Comparaison entre l'estimation paramétrique et l'estimation non-paramétrique à noyau pour une densité normale et pour une densité normale contaminée

Thèse présentée à la Faculté des Sciences
de l'Université de Neuchâtel
pour l'obtention du grade de docteur ès sciences

par

Etienne KAELIN

ETIENNE KAELIN

Comparaison entre l'estimation paramétrique et l'estimation non-paramétrique à noyau pour une densité normale et pour une densité normale contaminée

IMPRIMATUR POUR LA THÈSE

Comparaison entre l'estimation paramétrique et l'estimation non-paramétrique à noyau pour une densité normale et pour une densité normale contaminée

de Monsieur Etienne Kaelin

UNIVERSITÉ DE NEUCHATEL

FACULTÉ DES SCIENCES

La Faculté des sciences de l'Université de Neuchâtel, sur le rapport des membres du jury,

Madame M. Graf, MM. P. Banderet, M. Lejeune (Lausanne) et Th. Gasser (Mannheim)

autorise l'impression de la présente thèse.

Neuchâtel, le 24 juillet 1987

Le doyen:

François Sigrist

François Sigrist

Je remercie sincèrement,

Madame Dr.Monique Graf-Jaccottet, directeur de thèse, pour ses conseils avisés et sa disponibilité qui m'ont été d'une grande motivation.

Le professeur Michel Lejeune, de l'Université de Lausanne, qui est à l'origine de ce travail, pour ses encouragements et son intérêt bienveillant.

Les professeurs Pierre Banderet, de l'Université de Neuchâtel, et Théo Gasser, de l'Université de Mannheim, qui ont accepté d'être membres du jury, pour leurs conseils.

Comparaison entre l'estimation paramétrique et l'estimation non-paramétrique à noyau pour une densité normale et pour une densité normale contaminée

Résumé

En se référant aux articles bibliographiques de Wertz et Schneider (1979) et de Collomb (1981), on constate que la littérature abonde d'articles et plus récemment de quelques livres traitant de l'estimation non-paramétrique, et plus particulièrement de l'estimation de fonctions de densité. Cet estimateur s'adapte à la densité étudiée et est moins restrictif que l'estimateur paramétrique, gaussien notamment, utilisé fréquemment dans la pratique et dépendant de la connaissance précise de la répartition et de ses paramètres. Peu d'auteurs mentionnent l'estimation paramétrique d'une fonction de densité (Wertz, 1978, Tapia et al., 1978, Durbin, 1980).

Le but de ce travail est d'analyser et de comparer les propriétés de ces deux catégories d'estimateurs pour les familles **F** et **G**, où

F = { $f(\mu,\sigma^2;x)$ fonction de densité normale, d'espérance μ et de variance σ^2, en un point donné x, notée $\mathbf{N}(\mu,\sigma^2;x)$ }

où $\mathbf{N}(\mu,\sigma^2;x) = (2\Pi)^{-1/2}\ \sigma^{-1}\ \exp[-(x-\mu)^2/(2\sigma^2)]$;

G = { contamination d'une densité normale par une autre densité normale en un point donné x }

= { $(1-\varepsilon)\ f_1(x) + \varepsilon\ f_2(x)$; $f_1(x) = \mathbf{N}(\mu,\sigma^2;x)$,

$f_2(x) = \mathbf{N}(\alpha,\beta^2;x)$, $0 \leq \varepsilon \leq 1$ et x un point donné }.

Soient X_1, X_2, ..., X_n, n variables aléatoires indépendantes et identiquement distribuées, possédant une fonction de densité f, $f \in F$ ou G. L'estimateur non-paramétrique choisi est l'estimateur à noyau suggéré par Rosenblatt (1956) et généralisé par Parzen (1962), noté $f_n(x)$.

$$f_n(x) = (nh_n)^{-1} \sum_{i=1}^{n} K[(x-X_i)/h_n]$$

K est une fonction appelée noyau de f_n et h_n est la largeur ou fenêtre du noyau.

Les noyaux choisis sont le noyau biquadratique K_1 et le noyau normal K_2.

$$K_1(u) = \begin{cases} 15/16\ (1-u^2)^2 & \text{si } |u| \leq 1 \\ 0 & \text{sinon} \end{cases}$$

$$K_2(u) = N(0,1;u) .$$

L'estimateur paramétrique, noté g_n, est défini par la fonction de densité normale, où le(s) paramètre(s) inconnu(s) a(ont) été remplacé(s) par les estimateurs sans biais et de variance minimale, à savoir la moyenne arithmétique m et la variance empirique s^2. Ces estimateurs sont :

$g_{1n}(m,\sigma^2;x) = N(m,\sigma^2;x)$, μ inconnu ;

$g_{2n}(\mu,(n-1)s^2/n;x) = N(\mu,(n-1)s^2/n;x)$, σ^2 inconnu ;

$g_{3n}(m,s^2;x) = N(m,s^2;x)$, μ et σ^2 inconnus .

Le critère global de la qualité de l'estimateur f_n par rapport à la densité exacte f est l'erreur quadratique moyenne intégrée, abrégé EQMI, et le critère de comparaison de deux estimateurs f_n et g_n est l'efficacité;

celle-ci est notée eff ;

$$EQMI(f_n) = \int_{-\infty}^{+\infty} EQM[f_n(x)]\, dx$$

où $EQM[f_n(x)] = E_f\{[f_n(x) - f(x)]^2\}$ au point x fixé ;

$$eff(f_n, g_n) = \frac{EQMI(f_n)}{EQMI(g_n)} .$$

A. Résultats obtenus pour l'estimation d'une densité normale

Pour l'estimation non-paramétrique, Parzen (1962) détermine une approximation de l'$EQMI(f_n)$ à l'aide d'un développement limité, ainsi qu'une largeur de fenêtre minimisant cette EQMI. Pour le noyau biquadratique, on obtient :

$$EQMI(f_{1n}) = 0.32141\ \sigma^{-1}\ n^{-4/5} + o(n^{-4/5}) .$$

Pour le noyau normal, la formule asymptotique de l'EQMI vaut :

$$EQMI(f_{2n}) = 0.3329\ \sigma^{-1}\ n^{-4/5} + o(n^{-4/5}) .$$

Mais, il est possible de calculer explicitement l'EQMI de f_{2n} et de déterminer numériquement la valeur de k_n minimisant cette EQMI, k_n désignant la fenêtre réduite (Anderson, 1969 et Fryer, 1976).

$$EQMI(f_{2n}) = (4\Pi\sigma^2)^{-1/2}\ \{\ 1 + n^{-1}\ [\ k_n^{-1} - (1+k_n^2)^{-1/2}\ ]$$

$$+ (1+k_n^2)^{-1/2} - 2\ [2/(2+k_n^2)]^{1/2}\ ]$$

où $k_n = h_n/\sigma$

et $k_{Fryer} = 1.31\ n^{-0.205}$.

Pour l'estimateur paramétrique, l'EQMI théorique est connue uniquement pour le cas où seule l'espérance est inconnue; un estimateur minimisant l'EQMI a été trouvé par Wertz et Guttmann (1976) et par Klebanov (1977).

La détermination analytique de l'EQMI lorsque la variance est estimée forme les principaux résultats de cette partie. On trouve :

$$EQMI(g_{1n}) = (\Pi\sigma^2)^{-1/2} \{ 1 - [2n/(2n+1)]^{1/2} \} .$$

$$EQMI(g_{2n}) = (4\Pi\sigma^2)^{-1/2} \{ (n/2)^{1/2} \Gamma[(n-1)/2]/\Gamma(n/2) - [1 + 3/(8n) - 55/(256n^2)] \} + O(n^{-3}).$$

$$EQMI(g_{3n}) = (4\Pi\sigma^2)^{-1/2} \{ [(n-1)/2]^{1/2} \Gamma[(n-2)/2]/\Gamma[(n-1)/2] - [1 - 1/(8n) - 31/(256n^2)] \} + O(n^{-3}).$$

Les conclusions reposent sur l'analyse des EQMI des estimateurs f_{jn} et g_{jn}, de l'efficacité du noyau normal et des simulations menées parallèlement. Elles se résument ainsi :

1) La formule asymptotique est mauvaise pour les deux noyaux considérés, dans un rapport de 1.5; de plus, le choix du noyau n'est pas primordial.
2) Une fenêtre proche de celle proposée par Fryer n'altère que faiblement l'$EQMI(f_{2n})$.
3) L'$EQMI(f_{2n})$ est une fonction décroissante de n, de l'ordre de $n^{-1/5}$.
4) L'$EQMI(g_{jn})$ est une fonction décroissante de n, qui est approchée correctement par son développement selon les puissances de n^{-1}.
5) La connaissance d'un paramètre réduit de moitié l'EQMI de g_{1n} ou de g_{2n} par rapport à celle de g_{3n}.
6) En analysant l'efficacité du noyau normal, force est de constater qu'elle est toujours inférieure à 1.0, décroissante de l'ordre de $n^{-1/5}$, quelque soit l'estimateur paramétrique utilisé.

Ainsi, l'estimateur paramétrique est préférable à l'estimateur à noyau, f_{1n} ou f_{2n}, si la densité à estimer est normale.

B. Résultats obtenus pour l'estimation d'une densité normale contaminée par une autre densité normale.

Lorsque la densité normale à estimer est contaminée par une autre répartition normale, $f \in G$, on suppose que le statisticien ignore l'existence d'une quelconque contamination; ainsi, il choisit une fenêtre réduite k_n ne dépendant que de la taille de l'échantillon et estime l'espérance par m, comme dans la partie A.

Pour le noyau biquadratique, la formule asymptotique de l'EQMI, dépendante de k_n, est déterminée par un raisonnement analogue à celui de la première partie et permet ainsi d'évaluer l'EQMI(f_{1n}) par substitution de valeurs de k_n. On obtient :

$$EQMI(f_{1n}) = 7^{-1}\sigma^{-1}[\ 5/(nk_n) + 3/(224\sqrt{\pi})\ k_n^4 \cdot S] + o[(nk_n)^{-1}+k_n^4]$$

où $S = [\ (1-\varepsilon)^2 + \varepsilon^2/r^5 + 8\sqrt{2}/3\ (1-\varepsilon)\ \varepsilon\ R\ P(d,1+r^2;0)\ ]$,

$P(u,y;x) = (2\pi)^{1/2}\ N(u,y;x)$,

$R = (1+r^2)^{-2}\ [\ d^4/(1+r^2)^2 - 6\ d^2/(1+r^2) + 3\ ]$,

$d = (\alpha-\mu)/\sigma$, $r^2 = \beta^2/\sigma^2$ et $k_n = h_n/\sigma$.

Pour le noyau normal, l'EQMI asymptotique dépend des mêmes variables que celle du noyau biquadratique et vaut :

$$EQMI(f_{2n}) = (4\pi)^{-1/2}\ \sigma^{-1}\ [\ 1/(nk_n) + 3/16\ k_n^4 \cdot S\ ] + o[(nk_n)^{-1}+k_n^4]\ .$$

Mais, il est possible de calculer exactement l'EQMI(f_{2n}) (Anderson, 1969 et Fryer, 1976). La formule suivante est alors obtenue :

$$EQMI(f_{2n}) = (4\Pi)^{-1/2}\ \sigma^{-1}[\ (nk_n)^{-1}$$

$$+ (1-\varepsilon)^2\ [1 + (1-1/n)/(1+k_n^{\,2})^{1/2} - 2/(1+k_n^{\,2}/2)^{1/2}]$$

$$+ \varepsilon^2/r\ [1 + (1-1/n)/(1+k_n^{\,2}/r^2)^{1/2} - 2/(1+k_n^{\,2}/(2r^2))^{1/2}]$$

$$+ 2\sqrt{2}\ (1-\varepsilon)\ \varepsilon\ [P(d,1+r^2;0) + (1-1/n)\ P(d,1+r^2+2k_n^{\,2};0)$$

$$- 2\ P(d,1+r^2+k_n^{\,2};0)]\ ]\ .$$

L'estimateur paramétrique est défini en supposant que la variance σ^2 est connue, par :

$$g_n(\mathbf{m};x) = N(\mathbf{m},\sigma^2;x)\ .$$

Pour calculer l'EQMI(g_n), la fonction de densité de $\mathbf{m}$ a été déterminée exactement. On obtient :

$$EQMI(g_n) = (4\Pi)^{-1/2}\ \sigma^{-1}\ \{\ 1 + (1-\varepsilon)^2 + \varepsilon^2/r$$

$$+ 2\sqrt{2}\ [(1-\varepsilon)\ \varepsilon\ P(d,1+r^2;0)$$

$$- (1-\varepsilon) \sum_{j=1}^{n} \binom{n}{j} (1-\varepsilon)^j\ \varepsilon^{n-j}\ P(d(n-j)/n,2+a_j;0)$$

$$- \varepsilon \sum_{j=1}^{n} \binom{n}{j} (1-\varepsilon)^j\ \varepsilon^{n-j}\ P(d\cdot j/n,1+r^2+a_j;0)]\ \}$$

où $a_j = (j + (n-j)r^2)/\ n^2$

et $\binom{n}{j} = n!/[(n-j)!j!]$.

Pour analyser les différentes formules, on s'est restreint à deux sortes particulières de contamination, d'une part une contamination asymétrique avec $\sigma^2 = \beta^2$, et d'autre part une contamination symétrique, c-à-d $d = 0.0$.

Pour l'estimateur paramétrique g, les conclusions sont les suivantes :

1) L'EQMI est très sensible à toute contamination lorsque la différence réduite des espérances est moyenne ou élevée, $d > 1.0$, et/ou lorsque le rapport des variances est différent de 1.0.
2) Si la contamination est proche du mode principal, $d \leq 1.0$, l'EQMI(g) ne dépend alors que de la taille de l'échantillon et non du taux de contamination. On peut donc l'ignorer.

Pour l'EQMI des estimateurs non-paramétriques, f_1 et f_2, les remarques suivantes peuvent être formulées :

1) La formule asymptotique est grossière et ne donne qu'une approximation sommaire de l'erreur commise pour les tailles d'échantillons étudiées.
2) L'EQMI est une fonction pratiquement constante du taux de contamination et ne dépend que de la taille de l'échantillon, sauf lorsque la contamination accentue le mode, c-à-d $r^2 < 1.0$.
3) La fenêtre proposée par Fryer obtient globalement les meilleurs résultats; toutefois, une fenêtre plus étroite est préférable lorsque le mode est fortement prononcé.
4) Le choix du noyau n'est pas prédominant; le noyau biquadratique estime pourtant plus précisément les densités présentant des modes accentués.

En conclusion, l'estimateur non-paramétrique est préférable à l'estimateur paramétrique, car le premier fournit une densité parente de la densité exacte.

De plus, pour aider le statisticien dans ses prises de décision, les graphes des taux critiques de contamination en fonction de la taille de l'échantillon sont tracés pour différentes valeurs des paramètres d et r^2.

SYMBOLES ET NOTATIONS

$a_j = (j \cdot \mu + (n-j)\ \alpha)/\ n$.

$b_j = (j \cdot \sigma^2 + (n-j)\ \beta^2)/\ n^2$.

Biais(t_n) : biais de l'estimateur t_n.

$c_j(\varepsilon) = \binom{n}{j} (1-\varepsilon)^j\ \varepsilon^{n-j}$.

d : différence réduite des espérances dans le modèle contaminée;

$$d = (\alpha-\mu)/\sigma \ .$$

$E_f(X)$: espérance mathématique de la variable aléatoire réelle X.

eff(g_1, f_2) : efficacité de l'estimateur f_2 par rapport à l'estimateur g_1.

$$\mathrm{eff}(g_1, f_2) = \frac{\mathrm{EQMI}(f_2)}{\mathrm{EQMI}(g_1)} \ .$$

effsim(g_1, f_2) : efficacité moyenne simulée;

$$\mathrm{effsim}(g_1, f_2) = \frac{\mathrm{EQDM}(g_1)}{\mathrm{EQDM}(f_2)} \ .$$

EQD(f_2) : écart quadratique discret de l'estimateur f_2 sur un échantillon;

$$\mathrm{EQD}(f_2) = \sum_{j=1}^{300} [f_2(x_j) - f(x_j)]^2 \ .$$

EQDM(f_2) : moyenne de 50 écarts quadratiques discrets de l'estimateur f_2 ; estimation de l'EQMI(f_2).

EQDS(f_2) : écart-type de 50 écarts quadratiques discrets de l'estimateur f_2.

EQM(t_n) : erreur quadratique moyenne de l'estimateur t_n;

$$\mathrm{EQM}(t_n) = E_f\{[t_n - \tau]^2\} \ .$$

$EQM[f_2(x)]$: erreur quadratique moyenne de l'estimateur f_2 au point x;

$$EQM[f_2(x)] = E_f\{[f_2(x) - f(x)]^2\} \text{ , x un point fixé.}$$

$EQMI(f_2)$: erreur quadratique moyenne intégrée de l'estimateur f_2;

$$EQMI(f_2) = \int_{-\infty}^{+\infty} EQM[f_2(x)]\, dx \ .$$

$f(\tau;.)$: fonction de densité à estimer, de paramètre τ.

$F(.)$: fonction de répartition associée à $f(\tau;.)$.

f_{1n}, f_1 : estimateur non-paramétrique à noyau biquadratique de la fonction de densité f.

f_{2n}, f_2 : estimateur non-paramétrique à noyau normal de la fonction de densité f.

g_{jn}, g_j : estimateur paramétrique de la fonction de densité f; $j = 1,2,3$.

i.i.d : indépendant et identiquement distribué.

h_n, h : fenêtre du noyau.

k_n, k : fenêtre réduite du noyau; $k = h/\sigma$.

K_1 : noyau biquadratique.

K_2 : noyau normal.

Kum(u,a,b) : fonction de Kummer généralisée.

m_n, m : moyenne arithmétique de l'échantillon.

n : taille de l'échantillon aléatoire.

$N(\mu,\sigma^2;.)$: fonction de densité normale de paramètres μ et σ^2;

$$N(\mu,\sigma^2;.) = (2\Pi)^{-1/2}\ \sigma^{-1}\ \exp[-(x-\mu)^2/(2\sigma^2)] \ .$$

$\binom{n}{j} = n!/[(n-j)!j!]$.

$o(n^\alpha)$: $a_n = o(n^\alpha)$ si $\lim_{n\to\infty} (a_n/n^\alpha) = 0$.

$O(n^\alpha)$: $a_n = O(n^\alpha)$ si $\lim_{n\to\infty} (a_n/n^\alpha) =$ constante finie.

$\mathcal{P}$: ensemble des fonctions de densités à estimer.

$P(w,y;.) = (2\Pi)^{1/2}\ N(w,y;.)$, $y > 0$.

$Q^{(j)}(t)$: $j^{ième}$ dérivée de Q par rapport à la variable t.

r^2 : rapport des variances dans le modèle contaminée; $r^2 = ß^2/\sigma^2$.

$\mathbb{R}$: ensemble des nombres réels.

$R(f_2)$: rapport entre l'EQMI et l'EQDM de l'estimateur f_2 pour un ε fixé;

$$R(f_2) = \frac{EQMI(f_2) \text{ pour } \varepsilon_0}{EQDM(f_2) \text{ pour } \varepsilon_0} .$$

$Rap(f_2)$: rapport entre les EQMI de l'estimateur f_2 pour $\varepsilon = 0.5$ et pour $\varepsilon = 0.0$;

$$Rap(f_2) = \frac{EQMI(f_2) \text{ pour } \varepsilon = 0.5}{EQMI(f_2) \text{ pour } \varepsilon = 0.0} .$$

u_n, u : variance empirique lorsque l'espérance est connue.

v_n, v : variance empirique lorsque l'espérance est inconnue.

$Var_f(X)$: variance de la variable aléatoire réelle X.

X_i, X : variables aléatoires réelles de fonction de densité f.

α : espérance de la densité normale.

$ß^2$: variance de la densité normale.

$\Gamma(u)$: fonction gamma.

ε : taux de contamination, $0 \leq \varepsilon \leq 1$.

ε_1 : taux critique de contamination.

μ : espérance de la densité normale.

Π : constante pi = 3.14159265...

σ^2 : variance de la densité normale.

τ : paramètre de la fonction de densité f.

$\Phi_X(t)$: fonction caractéristique de la variable aléatoire X.

$\Phi(t)$: fonction caractéristique de la densité normale N.

$X_n^2(u)$: fonction de densité de la répartition du chi-carré à n degrés de liberté.

TABLE DES MATIERES

CHAPITRE 1 : GENERALITES

Introduction

Un des problèmes majeurs en statistique est celui de l'estimation, notamment de fonctionnelles et de paramètres associés à une loi de probabilité P à partir de données expérimentales. Par exemple : la fonction de répartition, la fonction de densité, le mode, les moments de la loi P. Ces estimations peuvent être ponctuelles, la moyenne arithmétique par exemple, ou données par un intervalle de confiance. Dans tous les cas, elles aident le statisticien à interpréter au mieux les résultats et à prendre les décisions utiles en tenant compte des hypothèses de travail.

Il semble difficile, voire impossible d'estimer quelque paramètre que ce soit sans une connaissance à priori de l'ensemble des lois de probabilité P, ensemble noté $\mathbb{P}$. Et pourtant, tout le monde, et les non-statisticiens en premier, utilise la moyenne arithmétique de n valeurs observées comme un estimateur de la tendance centrale des mesures effectuées, sans supposer d'hypothèse préalable sur la loi P.

Le problème traité ici est celui d'estimer la fonction de densité associée à la loi P. La fonction de densité n'est pas une construction purement théorique, mais décrit d'une manière complète la répartition de la variable étudiée : elle permet de calculer entre autre des paramètres tel que les moments, le(s) mode(s), et, plus concrètement, la probabilité de tout événement concernant cette variable. Son rôle est ainsi prédominant en statistique.

Aussi paradoxal que cela puisse paraître, il est possible d'estimer cette fonction sans aucune connaissance sur le modèle; c'est ce que le statisticien fait en visualisant les données sous forme d'un graphe particulier, appelé histogramme.

En revanche, si des hypothèses sont posées initialement sur l'ensemble $\mathbb{P}$, par exemple, la loi P est une répartition normale, alors la statistique classique fournit méthodes et résultats pour estimer paramètres et fonctionnelles découlant de P.

On constate que les deux approches énumérées sont différentes et méritent une définition précise qui est le but de ce paragraphe.

Définition I.1

Si la famille $\mathbb{P}$ est formée de fonctions connues et dépendant d'un nombre fini de paramètres inconnus, alors le problème d'estimation est appelé paramétrique et la famille $\mathbb{P}$ est dite paramétrisable.

Sinon, le problème est dit non-paramétrique.

Exemples :

$\mathbb{P}$ = { fonctions de densité possédant un premier moment }

Soit m_n la moyenne arithmétique : $m_n = n^{-1} \sum_{i=1}^{n} X_i$.

m_n est un estimateur du 1er moment, mais m_n est considéré dans ce problème comme un estimateur non-paramétrique, car la famille $\mathbb{P}$ ne peut être indexée par un nombre fini de paramètres.

Par contre, si

$\mathbb{P}$ = { fonctions de densité normales, $N(\mu, \sigma^2; x)$ } ,

alors m_n devient un estimateur paramétrique du 1er moment.

Ainsi, tout estimateur basé sur une connaissance à priori de l'ensemble $\mathbb{P}$ et de ses paramètres est appelé

paramétrique, tandis qu'un estimateur ne dépendant que des données initiales sans aucune contrainte sur la famille $\mathbb{P}$ est non-paramétrique.

Il est évident que certaines conditions initiales sont nécessaires pour pouvoir orienter la recherche et aussi pour calculer les caractéristiques du problème posé. Mais, plus elles sont précises, plus elles sont contraignantes et restreignent son champ d'applications. Aussi, il est important d'analyser les conséquences sur les résultats obtenus si l'on s'écarte des hypothèses formulées (robustesse). Sous cet angle, un estimateur non-paramétrique aura certainement tendance à être plus robuste qu'un estimateur paramétrique.

Dans le chapitre 2, l'estimation de la densité normale est examinée et comparée pour les deux classes d'estimateurs définis, paramétrique et non-paramétrique.

Dans le chapitre 3, le comportement de ces estimateurs est analysé pour une densité normale contaminée par une autre densité normale, puis leurs performances sont comparées.

Mais, il est nécessaire tout d'abord de définir les hypothèses, les critères de comparaison entre les deux méthodes, les estimateurs utilisés et les principaux résultats connus et nécessaires ultérieurement.

1.1 Hypothèses sur l'échantillon et sur la densité

(H1) Hypothèses sur l'échantillon

$X_1, X_2, \ldots, X_n$, n variables aléatoires réelles (v.a.r) indépendantes et identiquement distribuées (i.i.d), de fonction de densité f.
Soit X une variable aléatoire de densité f, notée $X \sim f$;

$E_f(X)$ et $Var_f(X)$ désignent respectivement l'espérance mathématique et la variance de la variable aléatoire X.

$$E_f(X) = \int_{-\infty}^{+\infty} x\, f(x)\, dx \ , \qquad (1)$$

$$Var_f(X) = E_f\{[X - E_f(X)]^2\} \ . \qquad (2)$$

Dans la suite, f désigne toujours la fonction de densité à estimer.

(H2) Hypothèses sur la densité f

Soit $\mathbb{P} = \{ f(\tau;x) \ ; \ \tau \in \mathbb{R}^k \ , \ 1 \leq k < \infty \ , \ x \in \mathbb{R} \}$

où f est une fonction de densité réelle,
τ est le vecteur des paramètres de la fonction f,
x un point fixé du support de f.

Remarque

x étant un point donné, il ne figure pas comme variable (en première position), mais en deuxième place, comme paramètre.

Suivant la nécessité, f vérifie certaines des hypothèses suivantes :

(H2.1) f est continue.

(H2.2) f est de carré intégrable, notée $f \in L_2$,

c-à-d $\int_{-\infty}^{+\infty} f(x)^2\, dx$ existe et est finie.

(H2.3) f est deux fois continûment différentiable et $f^{(2)}$ est bornée et de carré intégrable.

Les hypothèses (H2.2) et (H2.3) sont utilisées pour déterminer les formules asymptotiques des estimateurs non-paramétriques (cf paragraphe 1.5).

1.2. Définition d'un estimateur de densité et de ses caractéristiques

Définition 1.2.1

Un estimateur d'une fonction de densité f est une fonction mesurable sur les boréliens, f_n, telle que pour $X_1, X_2, \ldots, X_n$ et x fixés, f_n est une fonction de densité sur $\mathbb{R}$, c-à-d

$$f_n : \mathbb{R}^n \times \mathbb{R} \longrightarrow \mathbb{R} ,$$

$$(X_1, X_2, \ldots, X_n, x) \longmapsto f_n(X_1, X_2, \ldots, X_n; x) ,$$

et, f_n vérifie les conditions suivantes :

$f_n(x) \geq 0$ pour tout $x \in \mathbb{R}$,

$$\int_{-\infty}^{+\infty} f_n(x)\, dx = 1 .$$

Pour simplifier l'écriture, nous posons :

$$f_n(x) = f_n(X_1, X_2, \ldots, X_n; x) . \qquad (3)$$

La valeur moyenne théorique de l'estimateur f_n au point x est définie par son espérance mathématique $E_f[f_n(x)]$.

$$E_f[f_n(x)] = \int_{-\infty}^{+\infty} f_n(x_1, x_2, \ldots, x_n; x)\, f(x_1) f(x_2) \ldots f(x_n)\, dx_1\, dx_2 \ldots dx_n \qquad (4)$$

lorsque les X_i sont i.i.d. et de densité f.

Le biais au point x est interprété comme une erreur systématique de l'estimateur f_n due à l'estimation, par opposition à l'erreur aléatoire due aux fluctuations d'échantillonnage. On le définit par :

$$\text{Biais}[f_n(x)] = E_f[f_n(x)] - f(x) \quad . \tag{5}$$

La tendance à l'écart à la moyenne au point x est définie par la variance de f_n, notée $\text{Var}[f_n(x)]$.

$$\text{Var}_f[f_n(x)] = E_f\{[f_n(x)-E_f(f_n(x)]^2\} \quad . \tag{6}$$

Afin de pouvoir juger de la qualité d'un estimateur f_n, des critères doivent être définis. Ils peuvent dépendre du point x où la fonction f est estimée, ou mesurer globalement l'erreur commise. Les plus couramment utilisés dans la littérature sont l'erreur quadratique moyenne au point x, notée $\text{EQM}[f_n(x)]$, et l'erreur quadratique moyenne intégrée, $\text{EQMI}(f_n)$. Ils sont intéressants parce qu'ils se laissent décomposer facilement à l'aide du biais et de la variance de l'estimateur f_n et qu'ils se prêtent très bien aux calculs algébriques.

Définition 1.2.2

L'erreur quadratique moyenne de l'estimateur f_n au point x par rapport à la fonction f, ou plus succintement EQM de f_n au point x, notée $\text{EQM}[f_n(x)]$, est définie par :

$$\text{EQM}[f_n(x)] = E_f\{[f_n(x)-f(x)]^2\} \quad . \tag{7}$$

Par simple développement du carré, on montre que :

$$\text{EQM}[f_n(x)] = \text{Biais}^2[f_n(x)] + \text{Var}_f[f_n(x)] \tag{8}$$

$$= E_f[f_n{}^2(x)] + f^2(x) - 2\, f(x)\, E_f[f_n(x)] \quad . \tag{9}$$

Définition 1.2.3

L'erreur quadratique moyenne intégrée de l'estimateur f_n par rapport à la fonction f, ou plus simplement, EQMI de f_n, notée $EQMI(f_n)$, est définie par :

$$EQMI(f_n) = \int_{-\infty}^{+\infty} EQM[f_n(x)]\, dx \ . \qquad (10)$$

Pourtant, les critères de l'erreur absolue moyenne, abrégée EAM, et de l'erreur absolue moyenne intégrée, EAMI, sont plus intuitifs, mais aussi plus difficiles à traiter théoriquement :

$$EAM[f_n(x)] = E_f[\ |f_n(x)-f(x)|\ ] \ ,$$

$$EAMI(f_n) = \int_{-\infty}^{+\infty} EAM[f_n(x)]\, dx \ .$$

Le critère de l'EAMI est facilement interprétable; il est lié à l'erreur maximale commise si la probabilité d'un borélien est estimée à l'aide de f_n. Scheffé (1947) a montré que :

$$\int_B |f_n - f| = 2 \sup_{B \in \mathcal{B}} \left| \int f_n - \int f \right| \ .$$

On se référera au livre de L.Devroye et L.Györfi (1985) qui traite de l'estimation non-paramétrique de densité selon la norme L_1.

M.Lejeune (1982) a montré , sur la base de simulations, que les critères de l'EAMI et de l'$EQMI^{1/2}$ varient de manière identique, les extrema de l'un coïncidant avec les extrema de l'autre. Ceci est une raison supplémentaire qui nous a incité au choix du critère de l'EQMI.

1.3. Définition de l'efficacité entre deux estimateurs

Définissons un critère global qui permette la comparaison entre deux estimateurs donnés f_n et g_n de la fonction f. La définition naturelle est le quotient du critère de l'EQMI des deux estimateurs, appelé efficacité et notée $\mathrm{eff}(f_n, g_n)$.

Définition 1.3.1

$$\mathrm{eff}(f_n, g_n) = \frac{\mathrm{EQMI}(f_n)}{\mathrm{EQMI}(g_n)} . \qquad (11)$$

Définition 1.3.2

L'estimateur f_n est dit meilleur que l'estimateur g_n, pour la fonction f, si

$\mathrm{eff}(f_n, g_n) < 1$ pour tout n.

Définition 1.3.3

L'estimateur f_n est dit asymptotiquement meilleur que l'estimateur g_n, si

$$\lim_{n \to \infty} \mathrm{eff}(f_n, g_n) < 1 .$$

Les principaux critères ayant été déterminés, nous allons donner les définitions exactes des estimateurs utilisés ainsi que les principaux résultats connus.

1.4 Définition des estimateurs utilisés

Supposons que la famille $\mathbb{P}$ est paramétrisable et soit τ le vecteur de $\mathbb{R}^k$ des paramètres de la densité, $k < \infty$. Soit t_n un estimateur du paramètre τ, c-à-d $t_n : \mathbb{R}^n \longrightarrow \mathbb{R}$ tel que $t_n(X_1, X_2, \ldots, X_n)$ est mesurable.

Si $x_1, x_2, \ldots, x_n$ est une réalisation de $X_1, X_2, \ldots, X_n$, alors $t_n(x_1, x_2, \ldots, x_n)$ est une estimation de τ.

Définition 1.4.1

L'estimateur paramétrique g_n au point x de la fonction f est égal à la valeur de $f(\tau;x)$ où le paramètre τ est remplacé par l'estimateur $t_n(X_1, X_2, \ldots, X_n)$. Il est donc de la forme suivante :

$$g_n(X_1, X_2, \ldots, X_n; x) = f[t_n(X_1, X_2, \ldots, X_n); x] \quad . \tag{12}$$

Pour simplifier l'écriture, nous noterons :

$$g_n(x) = g_n(X_1, X_2, \ldots, X_n; x) \quad . \tag{13}$$

Or $t_n(X_1, X_2, \ldots, X_n)$ peut être considéré comme une variable aléatoire. Alors,

$$E_f[g_n(x)] = E_{t_n}[g_n(x)] \tag{14}$$

où $E_{t_n}[g_n(x)]$ désigne l'espérance par rapport à la fonction de répartition de t_n. La même remarque est applicable à la variance.

$$Var_f[g_n(x)] = Var_{t_n}[g_n(x)] \tag{15}$$

On constate dès lors les limites d'un tel estimateur; en effet, la connaissance de f n'implique pas nécessairement la connaissance explicite de la densité de t_n. Et si elle est déterminée, les calculs engendrés par

$E_{t_n}[g_n(x)]$ et/ou par $Var_{t_n}[g_n(x)]$ sont généralement très compliqués, voire insolubles théoriquement. L'illustration de ce fait est donnée, par exemple, aux théorèmes 2.1.2 et 2.1.3.

Seuls des résulats avec des hypothèses très fortes sont connus. On pourrait supposer qu'un choix optimal de l'estimateur du paramètre τ nous conduit à un estimateur de f lui aussi optimal. Or cela n'est déjà pas vérifié pour la densité normale (Wertz, 1978 et paragraphe 2.1, l'estimateur g_{12} , formule (2.17)).

Définition 1.4.2

L'estimateur non-paramétrique proposé est construit à partir de l'histogramme mobile de Rosenblatt (1956), dont nous donnons un bref aperçu .

Supposons que l'on désire estimer la fonction de répartition $F(x)$, $F(x) = P(X < x)$, à partir d'un échantillon aléatoire de taille n. Alors, un estimateur logique et naturel de $F(x)$ est $F_n(x)$, défini par :

$$F_n(x) = \frac{\text{card}\ \{ X < x \}}{n} = n^{-1} \sum_{i=1}^{n} I\{X_i < x\}$$

où $I\{X < x\}$ est la fonction indicatrice de l'ensemble $\{X < x\}$.

Si F est absolument différentiable, alors f est la dérivée de la fonction F. Mais F_n n'est pas continue, ni dérivable en tout point. Comme estimateur de la densité f, posons alors :

$$f_n(X_1, X_2, \ldots, X_n; x) = \frac{F_n(x+h_n/2) - F_n(x-h_n/2)}{h_n}$$

$$= (nh_n)^{-1} \sum_{i=1}^{n} I\{X_i - h_n/2 < x \leq X_i + h_n/2\} .$$

$f_n(x)$ est appelé histogramme mobile; pour estimer la fonction de densité au point x, on dénombre le nombre des X_i contenus dans l'intervalle $]x-h_n/2, x+h_n/2]$, tandis que l'histogramme classique dépend non seulement de h_n, mais aussi du point initial x_0. Le critère de l'EQMI est de l'ordre $n^{-4/5}$ pour f_n contre $n^{-2/3}$ pour l'histogramme classique, ce qui démontre que f_n est un meilleur estimateur que l'histogramme classique au sens de l'EQMI.

Pour cet estimateur, f_n, il est possible de calculer son biais, son EQM et son EQMI. Comme nous analysons une généralisation de cet estimateur, nous renvoyons le lecteur à la littérature pour de plus amples détails (Rosenblatt, 1956, Parzen, 1962, Tapia et al., 1978).

Afin que l'estimateur f_n ait des propriétés mathématiques plus séduisantes (continuité, dérivablité), la fonction indicatrice a été remplacée par des fonctions plus régulières, c-à-d continues, dérivables, intégrables (hypothèses (H3) ci-dessous).

La définition générale de f_n est la suivante :

$$f_n(X_1, X_2, \ldots, X_n; x) = (nh_n)^{-1} \sum_{i=1}^{n} K[(x-X_i)/h_n] \quad . \tag{16}$$

Pour simplifier l'écriture, nous noterons :

$$f_n(x) = (nh_n)^{-1} \sum_{i=1}^{n} K[(x-X_i)/h_n] \quad . \tag{17}$$

Terminologie

La fonction K est le noyau de l'estimateur.
h_n est appelé largeur ou fenêtre du noyau K, tandis que f_n est dit estimateur à noyau, ou encore estimateur non-paramétrique à noyau, car sa définition est indépendante de la famille $\mathbb{P}$.

La variable unique et donc principale de cet estimateur est la largeur du noyau. Son rôle est de lisser plus ou moins fortement l'estimation obtenue et son choix est primordial, bien qu'arbitraire (cf paragraphes 2.3.2.2 et 3.3.2.3).

Même si $\mathbb{P}$ est supposé connu dans ce travail, cet estimateur sera tout de même appelé non-paramétrique, la connaissance de $\mathbb{P}$ étant nécessaire pour la détermination de l'erreur commise, l'EQMI(f_n).

(H3) Hypothèses sur le noyau K

(H3.1) $K(y) \geq 0$ pour tout $y \in \mathbb{R}$.

(H3.2) $\int_{-\infty}^{+\infty} K(y)\, dy = 1$.

(H3.3) $\sup_{-\infty < y < \infty} | K(y) | < \infty$.

(H3.4) $\lim_{y \to \infty} | y \cdot K(y) | = 0$.

(H3.5) $K(y) = K(-y)$.

(H3.6) $\int_{-\infty}^{+\infty} y^2 K(y)\, dy \neq 0$ et est finie.

Les conditions (H3.1) - (H3.5) sont nécessaires pour démontrer la consistance de l'estimateur f_n alors que l'hypothèse (H3.6) est utile pour déterminer les formules asymptotiques de l'EQM[$f_n(x)$] et de l'EQMI(f_n) (cf paragraphe 1.5).
Les conditions (H3.1) et (H3.2) garantissent que f_n est une fonction de densité; les hypothèses (H3.1) - (H3.3) impliquent que le noyau K est de carré intégrable.

L'hypothèse (H3.4) permet de donner moins de poids aux observations éloignées du point x.
La condition (H3.5) est nécessaire afin que les observations se situant de part et d'autre et à égale distance de zéro aient un poids identique.

Le choix du noyau K est très vaste. Mais le comportement de l'estimateur correspondant est peu à pas influencé par le noyau choisi (cf paragraphes 2.4 et 3.5).
Voici quelques exemples de noyaux fréquemment utilisés dans la pratique :

$$K(y) = \begin{cases} 0.5 & \text{si } |y| \leq 1 \\ 0 & \text{sinon} \end{cases} \qquad \text{noyau de Rosenblatt}$$

$$K(y) = \begin{cases} 1 - |y| & \text{si } |y| \leq 1 \\ 0 & \text{sinon} \end{cases} \qquad \text{noyau triangulaire}$$

$$K(y) = \begin{cases} 15/16\ (1 - y^2)^2 & \text{si } |y| \leq 1 \\ 0 & \text{sinon} \end{cases} \qquad \text{noyau biquadratique}$$

$$K(y) = (2\Pi)^{-1/2} \exp[-y^2/2] \qquad \text{noyau normal}$$

$$K(y) = \begin{cases} 3/4\ (1 - y^2) & \text{si } |y| \leq 1 \\ 0 & \text{sinon} \end{cases} \qquad \text{noyau d'Epanechnikov}$$

(H4) Hypothèses sur la fenêtre h_n

(H4.1) h_n est une suite de nombres positifs.

(H4.2) $\lim_{n\to\infty} h_n = 0$.

(H4.3) $\lim_{n\to\infty} (n \cdot h_n) = \infty$.

Ces hypothèses sont utilisées pour démontrer la consistance de f_n et déterminer les formules asymptotiques de l'EQM$[f_n(x)]$ et de l'EQMI(f_n) (cf paragraphe 1.5).

1.5 Résultats généraux pour les estimateurs à noyau

De nombreux livres et articles traitent des estimateurs à noyau, démontrant la consistance, calculant les critères cités, les meilleurs choix possibles pour la largeur du noyau, appliquant ce genre d'estimateurs à d'autres estimations que la fonction de densité, par exemple pour la régression, pour le taux de défaillance, pour les densités multivariées. On se référera aux bibliographies abondantes se trouvant dans les ouvrages de B.Prakasa Rao (1983) et de L.Devroye et L.Györfi (1985).

Quelques théorèmes importants sont cités et démontrés, afin de pouvoir calculer par la suite aisément les différents critères énoncés pour la densité normale et la densité normale contaminée. Ces théorèmes se trouvent notamment dans le livre de B.Prakasa Rao (1983) et ont été énoncés et démontrés par Rosenblatt (1956) et Parzen (1962).

Les démonstrations de ces théorèmes sont basées essentiellement sur le théorème (1.5.1) dû à Bochner (1955), énoncé pour des noyaux plus généraux que ceux définis au paragraphe précédent.

Théorème 1.5.1 (Parzen, 1962)

Soit K un noyau absolument intégrable, noté $K \in L_1$, satisfaisant les hypothèses (H3.3) et (H3.4) et h_n une suite positive vérifiant les hypothèses (H4.1) et (H4.2).
Pour une fonction q absolument intégrable, c-à-d $q \in L_1$, posons :

$$q_n(x) = h_n^{-1} \int_{-\infty}^{+\infty} K(y/h_n)\, q(x-y)\, dy \ . \qquad (18)$$

Si q est continue au point x, alors

$$\lim_{n\to\infty} q_n(x) = q(x) \int_{-\infty}^{+\infty} K(y)\, dy \ . \qquad (19)$$

Démonstration

Soit $\delta > 0$; alors,

$$\left| q_n(x) - q(x) \int_{-\infty}^{+\infty} K(y)dy \right| = \left| h_n^{-1} \int_{-\infty}^{+\infty} K(y/h_n)[q(x-y)-q(x)]dy \right|$$

$$\leq \sup_{|y|\leq\delta} |q(x-y)-q(x)| \int_{|z|\leq\delta/h_n} |K(z)|\, dz$$

$$+ \int_{|y|>\delta} [|q(x-y)|/y]\, y/h_n\, K(y/h_n)\, dy$$

$$+ |q(x)| \int_{|z|>\delta/h_n} |K(z)|\, dz$$

$$\leq \sup_{|y|\leq\delta} |\, q(x-y)-q(x)\,| \int_{-\infty}^{+\infty} |\, K(z)\,|\, dz$$

$$+\ \delta^{-1} \sup_{|z|>\delta/h_n} |\, zK(z)\,| \int_{-\infty}^{+\infty} |\, q(y)\,|\, dy$$

$$+\ |\, q(x)\,| \int_{|z|>\delta/h_n} |\, K(z)\,|\, dz$$

Lorsque n tend vers l'infini, chaque terme de cette dernière expression converge vers zéro :
pour le premier terme, q est continue en x et $K \in L_1$;
pour le second terme, la limite de zK(z) est zéro selon (H3.4) et $q \in L_1$;
pour le dernier terme, h_n tend vers zéro et $K \in L_1$.

Théorème 1.5.2 (Parzen, 1962)

Supposons que la fonction de densité f vérifie l'hypothèse (H2.1), que le noyau K vérifie les hypothèses (H3.1) - (H3.4) et que la fenêtre h_n vérifie les hypothèses (H4.1) et (H4.2).
Alors, pour un point x donné,

$$\lim_{n\to\infty} E_f[f_n(x)] = f(x) \qquad (20)$$

et

$$E_f[f_n(x)] = f(x) + o(1)\ . \qquad (21)$$

Si de plus, h_n vérifie l'hypothèse (H4.3), alors

$$\lim_{n\to\infty} Var_f[f_n(x)] = 0 \qquad (22)$$

et

$$Var_f[f_n(x)] = (nh_n)^{-1}\ f(x) \int_{-\infty}^{+\infty} K^2(y)\ dy + o[(nh_n)^{-1}]. \qquad (23)$$

Démonstration

Puisque f est une fonction de densité continue et que K est lui-même une densité, les conditions nécessaires pour appliquer le théorème 1.5.1 sont vérifiées.

$$E_f[f_n(x)] = E_f\{(nh_n)^{-1} \sum_{i=1}^{n} K[(x-X_i)/h_n]\}$$

$$= E_f\{h_n^{-1} K[(x-X_i)/h_n]\}$$

$$= h_n^{-1} \int_{-\infty}^{+\infty} K[(x-y)/h_n] f(y) dy$$

En appliquant le théorème 1.5.1 et l'hypothèse (H3.2), on obtient :

$$\lim_{n\to\infty} E_f[f_n(x)] = f(x)$$

et aussi

$$E_f[f_n(x)] = f(x) + o(1) .$$

Ainsi, f_n est un estimateur asymptotiquement sans biais.

$$Var_f[f_n(x)] = Var_f\{(nh_n)^{-1} \sum_{i=1}^{n} K[(x-X_i)/h_n]\}$$

$$= n^{-1} Var_f\{h_n^{-1} K[(x-X_i)/h_n]\}$$

$$= n^{-1} \{ E_f\{[h_n^{-1} K[(x-y)/h_n]]^2\} - E_f^2\{h_n^{-1} K[(x-y)/h_n]\} \}$$

$$\leq (nh_n)^{-1} \{h_n^{-1} \int_{-\infty}^{+\infty} K^2[(x-y)/h_n] f(y) dy\}$$

La fonction K^2 vérifie les hypothèses du théorème 1.5.1.

Ainsi,

$$\lim_{n\to\infty} \{h_n^{-1} \int_{-\infty}^{+\infty} K^2[(x-y)/h_n]\, f(y)\, dy\} = f(x) \int_{-\infty}^{+\infty} K^2(y)\, dy$$

qui existe et est finie selon les hypothèses (H3.2) et (H3.3).

Puisque nh_n tend vers l'infini lorsque n tend vers l'infini, selon (H4.3), on a donc :

$$\lim_{n\to\infty} \mathrm{Var}_f[f_n(x)] = 0$$

et aussi

$$\mathrm{Var}_f[f_n(x)] = (nh_n)^{-1}\, f(x) \int_{-\infty}^{+\infty} K^2(y)\, dy + o[(nh_n)^{-1}]. \quad (24)$$

Démonstration de (24)

$$\lim_{n\to\infty} [\mathrm{Var}_f[f_n(x)] - (nh_n)^{-1}\, f(x) \int_{-\infty}^{+\infty} K^2(y)\, dy\,]/(nh_n)^{-1}$$

$$\leq \{(nh_n)^{-1} [h_n^{-1} \int_{-\infty}^{+\infty} K^2[(x-y)/h_n]\, f(y)\, dy]$$

$$- (nh_n)^{-1}\, f(x) \int_{-\infty}^{+\infty} K^2(y)\, dy\,]/(nh_n)^{-1}$$

$$= \{h_n^{-1} \int_{-\infty}^{+\infty} K^2[(x-y)/h_n]\, f(y)\, dy\} - f(x) \int_{-\infty}^{+\infty} K^2(y)\, dy$$

Cette dernière expression tend vers zéro lorsque n tend vers l'infini en appliquant le théorème 1.5.1 à K^2.

Corollaire 1.5.3

Supposons que la fonction de densité f vérifie l'hypothèse (H2.1), que le noyau K vérifie les hypothèses (H3.1) - (H3.4) et que la fenêtre h_n vérifie les hypothèses (H4.1) - (H4.3).
Alors, l'estimateur non-paramétrique f_n est un estimateur consistant en erreur quadratique, c-à-d,
pour tout point x donné,

$$\lim_{n\to\infty} EQM[f_n(x)] = 0 \ . \qquad (25)$$

Ce corollaire est une conséquence immédiate du théorème 1.5.2 et de la définition de l'$EQM[f_n(x)]$.

Déterminons les formules asymptotiques pour l'$EQM[f_n(x)]$ et l'$EQMI(f_n)$. Elles sont obtenues à l'aide d'un développement limité de Taylor de la fonction à estimer f, ce qui nécessite des hypothèses supplémentaires pour f et pour K.

Théorème 1.5.4 (Rosenblatt, 1971)

Supposons que la fonction f, le noyau K et la fenêtre h_n vérifient respectivement les hypothèses (H2.1) - (H2.3), (H3.1) - (H3.6) et (H4.1) - (H4.3).
Alors, pour tout point x fixé,

$$EQM[f_n(x)] = (nh_n)^{-1} f(x) A(K)$$
$$+ h_n^4/4 \ | f^{(2)}(x)|^2 \ B^2(K) + o[(nh_n)^{-1}+h_n^4] \ . \qquad (26)$$

De plus,

$$EQMI(f_n) = (nh_n)^{-1} A(K)$$
$$+ h_n^4/4 \ B^2(K) \int_{-\infty}^{+\infty} | f^{(2)}(x)|^2 dx + o[(nh_n)^{-1}+h_n^4] \qquad (27)$$

où

$$A(K) = \int_{-\infty}^{+\infty} K^2(y)\, dy \qquad (28)$$

et

$$B(K) = \int_{-\infty}^{+\infty} y^2 K(y)\, dy\ . \qquad (29)$$

Démonstration

Calculons l'EQM[$f_n(x)$] en appliquant la formule (1.8). Grâce à la symétrie de K et au développement limité de Taylor de f au point x, les égalités suivantes se déduisent aisément.

$$\text{Biais}[f_n(x)] = E_f[f_n(x)] - f(x)$$

$$= \int_{-\infty}^{+\infty} [f(x-uh_n)-f(x)]\, K(u)\, du$$

$$= \int_{-\infty}^{+\infty} [f(x+vh_n)-f(x)]\, K(v)\, dv$$

$$= \int_{-\infty}^{+\infty} h_n\, f'(x)\, v\, K(v)\, dv$$

$$+ \int_{-\infty}^{+\infty} h_n^2 \int_0^1 (1-t)\, f^{(2)}(x+vh_n t)\, dt\, v^2 K(v) dv \qquad (30)$$

La première intégrale est nulle car K est une fonction paire, selon (H3.5).

Pour le second terme, déterminons la limite de la double intégrale lorsque n tend vers l'infini, en appliquant le théorème de la convergence dominée de Lebesgue, d'une

part, à la suite de fonctions

$$\int_0^1 (1-t)\ f^{(2)}(x+vh_n t)\ dt\ v^2K(v)\ ,$$

et, d'autre part, à la suite de fonctions

$$(1-t)\ f^{(2)}(x+vh_n t)\ .$$

La première suite possède comme fonction dominante la fonction

$$M/2\ v^2\ K(v)$$

qui est intégrable selon l'hypothèse (H3.6). En effet,

$$\int_0^1 (1-t)\ f^{(2)}(x+vh_n t)\ dt\ v^2K(v)$$

$$\leq M \int_0^1 (1-t)\ dt\ v^2K(v)$$

$$\leq M/2\ v^2K(v)$$

où $M = \sup_{x \in \mathbb{R}} |\ f^{(2)}(x)|$.

La deuxième suite possède comme fonction dominante la fonction

$$M\ (1-t)$$

qui est intégrable sur [0,1] .

Ainsi, par le théorème de la convergence dominée de Lebesgue, on peut permuter limite et intégrale.

$$\lim_{n\to\infty} \int_{-\infty}^{+\infty}\int_0^1 (1-t)\ f^{(2)}(x+vh_n t)\ dt\ v^2K(v)\ dv =$$

$$= \int_{-\infty}^{+\infty} \lim_{n\to\infty} \int_0^1 (1-t)\, f^{(2)}(x+vh_n t)\, dt\, v^2K(v)\, dv$$

$$= \int_{-\infty}^{+\infty}\int_0^1 (1-t) \lim_{n\to\infty} [f^{(2)}(x+vh_n t)]\, dt\, v^2K(v)\, dv$$

$$= f^{(2)}(x)/2 \int_{-\infty}^{+\infty} v^2K(v)\, dv \ . \qquad (31)$$

Ainsi, puisque h_n tend vers zéro lorsque n tend vers l'infini, selon (H4.2), on a démontré que :

$$\lim_{n\to\infty} \text{Biais}[f_n(x)] = 0$$

et aussi

$$\text{Biais}[f_n(x)] = h_n^2/2\ f^{(2)}(x) \int_{-\infty}^{+\infty} v^2K(v)\, dv + o(h_n^2) \ .$$

Du théorème 1.5.2, on sait que :

$$\text{Var}_f[f_n(x)] = (nh_n)^{-1}\ f(x) \int_{-\infty}^{+\infty} K^2(u)\, du + o[(nh_n)^{-1}] \ .$$

Ainsi, en sommant les deux dernières expressions selon la formule (1.8), on obtient le résultat asymptotique désiré pour l'EQM$[f_n(x)]$:

$$\text{EQM}[f_n(x)] = (nh_n)^{-1}\ f(x) \int_{-\infty}^{+\infty} K^2(u)\, du + o[(nh_n)^{-1}]$$

$$+ h_n^4/4\ |\ f^{(2)}(x)|^2 [\ \int_{-\infty}^{+\infty} v^2K(v)dv\ ]^2 + o(h_n^4) \ .$$

Déterminons maintenant une formule analogue pour l'EQMI(f_n) :

$$EQMI(f_n) = \int_{-\infty}^{+\infty} \{ \mathrm{Var}_f[f_n(x)] + \mathrm{Biais}^2[f_n(x)] \} \, dx \, .$$

L'intégrale de la $\mathrm{Var}_f[f_n(x)]$ s'obtient aisément, alors que les calculs se compliquent quelque peu lors de l'intégration du $\mathrm{Biais}^2[f_n(x)]$.

$$\int_{-\infty}^{+\infty} \mathrm{Var}_f[f_n(x)] \, dx = (nh_n)^{-1} \int_{-\infty}^{+\infty} E_f\{h_n^{-1} K^2[(x-y)/h_n]\} \, dx$$

$$- n^{-1} \int_{-\infty}^{+\infty} E_f^{\,2}\{h_n^{-1} K[(x-y)/h_n]\} \, dx$$

Calculons chaque intégrale séparément et commençons par la première. La permutation des intégrales est permise par le théorème de Fubini.

$$\int_{-\infty}^{+\infty} E_f\{h_n^{-1} K^2[(x-y)/h_n]\} \, dx = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x+uh_n) \, K^2(u) \, du \, dx$$

$$\overset{\text{Fubini}}{=} \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x+uh_n) \, dx \, K^2(u) \, du$$

$$= \int_{-\infty}^{+\infty} f(x) \, dx \int_{-\infty}^{+\infty} K^2(u) \, du$$

$$= \int_{-\infty}^{+\infty} K^2(u) \, du \, .$$

Ainsi,

$$\lim_{n\to\infty} (nh_n)^{-1} \int_{-\infty}^{+\infty} E_f\{h_n^{-1} K^2[(x-y)/h_n]\}\, dx = 0$$

et

$$(nh_n)^{-1} \int_{-\infty}^{+\infty} E_f\{h_n^{-1} K^2[(x-y)/h_n]\}\, dx$$

$$= (nh_n)^{-1} \int_{-\infty}^{+\infty} K^2(u)\, du + o[(nh_n)^{-1}] .$$

Déterminons une approximation de la seconde intégrale en utilisant premièrement l'inégalité de Minkowski, puis celle de Cauchy (2x) et enfin le théorème de Fubini.

$$\int_{-\infty}^{+\infty} E_f{}^2\{h_n^{-1} K[(x-y)/h_n]\}\, dx =$$

$$= \int_{-\infty}^{+\infty} \{ \int_{-\infty}^{+\infty} [f(x) K(v) + h_n f'(x) vK(v)$$

$$+ h_n{}^2 \int_0^1 (1-t) f^{(2)}(x+vh_n t)dt\, v^2K(v)]dv \}^2 dx$$

Minkowski

$$\leq \int_{-\infty}^{+\infty} [\int_{-\infty}^{+\infty} f(x) K(v)\, dv]^2 dx$$

$$+ \int_{-\infty}^{+\infty} [\int_{-\infty}^{+\infty} h_n{}^2 \int_0^1 (1-t) f^{(2)}(x+vh_n t)dt\, v^2K(v)dv]^2 dx$$

$$\leq \int_{-\infty}^{+\infty} f^2(x)\ dx \left[\int_{-\infty}^{+\infty} K(v)dv\right]^2$$

Cauchy(2x)
et Fubini

$$+ h_n^4/4 \int_{-\infty}^{+\infty} [f^{(2)}(x)]^2 dx \left[\int_{-\infty}^{+\infty} v^2 K(v)dv\right]^2$$

$$\leq \int_{-\infty}^{+\infty} f^2(x)\ dx + h_n^4/4 \int_{-\infty}^{+\infty} [f^{(2)}(x)]^2 dx \left[\int_{-\infty}^{+\infty} v^2 K(v)dv\right]^2$$

Chaque intégrale existe selon les hypothèses (H2.2), (H2.3) et (H3.6) , si bien que

$$\int_{-\infty}^{+\infty} E_f{}^2\{h_n^{-1}\ K[(x-y)/h_n]\}\ dx$$

existe, est finie et positive. Ainsi, on a prouvé :

$$\lim_{n\to\infty}\ n^{-1} \int_{-\infty}^{+\infty} E_f{}^2\{h_n^{-1}\ K[(x-y)/h_n]\}\ dx = 0$$

et

$$n^{-1} \int_{-\infty}^{+\infty} E_f{}^2\{h_n^{-1}\ K[(x-y)/h_n]\}\ dx = O(n^{-1}) = o[(nh_n)^{-1}]\ .$$

Comme $Var_f[f_n(x)]$ est toujours positive, on a :

$$0 \leq \int_{-\infty}^{+\infty} Var_f[f_n(x)]\ dx \leq (nh_n)^{-1} \int_{-\infty}^{+\infty} E_f\{h_n^{-1} K^2[(x-y)/h_n]\}\ dx$$

$$= (nh_n)^{-1} \int_{-\infty}^{+\infty} K^2(u)\ du\ .$$

Lorsque n tend vers l'infini, nh_n tend vers l'infini selon (H4.3), et l'on a donc :

$$\lim_{n\to\infty} \int_{-\infty}^{+\infty} \mathrm{Var}_f[f_n(x)]\, dx = 0$$

et

$$\int_{-\infty}^{+\infty} \mathrm{Var}_f[f_n(x)]\, dx = (nh_n)^{-1} \int_{-\infty}^{+\infty} K^2(u)\, du + o[(nh_n)^{-1}] \quad (32)$$

Calculons maintenant une approximation de l'intégrale du Biais². Or, selon (30), on sait que :

$$\mathrm{Biais}[f_n(x)] = h_n^2 \int_{-\infty}^{+\infty} \int_0^1 (1-t)\, f^{(2)}(x+vh_n t)dt\; v^2K(v)dv \ .$$

Donc,

$$\int_{-\infty}^{+\infty} \mathrm{Biais}^2[f_n(x)]\, dx =$$

$$= \int_{-\infty}^{+\infty} [\, h_n^2 \int_{-\infty}^{+\infty} \int_0^1 (1-t)f^{(2)}(x+vh_n t)dt\; v^2K(v)dv]^2\, dx$$

$$= h_n^4 \int_{-\infty}^{+\infty} [\int_{-\infty}^{+\infty} \int_0^1 (1-t)f^{(2)}(x+vh_n t)dt\; v^2K(v)dv]^2\, dx \ .$$

Montrons que cette triple intégrale converge, lorsque n tend vers l'infini, vers

$$1/4 \int_{-\infty}^{+\infty} |f^{(2)}(x)|^2 dx \, [\int_0^{+\infty} v^2\, K(v)dv]^2 \ .$$

On utilise l'inégalité de Cauchy pour éliminer le carré de l'intégrale, en scindant une première fois $v^2K(v)$ en $vK^{\frac{1}{2}}(v)\cdot vK^{\frac{1}{2}}(v)$, et une deuxième fois $(1-t)f^{(2)}(x+vht)$ en $(1-t)^{\frac{1}{2}}\cdot(1-t)^{\frac{1}{2}}f^{(2)}(x+vht)$. Puis, en appliquant le théorème de Fubini, on permute les intégrales, puisque toutes les fonctions sont positives et qu'elles existent par hypothèse. Ainsi, on a la suite suivante d'égalités et d'inégalités :

$$\int_{-\infty}^{+\infty}\Big[\int_{-\infty}^{+\infty}\int_0^1 (1-t)f^{(2)}(x+vh_n t)dt\ v^2K(v)dv\Big]^2\ dx$$

$$\underset{\text{Cauchy}}{\leq} \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\Big[\int_0^1 (1-t)f^{(2)}(x+vh_n t)dt\Big]^2 v^2K(v)dv\Big[\int_{-\infty}^{+\infty} v^2K(v)dv\Big]dx$$

$$\underset{\text{Cauchy}}{\leq} \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}\Big[\int_0^1 (1-t)dt\Big]\Big\{\int_0^1 (1-t)\big[f^{(2)}(x+vh_n t)\big]^2dt\Big\}v^2K(v)dvdx\ B(K)$$

$$\underset{\text{Fubini}}{=} 1/2\int_{-\infty}^{+\infty}\int_0^1 (1-t)\ \Big\{\int_{-\infty}^{+\infty}\big[f^{(2)}(x+vh_n t)\big]^2dx\Big\}\ dt\ v^2K(v)dv\ B(K)$$

$$= 1/4\int_{-\infty}^{+\infty}\big[f^{(2)}(x)\big]^2dx\ \Big[\int_{-\infty}^{+\infty} v^2K(v)dv\Big]\ B(K)$$

$$= 1/4\int_{-\infty}^{+\infty}\big[f^{(2)}(x)\big]^2dx\ B^2(K) \qquad (33)$$

où $\quad B(K) = \int_{-\infty}^{+\infty} v^2K(v)dv\ .$

L'expression (33) existe, est finie selon les hypothèses (H2.3) et (H3.6), et indépendante de h_n.

Posons :

$$Q_n(x) = \int_{-\infty}^{+\infty} \int_0^1 (1-t)\, f^{(2)}(x+vh_n t)dt\, v^2K(v)dv \ .$$

Selon la formule (31), lorsque n tend vers l'infini, $Q_n(x)$ tend vers

$$Q(x) = f^{(2)}(x)/2 \int_{-\infty}^{+\infty} v^2K(v)dv$$

et donc $Q_n^2(x)$ tend vers $Q^2(x)$.

De plus, $\int_{-\infty}^{+\infty} Q_n^2(x)dx$ est toujours positive et inférieure à

$$1/4 \int_{-\infty}^{+\infty} [f^{(2)}(x)]^2dx\, B^2(K) \leq C < \infty \ .$$

Ainsi, en utilisant la formule (33) et le lemme de Fatou appliqué à la suite $Q_n^2(x)$, on a la suite suivante d'inégalités :

$$1/4 \int_{-\infty}^{+\infty} [f^{(2)}(x)]^2dx\, B^2(K) = \int_{-\infty}^{+\infty} \liminf_{n\to\infty} [Q_n^2(x)]\, dx$$

$$\overset{\text{Fatou}}{\leq} \liminf_{n\to\infty} \int_{-\infty}^{+\infty} Q_n^2(x)\, dx$$

$$\leq \limsup_{n\to\infty} \int_{-\infty}^{+\infty} Q_n^2(x)\, dx$$

$$\overset{(33)}{\leq} 1/4 \int_{-\infty}^{+\infty} [f^{(2)}(x)]^2dx\, B^2(K) \ .$$

Ainsi, on a démontré :

$$1) \lim_{n\to\infty} \int_{-\infty}^{+\infty} Q_n{}^2(x)\,dx = 1/4 \int_{-\infty}^{+\infty} [f^{(2)}(x)]^2 dx\; B^2(K) ,$$

$$2) \lim_{n\to\infty} \int_{-\infty}^{+\infty} Biais^2[f_n(x)]\,dx = \lim_{n\to\infty} [h_n{}^4 \int_{-\infty}^{+\infty} Q_n{}^2(x)\,dx] = 0 ,$$

$$3) \int_{-\infty}^{+\infty} Biais^2[f_n(x)]\,dx = h_n{}^4/4 \int_{-\infty}^{+\infty} [f^{(2)}(x)]^2 dx\; B^2(K) + o(h_n{}^4) . \quad (34)$$

En regoupant les expressions (32) et (34), la formule asymptotique désirée pour l'EQMI(f_n) est déterminée.

A partir des expressions (26) et (27), il est aisé de chercher une fenêtre minimisant chacun de ces deux critères; il suffit de déterminer les zéros des dérivées par rapport à h_n, puis de remplacer les largeurs trouvées, notées h_{EQM} et h_{EQMI}, dans les expressions correspondantes.

Dans ce travail nous étudions principalement le critère de l'EQMI pour différents noyaux; déterminons la fenêtre optimale minimisant l'EQMI et l'EQMI correspondante pour l'estimateur à noyau f_n.

Corollaire 1.5.5 (Rosenblatt, 1971)

Supposons que les hypothèses du théorème 1.5.4 soient vérifiées.
Alors, la fenêtre minimisant l'EQMI(f_n) vaut :

$$h_{EQMI} = [A(K)/B^2(K)]^{1/5} \left[\int_{-\infty}^{+\infty} |f^{(2)}(x)|^2 dx\right]^{-1/5} n^{-1/5} \quad (35)$$

et l'EQMI(f_n) minimale vaut :

$$EQMI(f_n) = 5/4\ [A^4(K)\cdot B^2(K)]^{1/5}\ \left[\int_{-\infty}^{+\infty} |f^{(2)}(x)|^2 dx\right]^{1/5}\ n^{-4/5} + o(n^{-4/5})\ . \quad (36)$$

On constate que l'EQMI optimale ne dépend que du noyau choisi et de la dérivée seconde de la densité à estimer f. Cette formule donne ainsi une borne pour l'EQMI(f_n).

Epanechnikov a déterminé un noyau minimisant l'EQMI(f_n) (cf p.11); mais de nombreux articles démontrent que le choix du noyau n'est pas prépondérant (Tapia et al. (1978) et paragraphes 2.4 et 3.5). En effet, si l'on substitue les valeurs de A(K) et B(K) obtenues pour différents noyaux, les EQMI correspondantes sont très voisines (cf théorème 2.2.1).

Remarque

Les théorèmes 1.5.2 à 1.5.5 peuvent être démontrés pour des noyaux qui ne sont pas nécessairement des densités. L'hypothèse (H3.1) est alors remplacée par $K \in L_1$, tandis que l'hypothèse (H3.6) est substituée, lorsque s est supérieur à 2 et pair, par :

$$\int_{-\infty}^{+\infty} v^j\ K(v)\ dv = 0 \qquad \text{pour } j = 1,\ldots, s-1\ ,$$

$$\int_{-\infty}^{+\infty} v^s\ K(v)\ dv \neq 0,$$

$$\int_{-\infty}^{+\infty} v^s\ |\ K(v)\ |\ dv \quad \text{est finie.}$$

L'hypothèse (H2.3) pour la densité f est remplacée elle par :

- f est s-fois continûment différentiable et $f^{(s)}$ est bornée et de carré intégrable.

Les démonstrations sont semblables à celles qui viennent d'être données (cf B.Prakasa Rao (1983)).

Dans notre travail, nous nous restreignons à deux noyaux, le noyau biquadratique K_1 et le noyau normal K_2.

$$K_1(y) = \begin{cases} 15/16\ (\ 1 - y^2)^2 & \text{si } |\ y\ | \leq 1 \\ 0 & \text{sinon} \end{cases} \qquad (37)$$

$$K_2(y) = (2\Pi)^{-1/2}\ \exp[-y^2/2] \qquad (38)$$

Ces noyaux sont les plus couramment utilisés dans la pratique d'une part, et ils possèdent de bonnes qualité de lissage d'autre part. Ils vérifient les hypothèses (H3) énoncées auparavant.

Le noyau normal possède encore un avantage non négligeable : il se prête très bien au calcul de l'estimation d'une densité normale grâce aux propriétés de multiplications et d'intégrations de densités normales (cf annexe 1). Ainsi, l'EQMI exacte peut être déterminée algébriquement.
Le noyau biquadratique possède lui un support compact et donc seules les observations les plus proches du point x contribuent à l'estimation de f. Sa formule asymptotique de l'EQMI est très voisine de celle du noyau normal (théorème 2.2.1) et ainsi la qualité de l'exactitude de la formule asymptotique peut être établie.

Pour ces deux noyaux, les constantes A(K) et B(K) valent respectivement :

pour le noyau biquadratique : $A(K_1) = 5/7$ (39)

$B(K_1) = 1/7$ (40)

pour le noyau normal :

$$A(K_2) = (4\Pi)^{-1/2} \qquad (41)$$

$$B(K_2) = 1.0 \qquad (42)$$

Conclusion

La présentation des estimateurs, des critères, des théorèmes nécessaires étant terminée, nous pouvons passer à l'analyse de l'estimation de la densité normale (chapitre 2), avant d'étudier une famille de densité plus générale, les densités normales contaminées par une autre densité normale (chapitre 3).

CHAPITRE 2 : ESTIMATION DE LA DENSITE NORMALE

Ce chapitre est consacré à l'étude des estimateurs paramétriques et non-paramétriques définis au chapitre 1 pour la famille $\mathbb{P}$ des densités normales :

$\mathbb{P}$ = { $f(\mu,\sigma^2;x)$ fonction de densité normale, d'espérance μ et de variance σ^2 en un point fixé x, notée $N(\mu,\sigma^2;x)$ }

où $N(\mu,\sigma^2;x) = (2\Pi)^{-1/2}\ \sigma^{-1}\ \exp[-(x-\mu)^2/(2\sigma^2)]$. (1)

La fonction $N(\mu,\sigma^2;x)$ vérifie les hypothèses (H2) et représente dans ce chapitre la fonction à estimer f.

Dans les deux premières parties, le critère de l'EQMI est calculé explicitement pour chaque estimateur. L'étude du comportement analytique de l'EQMI et de l'efficacité forme la troisième partie, tandis que le dernier paragraphe est consacré aux simulations.

2.1 EQMI des estimateurs paramétriques

La fonction de densité N à estimer, dépend au plus de deux paramètres. Les trois cas pouvant se présenter sont traités séparément et complètement, à savoir :

1. L'espérance seule est inconnue.
2. La variance seule est inconnue.
3. L'espérance et la variance sont inconnues.

De nombreux choix d'estimateurs pour μ et σ^2 s'offrent au statisticien :
les estimateurs . du maximum de vraisemblance (*MLE*),
. des moindres carrés,
. robustes (Huber),
. de rang,

. de Bayes,
. non biaisés,
. de variance minimale.

Certains estimateurs vérifient plusieurs de ces critères : exemple, la moyenne arithmétique. Les estimateurs non biaisés et de variance minimale se sont imposés à nos yeux, car ce sont les plus couramment utilisés d'une part et que leurs propriétés sont bien connues d'autre part, facilitant ainsi les calculs. La moyenne arithmétique, $\mathbf{m}_n$, et la variance empirique, u_n ou v_n, sont donc les estimateurs choisis pour l'espérance μ et pour la variance σ^2.

<u>1er cas</u> : $\mathbf{m}_n = n^{-1} \sum_{i=1}^{n} X_i$ σ^2 connu (2)

<u>2ème cas</u> : $u_n = n^{-1} \sum_{i=1}^{n} (X_i - \mu)^2$ μ connu (3)

<u>3ème cas</u> : $\mathbf{m}_n$ et $v_n = (n-1)^{-1} \sum_{i=1}^{n} (X_i - \mathbf{m}_n)^2$ (4)

Les estimateurs de la fonction de densité correspondant à ces trois cas sont respectivement :

$$g_{1n}(x) = \mathbf{N}(\mathbf{m}_n, \sigma^2; x) \ , \tag{5}$$

$$g_{2n}(x) = \mathbf{N}(\mu, u_n; x) \ , \tag{6}$$

$$g_{3n}(x) = \mathbf{N}(\mathbf{m}_n, v_n; x) \ . \tag{7}$$

<u>Conventions d'écriture</u>

L'indice n est omis si aucune confusion n'est possible pour les estimateurs, $\mathbf{m}_n$, u_n, v_n, g_{1n}, g_{2n}, g_{3n}.
$\mathbf{m}$, u, v, écrits en caractères gras, désignent les estimateurs, tandis qu'en caractères normaux, ils représentent une estimation.
Les bornes de Σ sont omises si elles varient de 1 à n.

Maintenant que les estimateurs des paramètres et que les estimateurs respectifs de la densité sont définis, on peut calculer les EQMI pour chaque cas. Les résultats lorsque l'espérance est inconnue ont été notamment démontrés par Guttmann et Wertz (1976) et sont déterminés dans le théorème suivant.

Théorème 2.1.1

Supposons que les hypothèses (H1) soient vérifiées, que seule l'espérance μ soit inconnue et qu'elle soit estimée par $\mathbf{m}$.
Alors l'estimateur paramétrique g_1 est défini par :

$$g_1(x) = \mathbf{N}(\mathbf{m},\sigma^2;x) \qquad \text{où} \qquad \mathbf{m} = n^{-1} \sum_i X_i \ ;$$

et l'on a :

$$EQMI(g_1) = (\Pi\sigma^2)^{-1/2} \{ 1 - [2n/(2n+1)]^{1/2} \} \qquad (8)$$

ou

$$EQMI(g_1) = (\Pi\sigma^2)^{-1/2} [1/(4n) - 3/(32n^2)] + O(n^{-3}). \qquad (9)$$

Démonstration

Les variables aléatoires $X_1, X_2, \ldots, X_n$ étant i.i.d. et normalement distribuées, l'estimateur $\mathbf{m}$ est une variable aléatoire distribuée selon une loi normale, de paramètres μ et σ^2/n. Les calculs des critères sont simples, car seules interviennent des convolutions de densités normales, puis leur intégration (cf annexe 1).

Pour un point x fixé, calculons les différents termes de l'$EQM[g_1(x)]$ selon la formule (1.9) :

$$EQM[g_1(x)] = E_f[g_1^2(x)] + \mathbf{N}^2(\mu,\sigma^2;x) - 2\,\mathbf{N}(\mu,\sigma^2;x)\,E_f[g_1(x)] \ . \qquad (10)$$

1) $E_f[g_1^2(x)] = E_m[N^2(m,\sigma^2;x)]$

$= E_m[(4\Pi\sigma^2)^{-1/2}\ N(m,\sigma^2/2;x)]$

$= \int_{-\infty}^{+\infty} (4\Pi\sigma^2)^{-1/2}\ N(m,\sigma^2/2;x)\ N(\mu,\sigma^2/n;m)\ dm$

$= (4\Pi\sigma^2)^{-1/2}\ N[\mu,(n+2)\sigma^2/(2n);x]$. (11)

2) $N^2(\mu,\sigma^2;x) = (4\Pi\sigma^2)^{-1/2}\ N(\mu,\sigma^2/2;x)$. (12)

3) $E_f[g_1(x)] = E_m[g_1(x)] = \int_{-\infty}^{+\infty} N(m,\sigma^2;x)\ N(\mu,\sigma^2/n;m)\ dm$

$= N[\mu,(n+1)\sigma^2/n;x]$.

Ainsi,

$N(\mu,\sigma^2;x)\ E_f[g_1(x)] = N(\mu,\sigma^2;x)\ N[\mu,(n+1)\sigma^2/n;x]$

$= (2\Pi)^{-1/2}\ [n\cdot\sigma^2/(2n+1)]^{1/2}\ N[\mu,(n+1)\sigma^2/(2n+1);x]$

$= (4\Pi\sigma^2)^{-1/2}\ [2n/(2n+1)]^{1/2}\ N[\mu,(n+1)\sigma^2/(2n+1);x]$. (13)

En substituant les expressions (11), (12) et (13) dans la formule (10), on obtient le résultat désiré :

$EQM[g_1(x)] = (4\Pi\sigma^2)^{-1/2}\ \{\ N[\mu,(n+2)\sigma^2/(2n);x] + N(\mu,\sigma^2/2;x)$

$- 2\ [2n/(2n+1)]^{1/2}\ N[\mu,(n+1)\sigma^2/(2n+1);x]\ \}$.

En intégrant cette dernière expression par rapport à x, on obtient l'$EQMI(g_1)$ selon la formule (1.10) :

$EQMI(g_1) = \int_{-\infty}^{+\infty} EQM[g_1(x)]\ dx$

$$= \int_{-\infty}^{+\infty} \{(4\Pi\sigma^2)^{-1/2}\ N[\mu,(n+2)\sigma^2/(2n);x] + N(\mu,\sigma^2/2;x)$$
$$- 2\ [2n/(2n+1)]^{1/2}\ N[\mu,(n+1)\sigma^2/(2n+1);x]\}\ dx$$

$$= (4\Pi\sigma^2)^{-1/2}\ \{\ 1 + 1 - 2\ [2n/(2n+1)]^{1/2}\ \}$$

$$= (\Pi\sigma^2)^{-1/2}\ \{\ 1 - [2n/(2n+1)]^{1/2}\ \}\ .$$

En développant selon les puissances de n^{-1} l'expression $[2n/(2n+1)]^{1/2}$, on obtient une approximation de l'EQMI(g_1) plus facilement interprétable :

$$EQMI(g_1) = (\Pi\sigma^2)^{-1/2}\ [\ 1/(4n) - 3/(32n^2)\ ] + O(n^{-3})\ .$$

Remarques

1) Il est possible de déterminer un estimateur de f dont le biais est nul. Il est noté g_{11} et défini par :

$$g_{11}(x) = N(m,(1-1/n)\sigma^2;x)\ . \qquad (14)$$

Un calcul identique au précédent conduit à la formule suivante pour son EQMI :

$$EQMI(g_{11}) = (4\Pi\sigma^2)^{-1/2}\ \{\ [n/(n-1)]^{1/2} - 1\ \} \qquad (15)$$

ou

$$EQMI(g_{11}) = (4\Pi\sigma^2)^{-1/2}\ [\ 1/n + 3/(4n^2)\ ] + O(n^{-3})\ . \qquad (16)$$

2) Guttmann et Wertz (1976) et Klebanov (1977) ont montré que la meilleure estimation paramétrique possible pour une densité normale, où seule l'espérance est inconnue, est donnée par :

$$g_{12}(x) = N[m,(n+1)\sigma^2/n;x]\ . \qquad (17)$$

Dans ce cas, l'EQMI(g_{12}) est calculée de manière identique que ci-dessus et l'on obtient :

$$EQMI(g_{12}) = (4\Pi\sigma^2)^{-1/2} \{ 1 - [n/(n+1)]^{1/2} \} \qquad (18)$$

ou

$$EQMI(g_{12}) = (\Pi\sigma^2)^{-1/2} [1/(4n) - 3/(16n^2)] + O(n^{-3}) . \qquad (19)$$

La différence entre les EQMI de g_1 et de g_{12} est minime comme l'on peut se rendre compte par les développements limités ou par les tabulations jointes (tableau 2.1). Les graphes des EQMI se confondent et ne sont donc pas tracés.
On constate de plus que si g_{11} est non biaisé, son EQMI est pourtant supérieure aux autres estimateurs proposés. Ces remarques confirment le choix effectué pour l'estimateur g_1, très proche du cas idéal, meilleur que l'estimateur sans biais et défini très logiquement.

Lorsque la variance est inconnue, il n'existe pas de résultats théoriques dans la littérature à notre connaissance. Les formules obtenues pour l'EQMI de g_2 et de g_3 représentent les résultats principaux de ce paragraphe.

Théorème 2.1.2

Supposons que les hypothèses (H1) soient vérifiées, que la variance σ^2 soit inconnue et qu'elle soit estimée par u, tandis que l'espérance est connue.
Alors, l'estimateur de la densité g_2 est défini par :

$$g_2(x) = N(\mu,u;x) \quad \text{où} \quad u = n^{-1} \Sigma_1 (X_1 - \mu)^2 ;$$

et l'on a :

$$EQMI(g_2) = (4\Pi\sigma^2)^{-1/2} \{ (n/2)^{1/2} \Gamma[(n-1)/2]/\Gamma(n/2)$$

$$- (1 + 3/(8n) - 55/(256n^2) \} + O(n^{-3}) \qquad (20)$$

ou

$$EQMI(g_2) = (\Pi\sigma^2)^{-1/2}[\ 3/(16n) + 255/(512n^2)\ ] + O(n^{-3})\ .\quad (21)$$

Démonstration

Les variables aléatoires $X_1, X_2, \ldots, X_n$ étant i.i.d et normalement distribuées, $n\cdot u/\sigma^2$ est une variable aléatoire distribuée selon une loi du chi-carré à n degrés de liberté. Ainsi, il est possible de déterminer la fonction de densité de **u** par un simple changement de variable; elle est notée l(u) :

$$l(u) = 1/\Gamma(n/2)\ [n/(2\sigma^2]^{n/2}\ u^{(n-2)/2}\exp[-n\cdot u/(2\sigma^2)]\quad (22)$$

où $u \geq 0$ et $\Gamma(n)$ représente la fonction Gamma.

Le calcul des intégrales intervenant dans l'$EQM[g_2(x)]$ et dans l'$EQMI(g_2)$ n'est plus aussi trivial qu'auparavant. Afin d'obtenir des résulats exacts, il a fallu intervertir l'ordre des intégrales dans chaque terme de l'EQMI, en appliquant le théorème de Fubini.
Selon la définition de l'EQMI, formules (1.9) et (1.10), on a :

$$EQMI(g_2) = \int_{-\infty}^{+\infty} [\ E_f[g_2{}^2(x)] + N^2(\mu,\sigma^2;x) - 2\ N(\mu,\sigma^2;x)\ E_f[g_2(x)]\ ]\ dx\ .\quad (23)$$

Or,

$$E_f[g_2(x)] = E_u[g_2(x)] = E_u[N(\mu,u;x)]$$

et

$$E_f[g_2{}^2(x)] = E_u[g_2{}^2(x)] = E_u[(4\Pi u)^{-1/2}\ N(\mu,u/2;x)].$$

Ainsi,

$$1) \int_{-\infty}^{+\infty} E_u[g_2{}^2(x)]\ dx = \int_{-\infty}^{+\infty}\int_{0}^{+\infty} (4\Pi u)^{-1/2}\ N(\mu,u/2;x)]\ l(u)\ du\ dx$$

$$= \int_0^{+\infty} (4\Pi u)^{-1/2} \int_{-\infty}^{+\infty} N(\mu, u/2; x)\, dx\, 1(u)\, du$$

$$= \int_0^{+\infty} (4\Pi u)^{-1/2}\, 1(u)\, du$$

$$= \int_0^{+\infty} (4\Pi u)^{-1/2}\, 1/\Gamma(n/2)[n/(2\sigma^2)]^{n/2}\, u^{(n-2)/2} \cdot \exp[-n\cdot u/(2\sigma^2)]\, du \qquad (24)$$

en posant $t = n\cdot u/(2\sigma^2)$, on obtient :

$$= (4\Pi\sigma^2)^{-1/2}\, (n/2)^{1/2}\, 1/\Gamma(n/2) \int_0^{+\infty} t^{(n-1)/2-1}\, \exp(-t)dt$$

$$= (4\Pi\sigma^2)^{-1/2}\, (n/2)^{1/2}\, \Gamma[(n-1)/2]/\Gamma(n/2)\ . \qquad (25)$$

table [1]
p.255 (6.1.1)

$$2) \int_{-\infty}^{+\infty} N^2(\mu, \sigma^2; x)\, dx = \int_{-\infty}^{+\infty} (4\Pi\sigma^2)^{-1/2}\, N(\mu, \sigma^2/2; x)\, dx$$

$$= (4\Pi\sigma^2)^{-1/2}\ . \qquad (26)$$

$$3) \int_{-\infty}^{+\infty} N(\mu, \sigma^2; x)\, E_u[g_2(x)]\, dx$$

$$= \int_{-\infty}^{+\infty} N(\mu, \sigma^2; x) \int_0^{+\infty} N(\mu, u; x)\, 1(u)\, du\, dx$$

$$= \int_0^{+\infty} \int_{-\infty}^{+\infty} N(\mu, \sigma^2; x)\, N(\mu, u; x)\, dx\, 1(u)\, du$$

$$= \int_0^{+\infty} (2\Pi)^{-1/2} \ (\sigma^2+u)^{-1/2} \ 1(u) \ du$$

$$= \int_0^{+\infty} (2\Pi\sigma^2)^{-1/2} \ (1+u/\sigma^2)^{-1/2} \ 1(u) \ du \qquad (27)$$

en posant $t = u/\sigma^2$, on obtient :

$$= (2\Pi\sigma^2)^{-1/2}(n/2)^{n/2} \ 1/\Gamma(n/2)$$

$$\cdot \int_0^{+\infty} (1+t)^{-1/2} \ t^{n/2-1} \ \exp(-n\cdot t/2)dt$$

$$= (2\Pi\sigma^2)^{-1/2} \ (n/2)^{n/2} \ \text{Kum}[n/2,(n+1)/2,n/2] \ . \qquad (28)$$

table [1]
p.505 (13.2.5)

Kum représente la fonction de Kummer généralisée. Il n'existe pas de formule satisfaisante à notre connaissance pour approximer cette fonction lorsque la variable n est supérieure à 10.
Pour pallier à cet inconvénient, l'intégrale (27) a été transformée en posant $t = n\cdot u/\sigma^2$; on obtient alors :

$$3) = \int_0^{+\infty} (2\Pi\sigma^2)^{-1/2} \ (1+t/n)^{-1/2} \ X_n^{\,2}(t) \ dt$$

où $X_n^{\,2}(t)$ est la fonction de densité de la répartition du chi-carré à n degrés de liberté.

Posons $P(t) = (1+t/n)^{-1/2}$ et développons la fonction P(t) en série de Taylor autour du point $t_0 = n$, qui correspond à la valeur moyenne théorique de la fonction du $X_n^{\,2}(t)$.

Si $P^{(j)}$ désigne la $j^{ième}$ dérivée de P par rapport à t, alors,

$$1/j!\ P^{(j)}(t) = D_j\ (1+t/n)^{-(2j+1)/2}\ n^{-j}$$

$$\text{où} \quad D_j = (-1)^j\ \frac{1\ 3\ 5 \ldots (2j-1)}{2\ 4\ 6 \ldots (2j)} \qquad j = 1,2, \ldots$$

En prenant les cinq premiers termes de la série de Taylor et en intégrant terme à terme, les termes de la série ainsi obtenue sont de la forme :

$$\alpha_j \cdot D_j \cdot 2^{-(2j+1)/2}\ n^{-j} \qquad j = 1,\ldots,4$$

où α_j représente le moment centré d'ordre j de la répartition du chi-carré qui est un polynôme en n de degré [j/2] ([x] désigne la partie entière de x).
En regroupant les termes de même puissance, on obtient finalement (cf annexe 2 pour les calculs détaillés) :

$$3) = (4\Pi\sigma^2)^{-1/2}\ [\ 1 + 3/(16n) - 55/(512n^2)\ ] + O(n^{-3})\ . \qquad (29)$$

$EQMI(g_2)$ découle de la substitution des termes (25), (26) et (29) dans la formule (23).

$$EQMI(g_2) = (4\Pi\sigma^2)^{-1/2}\ \{\ 1 + (n/2)^{1/2}\ \Gamma[(n-1)/2]/\Gamma(n/2)$$

$$- [\ 2 + 3/(8n) - 55/(256n^2)]\ \} + O(n^{-3})$$

$$= (4\Pi\sigma^2)^{-1/2}\ \{\ (n/2)^{1/2}\ \Gamma[(n-1)/2]/\Gamma(n/2)$$

$$- [\ 1 + 3/(8n) - 55/(256n^2]\ ] + O(n^{-3})\ .$$

Il est possible de développer en série de Taylor autour du même point $t_0 = n$ la fonction $u^{-1/2}$ intervenant dans l'intégrale (24); on obtient ainsi un résultat approximatif de l'intégrale (24) et donc aussi de l'$EQMI(g_2)$ (cf annexe 2) :

$$EQMI(g_2) = (4\Pi\sigma^2)^{-1/2} \left[3/(8n) + 255/(256n^2) \right] + O(n^{-3}).$$

Il reste à déterminer l'EQMI pour le cas le plus général où les deux paramètres sont inconnus.

Théorème 2.1.3

Supposons que les hypothèses (H1) soient vérifiées et que l'espérance et la variance soient inconnues. Ils sont estimés respectivement par m et v.
Alors l'estimateur de la densité g_3 est défini par :

$$g_3(x) = N(m,v;x)$$

où $\quad m = n^{-1} \Sigma_i X_i \quad$ et $\quad v = (n-1)^{-1} \Sigma_i (X_i - m)^2$;

l'$EQMI(g_3)$ vaut :

$$EQMI(g_3) = (4\Pi\sigma^2)^{-1} \left\{ [(n-1)/2]^{1/2} \Gamma[(n-2)/2]/\Gamma[(n-1)/2] \right.$$
$$\left. - [\, 1 - 1/(8n) - 31/(256n^2)] \right\} + O(n^{-3}) \qquad (30)$$

ou

$$EQMI(g_3) = (\Pi\sigma^2)^{-1/2} \left[7/(16n) + 423/(512n^2) \right] + O(n^{-3}). \qquad (31)$$

Démonstration

La façon de calculer l'$EQMI(g_3)$ est identique à celle de la proposition précédente, même si des intégrales triples interviennent. Par le théorème de Fubini, l'ordre d'intégration a été interverti, de telle manière que la variable v soit prise en considération en dernier lieu.

Les variables aléatoires $X_1, X_2, \ldots, X_n$ étant i.i.d et normalement distribuées, m et $(n-1)v/\sigma^2$ sont deux variables aléatoires indépendantes, de répartition normale, d'espérance μ et de variance σ^2/n pour m et de

répartition du chi-carré à (n-1) degrés de liberté pour $(n-1)v/\sigma^2$. Par changement de variable, on obtient aisément la densité de v, notée l(v) :

$$l(v) = 1/\Gamma[(n-1)/2]\ [(n-1)/(2\sigma^2)]^{(n-1)/2}\ v^{(n-3)/2} \cdot \exp[-(n-1)v/(2\sigma^2)]\ ,\ v \geq 0\ . \qquad (32)$$

Constatons tout d'abord que :

$$E_f[g_3(x)] = E_{m,v}[g_3(x)]$$

$$= \int_0^{+\infty} \int_{-\infty}^{+\infty} N(m,v;x)\ N(\mu,\sigma^2/n;m)\ l(v)\ dm\ dv$$

et

$$E_f[g_3^2(x)] = E_{m,v}[g_3^2(x)]$$

$$= \int_0^{+\infty} \int_{-\infty}^{+\infty} (4\Pi\sigma^2)^{-1/2}\ N(m,v/2;x)\ N(\mu,\sigma^2/n;m)\ l(v)\ dm\ dv.$$

Or, par définition,

$$EQMI(g_3) = \int_{-\infty}^{+\infty} \{\ E_f[g_3^2(x)] + N^2(\mu,\sigma^2;x) - 2\ N(\mu,\sigma^2;x)\ E_{m,v}[g_3(x)]\ \}\ dx\ . \qquad (33)$$

Déterminons pas à pas chaque terme de l'EQMI(g_3).

$$1)\ \int_{-\infty}^{+\infty} E_f[g_3^2(x)]\ dx = \int_{-\infty}^{+\infty} \int_0^{+\infty} \int_{-\infty}^{+\infty} (4\Pi v)^{-1/2} N(m,v/2;x) \cdot N(\mu,\sigma^2;m)\ dm\ dv\ dx$$

$$= \int_0^{+\infty} (4\pi v)^{-1/2} \left[\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} N(m, v/2; x)\, N(\mu, \sigma^2/n; m)\, dm\, dx \right] \cdot 1(v)\, dv$$

$$= \int_0^{+\infty} (4\pi v)^{-1/2} \int_{-\infty}^{+\infty} N(\mu, \sigma^2/n+v/2; x)\, dx\, 1(v)\, dv$$

$$= \int_0^{+\infty} (4\pi v)^{-1/2}\, 1(v)\, dv \qquad (34)$$

$$= (4\pi\sigma^2)^{-1/2} [(n-1)/2]^{1/2} \Gamma[(n-2)/2] / \Gamma[(n-1)/2]. \qquad (35)$$

table [1]
p.255 (6.1.1)

2) $$\int_{-\infty}^{+\infty} N^2(\mu, \sigma^2; x)\, dx = \int_{-\infty}^{+\infty} (4\pi\sigma^2)^{-1/2}\, N(\mu, \sigma^2/2; x)\, dx$$

$$= (4\pi\sigma^2)^{-1/2} . \qquad (36)$$

3) $$\int_{-\infty}^{+\infty} N(\mu, \sigma^2; x) \int_0^{+\infty} \int_{-\infty}^{+\infty} N(m, v; x)\, N(\mu, \sigma^2/n; m)\, dm\, 1(v)\, dv\, dx$$

$$= \int_0^{+\infty} \int_{-\infty}^{+\infty} N(\mu, \sigma^2; x)\, N(\mu, v+\sigma^2/n; x)\, dx\, 1(v)\, dv$$

$$= \int_0^{+\infty} (2\pi)^{-1/2} [v+(n+1)\sigma^2/n]^{-1/2}\, 1(v)\, dv$$

$$= \int_0^{+\infty} (2\pi\sigma^2)^{-1/2} [v/\sigma^2+(n+1)/n]^{-1/2}\, 1(v)\, dv \qquad (37)$$

$$= \int_0^{+\infty} (2\Pi\sigma^2)^{-1/2}\ Q(v)\ 1(v)\ dv$$

où $Q(v) = [v/\sigma^2+(n+1)/n]^{-1/2}$.

En mettant en évidence $[(n+1)/n]^{-1/2}$, puis en posant $t = n\cdot v/[(n+1)\sigma^2]$, on obtient dans le résultat à nouveau la fonction de Kummer.

$$3) = (2\Pi\sigma^2)^{-1/2}\ [n/(n+1)]^{1/2}\ [(n^2-1)/(2n)]^{(n-1)/2}$$

$$\cdot\ \mathrm{Kum}[(n-1)/2, n/2, (n^2-1)/(2n)]. \quad (38)$$

Afin que la dernière expression soit plus explicite et plus maniable, posons $s = (n-1)v/\sigma^2$ dans l'intégrale (37); alors l'expression devient :

$$3) = \int_0^{+\infty} (2\Pi\sigma^2)^{-1/2}\ R(s)\ X_{n-1}{}^2(s)\ ds$$

où $R(s) = [(n+1)/n + s/(n-1)]^{-1/2}$.

Si $R^{(j)}$ désigne la $j^{ième}$ dérivée de R(s) par rapport à s, alors,

$$1/j!\ R^{(j)}(s) = D_j \cdot [(n+1)/n+s/(n-1)]^{-(2j+1)/2}(n-1)^{-j}$$

où $D_j = (-1)^j \dfrac{1\ 3\ 5 \ldots (2j-1)}{2\ 4\ 6 \ldots (2j)}$ $\qquad j = 1,2, \ldots$

Développons en série de Taylor au point $s_0 = n-1$ la fonction R(s), et intégrons terme à terme. Les termes de la série obtenue sont de la forme :

$$ß_j \cdot D_j \cdot 2^{-1/2}[2(n-1)]^{-j}\ [1+1/(2n)]^{-(2j+1)/2} \qquad j = 1,\ldots,4$$

où $ß_j$ représente le moment centré d'ordre j de la répartition $X_{n-1}{}^2$. Puis on a développé les expressions

$[2(n-1)]^{-j}$ et $[1+1/(2n)]^{-(2j+1)/2}$ selon les puissances de n^{-1}, effectué les multiplications nécessaires et regroupé les termes de même degré jusqu'à l'ordre 3 (cf annexe 3). On obtient finalement :

$$3) = (4\Pi\sigma^2)^{-1/2} \left[1 - 1/(16n) - 31/(512n^2) \right] + O(n^{-3}) . \quad (39)$$

Alors, la formule désirée de l'EQMI est obtenue :

$$EQMI(g_3) = (4\Pi\sigma^2)^{-1} \{ 1 + [(n-1)/2]^{1/2} \Gamma[(n-2)/2]/\Gamma[(n-1)/2] - [2 - 1/(8n) - 31/(256n^2)] \} + O(n^{-3})$$

$$= (4\Pi\sigma^2)^{-1} \{ [(n-1)/2]^{1/2} \Gamma[(n-2)/2]/\Gamma[(n-1)/2] - [1 - 1/(8n) - 31/(256n^2)] \} + O(n^{-3}) .$$

En utilisant le même artifice que ci-dessus dans l'intégrale (34), on obtient une approximation du premier terme de l'accolade selon les puissances de n^{-1} et donc aussi de l'EQMI, après regroupement des termes semblables (cf annexe 3).

$$EQMI(g_3) = (4\Pi\sigma^2)^{-1/2} \left[7/(8n) + 423/(256n^2) \right] + O(n^{-3}).$$

Avant d'analyser et de comparer les formules obtenues pour les différents estimateurs paramétriques, nous allons déterminer les formules des EQMI pour les estimateurs non-paramétriques.

2.2 EQMI des estimateurs non-paramétriques

Comme il a été précisé au chapitre 1, les deux noyaux que nous retenons pour notre étude sont le noyau biquadratique, K_1, et le noyau normal, K_2.

L'estimateur non-paramétrique construit à l'aide du noyau biquadratique K_1 est noté f_{1n}, tandis que f_{2n} représente l'estimateur utilisant le noyau normal K_2.

$$f_{jn}(x) = (nh_n)^{-1} \sum_{i=1}^{n} K_j[(x-X_i)/h_n] \qquad j = 1,2 \qquad (40)$$

$$\text{où} \quad K_1(y) = \begin{cases} 15/16\ (1-y^2)^2 & \text{si } |y| \leq 1 \\ 0 & \text{sinon} \end{cases} \qquad (41)$$

$$K_2(y) = N(0,1;y)\ . \qquad (42)$$

Rappel

h_n est une suite de nombres positifs vérifiant les hypothèses (H4), c-à-d

$$\lim_{n\to\infty} h_n = 0 \qquad \text{et} \qquad \lim_{n\to\infty} nh_n = \infty\ .$$

Conventions d'écriture

L'indice n sera omis si aucune confusion n'est possible dans les expressions f_{1n}, f_{2n}, h_{1n}, h_{2n}. L'indice 1 se rapporte au noyau biquadratique, tandis que l'indice 2 au noyau normal.
Les bornes de Σ sont omises si elles varient de 1 à n.

Nous déterminons tout d'abord la fenêtre minimisant l'EQMI asymptotique et l'EQMI correspondante pour les deux noyaux cités, en utilisant les formules établies au corollaire 1.5.5.

Théorème 2.2.1 (Rosenblatt, 1971)

Supposons que les hypothèses (H1) et (H2) soient vérifiées.
Alors la largeur réduite k minimisant l'EQMI asymptotique vaut suivant les noyaux :

$$k_1 = 2.778\ n^{-1/5} \tag{43}$$

et

$$k_2 = 1.059\ n^{-1/5} \ . \tag{44}$$

Et l'EQMI correspondante vaut :

$$EQMI(f_1) = 0.32141\ \sigma^{-1}\ n^{-4/5} + o(n^{-4/5}) \tag{45}$$

et

$$EQMI(f_2) = 0.33290\ \sigma^{-1}\ n^{-4/5} + o(n^{-4/5}) \ . \tag{46}$$

Démonstration

Les choix proposés pour k_1 et k_2 remplissent les hypothèses (H4). Ces résultats découlent immédiatement des formules (1.35) et (1.36), en calculant les différents termes pour la densité normale et les noyaux K_1 et K_2.
Déterminons tout d'abord h_{opt}, selon la formule (1.35) :

$$h_{opt} = [A(K)/B^2(K)]^{1/5} \left\{ \int_{-\infty}^{+\infty} |f''(x)|^2\ dx \right\}^{-1/5} n^{-1/5} \ .$$

Pour la densité normale $N(\mu,\sigma^2;x)$, on a :

$$\int_{-\infty}^{+\infty} |f''(x)|^2\ dx = 3/(8\sigma^5\sqrt{\Pi}) \ . \tag{47}$$

Pour les noyaux biquadratique et normal, l'expression $A(K)/B^2(K)$ se calcule à partir des formules (1.39) à (1.42) et vaut respectivement :

$$A(K_1)/B^2(K_1) = 35.0 \ , \qquad (48)$$

$$A(K_2)/B^2(K_2) = (4\Pi)^{-1/2} \ . \qquad (49)$$

En substituant les constantes (47), (48) et (49) dans la formule de h_{opt} et en posant $k = h/\sigma$, on obtient le résultat désiré pour la fenêtre optimale de chaque noyau, à savoir :

$$k_1 = 2.778 \ n^{-1/5}$$

et

$$k_2 = 1.059 \ n^{-1/5} \ .$$

Cette fenêtre est appelée fenêtre ou largeur réduite et elle vérifie aussi les hypothèses (H4). Elle contient automatiquement le facteur d'échelle σ et le problème revient donc à estimer une densité normale $N(\mu,1;x)$.

L'EQMI asymptotique correspondant à la largeur optimale est obtenue de manière analogue en utilisant la formule asymptotique (1.36) et les constantes (1.39) à (1.42) et (47).

Pour étudier le comportement de l'$EQMI(f_1)$ en fonction de la largeur réduite k, les constantes (47), (48) et (49) sont substituées dans la formule asymptotique générale de l'EQMI(f) (1.27). La formule suivante, dépendante de k, est obtenue :

$$EQMI(f_1) = 7^{-1} \ \sigma^{-1} \ \{ \ 5/(nk) + 3/(224\sqrt{\Pi}) \ k^4 \}$$

$$+ \ o[(nk)^{-1}+k^4] . \qquad (50)$$

Cependant, pour l'estimateur f_2, il est possible de calculer explicitement et exactement le critère de l'EQMI. Le résultat obtenu dépend de la fenêtre réduite k. La formule a été calculée notamment par Anderson (1969), Guttmann et Wertz (1976) et Fryer (1976) et est déterminée dans le théorème suivant.

Théorème 2.2.2

Supposons que les hypothèses (H1) soient vérifiées.
Si le noyau normal est utilisé, alors la formule exacte de l'EQMI(f_2) vaut :

$$EQMI(f_2) = (4\Pi\sigma^2)^{-1/2} \{ 1 + n^{-1} [k^{-1} - (1+k^2)^{-1/2}] + (1+k^2)^{-1/2} - 2 [2/(2+k^2)]^{1/2} \} . \quad (51)$$

Démonstration

Les variables $X_1, X_2, \ldots, X_n$ étant i.i.d, l'estimateur $f_2(x)$ peut être considéré comme une somme de n variables aléatoires indépendantes, de la forme :

$$f_2(x) = n^{-1} \Sigma_i \; N(X_i, h^2; x)$$

Calculons les différents termes composants l'EQMI(f_2) en utilisant les formules (1.8) et (1.10) :

$$EQMI(f_2) = \int_{-\infty}^{+\infty} \{ Biais^2[f_2(x)] + Var_f[f_2(x)] \} dx .$$

Soit x un point donné;

1) $E_f[f_2(x)] = E_f[n^{-1} \Sigma_i \; N(X_i, h^2; x)]$

$$= E_f[N(y, h^2; x)] = \int_{-\infty}^{+\infty} N(y, h^2; x) \; N(\mu, \sigma^2; y) \; dy$$

$$= N(\mu,\sigma^2+h^2;x) \ .$$

Ainsi,

$$\text{Biais}[f_2(x)] = N(\mu,\sigma^2;x) - N(\mu,\sigma^2+h^2;x) \ ,$$

$$\text{Biais}^2[f_2(x)] = [N(\mu,\sigma^2;x) - N(\mu,\sigma^2+h^2;x)]^2$$

$$= (4\Pi\sigma^2)^{-1/2} \ N(\mu,\sigma^2/2;x) - 2 \ N(\mu,\sigma^2;x) \ N(\mu,\sigma^2+h^2;x)$$

$$+ [4\Pi(\sigma^2+h^2)]^{-1/2} \ N[\mu,(\sigma^2+h^2)/2;x]$$

$$= (4\Pi)^{-1/2} \ \{\sigma^{-1}N(\mu,\sigma^2/2;x) + (\sigma^2+h^2)^{-1/2}N[\mu,(\sigma^2+h^2)/2;x]$$

$$- 2\sqrt{2} \ (2\sigma^2+h^2)^{-1/2} \ N[\mu,\sigma^2(\sigma^2+h^2)/(2\sigma^2+h^2);x]\}. \quad (52)$$

En intégrant par rapport à x l'expression (52), on obtient le premier terme de l'EQMI(f_2) :

$$\int_{-\infty}^{+\infty} \text{Biais}^2[f_2(x)] \ dx = (4\Pi)^{-1/2} \ [\ \sigma^{-1} + (\sigma^2+h^2)^{-1/2}$$

$$- 2\sqrt{2}/(2\sigma^2+h^2)^{1/2} \ ] \ . \quad (53)$$

2) $\text{Var}_f[f_2(x)] = \text{Var}_f[n^{-1} \ \Sigma_i \ N(X_i,h^2;x)]$

$$= n^{-1} \ \text{Var}_f \ [N(y,h^2;x)]$$

$$= n^{-1} \ \{ \ E_f[N^2(y,h^2;x)] - [E_f[N(y,h^2;x)]]^2]$$

$$= n^{-1} \ \{ \int_{-\infty}^{+\infty} N^2(y,h^2;x) \ N(\mu,\sigma^2;x) \ dy - N^2(\mu,\sigma^2+h^2;x) \ \}$$

$$= n^{-1} \left\{ \int_{-\infty}^{+\infty} (4\Pi h^2)^{-1/2} \; N(y, h^2/2; x) \; N(\mu, \sigma^2; y) \; dy \right.$$

$$\left. - [4\Pi(\sigma^2+h^2)]^{-1/2} \; N[\mu, (\sigma^2+h^2)/2; x] \right\}$$

$$= (4\Pi)^{-1/2} \; n^{-1} \left\{ h^{-1} \; N(\mu, \sigma^2+h^2/2; x) \right.$$

$$\left. - (\sigma^2+h^2)^{-1/2} \; N[\mu, (\sigma^2+h^2)/2; x] \right\} . \qquad (54)$$

En intégrant l'expresssion (54) par rapport à x, on obtient le second terme de l'EQMI(f_2) :

$$\int_{-\infty}^{+\infty} Var_f[f_2(x)] \; dx = (4\Pi)^{-1/2} \; n^{-1} \left[h^{-1} - (\sigma^2+h^2)^{-1/2} \right]. \qquad (55)$$

En additionnant les expressions (53) et (55), on obtient l'expression désirée pour l'EQMI(f_2) :

$$EQMI(f_2) = (4\Pi)^{-1/2} \left\{ n^{-1} \left[h^{-1} - (\sigma^2+h^2)^{-1/2} \right] + \sigma^{-1} \right.$$

$$\left. + (\sigma^2+h^2)^{-1/2} - 2 \left[2/(2\sigma^2+h^2)\right]^{1/2} \right\}$$

$$= (4\Pi\sigma^2)^{-1/2} \left\{ 1 + n^{-1} \left[k^{-1} - (1+k^2)^{-1/2} \right] \right.$$

$$\left. + (1+k^2)^{-1/2} - 2 \left[2/(2+k^2)\right]^{1/2} \right\}$$

où $k = h/\sigma$.

L'EQMI dépend de deux variables, n et k. Il n'existe pas de valeur explicite minimisant cette dernière expression. Cependant, pour chaque taille d'échantillon, il est possible de déterminer numériquement une fenêtre minimisant l'EQMI(f_2). Si l'on cherche une fonction de la forme $k = C \cdot n^{-\alpha}$, on obtient alors la fenêtre de Fryer, par exemple :

$$k_{Fryer} = 1.31 \; n^{-0.205} . \qquad (56)$$

Cette valeur sert de référence dans la première partie de l'analyse de l'EQMI(f_2) (cf paragraphe 2.3.2.1). Dans le paragraphe 2.3.2.2, une analyse du comportement de l'EQMI(f_2) en fonction du choix de la fenêtre k est menée pour le noyau normal.

Les formules théoriques des EQMI ayant été établies pour chaque catégorie d'estimateurs, paramétriques et non-paramétriques, nous allons donc étudier maintenant le comportement de ces estimateurs.

2.3 Analyse des EQMI et des efficacités

Avant de comparer l'estimateur paramétrique à l'estimateur à noyau, il est utile d'analyser les performances de chaque catégorie d'estimateurs.
Le traitement analytique des formules trouvées n'est pas des plus aisés. Afin de faciliter ces études, les valeurs des EQMI ont été tabulées pour différentes tailles d'échantillons, et leurs graphes ont été tracés en fonction de la taille de l'échantillon.

2.3.1 Analyse de l'EQMI des estimateurs paramétriques

Pour chaque estimateur g_j, deux formules pour l'EQMI ont été données, la seconde étant une série de puissances de n^{-1}. Dans un premier temps, la différence entre chaque couple de valeurs a été analysée pour juger la qualité des séries, dont l'ordre de grandeur est plus parlant. Puis, l'EQMI exacte de g_1 selon (2.8) et les EQMI approchées de g_2 selon (2.20) et de g_3 selon (2.30) sont comparées entre elles. Sur la base de ces considérations, les remarques suivantes peuvent être formulées :

1. Les écarts entre les formules des EQMI et leurs séries respectives sont minimes et disparaissent pratiquement pour des valeurs de n supérieures à 20 (tableaux

2.1 et 2.2). Uniquement les graphes des EQMI des formules initiales (2.8), (2.20), (2.30) ont dès lors été tracés, les graphes des séries se confondant aux graphes précédents (figure 2.1). Les séries permettent néanmoins de retenir plus facilement l'ordre de grandeur des EQMI de chaque estimateur paramétrique.

2. Les séries des EQMI étant des séries de puissances de n^{-1}, l'EQMI des estimateurs g_1, g_2, g_3 décroît comme la fonction n^{-1}, c-à-d fortement pour des petites valeurs de n et plus lentement pour des grandes tailles d'échantillon (figure 2.1). Ainsi, un échantillon de grande taille permet une approximation sensiblement meilleure de la fonction de densité et la qualité de l'estimation est proportionnelle à la taille de l'échantillon : par exemple, pour n = 200, l'EQMI d'un estimateur g_j est quatre fois plus petite que pour n = 50.

3. Lorsque l'espérance ou la variance est inconnue, les EQMI respectives sont voisines, avec un avantage pour g_1 si n est petit, $n \leq 14$, et pour g_2 dès que $n \geq 15$ (figure 2.1). Il est donc légèrement plus facile d'estimer une densité normale lorsque l'espérance est connue que dans le cas où la variance est connue. En se représentant les deux estimateurs, on comprend qu'une erreur d'estimation de l'espérance entraîne une translation de toute la courbe, ce qui influence davantage l'EQMI qu'une contraction ou une dilatation de la courbe autour de son mode (erreur d'estimation de σ^2).

4. Si l'espérance et la variance sont inconnues, l'EQMI de g_3 selon (2.30) est environ le double de celle des deux autres estimateurs. Il est donc intéressant de pouvoir connaître au moins un des paramètres afin d'obtenir une meilleure estimation de la fonction de densité, ou afin d'utiliser un échantillon de taille plus petite pour une erreur du même ordre (tableau 2.2 et figure 2.1).

En résumé,

les séries de puissances des EQMI des estimateurs paramétriques décrivent correctement les EQMI correspondantes et la connaissance d'un paramètre réduit l'EQMI de g_1 ou de g_2 de moitié par rapport à l'EQMI(g_3).

2.3.2 Analyse de l'EQMI des estimateurs non-paramétriques

2.3.2.1 EQMI pour une fenêtre donnée

Les coefficients des formules asymptotiques optimales (2.45) et (2.46) étant très proches pour les deux noyaux, seule la formule asymptotique du noyau biquadratique (2.45) a été retenue pour l'analyse de la formule asymptotique, car elle est la plus précise des deux. Pour le noyau normal, la formule exacte (2.51) a été étudiée avec la fenêtre réduite proposée par Fryer (2.56). Les conclusions sont les suivantes :

1. L'EQMI exacte de f_2 est toujours nettement inférieure à l'EQMI asymptotique de f_1 (tableau 2.3 et figure 2.2). Le rapport entre les deux EQMI varie de la manière suivante :

n	10	20	50	100
$\frac{EQMI(f_1)}{EQMI(f_2)}$	2.06	1.80	1.56	1.42

Deux remarques principales découlent de cette première constatation :

1.1 Puisque les deux formules asymptotiques (2.45) et (2.46) sont équivalentes, on en déduit que la formule asymptotique de l'EQMI pour le noyau normal (2.46) est mauvaise . Elle ne donne qu'une approximation très grossière de l'EQMI(f_2) pour la densité normale.

1.2 Sur la base des valeurs de l'EQMI asymptotique de f_1 et de l'EQMI exacte de f_2, le noyau normal semble apparemment préférable au noyau biquadratique. Mais, les simulations (cf paragraphe 2.4) démontreront que la formule asymptotique pour le noyau biquadratique est elle aussi très grossière et même mauvaise, et que les noyaux biquadratique et normal obtiennent finalement des résultats équivalents.

Ainsi, les remarques énoncées pour le noyau normal peuvent donc être considérées aussi valables pour le noyau biquadratique. L'analyse détaillée de l'EQMI asymptotique et de l'efficacité du noyau biquadratique n'est donc pas opportune et a été supprimée.

2. L'EQMI exacte de f_2 est une fonction fortement décroissante pour des valeurs de n petites, $n \leq 20$, puis cette décroissance s'atténue; elle est inférieure à $n^{-4/5}$ (figure 2.2). Un échantillon de taille élevée permet donc d'améliorer la qualité et l'exactitude de l'estimateur non-paramétrique f_2, mais dans une moindre mesure que dans la situation paramétrique (figure 2.2).

2.3.2.2 EQMI du noyau normal pour différentes fenêtres

L'EQMI exacte de f_2 est une fonction dépendante non seulement de la taille de l'échantillon n, mais aussi de la largeur réduite k du noyau. La formule analysée auparavant est qualifiée d'optimale, la fenêtre de Fryer devant minimiser l'EQMI exacte de cet estimateur.

Mais, comment varie l'EQMI si le statisticien utilise une largeur non-optimale? Pour déterminer les conséquences d'un tel changement sur l'EQMI exacte de f_2, d'autres valeurs ont été attribuées à la fenêtre et substituées dans la formule de l'EQMI exacte du noyau normal (2.51). Rappelons que :

$$k_{Fryer} = 1.31\, n^{-0.205}.$$

Les conclusions sont les suivantes :

1. Des EQMI légèrement inférieures peuvent être atteintes, si la fenêtre est plus petite que celle indiquée par Fryer. Mais les différences des EQMI sont minimes, voire à peine visibles sur les graphes correspondants des EQMI (figure 2.3). La valeur "optimale" de la fenêtre se situe entre la valeur de Fryer et la valeur asymptotique de k_2 (2.44).

2. Un choix proche de la largeur de Fryer n'altère pas les résultats d'une manière conséquente. Néanmoins, on constate qu'une forte sous-évaluation de la fenêtre "optimale" influence moins l'EQMI qu'une nette surévaluation de celle-ci; mais les écarts ne sont pas énormes et ne portent pas à conséquence pour l'interprétation des résultats (figure 2.3).

En résumé,

la fenêtre minimisant l'EQMI exacte de f_2 est très légèrement inférieure à celle proposée par Fryer, mais une largeur proche de la largeur optimale n'altère pas de manière sensible l'EQMI exacte de f_2.
L'EQMI exacte de f_2 s'améliore si la taille de l'échantillon augmente.
La formule asymptotique de l'EQMI de f_2 est mauvaise pour une densité normale.

2.3.3 Analyse des efficacités

La comparaison proprement dite entre les deux classes d'estimateurs peut être envisagée, les EQMI de chaque estimateur ayant été présentées séparément et analysées d'une manière complète. Les formules (2.8), (2.20) et (2.30) des EQMI des estimateurs paramétriques g_1, g_2 et g_3 sont confrontées à l'EQMI exacte de f_2 (2.51) et la fenêtre de Fryer (2.56).

Pour chaque estimateur paramétrique g_j, l'efficacité du noyau normal, notée $eff(g_j, f_2)$, a été calculée numériquement et son graphe a été tracé (figure 2.5).

$$eff(g_j, f_2) = \frac{EQMI(g_j)}{EQMI(f_2)} \qquad j = 1,2,3$$

Comment peut-on interpréter une efficacité inférieure à 1.0 ? Dit simplement, une telle efficacité signifie que :
- l'estimateur g_j est meilleur que f_2, au sens de l'EQMI (chapitre 1, définition 1.3.2).
- pour obtenir une précision donnée, il faut moins d'observations pour l'estimateur g_j que pour l'estimateur f_2, et la proportion est déterminée par la valeur de l'efficacité : par exemple, si l'efficacité vaut 0.5, il faudrait le double d'observations avec f_2, pour obtenir une EQMI du même ordre que celle de g_j.

Les résultats sont analysés de la manière suivante : on détermine tout d'abord quel estimateur est le plus efficace dans chacune des trois situations possibles entre les estimateurs paramétriques et non-paramétrique (figure 2.5). Puis l'on quantifie plus finement ces résultats en étudiant les valeurs des efficacités (tableau 2.4).

Les remarques suivantes peuvent être formulées :

1. Quel que soit l'estimateur paramétrique utilisé, g_j, l'$EQMI(g_j)$ est presque toujours inférieure à l'$EQMI(f_2)$ (figure 2.4). Ainsi, l'efficacité est toujours inférieure à 1.0, sauf pour l'estimateur g_3 si la taille de l'échantillon est inférieure à 15. Les estimateurs paramétriques g_j sont donc toujours préférables aux estimateurs non-paramétriques (tableau 2.4 et figure 2.5).

2. Le comportement des efficacités par rapport aux estimateurs paramétriques est semblable et seul le pire des cas, g_3, est analysé plus précisément. L'efficacité décroît assez rapidement jusqu'à $n = 50$, puis plus modérément. Elle est de l'ordre de $n^{-1/5}$ environ. Un grand

nombre d'observations permet donc d'abaisser l'efficacité du noyau normal aussi proche de zéro qu'on le désire; mais le prix est chèrement payé. A titre indicatif, une efficacité de 0.5 est atteinte pour un échantillon de taille 60 environ (tableau 2.4).

3. Si l'efficacité est de 0.50, il faut en réalité plus du double d'observations pour que l'estimateur f_2 arrive à la précision de g_3, car la décroissance de l'EQMI(g_3) est plus rapide que celle de l'EQMI(f_2) ce qui favorise encore davantage l'estimateur g_3 (tableaux 2.2 et 2.3).

4. La qualité de l'efficacité dépend du nombre des paramètres à estimer. Lorsque les deux paramètres sont inconnus, c-à-d g_3, l'efficacité double pratiquement par rapport à celle de g_1 ou de g_2.

Exemple : (tiré du tableau 2.4)

n	20	50	100	200
eff(g_3,f_2)	0.84	0.57	0.44	0.36
eff(g_1,f_2)	0.43	0.31	0.25	0.20

Lorsqu'un des paramètres est connu, alors l'efficacité est pratiquement toujours inférieure à 0.5, et diminue encore pour n augmentant (tableau 2.4).
Ainsi, l'estimateur paramétrique accentue encore son avantage et est donc nettement préféré à l'estimateur à noyau normal. Pour les estimateurs g_1 et g_2, l'efficacité vaut 0.5 pour des échantillons de taille 20 déjà!

En résumé,

dans le cas étudié dans ce chapitre, la densité est normale, l'estimateur paramétrique g_1, est préférable à un estimateur à noyau, biquadratique ou normal, et dans une très forte mesure.

2.4 Simulations

Le but de ce paragraphe est non seulement de confirmer les résultats théoriques pour les estimateurs paramétriques et pour l'estimateur à noyau normal, mais aussi de démontrer que la formule asymptotique de l'EQMI(f_1) (2.45) est très approximative et que, de plus, le noyau n'est pas prédominant dans l'EQMI des estimateurs non-paramétriques.
Une estimation non biaisée de l'EQMI d'un estimateur est la moyenne des écarts quadratiques discrets sur 50 échantillons par exemple, notée EQDM.

Définitions 2.4.1

Soit f_n un estimateur d'une fonction de densité. Alors, l'écart quadratique discret de f_n est défini par :

$$EQD(f_n) = \sum_{j=1}^{300} [f_n(x_j) - f(x_j)]^2$$

où les x_j représentent une discrétisation en 300 points équidistants du support de f.

La moyenne des $EQD(f_n)$, notée EQDM, est définie par :

$$EQDM(f_n) = 50^{-1} \sum_{l=1}^{50} EQD(f_{nl})$$

où f_{nl} représente une estimation de la fonction f.

De manière analogue, on définit l'écart-type des $EQD(f_n)$ noté EQDS :

$$EQDS = \{ 49^{-1} \sum_{l=1}^{50} [EQD(f_{nl}) - EQDM(f_n)]^2 \}^{1/2} .$$

Remarque

L'indice l varie de 1 à 50 et sera omis s'il n'y a aucune confusion.

Les échantillons ont été générés à l'aide du générateur de nombres pseudo-aléatoires de "Mathematical Software ACM vol.2 p.132-138" et les observations d'une répartition de Gauss ont été obtenues à l'aide des fonctions congruentes. Pour la densité normale, le support a été réduit à l'intervalle [-5.0, 5.0]. Les tailles des échantillons retenues sont 20, 50 et 100.

Pour chaque échantillon, les estimateurs m , u et v ont été calculés et substitués dans la formule de la densité normale, tandis que les estimateurs non-paramétriques ont été déterminés en tous points de discrétisation x_j.

Le critère de l'EQD a été calculé pour chaque échantillon et pour chaque estimateur , ainsi que leur moyenne arithmétique et leur écart-type sur les 50 échantillons générés. Les résultats sont analysés en trois parties :

premièrement, la comparaison des EQDM et des EQMI de chaque estimateur; ensuite, l'influence d'un choix différent de la largeur de Fryer sur l'EQDM de l'estimateur non-paramétrique f_2; enfin, le comportement des efficacités simulées.

2.4.1 Analyse de l'EQDM des estimateurs paramétriques

Pour calculer l'EQDM de chaque estimateur paramétrique, g_1, g_2 et g_3, un autre ensemble d'échantillons a été généré, si bien que trois valeurs de l'EQDM sont déterminées pour les estimateurs à noyau (tableau 2.6).

Les valeurs des EQDM sont toutes légèrement inférieures aux valeurs théoriques correspondantes, à une exception près (n = 20, μ et σ^2 inconnnus). Mais, dans les deux cas, les écarts ne sont pas conséquents. Par contre, les écarts-types peuvent être qualifiés d'élevés, car ils sont supérieurs à l'EQDM ce qui indique que de grandes fluctuations apparaissent entre les EQD (tableau 2.5).

2.4.2 Analyse de l'EQDM des estimateurs non-paramétriques

2.4.2.1 EQDM pour une fenêtre donnée

Pour le noyau biquadratique, la fenêtre minimisant l'EQMI asymptotique (2.53) a été retenue pour cette étude.
Pour cet estimateur, de très grandes différences sont remarquées entre les valeurs théoriques et simulées, différences ne pouvant être expliquées uniquement par l'échantillonnage. Les EQDM sont très nettement inférieurs aux EQMI, dans un rapport de 2.0 environ pour $n = 20$ et $n = 50$, et de 1.5 pour $n = 100$. Les écarts-types sont eux plus petits que les EQDM, ce qui montre une bonne stabilité de l'EQD. Ces simulations démontrent que la formule asymptotique (1.36) n'est pas bonne pour le noyau biquadratique si la densité que l'on estime est normale (tableau 2.6 et figure 2.6).

Pour le noyau normal, la fenêtre choisie est celle proposée par Fryer (2.56).
Sauf dans deux cas, les valeurs des EQDM sont inférieures aux valeurs correspondantes des EQMI; mais, à nouveau, les écarts obtenus sont acceptables et peuvent être attribués à des fluctuations d'échantillonnage. Les écarts-types sont eux nettement inférieurs aux EQDM, sauf dans un cas où il est semblable à l'EQDM ce qui indique que les EQD ne varient pas énormément d'un échantillon à l'autre (tableau 2.6).

Etant donné les différences entre les valeurs théoriques et simulées pour l'estimateur f_1, il est intéressant de le comparer à nouveau à l'estimateur f_2. Alors, la constation suivante peut être formulée :

> Le noyau biquadratique obtient des EQDM inférieurs à ceux du noyau normal. Les différences restent toutefois négligeables, le rapport des EQDM étant proche de 0.95, valeur identique à celle des EQMI asymptotiques (2.45) et (2.46). Les écarts-types sont pratiquement du même ordre de grandeur pour les deux noyaux.

Ainsi, le noyau biquadratique est légèrement meilleur que le noyau normal pour estimer une densité normale, même si les écarts sont minimes.

2.4.2.2 EQDM pour différentes fenêtres

Pour chaque échantillon, quatre valeurs ont été attribuées à la largeur du noyau, afin d'analyser les effets sur les EQDM des estimateurs choisis : une fenêtre étroite, la fenêtre asymptotique, la fenêtre de Fryer et une fenêtre plus large. Le rapport entre deux fenêtres consécutives est constant. D'autres échantillons ont été générés pour cette étude, si bien que d'autres valeurs de l'EQDM sont obtenues pour des largeurs identiques à celles du paragraphe précédent.

Pour le noyau normal, les EQDM sont proches des valeurs théoriques, parfois supérieurs et parfois plus petits, mais jamais d'une manière excessive. Pour une fenêtre plus étroite ou plus large que celle de Fryer, l'EQDM de f_2 s'éloigne très légèrement de la valeur optimale. L'EQDM de f_2 est peu sensible à une fenêtre nettement trop étroite ou trop large, tandis que l'EQDS de f_2 augmente notablement dans le premier cas et diminue dans le deuxième cas. La fenêtre de Fryer obtient des résultats voisins des meilleurs possibles et leurs différences sont négligeables (tableau 2.7).

Pour le noyau biquadratique, les valeurs obtenues par simulations sont nettement inférieures aux valeurs théoriques et confirment l'inexactitude de la formule asymptotique. La fenêtre asymptotique donne toujours les meilleurs résultats; une fenêtre légèrement supérieure est moins appropriée, mais les différences ne sont pas très marquées. Par contre, l'EQDM se détériore notablement pour une fenêtre étroite et encore davantage pour une fenêtre large (tableau 2.7).

Pour des fenêtres équivalentes, l'EQDM du noyau biquadratique est inférieur à celui du noyau normal, sauf si

la fenêtre est trop large (tableau 2.7). Ainsi, l'estimateur à noyau biquadratique f_1 est préférable à l'estimateur à noyau normal f_2, en remarquant que les différences entre eux sont minimes.

En résumé,

la formule asymptotique pour les noyaux, (2.45) et (2.46), est très grossière et n'est pas bonne pour estimer l'EQMI d'une densité normale.
De plus, les noyaux sont équivalents, avec un léger avantage au noyau biquadratique s'il en est un.
Une fenêtre pas trop éloignée des formules asymptotiques et de Fryer n'influence pas de manière conséquente les EQDM.

2.4.3 Analyse des efficacités simulées

La moyenne des efficacités n'est pas une bonne caractéristique de la tendance centrale. Pour estimer l'efficacité théorique, on utilise le quotient des EQDM des estimateurs considérés, appelé efficacité simulée et abrégé effsim.

$$\text{effsim}(g_k, f_j) = \frac{\text{EQDM}(g_k)}{\text{EQDM}(f_j)} = \frac{\Sigma_1 \ \text{EQD}(g_{k1})}{\Sigma_1 \ \text{EQD}(f_{j1})} \qquad \begin{matrix} k = 1,2,3 \\ j = 1,2 \end{matrix}$$

Les EQDM du noyau biquadratique étant trop éloignés des EQMI théoriques correspondants, une comparaison entre les efficacités simulées et théoriques pour ce noyau n'est pas judicieuse. L'efficacité simulée du noyau normal est tout d'abord comparée à son efficacité théorique, puis à l'efficacité simulée du noyau biquadratique.

2.4.3.1 Efficacité du noyau normal

Les efficacités simulées sont très proches des efficacités théoriques, voire égales, sauf dans trois cas (tableau 2.8) :
- lorsque l'espérance est inconnue et n = 20, alors l'efficacité simulée est nettement inférieure à l'efficacité théorique dans un rapport de 1.25.
- lorsque l'espérance et la variance sont inconnues, la qualité de l'efficacité simulée dépend de la taille de l'échantillon. En effet, si n = 20, l'efficacité simulée est supérieure à l'efficacité théorique, tandis que si n = 100, la remarque contraire est correcte. Pour n = 50, les valeurs simulées et théoriques correspondent.

Les exceptions s'expliquent parce que les EQDM des estimateurs réagissent dans ces cas de manière oppposée, c-à-d que l'EQDM est plus petit que l'EQMI pour l'un des estimateurs et l'inégalité est inversée pour l'autre estimateur. Ainsi, ces différences ressortent nettement lorsque le quotient des EQDM est effectué.

De plus, lorsque la variance est inconnue, les efficacités sont de même grandeur, tout en constatant que l'efficacité simulée est toujours supérieure à la valeur théorique correspondante, car les différences entre les EQMI et les EQDM sont un peu plus marquées pour l'estimateur non-paramétrique f_2 que pour l'estimateur paramétrique g_2.

D'une manière globale, les efficacités simulées correspondent aux résultats théoriques.

2.4.3.2 Comparaison entre les efficacités des noyaux biquadratique et normal

Comparons les noyaux biquadratique et normal en analysant leur efficacité simulée.

L'efficacité simulée de f_1 est toujours inférieure à celle de f_2, mais les écarts sont très proches et peuvent être qualifiés de négligeables. Ce résultat est logique, puisque l'EQDM(f_1) est toujours inférieur à l'EQDM(f_2).
Ceci démontre, s'il en est encore besoin, que le noyau biquadratique est légèrement plus apte à estimer une densité normale que le noyau normal.

Conclusion

Sur la base des résultats théoriques et des simulations, on peut donc affirmer que l'estimateur paramétrique est plus apte à estimer une densité normale, même si les deux paramètres sont inconnus.
Quant au noyau à utiliser, un léger avantage est donné au noyau biquadratique, bien que leur comportement soit identique et proche.

Mais, ces résultats sont obtenus avec des hypothèses fortes, la densité à estimer est normale.

Alors, une question immédiate se pose : quelles sont les conséquences si la densité à estimer n'est pas tout à fait une densité normale? Le chapitre 3 apportera des réponses à cette question.

CHAPITRE 3 : ESTIMATION D'UNE DENSITE NORMALE CONTAMINEE PAR UNE DENSITE NORMALE

Introduction

Dans ce chapitre, nous nous intéressons à l'estimation des fonctions de densité appartenant à la famille $\mathcal{P}$ des densités normales contaminées par une autre densité normale. Ce phénomène se rencontre couramment dans la nature dans des domaines aussi nombreux que variés : en économie (Quandt), en médecine (Brownie et Robson), en biologie (Bhattacharia), en ingéniérie (Yong).

Deux interprétations différentes de la contamination peuvent être données.

Premièrement, des individus étrangers à la population mère et perturbateurs se sont glissés dans l'échantillon, individus provenant d'une autre répartition supposée normale. Deux cas peuvent se présenter :

1. On constate des valeurs éloignées de la tendance centrale; on parle alors de valeurs aberrantes. Ce cas se présente lorsque les moyennes théoriques des répartitions considérées diffèrent passablement (figure 3.21).

2. La contamination est "cachée", c-à-d que les moyennes théoriques sont proches. On aura soit un aplatissement de la courbe de la fonction de densité, soit une contraction de celle-ci suivant le rapport de leur variance (figure 3.22).

Deuxièmement, on peut aussi considérer une contamination comme une famille plus étendue, plus large qu'uniquement la famille des répartitions normales, permettant ainsi de traiter des échantillons d'espèces différentes. Rien, en effet, ne permet d'affirmer que telle valeur n'appartienne pas à l'échantillon considéré, même si elle paraît douteuse pour certains statisticiens. Ainsi, l'al-

longement de la queue de la répartition, ou l'apparition d'une bosse, ou encore l'écrasement ou l'accentuation du mode peuvent être une caractéristique propre et particulière de la population étudiée.

Décrivons le modèle mathématique associé à une répartition normale contaminée par une autre répartition normale.

Le taux de contamination est noté ε, avec $0 \leq \varepsilon \leq 1$.

Soient $Y_1, Y_2, \ldots, Y_n$, n variables aléatoires réelles suivant la répartition F_1 avec une probabilité de $1-\varepsilon$, et la répartition F_2 avec une probabilité de ε, c-à-d si U est une variable uniforme [0,1], alors :

$$F_{Y/U}(y/u) = \begin{cases} F_1(y) & \text{si } u \geq \varepsilon \\ F_2(y) & \text{si } u < \varepsilon \end{cases} \qquad (1)$$

où $F_{Y/U}$ désigne la répartition conditionnelle de Y connaissant U.

Une variable X provenant de ce modèle a pour répartition la fonction F définie par :

$$F(x) = (1-\varepsilon)\, F_1(x) + \varepsilon\, F_2(x) \ . \qquad (2)$$

Si F_i est absolument différentiable, alors la fonction de densité de la variable X, notée f, est définie par :

$$f(\tau_1, \tau_2; x) = (1-\varepsilon)\, f_1(\tau_1; x) + \varepsilon\, f_2(\tau_2; x) \qquad (3)$$

où f_i représente la fonction de densité associée à F_i,
τ_i le vecteur des paramètres de la densité f_i,
x un point fixé.

Comme nous ne traitons que des densités normales, la famille $\mathcal{P}$ est donc définie par :

$\mathbb{P} = \{ f = (1-\varepsilon)\, f_1 + \varepsilon\, f_2 \ , \ f_i$ densité normale $\}$

où $f_1(\tau_1;x) = N(\mu,\sigma^2;x)$ et $f_2(\tau_2;x) = N(\alpha,\beta^2;x)$.

L'espérance et la variance de X se déterminent aisément et valent respectivement :

$$E_f(X) = (1-\varepsilon)\,\mu + \varepsilon\,\alpha$$

et

$$Var_f(X) = (1-\varepsilon)\,\sigma^2 + \varepsilon\,\beta^2 + (1-\varepsilon)\,\varepsilon\,(\mu-\alpha)^2 \ .$$

Deux remarques doivent être dites au sujet de ce modèle:

1) Il est évident que si $\varepsilon > 0.5$, la répartition perturbatrice est en fait la répartition perturbée; ainsi, on suppose dans ce travail que $\varepsilon \leq 0.5$.
2) Sans restreindre la généralité, on pourrait poser $\mu = 0.0$ et $\sigma^2 = 1.0$.

   En effet, il suffit alors d'effectuer la transformation affine suivante:

   $$Z = (X - \mu)/\sigma \ .$$

   On obtient alors:

   $$f(z) = (1-\varepsilon)\, N(0,1;z) + \varepsilon\, N(d,r^2;z)$$

   où $d = (\alpha-\mu)/\sigma$ et $r^2 = \beta^2/\sigma^2$.

   d correspond à la différence réduite des espérances. Néanmoins, les formules des EQMI des estimateurs seront explicitées pour la forme la plus générale de contamination.

La fonction de densité que l'on estime au cours de ce chapitre est la fonction f. Elle vérifie les hypothèses (H2) du chapitre 1.

Nous allons procéder de la même manière qu'au chapitre 2; les EQMI de chaque estimateur sont tout d'abord calculées, puis analysées séparément. Ensuite, les deux méthodes sont comparées entre elles à l'aide des résultats théoriques et des simulations menées parallèlement.
Enfin, sur la base des valeurs et des graphes des taux critiques de contamination, des règles pratiques peuvent être déduites pour déterminer lequel des estimateurs est le plus approprié.

3.1 EQMI de l'estimateur paramétrique

Le but n'est pas d'estimer tous les paramètres intervenant dans le modèle présenté. En fait, on suppose que le statisticien ignore totalement l'existence d'une contamination. Il estime donc la moyenne arithmétique $\mathbf{m}_n$ et la substitue dans l'expression de la densité normale f_1. La variance de f_1, σ^2, est supposée connue. Ainsi, l'estimateur paramétrique g_n de la fonction f au point x est défini par :

$$g_n(x) = N(\mathbf{m}_n, \sigma^2; x) \qquad \text{où} \qquad \mathbf{m}_n = n^{-1} \sum_{i=1}^{n} X_i \ . \tag{4}$$

Conventions d'écriture

Si aucune confusion n'est possible, l'indice n est omis dans g_n et $\mathbf{m}_n$.
De plus, $\mathbf{m}$ écrit en caractères gras désigne l'estimateur, tandis que dans les formules, m représente une estimation.
Les bornes de Σ sont omises, si elles varient de 1 à n.

Avant de déterminer l'EQMI(g_n), nous allons tout d'abord calculer la fonction de densité de $\mathbf{m}$, car elle est utilisée dans les calculs de l'EQMI(g).

Lemme 3.1.1

Supposons que les hypothèses (H1) soient vérifiées.
Alors la fonction de densité de m, notée l(m), vaut :

$$l(m) = \Sigma_j \binom{n}{j} (1-\varepsilon)^j \varepsilon^{n-j} N(a_j, b_j; m) \qquad (5)$$

où $a_j = [\, j\cdot\mu + (n-j)\alpha \,]/n$, (6)

$b_j = [\, j\cdot\sigma^2 + (n-j)\beta^2]/n^2$, (7)

$\binom{n}{j} = n!/[(n-j)!\, j!]$ coefficient binomial.

Démonstration

Les calculs sont basés sur les propriétés des fonctions caractéristiques pour des variables i.i.d.
Désignons par $\Phi_X(t)$, $\Phi_i(t)$ les fonctions caractéristiques de la variable X et des densités normales f_i.
Alors,

$$\Phi_X(t) = (1-\varepsilon)\, \Phi_1(t) + \varepsilon\, \Phi_2(t) .$$

Si $Z = \Sigma_j X_j$, alors

$$\Phi_Z(t) = [\Phi_X(t)]^n$$

$$= \Sigma_j \binom{n}{j} (1-\varepsilon)^j \varepsilon^{n-j} \Phi_1{}^j(t)\cdot\Phi_2{}^{n-j}(t) .$$

Or $\Phi_1{}^j(t)\cdot\Phi_2{}^{n-j}(t)$ est la fonction caractéristique d'une densité normale de paramètres $j\cdot\mu+(n-j)\alpha$ et $j\cdot\sigma^2+(n-j)\beta^2$.

Par simple changement de variable, on obtient la fonction de densité de m, l(m); elle sera écrite de la façon suivante :

$$l(m) = \Sigma_j\, c_j(\varepsilon)\, N(a_j, b_j; m) \qquad (8)$$

où $c_j(\varepsilon) = \binom{n}{j} (1-\varepsilon)^j \varepsilon^{n-j}$. (9)

Théorème 3.1.2

Supposons que les hypothèses (H1) soient vérifiées, que l'espérance soit estimée par m et que la variance σ^2 soit connue. Alors, l'estimateur paramétrique g est défini par :

$$g(x) = N(m,\sigma^2;x) \qquad \text{où} \qquad m = n^{-1} \Sigma_i X_i \ .$$

L'EQMI de g vaut :

$$\begin{aligned} EQMI(g) = (4\Pi\sigma^2)^{-1/2}\{ & \varepsilon^2 [1 + 1/r - 2\sqrt{2}\, P(d,1+r^2;0)] \\ & - 2\varepsilon [1 - \sqrt{2}\, P(d,1+r^2;0) \\ & \qquad - \sqrt{2}\, \Sigma_j c_j(\varepsilon)\, P((n-j)d/n,2+b_j';0) \\ & \qquad + \sqrt{2}\, \Sigma_j c_j(\varepsilon)\, P(j\cdot d/n,1+r^2+b_j';0)] \\ & + 2 [1 - \sqrt{2}\, \Sigma_j c_j(\varepsilon)\, P((n-j)d/n,2+b_j';0)]] \end{aligned} \qquad (10)$$

où $P(u,v;x) = (2\Pi)^{1/2}\, N(u,v;x)$, (11)

$b_j' = [j + (n-j)r^2]/n^2$, (12)

$d = (\alpha - \mu)/\sigma$, (13)

$r^2 = \beta^2/\sigma^2$. (14)

Démonstration

La démonstration est identique à celle du théorème 2.1.1 avec des formules plus fastidieuses, la densité de m étant une somme pondérée de densités normales.
Calculons les différents termes de l'EQMI(g) selon les formules (1.9) et (1.10).
Pour un point donné x, on sait que :

$$f(\tau_1,\tau_2;x) = (1-\varepsilon)\, N(\mu,\sigma^2;x) + \varepsilon\, N(\alpha,\beta^2;x) \ .$$

1) $E_f[g^2(x)] = E_m[N^2(m,\sigma^2;x)]$

$= E_m[(4\Pi\sigma^2)^{-1/2}\ N_x(m,\sigma^2/2)]$

$= \int_{-\infty}^{+\infty} (4\Pi\sigma^2)^{-1/2}\ N(m,\sigma^2/2;x)\ [\ \Sigma_j\ c_j(\varepsilon)\ N(a_j,b_j;m)\ ]\ dm$

$= (4\Pi\sigma^2)^{-1/2}\ \Sigma_j\ c_j(\varepsilon)\ N(a_j,\sigma^2/2+b_j;x)\ .$

Alors, en intégrant cette dernière expression par rapport à x, on obtient le premier terme de l'EQMI(g);

$$\int_{-\infty}^{+\infty} E_f[g^2(x)]\ dx = (4\Pi\sigma^2)^{-1/2}\ \Sigma_j\ c_j(\varepsilon) \int_{-\infty}^{+\infty} N(a_j,\sigma^2/2+b_j;x)\ dx$$

$$= (4\Pi\sigma^2)^{-1/2}\ \Sigma_j\ c_j(\varepsilon)$$

$$= (4\Pi\sigma^2)^{-1/2}\ . \qquad (15)$$

2) $f^2(\tau_1,\tau_2;x) = (1-\varepsilon)^2\ N^2(\mu,\sigma^2;x) + \varepsilon^2\ N^2(\alpha,\beta^2;x)$

$+ 2\ (1-\varepsilon)\ \varepsilon\ N(\mu,\sigma^2;x)\ N(\alpha,\beta^2;x)$

$= (1-\varepsilon)^2\ (4\Pi\sigma^2)^{-1/2}\ N(\mu,\sigma^2/2;x)$

$+ \varepsilon^2\ (4\Pi\beta^2)^{-1/2}\ N(\alpha,\beta^2/2;x)$

$+ 2\ (1-\varepsilon)\ \varepsilon\ N(\alpha,\sigma^2+\beta^2;\mu)$

$\cdot\ N[(\mu\beta^2+\alpha\sigma^2)/(\sigma^2+\beta^2),\ \sigma^2\beta^2/(\sigma^2+\beta^2);x]\ .$

Alors, en intégrant cette expression par rapport à x, le second terme de l'EQMI(g) est déterminé;

$$\int_{-\infty}^{+\infty} f^2(\tau_1,\tau_2;x)\,dx = (1-\varepsilon)^2\,(4\Pi\sigma^2)^{-1/2} + \varepsilon^2\,(4\Pi\beta^2)^{-1/2} + 2\,(1-\varepsilon)\,\varepsilon\,N(\alpha,\sigma^2+\beta^2;\mu)$$

$$= (4\Pi)^{-1/2}\,[\,(1-\varepsilon)^2\,\sigma^{-1} + \varepsilon^2\,\beta^{-1}\,] + 2\,(1-\varepsilon)\,\varepsilon\,N(\alpha,\sigma^2+\beta^2;\mu)\ . \qquad (16)$$

$$3)\ E_f[g(x)] = E_m[g(x)] = \int_{-\infty}^{+\infty} N(m,\sigma^2;x)\,[\Sigma_j c_j(\varepsilon)N(a_j,b_j;m)]\,dm$$

$$= \Sigma_j\ c_j(\varepsilon)\ N(a_j,\sigma^2+b_j;x)\ .$$

$$f(\tau_1,\tau_2;x)\ E_f[g(x)] = [\,(1-\varepsilon)\ N(\mu,\sigma^2;x) + \varepsilon\ N(\alpha,\beta^2;x)] \cdot [\ \Sigma_j\ c_j(\varepsilon)\ N(a_j,\sigma^2+b_j;x)]$$

$$= (1-\varepsilon)\ [\ \Sigma_j\ c_j(\varepsilon)\ N(a_j,\sigma^2+b_j;x)]\ N(\mu,\sigma^2;x) + \varepsilon\ [\cdot\ \Sigma_j\ c_j(\varepsilon)\ N(a_j,\sigma^2+b_j;x)]\ N(\alpha,\beta^2;x)]$$

Alors, le dernier terme de l'EQMI(g) s'obtient par intégration de cette dernière expression;

$$\int_{-\infty}^{+\infty} f(\tau_1,\tau_2;x)\ E_f[g(x)]\,dx = (1-\varepsilon)\ \Sigma_j\ c_j(\varepsilon)\ N(a_j,2\sigma^2+b_j;\mu) + \varepsilon\ \Sigma_j\ c_j(\varepsilon)\ N(a_j,\sigma^2+\beta^2+b_j;\alpha)$$

$$= \Sigma_j\ c_j(\varepsilon)\ [\ (1-\varepsilon)\ N(a_j,2\sigma^2+b_j;\mu) + \varepsilon\ N(a_j,\sigma^2+\beta^2+b_j;\alpha)]\ . \qquad (17)$$

En regroupant les expressions (15), (16) et (17) selon les formules (1.9) et (1.10), on obtient le résultat désiré :

$$\begin{aligned} EQMI(g) = (4\Pi)^{-1/2} \{ & \sigma^{-1} + (1-\varepsilon)^2 \sigma^{-1} + \varepsilon^2 \beta^{-1} \} \\ & + 2 (1-\varepsilon) \varepsilon \mathbf{N}(\alpha, \sigma^2+\beta^2; \mu) \\ & - 2 \Sigma_j c_j(\varepsilon) [(1-\varepsilon) \mathbf{N}(a_j, 2\sigma^2+b_j; \mu) \\ & \qquad + \varepsilon \mathbf{N}(a_j, \sigma^2+\beta^2+b_j; \alpha)] . \quad (18) \end{aligned}$$

Reparamétrisons l'EQMI obtenue à l'aide des variables suivantes :

$$d = (\alpha-\mu)/\sigma \qquad \text{et} \qquad r^2 = \beta^2/\sigma^2 .$$

De plus, posons :

$$\begin{aligned} \mathbf{P}(u,v;x) &= (2\Pi)^{1/2} \mathbf{N}(u,v;x) \\ &= v^{-1/2} \exp[-(u-x)^2/(2v)] \quad , v > 0 . \end{aligned}$$

L'EQMI(g) s'écrit alors :

$$\begin{aligned} EQMI(g) = (4\Pi\sigma^2)^{-1/2} \{ & \varepsilon^2 [1 + 1/r - 2\sqrt{2}\, \mathbf{P}(d, 1+r^2; 0)] \\ & - 2 \varepsilon [1 - \sqrt{2}\, \mathbf{P}(d, 1+r^2; 0) \\ & \qquad - \sqrt{2} \Sigma_j c_j(\varepsilon) \mathbf{P}((n-j)d/n, 2+b_j'; 0) \\ & \qquad + \sqrt{2} \Sigma_j c_j(\varepsilon) \mathbf{P}(j \cdot d/n, 1+r^2+b_j'; 0)] \\ & + 2 [1 - \sqrt{2} \Sigma_j c_j(\varepsilon) \mathbf{P}((n-j)d/n, 2+b_j'; 0)] \} \end{aligned}$$

où $b_j' = b_j/\sigma^2 = [j + (n-j)r^2]/n^2$.

Cette variable b_j' dépend de r^2, n et j.

L'étude analytique du comportement de l'EQMI(g) n'est pas évidente. Elle dépend des quatre variables suivantes :

- le taux de contamination ε ,
- la différence réduite des espérances d ,
- le rapport des variances r^2 ,
- la taille de l'échantillon n .

On constate tout d'abord que l'EQMI(g) est une fonction continue en chacune des variables et qu'elle est symétrique en la variable d.

Pour faciliter l'analyse de l'EQMI(g), nous la restreignons à deux sortes précises de contamination:

(1) une contamination sur l'une des extrémités de la répartition, avec les variances σ^2 et β^2 égales (figure 3.21).
(2) une contamination centrale et symétrique, c-à-d les espérances μ et α sont égales (figure 3.22).

Pour ces deux cas particuliers de contamination, déterminons l'expression correspondante de l'EQMI(g).

Corollaire 3.1.3

Supposons que les hypothèses du théorème 3.1.2 soient vérifiées et que, de plus, les variances σ^2 et β^2 soient égales. Alors,

$$\begin{aligned} EQMI(g) = (\Pi\sigma^2)^{-1/2}\{ & \varepsilon^2 [1 - \sqrt{2}\, P(d,2;0)] \\ & - \varepsilon [1 - \sqrt{2}\, P(d,2;0) \\ & \qquad - \sqrt{2}\, \Sigma_j\, c_j(\varepsilon)\, P((n-j)d/n, 2+b_j';0) \\ & \qquad + \sqrt{2}\, \Sigma_j\, c_j(\varepsilon)\, P(j\cdot d/n, 2+b_j';0)] \\ & + [1 - \sqrt{2}\, \Sigma_j\, c_j(\varepsilon)\, P((n-j)d/n, 2+b_j';0)] \} \end{aligned} \quad (19)$$

où $b_j' = n^{-1}$. (20)

Il suffit de remplacer r^2 par 1.0 et de constater alors que :

$$1 + r^2 + b_j' = 2 + b_j' \qquad \text{et} \qquad b_j' = n^{-1} \ .$$

Corollaire 3.1.4

Supposons que les hypothèses du théorème 3.1.2 soient vérifiées et que, de plus, les espérances μ et α soient égales. Alors,

$$\begin{aligned} EQMI(g) = (4\pi\sigma^2)^{-1/2}\{ \varepsilon^2 [\ 1 + 1/r - 2\sqrt{2}\ (1+r^2)^{-1/2}\ ] \\ - 2\varepsilon [\ 1 - \sqrt{2}\ (1+r^2)^{-1/2} - \sqrt{2}\ \Sigma_j\ c_j(\varepsilon)\ (2+b_j')^{-1/2} \\ + \sqrt{2}\ \Sigma_j\ c_j(\varepsilon)\ (1+r^2+b_j')^{-1/2}\ ] \\ + 2 [\ 1 - \sqrt{2}\ \Sigma_j\ c_j(\varepsilon)\ (2+b_j')^{-1/2}\ ]\ \} \end{aligned} \qquad (21)$$

où $\quad b_j' = [\ j + (n-j)r^2]/n^2 \ . \qquad (22)$

Il suffit de remplacer d par 0.0 et de constater alors que

$$P(0,v;0) = v^{-1/2} \ , \quad v > 0 \ .$$

Avant de passer à l'analyse qualitative des formules théoriques de l'EQMI(g), nous allons déterminer les formules analogues pour l'EQMI des estimateurs non-paramétriques f_1 et f_2 pour la contamination la plus générale et pour les deux sortes particulières de contamination.

3.2 EQMI des estimateurs non-paramétriques

Les estimateurs non-paramétriques considérés sont les mêmes qu'au chapitre précédent et les notations restent identiques.
L'estimateur non-paramétrique construit à l'aide du noyau biquadratique K_1 est noté f_{1n}, tandis que f_{2n} représente l'estimateur formé à l'aide du noyau normal K_2.

$$f_{jn}(x) = (nh_n)^{-1} \sum_{i=1}^{n} K_j[(x-X_i)/h_n] \qquad j = 1,2$$

où $$K_1(y) = \begin{cases} 15/16\ (1-y^2)^2 & \text{si } |y| \leq 1 \\ 0 & \text{sinon} \end{cases}$$

$$K_2(y) = N(0,1;y) .$$

Conventions d'écriture

L'indice n sera omis si aucune confusion n'est possible dans les expressions f_{1n}, f_{2n}, h_n.
Les bornes de Σ sont omises si elles varient de 1 à n.

La contamination étant ignorée par le statisticien, la formule utilisant la fenêtre optimale (1.35) ne sera pas appliquée, car elle dépend indirectement de la connaissance de f. Le praticien choisit la valeur optimale de h déterminée pour une densité normale non contaminée et ne dépendant que de la taille de l'échantillon n. Pour estimer l'EQMI(f_j), la valeur de h sera substituée dans l'expression générale de l'EQMI (1.27).

Théorème 3.2.1

Supposons que les hypothèses (H1) et (H4) soient vérifiées. Alors, l'EQMI asymptotique des estimateurs non-paramétriques, f_1 et f_2, vaut respectivement :

$$EQMI(f_1) = 7^{-1}\ \sigma^{-1}\ [\ 5/(nk) + 3/(224\sqrt{\Pi})\ k^4 \cdot S\ ] + o[(nk)^{-1}+k^4] \qquad (23)$$

$$EQMI(f_2) = (4\Pi)^{-1/2}\ \sigma^{-1}\ [\ 1/(nk) + 3/16\ k^4 \cdot S\ ] + o[(nk)^{-1}+k^4] \qquad (24)$$

où $k = h/\sigma$ est la fenêtre réduite,

$$S = (1-\varepsilon)^2 + \varepsilon^2/r^5 + 8\sqrt{2}/3\ (1-\varepsilon)\ \varepsilon\ R\ P(d,1+r^2;0)\ , \qquad (25)$$

$$R = (1+r^2)^{-2}\ [\ d^4/(1+r^2)^2 - 6\ d^2/(1+r^2) + 3\ ]\ . \qquad (26)$$

Démonstration

Cette formule découle de l'expression générale (1.27). Il nous faut donc calculer $I(f'')$, seul facteur inconnu;

$$I(f'') = \int_{-\infty}^{+\infty} |\ f''(\tau_1,\tau_2;x)\ |^2\ dx$$

où $f''(\tau_1,\tau_2;x) = \dfrac{\delta^2}{\delta x^2}\ [\ (1-\varepsilon)\ N(\mu,\sigma^2;x) + \varepsilon\ N(\alpha,\beta^2;x)\ ]$

$$= (1-\varepsilon)\ \sigma^{-4}\ [(x-\mu)^2-\sigma^2]\ N(\mu,\sigma^2;x) + \varepsilon\ \beta^{-4}\ [(x-\alpha)^2-\beta^2]\ N(\alpha,\beta^2;x)\ .$$

Donc,

$$| f''(\tau_1,\tau_2;x) |^2 = (1-\varepsilon)^2\ \sigma^{-8}\ [(x-\mu)^2-\sigma^2]^2\ N^2(\mu,\sigma^2;x)$$

$$+ \varepsilon^2\ \beta^{-8}\ [(x-\alpha)^2-\beta^2]^2\ N^2(\alpha,\beta^2;x)$$

$$+ 2\ (1-\varepsilon)\ \varepsilon\ (\sigma\beta)^{-4}\ Q(x)$$

où $Q(x) = [(x-\mu)^2-\sigma^2]\cdot[(x-\alpha)^2-\beta^2]\ N(c,u^2;x)\ N(\alpha,\sigma^2+\beta^2;\mu)$, (27)

$$c = (\mu\beta^2+\alpha\sigma^2)/(\sigma^2+\beta^2)\ ,$$

$$u^2 = \sigma^2\beta^2/(\sigma^2+\beta^2)\ .$$

Le calcul de l'intégrale des termes en $(1-\varepsilon)^2$ et en ε^2 ne pose aucun problème (théorème 2.2.1). En effet,

$$\int_{-\infty}^{+\infty} [(x-\mu)^2-\sigma^2]\ N^2(\mu,\sigma^2;x)\ dx = 3/(8\sqrt{\pi}\sigma^5)\ .$$

Par contre, le calcul de l'intégrale de Q(x) est long et fastidieux. Pour obtenir un résultat ne dépendant que de d et r^2 pour les paramètres μ, α, σ^2 et β^2, il faut faire apparaître les moments centrés de $N(c,u^2;x)$.
Ecrivons l'expression entre crochets dans Q(x) de la manière suivante :

$$[(x-\mu)^2-\sigma^2]\cdot[(x-\alpha)^2-\beta^2] = \{[(x-c)+(c-\mu)]^2-\sigma^2\}$$

$$\cdot\ \{[(x-c)+(c-\alpha)]^2-\beta^2\}$$

$$= (x-c)^4 + 2\ (x-c)^3\ [(c-\alpha)+(c-\mu)]$$

$$+ (x-c)^2\ [(c-\alpha)^2-\beta^2+(c-\mu)^2-\sigma^2+4(c-\alpha)(c-\mu)]$$

$$+ 2\ (x-c)\ \{\ (c-\alpha)[(c-\mu)^2-\sigma^2]+(c-\mu)[(c-\alpha)^2-\beta^2]\ \}$$

$$+ \sigma^2\beta^2 + (c-\alpha)^2(c-\mu)^2 - \sigma^2(c-\alpha)^2 - \beta^2(c-\mu)^2\ .$$

Lors de l'intégration de Q(x), seuls les moments pairs sont différents de zéro. Les termes intervenants peuvent être alors écrits en fonction de d et r^2 uniquement. En effet,

$$c - \alpha = (\mu\beta^2+\alpha\sigma^2)/(\sigma^2+\beta^2) - \alpha = -\sigma\cdot r^2\, d/(1+r^2) \ , \qquad (28)$$

$$c - \mu = (\mu\beta^2+\alpha\sigma^2)/(\sigma^2+\beta^2) - \mu = \sigma\cdot d\ /(1+r^2) \ . \qquad (29)$$

La variance u^2 s'écrit :

$$u^2 = (\sigma^2\beta^2)/(\sigma^2+\beta^2) = \sigma^2\cdot r^2/(1+r^2) \ . \qquad (30)$$

En utilisant les expressions (28), (29) et (30), le terme obtenu pour le deuxième moment vaut :

$$\sigma^2\ [\ d^2(r^4-4r^2+1)/(1+r^2)^2 - (1+r^2)\ ] \ ,$$

tandis que le terme constant vaut :

$$\sigma^4\cdot r^2\ [\ r^2\cdot d^4/(1+r^2)^4 - d^2/(1+r^2)^2 + 1\ ] \ .$$

Finalement, on obtient pour l'intégrale cherchée :

$$\int_{-\infty}^{+\infty} Q(x)\,dx = 3\ \sigma^4\cdot r^4/(1+r^2)^2$$

$$+\ \sigma^2\cdot r^2/(1+r^2)\ \{\sigma^2[d^2(r^4-4r^2+1)/(1+r^2)^2-(1+r^2)]\}$$

$$+\ \sigma^4\cdot r^2\ [\ r^2d^4/(1+r^2)^4-d^2/(1+r^2)^2+1\ ]$$

$$\cdot\ N(\alpha,\sigma^2+\beta^2;\mu)$$

$$= \sigma^4\cdot r^4/(1+r^2)^2\ [\ d^4/(1+r^2)^2 - 6\ d^2/(1+r^2) + 3\ ]$$

$$\cdot\ N(\alpha,\sigma^2+\beta^2;\mu) \ .$$

Désignons par R_1 la première partie de cette dernière expression :

$$R_1 = \sigma^4 \cdot r^4/(1+r^2)^2 \; [\; d^4/(1+r^2)^2 - 6\, d^2/(1+r^2) + 3 \;] \; .$$

Ainsi, la formule complète de $I(f'')$ s'écrit :

$$\begin{aligned} I(f'') &= 3/(8\sqrt{\pi})\; \sigma^{-5}\; (1-\varepsilon)^2 + 3/(8\sqrt{\pi})\; \beta^{-5}\; \varepsilon^2 \\ &\qquad + 2\; (1-\varepsilon)\; \varepsilon\; (\sigma\cdot\beta)^{-4}\; R_1\; N(\alpha, \sigma^2+\beta^2; \mu) \\ &= 3/(8\sqrt{\pi})\; \sigma^{-5}\; [\; (1-\varepsilon)^2 + \varepsilon^2/r^5 \\ &\qquad + 8\sqrt{2}/3\; (1-\varepsilon)\; \varepsilon\; R\; P(d, 1+r^2; 0)\;] \end{aligned} \qquad (31)$$

où $R = (1+r^2)^{-2}\; [\; d^4/(1+r^2)^2 - 6\, d^2/(1+r^2) + 3 \;] \; .$

Pour simplifier, nous posons le terme entre accolades égal à S :

$$S = (1-\varepsilon)^2 + \varepsilon^2/r^5 + 8\sqrt{2}/3\; (1-\varepsilon)\; \varepsilon\; R\; P(d, 1+r^2; 0) \; .$$

La formule asymptotique de l'$EQMI(f_1)$ vaut donc :

$$\begin{aligned} EQMI(f_1) &= [\; 5/(nh) + 3/(224\sqrt{\pi})\; h^4/\sigma^5 \cdot S \;] \;/\; 7.0 \\ &\qquad + o[(nh)^{-1} + h^4] \\ &= 7^{-1}\; \sigma^{-1}\; [\; 5/(nk) + 3/(224\sqrt{\pi})\; k^4 \cdot S \;] \\ &\qquad + o[(nk)^{-1} + k^4] \; , \end{aligned}$$

tandis que celle de l'EQMI(f_2) vaut :

$$EQMI(f_2) = (4\Pi)^{-1/2}\ \sigma^{-1}\ [\ 1/(nk) + 3/16\ k^4 \cdot S\ ] + o[(nk)^{-1}+k^4]$$

où $k = h/\sigma$ représente la fenêtre réduite.

Cette fenêtre k vérifie les hypothèses (H4) par construction.

L'EQMI asymptotique dépend de cinq variables :
- le taux de contamination ε,
- la fenêtre réduite du noyau k,
- la différence réduite des espérances d,
- le rapport des variances r^2,
- la taille de l'échantillon n.

C'est une fonction continue en chacune des variables et elle est symétrique en la variable d.

L'étude analytique n'est pourtant pas aisée. Déterminons l'EQMI(f_2) pour les deux sortes de contamination nous intéressant. Seules les constantes R et S dépendent directement de la contamination. Ainsi, il suffit de les calculer pour $r^2 = 1.0$ et pour $d = 0.0$.

Corollaire 3.2.2

Si les hypothèses du théorème 3.2.1 sont vérifiées, et que, de plus, les variances sont égales, alors,

$$S = 1 - 2\ (1-\varepsilon)\ \varepsilon\ [1 - \sqrt{2}\ (1-d^2+d^4/12)\ P(d,2;0)]\ . \quad (32)$$

Démonstration

Cette formule se déduit immédiatement des formules (25) et (26) en substituant r^2 par 1.0. En effet, on obtient pour R :

$$R = 3/4 \left(1 - d^2 + d^4/12 \right) . \qquad (33)$$

Et par substitution de R dans S, on a l'expression désirée.

Corollaire 3.2.3

Si les hypothèses du théorème 3.2.1 sont vérifiées, et que, de plus, les espérances sont égales, alors,

$$S = (1-\varepsilon)^2 + \varepsilon^2/r^5 + 8\sqrt{2}\,(1-\varepsilon)\,\varepsilon\,(1+r^2)^{-5/2} . \qquad (34)$$

Démonstration

Evident, en posant d = 0.0 dans les formules (25) et (26).

Cependant, pour l'estimateur f_2, il est possible de calculer explicitement le critère de l'EQMI. Cette formule a été calculée notamment par Anderson (1969) et par Fryer (1976), et est déterminée dans le théorème suivant.

Théorème 3.2.4

Supposons que les hypothèses (H1) et (H4) soient vérifiées. Alors, l'EQMI de l'estimateur à noyau normal, f_2, vaut :

$$EQMI(f_2) = (4\Pi\sigma^2)^{-1/2}\{ (nk)^{-1}$$

$$+ (1-\varepsilon)^2 [1 + (1-1/n)/(1+k^2)^{1/2} - 2/(1+k^2/2)^{1/2}]$$

$$+ \varepsilon^2/r [1 + (1-1/n)/(1+k^2/r^2)^{1/2} - 2/(1+k^2/(2r^2))^{1/2}]$$

$$+ 2\sqrt{2} (1-\varepsilon) \varepsilon [P(d,1+r^2;0) + (1-1/n) P(d,1+r^2+2k^2;0)$$

$$- 2 P(d,1+r^2+k^2;0)] \} . \quad (35)$$

Démonstration

La démonstration est identique à celle du théorème 2.2.2, la fonction $N(\mu,\sigma^2;x)$ étant remplacée par la fonction $(1-\varepsilon)\ N(\mu,\sigma^2;x) + \varepsilon\ N(\alpha,\beta^2;x)$. Ainsi, les formules obtenues sont plus longues et plus compliquées.

Les variables $X_1, X_2, \ldots, X_n$ étant i.i.d , l'estimateur $f_2(x)$ peut être considéré comme une somme de n variables aléatoires indépendantes, de la forme :

$$f_2(x) = n^{-1} \Sigma_i\ N(X_i,h^2;x) .$$

Calculons les différents termes composant l'$EQMI(f_2)$ à l'aide des formules (1.8) et (1.10) :

$$EQMI[f_2(x)] = \int_{-\infty}^{+\infty} \{ Biais^2[f_2(x)] + Var_f[f_2(x)] \}\ dx .$$

Soit x un point donné.

1) $E_f[f_2(x)] = E_f[N(y,h^2;x)]$

$$= \int_{-\infty}^{+\infty} N(y,h^2;x) [(1-\varepsilon)\ N(\mu,\sigma^2;x) + \varepsilon\ N(\alpha,\beta^2;x)]\ dy$$

$$= (1-\varepsilon)\ N(\mu,\sigma^2+h^2;x) + \varepsilon\ N(\alpha,\beta^2+h^2;x) .$$

Alors,

$$\text{Biais}[f_2(x)] = (1-\varepsilon)\, N(\mu,\sigma^2;x) + \varepsilon\, N(\alpha,\beta^2;x)$$

$$- [(1-\varepsilon)\, N(\mu,\sigma^2+h^2;x) + \varepsilon\, N(\alpha,\beta^2+h^2;x)] \ . \quad (36)$$

En élevant l'expression (36) au carré et en intégrant le résultat par rapport à x, on obtient le premier terme de l'EQMI(f_2);

$$\int_{-\infty}^{+\infty} \text{Biais}^2(f_2)\, dx = (4\Pi)^{-1/2}\{\ \varepsilon^2[\ \beta^{-1} + (\beta^2+h^2)^{-1/2}$$

$$- 2\sqrt{2}/(2\beta^2+h^2)^{1/2}\ ]$$

$$+ (1-\varepsilon)^2\ [\ \sigma^{-1} + (\sigma^2+h^2)^{-1/2} - 2\sqrt{2}/(2\sigma^2+h^2)^{1/2}\ ]$$

$$+ 2\sqrt{2}\ (1-\varepsilon)\ \varepsilon\ [\ P(\alpha,\sigma^2+\beta^2;\mu) + P(\alpha,\sigma^2+\beta^2+2h^2;\mu)$$

$$- 2\ P(\alpha,\sigma^2+\beta^2+h^2;\mu)\ ]\ \}\ . \quad (37)$$

2) Déterminons la variance de f_2 au point x.

$$\text{Var}_f[f_2(x)] = n^{-1}\ \text{Var}_f[N(y,h^2;x)]$$

$$= n^{-1}\ \{\ E_f[N^2(y,h^2;x)] - [E_f[N(y,h^2;x)]]^2\}$$

$$= n^{-1}\ \{\ \int_{-\infty}^{+\infty} N^2(y,h^2;x)\ [(1-\varepsilon)N(\mu,\sigma^2;x)+\varepsilon\ N(\alpha,\beta^2;x)]dy$$

$$- [(1-\varepsilon)\ N(\mu,\sigma^2+h^2;x) + \varepsilon\ N(\alpha,\beta^2+h^2;x)]^2\}$$

$$= n^{-1}(4\Pi h^2)^{-1/2}\{\int_{-\infty}^{+\infty} N(y,h^2/2;x)[(1-\varepsilon)N(\mu,\sigma^2;x)+\varepsilon N(\alpha,\beta^2;x)]dy$$

$$- n^{-1}\ [\ (1-\varepsilon)^2\ N^2(\mu,\sigma^2+h^2;x) + \varepsilon^2\ N^2(\alpha,\beta^2+h^2;x)$$

$$+ 2\ (1-\varepsilon)\ \varepsilon\ N(\mu,\sigma^2+h^2;x)\ N^2(\alpha,\beta^2+h^2;x)\ ]$$

$$= n^{-1}(4\Pi h^2)^{-1/2}[(1-\varepsilon)N(\mu,\sigma^2+h^2/2;x)+\varepsilon N(\alpha,\beta^2+h^2/2;x)]$$

$$- n^{-1}\ [\ (1-\varepsilon)^2\ N^2(\mu,\sigma^2+h^2;x) + \varepsilon^2\ N^2(\alpha,\beta^2+h^2;x)$$

$$+ 2\ (1-\varepsilon)\ \varepsilon\ N(\mu,\sigma^2+h^2;x)\ N(\alpha,\beta^2+h^2;x)\ ]\ .\quad (38)$$

En intégrant l'expression (38) par rapport à x, on obtient le second membre de l'EQMI(f_2) :

$$\int_{-\infty}^{+\infty} Var_f[f_2(x)]\,dx = (4\Pi)^{-1/2}\ n^{-1}\ [\ h^{-1} - (1-\varepsilon)^2/(\sigma^2+h^2)^{1/2}$$

$$- \varepsilon^2/(\beta^2+h^2)^{1/2} - 2\sqrt{2}(1-\varepsilon)\ \varepsilon\ P(\alpha,\sigma^2+\beta^2+2h^2;\mu)\ ].\quad (39)$$

En sommant les expressions (37) et (39), l'expression désirée pour l'EQMI(f_2) est obtenue :

$$EQMI(f_2) = (4\Pi)^{-1/2}\ n^{-1}\ [\ h^{-1} - (1-\varepsilon)^2/(\sigma^2+h^2)^{1/2}$$

$$- \varepsilon^2/(\beta^2+h^2)^{1/2} - 2\sqrt{2}\ (1-\varepsilon)\ \varepsilon\ P(\alpha,\sigma^2+\beta^2+2h^2;\mu)\ ]$$

$$+ \{\ (1-\varepsilon)^2\ [\ \sigma^{-1} + (\sigma^2+h^2)^{-1/2} - 2\sqrt{2}/(2\sigma^2+h^2)^{1/2}\ ]$$

$$+ \varepsilon^2\ [\ \beta^{-1} + (\beta^2+h^2)^{-1/2} - 2\sqrt{2}/(2\beta^2+h^2)^{1/2}\ ]$$

$$+ 2\sqrt{2}\ (1-\varepsilon)\ \varepsilon\ [\ P(\alpha,\sigma^2+\beta^2;\mu) + P(\alpha,\sigma^2+\beta^2+2h^2;\mu)$$

$$- 2\ P(\alpha,\sigma^2+\beta^2+h^2;\mu)\ ]\ \}\ .\quad (40)$$

Ecrivons l'EQMI(f_2) à l'aide de la fenêtre réduite k d'une part, et des variables d et r^2, d'autre part :

$k = h/\sigma$,

$d = (\alpha-\mu)/\sigma$ et

$r^2 = \sigma^2/\beta^2$.

$$EQMI(f_2) = (4\pi\sigma^2)^{-1/2}\{ (nk)^{-1}$$

$$+ (1-\varepsilon)^2 [1 + (1-1/n)/(1+k^2)^{1/2} - 2/(1+k^2/2)^{1/2}]$$

$$+ \varepsilon^2/r [1 + (1-1/n)/(1+k^2/r^2)^{1/2} - 2/(1+k^2/(2r^2))^{1/2}]$$

$$+ 2\sqrt{2} (1-\varepsilon) \varepsilon [P(d,1+r^2;0) + (1-1/n)P(d,1+r^2+2k^2;0)$$

$$- 2 P(d,1+r^2+k^2;0)] \} .$$

On constate que l'EQMI(f_2) dépend des variables suivantes :

- le taux de contamination ε,
- la fenêtre réduite du noyau k,
- la différence réduite des espérances d,
- le rapport des variances r^2,
- la taille de l'échantillon n.

L'EQMI(f_2) est une fonction continue en chacune des variables et symétrique en la variable d. L'étude analytique reste pourtant complexe.

Calculons l'expression de l'EQMI pour les deux cas particuliers de contamination nous intéressant, $r^2 = 1.0$ et $d = 0.0$.

Corollaire 3.2.5

Si les hypothèses du théorème 3.2.4 sont vérifiées, et que, de plus, les variances sont égales, alors,

$$EQMI(f_2) = (4\pi\sigma^2)^{-1/2}\{ (nk)^{-1} + [2\varepsilon^2 - 2\varepsilon + 1]$$

$$\cdot [1 + (1-1/n)/(1+k^2)^{1/2} - 2/(1+k^2/2)^{1/2}]$$

$$+ 2\sqrt{2} (1-\varepsilon) \varepsilon [P(d,2;0) + (1-1/n) P(d,2+2k^2;0)$$

$$- 2 P(d,2+k^2;0)] \} . \qquad (41)$$

Ce corollaire est évident, en posant $r^2 = 1.0$ dans l'expression (35).

Corollaire 3.2.6

Si les hypothèses du théorème 3.2.4 sont vérifiées, et que, de plus, les espérances sont égales, alors,

$$EQMI(f_2) = (4\Pi\sigma^2)^{-1/2} \{ (nk)^{-1}$$

$$+ (1-\varepsilon)^2 [1 + (1-1/n)/(1+k^2)^{1/2} - 2/(1+k^2/2)^{1/2}]$$

$$+ \varepsilon^2/r [1 + (1-1/n)/(1+k^2/r^2)^{1/2} - 2/(1+k^2/(2r^2))^{1/2}]$$

$$+ 2\sqrt{2} (1-\varepsilon) \varepsilon [(1+r^2)^{-1/2} + (1-1/n)/(1+r^2+2k^2)^{1/2}$$

$$- 2 (1+r^2+k^2)^{-1/2}] \} . \qquad (42)$$

Cette expression de l'$EQMI(f_2)$ découle aisément de la formule (35), en posant $d = 0.0$ et en remarquant que : $P(0,y;x) = y^{-1/2}$, $y > 0$.

Maintenant que les formules théoriques des EQMI des estimateurs paramétrique et non-paramétriques sont déterminées pour les répartitions normales contaminées par une seconde répartition normale, une analyse comparative des résultats obtenus peut débuter.

3.3 Analyse des EQMI

Afin de mieux comprendre le comportement et les propriétés de l'efficacité entre les estimateurs utilisés, il est nécessaire d'analyser les caractéristiques des EQMI de chacun des estimateurs.
Une étude purement analytique s'avère très difficile, car l'EQMI de l'estimateur paramétrique g dépend de quatre variables, ε, n, d et r^2 et celle des estimateurs

non-paramétriques d'une cinquième variable, k.
Le taux de contamination ε est sans aucun doute la principale variable, car sans lui le problème n'existerait pas. L'EQMI est donc toujours étudiée par rapport à cette variable, les autres étant considérées comme des paramètres.

Si l'on analyse l'influence sur l'EQMI de chaque paramètre, avec cinq valeurs, en considérant les autres comme des constantes, auquelles trois valeurs sont données, on disposera d'un nombre considérable (5x3x3) de tableaux et de graphes dont il ne sera pas possible de tirer des caractéristiques et des conclusions nettes, l'interdépendance entre les paramètres pouvant être difficilement décelable.

Nous nous restreignons à deux sortes particulières de contamination qui permettent de schématiser et de comprendre mieux les contaminations plus générales :

(1) la répartition est contaminée sur l'une de ses extrémités et les variances sont égales, c-à-d $d \neq 0.0$ et $r^2 = 1.0$ (figure 3.21) ;
(2) une contamination cachée, symétrique et centrale, c-à-d $d = 0.0$ et $r^2 \neq 1.0$ (figure 3.22).

Ainsi, les EQMI deviennent des fonctions de trois ou de quatre variables : ε, n, d ou r^2, k. L'analyse des EQMI est menée en deux étapes pour chaque sorte de contamination.
Tout d'abord, le rôle de la taille de l'échantillon n est déterminé pour deux valeurs particulières du paramètre étudié; puis, l'influence détaillée de ce paramètre sur l'EQMI est analysée, la taille de l'échantillon n étant fixée à 50.
Pour l'estimateur non-paramétrique f_2, dans une troisième partie, différentes valeurs ont été attribuées en plus à la fenêtre réduite k pour juger les conséquences sur son EQMI .

Pour examiner la qualité de la croissance (décroissance) de l'EQMI d'un estimateur donné f_n, en fonction de ε,

le rapport entre l'EQMI pour $\varepsilon = 0.5$ et l'EQMI pour $\varepsilon = 0.0$ a été calculé; il est noté $Rap(f_n)$.

$$Rap(f_n) = \frac{EQMI(f_n) \text{ pour } \varepsilon = 0.5}{EQMI(f_n) \text{ pour } \varepsilon = 0.0}$$

Ce critère indique en fait combien de fois la valeur initiale a été multipliée; il s'avère meilleur que la différence entre les EQMI, car celle-ci est toujours relativement petite et donc pas très significative.

Si $Rap(f_n)$ est supérieur à 3.0, la croissance de l'EQMI est qualifiée de forte, tandis que si $Rap(f_n)$ est compris entre 0.9 et 1.1, l'EQMI est jugée de légèrement décroissante (croissante), voire constante.

3.3.1 Analyse de l'EQMI de l'estimateur paramétrique

L'analyse des EQMI est basée sur les résultats obtenus aux corollaires 3.1.3 et 3.1.4. L'étude est principalement conduite à partir des graphes des EQMI (figures 3.1 à 3.4) et des tableaux 3.1 et 3.5.

3.3.1.1 Analyse selon la différence des espérances

(Figures 3.1 et 3.2)

Déterminons le rôle de la taille de l'échantillon n pour deux valeurs de d (tableau 3.1).

Si $d = 1/2$, les EQMI sont des fonctions constantes, voire légèrement croissantes de ε (figure 3.1). Le taux de contamination ε n'influence pratiquement pas l'EQMI(g), qui est par contre fortement dépendante de la taille de l'échantillon n (tableau 3.1).
Pour $d = 3.0$, les EQMI sont des fonctions fortement croissantes du taux de contamination ε (figure 3.1). Les valeurs des EQMI pour $\varepsilon = 0.5$ sont semblables et

indépendantes de la taille de l'échantillon n, au contraire des EQMI pour $\varepsilon = 0.0$ où les différences relatives sont grandes. Cette constatation explique pourquoi Rap(g) est approximativement proportionnel à n. Le rôle de la taille de l'échantillon n est donc de fixer l'EQMI(g) en $\varepsilon = 0.0$; son importance diminue à mesure que le taux de contamination augmente et disparaît pratiquement pour $\varepsilon = 0.5$ (tableau 3.1).

Dans les deux cas, la taille de l'échantillon sert à déterminer l'EQMI(g) pour $\varepsilon = 0.0$ et ainsi à améliorer la qualité de l'estimateur g.

La contamination influence plus ou moins fortement l'EQMI(g) et une étude plus détaillée est nécessaire pour tirer des conclusions précises sur le rôle du paramètre d. Fixons la taille de l'échantillon à 50 pour cette analyse.

Les EQMI sont des fonctions monotones croissantes du taux de contamination ε, à tangente horizontale au point $\varepsilon = 0.5$ par symétrie (figure 3.2). Parmi les valeurs données à d, deux cas se distinguent très nettement : premièrement, $d > 1.0$ et deuxièmement $d \leq 1.0$ (tableau 3.5).
Dans le premier cas, l'EQMI est une fonction fortement croissante de ε (Rap(g) > 2.0), croissance s'accentuant encore plus avec d augmentant (Rap(g) > 40 pour $d = 3.0$). Ainsi, l'estimateur paramétrique g est instable à toute contamination si la différence des espérances μ et α est prononcée, c-à-d $d > 1.0$. Cette constatation n'est pourtant pas surprenante; la moyenne arithmétique des observations m est une bonne estimation de $E_f(X)$, mais une très mauvaise de μ. La densité ainsi obtenue est translatée de manière notoire sur la gauche ou la droite de l'espérance μ, tandis que la densité exacte s'éloigne de plus en plus d'une densité normale (provoquant une augmentation très forte de l'erreur commise) (figure 3.21).
Par contre, dans le second cas, $d \leq 1.0$, l'EQMI est pratiquement une fonction constante de ε, Rap(g) étant proche de 1.05 (tableau 3.5 et figure 3.2). L'estimateur

paramétrique g est donc insensible à une contamination voisine de μ. La densité estimée est translatée légèrement et demeure une bonne approximation de la vraie densité (figure 3.21).

3.3.1.2 Analyse selon le rapport des variances

(Figures 3.3 et 3.4)

Le rôle de la taille de l'échantillon est le même suivant les valeurs données à r^2.

Pour $r^2 = 1/4$, les EQMI sont des fonctions fortement croissantes de ε (figure 3.3). Mais, pour les valeurs traitées de n, les différences entre les valeurs des EQMI restent constantes et donc indépendantes de ε. Ainsi, la croissance des EQMI n'est due qu'au taux de contamination ε. L'influence de la taille de l'échantillon n est d'améliorer la qualité de notre estimateur (tableau 3.1 et figure 3.3).
Pour $r^2 = 4.0$, les EQMI sont des fonctions aussi fortement croissantes de ε (figure 3.3). Les différences entre les valeurs des EQMI pour les trois valeurs de n étudiées sont pratiquement constantes. Donc, une taille d'échantillon plus élevée augmente la précision de l'estimateur g (tableau 3.1 et figure 3.3).
La taille de l'échantillon occupe donc une place importante puisqu'elle permet d'abaisser notablement l'EQMI(g) et donc d'améliorer l'estimation de f.

Afin de cerner le comportement de l'EQMI suivant le paramètre r^2, quatre valeurs lui ont été attribuées, la taille de l'échantillon étant fixée à 50.

Pour chaque valeur de r^2, l'EQMI est une fonction monotone croissante du taux de contamination ε, sans aucune symétrie. Cette croissance est très marquée d'une part, Rap(g) > 7.0, et d'autre part elle s'accentue de plus en plus si r^2 s'éloigne de 1.0 et surtout si r^2 tend vers zéro (figure 3.4 et tableau 3.5). Cet estimateur est

donc instable à toute contamination centrale. Intuitivement, la moyenne estime correctement μ; par contre, si r^2 est inférieur à 1.0, l'estimateur obtenu sous-estime le mode et surestime les extrémités de la vraie répartition, tandis que le phénomène inverse se produit si r^2 est supérieur à 1.0 (figure 3.22). L'EQMI est donc aussi une fonction dépendant fortement du paramètre r^2.

En comparant la croissance des EQMI selon d et r^2, force est de constater que l'EQMI est plus sensible à des variations élevées des espérances que des variances (tableau 3.5 et figures 3.2 et 3.4). Une erreur d'estimation de μ entraîne une translation de toute la courbe de densité estimée et provoque ainsi une erreur supérieure à une simple contraction ou à un aplatissement de la densité exacte.

En résumé,

l'EQMI de l'estimateur paramétrique est très sensible :
- à toutes variations de r^2, si $d = 0.0$, et,
- à des différences élevées de d, $d > 1.0$, si $r^2 = 1.0$.

Par contre, l'EQMI(g) est indépendante du taux de contamination ε uniquement lorsque $r^2 = 1.0$ et $d \leq 1.0$.
La taille de l'échantillon permet d'améliorer la précision de l'estimateur g, sauf si d est supérieur à 1.0 et que le taux de contamination est élevé.

3.3.2 Analyse de l'EQMI des estimateurs non-paramétriques

Pour ces estimateurs, f_1 et f_2, l'analyse est conduite d'une manière quelque peu différente.
Tout d'abord, l'équivalence entre les formules asymptotiques des noyaux biquadratique et normal est démontrée pour des cas particuliers de n, d et r^2 avec la fenêtre optimale pour $\varepsilon = 0.0$.
Ensuite, les formules asymptotique et exacte de l'EQMI de f_2 sont comparées entre elles et la mauvaise qualité de l'EQMI asymptotique de f_2 est établie.
Enfin, en se référant à quelque conjecture et surtout en

étudiant les simulations (paragraphe 3.5), la formule asymptotique de l'EQMI(f_1) s'avère elle aussi peu performante.
Ainsi, seule la formule de l'EQMI exacte de f_2 est analysée sur le modèle de l'EQMI paramétrique, avec une partie complémentaire sur le rôle que joue la fenêtre dans l'EQMI(f_2).

3.3.2.1 Abandon de la formule asymptotique de l'EQMI

La variable jouant le rôle le plus important après le taux de contamination ε est sans aucun doute la fenêtre (réduite) k. Comme l'on suppose que le statisticien ignore la contamination, celui-ci choisit donc une largeur indépendante du taux de contamination ε et basée principalement sur la taille de l'échantillon n et sur le noyau utilisé. Pour ces formules asymptotiques, la fenêtre minimisant l'EQMI asymptotique lorsque la contamination est inexistante s'impose d'elle-même (2.43 et 2.44).

$k_1 = 2.778\ n^{-1/5}$ pour K_1.

$k_2 = 1.059\ n^{-1/5}$ pour K_2.

Afin d'examiner les EQMI, les valeurs de k calculées pour différentes tailles d'échantillon ont été substituées dans les expressions de l'EQMI asymptotique obtenues dans les corollaires 3.2.2 et 3.2.3.

Les cas particuliers étudiés précédemment sont repris dans ce paragraphe, à savoir :

<u>si $r^2 = 1.0$</u> : $d = 1/2$, $d = 3.0$;

<u>si $d = 0.0$</u> : $r^2 = 1/4$, $r^2 = 4.0$,

et les trois tailles d'échantillons, 20, 50 et 100.

En analysant les valeurs des EQMI asymptotiques pour les deux noyaux (tableaux 3.2 et 3.3), les remarques suivantes peuvent être formulées sans équivoque :

1. Les valeurs des EQMI pour le noyau biquadratique sont toujours légèrement inférieures à celles du noyau normal. Certes, les variations obtenues sont minimes et s'amenuisent lorsque la taille de l'échantillon augmente.
2. Le rapport entre les deux EQMI est pratiquement constant, quel que soit les valeurs de ε, n, d ou r^2, et est identique à celui des formules asymptotiques sans contamination ($EQMI(f_1)/EQMI(f_2) \approx 0.965$).

n	20	50	100
Différences	0.001	0.0005	0.0003
Rapport	0.96	0.96	0.96

En résumé,

les formules asymptotiques des noyaux biquadratique et normal sont équivalentes, avec un léger avantage au premier nommé, s'il en existe un.

Analysons la précision de la formule asymptotique de l'EQMI pour le noyau normal. En substituant la même largeur du noyau dans la formule exacte (3.41 ou 3.42) et dans la formule asymptotique (3.32 ou 3.34), les valeurs des EQMI correspondantes ont pu être calculées et reportées sur des graphes (figures 3.5 et 3.6).
La conclusion suivante peut être énoncée :

la formule asymptotique de l'EQMI est très grossière et donne un ordre imprécis de l'erreur commise.
En effet, les différences entre les valeurs exactes et les valeurs asymptotiques sont énormes et ne diminuent que très lentement lorsque n augmente :

par exemple : EQMI EX/EQMI AS ≃ 0.60 pour n = 50 et
EQMI EX/EQMI AS ≃ 0.70 pour n = 500.

Ces grandes variations ne peuvent être qualifiées de négligeables et démontrent ainsi l'inexactitude et la faiblesse de la formule asymptotique pour le noyau normal.

Les simulations confirmeront cette remarque et démontreront en plus que la formule asymptotique du noyau biquadratique est aussi mauvaise et très approximative (paragraphe 3.5, figures 3.17 - 3.20, tableaux 3.19 à 3.23).

En résumé,

les formules asymptotiques de l'EQMI des estimateurs non-paramétriques sont mauvaises et donc abandonnées pour l'analyse théorique qui suit.
Seule l'EQMI exacte du noyau normal est étudiée en détail selon le modèle de l'estimateur paramétrique dans les deux prochains paragraphes.

3.3.2.2 EQMI du noyau normal pour une fenêtre donnée

Le paragraphe 3.3.2.3 traite des variations de l'EQMI selon différentes fenêtres. Aussi, dans ces deux premiers paragraphes, la largeur de Fryer a été utilisée et substituée dans les expressions obtenues dans les corollaires 3.2.5 et 3.2.6. Rappelons que :

$k_{Fryer} = 1.31\ n^{-0.205}$ pour K_2.

Les constatations sont basées principalement sur l'analyse des graphes des EQMI (figures 3.7 - 3.10) et des tableaux 3.4 et 3.5.

3.3.2.2.1 Analyse selon la différence des espérances

(Figures 3.7 et 3.8)

Déterminons le rôle joué par la taille de l'échantillon n dans l'EQMI(f_2).

Pour les deux valeurs de d traitées, d = 1/2 et d = 3.0, l'EQMI(f_2) est une fonction pratiquement constante de ε (figure 3.7). Les différences entre les EQMI pour un taux de contamination donné sont elles aussi constantes et montrent ainsi que la taille de l'échantillon exerce une influence primordiale, puisqu'elle fixe l'EQMI(f_2) et donc la qualité de l'estimation (tableau 3.4).

Pour n = 50, précisons quelque peu le rôle du paramètre d.

L'EQMI(f_2) est une fonction légèrement décroissante, voire constante de la contamination ε et à tangente horizontale pour ε = 0.5 par symétrie (tableau 3.5 et figure 3.8). Le changement de courbure, Rap(f_2) > 1.0, qui a lieu si la différence des espérances est élevée, est essentiellement dû au choix de la fenêtre (paragraphe 3.3.2.3) et ne modifie pas les caractéristiques globales de l'EQMI de f_2. Cet estimateur nous rend donc l'allongement de la queue, et le cas échéant, le second mode de la vraie densité, et le sous-estime généralement (figure 3.21).

3.3.2.2.2 Analyse selon le rapport des variances

(Figures 3.9 et 3.10)

La taille de l'échantillon a pratiquement le même effet sur l'EQMI(f_2) suivant les deux valeurs données à r^2.

Si r^2= 1/4, l'EQMI(f_2) est une fonction fortement croissante du taux de contamination ε (figure 3.9). Les écarts entre les valeurs des EQMI pour un taux de contamination fixé augmentent eux aussi légèrement lorsque n croît, sans pour autant que les graphes

respectifs des EQMI semblent diverger. La taille de l'échantillon occupe donc un rôle primordial puisqu'elle permet d'abaisser l'EQMI(f_2) non seulement en $\varepsilon = 0.0$, mais aussi de plus en plus lorsque le taux de contamination augmente (tableau 3.4).
Si $r^2 = 4.0$, l'EQMI(f_2) est une fonction légèrement décroissante de ε. Les différences entre les valeurs des EQMI pour un ε fixé diminuent un peu lorsque n augmente, ce qui indique que l'influence du taux de contamination diminue quelque peu avec n. Mais, d'une manière globale, une taille d'échantillon élevée permet d'améliorer l'EQMI(f_2) (tableau 3.4).
Dans les deux cas traités, la taille de l'échantillon occcupe une place importante en bonifiant l'estimation, gain plus marqué pour $r^2 = 1/4$ que pour $r^2 = 4.0$.

Une analyse plus fine est nécessaire pour déterminer le rôle exact de r^2 dans l'EQMI, n étant fixé à 50.

Deux cas se distinguent très nettement en étudiant le tableau 3.5 et/ou les graphes des EQMI (figure 3.10) : $r^2 > 1.0$ et $r^2 < 1.0$.
Dans le premier cas, l'EQMI est une fonction monotone décroissante de ε, décroissance moins prononcée si r^2 augmente :

$r^2 = 4.0$, $Rap(f_2) \simeq 0.82$ et $r^2 = 9.0$, $Rap(f_2) \simeq 0.87$.

L'estimateur f_2 s'adapte donc à un mode plus aplati que celui de la densité normale (figure 3.22).
Par contre, si r^2 est inférieur à 1.0, l'EQMI devient une fonction monotone croissante de ε, croissance plus forte pour des valeurs de r^2 s'approchant de 0.0 :

$r^2 = 1/4$, $Rap(f_2) \simeq 2.88$ et $r^2 = 1/9$, $Rap(f_2) \simeq 6.95$.

Cet estimateur a donc des difficultés à estimer une densité dont le mode est fortement prononcé (figure 3.22).
De plus, on constate que la décroissance de l'EQMI(f_2) est modérée alors que la croissance est très forte.

En résumé,

l'estimateur f_2 s'adapte à la situation si, d'une part, la contamination est asymétrique ($d \neq 0.0$ et $r^2 = 1.0$) ou si, d'autre part, la contamination aplatit le mode de la densité normale ($d = 0.0$ et $r^2 > 1.0$).
Par contre, il estime difficilement une répartition fortement concentrée en son centre ($d = 0.0$ et $r^2 < 1.0$).
La taille de l'échantillon joue un rôle important, car elle améliore sensiblement l'EQMI(f_2) et donc la qualité de l'estimateur.

Une étude concernant les variations de l'EQMI suivant les valeurs de la fenêtre semble apppropriée et nécessaire pour compléter les résultats obtenus.

3.3.2.3 EQMI du noyau normal pour différentes fenêtres

Il se peut que la largeur de Fryer ne soit plus aussi performante que lorsque la contamination est inexistante. Nous avons donc substitué quatre largeurs tirées du tableau 2.7 dans la formule (3.41) ou (3.42) et analysé les nouvelles valeurs obtenues pour l'EQMI(f_2). Ces largeurs sont notées k_1, k_2, k_3 et k_4, k_2 désignant la fenêtre asymptotique et k_3 la fenêtre de Fryer. Le rapport entre les différentes largeurs est conservé.

Pour ne pas trop charger les tableaux et les graphes, nous nous sommes restreints à trois valeurs pour d,

$d = 1/3$, $d = 1.0$, $d = 5.0$,

et deux valeurs de part et d'autre de 1.0 pour r^2,

$r^2 = 1/9$, $r^2 = 9.0$.

L'analyse est conduite de la même manière que précédemment, c-à-d que, pour chaque sorte de contamination, le rôle de la taille de l'échantillon est déterminé avant

de disséquer l'influence de la fenêtre suivant les différentes valeurs du paramètre d ou r^2.
Elle est basée sur les graphes (figures 3.11 et 3.12) et sur les valeurs des EQMI pour les valeurs des paramètres n, d ou r^2, k (tableau 3.6).

3.3.2.3.1 Analyse selon la différence des espérances

(Figure 3.11)

Avant de caractériser le rôle précis de la fenêtre k, déterminons celui tenu par la taille de l'échantillon pour différentes valeurs de k et de d.

Pour des valeur fixées de k et de d, le rôle premier de la taille de l'échantillon est de déterminer l'EQMI(f_2) en $\varepsilon = 0.0$ (figure 3.11). Comme l'EQMI est une fonction pratiquement constante de ε, l'influence de n est d'améliorer sensiblement la qualité de l'estimation de f. Pour des taux élevés de contamination, cet effet peut être renforcé, par exemple pour d = 5.0 et k_1, k_2 et k_3, ou, au contraire, amoindri, par exemple pour d = 1.0 et k_3 et k_4 (tableau 3.6).
Ainsi, le rôle de la taille de l'échantillon est très important, puisque celle-ci fixe pratiquement seule l'EQMI(f_2).

Si l'on examine globalement une ligne de la figure 3.11, ou directement la formule (3.41) de l'EQMI(f_2), on constate alors que la figure de n = 20 est un agrandissement de celle obtenue pour n = 50, et elle-même un agrandissement de celle obtenue pour n = 100. En fait, l'EQMI pour n = 50 peut se calculer par une transformation affine de l'EQMI pour n = 100, dilatation et translation vers le haut, ou de n = 20, contraction et translation vers le bas.
Mais, les propriétés globales de l'EQMI(f_2) restent inchangées suivant la taille de l'échantillon.

Exemples : Rap(f_2) pour différentes fenêtres.
(tirés du tableau 3.6)

n	fenêtre	d = 1/3	d = 1.0	d = 5.0
20	0.71	0.98	0.86	1.13
50	0.59	0.98	0.84	1.05
100	0.51	0.98	0.84	1.00

Cernons maintenant le rôle précis de la fenêtre suivant les valeurs de d, si la taille de l'échantillon est fixée à 50.

L'influence de la largeur est prépondérante sur l'EQMI et n'est pas la même suivant la valeur de d (figure 3.11, colonne centrale).

Si d = 1/3, l'EQMI(f_2) est une fonction pratiquement constante de ε pour les largeurs étudiées et ne dépend donc pas du taux de contamination ε (Rap(f_2) $\simeq$ 0.98). La fenêtre joue un rôle essentiel en déterminant le niveau de l'EQMI(f_2) à l'origine. Pour une largeur inférieure à celle de Fryer, k_2, une EQMI plus petite est obtenue, mais les écarts sont minimes. Une fenêtre plus large, k_4, ou plus étroite, k_1, n'influence pas énormément l'EQMI(f_2) (tableau 3.6).
Si d = 1.0, l'EQMI(f_2) est une fonction constante, voire légèrement décroissante du taux de contamination ε : Rap(f_2) est compris entre 0.79 et 1.0 (tableau 3.6). La largeur de Fryer obtient l'EQMI la plus faible, sauf si le taux de contamination est faible ($\varepsilon \leq 0.15$) où une fenêtre plus étroite, k_2, s'avère alors meilleure. Mais les différences restent toutefois petites. Si la largeur est trop grande, k_4, alors l'EQMI(f_2) devient une fonction plus fortement décroissante de ε, et reste supérieure, quel que soit ε, à l'EQMI obtenue par la largeur de Fryer.
Si d = 5.0, l'EQMI(f_2) est une fonction légèrement croissante de ε, sauf si la fenêtre est large, k_4 : Rap(f_2) est compris entre 0.96 et 1.20 (tableau 3.6). La largeur de Fryer n'est devancée que par une fenêtre plus

étroite, k_2 si la contamination est faible ($\varepsilon < 0.12$); mais, les écarts sont négligeables. Pour une fenêtre étroite, k_1, la croissance de l'EQMI(f_2) par rapport à ε est plus marquée. L'EQMI(f_2) devient une fonction décroissante de ε si la fenêtre est large, k_4, dont les valeurs se rapprochent de celles obtenues par la fenêtre de Fryer. En prenant une fenêtre encore plus large, l'EQMI(f_2) deviendrait alors certainement inférieure à celle de la fenêtre de Fryer à partir d'un taux de contamination élevé et aux dépens d'une EQMI très forte pour des taux faibles à moyens.

En résumé,

Une sous-évaluation de la fenêtre par rapport à celle proposée par Fryer augmente l'EQMI(f_2) qui devient une fonction très légèrement croissante de ε, i.e. Rap(f_2) $\geq$ 1.0, alors qu'une surévaluation peut être adéquate si le taux de contamination est élevé et d supérieur à 1.0.
La fenêtre asymptotique et celle de Fryer obtiennent des résultats analogues, la première étant plus adaptée pour de faibles taux de contamination et la seconde pour des taux plus élevés.
Le rôle principal de la taille de l'échantillon est d'améliorer la qualité de l'estimation de f, qualité affinée par un choix judicieux de la fenêtre.

3.3.2.3.2 Analyse selon le rapport des variances

(Figure 3.12)

L'influence du choix de la fenêtre est étudiée après que celle de la taille de l'échantillon a été déterminée.

La taille de l'échantillon améliore considérablement la qualité de l'estimation de f pour une fenêtre fixée (figure 3.12).
Si $r^2 = 1/9$, cette amélioration est encore plus nette pour des taux élevés de contamination, car les différences entre les EQMI augmentent légèrement pour k et ε

fixés (tableau 3.6).
Si $r^2 = 9.0$, le rôle de la taille de l'échantillon peut être légèrement accentué ou diminué suivant la fenêtre utilisée. Mais l'augmentation et la diminution de l'erreur commise sont minimes entre les différentes fenêtres.

Si la figure 3.12 est analysée en ligne ou si l'on écrit la formule (3.42) par rapport à n^{-1} , on remarque à nouveau que la figure pour n = 50 est obtenue par une transformation affine de celle pour n = 100, dilatation et translation vers le haut, ou de celle de n = 20, contraction et translation vers le bas.
Ainsi, les propriétés globales de l'EQMI ne varient pas suivant la taille de l'échantillon.

Exemples : $Rap(f_2)$ pour différentes fenêtres.
(tirés du tableau 3.6)

n	fenêtre	$r^2 = 1/9$	$r^2 = 9.0$
20	0.71	5.38	0.92
50	0.59	6.98	0.86
100	0.51	8.39	0.84

Pour n = 50, étudions le rôle spécifique de la fenêtre (figure 3.12, colonne centrale).

Si $r^2 = 1/9$, l'$EQMI(f_2)$ est une fonction fortement croissante de ε. Une fenêtre étroite, k_1, obtient les résultats les meilleurs, dès que le taux de contamination est supérieur à 0.12. L'EQMI obtenue avec cette fenêtre est fortement croissante en ε dès que $\varepsilon \geq 0.20$, mais reste la meilleure. Pour la largeur de Fryer, ou une plus grande, k_4, l'$EQMI(f_2)$ est encore plus fortement croissante en ε (figure 3.12). Ainsi, pour améliorer les performances de cet estimateur, il suffit de choisir une fenêtre plus étroite (que celle de Fryer) si la densité présente un mode plus pointu (que la normale).

Si $r^2 = 9.0$, la fenêtre de Fryer obtient globalement les meilleurs résultats, dépassée par une largeur plus étroite, k_2, pour un faible taux de contamination ($\varepsilon < 0.10$), et par une fenêtre plus large, k_4, pour un taux élevé ($\varepsilon > 0.30$). Mais dans tous les cas, les différences sont minimes entre l'EQMI de la fenêtre de Fryer et la meilleure possible. Une fenêtre plus étroite, k_1, k_2, provoque un accroissement de l'EQMI qui devient alors une fonction légèrement croissante de ε, $Rap(f_2) \geq 1.0$. Si la fenêtre est plus large, k_4, que celle de Fryer, alors l'EQMI reste une fonction décroissante de ε, dont les valeurs peuvent devenir légèrement inférieures à celles obtenues par la largeur de Fryer (tableau 3.6).

En résumé,

la fenêtre de Fryer obtient de bons résultats si la densité à estimer présente un mode aplati ($r^2 > 1.0$).
Une fenêtre plus étroite, k_1, est nettement plus appropriée si la densité présente un pic, mais l'EQMI(f_2) reste alors une fonction fortement croissante du taux de contamination ε ($r^2 < 1.0$).
La taille de l'échantillon permet d'abaisser l'EQMI(f_2), et donc, d'améliorer l'estimation de f, amélioration augmentée par un choix judicieux de la fenêtre k.

En conclusion, (du paragraphe 3.3.2.3)

la qualité de l'estimation dépend principalement de la taille de l'échantillon et peut être encore améliorée par une fenêtre adéquate.
La fenêtre de Fryer, ou une valeur proche de celle-là, peut être considérée comme une largeur tout à fait performante, sauf si la répartition à estimer présente un mode très fortement prononcé. Dans ce cas, une fenêtre plus étroite est préférable pour estimer la densité f.

La fenêtre de Fryer est retenue pour l'analyse de l'efficacité du noyau normal qui suit.

Maintenant que les EQMI de g et de f_2 ont été analysées soigneusement, il nous reste à comparer les deux classes d'estimateurs à l'aide des efficacités et de leurs représentations graphiques.

3.4 Analyse des efficacités

L'efficacité est définie par les EQMI; les raisonnements suivants découlent donc des propriétés décrites des EQMI des estimateurs paramétrique et non-paramétrique, étudiées aux paragraphes précédents.
Seule l'efficacité du noyau normal est traitée, puisque l'estimateur à noyau biquadratique a été écarté pour l'étude théorique jusqu'aux simulations.

$$eff(g,f_2) = \frac{EQMI(g)}{EQMI(f_2)}$$

La variable principale est toujours le taux de contamination ε, les autres variables étant considérées comme des paramètres.
Une efficacité de 1.0 et le taux de contamination correspondant sont appelés respectivement efficacité et taux critique (de contamination). Ce dernier détermine à partir de quel degré de contamination l'estimateur non-paramétrique f_2 devient préférable à l'estimateur paramétrique g; il est noté ε_1.

L'analyse est conduite selon le schéma de l'étude de l'EQMI; pour chaque sorte de contamination, le rôle de la taille de l'échantillon est déterminé pour des valeurs particulières de d (ou de r^2) avant d'approfondir le rôle de d (ou de r^2) sur l'efficacité.

3.4.1 Efficacité du noyau normal

Vu les bons résultats globaux obtenus par la fenêtre de Fryer, celle-ci a été retenue pour notre analyse.

3.4.1.1 Efficacité selon la différence des espérances

(Figures 3.13 - 3.14)

Le rôle de la taille de l'échantillon n n'est pas du tout négligeable, puisque les EQMI des deux estimateurs en sont fortement dépendantes (figure 3.13).

Si d = 1/2, les EQMI des estimateurs g et f_2 sont pratiquement constantes et l'EQMI(g) est toujours inférieure à l'EQMI(f_2). Ainsi, l'efficacité est toujours inférieure à 1.0, et même à 0.5 (figure 3.13). L'efficacité diminue si n augmente et donc une taille élevée favorise encore davantage l'estimateur paramétrique, qui est donc préféré à l'estimateur non-paramétrique.
Si d = 3.0, l'EQMI est une fonction pratiquement constante de ε pour f_2 et fortement croissante de ε pour g. L'efficacité est ainsi une fonction fortement croissante de ε, croissance plus marquée lorsque n augmente (figure 3.13). Le taux critique ε_1 est atteint pour de faibles taux de contamination, valeur diminuant légèrement si n augmente. L'estimateur non-paramétrique est donc préférable, et une taille élevée d'échantillon accentue encore l'avantage de l'estimateur f_2.
Dans les deux cas traités, la taille de l'échantillon favorise l'estimateur le meilleur, à savoir l'estimateur paramétrique si d = 1/2 et l'estimateur non-paramétrique si d = 3.0. Les propriétés globales de l'efficacité restent identiques pour chaque taille d'échantillon.

	n	ε_1	Eff.pour ε = 0.5
d = 1/2 :	20	-----	0.46
	50	-----	0.35
	100	-----	0.30
d = 3.0 :	20	0.069	7.40
	50	0.061	13.96
	100	0.052	22.79

Le paramètre d semble influencer fortement l'efficacité et mérite donc une étude approfondie, la taille de l'échantillon étant fixé à 50.

A part la symétrie par rapport à $\varepsilon = 0.5$, les propriétés des efficacités sont totalement opposées suivant les valeurs de d, $d \leq 1.0$ et $d > 1.0$ (figure 3.14).
Si d est supérieur à 1.0, l'EQMI(g) est très sensible au taux de contamination ε, tandis que l'EQMI(f_2) est elle indépendante de ε (figure 3.14). Ainsi,l'estimateur paramétrique g se détériore dans une large mesure et perd rapidement son avantage initial. L'efficacité critique est atteinte pour des valeurs de ε de plus en plus faibles si d augmente, alors que le maximum de l'efficacité lui augmente.
Si d est plus petit que 1.0, l'efficacité est pratiquement constante et inférieure à 1.0, car les EQMI des deux estimateurs sont en fait pratiquement constantes et l'EQMI(g) est nettement inférieure à celle de f_2. La valeur $d = 1.0$ est le seuil où l'efficacité change de comportement pour $n = 50$ et devient une fonction croissante de ε.

n = 50	d	ε_1	Eff.pour $\varepsilon = 0.5$
d < 1.0	1/3	-----	0.32
	1/2	-----	0.35
d = 1.0	1.0	-----	0.77
d > 1.0	3.0	0.061	13.96
	5.0	0.033	31.31

En résumé,

pour la taille d'échantillon fixée à 50,
si $d < 1.0$, l'estimateur paramétrique g est meilleur que l'estimateur non-paramétrique f_2, $eff(g,f_2) \approx 0.3$, quel que soit le taux de contamination ε.
Si $d \geq 1.0$, l'estimateur non-paramétrique est préférable dès que $\varepsilon \geq 0.05$, et cette préférence est encore plus

nette si ε ou d augmente.
Si la taille de l'échantillon n augmente, alors elle favorise encore plus l'estimateur déjà préféré. De plus, elle détermine la valeur de d qui marque le changement du comportement de l'efficacité; si n augmente, alors le seuil critique de d diminue.

3.4.1.2 Efficacité selon le rapport de variances

(Figures 3.15 - 3.16)

Déterminons tout d'abord le rôle précis de la taille de l'échantillon, puisque celle-ci détermine grandement l'EQMI de chaque estimateur.

Si $r^2 = 1/4$, les EQMI des deux estimateurs sont des fonctions croissantes de ε, où l'influence de n est directement liée à la qualité de chaque estimateur. L'efficacité est malgré tout une fonction croissante de ε, croissance encore plus prononcée lorsque n augmente (figure 3.15). Le taux critique ε_1 est donc atteint aussi de plus en plus rapidement, tout en restant élevé ($\varepsilon \geq 0.25$). L'estimateur non-paramétrique est donc préféré si le taux de contamination est élevé.
Si $r^2 = 4.0$, l'EQMI(f_2) est constante, tandis que l'EQMI(g) est elle une fonction croissante de ε. L'efficacité est ainsi une fonction fortement croissante de ε, croissance encore plus nette lorsque n augmente (figure 3.15). Le taux critique ε_1 est atteint pour des valeurs élevées de ε ($\varepsilon \geq 0.2$). L'estimateur non-paramétrique est donc préférable si le taux de contamination est élevé.

Ainsi, dans les deux cas étudiés, l'estimateur non-paramétrique est préférable si le taux de contamination est élevé. Cette préférence s'accentue si n augmente, mais le comportement global de l'efficacité reste inchangé.

	n	ε_1	Eff.pour $\varepsilon = 0.5$
$r^2 = 1/4$:	20	----	0.96
	50	0.34	1.37
	100	0.25	1.94
$r^2 = 4.0$:	20	0.27	1.96
	50	0.25	2.82
	100	0.22	4.12

Analysons plus en détail le comportement de l'efficacité suivant le paramètre r^2, la taille étant fixée à 50.

Pour une contamination symétrique, d = 0, l'efficacité est une fonction croissante de ε, parvenant plus ou moins rapidement au taux critique ε_1 (figure 3.16). On constate de plus que certaines courbes se coupent et donc que la croissance n'évolue pas de la même manière.
Si $r^2 > 1.0$, le comportement des EQMI de g et de f_2 est totalement opposé, l'EQMI(g) étant fortement croissante, en ε, et celle de f_2 légèrement décroissante (figure 3.16). L'efficacité est donc elle aussi une fonction fortement croissante. Plus r^2 augmente, plus la croissance est forte et plus le taux critique est atteint rapidement, bien qu'assez élevé ($\varepsilon \geq 0.15$).
Si r^2 est inférieur à 1.0, les EQMI de g et de f_2 sont des fonctions croissantes de ε, croissance plus marquée pour g que pour f_2, mais diminuant légèrement lorsque ε augmente. Ainsi, l'efficacité est une fonction croissante de ε, dont la courbure change de sens (figure 3.16). Plus r^2 se rapproche de 0.0, plus l'efficacité est une fonction croissante de ε, et donc, plus le taux critique est faible, mais tout de même élevé ($\varepsilon \geq 0.25$).

n = 50	r^2	ε_1	Eff.pour $\varepsilon = 0.5$
$r^2 < 1.0$	1/9	0.23	1.51
	1/4	0.34	1.37
$r^2 > 1.0$	4.0	0.25	2.82
	9.0	0.16	4.91

En résumé,

pour une taille d'échantillon fixé à 50,
l'estimateur non-paramétrique est d'autant plus préférable que r^2 s'éloigne de 1.0. Cette tendance est plus marquée pour des densités aplaties ($r^2 > 1.0$), que pour des densités présentant un pic ($r^2 < 1.0$). Il faut cependant atteindre des taux de contamination non négligeables ($\varepsilon \geq 0.15$) pour que l'efficacité soit supérieure à 1.0.
Si la taille de l'échantillon augmente, alors elle favorise la supériorité de l'estimateur non-paramétrique.

Globalement, la supériorité de l'estimateur non-paramétrique est moins prononcée que lorsque la contamination était symétrique.

En conclusion, (du paragraphe 3.4)

Si la contamination est asymétrique et éloignée du mode principal, $d \geq 1.0$, alors l'estimateur non-paramétrique à noyau normal est préférable à l'estimateur paramétrique, tandis que pour une contamination proche du mode, l'estimateur paramétrique est meilleur que son concurrent.
Lorsque la contamination est symétrique, l'estimateur non-paramétrique est plus apte à estimer une densité dont le mode est soit aplati, soit fortement prononcé.

Une énorme différence caractérise les deux estimateurs concurrents. L'estimateur non-paramétrique s'adapte à la situation, alors que l'estimateur paramétrique est lié au modèle non contaminé par définition.

3.5 Simulations

L'objectif de ce paragraphe n'est pas uniquement de confirmer les résultats théoriques établis jusqu'ici. Cette partie nous donne un aperçu de la variance de l'erreur d'estimation et, en plus, elle nous permet d'évaluer précisément les performances du noyau biquadratique, pour lui-même et par rapport aux autres estimateurs.

Les simulations ont été conduites de la même manière qu'au chapitre précédent; la moyenne arithmétique, EQDM, et l'écart-type, EQDS, de 50 écarts quadratiques discrets ont été calculés pour les différents cas de contamination étudiés. Pour générer un échantillon aléatoire provenant d'une répartition normale contaminée, le modèle mathématique décrit au début de ce chapitre est utilisé :

1. Générer une variable pseudo-aléatoire uniforme entre 0.0 et 1.0, notée u_j.
2. Si u_j est inférieure au taux de contamination ε, alors générer une variable normale de paramètres α et $ß^2$. Si u_j est supérieure ou égale à ε, alors générer une variable normale de paramètres μ et σ^2.
3. Recommencer n fois le premier point, n représentant la taille de l'échantillon.

Les simulations sont menées sur des échantillons de taille 20, 50 et 100; quant aux paramètres d et r^2, deux valeurs leur ont été attribuées :

si $r^2 = 1.0$: $d = 1/2$, $d = 3.0$;

si $d = 0.0$: $r^2 = 1/4$, $r^2 = 4.0$.

La variable principale, le taux de contamination, prend cinq valeurs échelonnées entre 0.05 et 0.5 :

ε : 0.05 , 0.10 , 0.20 , 0.30 , 0.50 .

Pour les estimateurs g et f_2, les valeurs des EQDM n'ont pas été reportées sur les graphes ,correspondant des EQMI car les écarts entre les valeurs théoriques et simulées sont petits et les points représentant les EQDM se confondraient avec ceux des EQMI.
Par contre, pour le noyau biquadratique, une représentation simultanée des EQDM et des EQMI est possible et source de nombreux renseignements (figures 3.17 - 3.20). Les résultats numériques de ces simulations sont contenues dans les tableaux 3.7 - 3.28 qui nous fournissent la substance essentielle de ce paragraphe.

Soit f_n un estimateur de densité. Posons :

$$R(f_n) = \frac{EQMI(f_n)}{EQDM(f_n)}$$

Ce rapport entre l'EQMI et l'EQDM nous permet de mesurer et de quantifier les différences entre les résultats théoriques et simulés.

3.5.1 Analyse de l'EQDM de l'estimateur paramétrique

Les valeurs des EQDM simulées sont proches des valeurs des EQMI, avec bien entendu quelques fluctuations dues principalement à l'échantillonnage. D'une manière globale, les simulations confirment les résultats théoriques : même pour de faibles taux de contamination, l'EQDM(g) est très sensible à toutes variations de r^2 et de d, si celui-ci est supérieur à 1.0. Si d est inférieur à 1.0, alors l'estimateur g estime correctement la densité f.
L'EQDS est inférieur à l'EQDM en général, augmente jusqu'à un taux de contamination de 0.25 - 0.30 pour décroître ensuite. En fait, plus l'estimation est médiocre, plus l'EQDS est faible, car l'estimation est mauvaise pour tous les échantillons. Si d = 1/2, les EQDS sont légèrement supérieurs aux EQDM, car, dans ce cas, ce sont les EQDM qui sont peu élevés.

3.5.2 Analyse de l'EQDM des estimateurs non-paramétriques

En supposant toujours que le statisticien ignore la contamination, dans une première étape, la fenêtre asymptotique et de Fryer a été choisie respectivement pour le noyau biquadratique (K_1) et normal (K_2). Dans un second temps, d'autres valeurs ont été attribuées aux fenêtres pour étudier leur influence sur les EQDM principalement.

3.5.2.1 EQDM pour une fenêtre donnée

Pour le noyau biquadratique, les EQDM obtenus sont nettement inférieurs aux valeurs théoriques correspondantes. Ces différences ne doivent pas être uniquement imputées à des erreurs d'échantillonnage, mais plutôt à l'inexactitude de la formule asymptotique utilisée pour cet estimateur. $R(f_1)$ est nettement supérieur à 1.0 et dépend principalement de la taille de l'échantillon n.

	d varie	r^2 varie
n = 20	$R(f_1) \geq 1.8$	$R(f_1) \geq 2.0$
n = 50	$R(f_1) \approx 1.6$	$R(f_1) \approx 1.5$
n = 100	$R(f_1) \approx 1.7$	$R(f_1) \approx 1.3$

Ainsi, la formule asymptotique (3.23) n'est pas appropriée pour l'ensemble des densités normales contaminées par une autre densité normale. L'EQMI théorique surestime donc nettement l'erreur commise. Néanmoins, le comportement des EQDM est semblable à celui de l'EQMI, c-à-d stabilité par rapport aux variations de d et de r^2, si ce dernier est supérieur à 1.0, et forte sensibilité de l'EQDM si r^2 est inférieur à 1.0 (figures 3.17 - 3.20). Les EQDS dépendent principalement de la taille de l'échantillon et sont légèrement inférieurs aux EQDM, sauf si r^2 est inférieur à 1.0, car l'EQDM est lui très élevé dans ce dernier cas.

Pour le noyau normal, les résultats des EQDM sont proches des valeurs des EQMI correspondantes, tantôt supérieurs, tantôt inférieurs, mais toujours dans un intervalle acceptable. Les EQDS sont eux aussi inférieurs aux EQDM comme pour le noyau biquadratique et dépendent de la taille de l'échantillon avec toujours la même exception, et pour la même raison, $r^2 < 1.0$.

Ces premières constatations nous conduisent à la question suivante : vu les résultats obtenus par le noyau biquadratique lors de ces simulations, l'un des noyaux est-il préférable ?

Si la contamination est asymétrique et d est inférieur à 1.0, alors les EQDM de f_1 sont toujours plus petits que ceux de f_2, tandis que la constatation inverse est vraie pour les EQDS. Mais les différences sont minimes et les noyaux peuvent être considérés comme équivalents.

Si d est supérieur à 1.0, les valeurs obtenues pour les EQDM sont très semblables pour les deux noyaux, le noyau biquadratique étant meilleur pour de faibles taux de contamination et le noyau normal, pour des taux élevés. La fenêtre avantage l'un ou l'autre noyau suivant le degré de contamination. A part une ou deux exceptions, les EQDS de f_2 sont légèrement inférieurs à ceux de f_1, mais les différences restent minimes.

Si la contamination est symétrique et r^2 inférieur à 1.0, alors les EQDM du noyau biquadratique sont toujours inférieurs à ceux du noyau normal, avec des différences notables et plus marquées encore pour des taux élevés de contamination. Ces écarts importants sont dus au choix de la fenêtre qui est mieux adaptée pour le noyau biquadratique que pour le noyau normal dans cette situation (cf paragraphe 3.5.2.2). Les EQDS de f_2 sont très légèrement inférieurs aux EQDS de f_1, mais là aussi, les écarts sont minimes.

Si r^2 est supérieur à 1.0, alors les EQDM sont très proches les uns des autres, l'avantage étant au noyau biquadratique pour de faibles taux de contamination et au noyau normal pour des taux plus élevés. La fenêtre, plus ou moins bien appropriée suivant le taux de contamination, influence directement le comportement des estimateurs. Les EQDS sont eux très proches et aucune règle précise ne ressort des résultats.

En résumé,

avec les fenêtres proposées, le noyau biquadratique estime plus précisément une densité pointue ($r^2 < 1.0$), tandis que le noyau normal semble mieux adaptée pour une densité ayant un mode aplati ($r^2 > 1.0$).
Si la contamination est asymétrique, les noyaux sont équivalents.
Dans les deux cas, les EQDS sont du même ordre de grandeur.
La fenêtre joue un rôle important, puisqu'elle favorise l'un ou l'autre noyau.

Une étude plus complète est nécessaire pour déterminer l'influence exacte de la fenêtre suivant les cas.

3.5.2.2 EQDM pour différentes fenêtres

Les paramètres d et r^2 prennent les mêmes valeurs que lors de l'analyse théorique de l'EQMI(f_2), à savoir :

si $r^2 = 1.0$: $d = 1/3$, $d = 1.0$, $d = 5.0$;

si $d = 0.0$: $r^2 = 1/9$, $r^2 = 9.0$.

Pour chaque échantillon, quatre fenêtres différentes ont été testées sur chaque noyau (cf paragaphe 3.3.2.3). La moyenne et l'écart-type sur 50 échantillons ont été calculés uniquement pour des échantillons de taille 50 et les résultats se trouvent dans les tableaux 3.19 - 3.28.

Pour le noyau normal, les résultats obtenus des EQDM correspondent très exactement aux valeurs des EQMI, parfois supérieurs, parfois inférieurs; mais, dans tous les cas, les différences sont négligeables et le comportement des EQDM est donc semblable à celui des EQMI (tableaux 3.24 - 3.28).

La fenêtre de Fryer obtient globalement les résultats les meilleurs. Elle est dépassée par une fenêtre plus étroite pour une contamination proche du mode, $d < 1.0$, ou pour une contamination accentuant le mode , $r^2 < 1.0$. Une fenêtre plus large est préférable si le mode est aplati, $r^2 > 1.0$ et $\varepsilon \geq 0.3$. Mais, les différences entre les résultats obtenus par la fenêtre de Fryer et la fenêtre la meilleure possible sont minimes.

Pour le noyau biquadratique, seule une comparaison avec le noyau normal est possible, aucune étude théorique n'ayant été menée sur les EQMI de ce noyau au vue des différences entre les résultats théoriques et les premières simulations (tableaux 3.19 - 3.23).

Pour les quatres largeurs étudiées, les valeurs des EQDM obtenues par le noyau biquadratique sont légèrement inférieures à celles du noyau normal, sauf si la fenêtre est large, ou, si le mode de la densité est fortement prononcé ($r^2 < 1.0$).
Les fenêtres obtenant l'EQDM le plus petit correspondent pour les deux noyaux; les différences restent toutefois minimes entre les EQDM des deux noyaux et ne défavorisent pas le noyau normal.
Les EQDS sont semblables pour les deux noyaux, bien que ceux obtenus par le noyau biquadratique sont toujours très légèrement inférieurs.

En résumé,

la largeur de Fryer, ou proche de celle-ci, et son équivalent pour le noyau biquadratique obtiennent globalement de bons résultats, une fenêtre plus étroite ou plus large pouvant leur être préférée si la densité présente un pic ou si la différence des espérances est faible.

A la vue des simulations, le noyau biquadratique est aussi bon si ce n'est meilleur que le noyau normal, démontrant ainsi que la formule asymptotique de l'EQMI est mauvaise et pas appropriée pour les densités normales contaminées par une autre densité normale.

Il nous reste à analyser les performances des estimateurs non-paramétriques par rapport à celles de l'estimateur paramétrique lors de ces simulations.

3.5.3 Analyse des efficacités simulées

La fenêtre choisie est celle de Fryer pour le noyau normal et la largeur asymptotique pour le noyau biquadratique.

Vu les différences entre les EQMI et les EQDM pour le noyau biquadratique, l'analyse entre les efficacités théoriques et simulées a été abandonnée pour ce noyau. Etudions donc les résultats obtenus par le noyau normal avant de les comparer avec les efficacités simulées du noyau biquadratique. Les résultats sont regroupés avec ceux des EQDM et se trouvent donc dans les tableaux 3.7 à 3.18.

Pour le noyau normal, les efficacités simulées correspondent aux valeurs théoriques, ce qui n'est pas surprenant puisque les EQDM des deux estimateurs considérés correspondent aux EQMI. Les différences sont dues à l'échantillonnage et non à des erreurs des valeurs théoriques. Elles peuvent être aussi plus élevées que pour les EQDM, car les EQDM des deux estimateurs peuvent réagir de manière opposée. L'estimateur non-paramétrique f_2 est préférable à l'estimateur paramétrique pour des taux élevés de contamination si celle-ci est symétrique (d = 0.0); cette préférence est plus marquée pour $r^2 = 4.0$ que pour $r^2 = 1/4$, et augmente encore lorsque n augmente.

Si d = 3.0, l'estimateur non-paramétrique f_2 est meilleur que l'estimateur g quel que soit le taux de contamination.
Si d = 1/2, l'estimateur paramétrique g est nettement plus performant que l'estimateur non-paramétrique f_2; l'efficacité simulée est inférieure à 0.5.

Les efficacités simulées du noyau biquadratique sont tantôt supérieures, tantôt inférieures à celles du noyau normal et suivent fidèlement l'évolution des EQDM respectifs (cf paragraphe 3.5.2.1). Mais les écarts sont insignifiants et dus au choix de la fenêtre qui avantage l'un ou l'autre estimateur suivant le taux de contamination.

En résumé,

les estimateurs non-paramétriques, f_1 et f_2, sont préférables à l'estimateur paramétrique g, sauf si la contamination est proche du mode principal.

En conclusion,

les simulations confirment les résultats théoriques pour le noyau normal, le comportement des EQDM et des efficacités simulées étant identique à celui des EQMI et des efficacités théoriques.
Pour le noyau biquadratique, de très grandes différences entre les valeurs simulées et théoriques démontrent que la formule asymptotique n'est pas adaptée et est imprécise.
De plus le choix du noyau n'est pas prédominant pour estimer une densité normale contaminée par une seconde densité normale.

CONCLUSIONS

Tout au long de ce travail, nous avons parlé de fonctions de densité et d'erreurs associées aux estimateurs utilisés, sans pour autant nous intéresser directement à leur représentation graphique.

Afin de donner un aperçu de la variété des densités traitées, f, et des performances des estimateurs paramétriques, g, et non-paramétriques, f_1 et f_2, quelques graphes de la densité exacte et d'estimations ont été tracés à partir d'un échantillon aléatoire de taille 50 et d'une discrétisation en 300 points équidistants de l'intervalle $[\mu - 5.0, \alpha + 5.0]$ (figures 3.21 et 3.22). Le taux de contamination ε et les paramètres d, r^2 et k prennent les mêmes valeurs que lors des analyses des EQMI et des simulations, à savoir :

pour ε : $\varepsilon = 0.1$, $\varepsilon = 0.3$, $\varepsilon = 0.5$;

si $r^2 = 1.0$: $d = 1/2$, $d = 3.0$;

si $d = 0.0$: $r^2 = 1/4$, $r^2 = 4.0$;

pour la fenêtre : $k_1 = 1.27$, $k_2 = 0.59$.

Si la contamination est proche du mode, $d = 1/2$, alors ce genre de contamination est difficilement décelable et la densité exacte s'apparente à une densité normale (figure 3.21). Les estimateurs à noyau, biquadratique et normal, ont beaucoup de difficultés à redonner convenablement le mode et le sous-estiment. C'est pour ce genre de contamination que l'estimateur paramétrique g est préférable aux estimateurs à noyau.

Pour les autres sortes de contamination, $d = 3.0$ ou $d = 0.0$ avec $r^2 = 1/4$ ou $r^2 = 4.0$, l'estimateur paramétrique g paraît satisfaisant pour de faibles taux de contamination, $\varepsilon \leq 0.1$; mais il devient totalement fantaisis-

te et non-représentatif de la densité exacte lorsque le taux de contamination augmente, $\varepsilon = 0.3$ et $\varepsilon = 0.5$. Les estimateurs non-paramétriques s'adaptent tant bien ($d = 3.0$ et $r^2 = 4.0$) que mal ($r^2 = 1/4$) aux changements de forme de la densité à estimer. D'une manière générale, ils sous-estiment le mode de cette dernière, sous-évaluation d'une part moins marquée pour le noyau biquadratique que pour le noyau normal, et d'autre part, plus nette lorsque le mode est fortement prononcé, $r^2 = 1/4$.

Sur la base des analyses théoriques , des simulations et des graphes des densités exactes et des estimations , on peut conclure que les estimateurs non-paramétriques, f_j, sont préférables à l'estimateur paramétrique,g , car ils s'adaptent avec plus ou moins de succès aux changements de forme de la densité exacte. Ils fournissent toujours une densité parente de la densité exacte, bien qu'ils en sous-estiment généralement le ou les modes.

Par construction, l'estimateur paramétrique ne peut fournir qu'une densité normale non contaminée. Son EQMI est donc sensible à tout genre de contamination.

Quant au noyau, son choix n'est pas prédominant si l'on ne s'intéresse qu'à l'erreur globale commise, l'EQMI. Mais, le noyau biquadratique estime plus correctement le mode de la densité et pourra donc être adopté si le statisticien a besoin d'une estimation plus précise au mode et dans son voisinage.

Pour compléter ces conclusions et aider le statisticien dans sa prise de décision, les taux critiques de contamination ε_1 pour le noyau normal ont été calculés numériquement à l'aide de la procédure ZREAL1 de la bibliothèque de IMSL pour les tailles d'échantillon inférieures à 100. Les couples (n,ε_1) ont été reportés dans un système d'axes et la courbe ainsi obtenue est appelée frontière des taux critiques; la région située au-dessus

de la frontière, région hachurée sur les graphes, correspond aux points où l'efficacité est supérieure à 1.0, c-à-d où l'estimateur non-paramétrique est préférable à son concurrent, l'estimateur paramétrique (figures 3.23 - 3.25).

Les paramètres d et r^2 prennent les valeurs suivantes :

pour d : 0.0 , 1/2 , 1.0 , 1.5 ou 2.0 , 3.0 et 5.0 ;

pour r^2 : 1/9 , 1/4 , 1.0 , 4.0 et 9.0 .

Ces graphes résument très clairement les idées de ce travail et montrent en plus la relation entre les deux paramètres d et r^2.

En plus du cas particulier $r^2 = 1.0$, on distingue deux formes différentes de graphes :

(1) $r^2 < 1.0$ et $d \leq 1.0$,
(2) $r^2 > 1.0$ et ($r^2 < 1.0$ et $d > 1.0$) .

Pour $r^2 = 1.0$, le déplacement de la frontière des taux critiques vers le bas est important. Dès que la différence réduite des espérances est élevée, $d \geq 2.0$, la taille de l'échantillon n'a plus aucune influence. Le taux critique est faible, $\varepsilon_1 \leq 0.15$ et l'estimateur non-paramétrique est rapidement préférable à l'estimateur paramétrique.

Si $r^2 < 1.0$ et $d \leq 1.0$, une taille d'échantillon élevée favorise nettement l'estimateur non-paramétrique qui devient préférable au sens de l'EQMI à l'estimateur paramétrique. Mais le taux critique est supérieur à 0.15, taux non négligeable. La frontière est translatée vers le bas, fortement lorsque r^2 diminue, et, plus modérément lorsque d augmente jusqu'à la valeur 1.0.

Dans les autres cas, la taille de l'échantillon ne joue qu'un rôle secondaire, car le taux critique ne dépend pratiquement que de d et r^2.
Pour $r^2 > 1.0$ et $0.0 \leq d \leq 1.0$, le taux critique est plus fortement dépendant de la croissance de r^2 que de celle de d. Néanmoins, il s'abaisse régulièrement et varie entre 0.15 et 0.30.
Si $r^2 > 1.0$ et $d > 1.0$, alors le taux critique n'est influencé que par la différence des espérances d, le rôle de r^2 étant totalement négligeable. Il est déjà atteint pour de faibles taux de contamination, $\varepsilon_1 \leq 0.1$.

Ainsi, le statisticien peut se référer à ces graphes pour une bonne prise de décisions lorsqu'il est confronté à une contamination de densités normales.

ANNEXES

Annexe 1

A.1.1 Formules de multiplication de densités normales.

1) $N(\mu,\sigma^2;x)\cdot N(\alpha,\beta^2;x) = N(c,d;x)\ N(\alpha,\sigma^2+\beta^2;\mu)$

où $c = (\mu\beta^2+\alpha\sigma^2)/(\sigma^2+\beta^2)$,

$d = \sigma^2\cdot\beta^2/(\sigma^2+\beta^2)$.

Cas particuliers

2) Si $\mu = \alpha$, alors

$N(\mu,\sigma^2;x)\cdot N(\mu,\beta^2;x) = [2\Pi(\sigma^2+\beta^2)]^{-1/2}\ N(\mu,d;x)$.

3) Si $\sigma^2 = \beta^2$, alors

$N(\mu,\sigma^2;x)\cdot N(\alpha,\sigma^2;x) = N[(\mu+\alpha)/2,\sigma^2/2;x]\ N(\alpha,2\sigma^2;\mu)$.

4) Si $\mu = \alpha$ et $\sigma^2 = \beta^2$, alors

$N(\mu,\sigma^2;x)\cdot N(\mu,\sigma^2;x) = (4\Pi\sigma^2)^{-1/2}\ N(\mu,\sigma^2/2;x)$.

A.1.2 Formules d'intégrations de produits de densités normales.

1) $$\int_{-\infty}^{+\infty} N(\mu,\sigma^2;x)\cdot N(\alpha,\beta^2;x)\,dx = N(\alpha,\sigma^2+\beta^2;\mu) \ .$$

2) Si $\mu = \alpha$, alors

$$\int_{-\infty}^{+\infty} N(\mu,\sigma^2;x)\cdot N(\alpha,\beta^2;x)\,dx = [2\pi(\sigma^2+\beta^2)]^{-1/2} \ .$$

3) Si $\sigma^2 = \beta^2$, alors

$$\int_{-\infty}^{+\infty} N(\mu,\sigma^2;x)\cdot N(\alpha,\beta^2;x)\,dx = N(\alpha,2\sigma^2;\mu) \ .$$

4) Si $\mu = \alpha$ et $\sigma^2 = \beta^2$, alors

$$\int_{-\infty}^{+\infty} N(\mu,\sigma^2;x)\cdot N(\mu,\sigma^2;x)\,dx = (4\pi\sigma^2)^{-1/2} \ .$$

Annexe 2 : approximation de $\int_0^{+\infty} P(t)\, X_n^2(t)\, dt$

où $X_n^2(t)$ est la fonction de densité de la répartition du chi-carré à n degrés de liberté,

$$P(t) = (1+t/n)^{-1/2} .$$

Pour une variable t répartie selon une loi du chi-carré, son espérance mathématique vaut :

$$E(t) = n ,$$

et ses différents moments centrés d'ordre j, notés α_j, valent respectivement :

$\alpha_1 = 0$;

$\alpha_2 = 2n$;

$\alpha_3 = 8n$;

$\alpha_4 = 12n(n+4)$;

$\alpha_5 = 32n(n+2)$.

D'une manière générale, α_j est un polynôme de degré $[j/2]$ où $[x]$ représente la partie entière de x.

En effet, la formule récurrente des moments est :

$\alpha_{j+1} = 2j\ (\alpha_j + n\cdot\alpha_{j-1})$ (cf M.Kendall et A.Stuart).

Si $P^{(k)}$ désigne la $k^{ième}$ dérivée de P par rapport à t, alors

$$1/k!\ P^{(k)}(t) = D_k\ (1+t/n)^{-(2k+1)/2}\ n^{-k}$$

où $D_k = (-1)^k \dfrac{1\ 3\ 5\ \ldots(2k-1)}{2\ 4\ 6\ \ldots\ 2k}$ $\qquad k = 1,2,\ldots$

Développons la fonction P(t) en série de Taylor au point $t_0 = n$, qui est donc l'espérance de la variable t. Or,

$$P(t_0) = P(n) = 2^{-1/2} ,$$

$$1/k!\ P^{(k)}(n) = D_k \quad 2^{-(2k+1)/2}\ n^{-k} \qquad\qquad k = 1,2,...$$

Alors,

$$\int_0^{+\infty} (2\Pi\sigma^2)^{-1/2}\ P(t)\ X_n^{\,2}(t)\ dt$$

$$= (2\Pi\sigma^2)^{-1/2}\ \Big[\int_0^{+\infty} P(n)\ X_n^{\,2}(t)\ dt$$

$$+ \sum_k 1/k!\ P^{(k)}(n) \int_0^{+\infty} (t-n)^k\ X_n^{\,2}(t)\ dt\ \Big]$$

$$= (2\Pi\sigma^2)^{-1/2}\ \Big[\ 1/\sqrt{2} + \sum_{k=1}^{4} D_k\ (2n)^{-k}\ \alpha_k\ \Big] + O(n^{-3})$$

$$= (4\Pi\sigma^2)^{-1/2}\ \Big[\ 1 + \frac{1\ 3}{2\ 4}\ (2n)^{-2}\ (2n) - \frac{1\ 3\ 5}{2\ 4\ 6}\ (2n)^{-3}\ (8n)$$

$$+ \frac{1\ 3\ 5\ 7}{2\ 4\ 6\ 8}\ (2n)^{-4}\ 12n(n+4)\ \Big] + O(n^{-3})$$

$$= (4\Pi\sigma^2)^{-1/2}\ \Big[\ 1 + 3/(16n) - 5/(16n^2)$$

$$+ 105/(512n^2)\ \Big] + O(n^{-3})$$

$$= (4\Pi\sigma^2)^{-1/2}[1 + 3/(16n) - 55/(512n^2)] + O(n^{-3}) \qquad (A2.1)$$

Cette dernière expression est la formule (2.29).

Pour obtenir une approximation de l'intégrale (2.24), on procède de manière identique en développant la fonction $Q(t) = t^{-1/2}$ en série de Taylor autour du point $t_0 = n$.

$$1/k!\ Q^{(k)}(t) = D_k\ t^{-(2k+1)/2}\ n^{-k}$$

$$\text{où} \quad D_k = (-1)^k\ \frac{1\ 3\ 5\ \ldots(2k-1)}{2\ 4\ 6\ \ldots\ 2k} \qquad k = 1,2,\ldots$$

Développons la fonction $Q(t)$ autour du point $t_0 = n$;

$$Q(t_0) = Q(n) = n^{-1/2}\ ,$$

$$1/k!\ Q^{(k)}(n) = D_k\ n^{-(2k+1)/2} \qquad k = 1,2,\ldots$$

Alors, l'intégrale (2.24) devient :

$$\int_0^{+\infty} (4\Pi\sigma^2)^{-1/2}\ n^{-1/2}\ Q(t)\ X_n^{\,2}(t)\ dt$$

$$= n^{1/2}\ (4\Pi\sigma^2)^{-1/2}\ \Big[\ \int_0^{+\infty} Q(n)\ X_n^{\,2}(t)\ dt$$

$$+\ \sum_k\ 1/k!\ Q^{(k)}(n) \int_0^{+\infty} (t-n)^k\ X_n^{\,2}(t)\ dt\ \Big]$$

$$= n^{1/2}\ (4\Pi\sigma^2)^{-1/2}\ \Big[\ n^{-1/2} + \sum_{k=1}^{4} D_k\ n^{-(2k+1)/2}\ \alpha_k\ \Big]$$

$$+\ O(n^{-3})$$

$$= (4\Pi\sigma^2)^{-1/2}\ \Big[\ 1 + \frac{1\ 3}{2\ 4}\ n^{-2}\ (2n) - \frac{1\ 3\ 5}{2\ 4\ 6}\ n^{-3}\ (8n)$$

$$+ \frac{1\ 3\ 5\ 7}{2\ 4\ 6\ 8}\ n^{-4}\ 12n(n+4)\ ] + O(n^{-3})$$

$$= (4\Pi\sigma^2)^{-1/2}\ [\ 1 + 3/(4n) - 5/(2n^2)$$

$$+ 105/(32n^2)\ ] + O(n^{-3})$$

$$= (4\Pi\sigma^2)^{-1/2}[1 + 3/(4n) + 25/(32n^2)] + O(n^{-3})\ .\ (A2.2)$$

Ceci est l'expression approximative de l'intégrale (2.24). En regroupant les deux termes trouvés (A2.1) et (A2.2) selon les formules (1.9) et (1.10), on obtient l'approximation de l'EQMI(g_2) donnée en (2.21).

$$EQMI(g_2) = (4\Pi\sigma^2)^{-1/2}\ [3/(8n) + 255/(256n^2)] + O(n^{-3})\ .$$

Annexe 3 : approximation de $\int_0^{+\infty} R(s)\, X_{n-1}^2(s)\, ds$

où $X_{n-1}^2(t)$ est la fonction de densité de la répartition du chi-carré à (n-1) degrés de liberté,

$$R(s) = [1+1/n+s/(n-1)]^{-1/2} .$$

Pour une variable s répartie selon une loi du chi-carré, son espérance mathématique vaut :

$$E(s) = n-1 ,$$

et ses différents moments centrés d'ordre j, notés $ß_j$, valent respectivement :

$$ß_1 = 0 ;$$

$$ß_2 = 2(n-1) ;$$

$$ß_3 = 8(n-1) ;$$

$$ß_4 = 12(n-1)(n-1+4) = 12(n-1)(n+3) ;$$

$$ß_5 = 32(n-1)(n-1+2) = 32(n-1)(n+1) .$$

D'une manière générale, $ß_j$ est un polynôme de degré [j/2] où [x] représente la partie entière de x.

En effet, $ß_{j+1} = 2j\,[ß_j + (n-1)\, ß_{j-1}]$.

Si $R^{(k)}$ désigne la $k^{ième}$ dérivée de R par rapport à s, alors

$$1/k!\; R^{(k)}(s) = D_k\, [1+1/n+s/(n-1)]^{-(2k+1)/2}\, (n-1)^{-k}$$

où $D_k = (-1)^k \dfrac{1\ 3\ 5\ \ldots(2k-1)}{2\ 4\ 6\ \ldots\ 2k}$ $\qquad k = 1,2,\ldots$

Développons la fonction R(s) en série de Taylor au point $s_0 = n-1$, qui est donc l'espérance de la variable s. Or,

$$R(s_0) = R(n-1) = (2+1/n)^{-1/2} \ ,$$

$$1/k!\ R^{(k)}(n-1) = D_k\ 2^{-1/2}\ [1+1/(2n)]^{-(2k+1)/2}\ [2(n-1)]^{-k}$$

$$k = 1,2,\ldots$$

Alors,

$$\int_0^{+\infty} (2\Pi\sigma^2)^{-1/2}\ R(s)\ X_{n-1}{}^2(s)\ ds$$

$$= (2\Pi\sigma^2)^{-1/2}\ \{ \int_0^{+\infty} R(n-1)\ X_{n-1}{}^2(s)\ ds$$

$$+ \sum_k 1/k!\ R^{(k)}(n-1) \int_0^{+\infty} [s-(n-1)]^k\ X_{n-1}{}^2(s)\ ds\ \}$$

$$= (2\Pi\sigma^2)^{-1/2}\ \{ (2+1/n)^{-1/2}$$

$$+ \sum_{k=1}^{4} D_k\ 2^{-1/2}\ [2(n-1)]^{-k}\ [1+1/(2n)]^{-(2k+1)/2}\ \beta_k \}$$

$$+ O(n^{-3})$$

$$= (4\Pi\sigma^2)^{-1/2}\ \{ [1+1/(2n)]^{-1/2}$$

$$+ \frac{1}{2}\,\frac{3}{4}\ [1+1/(2n)]^{-5/2}\ [2(n-1)]^{-2}\ 2(n-1)$$

$$- \frac{1\ 3\ 5}{2\ 4\ 6} [1+1/(2n)]^{-7/2} [2(n-1)]^{-3} 8(n-1)$$

$$+ \frac{1\ 3\ 5\ 7}{2\ 4\ 6\ 8} [1+1/(2n)]^{-9/2} [2(n-1)]^{-4} 12(n-1)(n+3)]$$

$$+ O(n^{-3}) . \qquad \text{(A3.1)}$$

Or,

$$[1+1/(2n)]^{-1/2} = 1 - 1/(4n) + 3/(32n^2) + O(n^{-3}) ;$$

$$[1+1/(2n)]^{-5/2} [2(n-1)]^{-1} = 1/n - 1/(4n^2) + O(n^{-3}) ;$$

$$[1+1/(2n)]^{-7/2} [2(n-1)]^{-2} = 1/n^2 + O(n^{-3}) ;$$

$$[1+1/(2n)]^{-9/2} [2(n-1)]^{-2} = 1/n^2 + O(n^{-3}) .$$

En substituant ces expressions dans la formule (A3.1), on obtient une approximation en puissances de n^{-1} :

$$= (4\pi\sigma^2)^{-1/2}\{ 1 - 1/(4n) + 3/(32n^2) + 3/16 [1/n - 1/(4n^2)]$$

$$- 5/(16n^2) + 105/(512n^2)] + O(n^{-3})$$

$$= (4\pi\sigma^2)^{-1/2} [1 - (1/4 - 3/16)/n$$

$$- (3/32 - 3/64 - 5/16 + 105/512)/n^2] + O(n^{-3})$$

$$= (4\pi\sigma^2)^{-1/2} [1 - 1/(16n) - 31/(512n^2)] + O(n^{-3}). \qquad \text{(A3.2)}$$

Cette dernière expression est la formule (2.39).

Par analogie à l'expression (A2.2), une approximation de l'expression (2.34) vaut :

$$[(n-1)/(4\Pi\sigma^2)]^{1/2} \int_0^{+\infty} v^{-1/2}\, X_{n-1}^2(v)\, dv$$

$$= (4\Pi\sigma^2)^{-1/2} \{ 1 + 3/[4(n-1)] + 25/[32(n-1)^2] \}$$

$$+ O(n^{-3}) .$$

En développant selon les puissances de n^{-1} les expressions $(n-1)^{-1}$ et $(n-1)^{-2}$, on obtient l'approximation désirée :

$$= (4\Pi\sigma^2)^{-1/2} [1 + 3/(4n) + 49/(32n^2)] + O(n^{-3}) . \quad (A3.3)$$

En substituant les expressions (A3.2) et (A3.3) dans les formules (1.9) et (1.10) de l'EQMI, on obtient le résultat énoncé dans la formule (2.31) :

$$EQMI(g_3) = (4\Pi\sigma^2)^{-1/2} [7/(8n) + 423/(256n^2)] + O(n^{-3}).$$

BIBLIOGRAPHIE

1. Abramowitz, M., Stegun, I.A. (1965). Handbook of Mathematical Functions. Dover Publication, New-York.

2. Anderson, G.D. (1969). A comparison of methods for estimating a probability density function. Thèse de doctorat. Université de Washington.

3. Ash, R.B. (1972). Real Analysis and Probability. Academic Press, New-York.

4. Bauer, W. (1975). On pointwise nonparametric estimation of a density function. Period.Math.Hungar. (6), 59-67.

5. Bhattacharya, P.K. (1967). Estimation of a probability density function and its derivates. Sankhya Ser.A (29), 373-382.

6. Bickel, P., Rosenblatt, M. (1973). On some global measure of the deviation of density function estimates. Ann.Statist (1), 1071-1095. Correction (1975). Ann.Statist. (3), 1370.

7. Deheuvels, P., Hominal, P. (1980). Estimation automatique de la densité. Rev.Statist.Appl. (25), 5-42.

8. Devroye, L., Györfy, L. (1985). Nonparametric Density Estimation : The L_1 View. Wiley, New-York.

9. Devroye, L., Penrod, C.S. (1984). The consistency of automatic kernel density estimates. Ann.Statist. (12), 1231-1249.

10. Dodge, Y., Lejeune, M. (1982). Some difficulties involving the nonparametric estimation of a density function. Rapport technique au FNRS. Univ.Neuchâtel.

11. Durbin, J. (1980). Approximation for densities of sufficient estimators. Biometrika (69), 29-46.

12. Epanechnikov, V.A. (1969). Nonparametric estimation of a multidimensional probability density. Theory Prob.Appl. (14), 153-158.

13. Feller, W. (1968). An Introduction to Probability Theory and its Applications. Vol. I-II. Wiley, New-York.

14. Földes, A., Révéz, P. (1974). A general method for density estimation. Studia Sci.Math.Hungar. (9), 443-452.

15. Fryer, M.J. (1976). Some errors associated with the non-parametric estimation of density functions. J.Inst.Math.Appl. (18), 371-380.

16. Gasser, T., Müller, H.G. (1978). Kernel estimation of regression functions. Lectures Notes in Mathematics N° 757. Springer-Verlag.

17. Gasser, T., Härdle, W. (1984). Robust nonparametric function fitting. J.R.Statist.Soc.Ser.B (46), 42-51.

18. Geman, S., Hwang, C.-R. (1982). Nonparametric maximum likelihood estimation by the method of sieves. Ann.Statist. (10), 401-414.

19. Guttmann, H., Wertz, W. (1976). Note on estimating normal densities. Sankhya Ser.B (38), 231-236.

20. Hall. P (1983). Large sample optimality of least squares cross-validation in density estimation. Ann. Statist. (11), 1156-1174.

21. Hall, P. (1984). On optimal property of kernel estimators of a probability density. J.R.Statist.Soc.Ser.B (46), 134-138.

22. Huber, P.J. (1981). Robust Statistics. Wiley, New-York.

23. IMSL Library (1985). International mathematical and statistical libraries. Inc. Houston, Texas.

24. Johnson, N.L., Kotz, S. (1970). Distributions in Statistics. Vol. I-II-III. Wiley, New-York.

25. Kendall, M., Stuart, A. (1977). The Advanced Theory of Statistics. Vol. I-II-III. Griffin, London.

26. Klebanov, L.B. (1977). Parametric estimates of density function and a characterisation of the families of distributions with a location parameter admitting sufficient statistics (in russian). Zap. Nauch.Sem. (LOMI).

27. Kumar, T.K., Markmann J.M. (1975). Estimation of probability density functions : a Monte-Carlo comparison of parametric and nonparametric methods (Preprint).

28. Lang, S. (1969). Analysis I. Addison-Wesley.

29. Lejeune, M.(1982).Estimation de densité à noyau variable. Rapport technique au FNRS. Université de Neuchâtel.

30. Londhe, A.R., Gentle, J.E. (1979). Density estimation using kernels over variable size windows. Proc.of ASA (1979). Statist.Comp.Section , 354-358.

31. Nadaraja, E. (1965). On nonparametric estimation of density function and regression. Theory Prob.Appl. (10), 139-142.

32. Parzen, E. (1962). On estimation of a probability function and mode. Ann.Math.Statist. (33), 1065-1076.

33. Prakasa Rao , B.L.S. (1983). Nonparametric Functional Estimation. Academic Press, Orlando.

34. Quandt, R.E., Ramsey, J.B. (1978). Estimating mixtures of normal distributions and switching regressions. J.Amer.Statist.Ass. (73), 730-752.

35. Randles, R.H. (1982). On the asymptotic normality of statistics with estimated parameters. Ann. Statist. (10), 462-474.

36. Revesz, P. (1972). On empirical density function. Period.Math.Hungar. (2), 85-110.

37. Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. Ann.Math. Statist. (27), 832-837.

38. Rosenblatt, M. (1971). Curves estimates. Ann.Math.Statist. (42), 1-14.

39. Scott, D.W., Factor, L.E. (1981). Monte-Carlo study of three data-based nonparametric probability density function. J.Amer.Statist.Ass. (76), 9-15.

40. Silverman, B.W. (1978). Choosing the window width when estimating a density. Biometrika (65), 1-11.

41. Silverman, B.W. (1986). Density Estimation for Statistics and Data Analysis. Chapman and Hall, New-York.

42. Stone, C.J. (1984). An asymptotically optimal window selection rule for kernel density estimates. Technical Report N° 31, Berkeley Univ.

43. Tapia, R.A., Thompson, J.R. (1978). Nonparametric Probability Density Estimation. Johns Hopkins Press, Baltimore, Maryland.

44. Van Ryzin, J. (1969). On strong consistency of density estimates. Ann.Math.Statist. (42), 1870-86.

45. Walter, G., Blum, J. (1979). Probability estimation using delta sequences. Ann.Statist. (7), 328-340.

46. Wegman, E.J. (1972). Nonparametric probability density estimation I : a summary of available methods. Technometrics (14), 533-545.

47. Wegman, E.J. (1972). Nonparametric probability density estimation II : a comparison of density estimation methods. J.Statist.Comp.Simul. (1), 225-245.

48. Wertz, W. (1978). Statistical Density Estimation : A Survey. Vandenhoeck & Ruprecht, Göttingen.

49. Wertz, W. (1978). Optimal estimation of functions of probability densities. Studia Sci.Math.Hungar. (13), 377-383.

50. Wertz, W., Schneider, B. (1979). Statistical density estimation : a bibliography. Int.Stat.Rev. (47), 155-175.

51. Whittaker, E.F, Watson, G.N. (1952). Modern Analysis. 4 th. ed. Univ. Press, Cambridge.

52. Zacks, S. (1971). The Theory of Statistical Inference. Wiley, New-York.

TABLEAUX

LISTE DES TABLEAUX

Tableaux

TAILLE	EQMI(G1) [Th.2.1.1]	EQMI(G12) [fct.Wertz]	EQMI(G1) [fct(1/n)]	QUOTIENTS	
N	(1)	(2)	(3)	(1)/(2)	(1)/(3)
20	0.00692	0.00692	0.00680	1.01829	1.00039
30	0.00464	0.00464	0.00459	1.01229	1.00017
50	0.00280	0.00280	0.00270	1.00742	1.00006
75	0.00187	0.00187	0.00186	1.00497	1.00003
100	0.00141	0.00141	0.00140	1.00375	1.00003
200	0.00070	0.00070	0.00070	1.00184	0.99999
500	0.00028	0.00028	0.00028	1.00072	0.99996
1000	0.00014	0.00014	0.00014	1.00048	1.00006
5000	0.00003	0.00003	0.00003	1.00000	1.00024
10000	0.00001	0.00001	0.00001	1.00119	1.00140

Tableau 2.1 : EQMI de l'estimateur paramétrique G1 (mu inconnu).

TAILLE	EQMI(G2) [Th.2.1.2]	EQMI(G2) [fct(1/n)]	QUOT.	EQMI(G3) [Th.2.1.3]	EQMI(G3) [fct(1/n)]	QUOT.
N	(1)	(2)	(1)/(2)	(3)	(4)	(3)/(4)
20	0.00602	0.00599	1.00507	0.01363	0.01351	1.00919
30	0.00385	0.00384	1.00231	0.00878	0.00875	1.00404
50	0.00223	0.00223	1.00091	0.00513	0.00512	1.00144
75	0.00146	0.00146	1.00037	0.00338	0.00337	1.00066
100	0.00109	0.00109	1.00031	0.00252	0.00251	1.00028
200	0.00054	0.00054	1.00038	0.00125	0.00125	1.00003
500	0.00021	0.00021	1.00126	0.00050	0.00050	0.99952
1000	0.00011	0.00011	1.00317	0.00025	0.00025	0.99891
5000	0.00002	0.00002	1.04140	0.00005	0.00005	0.98163
10000	0.00001	0.00001	1.22152	0.00002	0.00002	0.90476

Tableau 2.2 : EQMI des estimateurs paramétriques G2 et G3 .
G2 : sigma inconnu ; G3 : mu et sigma inconnus .

TAILLE	EQMI(F1) [Th.2.2.1]	EQMI(F2) [Th.2.2.1]	EQMI(F2) [Th.2.2.2]	QUOTIENTS	
N	(1)	(2)	(3)	(3)/(1)	(3)/(2)
20	0.02926	0.03030	0.01625	0.5554	0.5362
30	0.02115	0.02191	0.01259	0.5953	0.5746
50	0.01406	0.01456	0.00904	0.6430	0.6208
75	0.01016	0.01053	0.00690	0.6791	0.6555
100	0.00807	0.00836	0.00568	0.7038	0.6789
200	0.00464	0.00480	0.00351	0.7565	0.7309
500	0.00223	0.00231	0.00182	0.8161	0.7900
1000	0.00128	0.00133	0.00110	0.8594	0.8277
5000	0.00035	0.00037	0.00033	0.9429	0.8947
10000	0.00020	0.00021	0.00019	0.9501	0.9167

Tableau 2.3 : EQMI des estimateurs non-paramétriques, F1 et F2.
F1 : noyau biquadratique ; F2 : noyau normal.

TAILLE	EFF(G1,F2)	EFF(G2,F2)	EFF(G3,F2)
20	0.426	0.371	0.839
30	0.369	0.306	0.698
50	0.310	0.247	0.568
75	0.271	0.212	0.489
100	0.248	0.191	0.443
200	0.201	0.153	0.355
500	0.155	0.117	0.272
1000	0.129	0.097	0.225
5000	0.086	0.067	0.148
10000	0.073	0.067	0.116

Tableau 2.4 : Efficacité de l'estimateur non-paramétrique F2 .

TAILLE	EQMI(G1)	EQDM(G1) EQDS(G1)	EQMI(G2)	EQDM(G2) EQDS(G2)	EQMI(G3)	EQDM(G3) EQDS(G3)
20	0.00692	0.00595 0.00677	0.00602	0.00576 0.01301	0.01363	0.01474 0.02118
50	0.00280	0.00240 0.00341	0.00223	0.00208 0.00253	0.00513	0.00469 0.00517
100	0.00141	0.00124 0.00223	0.00109	0.00109 0.00134	0.00252	0.00203 0.00201

Tableau 2.5 : EQMI et EQDM des estimateurs paramétriques G1, G2 et G3.

TAILLE	EQMI(F1) ASYMPT.	EQDM(F1) EQDS(F1)			EQMI(F2) EXACTE	EQDM(F2) EQDS(F2)		
20	0.02926	0.01625 0.01121	0.01387 0.01141	0.01461 0.01521	0.01626	0.01744 0.01078	0.01468 0.01067	0.01527 0.01417
50	0.01406	0.00707 0.00537	0.00677 0.00452	0.00740 0.00558	0.00906	0.00764 0.00538	0.00774 0.00461	0.00806 0.00592
100	0.00807	0.00508 0.00385	0.00471 0.00335	0.00572 0.00497	0.00568	0.00561 0.00376	0.00518 0.00356	0.00627 0.00493

Tableau 2.6 : EQMI et EQDM des estimateurs non-paramétriques, F1 et F2, pour la fenêtre optimale.

TAILLE	Fenêtre EQMI(F1) EQDM(F1) EQDS(F1)				Fenêtre EQMI(F2) EQDM(F2) EQDS(F2)			
20	1.24	1.53	1.87	2.29	0.47	0.58	0.71	0.87
	0.03135	0.02926	0.03230	0.04528	0.01925	0.01624	0.01626	0.01995
	0.01508	0.01267	0.01354	0.01863	0.01612	0.01329	0.01350	0.01747
	0.01281	0.00978	0.00847	0.00804	0.01335	0.01022	0.00875	0.00839
50	1.03	1.27	1.56	1.96	0.39	0.48	0.59	0.73
	0.01500	0.01406	0.01555	0.02322	0.01023	0.00882	0.00906	0.01177
	0.01106	0.00952	0.00971	0.01303	0.01163	0.00992	0.00983	0.01219
	0.01105	0.01019	0.00957	0.00880	0.01124	0.01033	0.00972	0.00914
100	0.90	1.11	1.35	1.64	0.34	0.42	0.51	0.62
	0.00865	0.00807	0.00888	0.01216	0.00624	0.00545	0.00568	0.00729
	0.00467	0.00396	0.00427	0.00610	0.00497	0.00417	0.00434	0.00588
	0.00341	0.00301	0.00305	0.00342	0.00347	0.00306	0.00307	0.00342

Tableau 2.7 : EQMI et EQDM des estimateurs non-paramétriques, F1 et F2, pour différentes fenêtres.

TAILLE	EFFICACITE BIQUADRATIQUE (G1/F1) (G2/F1) (G3/F1)			EFFICACITE NORMALE (G1/F2) (G2/F2) (G3/F2)		
N	Efficacité théorique Efficacité simulée			Efficacité théorique Efficacité simulée		
20	0.237 0.366	0.206 0.415	0.466 1.009	0.426 0.341	0.371 0.392	0.839 0.965
50	0.199 0.339	0.159 0.307	0.365 0.634	0.310 0.314	0.247 0.269	0.560 0.582
100	0.174 0.244	0.135 0.231	0.312 0.355	0.248 0.221	0.191 0.210	0.443 0.324

Tableau 2.0 : Efficacités théoriques et simulées.

TAILLE	EQMI(G)	EQMI(eps=0.5) EQMI(eps=0.5)/EQMI(eps=0.0)			
	Eps=0.0	D = 1/2	D = 3.0	R2= 1/4	R2= 4.0
20	0.00692	0.00721 1.04	0.11550 16.69	0.03971 5.74	0.02718 3.93
50	0.00280	0.00304 1.09	0.11599 41.43	0.03581 12.79	0.02093 7.48
100	0.00141	0.00163 1.16	0.11624 02.44	0.03449 24.46	0.01877 13.31

Tableau 3.1 : EQMI de l'estimateur paramétrique G pour epsilon = 0.0 et pour epsilon = 0.5 et rapport de ces EQMI lorsque n varie.

TAILLE	Eps	EQMI ASYMPTOTIQUES					
		BIQUAD	NORMAL	BIQ/NOR	BIQUAD	NORMAL	BIQ/NOR
		D = 1/2			D = 3.0		
20	0.05	0.02909	0.03014	0.96535	0.02862	0.02966	0.96498
	0.10	0.02895	0.02999	0.96524	0.02805	0.02908	0.96452
	0.20	0.02871	0.02975	0.96505	0.02712	0.02814	0.96372
	0.30	0.02854	0.02957	0.96491	0.02645	0.02746	0.96312
	0.50	0.02040	0.02943	0.96480	0.02591	0.02692	0.96261
50	0.05	0.01398	0.01448	0.96513	0.01375	0.01426	0.96458
	0.10	0.01391	0.01441	0.96496	0.01348	0.01399	0.96390
	0.20	0.01380	0.01430	0.96460	0.01304	0.01354	0.96271
	0.30	0.01371	0.01422	0.96448	0.01272	0.01323	0.96181
	0.50	0.01365	0.01415	0.96432	0.01247	0.01297	0.96106
100	0.05	0.00803	0.00832	0.96531	0.00790	0.00819	0.96482
	0.10	0.00799	0.00828	0.96516	0.00774	0.00803	0.96421
	0.20	0.00792	0.00821	0.96491	0.00748	0.00777	0.96316
	0.30	0.00787	0.00816	0.96473	0.00729	0.00758	0.96236
	0.50	0.00784	0.00812	0.96459	0.00715	0.00743	0.96169

Tableau 3.2 : EQMI asymptotiques des estimateurs non-paramétriques F1 et F2 lorsque n varie et R2 = 1.0.

TAILLE	Eps	EQMI ASYMPTOTIQUES					
		BIQUAD	NORMAL	BIQ/NOR	BIQUAD	NORMAL	BIQ/NOR
		R2 = 1/4			R2 = 4.0		
20	0.05	0.03097	0.03204	0.96671	0.02874	0.02978	0.96507
	0.10	0.03347	0.03457	0.96828	0.02824	0.02928	0.96468
	0.20	0.04083	0.04201	0.97182	0.02733	0.02835	0.96391
	0.30	0.05132	0.05263	0.97513	0.02651	0.02752	0.96317
	0.50	0.00172	0.00339	0.97996	0.02517	0.02616	0.96187
50	0.05	0.01487	0.01538	0.96716	0.01381	0.01432	0.96472
	0.10	0.01606	0.01656	0.96953	0.01358	0.01408	0.96413
	0.20	0.01955	0.02005	0.97485	0.01314	0.01365	0.96298
	0.30	0.02453	0.02503	0.97988	0.01275	0.01326	0.96189
	0.50	0.03896	0.03946	0.98728	0.01211	0.01262	0.95996
100	0.05	0.00055	0.00804	0.96710	0.00793	0.00822	0.96494
	0.10	0.00924	0.00954	0.96918	0.00779	0.00808	0.96442
	0.20	0.01120	0.01150	0.97386	0.00754	0.00783	0.96340
	0.30	0.01419	0.01450	0.97824	0.00731	0.00760	0.96244
	0.50	0.02261	0.02296	0.98465	0.00694	0.00722	0.96072

Tableau 3.3 : EQMI asymptotiques des estimateurs non-paramétriques F1 et F2 lorsque n varie et D = 0.0.

TAILLE	EQMI(F2)	EQMI(eps=0.5) EQMI(eps=0.5)/EQMI(eps=0.0)			
	Eps=0.0	D = 1/2	D = 3.0	R2= 1/4	R2= 4.0
20	0.01626	0.01555 0.96	0.01561 0.96	0.04140 2.55	0.01386 0.85
50	0.00906	0.00862 0.95	0.00831 0.92	0.02621 2.89	0.00741 0.82
100	0.00568	0.00539 0.95	0.00510 0.90	0.01777 3.13	0.00456 0.80

Tableau 3.4 : EQMI de l'estimateur non-paramétrique F2 pour epsilon = 0.0 et pour epsilon = 0.5 et rapport de ces EQMI lorsque n varie.

Eps = 0.5		EQMI(G) EQMI(F2) EQMI(eps=0.5)/EQMI(eps=0.0)	
Eps = 0.0		0.00280	0.00906
D	1/3	0.00286 1.02	0.00886 0.98
	1/2	0.00304 1.09	0.00862 0.95
	1.0	0.00586 2.09	0.00765 0.84
	3.0	0.11599 41.43	0.00831 0.92
	5.0	0.29650 105.90	0.00947 1.05
R2	1/9	0.09552 34.11	0.06325 6.98
	1/4	0.03581 12.79	0.02621 2.89
	4.0	0.02093 7.48	0.00741 0.82
	9.0	0.03837 13.70	0.00782 0.86

Tableau 3.5 : EQMI des estimateurs G et F2 pour epsilon = 0.0 et pour epsilon = 0.5 et rapport de ces EQMI pour n = 50.

TAILLE	EQMI(F2) Fenêtre	EQMI(eps=0.5) EQMI(eps=0.5)/EQMI(eps=0.0)				
	Eps=0.0	D = 1/3	D = 1.0	D = 5.0	R2= 1/9	R2= 9.0
20	0.01925 0.47	0.01927 1.00	0.01956 1.02	0.02466 1.20	0.05115 2.66	0.02310 1.20
	0.01624 0.58	0.01612 0.99	0.01555 0.96	0.02039 1.26	0.06666 4.10	0.01815 1.12
	0.01626 0.71	0.01593 0.98	0.01396 0.86	0.01834 1.13	0.08740 5.38	0.01493 0.92
	0.01995 0.87	0.01929 0.97	0.01520 0.76	0.01866 0.94	0.11335 5.68	0.01333 0.67
50	0.01023 0.39	0.01023 1.00	0.01029 1.00	0.01237 1.21	0.03226 3.15	0.01171 1.14
	0.00882 0.48	0.00875 0.99	0.00836 0.95	0.01035 1.17	0.04508 5.11	0.00935 1.06
	0.00906 0.59	0.00886 0.98	0.00765 0.84	0.00947 1.05	0.06325 6.98	0.00782 0.86
	0.01177 0.73	0.01134 0.96	0.00868 0.74	0.01011 0.86	0.08747 7.43	0.00722 0.61
100	0.00624 0.34	0.00624 1.00	0.00623 1.00	0.00728 1.17	0.02228 3.57	0.00693 1.11
	0.00545 0.42	0.00540 0.99	0.00511 0.94	0.00613 1.12	0.03311 6.08	0.00556 1.02
	0.00568 0.51	0.00555 0.98	0.00476 0.84	0.00570 1.00	0.04767 8.39	0.00476 0.84
	0.00729 0.62	0.00701 0.96	0.00535 0.73	0.00612 0.84	0.06692 9.18	0.00447 0.61

Tableau 3.6 : EQMI de l'estimateur non-paramétrique F2 pour epsilon = 0.0 et epsilon = 0.5 et rapport de ces EQMI si la fenêtre et n varient.

Eps	EQMI(G) EQDM(G) EQDS(G)	EQMI(F1) EQDM(F1) EQDS(F1)	EQMI(F2) EQDM(F2) EQDS(F2)	EFF(G,F1) EFFSIM	EFF(G,F2) EFFSIM
0.05	0.00695 0.00541 0.00856	0.02909 0.01531 0.01319	0.01613 0.01555 0.01231	0.24 0.35	0.43 0.35
0.10	0.00699 0.00541 0.00782	0.02895 0.01335 0.01130	0.01600 0.01356 0.01008	0.24 0.41	0.44 0.40
0.20	0.00707 0.00875 0.01283	0.02871 0.01813 0.01529	0.01580 0.01798 0.01404	0.25 0.48	0.45 0.49
0.30	0.00714 0.00776 0.00926	0.02854 0.01578 0.01327	0.01566 0.01632 0.01287	0.25 0.49	0.46 0.48
0.50	0.00721 0.00703 0.00743	0.02840 0.01568 0.01651	0.01555 0.01601 0.01467	0.25 0.45	0.46 0.44

Tableau 3.7 : Valeurs théoriques et simulées des EQMI et des efficacités pour D = 1/2 , R2 = 1.0 et N = 20.

Eps	EQMI(G) EQDM(G) EQDS(G)	EQMI(F1) EQDM(F1) EQDS(F1)	EQMI(F2) EQDM(F2) EQDS(F2)	EFF(G,F1) EFFSIM	EFF(G,F2) EFFSIM
0.05	0.00282 0.00259 0.00291	0.01390 0.00922 0.00752	0.00898 0.00973 0.00748	0.20 0.28	0.31 0.27
0.10	0.00284 0.00223 0.00271	0.01391 0.00681 0.00451	0.00891 0.00749 0.00464	0.20 0.33	0.32 0.30
0.20	0.00291 0.00221 0.00324	0.01380 0.00722 0.00580	0.00878 0.00789 0.00602	0.21 0.31	0.33 0.28
0.30	0.00297 0.00278 0.00347	0.01371 0.00852 0.00548	0.00869 0.00935 0.00560	0.22 0.33	0.34 0.30
0.50	0.00304 0.00238 0.00276	0.01365 0.00764 0.00598	0.00862 0.00818 0.00617	0.22 0.31	0.35 0.29

Tableau 3.8 : Valeurs théoriques et simulées des EQMI et des efficacités pour D = 1/2 , R2 = 1.0 et N = 50.

Eps	EQMI(G) EQDM(G) EQDS(G)	EQMI(F1) EQDM(F1) EQDS(F1)	EQMI(F2) EQDM(F2) EQDS(F2)	EFF(G,F1) EFFSIM	EFF(G,F2) EFFSIM
0.05	0.00142 0.00130 0.00143	0.00803 0.00608 0.00365	0.00563 0.00685 0.00394	0.18 0.21	0.25 0.19
0.10	0.00144 0.00104 0.00147	0.00799 0.00448 0.00284	0.00558 0.00489 0.00298	0.18 0.23	0.26 0.10
0.20	0.00150 0.00157 0.00218	0.00792 0.00558 0.00370	0.00550 0.00584 0.00382	0.19 0.28	0.27 0.27
0.30	0.00156 0.00162 0.00302	0.00787 0.00444 0.00332	0.00544 0.00469 0.00359	0.20 0.37	0.29 0.35
0.50	0.00163 0.00168 0.00195	0.00784 0.00561 0.00410	0.00539 0.00573 0.00390	0.21 0.30	0.30 0.29

Tableau 3.9 : Valeurs théoriques et simulées des EQMI et des efficacités pour D = 1/2 , R2 = 1.0 et N = 100.

Eps	EQMI(G) EQDM(G) EQDS(G)	EQMI(F1) EQDM(F1) EQDS(F1)	EQMI(F2) EQDM(F2) EQDS(F2)	EFF(G,F1) EFFSIM	EFF(G,F2) EFFSIM
0.05	0.01247 0.01433 0.01539	0.02862 0.01742 0.01173	0.01614 0.01796 0.01068	0.44 0.02	0.77 0.80
0.10	0.02294 0.02501 0.02519	0.02805 0.01680 0.01263	0.01603 0.01606 0.01119	0.02 1.49	1.43 1.56
0.20	0.05253 0.05187 0.03377	0.02712 0.01778 0.01187	0.01585 0.01632 0.01092	1.94 2.92	3.31 3.18
0.30	0.08385 0.07376 0.02429	0.02645 0.01512 0.00914	0.01572 0.01333 0.00829	3.17 4.88	5.33 5.53
0.50	0.11550 0.11552 0.00100	0.02591 0.01499 0.01305	0.01561 0.01323 0.01149	4.46 7.70	7.40 8.73

Tableau 3.10 : Valeurs théoriques et simulées des EQMI et des efficacités pour D = 3.0 , R2 = 1.0 et N = 20.

Eps	EQMI(G) EQDM(G) EQDS(G)	EQMI(F1) EQDM(F1) EQDS(F1)	EQMI(F2) EQDM(F2) EQDS(F2)	EFF(G,F1) EFFSIM	EFF(G,F2) EFFSIM
0.05	0.00718 0.00450 0.00496	0.01375 0.00769 0.00680	0.00892 0.00779 0.00688	0.52 0.58	0.80 0.58
0.10	0.01718 0.01650 0.01271	0.01348 0.00773 0.00565	0.00879 0.00758 0.00519	1.27 2.13	1.95 2.18
0.20	0.04763 0.04767 0.01870	0.01304 0.00952 0.00613	0.00858 0.00906 0.00602	3.65 5.01	5.55 5.26
0.30	0.08124 0.08158 0.01798	0.01272 0.00914 0.00792	0.00843 0.00845 0.00712	6.39 8.93	9.64 9.66
0.50	0.11599 0.11611 0.00048	0.01247 0.00866 0.00466	0.00831 0.00778 0.00419	9.30 13.40	13.96 14.92

Tableau 3.11 : Valeurs théoriques et simulées des EQMI et des efficacités pour D = 3.0 , R2 = 1.0 et N = 50.

Eps	EQMI(G) EQDM(G) EQDS(G)	EQMI(F1) EQDM(F1) EQDS(F1)	EQMI(F2) EQDM(F2) EQDS(F2)	EFF(G,F1) EFFSIM	EFF(G,F2) EFFSIM
0.05	0.00537 0.00625 0.00535	0.00790 0.00476 0.00405	0.00557 0.00505 0.00386	0.68 1.31	0.96 1.24
0.10	0.01519 0.01343 0.00694	0.00774 0.00522 0.00336	0.00547 0.00538 0.00357	1.96 2.57	2.78 2.50
0.20	0.04593 0.04714 0.01622	0.00748 0.00535 0.00357	0.00531 0.00512 0.00335	6.14 8.80	8.65 9.21
0.30	0.08035 0.08135 0.01684	0.00729 0.00532 0.00270	0.00519 0.00505 0.00275	11.02 15.30	15.48 16.10
0.50	0.11624 0.11625 0.00036	0.00715 0.00563 0.00297	0.00510 0.00521 0.00270	16.26 20.65	22.79 22.31

Tableau 3.12 : Valeurs théoriques et simulées des EQMI et des efficacités pour D = 3.0 , R2 = 1.0 et N = 100.

Eps	EQMI(G) EQDM(G) EQDS(G)	EQMI(F1) EQDM(F1) EQDS(F1)	EQMI(F2) EQDM(F2) EQDS(F2)	EFF(G,F1) EFFSIM	EFF(G,F2) EFFSIM
0.05	0.00733 0.00708 0.00739	0.03097 0.01523 0.01370	0.01743 0.01652 0.01363	0.24 0.46	0.42 0.43
0.10	0.00838 0.00688 0.00678	0.03347 0.01612 0.01115	0.01890 0.01810 0.01068	0.25 0.43	0.44 0.38
0.20	0.01239 0.01221 0.00816	0.04083 0.02090 0.01617	0.02273 0.02392 0.01465	0.30 0.58	0.55 0.51
0.30	0.01895 0.01947 0.00892	0.05132 0.02237 0.01356	0.02776 0.02758 0.01340	0.37 0.87	0.68 0.71
0.50	0.03971 0.03955 0.00739	0.08171 0.03280 0.01551	0.04140 0.04154 0.01555	0.49 1.21	0.96 0.95

Tableau 3.13 : Valeurs théoriques et simulées des EQMI et des efficacités pour D = 0.0 , R2 = 1/4 et N = 20.

Eps	EQMI(G) EQDM(G) EQDS(G)	EQMI(F1) EQDM(F1) EQDS(F1)	EQMI(F2) EQDM(F2) EQDS(F2)	EFF(G,F1) EFFSIM	EFF(G,F2) EFFSIM
0.05	0.00316 0.00365 0.00513	0.01487 0.00956 0.00682	0.00981 0.01090 0.00673	0.21 0.38	0.32 0.33
0.10	0.00418 0.00455 0.00401	0.01606 0.01037 0.00769	0.01078 0.01173 0.00768	0.26 0.44	0.39 0.39
0.20	0.00817 0.00790 0.00319	0.01955 0.01123 0.00801	0.01335 0.01362 0.00832	0.42 0.70	0.61 0.58
0.30	0.01477 0.01504 0.00572	0.02453 0.01323 0.00818	0.01678 0.01693 0.00866	0.60 1.14	0.88 0.89
0.50	0.03581 0.03592 0.00331	0.03896 0.01922 0.00951	0.02621 0.02622 0.00938	0.92 1.87	1.37 1.37

Tableau 3.14 : Valeurs théoriques et simulées des EQMI et des efficacités pour D = 0.0 , R2 = 1/4 et N = 50.

Eps	EQMI(G) EQDM(G) EQDS(G)	EQMI(F1) EQDM(F1) EQDS(F1)	EQMI(F2) EQDM(F2) EQDS(F2)	EFF(G,F1) EFFSIM	EFF(G,F2) EFFSIM
0.05	0.00175 0.00141 0.00152	0.00855 0.00539 0.00401	0.00618 0.00617 0.00415	0.20 0.26	0.28 0.23
0.10	0.00276 0.00259 0.00165	0.00924 0.00460 0.00318	0.00604 0.00547 0.00343	0.30 0.56	0.40 0.47
0.20	0.00674 0.00675 0.00282	0.01128 0.00672 0.00404	0.00663 0.00854 0.00425	0.60 1.00	0.78 0.79
0.30	0.01336 0.01320 0.00121	0.01419 0.00795 0.00471	0.01105 0.01052 0.00514	0.94 1.66	1.21 1.25
0.50	0.03449 0.03424 0.00119	0.02261 0.01342 0.00660	0.01777 0.01846 0.00700	1.53 2.55	1.94 1.86

Tableau 3.15 : Valeurs théoriques et simulées des EQMI et des efficacités pour D = 0.0 , R2 = 1/4 et N = 100.

Eps	EQMI(G) EQDM(G) EQDS(G)	EQMI(F1) EQDM(F1) EQDS(F1)	EQMI(F2) EQDM(F2) EQDS(F2)	EFF(G,F1) EFFSIM	EFF(G,F2) EFFSIM
0.05	0.00781 0.00521 0.00686	0.02074 0.01114 0.00975	0.01593 0.01141 0.01004	0.27 0.47	0.49 0.46
0.10	0.00894 0.00900 0.01273	0.02024 0.01579 0.01200	0.01561 0.01616 0.01137	0.32 0.57	0.57 0.56
0.20	0.01197 0.01095 0.01071	0.02733 0.01705 0.01368	0.01505 0.01601 0.01319	0.44 0.64	0.80 0.68
0.30	0.01602 0.01896 0.01943	0.02651 0.01673 0.01284	0.01457 0.01589 0.01171	0.60 1.13	1.10 1.19
0.50	0.02718 0.02694 0.01714	0.02517 0.01351 0.00855	0.01386 0.01185 0.00817	1.08 1.99	1.96 2.27

Tableau 3.16 : Valeurs théoriques et simulées des EQMI et des efficacités pour D = 0.0 , R2 = 4.0 et N = 20.

Eps	EQMI(G) EQDM(G) EQDS(G)	EQMI(F1) EQDM(F1) EQDS(F1)	EQMI(F2) EQDM(F2) EQDS(F2)	EFF(G,F1) EFFSIM	EFF(G,F2) EFFSIM
0.05	0.00326 0.00307 0.00370	0.01381 0.00022 0.00616	0.00804 0.00089 0.00611	0.24 0.37	0.37 0.34
0.10	0.00402 0.00383 0.00486	0.01350 0.00900 0.00601	0.00863 0.00929 0.00581	0.30 0.43	0.47 0.41
0.20	0.00645 0.00620 0.00495	0.01314 0.00794 0.00550	0.00825 0.00795 0.00539	0.49 0.79	0.78 0.79
0.30	0.01007 0.00926 0.00437	0.01275 0.00842 0.00604	0.00792 0.00832 0.00575	0.79 1.10	1.27 1.11
0.50	0.02093 0.02085 0.00534	0.01211 0.00797 0.00636	0.00741 0.00699 0.00607	1.73 2.62	2.82 2.98

Tableau 3.17 : Valeurs théoriques et simulées des EQMI et des efficacités pour D = 0.0 , R2 = 4.0 et N = 50.

Eps	EQMI(G) EQDM(G) EQDS(G)	EQMI(F1) EQDM(F1) EQDS(F1)	EQMI(F2) EQDM(F2) EQDS(F2)	EFF(G,F1) EFFSIM	EFF(G,F2) EFFSIM
0.05	0.00172 0.00159 0.00266	0.00793 0.00445 0.00304	0.00553 0.00480 0.00325	0.22 0.36	0.31 0.33
0.10	0.00235 0.00269 0.00280	0.00779 0.00465 0.00385	0.00539 0.00460 0.00300	0.30 0.58	0.44 0.58
0.20	0.00456 0.00509 0.00313	0.00754 0.00564 0.00383	0.00513 0.00554 0.00383	0.60 0.90	0.89 0.92
0.30	0.00803 0.00798 0.00330	0.00731 0.00531 0.00500	0.00491 0.00520 0.00492	1.10 1.50	1.64 1.51
0.50	0.01877 0.01813 0.00270	0.00694 0.00489 0.00257	0.00456 0.00446 0.00252	2.70 3.71	4.12 4.06

Tableau 3.18 : Valeurs théoriques et simulées des EQMI et des efficacités pour D = 0.0 , R2 = 4.0 et N = 100.

Fenêtre	EQDM(F1) EQDS(F1)				
Eps	0.05	0.10	0.20	0.30	0.50
1.03	0.01096 0.00906	0.00863 0.00438	0.00964 0.00626	0.00935 0.00612	0.00883 0.00602
1.27	0.00959 0.00825	0.00722 0.00403	0.00822 0.00544	0.00813 0.00594	0.00731 0.00504
1.56	0.00992 0.00783	0.00771 0.00422	0.00872 0.00558	0.00865 0.00607	0.00768 0.00485
1.96	0.01331 0.00750	0.01138 0.00452	0.01242 0.00603	0.01220 0.00618	0.01119 0.00505

Tableau 3.19 : Valeurs des EQDM de l'estimateur non-paramétrique F1 pour différentes fenêtres si D = 1/3 , R2 = 1.0 et N = 50.

Fenêtre	EQDM(F1) EQDS(F1)				
Eps	0.05	0.10	0.20	0.30	0.50
1.03	0.01110 0.00827	0.00906 0.00684	0.00917 0.00684	0.00871 0.00515	0.00961 0.00620
1.27	0.00936 0.00764	0.00716 0.00603	0.00731 0.00619	0.00720 0.00403	0.00749 0.00514
1.56	0.00940 0.00742	0.00702 0.00578	0.00701 0.00602	0.00711 0.00342	0.00672 0.00481
1.96	0.01251 0.00737	0.00982 0.00581	0.00926 0.00606	0.00919 0.00349	0.00812 0.00495

Tableau 3.20 : Valeurs des EQDM de l'estimateur non-paramétrique F1 pour différentes fenêtres si D = 1.0 , R2 = 1.0 et N = 50.

Fenêtre	EQDM(F1) EQDS(F1)				
Eps	0.05	0.10	0.20	0.30	0.50
1.03	0.00796 0.00563	0.01000 0.00685	0.01116 0.00742	0.01202 0.00754	0.01105 0.00558
1.27	0.00674 0.00503	0.00883 0.00620	0.00948 0.00696	0.00999 0.00666	0.00927 0.00531
1.56	0.00714 0.00489	0.00918 0.00596	0.00915 0.00667	0.00927 0.00595	0.00873 0.00515
1.96	0.01042 0.00502	0.01213 0.00588	0.01107 0.00639	0.01062 0.00533	0.01004 0.00499

Tableau 3.21 : Valeurs des EQDM de l'estimateur non-paramétrique F1 pour différentes fenêtres si D = 5.0 , R2 = 1.0 et N = 50.

Fenêtre	EQDM(F1) EQDS(F1)				
Eps	0.05	0.10	0.20	0.30	0.50
0.98	0.01135 0.00774	0.01264 0.00981	0.01147 0.00577	0.02209 0.01236	0.03428 0.01256
1.27	0.01048 0.00722	0.01221 0.00912	0.01428 0.00595	0.02735 0.01146	0.04988 0.01187
1.48	0.01167 0.00709	0.01410 0.00864	0.02006 0.00614	0.03622 0.01039	0.07111 0.01083
1.68	0.01644 0.00704	0.02010 0.00813	0.03099 0.00616	0.05104 0.00899	0.10084 0.00908

Tableau 3.22 : Valeurs des EQDM de l'estimateur non-paramétrique F1 pour différentes fenêtres si D = 0.0 , R2 = 1/9 et N = 50.

Fenêtre	EQDM(F1) EQDS(F1)				
Eps	0.05	0.10	0.20	0.30	0.50
1.03	0.01047 0.00755	0.01132 0.00680	0.01156 0.00610	0.00908 0.00628	0.01034 0.00799
1.27	0.00915 0.00701	0.00997 0.00643	0.00965 0.00536	0.00727 0.00542	0.00829 0.00698
1.56	0.00949 0.00712	0.01009 0.00648	0.00912 0.00499	0.00653 0.00484	0.00703 0.00612
1.96	0.01268 0.00741	0.01271 0.00664	0.01081 0.00491	0.00739 0.00449	0.00671 0.00547

Tableau 3.23 : Valeurs des EQDM de l'estimateur non-paramétrique F1 pour différentes fenêtres si D = 0.0 , R2 = 9.0 et N = 50.

Fenêtre	EQMI(F2) EQDM(F2) EQDS(F2)				
Eps	0.05	0.10	0.20	0.30	0.50
0.39	0.01023 0.01150 0.00921	0.01023 0.00925 0.00455	0.01023 0.01033 0.00654	0.01023 0.00993 0.00622	0.01023 0.00947 0.00626
0.48	0.00881 0.00997 0.00840	0.00880 0.00766 0.00412	0.00877 0.00060 0.00559	0.00876 0.00853 0.00602	0.00875 0.00777 0.00522
0.59	0.00903 0.01001 0.00795	0.00899 0.00782 0.00427	0.00893 0.00005 0.00561	0.00889 0.00875 0.00614	0.00886 0.00700 0.00494
0.73	0.01169 0.01246 0.00769	0.01162 0.01048 0.00461	0.01150 0.01154 0.00607	0.01141 0.01136 0.00632	0.01134 0.01032 0.00511

Tableau 3.24 : Valeurs théoriques et simulées des EQMI de l'estimateur non-paramétrique F2 pour différentes fenêtres si D = 1/3 , R2 = 1.0 et N = 50.

Fenêtre	EQMI(F2) EQDM(F2) EQDS(F2)				
Eps	0.05	0.10	0.20	0.30	0.50
0.39	0.01024 0.01175 0.00834	0.01025 0.00970 0.00700	0.01027 0.00981 0.00699	0.01028 0.00919 0.00529	0.01029 0.01025 0.00644
0.48	0.00873 0.00985 0.00774	0.00865 0.00765 0.00617	0.00852 0.00778 0.00630	0.00843 0.00764 0.00420	0.00836 0.00799 0.00532
0.59	0.00880 0.00960 0.00754	0.00855 0.00721 0.00588	0.00816 0.00722 0.00608	0.00787 0.00729 0.00355	0.00765 0.00702 0.00489
0.73	0.01118 0.01179 0.00755	0.01066 0.00912 0.00594	0.00979 0.00873 0.00614	0.00917 0.00874 0.00352	0.00868 0.00785 0.00499

Tableau 3.25 : Valeurs théoriques et simulées des EQMI de l'estimateur non-paramétrique F2 pour différentes fenêtres si D = 1.0 , R2 = 1.0 et N = 50.

Fenêtre	EQMI(F2) EQDM(F2) EQDS(F2)				
Eps	0.05	0.10	0.20	0.30	0.50
0.39	0.01064 0.00845 0.00580	0.01100 0.01053 0.00710	0.01160 0.01171 0.00749	0.01203 0.01262 0.00765	0.01237 0.01164 0.00561
0.48	0.00911 0.00708 0.00517	0.00937 0.00919 0.00638	0.00980 0.00990 0.00702	0.01011 0.01045 0.00670	0.01035 0.00971 0.00534
0.59	0.00914 0.00720 0.00499	0.00921 0.00928 0.00608	0.00932 0.00934 0.00674	0.00941 0.00950 0.00607	0.00947 0.00896 0.00518
0.73	0.01146 0.00958 0.00513	0.01118 0.01143 0.00604	0.01071 0.01061 0.00656	0.01038 0.01025 0.00551	0.01011 0.00974 0.00507

Tableau 3.26 : Valeurs théoriques et simulées des EQMI de l'estimateur non-paramétrique F2 pour différentes fenêtres si D = 5.0 , R2 = 1.0 et N = 50.

Fenêtre	EQMI(F2) EQDM(F2) EQDS(F2)				
Eps	0.05	0.10	0.20	0.30	0.50
0.39	0.01050 0.01193 0.00787	0.01134 0.01314 0.01000	0.01410 0.01135 0.00604	0.01850 0.02144 0.01208	0.03226 0.03135 0.01318
0.48	0.00950 0.01083 0.00736	0.01097 0.01237 0.00939	0.01567 0.01347 0.00624	0.02292 0.02553 0.01229	0.04508 0.04414 0.01268
0.59	0.01044 0.01163 0.00729	0.01272 0.01375 0.00897	0.01996 0.01030 0.00651	0.03080 0.03299 0.01152	0.06325 0.06230 0.01189
0.73	0.01407 0.01520 0.00737	0.01754 0.01024 0.00861	0.02799 0.02688 0.00662	0.04313 0.04482 0.01046	0.08747 0.08654 0.01064

Tableau 3.27 : Valeurs théoriques et simulées des EQMI de l'estimateur non-paramétrique F2 pour différentes fenêtres si D = 0.0 , R2 = 1/9 et N = 50.

Fenêtre	EQMI(F2) EQDM(F2) EQDS(F2)				
Eps	0.05	0.10	0.20	0.30	0.50
0.39	0.01041 0.01107 0.00774	0.01057 0.01188 0.00694	0.01069 0.01217 0.00621	0.01119 0.00961 0.00642	0.01171 0.01088 0.00806
0.48	0.00887 0.00956 0.00713	0.00893 0.01035 0.00649	0.00903 0.01013 0.00546	0.00914 0.00767 0.00554	0.00935 0.00871 0.00709
0.59	0.00009 0.00961 0.00717	0.00072 0.01022 0.00651	0.00842 0.00938 0.00508	0.00017 0.00674 0.00495	0.00782 0.00731 0.00626
0.73	0.01117 0.01193 0.00750	0.01060 0.01209 0.00673	0.00955 0.01044 0.00501	0.00864 0.00716 0.00461	0.00722 0.00678 0.00565

Tableau 3.28 : Valeurs théoriques et simulées des EQMI de l'estimateur non-paramétrique F2 pour différentes fenêtres si D = 0.0 , R2 = 9.0 et N = 50.

FIGURES

LISTE DES FIGURES

Figures

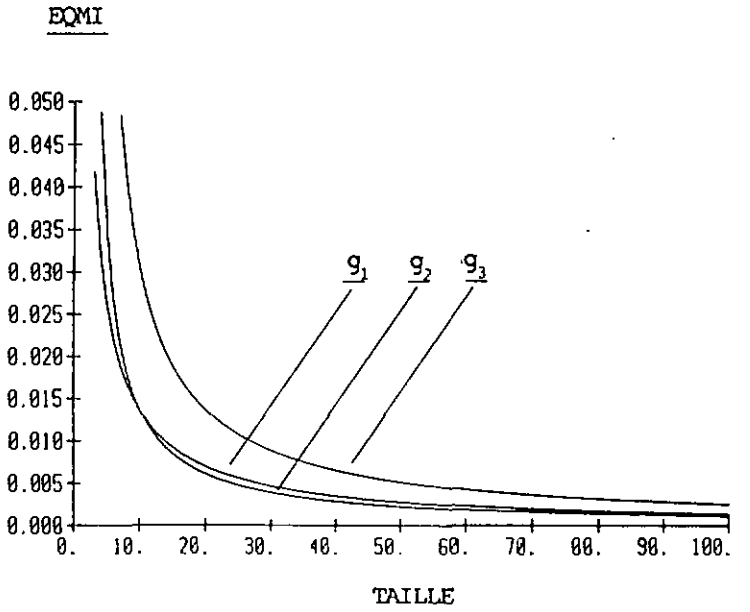


Figure 2.1 : EQMI des estimateurs paramétriques g_1, g_2 et g_3.

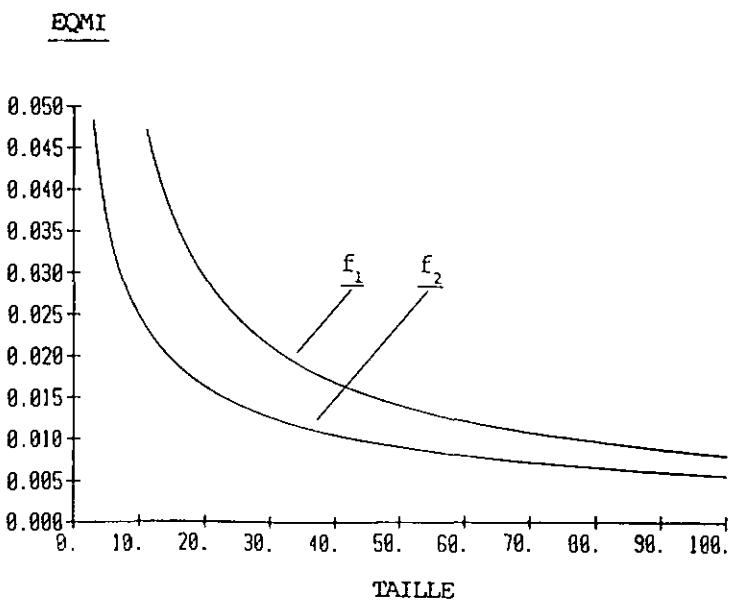


Figure 2.2 : EQMI asymptotique de l'estimateur non-paramétrique f_1 et EQMI exacte de l'estimateur non-paramétrique f_2.

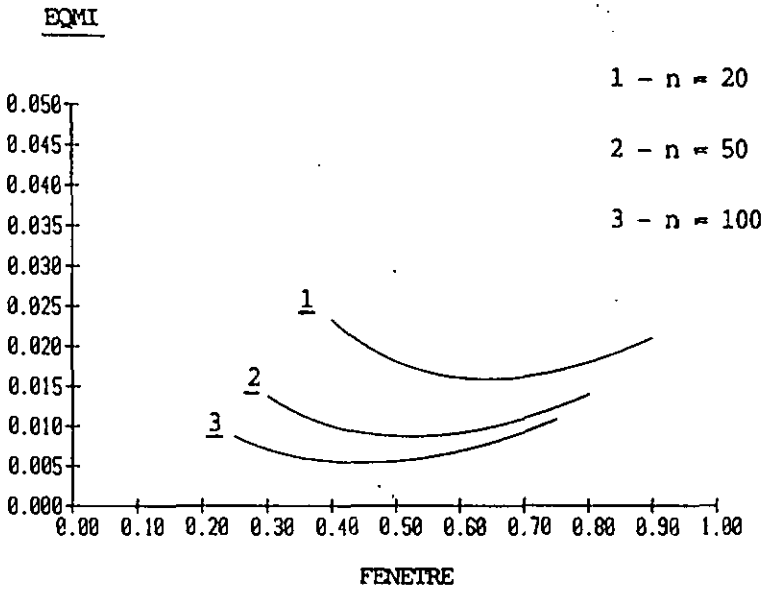


Figure 2.3 : EQMI exacte de l'estimateur non-paramétrique f_2 en fonction de la fenêtre lorsque n varie.

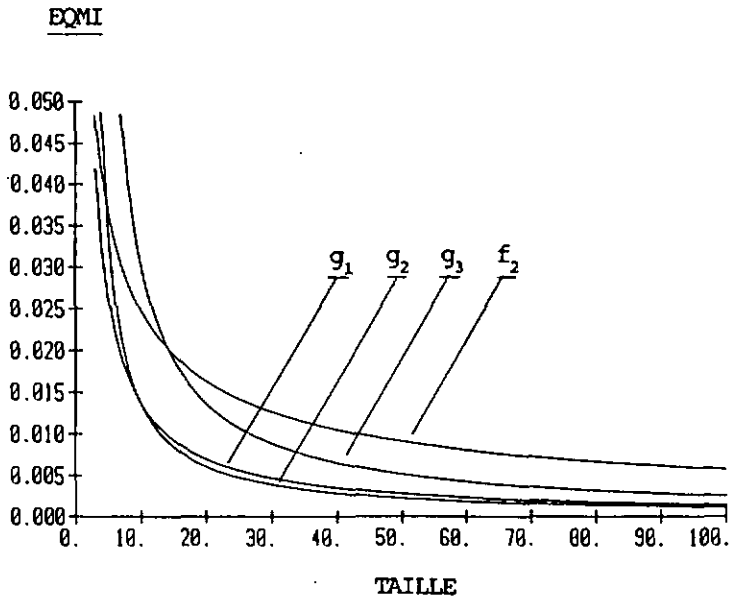


Figure 2.4 : EQMI des estimateurs paramétriques g_1, g_2 et g_3, et de l'estimateur non-paramétrique f_2.

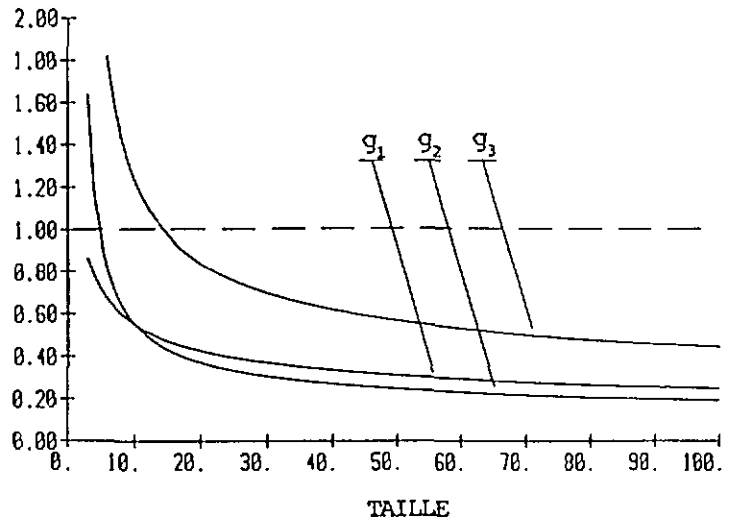


Figure 2.5 : Efficacité de l'estimateur non-paramétrique f_2.

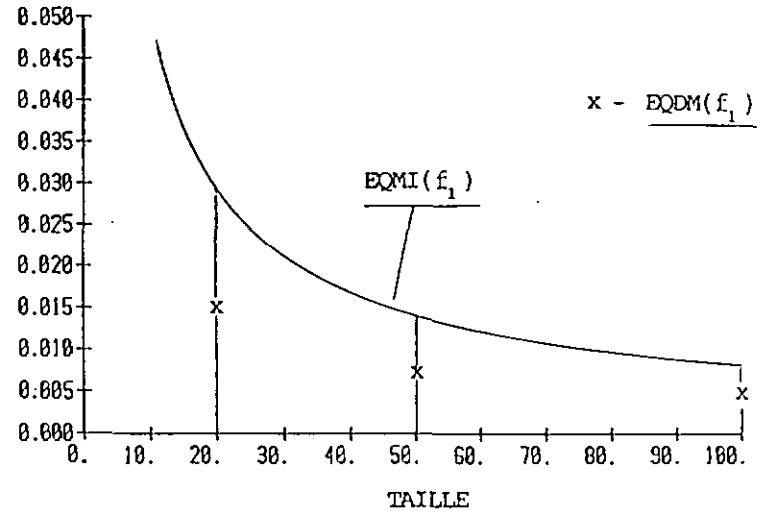


Figure 2.6 : EQMI et EQDM de l'estimateur non-paramétrique f_1 lorsque n varie.

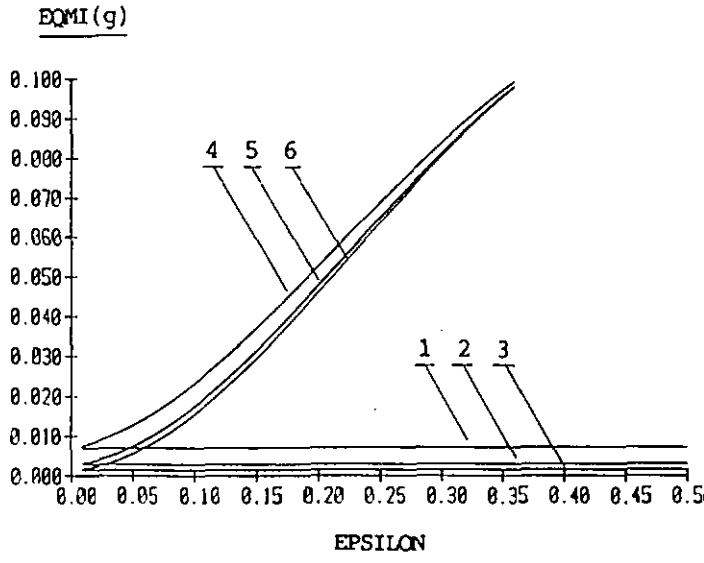


Figure 3.1 : EQMI de l'estimateur paramétrique g, lorsque n varie et d = 1/2 ou d = 3.0.

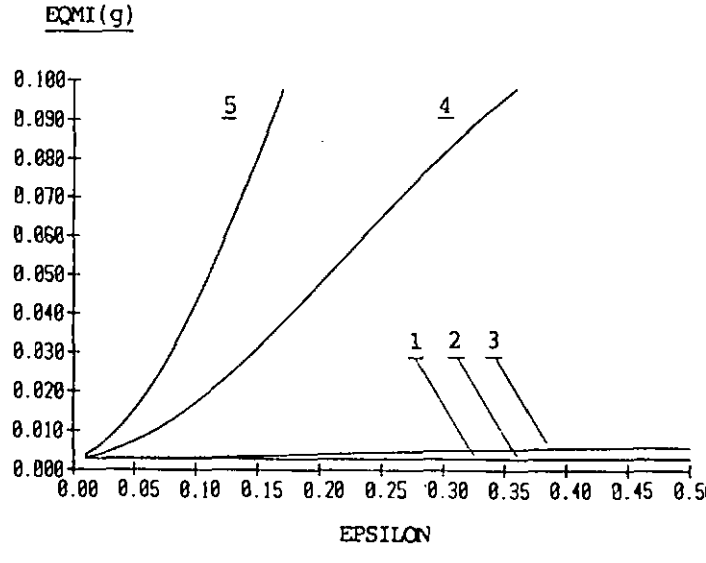


Figure 3.2 : EQMI de l'estimateur paramétrique g, lorsque d varie et n = 50.

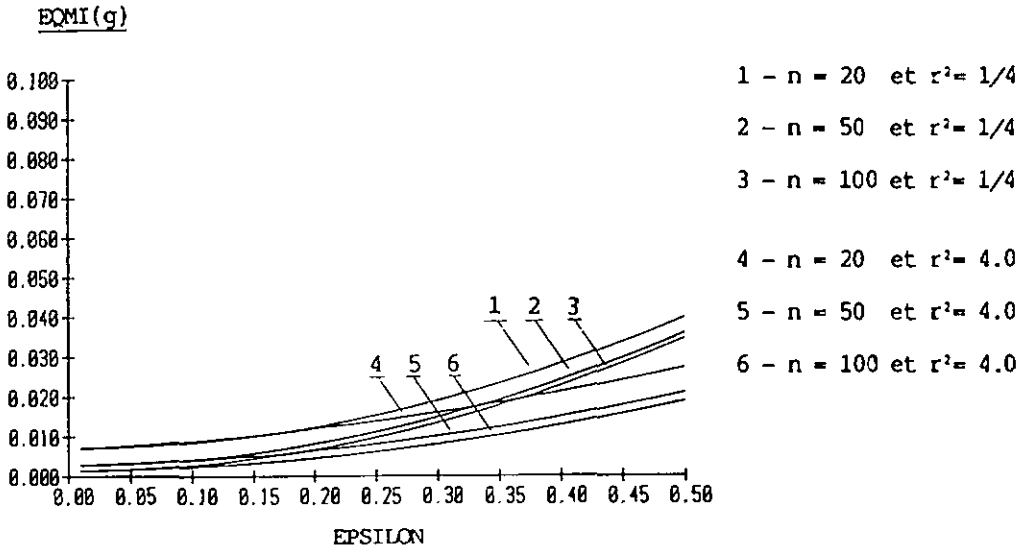


Figure 3.3 : EQMI de l'estimateur paramétrique g, lorsque n varie et $r^2 = 1/4$ ou $r^2 = 4.0$.

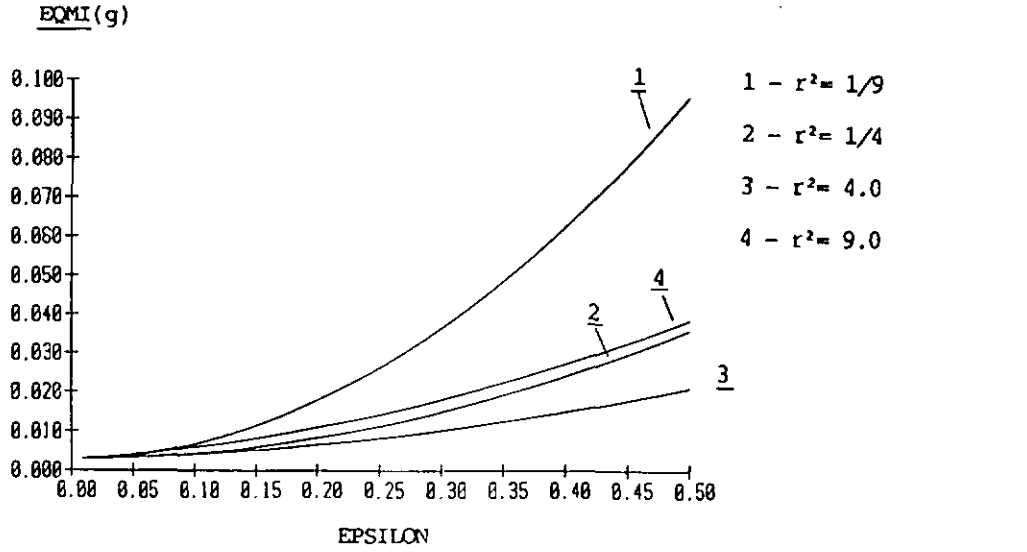


Figure 3.4 : EQMI de l'estimateur paramétrique g, lorsque r^2 varie et n = 50.

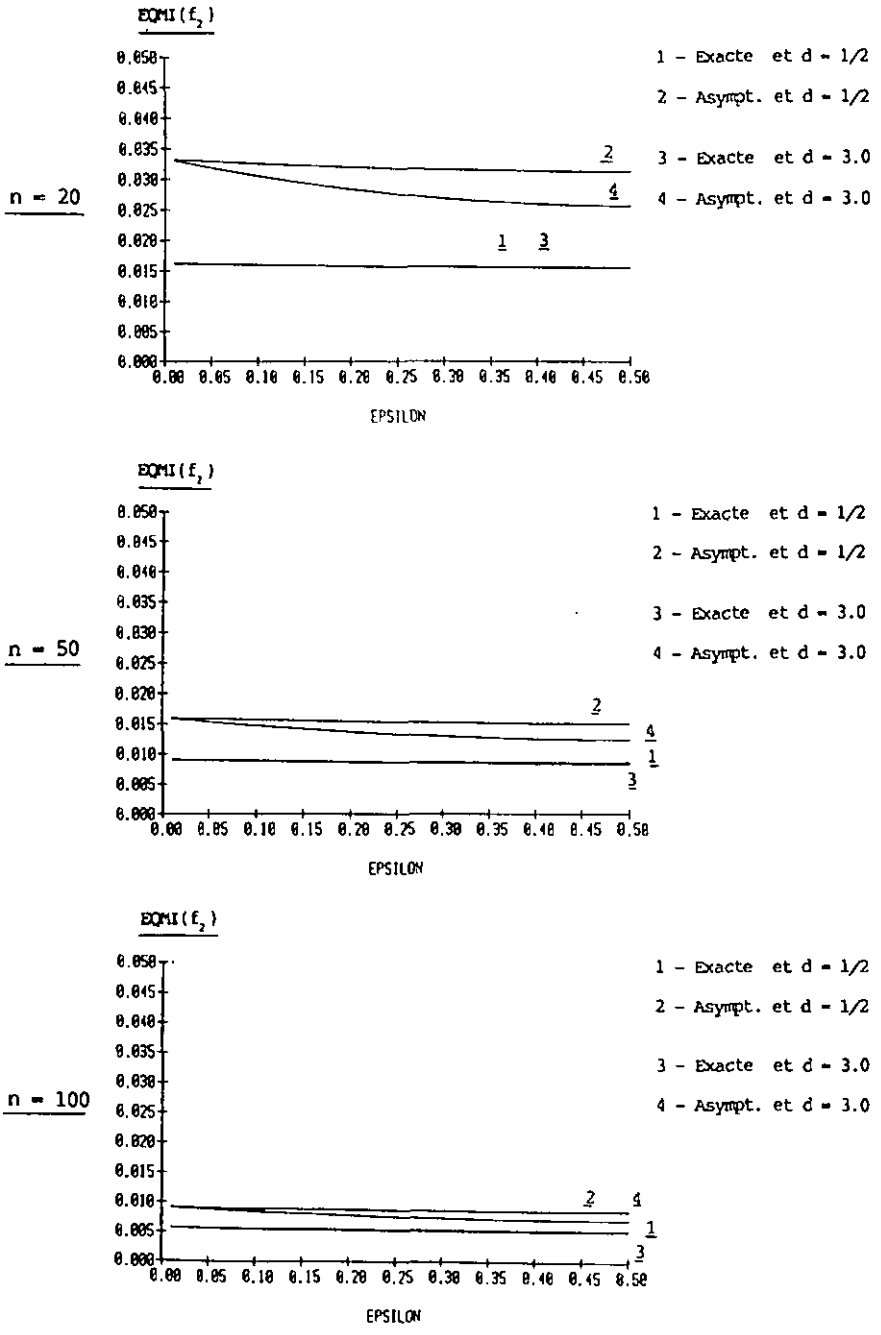
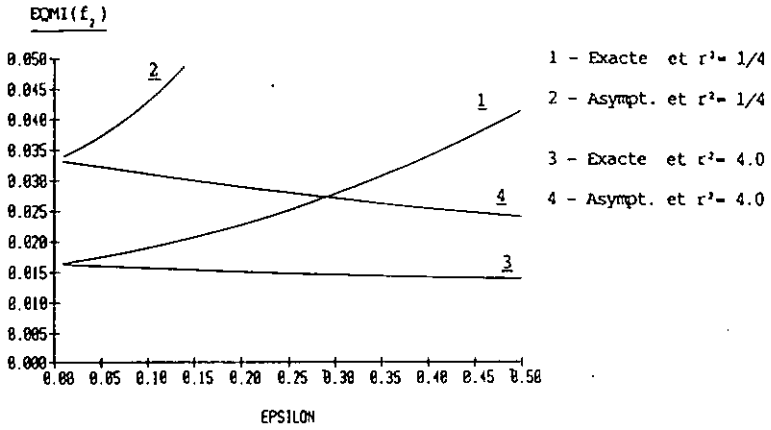
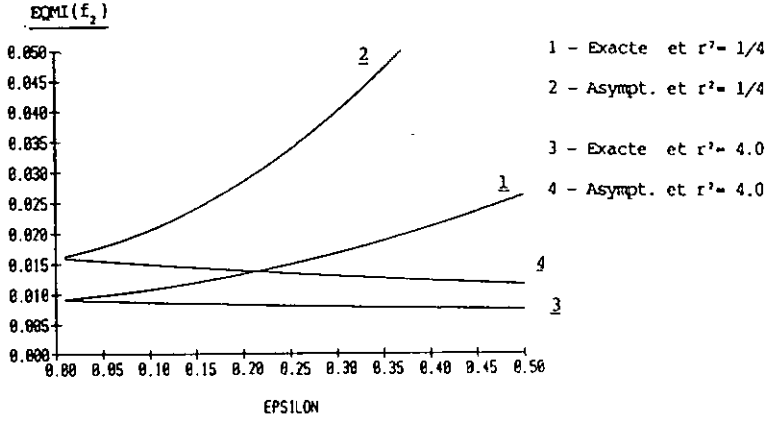


Figure 3.5 : EQMI asymptotique et exacte de l'estimateur non-paramétrique f_2, lorsque n et d varient.

n = 20



n = 50



n = 100

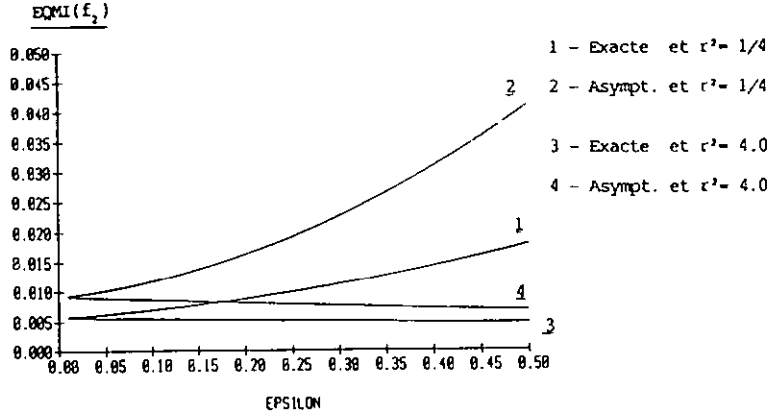


Figure 3.6 : EQMI asymptotique et exacte de l'estimateur non-paramétrique f_2, lorsque n et r^2 varient.

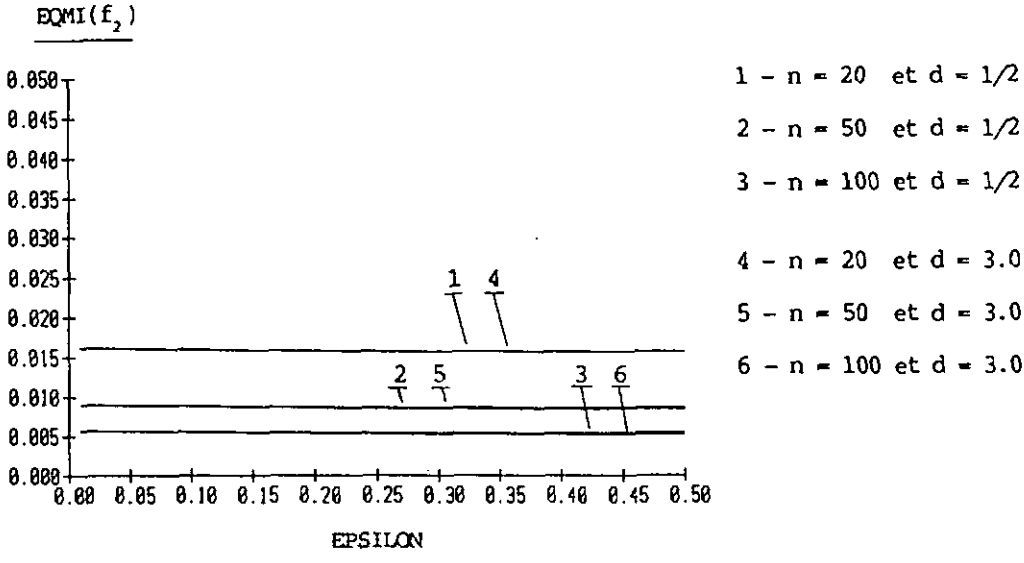


Figure 3.7 : EQMI de l'estimateur non-paramétrique f_2, lorsque n varie et d = 1/2 ou d = 3.0.

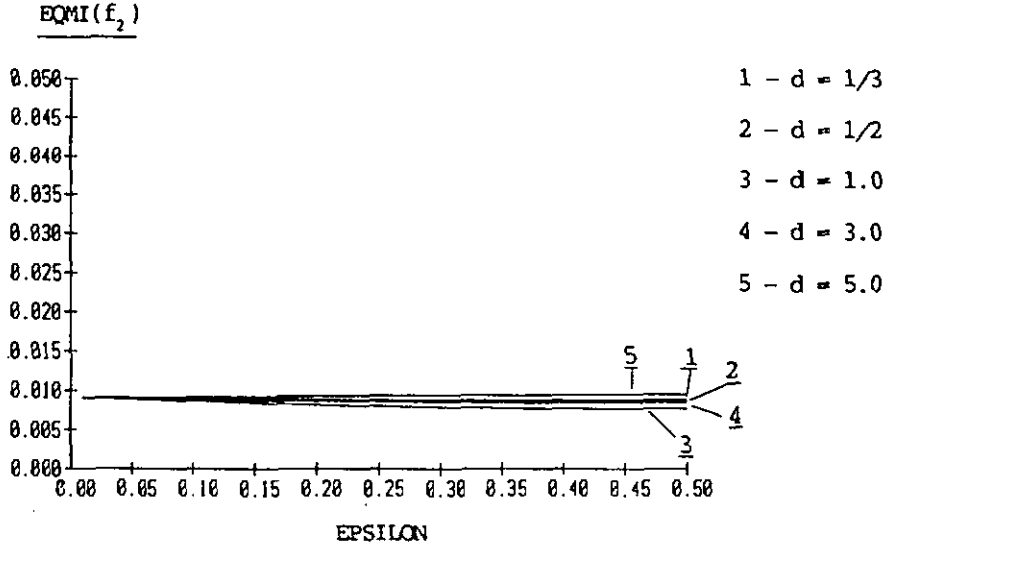


Figure 3.8 : EQMI de l'estimateur non-paramétrique f_2, lorsque d varie et n = 50.

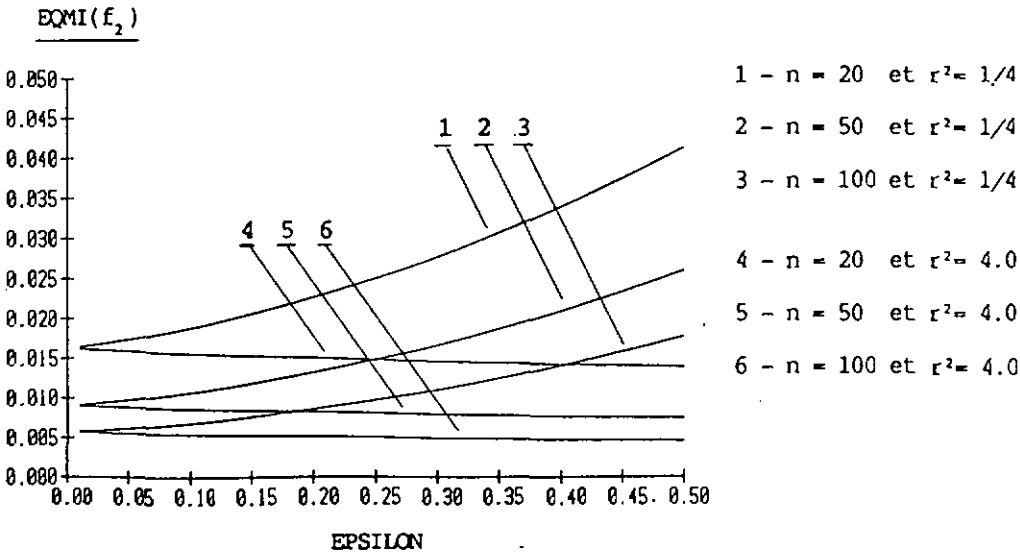


Figure 3.9 : EQMI de l'estimateur non-paramétrique f_2, lorsque n varie et $r^2 = 1/4$ ou $r^2 = 4.0$.

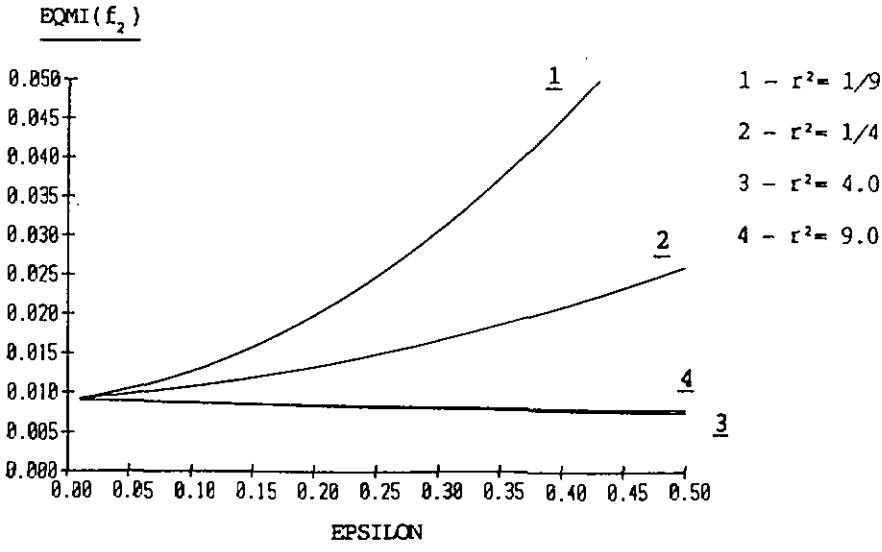


Figure 3.10 : EQMI de l'estimateur non-paramétrique f_2, lorsque r^2 varie et n = 50.

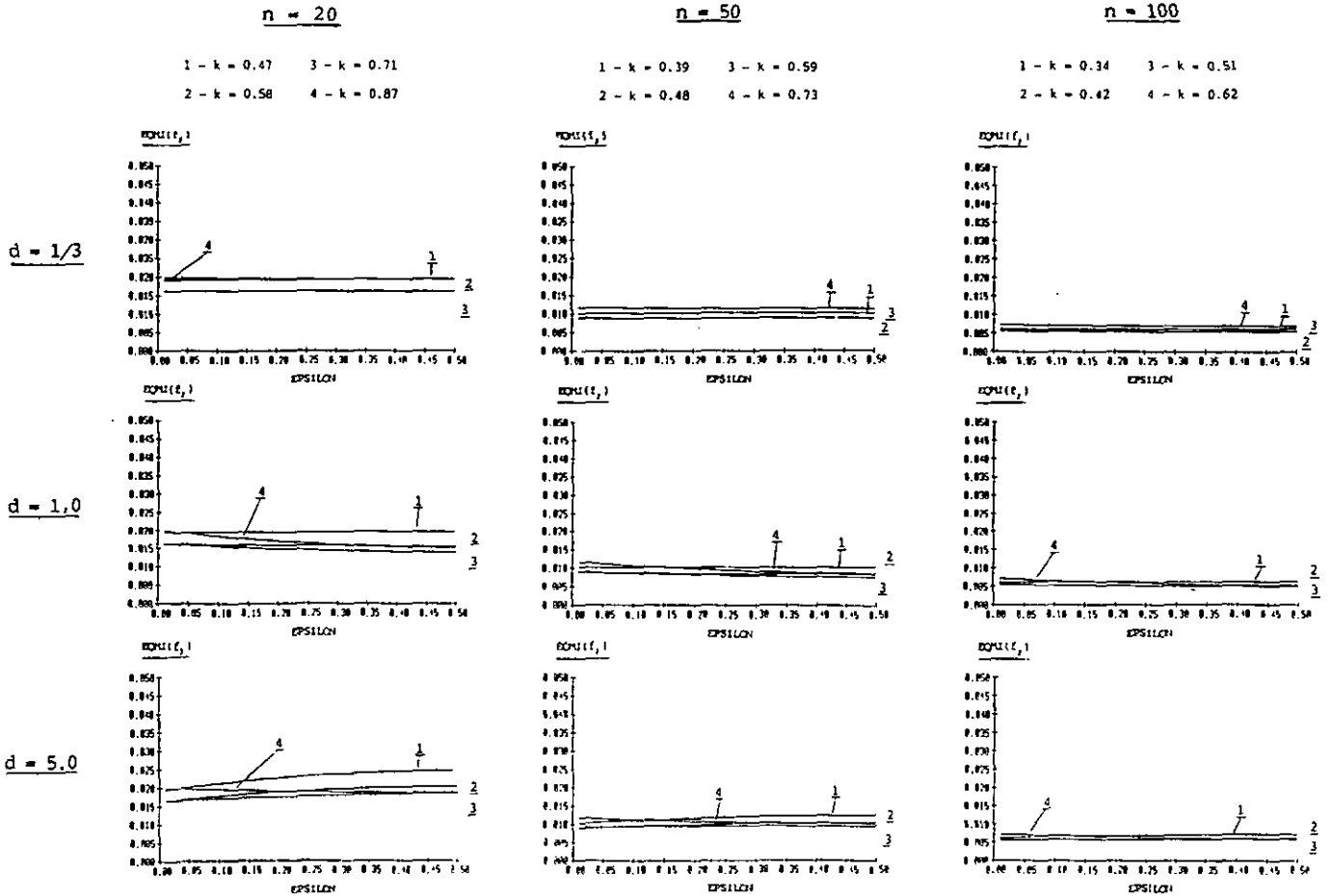
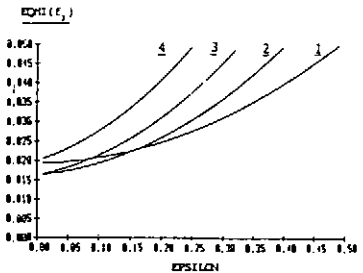
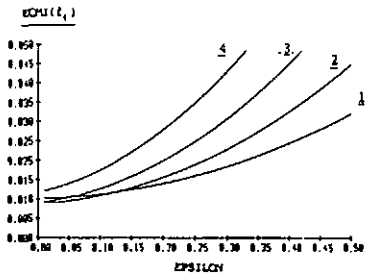


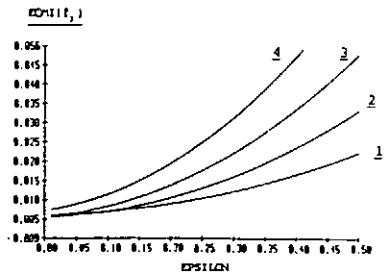
Figure 3.11 : EQMI de l'estimateur non-paramétrique f_2, pour différentes fenêtres k lorsque n et d varient.

n = 20		n = 50		n = 100	
1 - k = 0.47	3 - k = 0.71	1 - k = 0.39	3 - k = 0.59	1 - k = 0.34	3 - k = 0.51
2 - k = 0.58	4 - k = 0.87	2 - k = 0.48	4 - k = 0.73	2 - k = 0.42	4 - k = 0.62

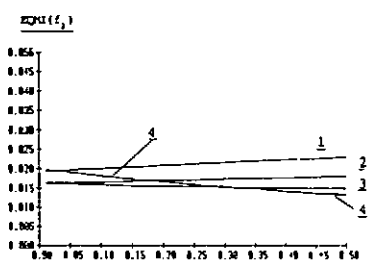
$r^2 = 1/9$

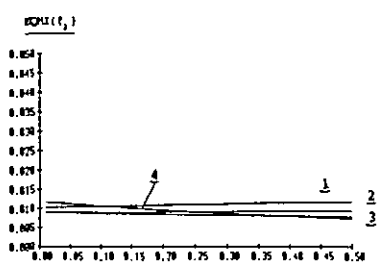






$r^2 = 9.0$





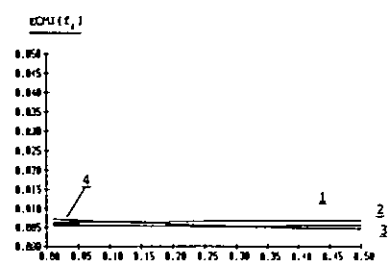


Figure 3.12 : EQMI de l'estimateur non-paramétrique f_2, pour différentes fenêtres k lorsque n et r^2 varient.

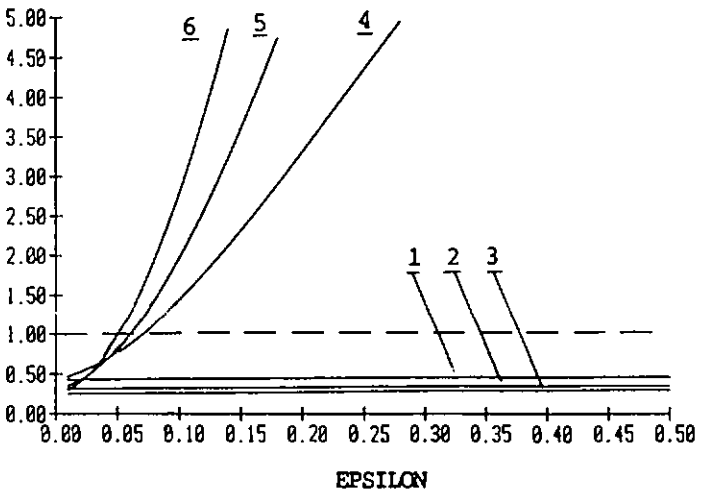


Figure 3.13 : Efficacité de l'estimateur non-paramétrique f_2, lorsque n varie et d = 1/2 ou d = 3.0.

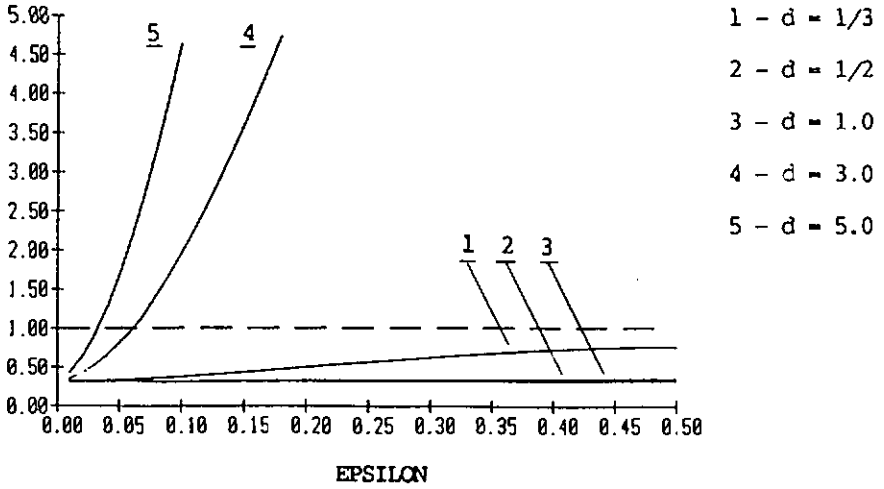


Figure 3.14 : Efficacité de l'estimateur non-paramétrique f_2, lorsque d varie et n = 50.

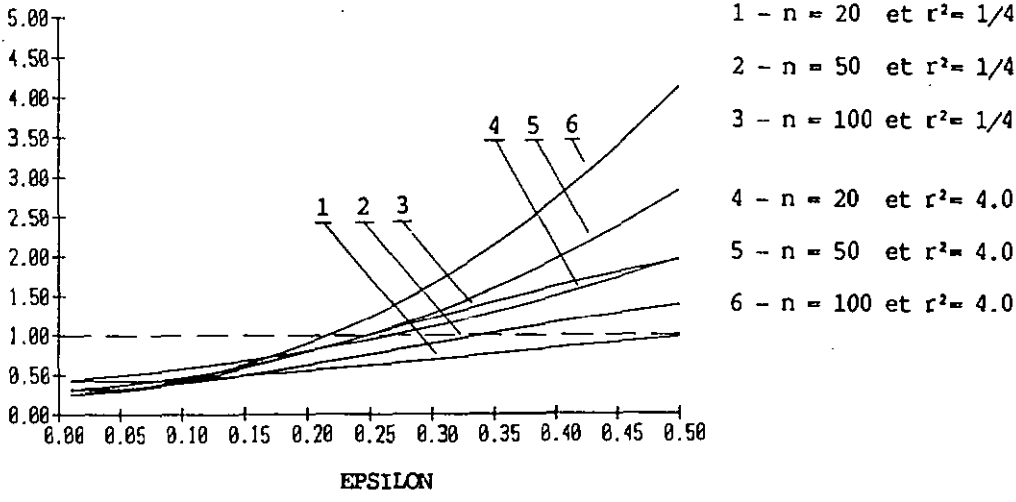


Figure 3.15 : Efficacité de l'estimateur non-paramétrique f_2, lorsque n varie et $r^2 = 1/4$ ou $r^2 = 4.0$.

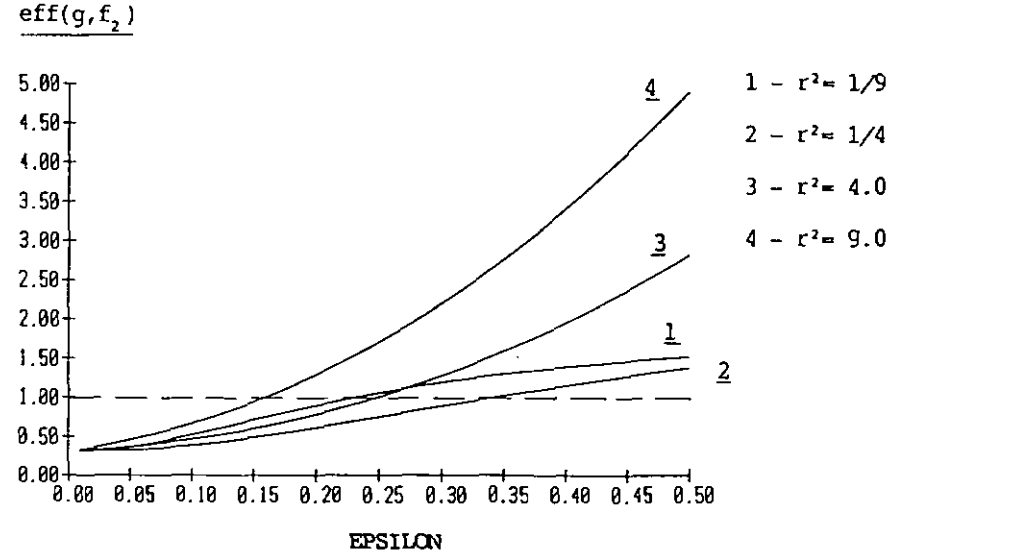


Figure 3.16 : Efficacité de l'estimateur non-paramétrique f_2, lorsque r^2 varie et n = 50.

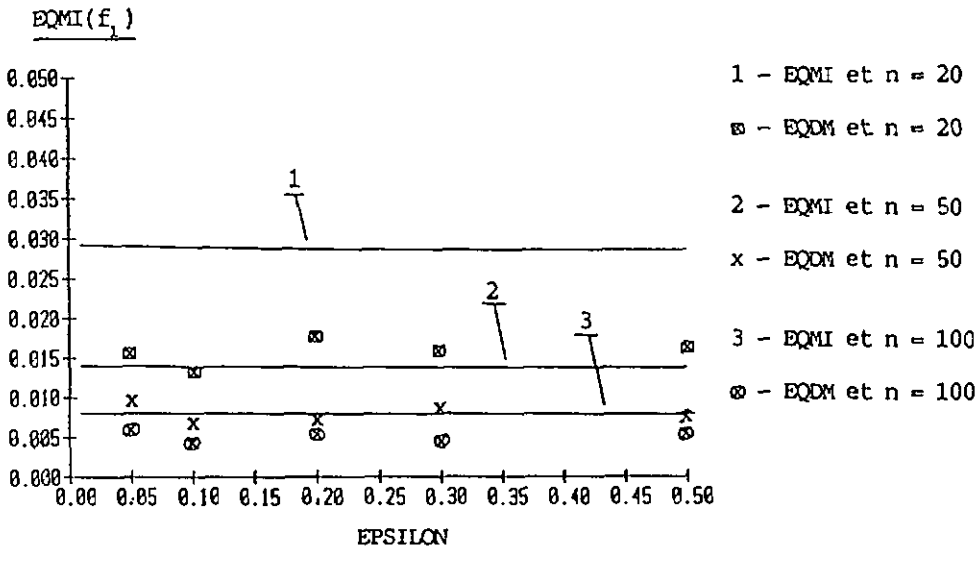


Figure 3.17 : EQMI et EQDM de l'estimateur non-paramétrique f_1, lorsque n varie et d = 1/2.

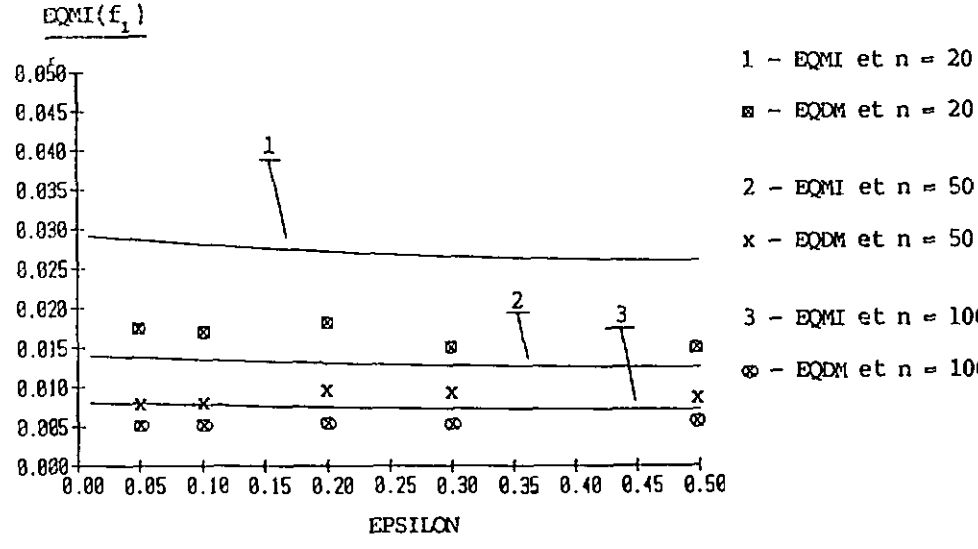


Figure 3.18 : EQMI et EQDM de l'estimateur non-paramétrique f_1, lorsque n varie et d = 3.0.

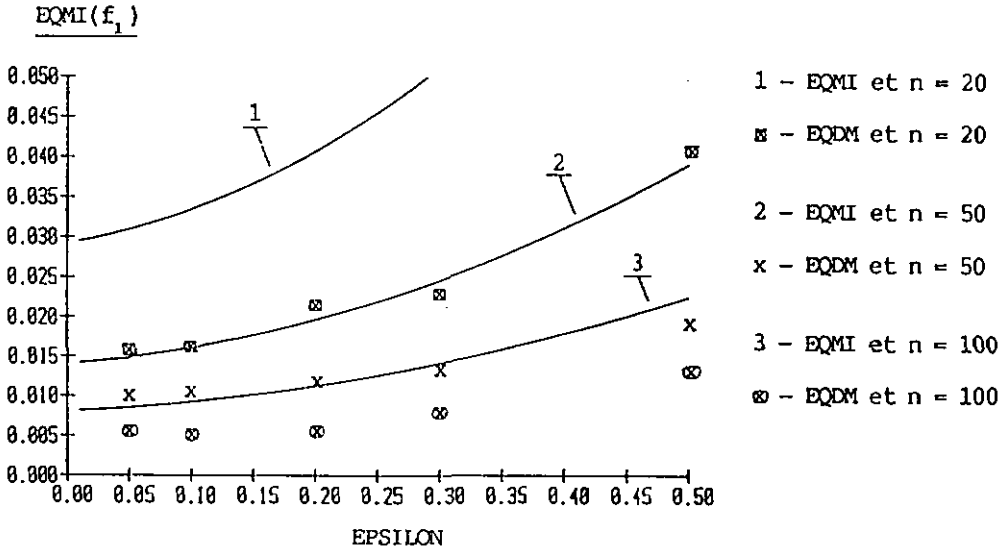


Figure 3.19 : EQMI et EQDM de l'estimateur non-paramétrique f_1, lorsque n varie et $r^2 = 1/4$.

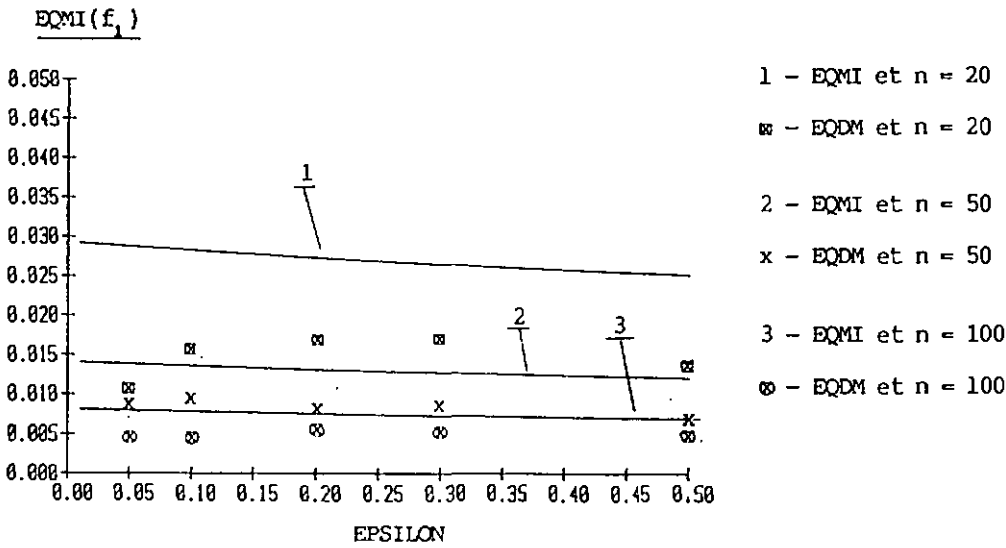
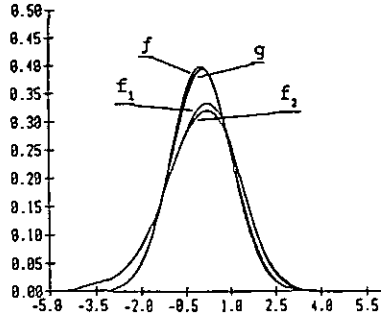
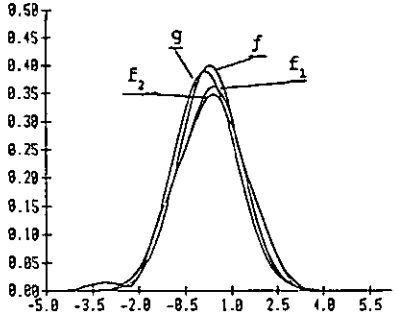


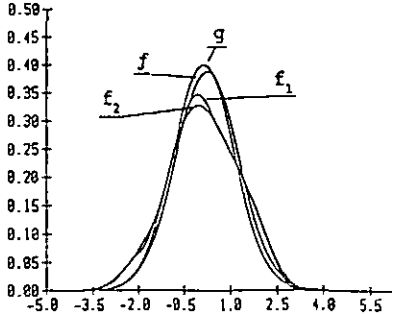
Figure 3.20 : EQMI et EQDM de l'estimateur non-paramétrique f_1, lorsque n varie et $r^2 = 4.0$.

epsilon = 0.1 | epsilon = 0.3 | epsilon = 0.5

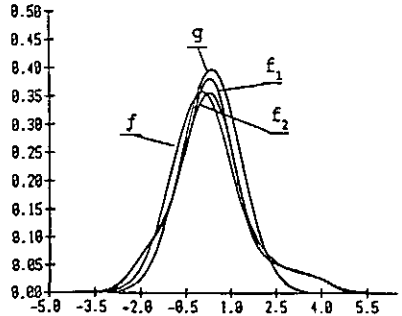
d = 1/2

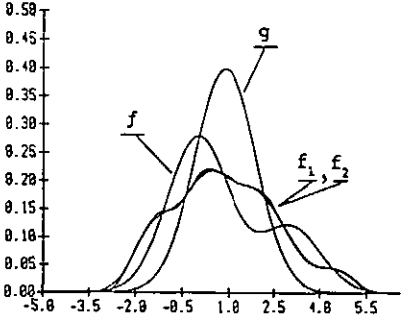






d = 3.0





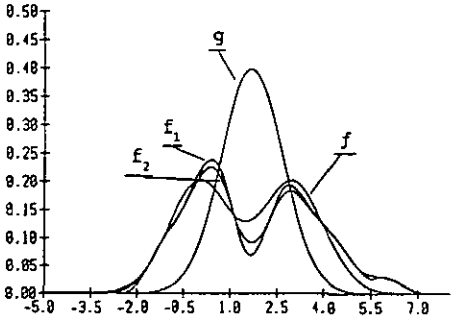


Figure 3.21 : Densités contaminées et estimations pour $r^2 = 1.0$ et $n = 50$.

f : densité exacte
g : estimateur paramétrique

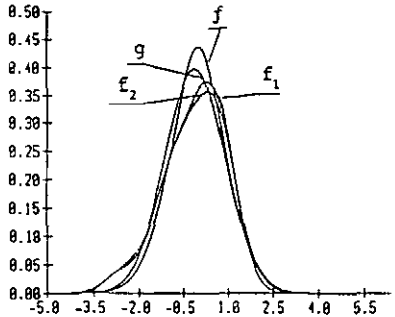
f_1 : estimateur à noyau biquadratique
f_2 : estimateur à noyau normal

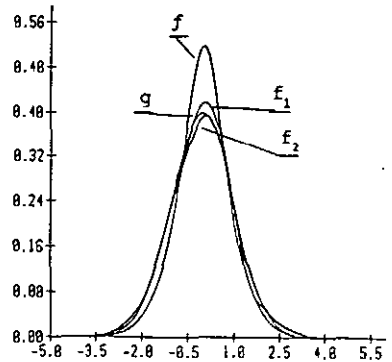
epsilon = 0.1

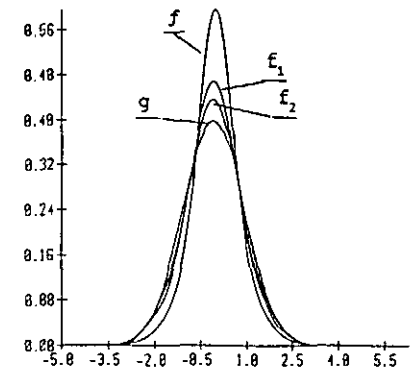
epsilon = 0.3

epsilon = 0.5

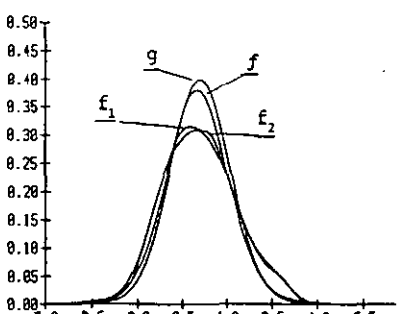
$r^2 = 1/4$

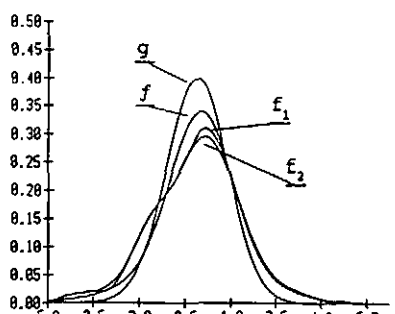






$r^2 = 4.0$





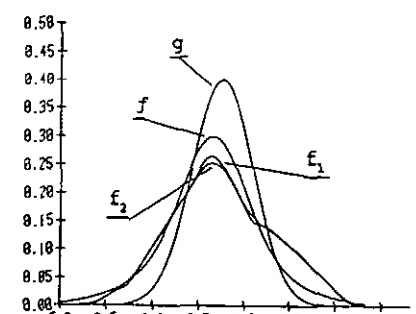


Figure 3.22 : Densités contaminées et estimations pour d = 0.0 et n = 50.

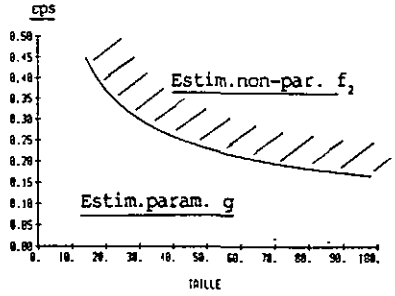
f : densité exacte
g : estimateur paramétrique

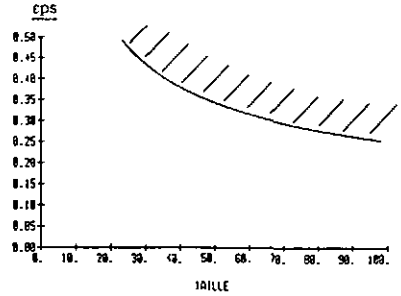
f_1 : estimateur à noyau biquadratique
f_2 : estimateur à noyau normal

$r^2 = 1/9$

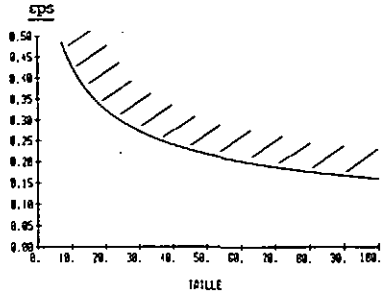
$r^2 = 1/4$

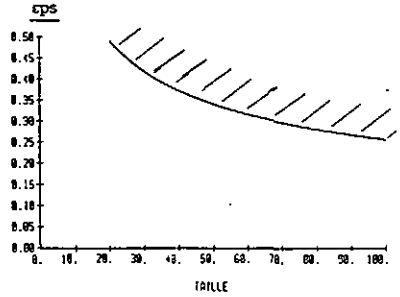
$d = 0.0$



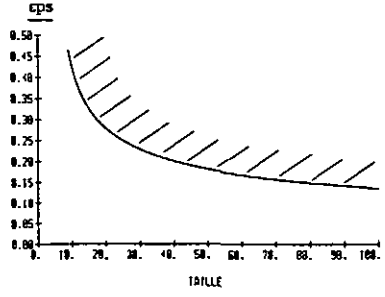


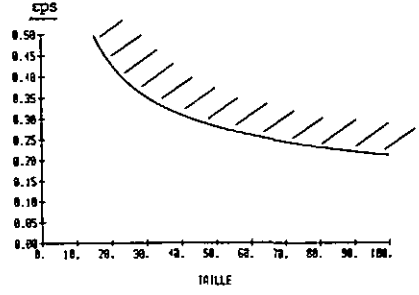
$d = 1/2$





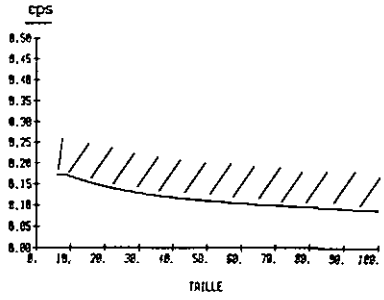
$d = 1.0$

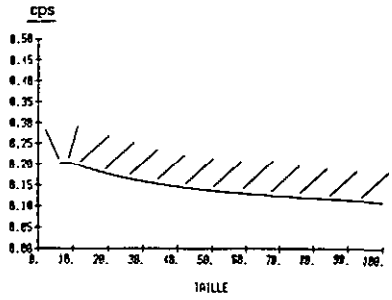




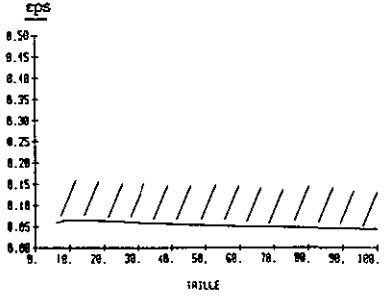
$r^2 = 1/9$ $r^2 = 1/4$

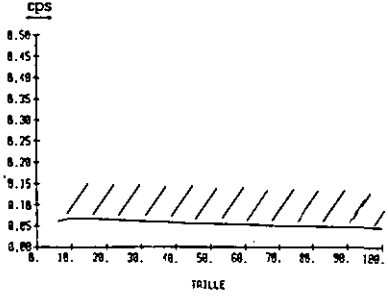
d = 1.5



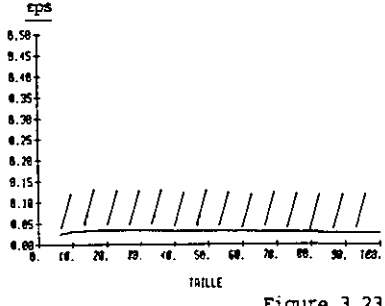


d = 3.0





d = 5.0



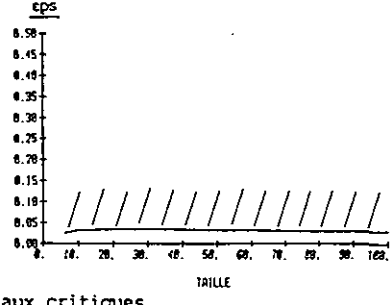
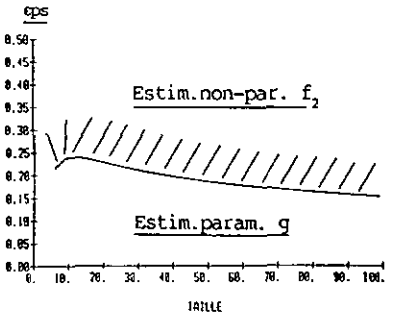


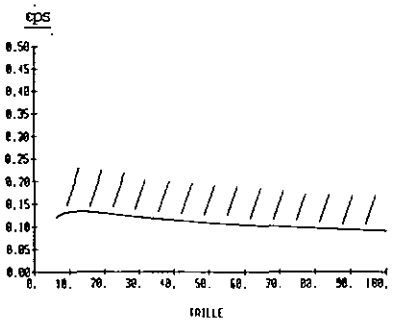
Figure 3.23 : Frontières des taux critiques pour $r^2 = 1/9$ et $r^2 = 1/4$.

Si $d \leq 1.0$, alors l'efficacité est toujours inférieure à 1.0.

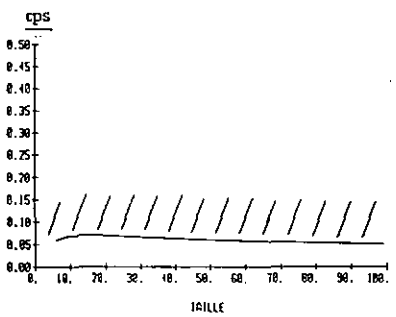
d = 1.5



d = 2.0



d = 3.0



d = 5.0

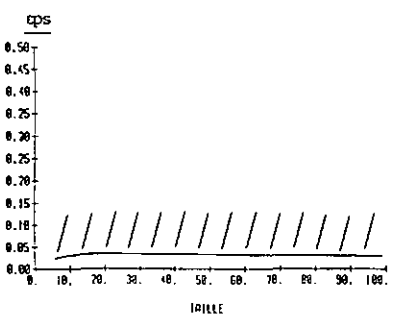
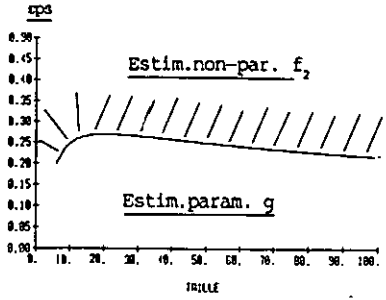


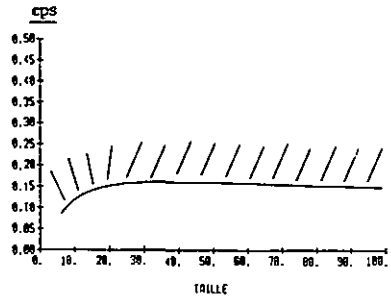
Figure 3.24 : Frontières des taux critiques pour $r^2 = 1.0$.

$r^2 = 4.0$

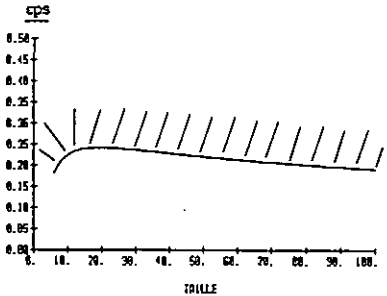
$r^2 = 9.0$

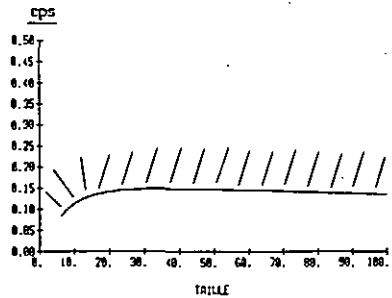
d = 0.0



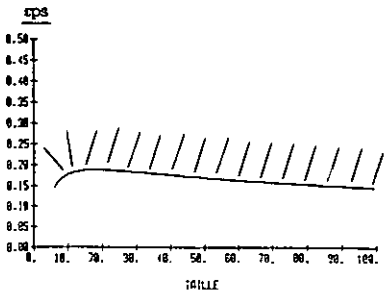


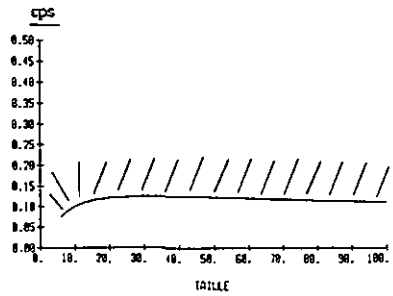
d = 1/2





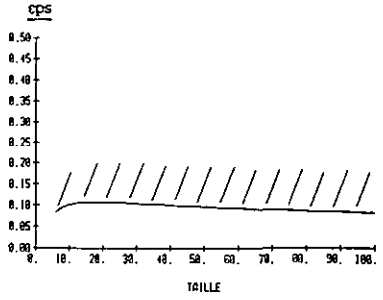
d = 1.0

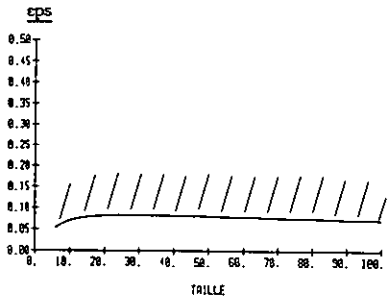




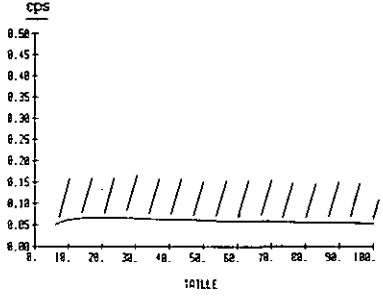
$r^2 = 4.0$ $r^2 = 9.0$

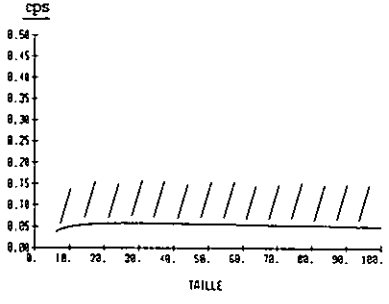
d = 2.0



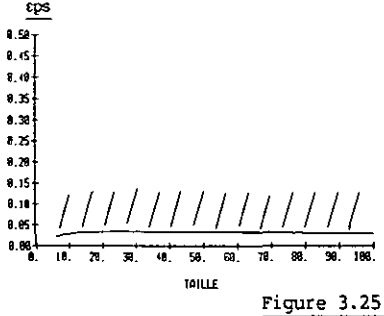


d = 3.0





d = 5.0



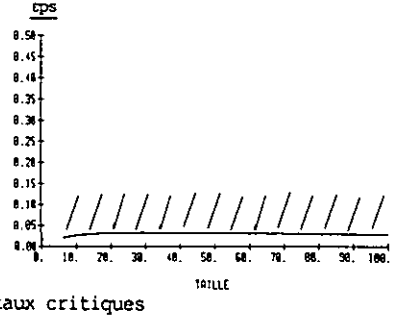


Figure 3.25 : Frontières des taux critiques
pour $r^2 = 4.0$ et $r^2 = 9.0$.