rendent compte des symétries de chaque composante. Cela est d'autant plus évident que les composantes sont bien séparées. Etant donné cette remarque, nous proposons la méthode suivante : ajuster f par un mélange de lois symétriques par exemple multinormales en procédant à une classification préalable des points de donnée, puis choisir une fenêtre par classe qui rendent compte de la forme de la classe. Nous obtenons l'estimation de la densité suivante ou C (1), ... , C(N) sont les classes obtenues.

$$f_n(x) = \sum_{h=1}^{N} \sum_{x_i \in C(k)} \frac{1}{m.h_n^p} \; K\left(T_n^k . \frac{(x - x_i)}{h_n}\right)$$

β) Estimateur de Breiman et généralisation

La division en classes proposée au α) peut cependant être artificielle. Présentons maintenant un estimateur de la densité tel que la fenêtre change en chaque point. Cet estimateur a été proposé dans 1 par Breiman, Meisel et Purcell.

$$f_n(x) = \frac{1}{n} \sum_{j=1}^{n} (\alpha_k d_{j,k})^{-P} K\left(\frac{x-X_j}{\alpha_k d_{j,k}}\right)$$

où $d_{j,k}$ est la distance de Xj à son k-ième plus proche voisin α_k est une constante multiplicative dépendant de k .

Le choix proposé pour k et α_k est fondé sur une statistique d'ajustement multivarié que nous présenterons au chapitre II.

Des simulations de mélange de lois normales bivariées ont été effectuées et ont donné des résultats significativement meilleurs que l'estimation par les noyaux de Parzen usuels.

Cependant aucun résultat de convergence théorique n'est donné.

Nous en proposons maintenant une modification dans l'esprit du théorème 1 - 2

$$f_n(x) = \frac{1}{n} \sum_{j=1}^{n} (\alpha_k d_{j,k})^{-P} K\left(T_{j,k}\left(\frac{x-X_j}{\alpha_k, d_{j,k}}\right)\right)$$

où $T_{j,k}$ est une matrice p x p telle que

$$(T_{j,k}\ T_{j,}{}^{t}{}_{k})^{-1} = \frac{C_{j,k}}{(\det C_{j,k})^{1/P}}$$

$C_{j,k}$ est la matrice de covariance empirique des k plus proches voisins de X_j. L'estimateur ainsi défini introduit localement la forme de la densité ce que l'extimateur du théorène 1 - 2 fait globalement.

En utilisant la notation définie au lc mettant en évidence les 2 premiers moments de chaque noyau, nous pouvons écrire

$$f_n(x) = \frac{1}{n} \sum_{i=1}^{n} \Delta\left(x \mid x_i, \frac{d_k\, d_{j,k}}{\det(C_{j,k})^{1/p}}\ C_{j,k}\right)$$

3 - Choix de la fenêtre par le critère du MISE

Dans ce paragraphe, nous proposons un algorithme pour obtenir h_n asymptotiquement optimale pour le critère du MISE (mean integrated square error : erreur

quadratique moyenne intégrée) dans un cadre restreint aux noyaux produits gaussiens.

Nous n'utiliserons donc pas les résultats les plus généraux obtenus notamment par Deheuvels 2 .

a)Résultats

Nous donnons ici un résultat d'Epanechmikov 5 que nous avons légèrement développé dans le cas de noyaux gaussiens

Théorème 1 - 3

soit $f : R^p \mapsto R$ une densité multivariée admettant des dérivées du premier et du second ordre continues.

soit $X_1, \ldots, X_n$ un échantillon issu de f

$$X_i = \begin{bmatrix} X_i(1) \\ \\ X_i(p) \end{bmatrix}$$

soit $f_n(x) = \frac{1}{n\, h_n^p} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h_n}\right)$

où K est une densité de probabilité paire admettant des moments du second ordre non nuls

soit $D^2 f$ la matrice p x p des dérivées secondes de f

$$M^2(K,f)(x) = \int \cdots \int K(y)\, y^t\, D^2 f\, y\, dy$$

si $h_n \to 0$ et $n\, h_n \to \infty$

$n \to \infty$ $\quad$ $n \to \infty$

alors l'écart quadratique moyen (mean square error)

$$MSE(x) = E\,(f_n(x) - f(x))^2$$

$$= \frac{1}{n.h_n^p}\, \tilde{f}(x) \int K^2(y)\, dy + \frac{h_n^4}{4} . M^4 (K,f)(x) + o\left(\frac{1}{n.h_n^p} + h_n^4\right)$$

Démonstration

Comme f admet des dérivées premières et secondes continues, on peut écrire son développement de Taylor au voisinage de chaque point.

$$f(x+\epsilon y) = f(x) + \epsilon D_{f.y} + \frac{\epsilon^2}{2}\ y^t D^2 f\ y + o(\epsilon^2)$$

$$E\left(K\left(\frac{x-X}{h}\right)\right) = \int K\left(\frac{x-y}{h}\right) f(y)\, dy$$

$$= h^p \int K(y)\, f(x+h_y)\, d_y$$

$$= h^p \int f(x)\quad K(y) d_y + hD_f \int y\, K(y) d_y + \frac{h^2}{2} \int K(y) y^t D^2 f\ y\, dy + o(h^2)$$

Comme K est paire le moment d'ordre 1 est nul

$$\text{donc } E\left(K\left(\frac{x-Xj}{h}\right)\right) = h^p \left(f(x) + \frac{h^2}{2} M^2(K,f)(x) + o(h^2)\right)$$

$$\text{d'où } E(f_n(x)) = \frac{1}{n h_n^p} \sum_{i=1}^{n} E\left(K\left(\frac{x-Xi}{h_n}\right)\right) = f(x) + \frac{h^2_n}{2} M^2(K,f)(x) + o(h^2_n)$$

$$\text{et } E(f_n(x)) - f(x)) \sim \frac{h^2_n}{2} M^2(K,f)$$

$$MSE = E(f_n(x) - f(x))^2 = E(f_n^2(x)) - 2f(x).\, E(f_n(x)) + f(x)^2$$

$$= \frac{1}{n^2 h_n^{2p}} \left[n\, E\left(K^2\left(\frac{x-X}{h_n}\right)\right) + n(n-1)\, E\left(K\left(\frac{x-Xi}{h_n}\right).\, K\left(\frac{x-Xj}{h_n}\right)\right)\right]$$

$$- \frac{2 f(x)}{h_n} E\left(K\left(\frac{x-X}{h_n}\right)\right) + f(x)^2$$

$$\text{Posons } E f_n(x) - f(x) - \frac{h_n^2}{2} .\, M^2(K, f)(x) = L(x)$$

$$MSE(x) : \frac{f(x)}{n h_n^p} \left[\int K^2(y) dy + O(h^2_n)\right] + \frac{n-1}{n} \left[f(x)^2 + f(x).h_n^2\, M^2(K,f)(x)\right.$$

$$\left. + \frac{h_n^4}{4} M^4(K,f)(x) + 2f(x)\, L(x) + o(h_n^4)\right] - 2 f(x)^2 - f(x) h_n^2\, M^2(K,f)(x)$$

$$- 2 f(x)\, L(x) + f(x)^2$$

$$= \frac{1}{n h_n^p} f(x) \int K^2(y) dy + \frac{h_n^4}{4} M^4(K,f)(x) + o\left(\frac{1}{n h_n^p} + h_n^4\right)$$

Proposition 1

sous les hypothèses du théorème 1 - 3

$$MISE = \int E(f_n(x) - f(x))^2\, dx \sim \frac{1}{n h_n^p} \int K^2(y) dy + \frac{h_n^4}{4} \int M^4(K,f)(x) dx$$

Prosposition 2

1) Si $K(x) = k(x(1))\ldots k(x(p))$

tel que $\int_{\mathbb{R}} t^2 k(t)dt = 1$

alors $M^2(K,f) = \sum_{i=1}^{p} \frac{\partial^2 f}{\partial x(i)^2}$

2) si $K(x) = \frac{1}{(2\pi)^{p/2} \sqrt{\det C}} \exp\left\{-\frac{1}{2} x^t C x^{-1}\right\}$

alors $M^2(K,f) = tr(C.\ D^2 f)$

Remarque

Dans 1) les conditions $K(x) = k(x(1))\ldots k(x(p))$ et $\int_{\mathbb{R}^2} t^2 k(t)dt = 1$ fixent le paramètre de forme. Dans 2) ce paromètre dépend de la matrice C.

Démonstration

1) $M^2(K,f)(x) = \int K(y) y^t D^2 f\ y\ dy$

$$= \int \prod_{i=1}^{p} K(y(i)) \left(\sum_{j=1}^{p} \frac{\partial^2 f}{\partial x^2(j)} y(j)^2 + 2 \sum_{j<j'} \frac{\partial^2 f}{\partial x(j)\partial x(j')} y(j)y(j') \right) dy$$

comme k(x) est paire les termes en y(j), y(j') s'annulent

donc $M^2(K,f) = \sum_{i=1}^{p} \frac{\partial^2 f}{\partial x^2(i)} \int \prod_{j=1}^{p} k(y(j))\ y(i)^2\ dy(1)\ldots dy(p)$

$$= \sum_{i=1}^{p} \frac{\partial^2 f}{\partial x^2(i)} \prod_{j \neq i} \int k(y(j))dy(j) \int k(y(i))y(i)^2 dy(i)$$

$$= \sum_{i=1}^{p} \frac{\partial^2 f}{\partial x^2(i)}$$

2) $M^2(K,f)(x) = \int K(y)\ y^t\ D^2{}_f\ y\ dy$

$$= \frac{1}{(2\pi)^{p/2} \sqrt{\det C}} \int y^t D^2 f\ y \cdot \exp\left\{-\frac{1}{2}\ y^t C^{-1} y\right\} dy$$

soit T tel que $D^2{}_f = T^t T$, $y^t D^2{}_f y = (Ty)^t (Ty)$

posons $Y = T.y$

$$M^2(K,f)(x) = \frac{1}{|\det T|\,(2\pi)^{p/2}\sqrt{\det C}} \int y^t\, y \exp\left\{-\frac{1}{2}\, y^t C^{-1} y\right\} dy$$

ou $C'^{-1} = (T^{-1})^t\, C^{-1}\, T^{-1}$ $\qquad C' = T\, C\, T^t$

$$M^2(K,f)(x) = \frac{\sqrt{\det C'}}{|\det T|\sqrt{\det C}}\ \text{tr}\ C'$$

or $\det C' = (\det T)^2 \det C$

et $\text{tr}\ C' = \text{tr}\ (T.C.T.^t) = \text{tr}\ (T.T^t.C) = \text{tr}\ (C.D^2f)$

donc $M^2(K,f)(x) = \text{tr}\ (C.\ D^2f)$

Application

Si K est un noyau produit tel qu'il est défini dans la proposition 2, alors la fenêtre h_n minimisant asymptotiquement le critère du M I S E

$$\text{est } h_n = \left(\frac{p \int K^2(y)\, dy}{n \int \left(\sum_{i=1}^{p} \frac{\partial^2 f}{\partial (x(i))^2}\right)^2 dx}\right)^{1/p+4} \qquad (*)$$

démonstration

Il suffit de dériver l'expression du M I S E donnée par les propositions 1 et 2

B) Algorithme

α) Principe et justification

Dans [11] montre que pour r entier supérieur ou égal à 1 et p = 1 si les dérivées successives de K, $K^{(j)}$, j< r vérifient certaines conditions de régularité (qui sont vérifiées par le noyau gqussien)

si $f^{(r)}$ est uniformément continue

si $h_n \to 0$ et $\frac{\text{Log } h_n}{n h_n^{2r+1}} \to 0$ quand $n \to \infty$

alors $\sup_{x \in R} |f_n^{(r)}(x) - f(x)^{(r)}| \to 0$ presque sûrement et en probabilité

Donc sous certaines conditions, les dérivées de l'estimation de densité convergent uniformément presque sûrement vers les dérivées de la densité.

La convergence est uniforme donc si r=2 et si f vérifie les conditions ci-dessus, nous avons aussi

$\int f_n''(x)\,dx$ converge presque sûrement vers $\int f''(x)\,dx$.

La supposition heuristique à la base de l'algorithme suivante est que ces propriétés sont conservées pour des densités multivariées. Notons alors :

$$\beta(h) = \left(\int \left(\sum_{i=1}^{p} \frac{\partial^2 fn}{\partial x(i)^2} \right)^2 dx \right)$$

$$\alpha(K) = \int K^2(y)\,dy$$

On obtient l'algorithme itératif suivant fondé sur la relation (*)

1) <u>initialisation</u>

soit $h = h_o$

2) <u>Boucle</u>

FAIRE $h_{i+1} = \left(\frac{p\ \alpha(k)}{n.\beta(hi)} \right)^{1/p+4}$

JUSQU'A

3) <u>Test d'arrêt</u> $\frac{|h_{i+1} - h_i|}{h_i} < \epsilon$

κ

<u>Mise en oeuvre de l'algorithme pour K noyau gaussien réduit</u>

notons $\Delta(x|xi) = K\left(\frac{x-xi}{\sigma}\right)$

Nous avons remplacé h par σ , car c'est la notation usuelle pour un noyau gaussien isotrope.

<u>Proposition 3</u>

$$\alpha(K) = \left(\frac{1}{2\sqrt{\pi}} \right)^p$$

en posant $A_{ij} = \left|\left| \frac{x_i - x_j}{2} \right|\right|^2$ et $a_{ij}(k) = \left(\frac{x_i(k) - x_j(k)}{2\sigma}\right)^2$

alors $\beta(\sigma) = \frac{1}{n^2 (2\sqrt{\pi})^p \sigma^{4+p}} \left(\frac{3pn}{4} + 2 \sum_{i<j} e^{-A_{ij}} \left(\sum_{k=1}^{p} \left(\frac{3}{4} - 3\, a_{ij}(K) + a^2{}_{ij}(k) \right) \right.\right.$

$$\left.\left. + 2 \sum_{1 \leqslant k<k' \leqslant p} a_{ij}(k).\, a_{ij}(k') \right) \right)$$

<u>Démonstration</u>

a) $K(y) = \frac{1}{(2\pi)^{p/2}} \exp\left\{ -\frac{1}{2}\, y^t\, y \right\}$

o

$$\int K^2(y)\,dy = \frac{1}{(2\pi)^p}\exp\left\{-y^t y\right\} dy = \frac{1}{2^{p/2}(2\pi)^p}\exp\left\{-\frac{1}{2}x^t x\right\} dx \quad (\text{ou } x=\sqrt{2}y)$$

$$= \frac{(2\pi)^{p/2}}{2^{p/2}\,(2\pi)^p} = \frac{1}{(2\sqrt{\pi})^p}$$

$$b)\left(\sum_{k=1}^{p}\frac{\partial^2 f_n}{x(k)}\right)^2 = \frac{1}{n^2}\left(\sum_{i=1}^{n}\left(\sum_{k=1}^{p}\left(\frac{\partial^2\Delta(x|x_i)}{\partial^2 x(k)}\right) + 2\sum_{k<k'}\frac{\partial^2\Delta(x|x_i)}{\partial x(k)^2}\cdot\frac{\partial^2\Delta(x|x_j)}{\partial x(k')^2}\right)\right.$$

$$\left. + 2\sum_{1\leq i<j\leq n}\left(\sum_{k=1}^{p}\frac{\partial^2\Delta(x|xi)}{\partial x(k)^2}\cdot\frac{\partial^2\Delta(x|x_j)}{\partial x(k')^2} + \sum_{k\neq k'}\frac{\partial^2\Delta(x|x_i)}{\partial x(k)^2}\cdot\frac{\partial^2\Delta(x|x_j)}{\partial x(k')^2}\right)\right)$$

Pour connaître $\beta(\sigma)$ il suffit donc de pouvoir calculer l'expression suivante dans tous les cas de figure

$$\int\cdots\int\frac{\partial^2\Delta(x)|a)}{\partial x(k)^2}\cdot\frac{\partial^2\Delta(x|b)}{\partial x(k')^2}\,d_x(1)\ldots d_x(p)$$

$$\text{or}\quad \frac{\partial^2\Delta(x|a)}{\partial x(k)} = \frac{1}{\sigma^{p+2}(2\pi)^{p/2}}\left(\left(\frac{x(k)-a(k)^2}{\sigma}\right)-1\right)\exp\left\{-\frac{1}{2\sigma^2}(x-a)^t(x-a)\right\}$$

$\alpha)$ $k\neq k'$

$$\int\cdots\int\frac{\partial^2\Delta(x|a)}{\partial x(k)^2}\cdot\frac{\partial^2\Delta(x|b)}{\partial x(k')^2}\,d_x(1)\ldots d_x(p)$$

$$= \frac{1}{(2\pi)^{p/2}\cdot\sigma^{4+p}}\left(\prod_{\substack{r=1\\ r\neq k\\ r\neq k'}}^{p}\int\frac{1}{\sigma\sqrt{2\pi}}\exp\left\{-\frac{1}{2\sigma^2}(x(r)-a(r))^2+(x(r)-b(r))^2\right\}d_x(r)\right)$$

$$\times\int\frac{1}{\sigma\sqrt{2\pi}}\left(\frac{(x(k)-a(k))^2}{\sigma^2}-1\right)\exp\left\{-\frac{1}{2\sigma^2}\left[(x(k)-a(k))^2+(x(k)-b(k))^2\right]\right\}d_x(k)$$

$$\times\int\frac{1}{\sigma\sqrt{2\pi}}\left(\frac{(x(k')-a(k'))^2}{\sigma^2}-1\right)\exp\left\{-\frac{1}{2\sigma^2}\left[(x(k')-a(k'))^2+(x(k')-b(k'))^2\right]\right\}d_x(k')$$

$$\text{or}\quad (x(r)-a(r))^2+(x(r)-b(r))^2 = 2\left(\left(x(r)-\frac{a(r)+b(r)}{2}\right)^2+\frac{(a(r)-b(r))^2}{4}\right)$$

donc

$$\int \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma^2}\left[(x(r)-a(r))^2 + (x(r)-b(r))^2\right]\right\} d_x(r) = \frac{1}{2} \exp\left\{-\left(\frac{a(r)-b(r)}{2\sigma}\right)^2\right\}$$

et $\int \frac{1}{\sigma\sqrt{2\pi}} \left(\left(\frac{x(k)-a(k)}{\sigma}\right)^2 - 1\right) \exp\left\{-\frac{1}{2\sigma^2}\left[(x(k)-a(k))^2 + (x(k)-b(k))^2\right]\right\} d_x(k)$

$$= \frac{1}{\sqrt{2}} \exp\left\{-\left(\frac{a(k)-b(k)}{2\sigma}\right)^2\right\} \left(\frac{a(k)-b(k)}{2\sigma}\right)^2$$

En résumé si $k' \neq k$

$$(2)\int\cdots\int \frac{\partial^2 (x|a)}{\partial x(k)^2} \frac{\partial^2 (x|b)}{\partial x(k')^2} d_x(1)\ldots d_x(p) =$$

$$\frac{1}{(2\sqrt{\pi})^p \sigma^{4+p}} \exp\left\{-\frac{(a-b)^t(a-b)}{2\sigma^2}\right\} \left(\frac{a(k)-b(k)}{2\sigma}\right)^2 \left(\frac{a(k')-b(k')}{2\sigma}\right)^2$$

$\beta)$ $k = k'$

$$\int\cdots\int \frac{\partial^2 (x|a)}{\partial x(k)^2} \frac{\partial^2 (x|b)}{\partial x(k)} d_x(1)\ldots d_x(p)$$

$$= \frac{1}{(2\pi)^{p/2} \sigma^{4+p}} \prod_{\substack{r=1 \\ r\neq k}}^{p} \frac{1}{\sqrt{2}} \exp\left\{-\left(\frac{a(r)-b(r)}{2\sigma}\right)^2\right\} \cdot \int \frac{1}{\sigma\sqrt{2\pi}} \left(\left(\frac{x(k)-a(k)}{\sigma}\right)^2 - 1\right)$$

$$\times \left(\left(\frac{x(k)-b(k)}{\sigma}\right)^2 - 1\right) \exp\left\{-\frac{1}{2\sigma^2}(x(k)-a(k))^2 + (x(k)-b(k))^2\right\} d_x(k)$$

notons $m(k) = \frac{a(k) + b(k)}{2}$

alors

$$\left(\left(\frac{x(k) - a(k)}{\sigma}\right)^2 - 1\right)\left(\left(\frac{x(k) - b(k)}{\sigma}\right)^2 - 1\right) = \left(\frac{x-m(k)}{\sigma}\right)^4 - 2\left(1 + \frac{1}{4}\left(\frac{a(k)-b(k)}{\sigma}\right)^2\right)$$

$$\times \left(\frac{x-m(k)}{\sigma}\right)^2 + \left(1 - \frac{1}{2}\left(\frac{a(k)-b(k)}{\sigma}\right)^2 + \frac{1}{16}\left(\frac{a(k) - b(k)}{\sigma}\right)^4\right)$$

donc

$$\frac{1}{\sqrt{2}} \left(\left(\frac{x(k)-a(k)}{\sigma}\right)^2 - 1 \right) \left(\left(\frac{x(k)-b(k)}{\sigma}\right)^2 - 1 \right) \exp\left\{ \frac{-1}{2\sigma^2} (x(k)-a(k))^2 + (x(k)-b(k))^2 \right\} d_x(k)$$

$$= \frac{1}{\sqrt{2}} \exp\left\{ - \left(\frac{a(k)-b(k)}{2\sigma}\right)^2 \right\} \left(\frac{3}{4} - \left(1 + \frac{1}{4}\right) \left(\frac{a(k)-b(k)}{\sigma}\right)^2 + \left(1 - \frac{1}{2}\right) \left(\frac{a(k)-b(k)}{\sigma}\right)^2 + \frac{1}{16} \left(\frac{a(k)-b(k)}{\sigma}\right)^4 \right)$$

En résumé

$$\int \cdots \int \frac{\partial^2 \Delta(x|a)}{\partial x(k)^2} \frac{\partial^2 \Delta(x|b)}{\partial x(k)^2} d_x(1) \, d_x(p)$$

$$= \frac{1}{(2\sqrt{\pi})^p \sigma^{4+p}} \exp\left\{ - \frac{(a-b)^t (a-b)}{4\sigma^2} \right\} \left(\frac{3}{4} - 3 \left(\frac{(a(k) - b(k)^2}{2\sigma}\right) + \left(\frac{a(k)(b(k)}{2\sigma}\right)^4 \right) \qquad (3)$$

En remplaçant dans la formule (1) les expressions obtenues dans les formules (2) et (3), nous obtenons le résultat annoncé dans la proposition 3.

γ) Initialisation de l'algorithme

Comme initialisation de cet algorithme, nous utilisons la fenêtre asymptotiquement optimale au sens du MISE dans le cas où la densité à estimer f est celle de la loi normale centrée réduite φ.

$$\left(\sum_{k=1}^{p} \frac{\partial^2 \varphi}{x(k)^2} \right)^2 d_x(1) \ldots d_x(k) = \frac{3}{4 (2\sqrt{\pi})^p}$$

donc $h_o = \left(\frac{4p}{3n}\right)^{1/4+p}$

δ) Remarques

*Si on veut utiliser un autre noyau K' que le noyau normal K, on peut toujours appliquer l'algorithme précédent puis calculer ainsi que le suggère la relation (*)

$$h_{opt} = \left(\frac{\alpha(K')}{\alpha(K)}\right)^{1/p+4} \sigma_{opt}$$

ou h_{opt} désigne la fenêtre optimale pour le noyau K'

et σ_{opt} désigne la fenêtre optimale pour le noyau normal K

*Tapia et Thompson [13] proposent le principe de cet algorithme en dimension 1, sans donner de détails de mise en oeuvre ; Deheuvels et Hominal [3] proposent une version modifiée de cet algorithme en dimension 1, en utilisant des noyaux polynomiaux.

* Nous avons obtenu une relation fonctionnelle $\sigma_{i+1} = (\sigma_i)$ une amélioration de cet algorithme pourrait être d'appliquer une méthode de NEWTON à la recherche du point fixe.

BIBLIOGRAPHIE DU CHAPITRE I

[1] L. BREIMAN, W. MEISEL, E. PURCELL - Variable kernel estimates of multivariate densities technometrics (1977), V. 19, n° 2, 135-144.

[2] P. DEHEUVELS - Estimation non paramétrique de la densité par histogrammes généralisés. Pub. inst. stat. univ. de Paris (1977), XXII, 1-23.

[3] P. DEHEUVELS et P. HOMINAL - Estimation automatique de la densité. Revue de statistique appliquée (1980), XXVIII, n° 1.

[4] P. DEVROYE et T.J. WAGNER - The uniform consistency of kernel density estimates. P.R. Krishnaiah ed. multivariate analysis V, North Holland (1980), 59-77.

[5] V.A. EPANECHNIKOV - Non-parametric estimation of a multivariate density theory of probability and its applications (1969), XIV, 153-158.

[6] J. KIEFFER and J. WOLFOWITZ - On the deviationsof the empiric distribution function of vector chance variables. Annal of mathematical statistic (1958), XXIX, 173-186.

[7] E.A. NADARAYA - On the integral mean square error of some non parametric estimates for the density function.Theory prob. appl. (1974), XIX, n° 1, 133-141.

[8] E. PARZEN - On estimation of a probability density function and mode. Annals math. stat. (1962), XXXIII, 1065-1076.

[9] M. ROSENBLATT - Remarks on some non parametric estimates of a density function. Annals of math. stat. (1956), XXVII, 832-835.

[10] B.W. SILVERMAN - Choosing the window width when estimating a density. Biometrika (1978), 65, n° 1, 1-11.

[11] B.W. SILVERMAN - Weak and strong uniform consistency of the kernel estimate of a density and its derivatives. Annals of statistics (1978), VI, n° 1, 177-184.

[12] E. SCHUSTER - Note on the uniform convergence of density estimates. Annals math. stat. (1970), XLI, 1347-1348.

[13] R.A. TAPIA and J.R. THOMPSON - Non parametric probability density estimation. Johns Hopkins (1978).

[14] T.J. WAGNER - Non parametric estimates of probability densities. IEEE trans. info. theory (1975) IT-21, 438-440.

[15] W. WERTZ - Statistical density estimation : a survey. Vandenhock and Ruprecht (1978).

[16] W. WERTZ - Statistical density estimation : a bibliography. International statistical revue 47 (1979), 155-175.

Chapitre II : APPLICATIONS

A - Introduction : ensemble $D^p(\Omega, \mathcal{A}, P)$ et indicatrices floues)

Les notations définies dans ce paragraphe ont été proposées dans [17]

1) Ensemble $D^p(\Omega, \mathcal{A}, P)$

Définition :

Nous désignerons par $D^p(\Omega, \mathcal{A}, P)$ l'ensemble des variables aléatoires X définies sur $(\Omega, \mathcal{A}, P)$ à valeurs dans R^p et possédant une densité f_x satisfaisant aux 3 conditions ci-dessous :

1) f_x est de la forme suivante où $n \in N$ et $x_1, \ldots, x_n$ une suite de n points de R^p

$$f_x(x) = \sum_{i=1}^{n} p_i \ \Delta(x|x_i, C_i), \ \forall x \in R^p$$

2) Chaque fonction $\Delta(x|x_i, C_i)$ est une densité de probabilité de moyenne x_i et de matrice de covariance C_i

i e $\int x \ \Delta(x|x_i, C_i)\, dx = x_i$

$$\int (x-x_i)(x-x_i)^t \Delta(x|x_i, C_i) d_x = C_i$$

telle que $0 < \Delta(x|x_i, C_i) < \Delta(x_i|x_i, C_i) < \infty$

3) $p_i \geqslant 0$ et $\sum_{i=1}^{n} p_i = 1$

Dans la suite, nous étudierons les variables aléatoires de cet ensemble.

En effet, tous les estimateurs de la densité présentés au Chapitre I : estimateurs de Rosenblatt, Parzen, Deheuvels, Breiman et sa modification proposée ainsi que l'histogramme classique peuvent être mis sous la forme d'une densité d'une variable aléatoire de $D^p(\Omega, \mathcal{A}, P)$.

D'autre part, les théorèmes de convergence existants montrent qu'une densité régulière (les conditions de régularité dépendent du mode de convergence considéré : ex. continuité uniforme pour la convergence uniforme) peut être

approchée d'aussi près que l'on veut par une densité d'une variable aléatoire de $D^p(\Omega, \mathcal{A}, P)$.

Exemples de fonctions

a) Noyau constant

$$\Delta(x|x_i) = \frac{\Gamma(\frac{p}{2}+1)}{\sigma(x_i)^p \ \Pi^{p/2}} \qquad \text{si } \| x-x_i \| < \sigma(x_i)$$

$$= 0 \qquad \text{si non}$$

b) Noyau normal

$$\Delta(x|x_i, C_i) = \frac{(\Pi)^{-n/2}}{(\det C_i)^{1/2}} \exp\left\{-\frac{1}{2}(x-x_i)^t C_i^{-1}(x-x_i)\right\}$$

où C_i est une matrice carrée symétrique, définie, positive.

Souvent $C_i = \sigma_i^2 . I_p$ et I_p = matrice identité d'ordre p.

c) Noyaux polynomiaux

2) Propriétés de $D^p(\Omega, \mathcal{A}, P)$

Proposition

1) $D^p(\Omega, \mathcal{A}, P) \subset L^2(\Omega, \mathcal{A}, P)$

2) soit $X \in D^p$ telle que

$$f_x = \sum_{i=1}^{n} p_i \ \Delta(x|x_i, C_i)$$

ou x $E(X) = \sum_{i=1}^{n} p_i x_i$

$$E(XX^t) = \sum_{i=1}^{n} p_i \left\{C_i + x_i x_i^t\right\}$$

d'où $Var(X) = \sum_{i=1}^{n} p_i C_i + \left[\sum_{i=1}^{n} p_i x_i x_i^t - \left(\sum_{i=1}^{n} p_i x_i\right)\left(\sum_{i=1}^{n} p_i x_i^t\right)\right]$

Remarque

Ainsi, si la densité d'une variable aléatoire est la somme de n noyaux centrés en x_i, son espérance est la moyenne des x_i et sa matrice de covariance est égale à la moyenne des covariances des noyaux plus la matrice de covariance empirique des x_i.

Démonstration

$$f_X(x) = \sum_{i=1}^{n} p_i \ \Delta(x|x_i, C_i)$$

En appliquant la formule du transfert, on peut écrire

$$\int_\Omega \| X(\omega) \|^2 P(d\omega) = \int_{R^p} \| x \|^2 P_X(dx)$$

$$= \sum_{i=1}^{n} p_i \int_{R^p} \| x \|^2 \Delta(x|x_i, C_i)\, dx$$

or Δ admet des moments du second ordre donc la dernière expression a un sens

$$D^p(\Omega, \mathcal{A}, P) \subset L^2(\Omega, \mathcal{A}, P)$$

$$E(X) = \int x f_X(x)\, dx = \sum_{i=1}^{n} p_i \int_{R^p} x \Delta(x|x_i, C_i)\, dx = \sum_{i=1}^{n} p_i x_i$$

$$E(XX^t) = \int_{R^p} xx^t f_X(x) d_x = \sum_{i=1}^{n} p_i \int_{R^p} xx^t \Delta(x|x_i, C_i) d_x$$

$$= \sum_{i=1}^{n} p_i \left(\int_{R^p} (x-x_i)(x-x_i)^t \Delta(x|x_i, C_i)\, d_x + x_i x_i^t \right)$$

$$= \sum_{i=1}^{n} p_i \left\{ C_i + x_i x_i^t \right\}$$

3) Indicatrices floues

a) Rappel à propos des ensembles flous

Définition

Soit un ensemble E ; la donnée d'un sous-ensemble flou est la donnée d'une fonction $\varphi_A : E \longrightarrow [0,1]$

- $\varphi_A(x)$ est l'indice d'appartenance de x au sous-ensemble flou A

- soient A, B deux sous-ensembles flous de E on dit que

* $A \subset B \quad \longleftrightarrow \quad \varphi_A(x) < \varphi_B(x) \quad \forall x \in E$

* $A \cap B$ est défini par $\varphi_{A \cap B}(x) = \inf(\varphi_A(x), \varphi_B(x))$

$A \cup B$ $\varphi_{A \cup B}(x) = \sup(\varphi_A(x), \varphi_B(x))$

Remarque

$\varphi_A(x)$ est aussi appelé indicatrice floue de l'ensemble flou A

Pour des développements plus récents, voir [27].

b) Indicatrice floue d'un point

Définition

On appelle indicatrice floue de s toute fonction $\varphi_s(.)$ telle que

$0 < \varphi_s(x) < \varphi_s(s) < 1$

En général, on choisit $\varphi_s(.)$ de telle façon que la valeur $\varphi_s(x)$ puisse être interprétée comme un "indice de proximité" entre le point courant x et le point s.

c) Fonction aléatoire associée à une famille d'indicatrices floues

Soit une famille F de points dans R^p

Soit $\varphi = (\varphi_s(.) \quad s \in F)$ une famille d'indicatrices floues

Soit X une variable aléatoire à valeurs dans R^p

on appelle fonction aléatoire associée à la famille φ le processus φ_s^X tel que

$\varphi_s^X(\omega) = \varphi_s(X(\omega))$

$\varphi_s(X(\omega))$ représente l'indice de proximité entre un point s fixé dans F et un point aléatoire issu de X.

L'étude de φ_s^X est l'étude locale de la variable aléatoire X au voisinage du point s fixé dans F.

On peut dire que φ_s^X est un processus de comportement local de X.

d) Indicatrices floues disjonctives et canoniques

Soit $X \in D^p(\Omega, \mathcal{A}, P)$, la donnée d'une densité sous la forme $\sum_{i=1}^{n} p_i \Delta(x|x_i, C_i)$ induit naturellement deux familles d'indicatrices floues des points x_i.

Indicatrices disjonctives floues

Nous appellerons indicatrices disjonctives floues associées à X les n fonctions $I_{x_i}(x) = p_i \dfrac{\Delta(x|xi, Ci)}{f_X(x)} \quad \forall x \in R^p$

La fonction aléatoire associée sera notée $I^X_{x_i}$ et nous conviendrons de poser

$I^X(\omega) = [I^X_{x_1}(\omega), \ldots, I^X_{x_n}(\omega)]$

Le qualificatif de disjonctif sera justifié ultérieurement.

Indicatrices canoniques floues

Nous appellerons indicatrices canoniques floues associées à X les n fonctions $J_{x_i}(x)$ définies comme suit :

$* \; x \in R^p, \; i=1, \ldots, n \quad J_{x_i}(x) = \dfrac{\Delta(x|xi, Ci)}{\Delta(xi|xi, Ci)}$

La fonction aléatoire associée sera notée $J^X_{x_i}$

Nous conviendrons de poser

$$J^X(\omega) = [J^X_{x_i}(\omega), \ldots, J^X_{x_1}, \ldots, J^X_{x_n}(\omega)]$$

Remarque 1

La connaissance des fonctions $J_{x_i}(.)$ permet de reconstituer la densité de X d'où le qualificatif de canonique.

Remarque 2

$$J_{x_i}(x) = 1 \iff x = x_i$$

Nous pouvons aussi définir un processus complété de J^X_x que l'on note $\overline{J}^X_x$

soit s appartenant à l'enveloppe convexe de F, il existe des réels $q_1, \ldots, q_n$

tels que $s = \sum_{i=1}^{n} q_i x_i$

avec $q_i > 0$ et $\sum_{i=1}^{n} q_i = 1$

posons $C_s = \sum_{i=1}^{n} q_i C_i$

Alors $\bar{J}_s^X \left| \begin{array}{l} = J_s^X \quad \text{si } s = x_i \\ = \dfrac{\Delta(x|s, Cs)}{\Delta(x|s, Cs)} \quad \text{si } s \ F \end{array} \right.$

si $C_i = C \quad \forall i = 1, \ldots, n$

on peut définir $\bar{J}_s^X$ pour tout s appartenant à R^p

$$\bar{J}_s^X = \frac{\Delta(X|s, C)}{\Delta(s|s, C)}$$

4) Plan du chapitre II

Nous allons dans un premier temps étudier comment réduire le nombre de noyaux d'une densité de D^p et établir un lien canonique avec la classification floue. Nous obtenons plusieurs critères d'optimalité pour une classification.

Nous étudions ensuite différentes applications d'une densité à noyaux à la régression.

Nous examinons ensuite plus particulièrement les deux familles d'incatrices floues introduites précédemment et la fonction aléatoire qui leur est associée.

Finalement, nous proposons un nouveau coefficient que nous utiliserons pour étudier les correspondances entre variables et données.

B - Réduction du nombre des noyaux

Cette démarche a un double fondement :

- l'estimation de la densité par des noyaux est une méthode non paramétrique, son efficacité surtout dans le cas multivarié est assez faible, il faut donc un grand nombre de points de donnée pour obtenir une bonne approximation de la densité, le calcul de l'estimation de la densité revient donc cher au point de vue temps machine d'où l'intérêt de réduire le nombre de noyaux tout en conservant le maximum d'information.

- D'un autre point de vue, l'estimation obtenue par les noyaux est dépendante de l'échantillonnage, mais la distribution de la variable aléatoire n'en dépend pas. Il est donc naturel d'essayer d'ajuster la densité par un nombre fixé de noyaux liés intrinsèquement à la densité de la variable aléatoire observée. Enfin, il est préférable que l'étude locale par les indicatrices floues ne dépende pas de l'échantillon observé.

Nous chercherons donc une fonction
$g(x) = \sum_{k=1-}^{K} q_k \ \Delta(x|m_k, C_k)$ qui approche au plus près de l'estimation de la densité par les noyaux de Parzen
$f(x) = \sum_{i=1}^{n} \frac{1}{n} \ \Delta(x|xi, C)$.

1) Lien canonique entre une fonction $g(x) = \sum_{k=1}^{K} q_k \Delta(x|m_k, C_k)$ et une K-partition floue

a) Définition d'une partition floue

Une K-partition foue de $S = \{x_1, \ldots x_n\}$ est la donnée de K sous ensembles flous de S : $C_1, \ldots, C_k$
où $C_i = (x_1, p_i(x_1)), \ldots, (x_n, p_i,(x_n))$
pi est la fonction de S ⟶ [0,1] qui à chaque point x_i associe son degré d'appartenance à la Kième classe de la partition.

Pour rester cohérent avec la définition d'une partition "dure" (les indices d'appartenance étant des fonctions caractéristiques), nous exigeons d'autre part que :

$$j = 1, \ldots, n \quad \sum_{k=1}^{K} P_k(x_i) = 1$$

autrement dit : les fonctions $P_k(.)$ forment une partition de l'unité.

Ceci nous amène à préciser les notations suivantes :

b) Notations

* δ_x masse de Dirac en x

* μ distribution de masse sur R^p chargeant les points de S

$$\mu = \sum_{x_j \in S} \frac{1}{n} \delta_{x_j}$$

soit m_μ et C_μ les 2 premiers moments

* μ_k distribution de masse sur R^p chargeant les points de la k-ième classe de la partition floue

$$k = \sum_{x_j \in S} p_k(x_j) \delta_{x_j}$$

m_{μ_k} et $C\mu_k$ les 2 premiers moments

* f(x) est la densité de la distribution continue sur R^p engendrée par la méthode des noyaux de Parzen et les points de S

$$f(x) = \sum_{x_j \in S} \frac{1}{n} \Delta(x|x_j, C)$$

* $f_k(x)$ est la densité de la distribution continue sur R^p engendrée par la méthode des noyaux de Parzen par les points de la k-ième classe de la partition floue

$$f_k(x) = \frac{1}{\mu_k(R^p)} \quad p_k(x_j) \Delta(x|x_j, C)$$

c) Position du problème

Avec ces nouvelles notations, nous avons la propriété suivante :

Propriété

$$f(x) = \sum_{k=1}^{K} \frac{\mu_k (R^n)}{n} f_k(x)$$

or

$$g(x) = \sum_{k=1}^{K} q_k \Delta(x|m_k, C_k)$$

Sous cette forme si nous voulons rendre g(x) "proche" de f(x) nous sommes amenés à identifier les 2 fonctions terme par terme. En identifiant les moments des premiers et seconds ordres, nous avons les relations suivantes :

$$q_k = \frac{\mu_k (R^p)}{n} \qquad (0)$$

$$\int x \Delta(x|m_k, C_k)\, dx = \int x f_k(x)\, dx \qquad (1)$$

$$\int xx^t \Delta(x|m_k, C_k)\, dx = \int xx^t f_k(x)\, dx \qquad (2)$$

Proposition 2-1

1) Sous les conditions (0), (1), (2) les deux premiers moments de f et de g sont égaux

2) (1), (2) $\Longleftrightarrow$ $m_k = m_{\mu_k}$

$C_k = C + C_{\mu_k}$

Démontration

1) $$\int x f(x)\, dx = \frac{1}{n} \int x \left(\sum_{k=1}^{K} \mu_k(R^p) f_k(x) \right) dx = \sum_{k=1}^{K} \frac{\mu_k (R^p)}{n} \int x f_k(x)\, dx$$

$$\overset{(1)}{=} \sum_{k=1}^{n} q_k \int x \Delta(x|m_k, C_k)\, dx = \int x \left(\sum_{k=1}^{K} q_k \Delta(x|m_k, C_k) \right) dx$$

$$= \int x\, g(x)\, dx$$

de même pour le deuxième moment.

2) (1) $\Rightarrow m_k = \frac{1}{\mu_k(RP)} \sum_{x_j \in S} p_k(x_j) \int x \Delta(x|x_j, C)\, dx$

$= \frac{1}{\mu_k(RP)} \sum_{x_j \in S} p_k(x_j).\ x_j = m\mu_k$

(2) $\Rightarrow C_k + m_k m_k{}^t = \frac{1}{\mu_k(RP)} \sum_{x_j \in S} p_k(x_j) \int xx^t \Delta(x|x_j, C)\, dx$

$= \frac{1}{\mu_k(RP)} \sum_{x_j \in S} p_k(x_j)\ (x_j x_j{}^t + C)$

$= C + \frac{1}{\mu_k(RP)} \sum_{x_j \in S} p_k(x_j)\ x_j x_j{}^t$

donc $C_k = C + C\mu_k$

Nous avons ainsi à partir d'une K-partition floue donnée trouvé une fonction $g(x) = \sum_{k=1}^{K} q_k \Delta(x|m_k, C_k)$ de façon à ce que les deux premiers moments soient égaux globalement et localement.

Le problème consiste maintenant à obtenir une "meilleure" partition floue suivant des critères à définir.

2 - Différents critères pour une "meilleure" partition floue

a) Remarque préliminaire

Dans la forme générale d'une densité de D^p, la forme du noyau n'est pas précisée, cependant nous avons signalé que la qualité de l'estimation n'est pas très dépendante de cette forme. Pour la suite des calculs, nous n'hésiterons donc pas à particulariser la forme des noyaux quand les calculs le demandent en choisissant notamment des noyaux gaussiens.

b) Critère I_1 des moments d'ordre 3

Pour obtenir une bijection entre l'ensemble des partitions floues et l'ensemble des fonctions g(x), nous avons égalé les deux premiers moments des deux distributions définies par f(x) donnée par les noyaux de Parzen et par g(x).

Un critère naturel pour une meilleure partition floue est donc la norme de la différence des moments d'ordre 3

$$I_1 = \| M_3 \|$$

où $M_3 = \int xx^t x \, (f(x) - g(x)) \, dx$

Proposition 2-2

Si les fonctions sont des fonctions paires (c'est le cas pour tous les noyaux usuels et pour chaque noyau raisonnable), alors

$$M_3 = \sum_{i=1}^{n} \frac{1}{n} x_i x_i^t x_i - \sum_{k=1}^{K} q_k \left[2 C_{\mu_k} . m_k + tr (C_{\mu_k} + m_k m_k^t) . m_k \right]$$

Remarque 1

$C_k . m_k$ est le produit d'une matrice et d'un vecteur $tr (C_k + m_k . m_k^t)$. m_k est le produit d'un scalaire et d'un vecteur.

Remarque 2

$$tr (C_k + m_k m_k^t) = \frac{1}{q_k} \sum_{i=1}^{n} p_k(x_i) x_i^t x_i$$

Démonstration de la proposition 2-2

Nous citons d'abord le lemme suivant :

Lemme : soit X un vecteur aléatoire ayant une distribution symétrique (i.e $P(X \in A) = P(-X \in A)$ $\forall A$ de moyenne m

alors $E(XX^tX) = 2 (E(XX^t) - mm^t) m + mE(X^tX)$

$= 2 (E(X-m)(X-m)^t) m + mE(X^tX)$

donc $\int xx^t x\, \Delta(x|x_i, C)\, dx = 2\, C.x_i + x_i\, tr\,(C + x_i x_i^t)$

$\int xx^t x\, \Delta(x|m_k, C_k)dx = 2\, C_k m_k + m_k\, tr\,(C_k + m_k\, m_k^t)$

Posons

$$\Delta_1 = (C.m - \sum_{k=1}^{K} q_k\, C_k\, m_k)$$

$$\Delta_2 = \sum_{i=1}^{n} \frac{1}{n}\, x_i\, tr\,(C + x_i x_i^t) - \sum_{k=1}^{K} q_k\, m_k\, tr\,(C_k + m_k\, m_k^t)$$

Alors $M_3 = 2\Delta_1 + \Delta_2$

Calcul de Δ_1 : par la proposition 2-1 $C_k = C + C\mu_k$

et $\sum_{k=1}^{K} q_k\, m_k = \sum_{k=1}^{K} \sum_{j=1}^{n} p_k(x_j)\, x_j = \sum_j (\sum_k p_k(x_j)) x_j = m_\mu$

donc $\Delta_1 = Cm_\mu - \sum_{k=1}^{K} q_k\, (C + C\mu_k)\, m_k$

$$= - \sum_{k=1}^{K} q_k\, C\mu_k\, m_k$$

Calcul de Δ_2 :

$$\Delta_2 = \sum_{i=1}^{n} \frac{1}{n}\, x_i\, tr\,(C + x_i x_i^t) - \sum_{k=1}^{K} q_k\, m_k\, tr\,(C + C\mu_k + m_k\, m_k^t)$$

$$= m_\mu\ tr\ C - \sum_{k=1}^{K} q_k\, m_k\, (trC) + \frac{1}{n} \sum_{i=1}^{n} x_i x_i^t\, x_i - \sum_{k=1}^{K} q_k\, m_k\, tr\,(C\mu_k + m_k\, m_k^t)$$

$$= \sum_{i=1}^{n} \frac{1}{n}\, x_i x_i^t\, x_i - \sum_{k=1}^{K} q_k\, m_k\, tr\,(C\mu_k + m_k\, m_k^t)$$

d'où le résultat

Démontration du lemme

$E\,(X-m)(X-m)^t(X-m) = 0$ car la distribution est symétrique

or $(X-m)(X-m)^t(X-m) = XX^tX - XX^tm - Xm^tX - mX^tX + mm^tX + mX^tm + Xm^tm - mm^tm$

X^tm est un scalaire donc égal à m^tX

et $E(X) = m$

donc $E(X-m)(X-m)^t(X-m) = 0 = E(XX^tX) - 2\,E(XX^t)m - mE(X^tX) + 2\,mm^tm$

d'où le résultat.

Remarques

Le critère I_1 possède les avantages suivants

1) il ne dépend pas de la forme de noyau (si celui-ci est symétrique)

2) il ne dépend pas du paramètre de lissage C

3) Le volume de calcul utilisé est raisonnable

4) I_1 a le désavantage suivant :

$I_1 = 0$ n'implique pas que $f = g$

c) Critère I de l'écart quadratique intégré

I_2 est la norme au carré dans L^2 de $f - g$

$$I_2 = \int_{RP} (f(x) - g(x))^2 \, dx$$

Pour obtenir une expression de ce critère, nous devons préciser la forme des noyaux : nous les choisissons gaussiens.

On a alors

$$\Delta(x|x_1, C_1) = \frac{1}{(2\pi)^{p/2} \sqrt{\det C_1}} \exp \left\{ - \frac{1}{2} (x-x_1)^t \; C_1^{-1} (x-x_1) \right\}$$

Notons $\langle ., . \rangle$ le produit scalaire dans L^2

alors $I_2 = \int f(x)^2 \, dx - 2 \langle f, g \rangle + \int g(x)^2 \, dx$

$$I_2 = \sum_{i=1}^{n} \frac{1}{n^2} \langle \Delta(x_1, C), \Delta(x_i, C) \rangle + 2 \sum_{1 \leq i < j \leq n} \frac{1}{n^2} \langle \Delta(x_i, C), \Delta(x_j, C) \rangle$$

$$+ \sum_{k=1}^{K} q_k^2 \langle \Delta(m_k, C_k), \Delta(m_k, C_k) \rangle + 2 \sum_{1 \leq k < r \leq K} q_k q_r \langle \Delta(m_k, C_k), \Delta(m_r, C_r) \rangle$$

$$- 2 \sum_{i=1}^{n} \sum_{k=1}^{K} \frac{q_k}{n} \langle \Delta(m_k, C_k), \Delta(x_i, C) \rangle$$

Proposition 2-3

$$\langle \Delta(x_1, C_1), \Delta(x_2, C_2) \rangle$$

$$= \frac{1}{(2\pi)^{p/2} \sqrt{\det (C_1 + C_2)}} \exp \left\{ - \frac{1}{2} (x_1 - x_2)^t \; (C_1 + C_2)^{-1} \; (x_1 - x_2) \right\}$$

Démonstration de la proposition 2-3

Lemme d'algèbre linéaire

soit A et A' matrices symétriques telles que A+A' soit inversible

$$(x-m)^t A(x-m) + (x-m')^t A'(x-m')$$
$$= \left[x - (A+A')^{-1}(Am+A'm')\right]^t (A+A') \left[x-(A+A')^{-1}(Am+A'm')\right]$$
$$+ (m-m')^t A(A+A')^{-t} A' (m-m')$$

donc $\langle \Delta(x_1,C_1), \Delta(x_2,C_2) \rangle$

$$= \frac{\sqrt{\det (C_1^{-1} + C_2^{-1})^{-1}}}{(2\pi)^{p/2}\sqrt{\det C_1C_2}} \exp\left\{-\frac{1}{2}(x_1-x_2)^t C_1^{-1} \quad (C_1^{-1} + C_2^{-1})^{-1} C_2^{-1} (x_1-x_2)\right\}$$

or $C_1^{-1}(C_1^{-1} + C_2^{-1}) C_2^{-1} = \left[C_2 (C_1^{-1} + C_2^{-1}) C_1\right]^{-1}$

et $$\frac{\det (C_1^{-1} + C_2^{-1})^{-1}}{\det C_1 C_2} = \frac{1}{\det (C_1 + C_2)}$$

d'où le résultat

Démontration lemme

$$(x-m)^t A (x-m) + (x-m')^t A' (x-m')$$
$$= x^t (A+A')x - x^t (Am+A'm') - (m^tA + m'^tA')x + m^tAm + m'^tA'm'$$
$$= (x - (A+A')^{-1}(Am+A'm'))^t (A+A')(x-(A+A')^{-1}(Am+A'm'))$$
$$- (Am+A'm')^t (A+A')^{-1} (Am+A'm') + m^tAm + m'^tA'm'$$

posons $Z = - (Am+A'm')^t (A+A')^{-1}(Am+A'm') + m^tAm + m'^tA'm'$

$$= m^t(-A(A+A')^{-1}A+A)m - m^t (A(A+A')^{-1}A')m'$$
$$- m'(A'(A+A')^{-1}A) + m'^t (A' - A'(A+A')^{-1}A')m'$$

or $A - A(A+A')^{-1} A = A - (A+A')(A+A')^{-1}A + A'(A+A')^{-1}A$

$$= A' (A+A')^{-1}A$$
$$= A - A(A+A')^{-1}(A+A') + A(A+A')^{-1}A'$$
$$= A (A+A')^{-1}A'$$
$$= A' - A'(A+A')^{-1}A'$$

Donc $Z = (m-m')^t A(A+A')^{-t}A' (m-m')$

finalement en posant $B = A(A+A')^{-1}A'$

$$M = (A+A')^{-1}(Am+A'm')$$

$$(x-m)^t A (x-m) + (x-m')^t A' (x-m')$$

$$= (x-M)^t(A+A')(x-M) + (m-m')^t B (m-m')$$

ce qui était le résultat annoncé.

Commentaires

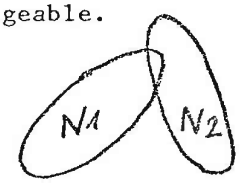
Ce critère a l'avantage d'avoir une signification intuitive et de vérifier la propriété $I_2 = 0 \Rightarrow f = g$ que ne vérifie pas le critère I_1, il a cependant un défaut majeur : il faut inverser $\frac{K(K+1)}{2} + K + 1$ matrice pxp et calculer $\frac{n(n+1)}{2} + \frac{K(K+1)}{2} + K_n$ exponentielles. Un autre désavantage de I_2 par rapport à I_1 est sa dépendance du paramètre C et de la forme gausienne du noyau.

On peut cependant réduire le volume du calcul de I_2

1) $f(x)^2 dx$ ne dépend pas de la partition il suffit de minimiser

$$g(x)^2 dx - 2 \int f(x).g(x) dx$$

2) Avant d'effectuer un produit de noyaux, faire un test préalable s'il n'est pas négligeable.



Dans ce cas de figure $\langle N_1, N_3\rangle << \langle N_1, N_2\rangle$

3) Remplacer I_2 par $I_2' = \sum_{k=1}^{n} q_k \int_{RP} (f_k(x) - (x|m_k, C_k))^2 dx$

4) Si on n'a pas besoin d'une forme fonctionnelle de I_2 on peut utiliser des méthodes numériques pour le calcul de I_2, le noyau n'a plus besoin d'être gaussien.

d) Critère d'ajustement multivarié de Breiman

La statistique employée par Breiman pour optimiser son estimateur (voir Ch. I) est la suivante :

si $w_j = \exp\left\{-n\ f(x_i).V\ (d_{j.1})\right\}$

où f est la densité que l'on ajuste et

$$V\ (d_{j.1}) = \frac{\Gamma(\frac{p}{2} + 1)}{d_{j.1}\ \pi^{p/2}}$$

$d_{j.1}$ distance de x_j à son plus proche voisin

D'après Breiman, les w_j suivent approximativement une loi uniforme.

En considérant l'échantillon ordonné

$w(1) < w(2) < \ldots < w\ (n)$

Le critère d'ajustement est

$$B(f) = \sum_{j=1}^{n} (w(j) - \frac{j}{n})^2$$

Nous disposons ainsi en prenant $I_3 = B(g)$ d'un critère à minimiser, mais aussi d'un seuil (B(fn)) à partir duquel l'ajustement peut être considéré comme satisfaisant.

Notons que ce critère pourrait être utilisé pour déterminer la fenêtre dans une estimation par les noyaux de Parzen.

e) Critères de l'écart quadratique pondéré intégré

$$I_4 = \int (f(x)-g(x))^2\ f(x)\ dx$$

$$I'_4 = \int (f(x)-g(x))^2\ g(x)\ dx$$

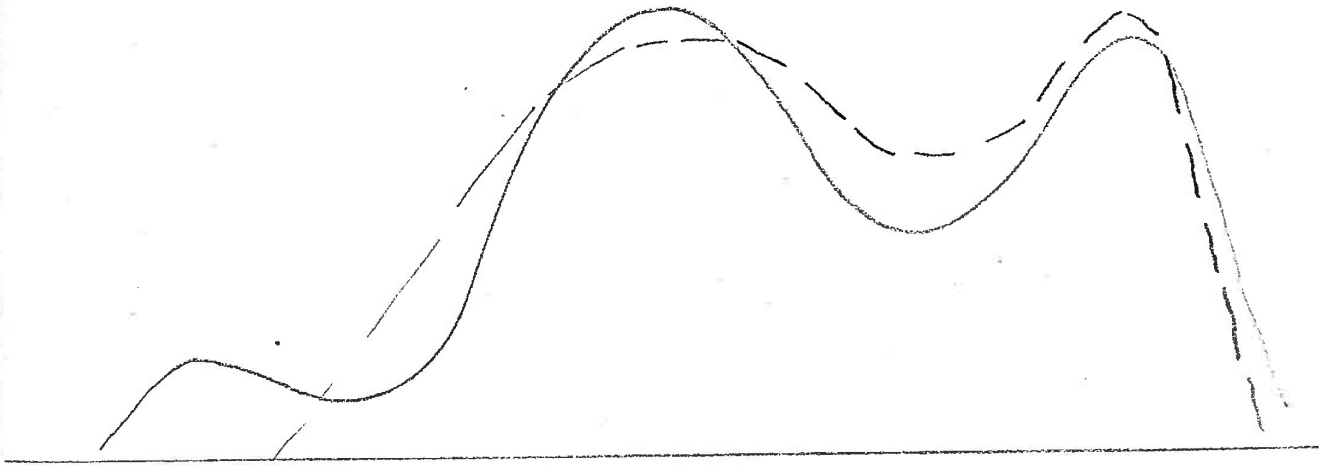
Ces critères ont à peu près les mêmes qualités et les mêmes défauts que I_2.

Leur calcul utilise les mêmes résultats que I_2 et n'est donc faisable que dans le cas gaussien. De plus ces critères exigent un temps de calcul encore nettement plus important que I_2.

Ces critères mettent l'accent sur l'ajustement des modes et donneront donc de bons résultats si les données sont fortement structurées et si l'on connait le nombre optimal de classes.

Cependant, si le nombre de modes est inconnu, ces critères risquent de ne pas être satisfaisants.

Exemple :



Si on pondère l'écart quadratique par la fonction bimodale, l'écart dû au troisième mode est éliminé.

3 - Algorithmes d'optimisations d'un critère I

a) Préliminaires

On peut représenter l'ensemble des partitions floues de l'ensemble des n points x_i en K classes par un simplexe compact de l'espace vectoriel de dimension Kxn des matrices $P = \left[p_k (x_i) \right]$.

Les limites de ce simplexe étant les hyperplans d'équation $p_k(x_i) = 1$ et $P_k(x_i) = 0$. Comme d'autre part les partitions floues vérifient les n équations $\sum_{k=1}^{K} P_k(x_i)=1$, on pourrait plonger l'ensemble des partitions floues dans un espace vectoriel de dimension (K-1).n.

b) Algorithmes spécifiques à la classification floue

L'ensemble des partitions floues considéré comme sous ensemble de $R^{n.k}$ est un compact car il est fermé borné ; tout critère continu admet donc un minimum sur cet ensemble, nous pouvons ainsi appliquer toutes les techniques usuelles d'analyse numérique à la recherche d'un minimum d'une fonction.

On peut par exemple utiliser la technique du gradient projeté en présentant le problème sous la forme d'une optimisation d'une fonction deux fois différentiable sous les contraintes mixtes suivantes :

$$0 < P_k(x_i) < 1 \qquad \text{pour tout k,i vérifiant} \quad 1 < k < K, \quad 1 < i < n$$

$$\text{et} \sum_{k=1}^{K} P_k(x_i) = 1 \qquad \text{pour tout i}$$

Cette méthode a été proposée par Ruspini [23] pour optimiser entre autres le critère suivant

$$J = \sum_{i=1}^{n} \sum_{j=i}^{n} \left\{ \left[\sum_{k=i}^{K} \sigma (P_k(x_i) - P_k(x_j))^2 \right] - d^2_{ij} \right\}$$

où σ est une constante réelle et d_{ij} un indice de dissimilarité entre x_i et x_j.

c) Remarques et proposition

Les méthodes développées ci-dessus ne présentent cependant guère plus qu'un

intérêt théorique. Hors des cas triviaux, la dimension de l'espace des partitions floues est trop grand pour pouvoir obtenir une convergence dans un temps de calcul raisonnable. De plus, l'ensemble des partitions floues ayant la puissance du continu, obtenir une partition floue optimale au sens d'un critère paraît plus difficile qu'obtenir une partition optimale "dure" dont il n'existe qu'un nombre fini.

Enfin, nous avons obtenu le résultat suivant :

Proposition 2-4

Si I est un critère convexe sur l'ensemble des partitions floues, la partition optimale au sens du critère I est une partition dure.

Il apparaît donc qu'il est nécessaire de chercher une partition dure optimisant les critères proposés, quitte à les affiner dans le cas flou.

Démonstration de la proposition 2-4

Soit une partition floue P

$$P \begin{bmatrix} P^1(x_1) & \cdots & P_1(x_n) \\ \vdots & & \vdots \\ P_K(x_1) & \cdots & P_K(x_n) \end{bmatrix}$$

Notons $P_{k_1, k_2, \ldots k_n}$ la partition dure suivante

$$P_{k_1,\ldots,k_n} = \delta_{j,i} \qquad \delta_{j,i} = 1 \text{ si } k_j = i$$
$$= 0 \text{ sinon}$$

$$P = \sum_{k_1=1}^{K} P_{k_1}(x_1) \begin{bmatrix} 0 & P_1(x_2) & \cdots & P_1(x_n) \\ 1 & & & \\ 0 & P_K(x_2) & \cdots & P_K(x_n) \end{bmatrix} k_1$$

$$= \sum_{k_1=1}^{K} P_{k_1}(x_1) \sum_{k_2=1}^{K} P_{k_2}(x_2) \begin{bmatrix} 0 & 0 & \cdots & P_1(x_n) \\ & 1 & & \\ 1 & & & \\ 0 & 0 & \cdots & P_K(x_n) \end{bmatrix}$$

finalement

$$P = \sum_{k_1=1}^{K} \cdots \sum_{k_n=1}^{K} P_{k_1}(x_1) \cdots P_{k_n}(x_n) \cdot P_{k_1,\ldots,k_n}$$

or $$\sum_{k_1=1}^{K} \cdots \sum_{k_n=1}^{K} P_{k_1}(x_1) \cdots P_{k_n}(x_n) = 1$$

Une partition floue quelconque peut donc être exprimée comme barycentre de partitions dures.

Comme I est convexe

$$I(P) \geq \sum_{k_1=1}^{K} \cdots \sum_{k_n=1}^{K} P_{k_1}(x_1) \cdots P_{k_n}(x_n) \quad I(P_{k_1,\ldots,k_n})$$

$$\geq \inf I \ (P_{k_1,\ldots,k_n})$$

Donc pour toute partition floue P, il existe une partition dure au moins aussi bonne au sens du critère convexe I

d) Algorithme de Marriott

Dans [20] F.H.C. Marriott propose l'algorithme suivant optimisant un critère sur l'ensemble des K partitions dures.

1 - Initialisation

Diviser l'ensemble des données en K groupes

2 - Boucle

Enlever chaque point à tour de rôle de l'ensemble des données.

L'assigner au groupe tel que I soit optimal.

3 - Test d'arrêt

Continuer jusqu'à ce qu'aucun point ne soit réassigné.

La partition ainsi obtenu est localement optimale dans le sens que le critère ne peut être amélioré en transférant un seul point.

Remarque : il n'est pas nécessaire de calculer le critère à chaque étape, il suffit de connaître la variation du critère si on ajoute un point x_i au groupe g.

Application au critère I_1

La variation du critère I_1 est connue dès que l'on connait la variation de C_{μ_k} et de m_k, si on rajoute le point x au groupe k.

Or, si on note n_k le cardinal de k-ième groupe et C^*_r et m^*_k les valeurs de la covariance et de la moyenne après ajout du point x, nous avons les relations suivantes :

$$m^*_k = \frac{1}{n_k+1} (n_k\ m_k+x)$$

$$C^*_{\mu_k} = C_{\mu_k} + \frac{n_k}{n_k+1} (x-m_k)(x-m_k)^t$$

e) Méthodes hiérarchiques

Nous pouvons adapter les méthodes classiques de classification hiérarchique à l'optimisation d'un critère. Au lieu de fusionner les 2 classes les "proches", nous fusionnons les 2 classes telles que le critère soit optimal.

4 - Approche alternative : estimation paramétrique d'un mélange

a) Dans ce paragraphe, nous proposons une approche alternative au problème de réduction du nombre des noyaux. Jusqu'à présent nous partions de l'estimation non-paramétrique f(x) de la densité par les noyaux de Parzen et cherchions à l'ajuster par une fonction g(x). Les méthodes proposées dans ce paragraphe sont des méthodes paramétriques : soit $g(x)= \sum_{k=1}^{K} q_k\ \Delta(x)|m_k,C_k)$ où $\Delta(x|m_k,C_k)$ est la densité multinormale de moyenne m_k et de matrice de covariance C_k. En supposant que g(x) est la densité dont sont issus les points de donnée, nous cherchons à estimer les paramètres m_k et C_k. Plusieurs méthodes ont été proposées

pour ce problème de reconnaissance d'un mélange. Citons notamment :

* les méthodes graphiques,

* approximation stochastique,

* déconvolution.

Signalons l'article de Cazes [6] qui donne une review de ces méthodes dans le cas unidimensionnel.

La méthode qui a été la plus étudiée est celle du maximum de vrai semblance, les premières références étant les articles de Day [9] et de Wolfe [26]. Dans ce qui suit, nous détaillerons cette méthode.

b) Estimation des paramètres de g(x) par le maximum de vrai semblance

La fonction de vrai semblance s'écrit :

$$L(x_1, \quad , x_n) = \sum_{i=1}^{n} g(x_i)$$

$$\text{ou} \quad g(x) = \sum_{k=1}^{K-1} q_k \, \Delta(x|m_k, C_k) + \left(1 - \sum_{k=1}^{K-1} q_k\right) \Delta(x|m_K, C_K)$$

En annulant le gradient de Log $L(x_1, \quad , x_n)$ par rapport aux paramètres q_k, m_k, C_k, nous obtenons les relations suivantes en posant

$$I_k(x_i) = \frac{q_k \cdot \Delta(x_i|m_k, C_k)}{g(x_i)}$$

$$\begin{cases} q_k = \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} I_k(x_i) \\ m_k = \dfrac{1}{\displaystyle\sum_{x_i} I_k(x_i)} \displaystyle\sum_{i=1}^{n} I_k(x_i) x_i \\ C_k = \dfrac{1}{\displaystyle\sum_{x_i} I_k(x_i)} \displaystyle\sum_{i=1}^{n} I_k(x_i)(x_i - m_k)(x_i - m_k)^t \end{cases}$$

Dans ces relations nous tirons l'algorithme d'estimation séquentielle suivant :

c) Algorithme

1 - initialisation

soit une matrice KXn initiale $\left[I_K^{(\circ)}(x_i) \right]$ telle que

$$\sum_{k=1}^{K} I_k^{(\circ)}(x_i)=1 \qquad i= 1, \quad ,n$$

2 - Boucle

FAIRE

i) $$q_k^{(m)}= \frac{1}{n} \sum_{x_i} I_k^{(m)}(x_i)$$

$$C_k^{(m)} = \frac{1}{\sum_{x_i} I_k^{(m)}(x_i)} \sum_{x_i} I_k^{(m)}(x_i)x_i$$

$$C_k^{(m)} = \frac{1}{\sum_{x_i} I_k^{(m)}(x_i)} \sum_{x_i} I_k^{(m)}(x_i)(x_i-m_k^{(m)})(x_i-m_k^{(m)})^t$$

ii) $$I_k^{(m+1)}(x_i) = \frac{q_k^{(m)} \Delta (x_i | m_k^{(m)}, C_k^{(m)})}{g^{(m)}(x_i)}$$

JUSQU'A

3 - Test d'arrêt

$$\| I^{(m)} - I^{(m+1)} \| < \varepsilon$$

d) Lien entre cette approche et la première

Wolfe [26] qui propose cet algorithme parle des quantités $P_k(x_i)$ comme probabilité d'appartenance de x_i à une k-ième classe, reliant ainsi ce problème à la classification automatique.

Bezdek et Dunn [3] proposent une approche voisine en posant $I_k(x_i)$ comme indice d'appartenance à la K-ième classe floue, établissant un autre bien canonique entre une partition floue et une fonction g(x). Notons qu'asymptotiquement les deux approches sont les mêmes (voir P.38) la fenêtre C tendant vers 0 quand n tend vers l'infini.

e)Remarques

i) la méthode du maximum de vraissemblance a une interprétation bayesienne: si on considère $I_k^{(m)}(x_i)$ comme une probabilité à priori d'appartenance à la k-ième classe, $I_k^{(m+1)}(x_i)$ peut être interprétée comme la probabilité à postériori.

ii) Dans le cadre des nuées dynamiques de Diday [11], Schroeder [24] obtient comme cas particulier un algorithme équivalent à celui qui précède avec la différence qu'à chaque itération, les points sont associés de façon dure à la classe de plus grande probabilité. (voir aussi [25]). Mais Br yant, généralisant les travaux de Marriott montre qu'une telle approche introduit des biais dont ne souffre l'algorithme du maximum de vraissemblance. (voir [5]).

5 - Conclusion

Un problème qui n'a pas été abordé est le choix du nombre K de noyaux (ou de classes). Dans le cas d'une classification optimisant un critère, la démarche suivante semble s'imposer : pour K fixé trouver la meilleure partition au sens du critère, pour faire varier K et prendre le K minimisant le critère. Cette démarche comporte néanmoins deux défauts : d'une part il se peut que le critère soit une fonction décroissante de K et d'autre part cette démarche est très couteuse au point de vue temps machine. Ce rep roche peut d'ailleurs être adressé à toutes les méthodes de classification présentées en particulier à la méthode du maximum de vrai semblance. On aura intérêt à utiliser des méthodes de classification rapides telles les méthodes du type nuées dynamiques (voir aussi [2], [16] et utiliser les critères présentés comme critère de validité.

C - Application à la régression

Dans ce paragraphe, nous allons examiner le problème de la régression dans le cas de variables aléatoires de la classe DP (Ω, A, P). Nous allons d'abord présenter une méthode d'analyse des données proposée notamment par Cazes [7] appelée régression par boules. Nous donnerons ensuite un théorème cité par Mallet [17] qui obtient l'expression de l'espérance conditionnelle pour des variables appartenant à un sous ensemble de DP (Ω,A,P). Nous en déduirons une justification de la régression par boules. Puis nous généraliserons ce théorème dans le cas de noyaux multinormaux que nous interpréterons comme une extension du modèle linéaire. En dernier lieu, nous examinerons la qualité de la regression.

1 - Regression par boules

soit n réalisations d'un vecteur aléatoire

$(Y^1, \ldots, Y^q, X^1, \ldots, X^p)$

notées

$(y^1_j, \ldots, y^q_j, x^1_j, \ldots, x^p_j)$ $\quad j = 1,n$

$x^1, \ldots, x^p$ sont des variables explicatives.

Pour prédire les valeurs des composantes au vecteur y correspondant à un point x de R^p, Cazes propose 2 méthodes

a) Régression avec nombre de points fixés

soit $l \in N$ fixé on recherche les l points de l'échantillon les plus proches de x dans l'espace des variables explicatives R^p. On calcule alors la moyenne arithmétique les vecteurs y_k correspondants. La précision de cette regression peut-être surveillée en calculant la matrice de covariance empirique des y_k.

Remarque

Cazes conseille un choix de l de l'ordre de 1 à 10 % de la taille de l'échantillon.

b) Régression avec rayon r fixé

soit $r > 0$ fixé, on recherche tous les points de l'échantillon qui sont à une distance inférieure à r dans l'espace des variables explicatives.

De même que précédemment, on calcule la moyenne arithmétique des y_k correspondants :

Remarque : aucun choix de r n'est donné

2 - Régression avec densité pseudo-factorisée

a) Théorème 2-1

soient X et Y deux vecteurs aléatoires à valeurs respectivement dans R^p et R^q, telles que le couple (X,Y) appartienne à $D^{p+q}(\Omega, A, P)$ en admettant une densité "pseudo-factorisée" du type suivant :

$$f_{XY}(x, y) = \sum_{k=1}^{n} p_k \ \Delta_X(x|x_k).\Delta_Y(y|y_k, C^Y{}_k)$$

Dans ces conditions

X appartient à $D^p(\Omega, A, P)$ et admet la densité suivante

$$f_X(x) = \sum_{k=1}^{n} p_k \ \Delta_X(x \mid x_k)$$

De plus si on pose $I_{x_k}(x) = p_k \ \dfrac{\Delta_X(x \mid x_k)}{f_X(x)}$

on a

$$E(Y \mid X) = \sum_{k=1}^{n} y_k \ I^X{}_{x_k}$$

ou I^X_s désigne le processus aléatoire associé à la famille des indicatrices floues disjonctives I_s.

si on note $m_Y(x) = \sum_{k=1}^{n} y_k \ I_{x_k}(x)$

on a

$$C_{YY^t|X}(x) = E\left\{ \left[Y-E^X(Y)\right]\left[Y-E^X(Y)\right]^t \mid X = x \right\}$$

$$= \sum_{k=1}^{n} \left\{ C^Y{}_k + \left(y_k-m_Y(x)\right)\left(y_k-m_y(x)\right)^t \right\} I_{x_k}(x)$$

Démonstration

Calculons d'abord la densite marginale de X

$$f_X(x) = \int_{R^q} \left(\sum_{k=1}^{n} p_k \ \Delta_X(x|x_k).\Delta_Y(y|y_k, C^Y{}_k) \right) dy$$

$$= \sum_{k=1}^{n} p_k \, \Delta_X(x|x_k) \int_{R^q} \Delta_Y(y|y_k, C^Y_k) \, dy$$

$$= \sum_{k=1}^{n} p_k \Delta_X(x|x_k)$$

La densité conditionnelle de Y par rapport à X s'écrit alors

$$f_{Y|X=x}(y) = \frac{f_{X,Y}(x,y)}{f_X(x)} = \sum_{k=1}^{n} \Delta(y|y_k, C^Y_k) \, I_{x_k}(x)$$

$$\text{donc } m_Y(x) = \int_{R^q} y. \, f_{Y|X=x}(y) \, dy$$

$$= \sum_{k=1}^{n} \left\{ \int_{R^q} y. \, \Delta_Y(y|y_k, C^Y_k) \, dy \right\} I_{x_k}(x)$$

$$= \sum_{k=1}^{n} y_k \, I_{x_k}$$

on vérifie d'autre part que $\forall a \in R^q$

$$\int_{R^q} (y-a)(y-a)^t \Delta_Y(y|y_k, C^Y_k) \, dy = C^Y_k + (y_k-a)(y_k-a)^t$$

donc

$$C_{YY^t|X}(x) = \int_{Rq} [y-m_Y(x)][y-m_Y(x)]^t \, f_{Y|X=x}(y) \, dy$$

$$= \sum_{k=1}^{n} I_{x_k}(x) \int_{R^q} (y-m_Y(x))(y-m_Y(x))^t \, \Delta_Y(y|y_k, C^Y_k) \, dy$$

$$= \sum_{k=1}^{n} I_{x_k}(x) \left\{ C^Y_k + (y_k-m_Y(x))(y_k-m_Y(x))^t \right\}$$

b) Commentaires

La condition de pseudo-factorisation de la densité peut sembler assez restrictive, elle signifie une espèce d'indépendance locale entre X et Y. C'est pourtant sous cette forme que se présente la plupart du temps l'estimation de la densité par les noyaux de Parzen : si les nombres des points de donnée est très grand on

peut faire l'approximation de cette indépendance locale. Ce théorème n'est pas applicable si on réduit le nombre de noyaux.

c) Lien avec la régression par boules

Considérons l'estimation de la densité par les noyaux de Parzen dans R^{p+q} avec le noyau cylindrique suivant

$$\Delta_{XY}(x,y|x_k,y_k) = \Delta_X(x|x_k).\Delta_Y(y|y_k,C^Y{}_k)$$

ou $\Delta_X(x|x_k) \left| \begin{array}{l} = \frac{1}{V(h)} \quad \text{si } \|x-x_k\|<h \\ = 0 \text{ sinon} \end{array} \right.$

ou V(h) est le volume de la sphère enclidienne de rayon h

$$f_X(x) = \frac{1}{n} \sum_{k=1}^{n} \Delta(x|x_k)$$

$$\text{donc } I_{x_k}(x) = \frac{1}{n} \frac{\frac{1}{V(h)}}{\frac{1}{nV(h)} (\#x_i,\ \|x_i-x\|<h)} \qquad \text{si } \|x-x_k\|<h$$

C'est le cas de la régression à boules avec rayon h fixé.
Cette interprétation de la régression à boules fournit des critères de choix pour h, en prenant celui-ci optimal du sens de l'estimation de la densité par les noyaux de Parzen.

Notons que l'estimateur de la régression donné par

$m_Y(x) = \sum_{k=1}^{n} I_{x_k}(x)y_k$ avec le noyau quelconque a été proposé dès 1976 par Collomb [8] et que les propriétés de convergence sont bien connues maintenant (voir [10]).

La régression à boules par les plus proches voisins peut aussi être placé dans le cadre du théorème 2-1 en considérant l'estimateur de la densité de Breiman (voir Chapitre I).

3 - Extension du modèle linéaire

a) Théorème 2-2

soient X et Y deux variables aléatoires à valeurs respectivement dans R^p et R^q telles que le couple (X,Y) appartiennent à $D^{p+q}(\Omega, A, P)$

et admette la densité de probabilité suivante :

$$f_{XY}(x,y) = \sum_{k=1}^{n} p_x \quad \Delta(x,y|m_k,C_k)$$

où Δ est une densité gaussienne

$$m_k = \begin{bmatrix} x_k \\ y_k \end{bmatrix} \qquad C_k = \begin{bmatrix} C_k^{11} & C_k^{12} \\ C_k^{21} & C_k^{22} \end{bmatrix}$$

alors

$$f_X(x) = \sum_{k=1}^{n} p_k \ \Delta(x|x_k, C_k^{11})$$

et si on pose $I_{x_k}(x) = p_k \ \dfrac{\Delta(x|x_k, C_k^{11})}{f_X(x)}$

on a

$$E(Y|X=x) = m_Y(x) = \sum_{k=1}^{n} I_{x_k}(x) \ \tilde{y}_k(x)$$

avec $\tilde{y}_k(x) = y_k - C_k^{21} (C_k^{11})^{-1} (x - x_k)$

de même

$$C_{YY^t|X}(x) = E\left\{ \left[Y - E^X(Y)\right]\left[Y - E^X(Y)\right]^t \ | X=x \right\}$$

$$= \sum_{k=1}^{n} I_{x_k}(x) \left[C_k^{22} - C_k^{21} (C_k^{11})^{-1} C_k^{12} + (\tilde{y}_k(x) - m_Y(x)) (\tilde{y}_k(x) - m_Y(x))^t \right]$$

Démonstration

Dans [1] Anderson donne les distributions conditionnelles d'une variable aléatoire multinormale.

si (X,Y) suit une loi normale multivariée de paramètre

$$m_o = \begin{bmatrix} x_o \\ y_o \end{bmatrix} \text{ et } C_o = \begin{bmatrix} C_o^{11} & C_o^{12} \\ C_o^{21} & C_o^{22} \end{bmatrix}$$

alors $E\left[Y|X=x\right] = y_o + C_o^{21} (C_o^{11})^{-1} (x-x_o) = \tilde{y}_o(x)$

$$E\left[(Y=\tilde{y}_o)\ (Y-\tilde{y}_o)^t \mid X=x\right] = C_o^{22} - C_o^{21}\ C_o^{11}\ C_o^{12}$$

donc si (X,Y) admet une densité

$$f_{XY}(x,y) = \sum_{k=1}^{n} p_k \Delta(x,y|m_k,C_k)$$

on a

$$m_Y(x) = \int y \quad f_{Y|X=x}(y)\, dy$$

$$= \int y \quad \frac{f_{XY}(x,y)}{f_X(x)}\, dy$$

$$= \sum_{k=1}^{n} \frac{p_k}{f_X(x)} \int y\ \Delta(x,y|m_k,C_k)\, dy$$

$$= \sum_{k=1}^{n} \frac{p_k\ \Delta(x|x_k,C_k^{11})}{f_X(x)} \int y \quad \frac{(x,y|m_k,C_k)}{(x|x_k,C_k^{11})}\, dy$$

$$= \sum_{k=1}^{n} I_{x_k}(x)\ \tilde{y}_k(x)$$

$$C_{YY^t|X}(x) = \int (y-m_y(x))\ (y-m_Y(x))^t\ \frac{f_{XY}(x,y)}{f_X(x)}\, dy$$

$$= \sum_{k=1}^{n} \frac{p_k\ (x|x_k,C_k^{11})}{f_X(x)} \int \left[(y-y_k)\ (y-y_k)^t + (y_k-m_Y)\ (y_k-m_Y)\right] \frac{\Delta(x,y|m_k,C_k}{\Delta(x|x_k,\ C_k^{11})}\, dy$$

$$= \sum_{k=1}^{n} I_{x_k}(x) \left\{(C_k^{22} - C_k^{21}\ (C_k^{11})^{-1}\ C_k^{12}) + (\tilde{y}_k(x)-m_Y(x))\ (\tilde{y}_k(x)-m_Y(x))^t\right\}$$

B) Application

Ce théorème contient comme cas particulier le théorème 2-1 avec noyaux normaux

il permet donc d'obtenir la régression non-paramétrique dans le cas de l'estimateur de la densité proposé par Deheuvels et en particulier proposé dans le théorème 1-2.

Notons que la généralisation proposée de l'estimateur de Breiman rentre aussi dans ce cadre.

L'application la plus importante vient de la remarque suivante : si n=1 et C_1 est la matrice de covariance empirique, la regression proposée est la régression dans le modèle linéaire (voir par exemple [12]).

Ainsi, si on considère la densité obtenu dans le cadre du II^B, réduction du nombre des noyaux par une méthode de classification ou comme analyse en mélange de composantes gaussiennes, nous obtenons une méthode de régression semi-paramétrique, qui contient à la fois le modèle linéaire et le modèle non-paramétrique de la regression à noyaux.

Cette méthode ne suppose pas l'hypothèse très restrictive de linéarité. Cependant, si le caractère de linéarité est apparent, il sera reconnu par une bonne classification. D'un autre côté, cette méthode ne demande pas le volume important de temps calcul exigé par la régression non-paramétrique.

4 - Qualité de la régression : calcul approché du rapport de corrélation

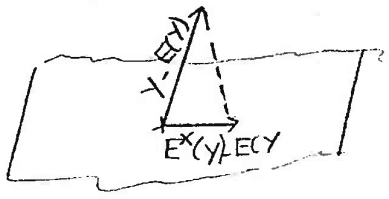
Rappel de la définition

soit Y une variable aléatoire scolaire (q=1)

Le rapport de corrélation de Y avec X est défini par

$$\eta(Y|X)=1 - \frac{E\left((Y-E^X(Y))^2\right)}{\sigma_Y^2} = \frac{E\,(E^X(Y)-E(Y))^2}{E\,(Y-E(Y))^2}$$

$$\eta(Y|X) = \frac{E\,(E^X(Y))^2 - E(Y)^2}{E\,(Y^2) - E(Y)^2}$$



C'est le carré du cosimus quand on considère l'espérance conditionnelle comme projection dans L^2.

Plaçons nous dans le cadre général du théorème 2-2

$$f_{XY}(x,y) = \sum_{k=1}^{n} P_k \quad \Delta(x,y|m_k,C_k)$$

$$E(Y) = \int y \sum_{k=1}^{n} P_k \Delta(y|y_k, C_k^{22})\,d_y$$

$$= \sum_{k=1}^{n} P_k\, y_k$$

$$E\,(Y^2) = \int y^2 \sum_{k=1}^{n} P_k \quad (y|y_k, C_k^{22})\,d_y$$

$$= \sum_{k=1}^{n} P_k \quad \left[C_k^{22} + y_k^2 \right]$$

$$E^X\,(Y) = \sum_{k=1}^{n} I_{x_k}(x).\, \overset{\sim}{y_k}(x)$$

On peut calculer $E(E^X(Y)^2)$ d'un point de vue numérique. En effet

$$E(E^X(Y)^2) = \int_{R^P} m_y(x)^2\, f_x(x)\, d_x$$

Il est possible d'en donner une expression approchée dans certains cas.

Si les I_{x_k} sont bien séparées, c'est à dire si $I_{x_k}(x)$ vaut 0 ou 1 on a

$I_{x_k}\, I_{x_{k'}} = \delta_{k,k}\, I_{x_k}$

ou δ est le symbole de Kronecker.

Ceci se produit en particulier dans le cas où le nuage de points de donnée est séparé en composantes normales, bien différencées et où on effectue une classification préalable.

$$\text{Alors } E(E^X(Y)^2) = \int \sum_{k=1}^{n} I_{x_k}(x)\, \tilde{y}^2(x) f_X(x) d_x$$

$$= \sum_{k=1}^{n} P_k \int \tilde{y}^2_k \, \Delta(x|x_k, C_k^{11}) d_x$$

donc si on pose

$$a_k = C_k^{12} (C_k^{22})^{-1} \text{ et } \bar{y} = \sum_{k=1}^{n} P_k y_k$$

$$E\,(E^X(Y)^2) = \sum_{k=1}^{n} P_k y_k^2 + \sum_{k=1}^{n} P_k \int_{a_k}^{t} (x-x_k)(x-x_k)^t a_k \, \Delta(x|x_k, C_k^{11}) d_x$$

$$= \sum_{k=1}^{n} P_k \left[y_k^2 + a_k^t \; C_k^{11} \; a_k \right]$$

En particulier : dans le cas pseudo-factorisé avec composantes disjointes

$$\eta(Y|X) = \frac{\sum_{k=1}^{n} P_k y_k^2 - \bar{y}^2}{\sum_{k=1}^{n} P_k (y_k^2 + C_k^{22}) - \bar{y}^2}$$

D - Indicatrices floues et fonctions aléatoires associées

Dans l'introduction de ce chapitre, nous avons présenté la notion d'indicatrice floue et de fonctions aléatoires correspondantes associées à une variable aléatoire X de DP(Ω , A, P) c'est à dire admettant une densité de la forme :

$$f(x) = \sum_{i=1}^{n} P_i \, \Delta \, (x|x_i, C_i)$$

Nous avons en particulier mis en évidence 2 familles d'indicatrices floues que nous allons étudier plus en détail (voir aussi [17]).

1 - Indicatrices floues disjonctives

a) Rappel de la définition et commentaires

Nous appelons indicatrices disjonctives associées à X les n fonctions suivantes :

$$I_{x_i}(x) = P_i \frac{\Delta(x|x_i,C_i)}{f_x(x)} \qquad \text{ou } x \in RP$$

Ces fonctions sont apparues plusieurs fois au courant de cette étude : comme indices d'appartenance dans la classification floue, comme fonctions pondérantes dans la régression...

Ces indicatrices ont donc une grande importance pratique d'où l'intérêt de leur étude.

b) Propriété 1 : justification de la disjonctivité

Propriété 1

Soit $A_k = \{ x \in RP,\ I_{x_k} > \frac{1}{2} \}$

Alors si $k' \neq k$ $\quad A_k \cap A_{k'} = \emptyset$

Commentaire : Les ensembles A_k ainsi définis peuvent être vides

Démonstration

$$f(x) = \sum_{i=1}^{n} P_i \ \Delta(x|x_i,C_i)$$

$$\text{donc} \quad \sum_{i=1}^{n} I_{x_i}(x) = \sum_{i=1}^{n} \frac{P_i \ \Delta(x|x_i,C_i)}{f(x)} = 1$$

soit $x_o \in A_k \cap A_{k'}$

$I_{x_k}(x_o) > \frac{1}{2}$ et $I_{x_{k'}}(x_o) > \frac{1}{2}$ impossible

donc $A_k \cap A_{k'} = \emptyset$

Remarques

* posons $B_k = \left\{ x \in RP,\ I_{x_k}(x) > I_{x_{k'}},\ \forall\, k' \neq k \right\}$

alors $A_k \subset B_k$ et $B_k \cap B_{k'} = \emptyset$

* si f(x) est une estimation par les noyaux de Parzen c'est à dire

$$f(x) = \frac{1}{n} \sum_{i=1}^{n} \Delta(x|x_i, C)$$

avec Δ noyau normal

alors $x_k \in B_k$ et $\bigcup_{k=1}^{n} B_k = RP - D$ ou $\lambda(D) = 0$

c) Invariance dans un changement de variable linéaire

Propriété 2

soit $X \in DP(\Omega, A, P)$

$I_{x_k}(x)$ sa famille d'indicatrices floues disjonctives associée

soit A une matrice carrée (pxp) inversible

et B un vecteur (px1)

si $Y = AX + B$ alors $Y \in DP(\Omega, A, P)$

La famille d'indicatrices floues disjonctives associée à Y vérifie

$I_{y_k}(y) = I_{x_k}(x)$ ou $y_k = A x_k + B$

et $y = Ax + B$

De plus $I^{X}_{x_k}(\omega) = I^{Y}_{y_k}(\omega)$ $\quad \forall \omega \in \Omega$

Démonstration

soit $f_X(x) = \sum_{i=1}^{n} P_i\, \Delta(x|x_i, C_i)$ la densité de X

Alors la densité de Y s'écrit

$f_Y(y) = |J(y)|\ f_x(A^{-1}(y-B))$ où $J(y)$ est le jacobien du changement de variable,

$J(y) = \dfrac{1}{\det A}$

posons $y_k = A_y + B$

$D_k = A\ C_k\ A^t$

alors $\Delta(y|y_k,D_k) = \dfrac{1}{\det A}\Delta\left(A^{-1}(y-B)|x_k,C_k\right)$

en effet

$$\int_y \Delta(y|y_k,D_k)dy = \frac{1}{\det A}\int_y \Delta(A^{-1}(y-B)|x_k,C_k)dy$$

$$= \int (A_y+B)\,\Delta(y|x_k,C_k)dy$$

$$= A\int y\ \Delta(y|x_k,C_k)dy + B = A_{x_k} + B$$

$$\int (y-y_k)(y-y_k)^t\ \Delta(y|y_k,D_k)dy$$

$$= \frac{1}{\det A}\int (y-y_k)(y-y_k)^t\ \Delta(A^{-1}(y-B)|x_k,C_k)dy$$

$$= \int (A_y+B-y_k)(A_y+B-y_k)^t\ \Delta(y|x_k,C_k)dy$$

or $B - y_k = -A_{x_k}$

$$= \int A(y-x_k)(y-x_k)^t\ A^t\ \Delta(y|x_k,C_k)dy$$

$$= A\ C_k\ A^t$$

Donc $f_Y(y) = \sum_{i=1}^{n} P_i\ \Delta(y|y_i,D_i)$ donc $Y \in DP(\Omega, A, P)$

$$I^Y_{y_k}(\omega) = \frac{P_k\ \Delta(Y(\omega)|y_k,D_k)}{f_Y(y)}$$

$$= P_k\frac{|J|\ \Delta(A^{-1}(Y(\omega)-B)|x_k,C_k}{|J|\ \sum_{i=1}^{n} P_i\ \Delta(A^{-1}(Y(\omega)-B)|x_i,C_i)}$$

$$= \frac{P_k \ \Delta(X(\omega)|x_k,C_k)}{\sum_{i=1}^{n} P_i \Delta(X(\omega)|x_i,C_i)} = I^X_{x_k}(\omega)$$

Remarque

Soit φ bijection contrairement différentiable non linéaire

$Y = \varphi(X)$

$$f_Y(y) = |J(y)| \sum_{i=1}^{n} P_i \ \Delta(\varphi^{-1}(y)|x_i,C_i)$$

Pour que $Y \in DP(\Omega,A,P)$ il faut que $\Delta(\varphi^{-1}(y)|x_i,C_i)$ admette des moments d'ordre 2

Alors $I^Y_{y_k} = I^X_{x_k}$ mais $y_k = |J| \int \varphi^{-1}(y) \Delta(y|x_k,C_k) d_y \neq \varphi(x_k)$

d) Propriétés de la fonction aléatoire associée

soit $X \in DP(\Omega,A,P)$ et I_{x_k} les indicatrices disjonctives associées

alors $I^X_{x_k} = I_{x_k}(X)$ est la fonction aléatoire associée

on a

$$\sum_{k=1}^{n} I^X_k = 1$$

$$E(I^X_k) = P_k$$

$$E(X.I^X_k) = P_k \ x_k$$

$$E(XX^t \ I^X_k) = P_k \ (C_k + x_k \ x^t_k)$$

Démonstration

La première propriété a été déjà montrée au courant de la démonstration de la propriété 1

$$E(I^X_k) = \int I_k(x) \ dP^X(x) = \int \frac{P_k \ \Delta(x|x_k,C_k)}{f(x)} \ f(x) \ d_x$$

$$= P_k \int \Delta(x|x_k,C_k)dx = q_k$$

$$E\ (X.I_k^X) = \int x.I_k(x)dP^X(x) = \int P_k x\ \Delta(x|x_k,C_k)dx = P_k\ x_k$$

$$E\ (XX^t\ I_k^X) = \int xx^t\ I_k(x)dP^X(x) = P_k \int xx^t\ (x|x_k,C_k)dx = P_k(C_k+x_k\ x_k^t)$$

2 - Indicatrices floues canoniques

a) Pour l'étude des indicatrices canoniques plaçons nous dans le cadre d'une estimation de la densité obtenue par les noyaux de Parzen avec noyau multinormal

$$f(x) = \frac{1}{n} \sum_{i=1}^{n} \Delta(x|x_i,C)$$

D'autre part considérons la densité g(x) obtenue par les techniques présentées au $\coprod$B

$$g(x) = \sum_{k=1}^{K} \Delta(x|m_k,C_k)$$

Nous obtenons ainsi deux familles d'indicatrices canoniques

$$J_{x_i}(x) = \frac{\Delta(x|x_i,C)}{\Delta(x_i|x_i,C)} = \exp\left\{-\frac{1}{2}(x-x_i^t)\ C^{-1}(x-x_i)\right\} \quad ; i=1,\ldots,n$$

$$J'_{m_k}(x) = \frac{\Delta(x|m_x,C_k)}{\Delta(m_k)m_k,C_k)} = \exp\left\{-\frac{1}{2}(x-m_k)^t C_k^{-1}(x-m_k)\right\} \quad ; k=1,\ldots,K$$

Pour examiner les propriétés de $J^X(x)$, mettons d'abord une structure algébrique sur l'ensemble des indicatrices canoniques.

b) Définition d'une indicatrice produit et propriétés algébriques

Remarque : une indicatrice canonique J est entièrement déterminée par une matrice symétrique, définie positive C et un vecteur m

$$J\ (x) = \exp\left\{-\frac{1}{2}(x-m)^t\ C^{-1}\ (x-m)\right\}$$

Définition d'une indicatrice produit

Soit J' (x) donnée par (C', m')

et J" (x) donnée par (C", m")

alors J = J' * J" est appelée indicatrice produit de J' et de J"

J est donné par (C, m)

ou $C = (C'^{-1} + C''^{-1})^{-1}$

$m = C\,(C'^{-1}m' + C''^{-1}\,m'')$

Propriété 1

$$J'(x).J''(x) = \exp\left\{-\frac{1}{2}(m' - m'')(C'+C'')^{-1}(m'-m'')\right\}\ (J' * J'')(x)$$

dem : c'est une conséquence immédiate du lemme d'albèbre linéaire de II B 3c

Propriété 2

L'ensemble des indicatrices canoniques muni de la loi * est un semi-groupe commutatif qui n'admet pas d'élément neutre

Démonstration

J' * J" = J" * J de façon immédiate

or

$$(J * J') * J'' = (C,m)*(C',m') * (C'',m'')$$

$$= (C^{-1}+C'^{-1})^{-1},\ (C^{-1}+C'^{-1})^{-1}(C^{-1}m+C'^{-1}m') * (C'',m'')$$

$$= (C^{-1}+C'^{-1}+C''^{-1})^{-1},\ (C^{-1}+C'^{-1})^{-1}\ (C^{-1}m + C'^{-1}m' + C''^{-1}m'')$$

d'où l'associativité.

Il n'y a pas d'élément neutre.

En effet, supposons que (C_o,m_o) soit un élément neutre alors $(C,m)*(C_o,m_o)=(C,m)$

$(C^{-1} + C_o^{-1})^{-1} = C \Rightarrow C_o^{-1} = 0$ ce qui est impossible car C_o doit être inversible.

On peut se poser la question suivante :

soit $J_o = (C_o,m_o)$ et $J_1 = (C_1,m_1)$

Dans quel cas existe-t-il J tel que $J_o * J = J_1$ (1)

$$(1) \Rightarrow \begin{cases} C_1 = \left[C_o^{-1} + C^{-1}\right]^{-1} \\ m_1 = C_1 \left[C_o^{-1} m_o + C^{-1} m\right] \end{cases}$$

$$\Rightarrow \begin{cases} C^{-1} = C_1^{-1} - C_o^{-1} \\ m = m_1 + C_1 C_o^{-1} (m_1 - m_o) \end{cases}$$

Donc si $C_1^{-1} - C_o^{-1}$ est une matrice définie positive alors (1) admet la solution unique

$$J = (C, m) = \left((C_1^{-1} - C_o^{-1})^{-1},\ m_1 + C_1 C_o^{-1} (m_1 - m_o)\right)$$

c) Propriétés des fonctions aléatoires associées

Nous allons donner quelques propriétés des fonctions aléatoires associées aux deux familles d'indicatrices canoniques présentées. Nous calculerons notamment les "moments locaux" de X ($E(X.J^X)$, et une proximité entre deux indicatrices moyenne au sens de X: $E(J_o^X J_1^X)$. Ces moments sont donnés avec les deux estimations de la densité f et g.

Propriété 3

Soit J_o et J_1, deux indicatrices définies respectivement par (C_o, m_o) et (C_1, m_1)

Notons $\alpha(J_o, J_1) = \dfrac{1}{\sqrt{\det\ (I_p + C_o C_1^{-1})}} \exp\left\{-\frac{1}{2}(m_o - m_1)^t (C_o + C_1)^{-1} (m_o - m_1)\right\}$

et $m\ (J_o, J_1) = C_o^{-1} + C_1^{-1}\ (C_o^{-1} m_o + C_1^{-1} m_1)$

Alors

1) $E^f\ (J_o^X) = \dfrac{1}{n} \sum_{i=1}^{n} \alpha(J_{x_i}, J_o)$

$E^f\ (X\ J_o^X) = \dfrac{1}{n} \sum_{i=1}^{n} \alpha(J_{x_i}, J_o)\ m\ (J_{x_i}, J_o)$

$E^f\ (J_o^X\ J_1^X) = \exp\left\{-\frac{1}{2}(m_o - m_1)^t\ (C_o + C_1)^{-1} (m_o - m_1)\right\} . E^f\left[(J_o * J_1)^X\right]$

$$2)\ \mathrm{Eg}\ (J_o^X) = \sum_{k=1}^{K} P_k\ \alpha(J'_{m_k}, J_o)$$

$$\mathrm{Eg}\ (X\ J_o^X) = \sum_{k=1}^{n} P_k\ \alpha(J'_{m_k}, J_o)\ m\ (J'_{m_k}, J_o)$$

$$\mathrm{Eg}\ (J_o^X\ J_1^X) = \exp\left\{-\frac{1}{2}(m_o-m_1)^t(C_o+C_1)^{-1}(m_o-m_1)\right\}\ \mathrm{Eg}\left[(J_o * J_1)^X\right]$$

Démonstration

Démontrons le 2)

$$\mathrm{Eg}\ (J_o^X) = \int \exp\left\{-\frac{1}{2}(x-m_o)^t C_o^{-1}(x-m_o)\right\} \sum_{k=1}^{K} \frac{P_k}{(2\pi)^{P/2}\sqrt{\det C_k}} \exp\left\{-\frac{1}{2}(x-m_k)^t C_k^{-1}(x-m_k)\right\} d_x$$

$$= \sum_{k=1}^{K} \frac{k}{(2\pi)^{P/2}\sqrt{\det C_k}} \int J_o(x)\ J'_{m_k}(x)\ d_x$$

$$= \sum_{k=1}^{K} \frac{P_k}{\sqrt{\det C_k}} \exp\left\{-\frac{1}{2}(m_o-m_k)(C_o+C_k)^{-1}(m_o-m_k)\right\} \frac{1}{(2\pi)^{P/2}} \int (J_o * J'_{m_k})(x) d_x$$

$$= \sum_{k=1}^{K} \frac{P_k}{\sqrt{\det(I_p+C_k C_o^{-1})}} \exp\left\{-\frac{1}{2}(m_o-m_k)(C_o=C_k)^{-1}(m_o-m_k)\right\}$$

$$\mathrm{Eg}\ (X\ J_o^X) = \sum_{=1}^{K} \frac{P_k}{\sqrt{\det C_k}} \exp\left\{-\frac{1}{2}(m_o-m_k)^t(C_o+C_k)^{-1}(m_o-m_k)\right\} \frac{1}{(2\pi)^{P/2}} \int x(J_o * J'_{m_k}) d_x$$

$$= \sum_{k=1}^{K} P_k\ \alpha(J'_{m_k}, J_o)\ m\ (J'_{m_k}, J_o)$$

$$J_o(x).\ J_1(x) = \exp\left\{-\frac{1}{2}(m_o-m_1)^t\text{-}C_o+C_1)^{-1}(m_o-m_1)\right\} \cdot (J_o * J_1)(x)$$

donc $\mathrm{Eg}\ (J_o^X\ J_1^X) = \exp\left\{-\frac{1}{2}(m_o-m_1)^t(C_o+C_1)^{-1}(m_o-m_1)\right\}\ \mathrm{Eg}\left[(J_o * J_1)^X\right]$

E - Applications heuristiques des indicatrices canoniques

1) Un coefficient de corrélation entre points

Ce coefficient a été proposé et étudié par Mallet dans [18]

Définition

$$R^X(a,b) = \frac{E(J_a^X \; J_b^X}{(E(J_a^X)^2 \; E(J_b^X)^2)^{1/2}}$$

Pour calculer ces espérances, on suppose que X admet la densité f (x) obtenue par la méthode des noyaux de Parzen et on utilise la propriété 3 du D.

Un choix possible des indicatrices est celui donné par le processus complété défini au ∏A3d

Propriétés

* $0 \leqslant R(a,b) \leqslant 1$

* R (a,b) est d'autant plus grand que la densité est grande entre a et b.

2) Un coefficient de corrélation entre variables et données

Ce coefficient et son application ont été présentés dans [19] soit n réalisations $x_1, \cdots, x_n$ du vecteur aléatoire X

$$X = \begin{bmatrix} X(1) \\ \\ X(p) \end{bmatrix} \quad \text{soit } f(x) = \frac{1}{n} \sum_{i=1}^{n} \Delta(x|x_i,C)$$

L'estimation de la densité de X par les noyaux de Parzen avec Δ gaussien.

soit $J_s(x) = \exp\left\{-\frac{1}{2}(x-s)^t \; C^{-1}(x-s)\right\}$, $s \in RP$

La famille d'indicatrices floues canoniques associées a f(x) soit J_s^X la fonction aléatoire associée.

Alors

Définition

$$R^X(X(i),s) = \frac{E(X(i).\; J_s^X)}{(Var(X(i).E(J_s^X)^2)^{1/2}}$$

est appelé coefficient de corrélation entre le point s et la variable X(i).

Remarques

* Le coefficient R^X (X(i),s) peut être considéré comme la moyenne normalisée de X(i) dans un domaine flou autour de s.

* Pour avoir une stabilité vis à vis d'un changement de variable linéaire, il faut centrer et réduire les données variable par variable.

3) Proposition d'une analyse des correspondances

soit $X = \begin{bmatrix} x_{11} & \cdots & x_{n1} \\ & & \vdots \\ x_{1p} & \cdots & x_{np} \end{bmatrix}$

Un tableau de données quantitatives.

La méthode proposée est la suivante :

a) Centrer et réduire le tableau de données variable par variable, soit X' le tableau obtenu.

b) Effectuer une analyse en composantes principales sur le tableau X'.

c) Estimer les densités marginales sur les plans factoriels.

d) Projeter les points de donnée sur les diagrammes de corrélation obtenus par les coefficients de corrélation définis précédemment.

Nous obtenons ainsi un moyen de projeter variables et données sur un même diagramme. En relevant les proximités, nous examinons les correspondances entre variables et données.

Un exemple sur données réelles sera présenté en annexe.

Conclusion

Dans ce travail, nous avons étudié un modèle non-paramétrique que nous avons entre autres approché par un modèle paramétrique souple, tenant compte des paramètres empiriques d'un ensemble de données (moments, moment locaux, plus proche voisins). Ce sont ces idées qui, peut-être en parallèle avec d'autres techniques (telles Jackknife ou Bootstrap), nous semblent déboucher sur une recherche future.

BIBLIOGRAPHIE CHAPITRE II

1. T.W. ANDERSON
   An introduction to multivariate statistical analysis
   (John Wiley, New York) (1958).

2. J.C. BEZDEK
   Pattern recognition with fuzzy objective function algorithms
   (Plenum Press) (1981).

3. J.C. BEZDEK et J.C. DUNN
   Optimal fuzzy partitions : a heuristic for estimating the parameters in a mixture of normal distributions.
   IEEE Trans. Comp. Vol C-24 (1975) 835-838.

4. D.A. BINDER
   Approximations to bayesian Clustering rules
   Biometrika (1981) 68, 1, 275-285.

5. P. BRYANT and J.A. WILLIAMSON
   Asymptotic behaviour of classification maximum likelyhood estimates.
   Biometrika (1978), 65, 2, 273-281.

6. P. CAZES
   Décomposition d'un histogramme en composantes gaussiennes.
   Revue de statistique appliquée (1976) XXIV n° 1, 63-82.

7. P. CAZES

8. G. COLLOMB
   Estimation non paramétrique de la régression : revue bibliographique.
   Internat. Statist. Review (1981) 49, 75-93.

9. N.E. DAY
   Estimating the components of a mixture of normal distributions.
   Biometrika (1969), 56, 3, 463-467.

10. L. DEVROYE
    On the almost everywhere convergence of non parametric regression function estimates.
    Annals of statistic (1981), 9, 6, 1310-1319.

11. E. DIDAY
    Optimisation en classification automatique.
    Revue française d'automatique, informatique et recherche opérationnelle (1972) V-3, 61-96.

12. E. DIDAY, J. LEMAIRE, J. POUGET, F. TESTU
    Eléments d'analyse des données (DUNOD) (1982).

13. R.O. DUDA et P.E. HART
    Pattern classification and scene analysis.
    (John Wiley interscience) (1973).

14. J.C. DUNN
Well-separated clusters and optimal fuzzy partitions
Journal of cybernetics (1974) 4, 1, 95-104.

15. D.E. GUSTAFSON et W. KESSEL
Fuzzy clustering with a fuzzy covariance matrix
Proc. IEEE-CDC Vol 2 (IEEE-Press, Pixataway) (1979) 761-766.

16. J. HARTIGNAN
Clustering algorithms
(J. Wiley) (1975).

17. J.L. MALLET
Propositions for fuzzy characteristic functions in data analysis.
Proc. Compstat 82. Physica-Verlag (1982) 324-329.

18. J.L. MALLET
Definition of a correlation coefficent between data points
Mathematical geology, 15, 2 (1983).

19. J.L. MALLET et P. WILD
An analogue to correspondence analysis with fuzzy characteristic functions.
(A paraître dans Sciences de la terre).

20. F.H.C. MARRIOTT
Optimization methods of cluster analysis
Biometrika (1982) 69, 2, 417-421.

21. K. MATUSITA et N. OHSUMI
A criterion for choosing the number of clusters in cluster analysis.
Recent developments in statistical inference and date analysis (1980).

22. L. REJTÖ et P. REVESZ
Density estimation and pattern classification
Problems of control and information theory, 2 (1973), 67-80.

23. E. RUSPINI
Numerical methods for fuzzy clustering
Information science Vol 2 (1970), 319-350.

24. A. SCHROEDER
Analyse d'un mélange de distributions de probabilité de même type.
RSA (1976) XXIV n° 1.

25. M.J. SYMONS
Clustering criteria and multivariate normal mixtures
Biometrics 37 (1981), 35-43.

26. J.H. WOLFE
Pattern clustering by multivariate mixture analysis
Multivariate behavioral research (1970) 5, 329-350.

27. L.A. ZADEH
Fuzzy sets as a basis for a theory of possibility.
Fuzzy sets and systems, 1, 3-28 (1978).

ANNEXE

Dans cette annexe nous avons testé les méthodes proposées dans notre travail. Les calculs ont été exécutés sur IRIS 80 à l'Institut Universitaire de Calcul Automatique de Lorraine.

Nous avons utilisé la bibliothèque de cartographie automatique "cartolab" développée par Monsieur MALLET. Le générateur de nombres au hasard ainsi que le programme de simulation d'une loi normale sont ceux proposés dans l'ouvrage "traitement des données statistiques" par Messieurs LEBART, MORINEAU et FENELON.

Les programmes ont été écrits en FORTRAN IV.

I. ESTIMATION AUTOMATIQUE DE LA DENSITE MULTIVARIEE

1) Nous avons testé dans cette partie les méthodes proposées dans le premier chapitre de notre travail ; le but de ce chapitre étant de passer d'un ensemble de données à une bonne estimation de la densité sous-jacente sans l'intervention de l'utilisateur. Ceci implique la détermination automatique du paramètre de lissage que nous avons décomposé en un paramètre taille et un paramètre forme de la "fenêtre".

Le paramètre de taille est obtenu par notre subroutine IMSE 1 qui applique l'algorithme présenté au I3. (Rappelons que cet algorithme généralise au cas multivarié un algorithme proposé par Tapia et Thompson([13]ch. I) optimisant le critère du MISE asymptotique d'Epanechnibrov.

Les paramètres de forme testés sont d'une part la matrice identité (correspondant aux noyaux produits usuels) et d'autre part le paramètre proposé en application du théorème 2 prenant en compte la covariance globale de l'échantillon. Nous parlerons alors de noyaux covariants.

2) Simulations

a) Robustesse du paramètre taille par rapport à l'échantillonage

Nous avons tiré au hasard 10 séries de 100 points au cours d'une simulation d'une loi normale bivariée de matrice de covariance identité.

Le tableau 1 contient les tailles des fenêtres correspondantes obtenues par la subroutine IMSE 1.

n° de l'échantillon	1	2	3	4	5	6	7	8	9	10
taille IMSE 1	.158	.148	.136	.164	.162	.158	.165	.175	.169	.178

b) Influence des paramètres taille et forme

Nous avons tiré au hasard 100 points au cours d'une simulation d'une loi normale bivariée de matrice de covariance

$$\begin{matrix} \frac{7}{16} & \frac{3\sqrt{3}}{16} \\ \frac{3\sqrt{3}}{16} & \frac{13}{16} \end{matrix}$$

La figure 1 représente le nuage des points. La figure 2 est un bloc-diagramme de la densité théorique correspondant à cette loi binormale que nous vous proposons d'estimer.

Les figures 3 montrent les estimations de cette densité, obtenues avec ces nuages par des noyaux produits (isotropes) avec différentes tailles de la fenêtre.

La figure 3a correspond à la taille obtenue par la subroutine IMSE 1 (soit $\sigma = .122$)

La figure 3b correspond à la taille double ($\sigma = .244$)

La figure 3c correspond à la taille quadruple ($\sigma = .488$)

Les figures 4 montrent les estimations de la densité obtenues pour des noyaux covariants (= anisotropes) avec les mêmes tailles que dans les figures 3.

3) Conclusion

a) Les résultats du tableau 1 montrent que la subroutine IMSE 1 donne des résultats stables par rapport à l'échantillonage. Cependant la valeur théorique étant $\sigma = .546$, celle-ci est nettement sous-estimée.

Cette tendance semble être confirmée par les figures 3 et 4. La meilleure taille de la fenêtre (quelle que soit la forme du noyau) est comprise entre 2 et 4 fois la valeur obtenue par l'algorithme IMSE 1.

La détermination d'une taille optimale reste donc un problème ouvert.

b) Le choix d'un noyau covariant semble avoir des conséquences heureuses : en comparant les figures 3 et 4 et en particulier la figure 3c avec la figure 4c, on constate qu'en dépit d'un trop grand lissage (qui est apparent si l'on compare avec la densité théorique de la figure 2) la figure 3c semble légèrement bimodale ce qui n'est plus le cas avec un noyau covariant. D'autre part le mode est **moins** sous-estimé dans le cas de la figure 4c que dans la figure 3c.

II. APPLICATIONS

A) Classification

Nous avons écrit un programme effectuant la classification au sens du maximum de vraisemblance présenté dans II B.

1) Pour sacrifier à la tradition, nous avons d'abord effectué la classification sur le fichier des iris de Fisher (cf [2]). Ce fichier est constitué de trois groupes de 50 fleurs dont on a mesuré la longueur et la largeur des pétales et des sépales.

Après classification seules les fleurs n° 69, 71, 73 et 84 sont mal classées. La fleur n° 78 est douteuse.

Le pourcentage d'erreur est donc inférieur à 3%.

2) Nous avons d'autre part simulé un mélange de 3 lois normales bivariées de poids respectifs : $\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$; de centres $\begin{bmatrix}-1.2\\0\end{bmatrix}, \begin{bmatrix}0\\.17\end{bmatrix}, \begin{bmatrix}1.2\\0\end{bmatrix}$ de matrices de covariances :

$$\begin{bmatrix}.156 & .094\\.094 & .156\end{bmatrix}, \begin{bmatrix}.25 & 0.\\0. & .0625\end{bmatrix}, \begin{bmatrix}.156 & -.094\\-.094 & .156\end{bmatrix}$$

Nous avons tiré 300 points au hasard issus de cette loi. (voir figure 6) Nous avons classé ces points en 3 groupes. Nous obtenons les paramètres suivants.

poids : .329 , .416 , .254

centres : $\begin{bmatrix}-1.15\\-0.02\end{bmatrix}, \begin{bmatrix}.12\\.18\end{bmatrix}, \begin{bmatrix}1.26\\-0.03\end{bmatrix}$

matrices de covariances :

$$\begin{bmatrix}.135 & .092\\.092 & .146\end{bmatrix}, \begin{bmatrix}.478 & .002\\.002 & .061\end{bmatrix}, \begin{bmatrix}.132 & -.083\\-.083 & .125\end{bmatrix}$$

Compte tenu de l'interpénétration de ces groupes (aucune classification à vue n'est possible), cette classification est très bonne.

3) Nous pouvons néanmoins regretter que cette méthode utilise un temps machine plus grand que par exemple les procédures de la famille des nuées dynamiques. De plus ce temps augmente sensiblement avec le nombre des points à

classer. Ainsi une série de 10 itérations de l'algorithme sur les 300 points du deuxième exemple utilise 25 secondes de temps de calcul d'un ordinateur IRIS 80.

B) Régression

Nous avons testé les méthodes de régression proposées au II C sur le même fichier de données simulées présentées précédemment (voir figure 6).

Etant donné que nous connaissons la densité sous-jacente nous avons pu effectuer le calcul exact de la moyenne conditionnelle, c'est cette courbe qui est tracée sur la figure 6. Remarquons que sur les bords de la figure, cette courbe ne correspond pas tout à fait à la solution intuitive que l'on a envie de tracer.

Résultats obtenus

a) Les figures 7 représentent les courbes de régression obtenues par la méthode du noyau présentée au paragraphe II C 2 :

- La première courbe représente la régression correspondant à une taille de fenêtre obtenue par la subroutine IMSE 1.
- La deuxième courbe correspond à trois fois cette taille.
- La troisième courbe correspond à six fois cette taille.

Nous constatons que la deuxième courbe est une bonne estimation de la regression ce qui confirme le point que la subroutine IMSE 1 sous-estime la taille du noyau. Les figures 7a et 7c nous montrent néanmoins que cette méthode de regression est peu robuste envers un mauvais choix de la taille de la fenêtre.

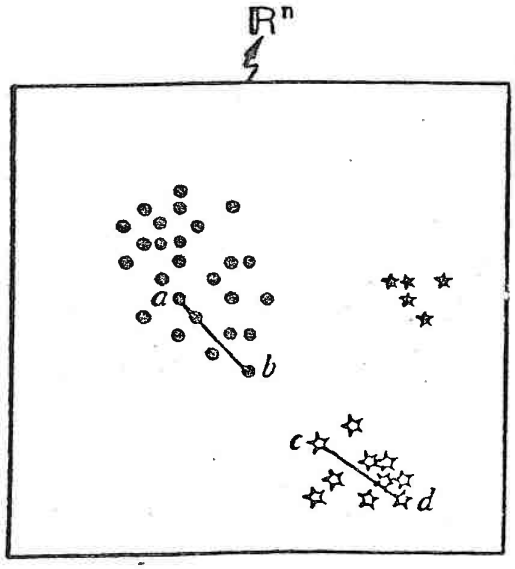
b) La figure 8 donne la régression obtenue par le théorème 2-2 avec classification préalable par le maximum de vraisemblance.

Notons que cette méthode ne dépend plus d'un paramètre de lissage et que le résultat obtenu est tout à fait satisfaisant.

Un point qui reste à étudier est la robustesse de cette méthode par rapport au choix du nombre de classes.

C) Corrélations entre points de données

Cette technique présentée au II E 1) et proposée par MALLET dans la référence [18] du chapitre II est illustrée par la figure 11. Notons qu'à l'aide de ces coefficients de corrélation nous pourrions définir des corrélations multiples entre groupes de points.



$R^X(a,b) > R^X(b,c)$

$R^X(c,d) > R^X(b,c)$

D) Corrélations entre variables et données

Pour montrer le comportement des "corrélations" entre variables et données, nous avons étudié celui-ci sur un ensemble de 20 points de données artificielles que nous avons centré et réduit (cf tableau 2).

Data	1	1.1	1.12	1.15	1.16	1.165	1.17	1.175	1.18	1.19	1.21	1.25	1.30	1.4	1.6	2	2.1	2.15	2.2	2.3
Centered reduced data	-1.09	-.81	-.76	-0.69	-0.67	-.66	-.65	-.64	-.62	-.60	-.55	-.46	-.34	-.11	.36	1.30	1.54	1.65	1.77	2.00

Les figures 9 et 10 représentent respectivement l'estimation de la densité univariée et la fonction y = R(X,x) où R(X,x) désigne la corrélation entre la variable aléatoire X et le point x telle qu'elle a été définie au II E 2. Notons que les extrema des deux fonctions coïncident.

E) Projection des corrélations

Dans ce paragraphe nous allons tester la méthode proposée au II E 3 sur un fichier proposé par Benzecri dans [1] qui donne le comportement hydrologique de 10 fleuves au Liban. (voir tableau 3)

Dans un premier examen de ce tableau de données, nous constatons que tous les fleuves ont à peu près le même comportement excepté El Assis dont le débit ne varie que légèrement au courant de l'année.

1) Dans un premier temps considérons les mois comme les points de donnée et les fleuves comme variables.

Après centrage et réduction nous obtenons le tableau (4). En effectuant l'analyse en composantes principales de ce tableau nous obtenons les diagrammes de corrélation de la figure sur lesquels nous avons projeté les points de données (les mois) par la méthode annoncée.

Excepté El Assis tous les fleuves sont très corrélés avec le premier facteur. On peut donc interpréter ce facteur comme étant caractéristique de la variation saisonnière du débit. De fait, les mois projetés sur ce diagramme sont divisés sur le premier facteur en mois secs et mois humides.

Le second facteur est très corrélé avec El Assis, en conséquence les mois projetés se divisent en mois pendant lesquels El Assis a un plus grand débit qu'en moyenne et les autres.

Proximités :
Janvier-Février sont proches de Demour, Litani et Beyrouth,
Avril-Mai sont proches de Abou-Ali et Ibrahim sur les deux diagrammes de corrélation.

2) Considérons maintenant les rivières comme point de donnée.

Après centrage et réduction nous obtenons le tableau (5).

La figure 12 montre les diagrammes de corrélation obtenu par l'analyse en composantes principales sur lesquels nous avons projeté les points de donnée (i e les fleuves).

Toutes les variables sont bien corrélées avec le premier facteur : les points de donnée correspondant se divisent en grands fleuves et petits fleuves : le rapport des fleuves entre eux est respecté quelle que soit la saison. Le second facteur divise les mois en 3 groupes :

- un groupe : été-automne correspondant aux débits faibles.
- un groupe : hiver-printemps correspondant aux débits élevés.
- deux mois intermédiaires.

Les rivières sont divisées dans les mêmes groupes que sur le premier facteur avec l'exception significative d'El Assis qui est plus proche des mois d'été : ce fleuve a un plus grand débit que les autres fleuves pendant ces mois.

Nous constatons d'autre part que les proximités signalées lors de l'analyse duale sont conservées.

3) Comparaison avec l'analyse des correspondances de Benzecri

Nous constatons dans un premier temps que les proximités signalées dans ce qui précède se retrouvent sur le diagramme de l'analyse des correspondances. (voir figure 13)

En particulier sur ce diagramme El Assis est très près des mois d'été, mais ainsi que nous l'avons signalé cette propriété n'est pas significative si l'on considère les fleuves comme variables : dans ce point de vue El Assis ne discrimine pas du tout entre les fleuves.

4) Conclusion

Nous avons ainsi obtenu une technique heuristique dont les résultats sont (du moins sur l'exemple considéré) cohérents avec les techniques existantes tout en les offinant.

BIBLIOGRAPHIE

[1] *BENZECRI* *La taxinomie*

[2] *FISHER R.A.* *The use of multiple measurements in taxinomic problems. Contributions to mathematical statistics (J. WILEY)*

Figure 2

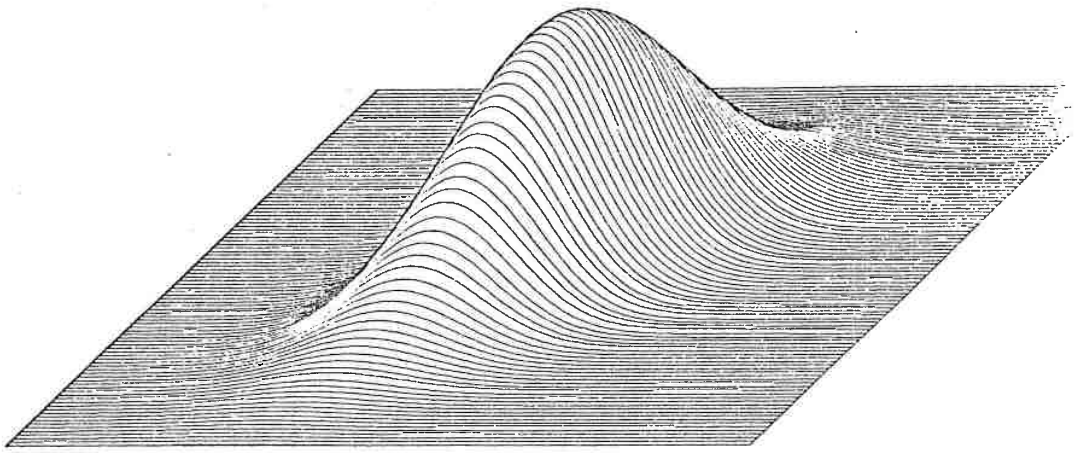
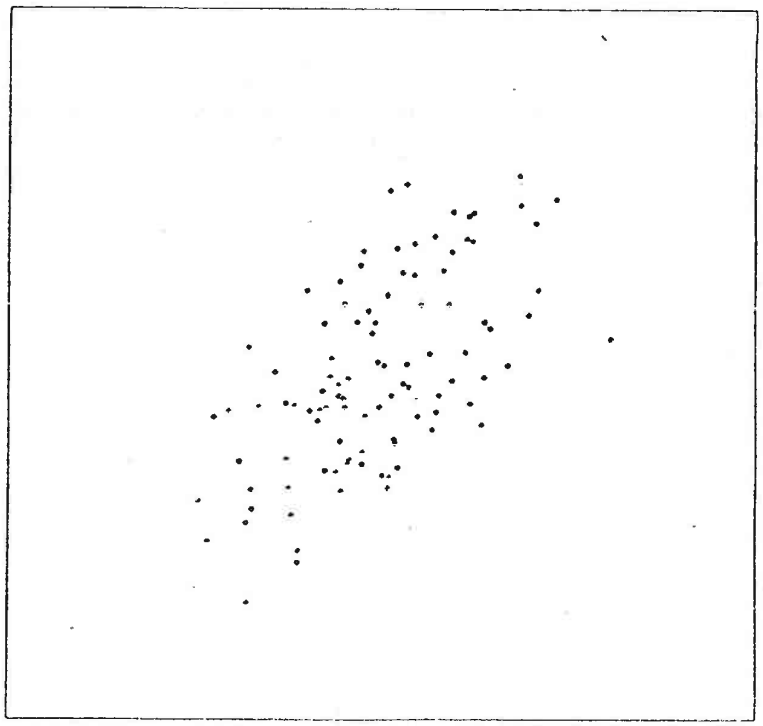
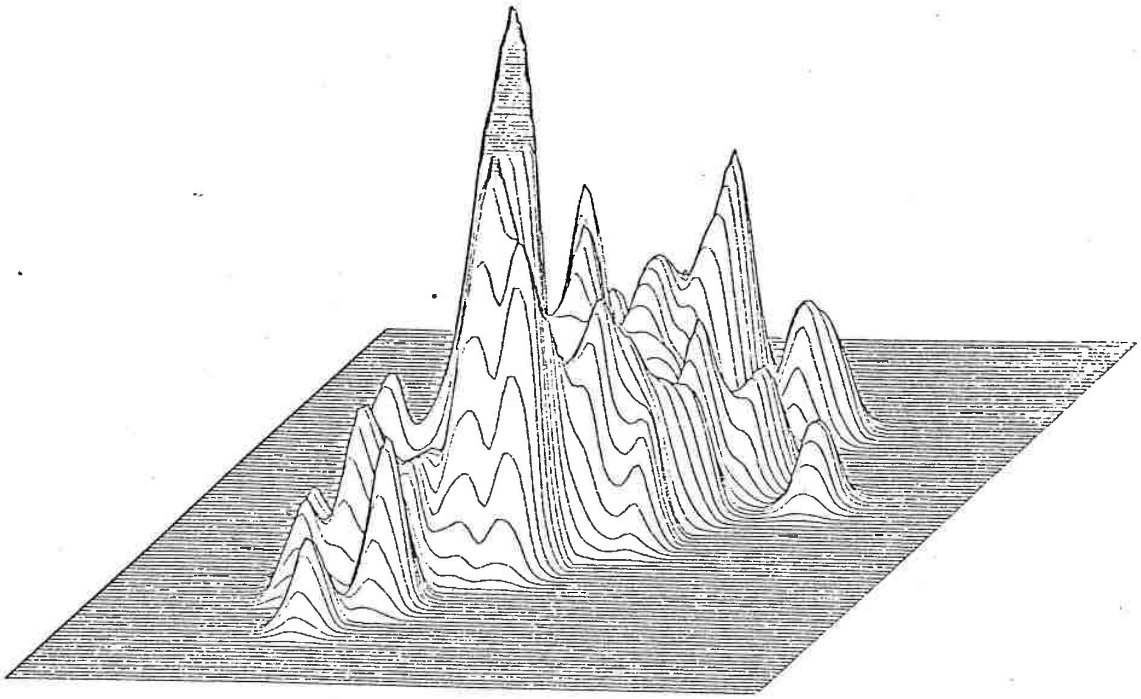


Figure 1

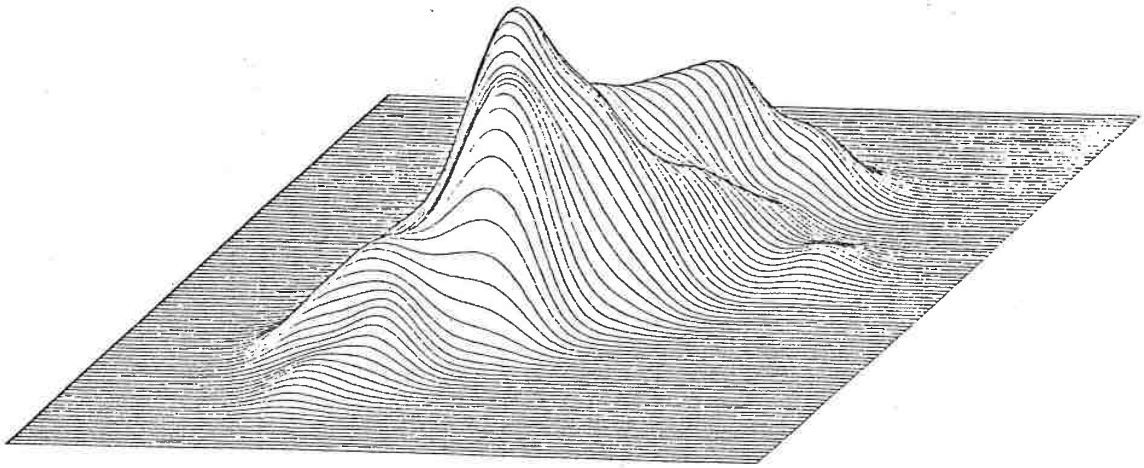


Figures 3

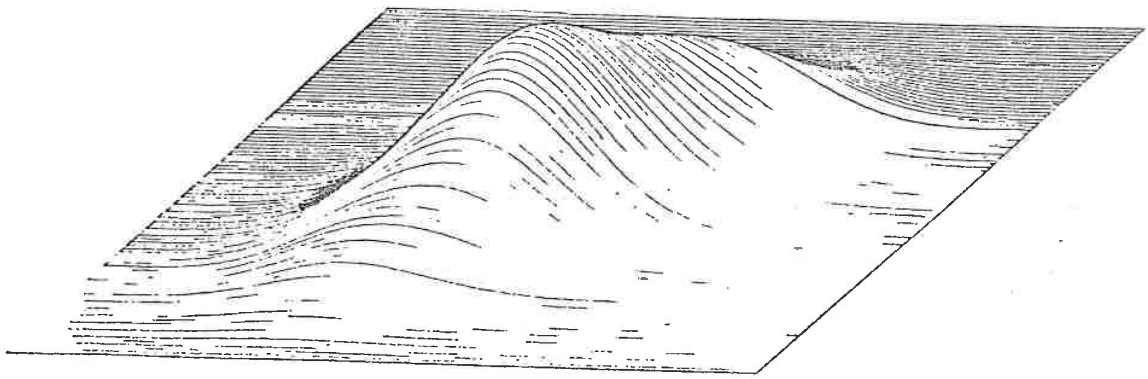
a)



b)

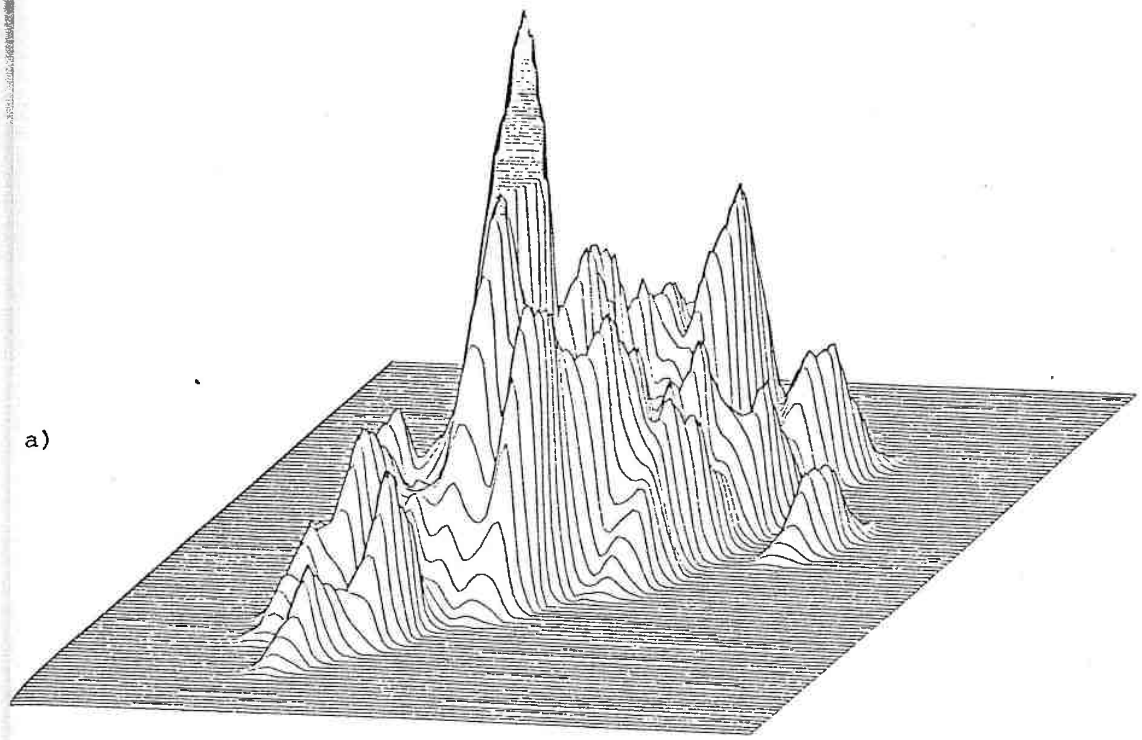


c)

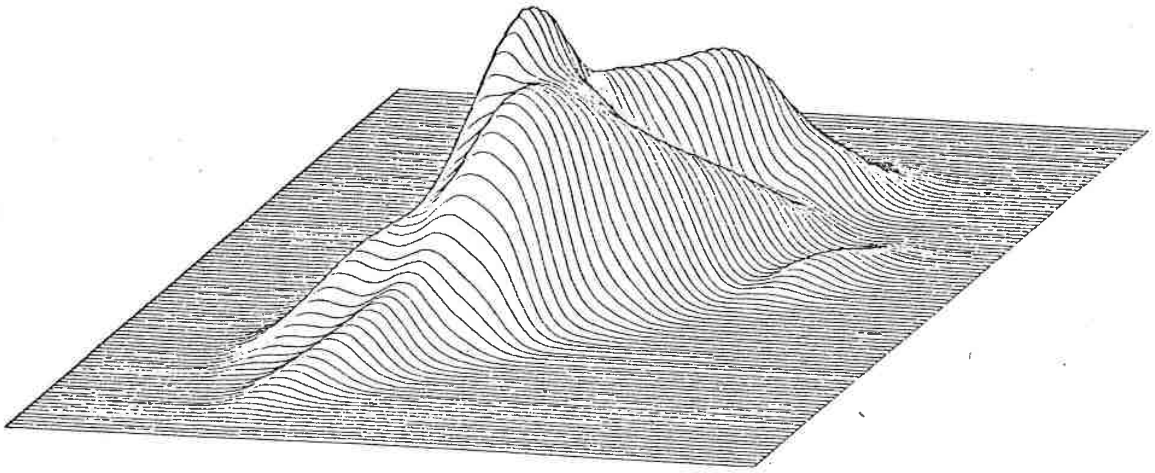


Figures 4

a)



b)



c)

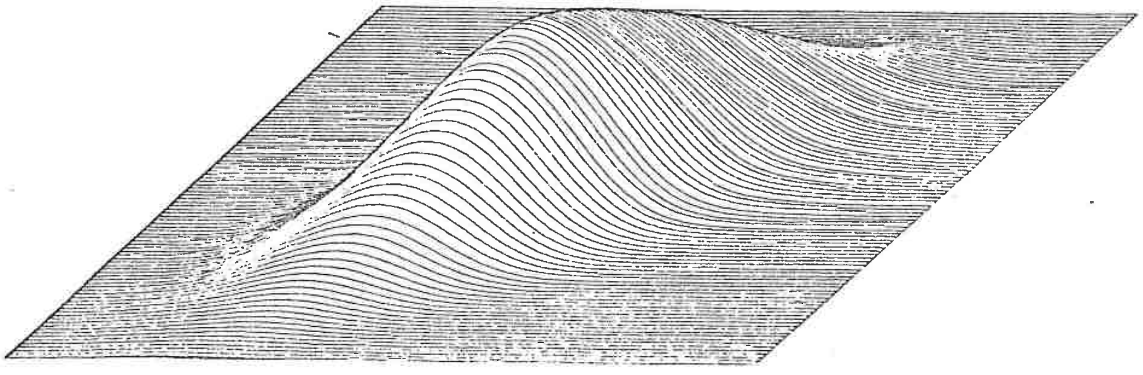


Figure 6

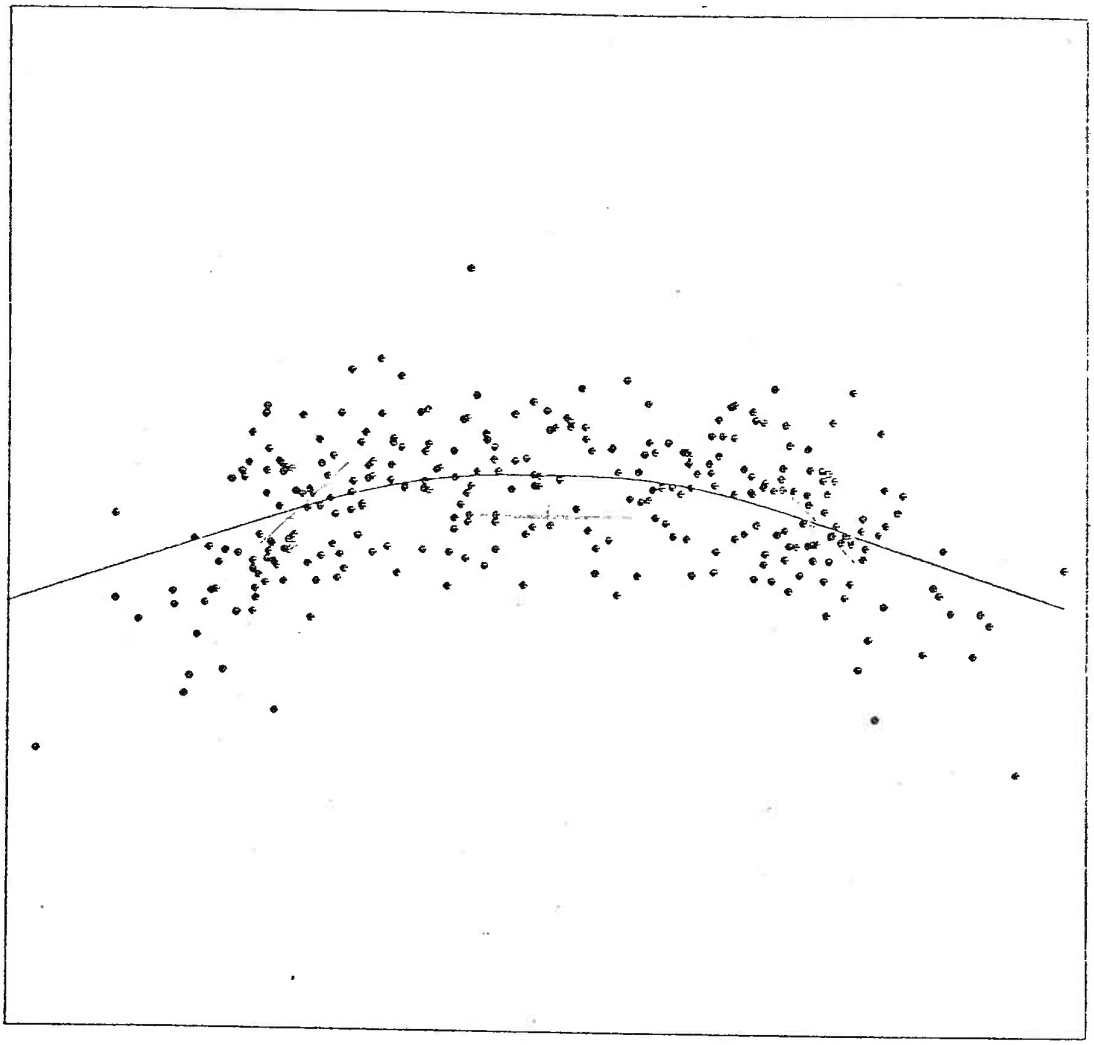


Figure 7a

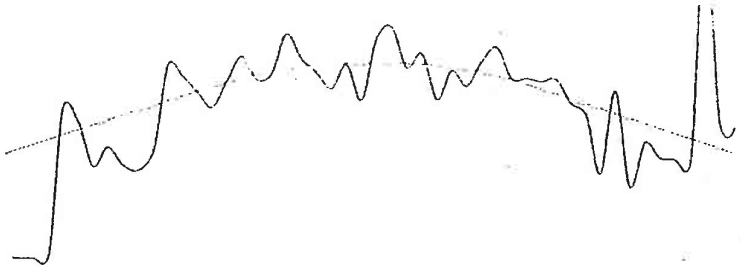


Figure 7b

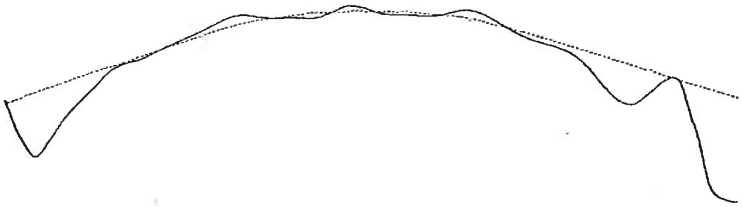


Figure 7c

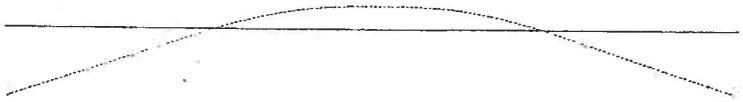


Figure 8

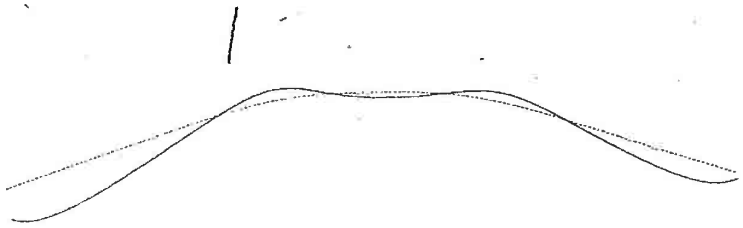


Figure 9

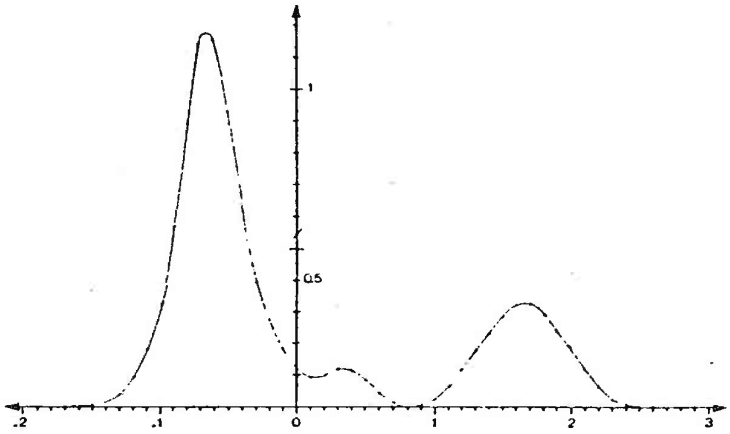


Figure 10

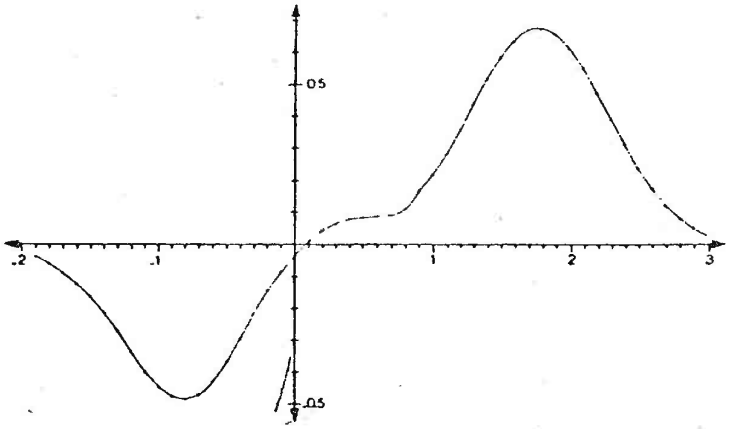


Tableau 3

Rivers	OUST.	AROA	EL-BARED	ABOU-ALI	IBRA.	EL-KELB	BEYR.	DEMOUR	EL-ASSIS	LITANI
SEPTEMBER	2.7	2.8	9.1	4.6	7.2	3.2	.6	1.6	37.5	12.1
OCTOBER	3.1	3.3	10.7	5.9	7.3	3.3	.8	2.2	37.2	14.9
NOVEMBER	4.8	4.9	15.3	9.1	10.	4.9	3.5	6.0	33.7	19.1
DECEMBER	5.3	5.4	17.4	21.1	14.8	15.0	9.7	18.3	34.7	32.6
JANUARY	7.9	8.0	25.5	27.8	38.8	31.7	20.9	33.0	35.0	85.5
FEBRUARY	7.6	7.8	24.2	47.9	46.5	41.0	28.3	38.0	33.9	89.1
MARCH	10.7	10.9	34.9	75.4	76.2	49.0	29.5	32.5	39.3	82.5
APRIL	10.6	10.7	36.2	83.4	112.4	42.3	29.3	20.2	40.2	56.9
MAY	9.6	9.7	34.9	80.1	93.4	27.0	6.0	7.8	42.7	36.8
JUNE	5.4	5.5	23.3	28.8	34.3	17.4	1.7	3.4	41.6	20.7
JULY	3.9	4.0	14.7	9.1	14.7	11.8	.9	2.6	42.4	13.9
AUGUST	3.1	3.3	10.7	5.8	9.2	4.2	.7	1.9	40.5	11.3

Tableau 4

Rivers	SEPT.	OCT.	NOV.	DEC.	JAN.	FEB.	MARCH	APRIL	MAY
OUSTOUWENE	-.252	-.224	-.098	-.072	.107	.073	.294	.289	.235
AROA	-.255	-.219	-.101	-.075	.105	.079	.299	.286	.233
EL-BARED	-.264	-.229	-.129	-.093	.073	.033	.260	.291	.285
ABOU-ALI	-.193	-.187	-.172	-.092	-.069	.080	.264	.325	.324
IBRAHIM	-.178	-.181	-.172	-.155	-.024	.019	.185	.415	.317
EL-KELB	-.220	-.223	-.212	-.083	.123	.244	.332	.243	.054
BEYROUTH	-.173	-.173	-.133	-.026	.161	.292	.296	.297	-.122
DEMOUR	-.186	-.179	-.125	.076	.299	.372	.264	.065	-.122
EL-ASSIS	-.052	-.073	-.322	-.252	-.225	-.297	.071	.128	.300
LITANI	-.185	-.167	-.147	-.055	.328	.342	.274	.086	-.043

Rivers	JUNE	JULY	AUG.
OUSTOUWENE	-.065	-.169	-.225
AROA	-.068	-.172	-.219
EL-BARED	.048	-.144	-.230
ABOU-ALI	-.035	-.173	-.189
IBRAHIM	-.031	-.143	-.170
EL-KELB	-.046	-.110	-.210
BEYROUTH	-.178	-.177	-.173
DEMOUR	-.175	-.175	-.181
EL-ASSIS	.235	.292	.158
LITANI	-.141	-.179	-.193

Tableau 5

Rivers	OUST.	AROA	EL-BARED	ABOU-ALI	IBRA.	EL-KELB	BEYROUTH	DEMOUR	EL-ASSIS	LITANI
SEPTEMBER	-.525	-.515	.093	-.342	-.091	-.477	-.728	-.631	2.883	.382
OCTOBER	-.561	-.541	.178	-.289	-.153	-.541	-.784	-.648	2.754	.586
NOVEMBER	-.709	-.697	.467	-.227	-.127	-.697	-.854	-.574	2.527	.892
DECEMBER	-1.274	-1.263	.003	.385	-.276	-.256	-.812	.091	1.813	1.593
JANUARY	-1.140	-1.135	.287	-.175	.358	.014	-.510	-.077	.174	2.622
FEBRUARY	-1.302	-1.293	.552	.518	.455	.206	-.367	.071	-.114	2.379
MARCH	-1.343	-1.335	.370	1.259	1.291	.197	-.587	-.466	-.193	1.545
APRIL	-1.092	-1.089	.260	1.272	2,214	-.062	-.485	-.780	-.131	.412
MAY	-.868	-.864	.003	1.560	2.017	-.269	-.992	-.930	.272	.069
JUNE	-.963	-.956	.383	.796	1.210	-.061	-.1242	-1.11	1.759	.187
JULY	-.695	-.686	.255	-.238	.255	.000	-.959	-.810	2.693	.185
AUGUST	-.540	-.522	.148	-.296	.012	-.441	-.758	-.649	2.845	.202

Figure 11

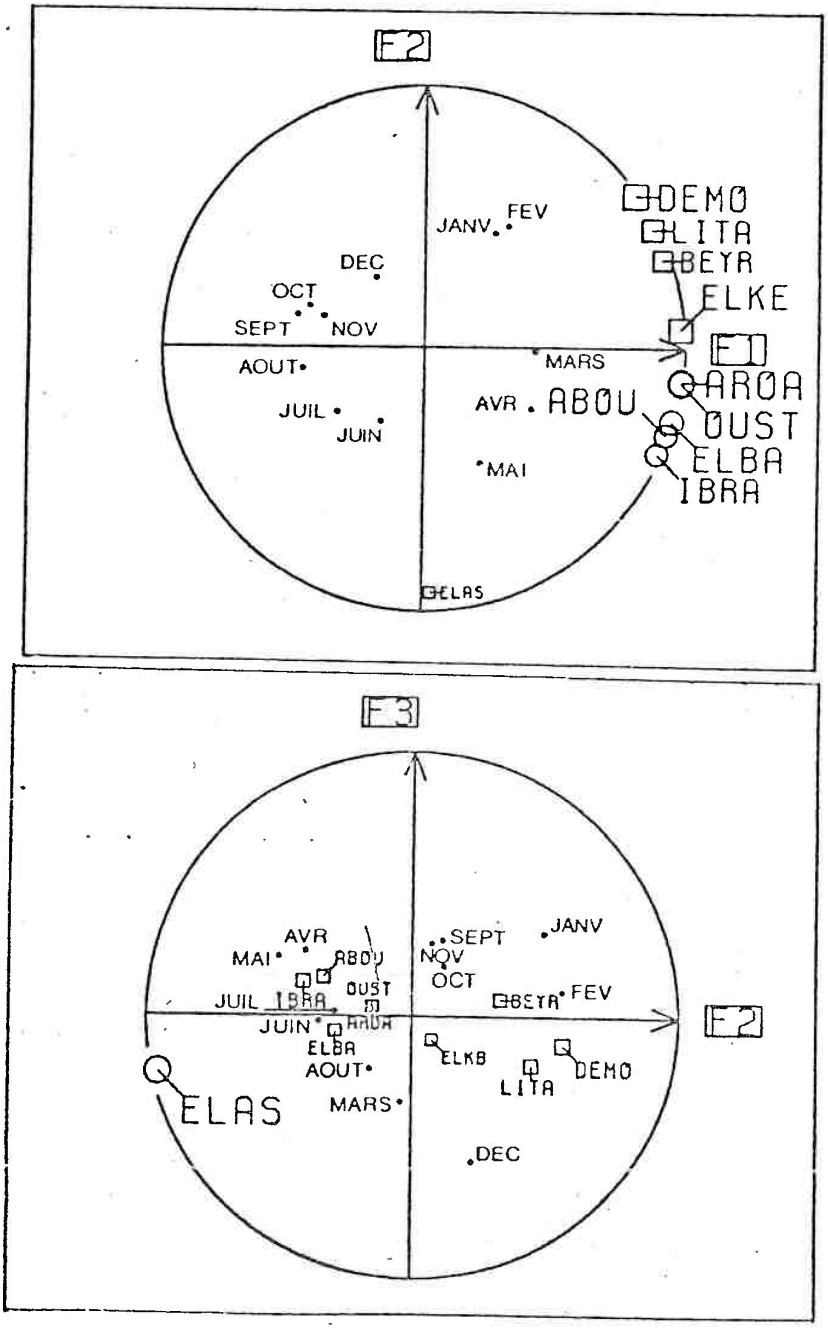

F2
JANV
FEV
DEMO
LITA
BETR
ELKE
DEC
OCT
SEPT
NOV
F1
MARS
AOUT
ARDA
AVR
ABOU
JUIL
OUST
JUIN
ELBA
MAI
IBRA
ELAS
F3
AVR
SEPT
JANV
MAI
ABOU
NOV
OCT
OUST
JUIL
IBRA
BETR
FEV
F2
JUIN
ARDA
ELBA
ELKB
DEMO
AOUT
LITA
ELAS
MARS
DEC

Figure 12

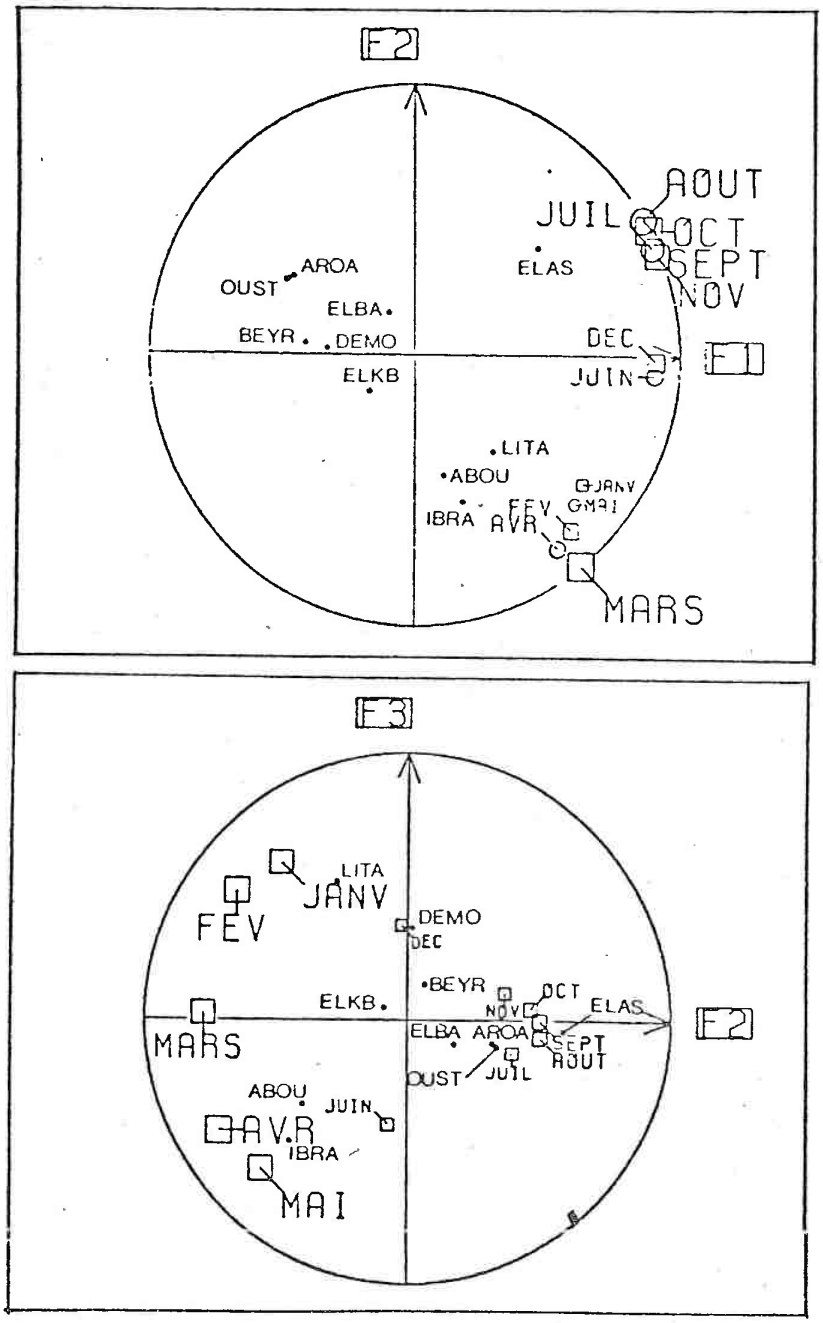

F2
AOUT
JUIL
OCT
SEPT
NOV
ELAS
AROA
OUST
ELBA
BEYR
DEMO
DEC
F1
ELKB
JUIN
LITA
ABOU
JANV
FEV
GMAI
IBRA
AVR
MARS
F3
LITA
JANV
FEV
DEMO
DEC
BEYR
OCT
NOV
ELAS
ELKB
F2
MARS
ELBA
AROA
SEPT
AOUT
OUST
JUIL
ABOU
JUIN
AVR
IBRA
MAI

Figure 13