50376
1980
228

N° d'ordre : 859

50376
1980
228

THÈSE

présentée à

l'UNIVERSITÉ DES SCIENCES ET TECHNIQUES DE LILLE I

pour obtenir

LE GRADE DE DOCTEUR 3ème CYCLE

par

Nadji RAHMANIA

ESTIMATION NON PARAMÉTRIQUE DE LA RÉGRESSION : RÉGRESSOGRAMME ET MÉTHODES SPLINE

BIBLIOTHÈQUE UNIVERSITAIRE
Section de SCIENCES
LILLE

Membres du Jury : MM. P. POUZET, Président
D. BOSQ, Rapporteur
P. JACOB, Examinateur
G. COLLOMB, Invité
(Université de Toulouse)

Soutenue le 26 novembre 1980

A ma mère,

A mes frères et soeur
Ahmed, Alima,
Eliassine et Hakim
et particulièrement
à mon cher Rachid,

A tous les miens .

Nous tenons à exprimer notre respectueuse gratitude à Monsieur le Professeur P. Pouzet qui nous a fait l'honneur de présider le jury de cette thèse.

Nous remercions vivement Monsieur le Professeur D. Bosq qui nous a proposé le sujet de cette thèse ; ses critiques stimulantes, ses encouragements, son attentive patience à notre égard et son extrême amabilité ont été pour nous très bénéfiques.

Nos remerciements vont aussi à Monsieur G. Collomb de l'Université de Toulouse ; il a accepté d'examiner notre travail avec rigueur ; nous lui en sommes très reconnaissant.

Monsieur P. Jacob nous a fait l'honneur de faire partie du jury de cette thèse ; nous lui exprimons notre profonde reconnaissance.

Nous tenons aussi à exprimer nos remerciements à Raymonde Bérat et Arlette Lengaigne qui ont assuré la réalisation matérielle de cette thèse ; ainsi que Françoise Wdowczyk, Monique Lloret, Albert Gournay et Michel Provost.

TABLE DES MATIERES.

Pages

INTRODUCTION. 1

CHAPITRE I - GENERALITES. 3

I.1. - Paramètres de régression. 3

I.2. - Exemples d'estimation non-paramétrique de la régression. 4

I.3. - Inexistence d'estimateur sans biais. 8

I.4. - Estimation par les fonctions spline. 9

CHAPITRE II - NOTATIONS, DEFINITIONS ET HYPOTHESES GENERALES. 12

II.1. - Construction de l'estimateur - spline (régrospline). 12

II.2. - Représentation de la régrospline par les B-splines. 16

II.3. - Hypothèses générales. 17

CHAPITRE III - ETUDE ASYMPTOTIQUE DU REGRESSOGRAMME $R_n(x)$, x FIXE. 19

III.1. - Convergence asymptotiquement sans biais. 19

III.1.1. - Résultats préliminaires.

III.1.2. - Limite et majoration asymptotique du biais (en valeur absolue).

III.2. - Majoration asymptotique de la variance de $R_n(x)$. 24

III.3. - Non-corrélation asymptotique de $R_n(x)$, $R_n(x')$, $x \neq x'$ fixés. 28

III.4. - Convergence en moyenne quadratique de $R_n(x)$. 31

III.5. - Convergence en moyenne quadratique intégrée. 32

III.5.1. - Introduction.

III.5.2. - Majoration asymptotique du MISE.

III.5.3. - Limite du MISE et vitesse de convergence.

./..

Pages

CHAPITRE IV - ETUDE ASYMPTOTIQUE DE L'ESTIMATEUR-SPLINE $\hat{R}_n(x)$, x FIXE. 35

IV.1. - Convergence asymptotiquement sans biais. 35

IV.1.1. - Résultats préliminaires.

IV.1.2. - Majoration asymptotique du biais (en valeur absolue).

IV.1.3. - Convergence asymptotiquement sans biais.

IV.2. - Comportement asymptotique de la variance de $\hat{R}_n(x)$. 41

IV.2.1. - Résultats préliminaires.

IV.2.2. - Majoration asymptotique de la variance de $\hat{R}_n(x)$.

IV.3. - Non-corrélation asymptotique de $\hat{R}_n(x)$ et $\hat{R}_n(x')$, $x \neq x'$ fixés. 45

IV.4. - Comportement de l'erreur quadratique de $\hat{R}_n(x)$. 46

IV.4.1. - Majoration asymptotique de l'erreur quadratique.

IV.4.2. - Convergence en moyenne quadratique.

IV.4.3. - Remarques.

IV.4.4. - Optimisation selon le critère de "lissage".

IV.5. - Convergence en moyenne quadratique intégrée. 49

IV.5.1. - Introduction.

IV.5.2. - Résultats préliminaires.

IV.5.3. - Limite du MISE.

IV.6. - Convergence uniforme presque-complète et en moyenne uniforme de $\hat{R}_n$. 57

IV.6.1. - Introduction.

IV.6.2. - Résultats préliminaires.

IV.6.3. - Convergence uniforme presque-complète de $\hat{R}_n$.

IV.6.4. - Convergence en moyenne uniforme.

CHAPITRE V - SIMULATION. 70

IV.1. - Introduction. 70

IV.2. - Exemples. 71

BIBLIOGRAPHIE. 96

INTRODUCTION.

L'estimation non-paramétrique de la régression revêt un grand intérêt en statistique mathématique et trouve diverses applications, notamment dans les problèmes de prévision.

A l'origine, il y avait une idée intuitive très évidente ; par exemple, pour prévoir le temps au lendemain du jour caractérisé par une certaine quantité x, il semblait naturel de considérer le temps au lendemain des jours caractérisés par des observations x_i , $i = 1,2,\ldots,n$, assez "voisines" de x.

Des méthodes d'estimation non-paramétriques de la régression ont été proposées principalement par Bhattacharya et Parthasarathy [5] ; ces derniers utilisent une méthode d'analyse graphique fractile, d'après une idée de Mahalanobis. Nadaraya [20], Watson [36] et Schuster [27] proposent, quant à eux, une méthode plus générale, bien connue lorsqu'elle est utilisée pour estimer une densité de probabilité et appelée habituellement "méthode du noyau". Plus récemment, Collomb [9] a commenté les études de ces trois derniers auteurs et a obtenu des propriétés de convergence intéressantes, dans la mesure où des préoccupations concrètes, telles que les problèmes de prévision météorologique, de reconnaissance des formes ou d'intelligence artificielle y trouvent un support mathématique certain. Bosq [7] a proposé une méthode plus générale encore, à notre avis, et qui permet d'estimer les fonctions de régression en partant de l'estimation de la densité conditionnelle et d'obtenir un cas particulier, le régressogramme, introduit par Tukey [32] et étudié principalement par Geffroy [15], Major [17] et Sabry [26].

Par ailleurs, Stone [31] a construit un estimateur qui possède un certain nombre de propriétés et qui rejoint, dans un certain sens, la méthode de Bosq.

L'objet de cette thèse est l'étude d'un nouvel estimateur de la régression, basé sur l'interpolation par les fonctions-spline. Ces dernières, introduites par Schoenberg [29] en 1946, jouent un rôle très important en analyse numérique et plus particulièrement dans la théorie de l'approximation. Leurs bonnes propriétés ont notamment contribué à la résolution d'équations fonctionnelles (équations intégrales, différentielles et aux dérivées partielles) ; par ailleurs, il est important de remarquer qu'il y a deux façons différentes d'étudier les splines : la première, de type algébrique, consiste à choisir ces fonctions comme des morceaux de polynômes ou de certaines fonctions qui se raccordent en des points convenablement choisis, appelés noeuds de la spline ; la deuxième, consiste à prendre les splines comme des fonctions qui minimisent une certaine fonctionnelle dans un espace approprié ; ce dernier point de vue s'est avéré extrêmement fécond et a permis une grande généralisation grâce, en particulier, aux travaux d'Atteia, P.J. Laurent et J.P. Aubin.

Leur première application en statistique mathématique a été faite, à notre connaissance, par Boneva, Kendall et Stephanov (B.K.S.) [6] ; ces derniers ont construit un estimateur de la densité de probabilité en interpolant convenablement la fonction de répartition empirique par une spline cubique.

G. Wahba [34] a montré que l'estimateur-spline proposé par B.K.S. converge ponctuellement en moyenne quadratique. Berlinet [3] a prouvé que cet estimateur appartient à la classe des estimateurs de la densité proposée par Bleuez et Bosq [4].

Dans ce qui suit, on estimera la régression en partant du même principe en procédant à un lissage du régressogramme.

CHAPITRE I

GENERALITES SUR L'ESTIMATION NON-PARAMETRIQUE DE LA REGRESSION.

Pour mieux situer notre travail, il nous semble utile de rappeler quelques résultats et définitions relatives à l'estimation non-paramétrique de la régression.

I.1. - *Paramètres de régression* (Geffroy [15]).

Soit (X,Y) un couple aléatoire à valeurs dans $\mathcal{X} \times \mathcal{Y}$ où $\mathcal{X}$ désigne un pavé de $\mathbb{R}^s$, s entier positif, et $\mathcal{Y}$ désigne $\mathbb{R}$ ou un intervalle de $\mathbb{R}$; on suppose que la loi P_{XY} du couple (X,Y) dans l'espace $(\mathbb{R}^{s+1}, \mathcal{B}_{\mathbb{R}^{s+1}})$ inconnue, mais appartenant à un ensemble donné $\mathcal{D}$, défini par des hypothèses de régularité mathématique, ceci justifiant l'emploi des méthodes non-paramétriques. On sait que Y admet une famille de lois conditionnelles $(P_Y^x)_{x \in \mathcal{X}}$, bien définie modulo P_X ; soit π un espace topologique et ϕ une application de l'ensemble $\{P_Y^x,\ x \in \mathcal{X},\ P \in \mathcal{D}\}$ dans π ; nous dirons que la fonction :

$$\bar{\omega}(P,x) = \phi(P_Y^x)$$

est un paramètre de régression pour Y.

Soit (X_1,Y_1), (X_2,Y_2) ,..., (X_n,Y_n) un échantillon de taille n de la loi P_{XY} ; un estimateur de $\bar{\omega}$ reposant sur cet échantillon est une fonction de la forme :

$$\bar{\omega}_n(\{X_1,Y_1\}\ ,\ \{X_2,Y_2\}\ ,\ldots,\ \{X_n,Y_n\}\ ;\ x)$$

qui est mesurable par rapport à l'ensemble des variables dont elle dépend et prend ses valeurs dans un espace topologique π' dont π est un sous-espace.

Exemples de paramètres de régression.

a) Si le support de chaque loi P_Y^X admet une borne supérieure $b(P,x)$, pour tout x de X et tout P de $\mathcal{D}$, la fonction b est un paramètre de régression.

b) Si toute loi P_Y^X admet une densité conditionnelle $g(P,x)$, pour tout x de X et tout P de $\mathcal{D}$, la fonction g est un paramètre de régression ; Bosq [7] donne un estimateur de g que nous évoquerons dans ce chapitre.

c) Si nous désignons par $\bar{\omega}(P,x)$ la médiane, convenablement précisée, de la loi P_Y^X, pour x de X et tout P de $\mathcal{D}$, la fonction $\bar{\omega}$ est un paramètre de régression.

d) Nous supposons que pour tout P de $\mathcal{D}$, EY existe. Pour P fixé dans $\mathcal{D}$, la régression de Y en x, définie comme une détermination de l'espérance conditionnelle $E(Y/X)$ est un paramètre de régression ; l'estimation de ce dernier est l'objet de notre étude.

Dans ce qui suit, on va donner quelques résultats relatifs à l'estimation de ce paramètre.

I.2. - Exemples d'estimation non-paramétrique de la régression.

Soit P_Y^X une version régulière de la probabilité conditionnelle de Y par rapport à X telle que :

$$(\forall x \in \mathbb{R}^s) , \qquad R(x) = \int y \, P_Y^X \, dy \qquad \text{existe} .$$

L'estimateur qui, selon nous, a suscité le plus de recherches est celui étudié par Nadaraya [20], Watson [36] et Schuster [27] ; il est connu sous le nom d'estimateur par la méthode du noyau ; nous le désignons par R_n^K et le définissons de la manière suivante : X étant un vecteur aléatoire de $\mathbb{R}^s$ et Y une v.a.r. , on se donne n observations indépendantes du couple (X,Y) , notées (X_i,Y_i), $i = 1,2,\ldots,n$; on suppose, par ailleurs, l'existence d'une densité p, strictement positive pour la loi P_{XY} , continue et telle que :

$$(\forall x \in \mathbb{R}^s), \qquad R(x) = \frac{\int y\, p(x,y)\, dy}{\int p(x,y)\, dy}$$

On pose alors :

$$(\forall x \in \mathbb{R}^s), \qquad R_n^K(x) = \begin{cases} \dfrac{\sum_{i=1}^{n} Y_i\, K\left(\frac{x-X_i}{h_n}\right)}{\sum_{i=1}^{n} K\left(\frac{x-X_i}{h_n}\right)} & \text{si} \quad \sum_{i=1}^{n} K\left(\frac{x-X_i}{h_n}\right) \neq 0 \\ 0 & \text{sinon} \end{cases}$$

où h_n désigne une suite réelle strictement positive, de limite nulle et K un noyau positif.

Collomb [9] a cité les études faites par Nadaraya, Watson et Schuster sur l'estimateur R_n^K ; il a obtenu des résultats relatifs à cet estimateur et à l'estimateur des dérivées de la régression. On peut souligner, aussi, que R_n^K est le rapport de deux estimateurs : le numérateur et le dénominateur sont respectivement les estimateurs, à une constante multiplicative près, de

$$\phi(x) = \int y\, p(x,y)\, dy \quad \text{et} \quad f(x) = \int p(x,y)\, dy \qquad \forall x \in \mathbb{R}^s ,$$

par la méthode du noyau.

Schuster et Yakowitz [28] ont donné des résultats relatifs à la convergence uniforme de $R_n^K(x)$, x dans $[0,1]$ et les vitesses de convergence correspondantes ; des résultats relatifs aux estimateurs des dérivées de R_n^K ont été, aussi, proposés par ces auteurs.

Benedetti [2] a également étudié l'estimation non-paramétrique de la régression en utilisant la méthode du noyau ; elle donne essentiellement la convergence asymptotiquement normale de $R_n^K(x)$, x dans $[0,1]$.

Bosq [7] propose une méthode qui nous paraît plus générale ; son estimateur est définie comme suit : étant donné un estimateur de la densité de probabilité , on lui associe un estimateur $g_n^{[x]}$ de la densité conditionnelle $g^{[x]}$, à partir de la densité univariée ; pour tout x dans $\mathbb{R}^s$, on pose :

$$R_n^B(x) = \int y\, g_n^{[x]}(y)\, dy$$

La construction de cet estimateur se justifie par le fait que la régression peut s'écrire :

$$(\forall x \in \mathbb{R}^s) , \qquad R(x) = \int y\, g^{[x]}(y)\, dy$$

Un cas particulier de la méthode de Bosq est l'estimateur proposé par Tukey [32] ; cet estimateur appelé le plus souvent "régressogramme" est construit de la façon suivante : soit $[u,v]$ un intervalle compact et soit une partition de $[u,v]$ en k intervalles de même longueur, k entier positif ; on estime la fonction de régression par la fonction en escalier qui, dans chaque intervalle de la partition est constante et est

égale à la moyenne arithmétique des ordonnées Y_i , $i = 1,2,\ldots,n$ des points de l'échantillon dont l'abcisse X_i appartient à l'intervalle considéré ; si aucun point X_i de l'échantillon ne prend sa valeur dans cet intervalle, on prend, alors, 0 comme valeur de la fonction.

La définition de cet estimateur s'étend facilement au cas où $s \geq 1$, en remplaçant $[u,v]$ par $[u,v]^s$ et en considérant k^s cubes de $\mathbb{R}^s$.

Geffroy [15] montre dans quelles conditions le régressogramme converge uniformément en probabilité et uniformément presque - complètement ; la condition fondamentale que doit vérifier la suite k est :

$$\lim_{n\to\infty} \left(\frac{1}{k^s} \times \frac{n}{\text{Log } n}\right) = +\infty$$

Major [17] donne un résultat sur la loi limite de cet estimateur et plus récemment Sabry [26] a étudié la convergence uniforme du régressogramme sur des intervalles dont la longueur tend vers l'infini avec la taille de l'échantillon; ceci est important car l'estimation de la régression se fait sur tout $\mathbb{R}$; par ailleurs, le même auteur a considéré le lissage du régressogramme par une ligne polygone dans un premier temps et par les polynômes de Bernstein dans un deuxième temps.

Une méthode qui utilise la méthode du noyau et la méthode de Bosq est proposée par T. Gasser et H.G. Müller [19] ; en effet, ces deux derniers utilisent, en quelque sorte, la méthode de Bosq et estiment la densité conditionnelle par la méthode du noyau.

Revesz [24] propose, pour sa part, un estimateur défini de la façon suivante :

$(\forall x \in [0,1])$, on pose :

$$\hat{r}_o(x) \equiv 0$$

$$\hat{r}_{n+1}(x) = \hat{r}_n(x) + \frac{1}{(n+1)h_{n+1}} K\left(\frac{X-X_{n+1}}{h_{n+1}}\right) \cdot (Y_{n+1} - \hat{r}_n(x))$$

où $h_n = \frac{1}{n^\alpha}$, $0 < \alpha < 1$; K un noyau et $(X_i, Y_i)_{i=1}^{i=n}$ un échantillon de taille n de la loi P_{XY} .

Revesz montre la convergence asymptotiquement normale de $\hat{r}_n(x)$. Un article de Wisotzki [37] nous semble intéressant d'être cité ; ce dernier estime la régression par un polynôme et montre que son estimateur est convergeant en moyenne quadratique intégrée. Enfin, une publication de Vitale et Pipkin [33] nous donne une caractérisation de la classe des fonctions de régression correspondant à la classe des lois P_{XY} dont les marges sont uniformes sur $[0,1]$.

Ces résultats généraux nous amènent à rappeler un point important relatif à l'

I.3. - *Inexistence d'un estimateur sans biais de la régression.*

Nous désignons par $\mathcal{D}$ l'ensemble des probabilités P_{XY} admettant une densité p par rapport à la mesure de Lebesgue sur $\mathbb{R}^{s+1}$, cette densité étant continue et telle que :

$$(\forall x \in \mathbb{R}^s) , \qquad f(x) = \int p(x,y)\, dy \neq 0 .$$

Proposition I.3.- Il n'existe pas d'estimateur non biaisé de la régression R, en ce sens qu'il n'existe pas d'estimateur $\hat{R}_n$ de R tel que : $(\forall x \in \mathbb{R}^s)$, $(\forall P_{XY} \in \mathcal{D})$, $E\,\hat{R}_n(x) = R(x)$.

Démonstration : Il suffit d'appliquer directement le théorème de Bickel-Lehmann ; on pourra se référer à Bosq [7] pour son énoncé et à Bosq [8] pour sa démonstration.

1.4. - *Estimation par les fonctions-spline.*

Toujours dans le souci de mieux situer notre travail, il est également utile de rappeler les récents travaux concernant l'estimation de la densité de probabilité par une spline quadratique ; on peut citer essentiellement un article de Wahba [34] ; partant d'une idée de Boneva, Kendall et Stephanov [6], elle pose comme estimateur de la densité $f(x)$, pour tout x fixé dans $[0,1]$, la fonction $\hat{f}_n(x)$, obtenue par dérivation de la spline cubique $\hat{F}_n(x)$, solution du problème suivant : étant donné l'espace de Sobolev $W_2^{(2)}[0,1]$, défini par :

$$W_2^{(2)}[0,1] = \{h,\ h^{(\mu)} \text{ abs.continues},\ \nu = 0,1,\ h^{(2)} \in L_2[0,1]\}.$$

On cherche l'unique fonction $\hat{F}_n(x)$ dans $W_2^{(2)}[0,1]$, telle que :

$$\text{a)}\quad \hat{F}_n(jh) = F_n(jh) \quad , \quad \forall j \ , \ j = 0,1,\ldots,\ell+1$$

$$\hat{F}_n'(0) = \hat{a}$$

$$\hat{F}_n'(1) = \hat{b}$$

où F_n est la fonction de répartition empirique associée à l'échantillon $(X_1,X_2,\ldots,X_n)$ de la v.a.r. X dont la densité de probabilité est f ; la suite (jh), $j = 0,1,\ldots, \ell+1$ une subdivision équidistante de $[0,1]$; $\ell = \ell(n)$ une suite d'entiers vérifiant

$$h(\ell+1) = 1 \quad \text{avec} \quad \lim_{n\to\infty} h = 0$$

et $\hat{a}$ et $\hat{b}$ des réels donnés.

b) $\hat{F}_n(x)$ minimise la quantité : $\int_0^1 [H''(t)]^2 \, dt$ parmi toutes les fonctions H de $W_2^{(2)}[0,1]$.

Wahba [34] montre que la spline quadratique $\hat{f}_n(x)$, x dans $[0,1]$, converge en moyenne quadratique vers $f(x)$; en outre, si f est dans $W_p^{(m)}[0,1]$, définie par :

$$W_p^{(m)}[0,1] = \{f,\ f^{(\nu)} \text{ abs. continues },\ \nu = 0,1,\ldots,m-1,\ f^{(m)} \in L_p[0,1]\}$$

elle montre que pour $m = 1,2,3$ et pour certaines valeurs de p, on a :

$$(\forall x \in [0,1]), \quad E[\hat{f}_n(x) - f(x)]^2 = O(n^{-\phi(m,p)})$$

où $$\phi(m,p) = \frac{2m - 2/p}{2m+1 - 2p} \quad .$$

Par ailleurs, en supposant que f appartient à l'espace :

$$W_p^{(m)}([0,1], M) = \{f,\ f \in W_p^{(m)}[0,1] \text{ et } ||f^{(n)}||_p \leq M\}$$

où $||.||$ désigne la norme dans L_p, elle obtient, grâce à un résultat de Farrell, la relation suivante :

$$(\forall x \in [0,1]), \quad \sup_{\substack{f \in W_p^{(n)}([0,1],M) \\ f \text{ densité}}} E_f [\hat{f}_n(x) - f(x)]^2 \leq D \cdot n^{-\phi(m,p)} \qquad (*)$$

où D désigne une constante indépendante de n, m, p et M, et ϕ la fonction définie plus haut.

Dans un autre article, Wahba [35] met l'accent sur une remarque assez intéressante : la relation (*) peut être généralisée aux estimateurs

non-paramétriques de la densité, en ce sens que si on suppose que f est dans $W_p^{(m)}([0,1], M)$ et si $\tilde{f}_n(x)$ désigne un estimateur de $f(x)$, x fixé dans $[0,1]$, obtenu par une des méthodes d'estimation non-paramétriques suivantes : méthode du noyau, lorsque ce dernier est bien choisi, méthode des fonctions orthogonales, l'histogramme ou, enfin, la méthode de l'algorithme polynomial, on obtient :

$$(\forall x \in [0,1]) \ , \quad \sup_{\substack{f \in W_p^{(m)}([0,1],M) \\ f \text{ densité}}} E_f\left[\tilde{f}_n(x) - f(x)\right]^2 \leq D\, n^{-\phi(m,p)}$$

Tout récemment, Quidel [23], en faisant de l'ajustement, suggère un estimateur $\psi_{h_n}(x)$ de $R(x)$, x dans $[0,1]$, basé sur les fonctions - spline et qui minimise la quantité

$$T(\psi_{h_n}) = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \psi_{h_n}(X_i))^2$$

où h_n est une suite strictement positive et de limite nulle - et (X_i, Y_i), $i = 1,2,\ldots,n$ un échantillon de taille n de (X,Y) ; Quidel montre que $\psi_{h_n}(x)$ converge presque - complètement uniformément vers $R(x)$ si la suite h_n vérifie :

$$\lim_{n\to\infty} \frac{n\, h_n}{\text{Log } n} = +\infty$$

CHAPITRE II

NOTATIONS, DEFINITIONS et HYPOTHESES GENERALES.

II.1. - Construction de l'estimateur-spline (régrospline).

Soit (X,Y) un couple aléatoire à valeurs dans $[0,1[\times \mathbb{R}$; on suppose que la loi P_{XY} du couple (X,Y) dans l'espace $([0,1[\times \mathbb{R}, B_{[0,1[\times\mathbb{R}})$ inconnue mais appartenant à un ensemble $\mathcal{D}$; on sait d'après un théorème de Jirina que Y admet une famille de lois conditionnelles $(P_Y^X)_{x\in[0,1[}$, bien définie modulo P_X ; on suppose que EY existe, auquel cas la régression de Y en X est une détermination de $E(Y/X)$.

Soit (X_i,Y_i), $i = 1,2,\ldots,n$ un échantillon de la loi P_{XY} et (Δ) la subdivision équidistante de $[0,1]$, définie comme suit :

$$\Delta : 0 = x_o < x_1 < \ldots < x_\ell < x_{\ell+1} = 1$$

où $x_j = jh$, $j = 0,1,\ldots,\ell+1$; avec $h = h(n)$ une suite strictement positive et de limite nulle et $\ell = \ell(n)$ une suite d'entiers vérifiant : $(\ell+1)h = 1$.

A partir des observations (X_i,Y_i), $i = 1,2,\ldots,n$ et la subdivision (Δ), on construit le régressogramme R_n de la façon suivante :

$(\forall\ j,\ j = 0,1,\ldots,\ell+1)$, $R_n(x_j)$ est constant et est égal à la moyenne arithmétique des observations Y_i dont les abcisses X_i appartiennent à l'intervalle $I_j = [x_{j-1},x_j[$, $j = 1,2,\ldots,\ell+1$. R_n vaut zéro si aucun X_i n'appartient à I_j ; autrement dit, si

on désigne par V_{nj} le nombre de points de l'échantillon $(X_1, X_2, \ldots, X_n)$ qui appartiennent à I_j et $\mathcal{I}_{nj}$ l'ensemble, aléatoire, de leurs indices, on a :

$$(\forall\, x \in I_j), \qquad R_n(x) = \frac{1}{V_{nj}} \sum_{i \in \mathcal{I}_{nj}} Y_i \qquad \text{si } V_{nj} \geq 1$$

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

on considère le problème suivant : "Chercher l'unique fonction G dans $C^2[0,1]$ qui vérifie le système suivant :

$$G(x_j) = R_n(x_j), \qquad j = 0,1,\ldots,\ell+1$$
$$G'(0) = \hat{a}$$
$$G'(1) = \hat{b}$$

et qui minimise la quantité $\int_0^1 [H''(t)]^2 dt$ parmi toutes les fonctions H de $C^2[0,1]$; $\hat{a}$ et $\hat{b}$ sont définis comme suit :

$$\hat{a} = \frac{R_n(x_1) - R_n(x_0)}{h}$$

$$\hat{b} = \frac{R_n(x_{\ell+1}) - R_n(x_\ell)}{h} \quad .$$

D'après Schoenberg [29], la solution de ce problème est une spline cubique aux joints (x_j), $j = 0,1,\ldots,\ell+1$ sur l'intervalle $[0,1]$; pour tout entier n, on appelera $\hat{R}_n$ la solution obtenue ; on sait que c'est une fonction de classe C^2 sur $[0,1]$, dont la restriction à I_j est un polynôme de degré 3 au plus ; on pose, pour tout j, $j = 0,1,\ldots,\ell+1$

$$N_j = \frac{d^2}{dx^2} \hat{R}_n(x) \Big|_{x=x_j} \quad .$$

Ces dernières quantités sont souvent appelés "moments" de la spline.

Expression de l'estimateur-spline $\hat{R}_n$ *en fonction de ses moments.*

$$(\forall\, x \in I_j), \qquad \hat{R}_n''(x) = N_{j-1}\,\frac{x_j - x}{h} + N_j\,\frac{x - x_{j-1}}{h} \quad .$$

En intégrant, il vient :

$$\hat{R}_n'(x) = -N_{j-1}\,\frac{(x_j-x)^2}{2h} + N_j\,\frac{(x-x_{j-1})^2}{2h} + A$$

$$\text{d'où :} \quad \hat{R}_n(x) = N_{j-1}\,\frac{(x_j-x)^3}{6h} + N_j\,\frac{(x-x_{j-1})^3}{6h} + Ax + B.$$

Pour déterminer les constantes A et B, on écrit que :

$$\hat{R}_n(x_j) = R_n(x_j)$$

$$\text{et} \qquad \hat{R}_n(x_{j-1}) = R_n(x_{j-1})$$

d'où :

$$(\forall\, x \in I_j), \quad \hat{R}_n(x) = N_{j-1}\,\frac{(x_j-x)^3}{6h} + N_j\,\frac{(x-x_{j-1})^3}{6h} + N_{j-1}\,\frac{h(x-x_j)}{6}$$

$$- N_j\,\frac{h(x-x_{j-1})}{6} + \frac{R_n(x_j)\times(x-x_{j-1}) - R_n(x_{j-1}) \times (x-x_j)}{h}$$

dès lors, on peut écrire :

$$(\forall\, x \in I_j), \quad \hat{R}_n'(x) = -N_{j-1}\,\frac{(x_j-x)^2}{2h} + N_j\,\frac{(x-x_{j-1})^2}{2h} + \frac{R_n(x_j) - R_n(x_{j-1})}{h}$$

$$- \frac{(N_j - N_{j-1})}{6}\,h.$$

En écrivant que :

$$\hat{R}_n'(x_j-) = \hat{R}_n'(x_j+), \qquad j = 1,2,\ldots,\ell$$

on obtient le système :

$$\frac{h}{6} N_{j-1} + \frac{2}{3} h N_j + \frac{h}{6} N_{j+1} = \frac{1}{h} \left[R_n(x_{j+1}) - 2R_n(x_j) + R_n(x_{j-1}) \right]$$

avec les deux conditions :

$$\hat{R}'_n(0) = \hat{a}$$
$$\hat{R}'_n(1) = \hat{b}$$

on obtient un système à $(\ell+2)$ équations à $(\ell+2)$ inconnues qui sont $(N_o, N_1, \ldots, N_{\ell+1})$. En mettant respectivement $x = 0$ et $x = 1$ dans l'expression ci-dessus de $\hat{R}'_n$ et compte tenu de la définition de $\hat{a}$ et $\hat{b}$, on obtient

$$2N_o + N_1 = 0$$
$$N_\ell + 2N_{\ell+1} = 0$$

et le système s'écrit :

$$\begin{bmatrix} 2 & 1 & & & & & \\ 1/2 & 2 & 1/2 & & & & \\ & 1/2 & 2 & & & & \\ & & & \ddots & & & \\ & & & & 2 & 1/2 & \\ & & & & 1/2 & 2 & 1/2 \\ & & & & & 1 & 2 \end{bmatrix} \begin{bmatrix} N_o \\ N_1 \\ \vdots \\ N_{\ell+1} \end{bmatrix} = \begin{bmatrix} d_o \\ d_1 \\ \vdots \\ d_{\ell+1} \end{bmatrix}$$

A

avec : $$d_j = \frac{6}{h}\left[R_n(x_{j+1}) - 2R_n(x_j) + R_n(x_{j-1})\right], \quad j = 1,2,\ldots,\ell$$

et $$d_o = d_{\ell+1} = 0.$$

Il suffit d'inverser la matrice A pour déterminer les moments N_j, $j = 0,1,\ldots,\ell+1$; aussi, si on désigne par $(A^{-1}_{i,j})$, $0 \leq i, j \leq \ell+1$, les éléments de la matrice inverse A^{-1}, il vient :

$$N_j = \sum_{k=0}^{\ell+1} A^{-1}_{j,k}\, d_k, \quad j = 0,1,\ldots,\ell+1.$$

D'après A.N.W. [1], page 38, on a, en posant $\sigma = \sqrt{3} - 2$

$$A^{-1}_{k,j} = \frac{\sigma^{j-k}(1+\sigma^{2k})(1-\sigma^{2(\sigma+1)-2j})}{(2+\sigma)(1-\sigma^{2(\ell+1)})}, \quad 0 < k \leq j \leq \ell+1$$

$$A^{-1}_{k,\ell+1} = \frac{\sigma^{\ell+1-k}(1+\sigma^{2k})}{(2+\sigma)(1-\sigma^{2(\ell+1)})}, \quad 0 < k \leq \ell+1$$

$$A^{-1}_{0,j} = \frac{2\sigma^{j}(1+\sigma^{2(\ell+1)-2j})}{(2+\sigma)(1-\sigma^{2(\ell+1)})} \quad 0 < j < \ell+1$$

$$A^{-1}_{0,\ell+1} = \frac{2\sigma^{\ell+1}}{(2+\sigma)(1-\sigma^{2(\ell+1)})}.$$

II.2. - Représentation de la régrospline par les B-splines.

Les B-splines, introduites par I.J. Schoenberg [30] ont été principalement étudiées par Deboor [12] et G. Cox [11] ; elles sont définies comme suit :

Soit $\pi = (t_i)_{i\in Z}$ une suite doublement infinie et croissante de nombres réels ; pour tout entier $m \geq 1$, on pose :

$$g_m(s,t) = (s-t)_+^{m-1} = \begin{cases} (s-t)^{m-1} & \text{si } s \geq t \\ 0 & \text{sinon.} \end{cases}$$

La B-spline $M_{i,m}(t)$ est définie comme différence divisée d'ordre m de $g_m(s,t)$ considérée comme fonction de s et basée sur les points $t_i, t_{i+1}, \ldots, t_{i+m}$, c'est-à-dire

$$M_{i,m}(t) = g_m(t_i, t_{i+1}, \ldots, t_{i+m}; t).$$

Si on note $S_k(\pi)$ l'espace vectoriel des fonctions-spline de degré k, ayant pour noeuds les points (t_i), $i \in \mathbb{Z}$ et définies sur R, toute spline S_k de cet espace peut s'écrire d'une manière unique de la façon suivante :

$$(\forall\, t \in \mathbb{R}), \qquad S_k(t) = \sum_{j \in \mathbb{Z}} A_j\, M_{k,k+1}(t)$$

où les (A_j) sont des constantes.

Aussi, notre estimateur peut s'écrire de manière unique

$$(\forall\, x \in [0,1]), \quad \hat{R}_n(x) = \sum_{j=0}^{\ell+1} C_j\, M_{j,4}(x)$$

où les $(c_j)_j$ sont des quantités qu'on détermine à partir des conditions d'interpolation.

II.3. - *Hypothèses générales.*

II.3.1. - *Hypothèses relatives à la loi* P_{XY}.

On supposera que la v.a.r X est à valeurs dans $[0,1[$ et que Y est bornée, sauf avis contraire ; on notera $p_j = P(X \in I_j)$, $j = 1, \ldots, \ell+1$ et $q_j = 1-p_j$; on supposera que $p_j > 0$ et on appelle $\mathcal{D}$ l'ensemble des probabilités P_{XY} admettant une densité p par rapport à la mesure de Lebesgue sur $[0,1[\times \mathbb{R}$ strictement positive.

II.3.2. - Hypothèses relatives à la régression.

On note par R la régression inconnue ; on supposera, sauf avis contraire, qu'elle est continue sur $[0,1[$; on notera $\omega_j(R,h)$ l'oscillation de R sur I_j et $\omega(R,h) = \max_{1\leq j\leq \ell+1} \omega_j(R,h)$.

On désigne par x un point fixé de $[0,1[$.

CHAPITRE III

ETUDE ASYMPTOTIQUE DU REGRESSOGRAMME EN x FIXÉ.

III.1. - Convergence asymptotiquement sans biais.

III.1.1. - Résultats Préliminaires.

Dans ce qui suit, nous donnons des résultats relatifs à l'espérance mathématique du régressogramme.

Proposition III.1.1.- Pour tout x fixé dans $[0,1[$, on a

$$E\, R_n(x) = E(Y/X \in I_j)(1 - q_j^n)$$

où I_j est l'intervalle contenant x.

Démonstration : On suivra une idée suggérée par Bosq.

Désignons par $[\nu_{n_j} = k]$, $k = 0,1,2,\ldots,n$, l'événement : "k points de l'échantillon $(X_1,X_2,\ldots,X_n)$ de la loi P_X appartiennent à l'intervalle I_j" ; il vient alors :

$$E\, R_n(x) = \sum_{k=0}^{n} \int_{[\nu_{n_j}=k]} R_n(x,\omega)\, dP(\omega)$$

or :

$$R_n(x) = \frac{1}{\nu_{n_j}} \sum_{X_i \in I_j} Y_i(\omega) \qquad \text{si} \quad \nu_{n_j} \geq 1$$

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

d'où :

$$E\, R_n(x) = \sum_{k=0}^{n} \frac{1}{k} \sum_{X_i \in I_j} \int_{[\nu_{n_j}=k]} Y_i(\omega)\, dP(\omega) \quad .$$

Par ailleurs,

$$\left[\nu_{n_j} = k\right] = \sum_{i_1,i_2,\ldots,i_k} \left[X_{i_1} \in I_j \; ; \; X_{i_2} \in I_j \; ; \ldots ; \; X_{i_k} \in I_j \; ; \; X_{i_{k+1}} \notin I_j \; ; \ldots ; \; X_{i_n} \notin I_j\right]$$

Aussi, il vient :

$$\int_{\left[\nu_{n_j}=k\right]} R_n(x,\omega)\, dP(\omega) =$$

$$\frac{1}{k} \sum_{i_1,\ldots,i_k} \int_{\left[X_{i_1} \in I_j ; \ldots ; X_{i_k} \in I_j ; X_{i_{k+1}} \notin I_j ; \ldots ; X_{i_n} \notin I_j\right]} \left(\sum_{\alpha=1}^{k} Y_{i_\alpha}(\omega)\right) dP(\omega)$$

$$\int_{\left[\nu_{n_j}=k\right]} R_n(x,\omega)\, dP(\omega) =$$

$$\frac{1}{k} \sum_{i_1,\ldots,i_k} \int \sum_{\alpha=1}^{k} Y_{i_\alpha}(\omega)\, 1_{\left[X_{i_1},\ldots,X_{i_k} \in I_j ; X_{i_{k+1}},\ldots,X_{i_n} \notin I_j\right]}(\omega)\, dP(\omega)$$

Pour tout k dans $\{1,2,\ldots,n\}$, posons :

$$A_k = \int Y_{i_\alpha}(\omega) \times 1_{\left[X_{i_1} \in I_j ; \ldots ; X_{i_k} \in I_j ; X_{i_{k+1}} \notin I_j,\ldots,X_{i_n} \notin I_j\right]}(\omega)\, dP(\omega) \; .$$

La loi du vecteur $\left[(X_1,Y_1),(X_2,Y_2),\ldots,(X_n,Y_n)\right]$ étant symétrique, A_k ne dépend ni des indices $i_1,i_2,\ldots,i_k$, ni de $\alpha = 1,2,\ldots,k$; par conséquent, cette intégrale vaut :

$$A_k = \int 1_{\left[X_1,X_2,\ldots,X_k \in I_j ; X_{k+1} \notin I_j,\ldots,X_n \notin I_j\right]}(\omega) \; . \; Y_1(\omega)\, dP(\omega)$$

d'où :

$$\int_{[\nu_{n_j}=k]} R_n(x,\omega)\, dP(\omega) =$$

$$C_n^k \int 1_{[X_1,X_2,\ldots,X_k \epsilon I_j ; X_{k+1},\ldots,X_n \notin I_j]}(\omega) \cdot Y_1(\omega)\, dP(\omega)$$

En regroupant tous les résultats, on obtient :

$$E\, R_n(x) = \sum_{k=0}^{n} \int_{[\nu_{n_j}=k]} R_n(x,\omega)\, dP(\omega) =$$

$$\sum_{k=1}^{n} C_k \int_{[X_1;\ldots;X_k \epsilon I_j ; X_{k+1},\ldots,X_n \notin I_j]} Y_1(\omega)\, dP(\omega) .$$

Comme les variables $(Y_1, X_1, X_2, \ldots, X_n)$ sont indépendantes, on a

$$\int_{[X_1,\ldots,X_k \epsilon I_j ; X_{k+1},\ldots,X_n \notin I_j]} Y_1(\omega)\, dP(\omega) = \left(\int_{[X_1 \epsilon I_j]} Y_1(\omega)\, dP(\omega)\right)$$

$$\times P(X \in I_j)^{k-1} \times P(X \notin I_j)^{n-k}$$

$$= E(Y/X \epsilon I_j)\, p_j^k \times q_j^{n-k}$$

et en définitive :

$$E\, R_n(n) = E(Y/X \epsilon I_j) \sum_{k=1}^{n} C_n^k\, p_j^k\, q_j^{n-k} = E(Y/X \epsilon I_j)(1-q_j^n)$$

Corollaire III.1.1.- Sous l'hypothèse suivante $(\exists A > 0)$, indépendant de h et de j, tel que

$$0 < Ah \leq p_j$$

on a : $$E\, R_n(x) = E(Y/X \epsilon I_j) + O(e^{-nh})$$

Démonstration : On sait que :

$$E\, R_n(x) = E(Y/X \in I_j) - E(Y/X \in I_j)\, q_j^n$$

comme : $$q_j^n = (1 - p_j)^n \leq (1 - Ah)^n$$

d'autre part Y étant supposée bornée, on a

$$q_j^n\, E(Y/X \in I_j) \leq \sup |Y| \times (1 - Ah)^n$$

comme $(1-Ah)^n$ est de l'ordre de e^{-nh} , on a :

$$E\, R_n(x) = E(Y/X \in I_j) + O(e^{-nh}) .$$

III.1.2. - Limite et majoration asymptotique du biais (en valeur absolue).

Le comportement asymptotique du biais du régressogramme est donné par la :

Proposition III.1.2.- Sous les hypothèses suivantes :

1) $(\exists A > 0)$, indépendant de h et de j, tel que : $0 < Ah \leq p_j$.

2) X admet une densité f par rapport à la mesure de Lebesgue avec $f(x) > 0$.

3) $\lim_{n\to\infty} nh = +\infty$

alors : $$\lim_{n\to\infty} \left[E\, R_n(x) - R(x)\right] = 0 .$$

Démonstration : Posons : $\delta_j = E(Y/X \in I_j) - R(x)$

l'hypothèse 2) nous permet d'écrire que :

$$\delta_j = \frac{\int_{I_j} R(y)\, f(y)\, dy}{\int_{I_j} f(y)\, dy} - \frac{R(x) \int_{I_j} f(y)\, dy}{\int_{I_j} f(y)\, dy}$$

$$\delta_j = \frac{1}{p_j} \left[\int_{I_j} \{R(y) - R(x)\}\, f(y)\, dy\right.$$

d'où $|\delta_j| \leq \omega_j(R,h)$.

La continuité de R implique que $\lim_{n\to\infty} \omega_j(R,h) = 0$; le corollaire III.1.1. et les hypothèses 1) et 2) nous permettent de conclure.

Remarque : On peut montrer que la continuité de R implique l'existence d'un nombre ξ_j dans I_j tel que :

$$E(Y/X \in I_j) = R(\xi_j) .$$

Corollaire III.1.2.- Sous les hypothèses de la proposition III.1.2., si de plus R est lipschitzienne et si h est de la forme $\frac{1}{n^\alpha}$ $0 < \alpha < 1$, alors on a :

$$E\, R_n(x) - R(x) = O(h) .$$

Démonstration : Lorsque R est lipschitzienne

$$\omega_j(R,h) = O(h) .$$

La condition $h = \frac{1}{n^\alpha}$, $0 < \alpha < 1$, implique que le terme en $O(e^{-nh})$ est absorbé par celui de $O(h)$; la proposition précédente nous permet

de conclure ; les suites h qui vérifient $\lim_{n\to\infty} \frac{e^{-nh}}{h} = +\infty$ donnent le même résultat.

III.2. - Majoration asymptotique de la variance de $R_n(x)$.

Proposition III.2.1.-

a) $$\text{Var } R_n(x) \leq 2 \frac{\text{Var}(Y/X\epsilon I_j)}{(n+1)p_j} (1-q_j^{n+1}) - 2 \text{ Var}(Y/X\epsilon I_j)\, q_j^n + M^2 q_j^n(1-q_j^n)$$

b) $$\text{Var } R_n(x) \geq 2^{-1} \frac{\text{Var}(Y/X\epsilon I_j)}{(n+1)p_j} (1-q_j^{n+1}) - 2 \text{ Var}(Y/X\epsilon I_j)\, q_j^n + M^2 q_j^n(1-q_j^n)$$

où $$M = \sup |Y|$$

Démonstration :

a) $$\text{Var } R_n(x) = \sum_{k=0}^{n} \text{Var}\left[R_n(x)/_{\nu_{n_j}=k}\right] \times P\left[\nu_{n_j}=k\right] .$$

En utilisant la définition du régressogramme, on a :

$$E\left[R_n^2(x)/\nu_{n_j}=k\right] = \frac{1}{P\left[\nu_{n_j}=k\right]} \int_{\left[\nu_{n_j}=k\right]} R_n^2(x,\omega)\, dP(\omega)$$

$$= \frac{1}{P\left[\nu_{n_j}=k\right]} \sum_{i_1,\ldots,i_k} \int_{\left[X_{i_1}\epsilon I_j,\ldots,X_{i_k}\epsilon I_j;X_{i_{k+1}},\ldots,X_{i_n}\notin I_j\right]} R_n^2(x,\omega)\, dP(\omega)$$

or :

$$\int_{\left[\nu_{n_j}=k\right]} R_n^2(x,\omega)\, dP(\omega) = C_n^k \int_{\left[X_1,\ldots,X_k\epsilon I_j;X_{k+1},\ldots,X_n\notin I_j\right]} \left(\frac{1}{k}\sum_{i=1}^{k} Y_i(\omega)\right)^2 dP(\omega)$$

Posons pour tout entier k appartenant à $\{1,2,\ldots,n\}$ et pour tout j dans $\{1,2,\ldots,\ell+1\}$:

$$A_{k,j} = [X_1 \in I_j,\ldots,X_k \in I_j \; ; \; X_{k+1} \notin I_j,\ldots,X_n \notin I_j]$$

Il vient alors :

$$\int_{[\nu_{n_j}=k]} R_n^2(x,\omega)dP(\omega) = \frac{1}{k^2} C_n^k \sum_{i=1}^{k} \int_{A_{k,j}} Y_i^2(\omega)\, dP(\omega) +$$

$$+ \frac{1}{k^2} C_n^k \sum_{i\neq i'} \int_{A_{k,j}} Y_i(\omega)\, Y_{i'}(\omega)\, dP(\omega)$$

$$= \frac{1}{k} C_n^k E(Y^2/X\in I_j)\, p_j^k\, q_j^{n-k} + \frac{k-1}{k} C^k\, E^2(Y/X\in I_j) p_j^n q_j^{n-1}$$

d'où :

$$E\, R_n^2(x) = E(Y^2/X\in I_j) \sum_{k=1}^{n} \frac{1}{k} C_n^k\, p_j^k\, q_j^{n-k} +$$

$$+ E^2(Y/X\in I_j) \sum_{k=1}^{n} (1 - \frac{1}{k}) C_n^k\, p_j^k\, q_j^{n-k}$$

$$= Var(Y/X\in I_j) \sum_{k=1}^{n} \frac{1}{k} C_n^k\, p_j^k\, q_j^{n-k} + E^2(Y/x\in I_j)\, (1-q_j^n) .$$

Par ailleurs, on a d'après la proposition III.1.1. que :

$$E^2\, R_n(x) = (1 - q_j^n)^2\, E^2(Y/X\in I_j)$$

d'où :

$$Var\, R_n(x) = Var(Y/X\in I_j) \sum_{k=1}^{n} \frac{1}{k} C_n^k\, p_j^k\, q_j^{n-k} + E^2(Y/X\in I_j)\, a_j(1-a_j)$$

où : $$a_j = 1 - q_j^n .$$

En remarquant que : $C_n^k = \frac{k+1}{n+1} C_{n+1}^{k+1}$, on a :

$$\gamma_j = \sum_{k=1}^{n} \frac{1}{k} C_n^k p_j^k q_j^{n-k} = \frac{1}{n+1} \sum_{k=1}^{n} \frac{k+1}{k} C_{n+1}^{k+1} p_j^k q_j^{n-k}$$

Comme $\forall k$ dans $\{1,2,\ldots,n\}$, on a : $\frac{k+1}{k} \leq 2$, il vient :

$$\gamma_j \leq \frac{2}{(n+1)p_j} \left[\sum_{\beta=2}^{n+1} C_{n+1}^{\beta} p_j^{\beta} q_j^{n+1-\beta} \right]$$

$$\gamma_j \leq \frac{2}{(n+1)p_j} (1 - q_j^{n+1}) - 2 q_j^n$$

d'où la majoration :

$$\text{Var } R_n(x) \leq 2 \frac{\text{Var}(Y/x \in I_j)}{(n+1)\ p_j} (1 - q_j^{n+1}) - 2\ \text{Var}(Y/X \in I_j)\ q_j^n + M^2(1-q_j^n)\ q_j^n .$$

b) γ_j peut se minorer de la façon suivante :

$$\gamma_j \geq \frac{1}{2(n+1)p_j} \sum_{k=1}^{n} C_{n+1}^{k+1} p_j^k q_j^{n-k}$$

$$\gamma_j \geq \frac{2^{-1}}{(n+1)p_j} (1 - q_j^{n+1}) - 2 q_j^n$$

et $\text{Var } R_n(x) \geq 2^{-1} \frac{\text{Var}(Y/X \in I_j)}{(n+1)p_j} (1 - q_j^{n+1}) - 2\ \text{Var}(Y/X \in I_j)\ q_j^n + M^2(1-q_j^n)\ q_j^n .$

Corollaire III.2.1.- Sous les hypothèses suivantes :

1) $(\forall P_{XY} \in \mathcal{D})$, $(\exists A > 0)$, indépendant de h et de j tel que :

$$0 < Ah \leq p_j$$

2) $h = \frac{1}{n^{\alpha}}$, $0 < \alpha < 1$

alors $$\text{Var } R_n(x) = O(\frac{1}{nh}) .$$

Démonstration : En utilisant la proposition précédente, on a :

$$\text{Var } R_n(x) \leq \frac{2 \text{ Var}(Y/X\epsilon I_j)}{(n+1)\ p_j} (1-q_j^{n+1}) - 2 \text{ Var}(Y/X\epsilon I_j)\ q_j^n + M^2(1-q_j^n)q_j^n$$

Y étant bornée, $\text{Var}(Y/X\epsilon I_j)$ l'est aussi. p_j étant minorée par Ah, on a : $q_j^n = O(e^{-nh})$.

Les termes en e^{-nh} étant absorbés par ceux en $\frac{1}{nh}$. Grâce à l'hypothèse 2) et en regroupant tous les calculs, on obtient le résultat.

Remarque 1 :

Sabry [26] , sous les hypothèses suivantes :

i) $\exists (\alpha,\beta) \in \mathbb{R}^+ \times \mathbb{R}$ tels que, $\forall (x,y) \in \mathbb{R}^2$, $|y| \geq |x|$,

on a : $$d[R(x),R(y)] \leq \alpha\ e^{\beta|y|}\ d(x,y)$$

où d désigne la distance euclidienne usuelle.

ii) $(\exists M > 0)$, $|R(0)| \leq M$

iii) $(\exists V > 0)$, $\text{Var}(Y/X\epsilon I_j) \leq V$

obtient que pour tout x fixé dans I_j .

$$\text{Var } R_n(x) \leq 2\frac{V}{np_j} + 2\ M^2(1-p_j)^n + 2\ \alpha^2\ x_j\ (1-p_j)^n\ e^{2\beta x_j} .$$

Remarque 2 :

Collomb [10] montre que s'il existe une détermination v de la variance conditionnelle de Y par rapport à X, i.e., de

$$E((Y - E(Y/X))^2/X)$$

qui est continue au point x, alors lorsque R est continue en x et n tend vers l'infini

$$\text{Var } R_n(x) \sim \frac{1}{nh_n} \frac{v(x)}{f(x)}$$

f étant la densité de X.

III.3. - Non-corrélation asymptotique de $R_n(x)$ *et* $R_n(x')$, $x \neq x'$ *fixés.*

La non-corrélation de $R_n(x)$ et $R_n(x')$ est assurée lorsque l'hypothèse du corollaire III.2.1. est vérifiée ; en effet, l'inégalité de Schwartz nous permet de l'affirmer ; dans la proposition qui suit, on donne la vitesse de convergence relative à la non-corrélation de $R_n(x)$ et $R_n(x')$.

Proposition III.3.1.- Sous l'hypothèse suivante :

$(\forall P_{XY} \in \mathcal{D})$, $(\exists A > 0)$, indépendant de h et de j tel que

$$0 < Ah \leq p_j$$

alors : pour x et x' fixés dans $[0,1[$, avec $x \neq x'$, on a :

$$\text{cov}[R_n(x), R_n(x')] = O(e^{-nh})$$

Démonstration : x et x' étant fixés dans $[0,1[$ et étant supposés distincts, il va exister k et k' dans $\{1,2,\ldots,\ell+1\}$, $k \neq k'$ tels que $x \in I_k$ et $x' \in I_{k'}$; en outre pour n assez grand, on a :

$$I_k \cap I_{k'} = \emptyset$$

Par ailleurs, on peut écrire :

$$\operatorname{cov}\left[R_n(x),\ R_n(x')\right] = E\ R_n(x)\ .\ R_n(x') - E\ R_n(x)\ .\ E\ R_n(x')\ .$$

a) Evaluation asymptotique de $E\ R_n(x)\ .\ R_n(x')$.

$$E\ R_n(x)\ .\ R_n(x') =$$

$$\int R_n(x,\omega)\ .\ R_n(x',\omega)\ dP(\omega)$$

$$= \sum_{j=0}^{n} \int_{\left[\nu_{n_j}=k\right]} R_n(x,\omega)\ .\ R_n(x',\omega)\ dP(\omega)$$

$$= \sum_{j=1}^{n} \sum_{i_1,i_2,\ldots,i_k} \int_{\left[X_{i_1}\in I_k,\ldots,X_{i_j}\in I_k;X_{i_{j+1}},\ldots,X_{i_n}\notin I_k\right]} \frac{1}{j} \sum_{\alpha=1}^{j} Y_{i_\alpha}(\omega).R_n(x',\omega)\ dP(\omega)$$

$$= \sum_{j=1}^{n} C_n^j \int_{\left[X_1,\ldots,X_j\in I_k;X_{j+1},\ldots,X_n\notin I_k\right]} Y_1(\omega)\ R_n(x',\omega)\ dP(\omega)$$

$$= \sum_{j=1}^{n} C_n^j \int 1_{\{X_1,\ldots,X_j\in I_k;X_{j+1},\ldots,X_n\notin I_k\}}(\omega)\ Y_1(\omega).R_n(x',\omega)\ dP(\omega)$$

$$= \sum_{j=1}^{n} C_n^j \sum_{j'=0}^{n-j} \sum_{i'_1,\ldots,i'_j} \int_{\left[X_{i'_1},\ldots,X_{i'_j}\in I_{k'};X_{i'_{j'+1}},\ldots,X_{i'_{n-j}}\notin I_{k'}\right]}$$

$$\frac{1}{j'} \sum_{\alpha=1}^{j'} Y_{i'_\alpha}(\omega)\ 1_{\{X_1,\ldots,X_j\in I_k;X_{j+1},\ldots,X_n\notin I_k\}}(\omega)\ Y_1(\omega)\ dP(\omega)$$

$$= \sum_{j=1}^{n} C_n^j \sum_{j'=0}^{n-j} C_{j'}^{n-j} \int Y_{j+1}(\omega)\ .\ Y_1(\omega)\ 1_{A_k}(\omega)\ .\ 1_{A_{k'}}(\omega)\ dP(\omega)$$

avec $\quad A_k = \left[X_1 \in I_j\ ,\ X_2 \in I_j\ ,\ldots\ldots,X_j\in I_k;X_{j+1},\ldots,X_n\notin I_j\right]$

$$A_{k'} = \left[X_{j+1} \in I_{k'}\ ;\ldots,\ X_{j'}\in I_{k'};X_{j'+1}\notin I_k,\ldots,X_{n-j}\notin I_{k'}\right]$$

d'où :

$$E\, R_n(x) \times R_n(x') = \sum_{j=1}^{n} C_n^j\, p_k^j\, (1 - p_k)^{n-j} \left[1 - (1 - p_{k'})^{n-j}\right] E_k \cdot E_{k'}$$

où $E_s = E(Y/X \in I_s), \quad s = k, k'$

$$E\, R_n(x) \cdot R_n(x') = E_k \cdot E_{k'} \left[1 - (1 - p_k)^n\right] - E_k \cdot E_{k'} \{\left[p_k + (1 - p_k)(1 - p_{k'})\right]^n - (1 - p_k)^n (1 - p_k)^n\} .$$

b) *Evaluation asymptotique de* $E\, R_n(x) \cdot E\, R_n(x')$.

D'après la proposition III.1.1., on a

$$E\, R_n(x) = E_k \left[1 - (1 - p_k)^n\right]$$

$$E\, R_n(x') = E_{k'} \left[1 - (1 - p_{k'})^n\right]$$

d'où : $E\, R_n(x)\, E\, R_n(x') = E_k\, E_{k'} \left[1 - (1 - p_k)^n\right] \times \left[1 - (1 - p_{k'})^n\right]$

En regroupant les calculs, il vient :

$$\operatorname{cov}\left[R_n(x), R_n(x')\right] = E_k \cdot E_{k'} (1 - p_{k'})^n - E_k\, E_{k'} \left[p_k + (1 - p_k)(1 - p_{k'})\right]^n$$

Sous notre hypothèse, on a :

$$(1 - p_{k'})^n = O(e^{-nh})$$

et

$$\left[p_k + (1 - p_k)(1 - p_{k'})\right]^n = \left[1 + p_k\, p_{k'} - p_{k'}\right]^n = O(e^{-nh}) .$$

Comme Y est bornée, E_k et $E_{k'}$ le sont aussi ; d'où le résultat.

III.4. - Convergence en moyenne quadratique de $R_n(x)$, x *fixé.*

La limite de l'erreur quadratique du régressogramme est donnée par la :

Proposition III.4.1.- Sous les hypothèses des corollaires III.1.2. et III.2.1. et si $\lim_{n\to\infty} nh = \infty$, on a :

$$\lim_{n\to\infty} E\left[R_n(x) - R(x)\right] = 0 \ .$$

Démonstration : Il suffit d'écrire que :

$$E\left[R_n(x) - R(x)\right]^2 = \left[E\ R_n(x) - R(x)\right]^2 + \text{Var}\ R_n(x) \ .$$

Les corollaires III.1.2. et III.2.1. nous permettent d'écrire :

$$E\left[R_n(x) - R(x)\right]^2 = O(h^2) + O(\frac{1}{nh})$$

comme $\lim_{n} nh = +\infty$, on obtient le résultat.

Corollaire III.4.1.- Sous les hypothèses de la proposition précédente et pour $h = O(n^{-1/3})$, on a

$$E\left[R_n(x) - R(x)\right]^2 = O(n^{-2/3})$$

Démonstration : Il suffit de voir que $h = O(n^{-1/3})$ minimise la fonction $O(h^2) + O(\frac{1}{nh})$.

Remarque : Collomb [10] obtient un résultat analogue en supposant que R est dérivable - voir page 24, proposition 3.1 de [10].

III.5. - Convergence en moyenne quadratique intégrée du régressogramme.

III.5.1. - Introduction.

Le "mise" a été introduit par Parzen [22] ; ce mode de convergence est utilisé essentiellement dans l'estimation non-paramétrique de la densité ; il permet notamment de comparer l'efficacité des diverses méthodes d'estimation.

Pour un paramètre fonctionnel g quelconque et son estimateur g_n , $n \in \mathbb{N}$, tous deux supposés de carré intégrable, le "mise" est défini par :

$$E \int_{\mathbb{R}} \left[g_n(x) - g(x)\right]^2 dx$$

III.5.2. - Majoration asymptotique du "mise" du régressogramme.

Elle est donnée par la :

Proposition III.5.2.- Sous l'hypothèse suivante :

$(\forall P_{XY} \in \mathcal{D})$, $(\exists A > 0)$, indépendant de h et de j, tel que

$$0 < Ah \leq p_j$$

on a : $$E \int_0^1 \left[R_n(x) - R(x)\right]^2 dx \leq K_1 h^2 + K_2 \frac{1}{nh} + K_3 e^{-2nh}$$

où K_1 , K_2 , K_3 sont des constantes indépendantes de n, h et j.

Démonstration : Pour x fixé dans I_j , on a :

$$\left[R_n(x) - R(x)\right]^2 \leq 2\left[R_n(x) - E\, R_n(x)\right]^2 + 2\left[E\, R_n(x) - R(x)\right]^2$$

$$\sup_{x \in I_j} \left[R_n(x) - R(x)\right]^2 \leq 2\left[R_n(x) - E\, R_n(x)\right]^2 + \sup_{x \in I_j} \left[E\, R_n(x) - R(x)\right]^2$$

car $\left[R_n(x) - E\, R_n(x)\right]$ est constant sur I_j .

$$\sup_{x\in I_j} \left[E\, R_n(x) - R(x)\right]^2 \leq 2 \sup_{x\in I_j} \left[E\, R_n(x) - E(Y/X\, I_j)\right]^2 + 2 \sup_{x\in I_j} \left[E(Y/X\in I_j) - R(x)\right]^2$$

$$\sup_{x\in I_j} \left[E\, R_n(x) - R(x)\right]^2 \leq 2 \sup_{x\in I_j} \left[E\, R_n(x) - E(Y/X\, I_j)\right]^2 + 2\, \omega_j^2(R,h)$$

Y étant bornée, d'après le corollaire III.1.1., il va exister une constante $K > 0$, indépendante de h et de j telle que

$$\left[E\, R_n(x) - E(Y/X\in I_j)\right] \leq K\, e^{-nh}$$

d'où :

$$\sup_{x\in I_j} \left[R_n(x) - R(x)\right]^2 \leq 2\left[R_n(x) - E\, R_n(x)\right]^2 + 4\, K^2\, e^{-2nh} + 4\, \omega^2(R,h)$$

en intégrant par rapport à x, il vient :

$$\int_{I_j} \left[R_n(x) - R(x)\right]^2 dx \leq 2h \left[R_n(\xi_j) - E\, R_n(\xi_j)\right]^2 + 4\, K^2\, h\, e^{-2nh} + 4\, h\, \omega^2(R,h)$$

où ξ_j est un point de I_j et en intégrant par rapport à la probabilité P, on a :

$$E \int_{I_j} \left[R_n(x) - R(x)\right]^2 dx \leq 2h\, \mathrm{Var}\, R_n(\xi_j) + 4h\, K^2\, e^{-2nh} + 4h\, \omega^2(R,h)$$

Or d'après l'étude asymptotique de la variance du régressogramme, il va exister une constante $V > 0$, indépendante de x, h et j telle que :

$$\mathrm{Var}\, R_n(\xi_j) \leq \frac{V}{nh}\ .$$

Aussi, on a :

$$E \int_{I_j} \left[R_n(x) - R(x)\right]^2 dx \leq 2 \frac{V}{n} + 4 K^2 h e^{-2nh} + 4h \omega^2(R,h)$$

d'où : $$E \int_0^1 \left[R_n(x) - R(x)\right]^2 dx \leq \sum_{j=1}^{\ell+1} \left[\frac{2V}{n} + 4 K^2 h e^{-2nh} + 4h \omega^2(R,h)\right]$$

ou bien :

$$E \int_0^1 \left[R_n(x) - R(x)\right]^2 dx \leq 2 \frac{V}{nh} + 4 K^2 e^{-2nh} + 4 \omega^2(R,h)$$

d'où le résultat.

III.5.3. - Limite du "mise" et vitesse de convergence.

Proposition III.5.3.- Sous l'hypothèse de la proposition précédente, alors si la suite h vérife la condition : $\lim_{n\to\infty} nh = +\infty$.

On a : $$\lim_{n\to\infty} E \int_0^1 \left[R_n(x) - R(x)\right]^2 dx = 0$$

Démonstration : En utilisant la majoration asymptotique de la proposition précédente, on a le résultat.

Corollaire III.5.3.- Sous l'hypothèse de la proposition III.5.3. et si de plus R est lipschitzienne, on a :

$$E \int_0^1 \left[R_n(x) - R(x)\right]^2 dx = O(n^{-2/3})$$

pour $$h = O(n^{-1/3}) .$$

Démonstration : Lorsque R est lipschitzienne, $\omega(R,h) = O(h)$. Sachant que le terme en e^{-2nh} de la majoration ci-dessus est absorbé par celui en $\frac{1}{nh}$, on a :

$$E \int_0^1 \left[R_n(x) - R(x)\right]^2 dx = O(\frac{1}{nh}) + O(h^2) .$$

En appliquant le lemme de Parzen [22], on a le résultat.

CHAPITRE IV

ETUDE ASYMPTOTIQUE DE $\hat{R}_n(x)$, x FIXÉ.

IV.1. - Comportement asymptotique du biais de $\hat{R}_n(x)$.

IV.1.1. - Résultats préliminaires.

$$E\,\hat{R}_n(x) = E\,N_{j-1}\,\frac{(x_j-x)^3 - h^2(x_j-x)}{6h} + E\,N_j\,\frac{(x-x_{j-1})^3 - h^2(x-x_{j-1})}{6h}$$

$$+ E\,R_n(x_j)\,\frac{(x-x_{j-1})}{h} - E\,R_n(x_{j-1})\,\frac{(x-x_j)}{h}$$

Proposition IV.1.1.-

1) L'espérance de l'estimateur spline $\hat{R}_n(x)$ est aussi une spline cubique qui interpole la suite $E\,R_n(x_j)$, $j = 0,1,\ldots,\ell+1$, aux joints $(x_j)_j$; la dérivée de l'espérance de $\hat{R}_n(x)$ est égale à l'espérance de la dérivée du même estimateur.

$$2)\quad \max_{0\leq j\leq \ell+1} |E\,N_j| \leq \max_{0\leq j\leq \ell+1} |E\,d_j|$$

avec :

$$1 \leq j \leq \ell\ ,\quad d_j = \frac{3}{h}\left[\frac{R_n(x_{j+1}) - R_n(x_j)}{h} - \frac{R_n(x_j) - R_n(x_{j-1})}{h}\right]$$

$$d_o = d_{\ell+1} = 0$$

Démonstration :

1) à partir de son expression, $E\,\hat{R}_n(x)$ est une spline cubique qui vérifie le système :

$$E\,\hat{R}_n(x_j) = E\,R_n(x_j)\ , \qquad j = 0,1,\ldots,\ell+1$$

$$E\,\hat{R}_n'(0) = E\,\hat{a}$$

$$E\,\hat{R}_n'(1) = E\,\hat{b}$$

avec : $$\hat{a} = \frac{R_n(h) - R_n(0)}{h} \quad ; \quad \hat{b} = \frac{R_n(1) - R_n(\ell h)}{h}$$

En dérivant $\hat{R}_n(x)$ puis en intégrant par rapport à la probabilité P, on a : $E \frac{d}{dx}\,\hat{R}_n(x) = \frac{d}{dx}\,E\,\hat{R}_n(x)$.

2) En écrivant que :

$$1 \leq j \leq \ell\ , \quad E\,\hat{R}_n'(x_j-) = E\,\hat{R}_n'(x_j+)$$

$$E\,\hat{R}_n'(0) = E\,\hat{a}$$

$$E\,\hat{R}_n'(1) = E\,\hat{b}$$

On obtient conformément à A.N.W. [1] le système $1 \leq j \leq \ell$:

$$\frac{h}{6}\,E\,N_{j-1} + \frac{2}{3}\,h\,E\,N_j + \frac{h}{6}\,E\,N_{j+1} = \frac{E\,R_n(x_{j+1}) - E\,R_n(x_j)}{h}$$

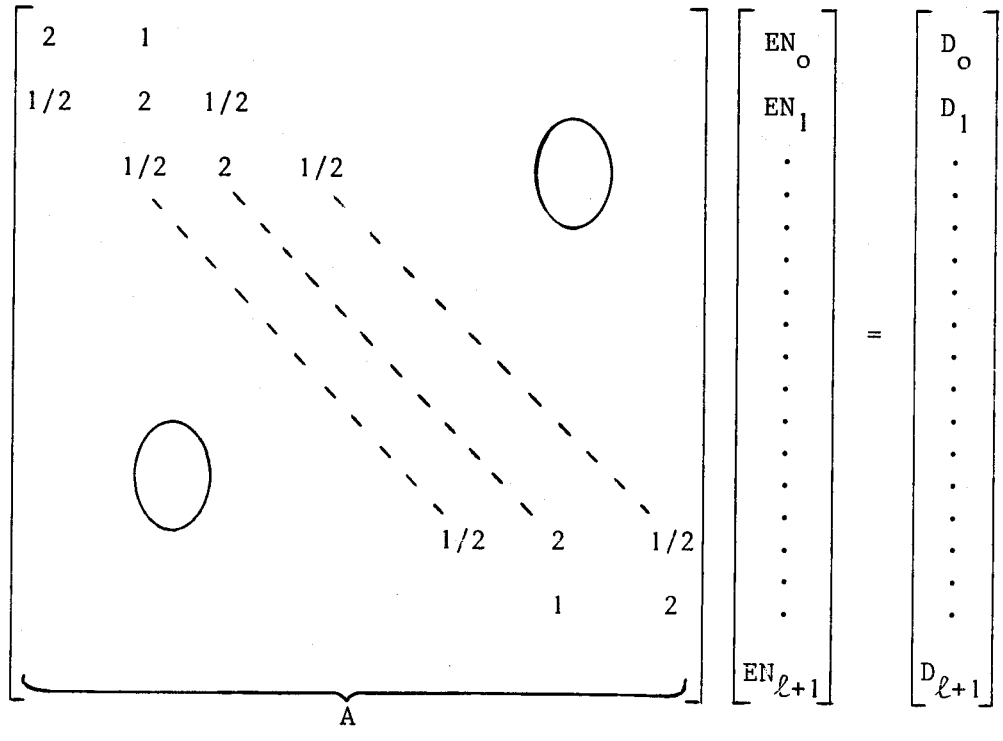
$$2\,E\,N_o + 2\,E\,N_1 = \frac{6}{h}\left[\frac{E\,R_n(x_1) - E\,R_n(x_o)}{h} - E\,\hat{a}\right] = 0$$

$$E\,N_\ell + 2\,E\,N_{\ell+1} = \frac{6}{h}\left[E\,\hat{b} - \frac{E\,R_n(x_{\ell+1}) - E\,R_n(x_\ell)}{h}\right] = 0$$

Posons :

$$(1 \leq j \leq \ell)\ , \quad D_j = \frac{3}{h}\left[\frac{E\,R_n(x_{j+1}) - E\,R_n(x_j)}{h} - \frac{E\,R_n(x_j) - E\,R_n(x_{j-1})}{h}\right]$$

et $$D_o = D_{\ell+1} = 0\ .$$



La matrice A est à diagonale dominante ; pour $\theta = (\theta_o, \theta_1, \ldots, \theta_{\ell+1}) \in \mathbb{R}^{\ell+2}$, posons

$$||\theta|| = \max_{0 \leq i \leq \ell+1} |\theta_i|$$

et $A = (a_{ij})$ $0 \leq i \leq \ell+1$; $0 \leq j \leq \ell+1$

on a : $$|a_{ii}| = 2 > \sum_{j \neq i} |a_{ij}|$$

Soit $t = (t_o, t_1, \ldots, t_{\ell+1}) \in \mathbb{R}^{\ell+2}$ tel que :

$$||t|| = |t|_k \quad , \qquad 0 \leq k \leq \ell+1$$

et posons : $z = A \cdot t$.

Aussi, on peut écrire :

$$||z|| = \max_i \left| \sum_{j=0}^{\ell+1} a_{ij}\, t_j \right|$$

$$||z|| \geq |a_{kk}\, t_k| - \left(\sum_{j=0}^{\ell+1} |a_{kj}| \; ||t|| \right)$$

ou bien $||At|| \geq \left(a_{kk} - \sum_{\substack{j=0 \\ j\neq k}}^{\ell+1} |a_{kj}|\right) ||t||$

d'où $$||A^{-1}|| = \sup_{z\neq 0} \frac{||A^{-1} z||}{||z||} \leq \frac{1}{\min_i \left(|a_{ii} - \sum_{j\neq i} |a_{ij}|\right)}$$

i.e. $||A^{-1}|| \leq 1$

d'où le résultat .

IV.1.2. - Majoration asymptotique du biais (en valeur absolue).

Pour tout x fixé dans I_j , on a :

$$E\,\hat{R}_n(x) - R(x) = E\,N_{j-1} \frac{(x_j - x)^3 - h^2(x_j - x)}{6h} + E\,N_j \frac{(x - x_{j-1})^3 - h^2(x - x_{j-1})}{6h}$$

$$+ \left[E\,R_n(x_{j-1}) - R(x)\right] \frac{(x - x_{j-1})}{h} - \frac{(x - x_j)}{h} \left[E\,R_n(x_{j-1}) - R(x)\right]$$

Proposition IV.1.2.- Sous les hypothèses suivantes :

1) $(\forall P_{XY} \in \mathcal{D})$, $(\exists A > 0)$ indépendant de h et de j, tel que :

$$0 < Ah \leq p_j$$

2) X admet une densité f par rapport à la mesure de Lebesgue avec $f(x) > 0$.

Alors on a :

$$|E\,\hat{R}_n(x) - R(x)| \leq (2 + \frac{2}{\sqrt{3}})\,\omega(R,h) + (\frac{4}{3\sqrt{3}} + 2)\,K\,e^{-nh}$$

où K est une constante positive, indépendante de h et de j.

Démonstration : Les coefficients de $E\,N_j$ et $E\,N_{j-1}$ sont majorés en valeur absolue par $\frac{h^2}{9\sqrt{3}}$.

D'après la proposition préliminaire précédente, on a besoin de majorer $|E\,d_j|$ pour avoir une majoration de $\max_{0\leq j\leq \ell+1} |E\,N_j|$

Si $1 \leq j \leq \ell$,

$$E\,d_j = \frac{3}{h}\left[\frac{E\,R_n(x_{j+1}) - E\,R_n(x_j)}{h} - \frac{E\,R_n(x_j) - E\,R_n(x_{j-1})}{h}\right]$$

$$|E\,d_j| \leq \frac{3}{h^2}\,[|E\,R_n(x_{j+1}) - R(x_{j+1})| + |R(x_{j+1}) - R(x_j)| +$$

$$2\,|R(x_j) - E\,R_n(x_j)| + |R(x_j) - R(x_{j-1})| +$$

$$|R(x_{j-1}) - E\,R_n(x_{j-1})|]$$

D'après l'étude asymptotique du biais du régressogramme, corollaire III.1.1., $(\exists K > 0)$, indépendante de h et de j telle que :

$$|E\,R_n(x) - E(Y/X\varepsilon I_j)| \leq K\,e^{-nh}$$

aussi, il vient :

$$|E\,R_n(x) - R(x)| \leq |E\,R_n(x) - E(Y/X\varepsilon I_j)| + |E(Y/X\,I_j) - R(x)|$$

$$|E\,R_n(x) - R(x)| \leq \omega_j(R,h) + K(e^{-nh})$$

$$\leq \omega(R,h) + K\,e^{-nh}\,,$$

c'est-à-dire $|E\,d_j| \leq \frac{3}{h^2}\,[6\,\omega(R,h) + 4\,K\,e^{-nh}]$

Aussi :

$$\left| E\,\hat{R}_n(x) - R(x) \right| \leq \left(2 + \frac{2}{\sqrt{3}}\right) \omega(R,h) + \left(\frac{4}{3\sqrt{3}} + 2\right) K\, e^{-nh}$$

IV.1.3. - Convergence asymptotiquement sans biais.

Proposition IV.1.3.- Sous les mêmes hypothèses de la proposition précédente alors :

$$\lim_{n\to\infty} \left[E\,\hat{R}_n(x) - R(x)\right] = 0$$

si $\lim_{n\to\infty} nh = +\infty$

Démonstration : La proposition précédente nous montre que si $\lim_{n\to\infty} nh = \infty$, $\hat{R}_n(x)$ est asymptotiquement sans biais.

Corollaire IV.1.3.- Sous les mêmes hypothèses de la proposition IV.1.2. et si R est lipschitzienne, on a, pour h de la forme $h = \frac{1}{n^\alpha}$, $0 < \alpha < 1$

$$E\,\hat{R}_n(x) - R(x) = O(h) \quad .$$

Démonstration : En effet, si R est lipschitzienne, $\omega(R,h) = O(h)$; en appliquant la majoration de la proposition IV.1.2., et en tenant compte du fait que si la suite h vérifie

$$\lim_{n\to\infty} \frac{e^{-nh}}{h} = 0 \quad , \quad \text{car} \quad h = \frac{1}{n^\alpha} \quad ,$$

le terme en e^{-nh} est absorbé par celui de h, on obtient le résultat.

IV.2. - Comportement asymptotique de la variance de $\hat{R}_n(x)$, x fixé.

IV.2.1. - Résultats préliminaires.

Pour x fixé dans I_j ,

$$\hat{R}_n(x) = N_{j-1} \frac{(x_j-x)^3 + h^2(x-x_j)}{6h} + N_j \frac{(x-x_{j-1})^3 + h^2(x_{j-1}-x)}{6h}$$

$$+ R_n(x_j) \frac{(x-x_{j-1})}{h} - \frac{R_n(x_{j-1})\ (x-x_j)}{h}$$

avec : $$N_j = \frac{6}{h^2} A^{-1}_{j,o} \left[R_n(x_o) - R_n(x_1)\right] - \frac{18}{h^2} \sum_{i=0}^{\ell+1} A^{-1}_{j,i}\ R_n(x_i)$$

$$+ \frac{6}{h^2} R_n(x_j) + \frac{6}{h^2} A^{-1}_{j,\ell+1} \left[R_n(x_{\ell+1}) - R_n(x_\ell)\right]$$

Lemme IV.2.1.- Il existe deux constantes c et q, indépendantes de h et de j telles que : $c > 0$, $0 < q < 1$ et

$$|A^{-1}_{i,j}| \leq c\ q^{|i-j|} \quad , \quad i\ ,\ j = 0,1,\ldots,\ \ell+1$$

Démonstration : C'est une conséquence immédiate d'un théorème de Domsta [13], une étude plus précise montre que : $q = 2 - \sqrt{3}$ et $c = \sqrt{3}$ conformément à Rosenblatt [asymptotic results on a spline estimate of a probability density , statistical conference. Academic press].

IV.2.2. - Majoration asymptotique de la variance de $\hat{R}_n(x)$.

Proposition IV.2.2.- Sous les hypothèses suivantes :

1) $(\forall P_{XY} \in \mathcal{D})$, $(\exists A > 0)$, indépendant de h et de j, tel que :

$$0 < Ah \leq p_j$$

2) Si $h = \frac{1}{n^{\alpha}}$, $0 < \alpha < 1$, on a :

$$\mathrm{Var}\ \hat{R}_n(x) = O\left(\frac{1}{nh}\right) .$$

Démonstration :

$$\mathrm{Var}\ \hat{R}_n(x) = \left[\frac{(x_j-x)^3 + h^2(x-x_j)}{6h}\right]^2 \mathrm{Var}\ N_{j-1} + \left[\frac{(x-x_{j-1})^3 + h^2(x_{j-1}-x)}{6h}\right]^2 \mathrm{Var}\ N_j$$

$$+ \frac{(x-x_{j-1})^2}{h^2} \mathrm{Var}\ R_n(x_j) + \frac{(x-x_j)^2}{h^2} \mathrm{Var}\ R_n(x_{j-1})$$

$$+ \frac{1}{18h^2}\left[(x_j-x)^3 + h^2(x-x_j)\right] \times \left[(x-x_{j-1})^3 + h^2(x_{j-1}-x)\right] \mathrm{cov}(N_j, N_{j-1})$$

$$+ \frac{1}{3h^2}\left[(x_j-x)^3 + h^2(x-x_j)\right] \times (x-x_{j-1})\ \mathrm{cov}(N_{j-1}, R_n(x_j))$$

$$- \frac{1}{3h^2}\left[(x_j-x)^3 + h^2(x-x_j)\right] \times (x-x_j)\ \mathrm{cov}(N_{j-1}, R_n(x_{j-1}))$$

$$+ \frac{1}{3h^2}\left[(x-x_{j-1})^3 + h^2(x_{j-1}-x)\right] \times (x-x_j)\ \mathrm{cov}\left[N_j, R_n(x_{j-1})\right]$$

$$- \frac{2}{h^2}(x-x_{j-1})(x-x_j)\ \mathrm{cov}\left[R_n(x_j), R_n(x_{j-1})\right]$$

Les coefficients de $\mathrm{Var}\ N_{j-1}$, $\mathrm{Var}\ N_j$ et $\mathrm{cov}(N_j, N_{j-1})$ sont de l'ordre de h^4 ; ceux de $\mathrm{Var}\ R_n(x_j)$, $\mathrm{Var}\ R_n(x_{j-1})$ et de $\mathrm{cov}\left[R_n(x_j), R_n(x_{j-1})\right]$ sont majorés, en valeur absolue par 1 ; ceux de $\mathrm{cov}\left[N_{j-1}, R_n(x_j)\right]$ et $\mathrm{cov}\left[N_{j-1}, R_n(x_{j-1})\right]$ sont de l'ordre de h^2 .

Majoration asymptotique de $\operatorname{Var} N_j$.

$$\operatorname{Var} N_j = \frac{36}{h^4} (A^{-1}_{j,o})^2 \left[\operatorname{Var} R_n(x_o) + \operatorname{Var} R_n(x_1) - 2 \operatorname{cov}\{R_n(x_o), R_n(x_1)\}\right]$$

$$+ \frac{18^2}{h^4} \left[\sum_{i=0}^{\ell+1} (A^{-1}_{j,i})^2 \operatorname{Var} R_n(x_i) + \sum_{\substack{i,k \\ i\neq k}} A^{-1}_{j,i} \; A^{-1}_{j,k} \operatorname{cov}(R_n(x_i), R_n(x_k)\right]$$

$$+ \frac{6^2}{h^4} \operatorname{Var} R_n(x_j) + \frac{6^2}{h^4} (A^{-1}_{j,\ell+1})^2 \left[\operatorname{Var} R_n(x_{\ell+1}) + \operatorname{Var} R_n(x_\ell)\right.$$

$$\left. - 2 \operatorname{cov}\{R_n(x_{\ell+1}), R_n(x_\ell)\}\right] - \frac{216}{h^4} A^{-1}_{j,o} \left[\sum_{i=0}^{\ell+1} A^{-1}_{j,i} \operatorname{cov}\{R_n(x_o), R_n(x_i)\}\right.$$

$$\left. - \sum_{i=0}^{\ell+1} A^{-1}_{j,i} \operatorname{cov}\{R_n(x_1) \; ; \; R_n(x_i)\}\right]$$

$$+ \frac{72}{h^4} A^{-1}_{j,o} \left[\operatorname{cov}\{R_n(x_1), R_n(x_j)\} - \operatorname{cov}\{R_n(x_1), R_n(x_j)\}\right]$$

$$+ \frac{72}{h^4} A^{-1}_{j,o} \; A^{-1}_{j,\ell+1} \left[\operatorname{cov}\{R_n(x_o), R_n(x_\ell)\} - \operatorname{cov}\{R_n(x_o), R_n(x_\ell)\}\right.$$

$$\left. - \operatorname{cov}\{R_n(x_1), R_n(x_{\ell+1})\} + \operatorname{cov}\{R_n(x_1), R_n(x_\ell)\}\right]$$

$$- \frac{216}{h^4} A^{-1}_{j,\ell+1} \left[\sum_{i=0}^{\ell+1} A^{-1}_{j,i} \operatorname{cov}\{R_n(x_{\ell+1}), R_n(x_i)\}\right.$$

$$\left. - \sum_{i=0}^{\ell+1} A^{-1}_{j,i} \operatorname{cov}\{R_n(x_i), R_n(x_\ell)\}\right] + \frac{72}{h^4} A^{-1}_{j,\ell+1}$$

$$\left[\operatorname{cov}\{R_n(x_j), R_n(x_{\ell+1})\} - \operatorname{cov}\{R_n(x_j), R_n(x_\ell)\}\right]$$

$$- \frac{216}{h^4} \sum_{i=0}^{\ell+1} A^{-1}_{j,i} \operatorname{cov}\left[R_n(x_i), R_n(x_j)\right]$$

En appliquant le lemme IV.2.1., chacun des coefficients $A^{-1}_{i,j}$ est majoré, en valeur absolue par c, la majoration est uniforme en i et j car $q < 1$; de plus, on a :

$$\sum_{i=0}^{\ell+1} |A^{-1}_{j,i}| \leq \sum_{i=0}^{\ell+1} c\, q^{|i-j|} = c \sum_{0\leq i\leq j} q^{j-i} + c \sum_{\ell+1\geq i>j} q^{i-j}$$

$$\sum_{i=0}^{\ell+1} |A^{-1}_{j,i}| \leq c\, \frac{q^{j+1} - 1}{q-1} + c\, q\, \frac{q^{+1-j} - 1}{q-1}$$

comme $0 < q < 1$,

$$\sum_{i=0}^{\ell+1} |A^{-1}_{j,i}| < c(1 + q)$$

de même :

$$\sum_{i=0}^{\ell+1} \left[A^{-1}_{j,i}\right]^2 \leq c^2 \sum_{i=0}^{\ell+1} q^{(i-j)^2} \leq c^2 \sum_{i=0}^{\ell+1} q^{|i-j|} < c^2(1 + q)$$

et

$$\sum_{\substack{i,j \\ i\neq k}} |A^{-1}_{j,i} \times A^{-1}_{j,k}| \leq c^2(1 + q)$$

En appliquant les résultats asymptotiques relatifs à la variance et la covariance du régressogramme donnés par le corollaire III.2.1. et la proposition III.3.1., on a :

$$\text{Var } R_n(x) = O\left(\frac{1}{nh}\right)$$

et pour x et x' fixés dans $[0,1[$, tels que : $x \neq x'$,

$$\text{cov}\{R_n(x), R_n(x')\} = O\left(\frac{1}{nh}\right) .$$

Chaque terme qui intervient dans l'expression de Var N_j est composé d'un coefficient de l'ordre $\frac{1}{h^4}$ et d'un autre de l'ordre $\frac{1}{nh}$; il s'en suit que :

$$\text{Var } N_j = O(\frac{1}{nh^5})$$

En utilisant l'inégalité de Schwartz, on a :

$$\text{cov}[N_j, N_{j-l}] = O(\frac{1}{nh^5})$$

$$\text{cov}[N_j, R_n(x)] = O(\frac{1}{nh^3})$$

Aussi, tous les termes qui composent l'expression de Var $\hat{R}_n(x)$ sont de l'ordre de $\frac{1}{nh}$, d'où le résultat.

IV.3. - Non-corrélation asymptotique de $\hat{R}_n(x)$ *et* $\hat{R}_n(x')$, x *et* x' *fixés dans* $[0,1[$, $x \neq x'$.

Proposition IV.3.1.- Sous les hypothèses de la proposition IV.2.2., $\hat{R}_n(x)$ et $\hat{R}_n(x')$, x et x' fixés dans $[0,1[$, avec $x \neq x'$, sont asymptotiquement non-corrélés lorsque $\lim_{\infty} nh = +\infty$.

Démonstration : Pour n assez grand, il va exister deux entiers k et k' , $k \neq k'$, dans $\{1,2,\ldots,\ell+1\}$ tels que $x \in I_k$ et $x' \in I_{k'}$. Les expressions de $\hat{R}_n(x)$ et $\hat{R}_n(x')$ n'étant pas simples, on peut obtenir le résultat en appliquant l'inégalité de Schwartz et la proposition IV.2.2., une démonstration directe du résultat consiste à considérer $E\,\hat{R}_n(x) \times \hat{R}_n(x')$; en groupant convenablement les

termes $E\ N_k\ N_{k'}$; $E\ N_k\ N_{k'-1}$; $E\ N_{k-1}\ N_{k'-1}$, $E\ N_k\ R_n(x_{k'})$, $E\ N_k\ R_n(x_{k'-1})$ $E\ N_{k-1}\ R_n(x_{k'})$; $E\ N_{k-1}\ R_n(x_{k'-1})$; $E\ R_n(x_k)\ R_n(x_{k'})$; $E\ R_n(x_k)\ R_n(x_{k'-1})$ et $E\ R_n(x_{k-1})\ R_n(x'_{k-1})$; en appliquant le lemme IV.2.1., on peut constater que $\mathrm{cov}\left[\hat{R}_n(x), \hat{R}_n(x')\right]$ est composé de termes de l'ordre de $\frac{1}{nh}$ - d'où le résultat.

IV.4. - *Comportement asymptotique de l'erreur quadratique de* $\hat{R}_n(x)$, x *fixé*.

IV.4.1. - *Majoration asymptotique de l'erreur quadratique*.

Elle est donnée par la :

Proposition IV.4.1.- Sous les hypothèses du corollaire IV.1.3. et de la proposition IV.2.2., on a :

$$E\left[\hat{R}_n(x) - R(x)\right]^2 = O(h^2) + O\left(\frac{1}{nh}\right) .$$

Démonstration : Elle est évidente lorsqu'on écrit que :

$$E\left[\hat{R}_n(x) - R(x)\right]^2 = \left[E\ \hat{R}_n(x) - R(x)\right]^2 + \mathrm{Var}\ \hat{R}_n(x) .$$

et en appliquant les majorations asymptotiques du biais et de la variance de $\hat{R}_n(x)$.

IV.4.2. - *Convergence en moyenne quadratique de* $\hat{R}_n(x)$.

Proposition IV.4.2.- Sous les mêmes hypothèses de la proposition précédente et si $\lim_{n\to\infty} nh = \infty$, $\hat{R}_n(x)$ converge en moyenne quadratique vers $R(x)$.

Démonstration : Elle est évidente lorsqu'on considère le résultat de la proposition précédente.

Corollaire IV.4.2.- Sous les mêmes hypothèses de la proposition IV.4.1., et pour $h = O(n^{-1/3})$, on a

$$E\left[\hat{R}_n(x) - R(x)\right]^2 = O(n^{-2/3})$$

Démonstration : Il suffit d'appliquer la proposition IV.4.1. et le lemme de Parzen pour avoir le résultat.

IV.4.3. - Remarques :

1) On peut utiliser la théorie des noyaux reproduisants pour démontrer que l'estimateur spline $\hat{R}_n(x)$ est convergeant en moyenne quadratique ; en effet, si on désigne par $\tilde{R}$ la spline cubique théorique qui vérifie le système :

$$\tilde{R}(x_j) = R(x_j) \qquad 0 \leq j \leq \ell+1$$

$$\tilde{R}'(0) = R'(0)$$

$$\tilde{R}'(1) = R'(1)$$

on peut écrire que :

$$E\left[\hat{R}_n(x) - R(x)\right]^2 \leq 2\left[R(x) - \tilde{R}(x)\right]^2 + 2\,E\left[\tilde{R}(x) - R_n(x)\right]^2 \qquad (*)$$

Le premier terme du 2ème membre se majore facilement grâce à un résultat classique relatif à la convergence des splines cubiques ; voir A.N.W. [1] par exemple ; le 2ème terme du 2ème membre de (*) est aléatoire et se majore en considérant l'espace $C^2[0,1]$ comme espace auto-reproduisant dont le noyau s'écrit :

$$(\forall (s,t) \in [0,1]^2) \ , \quad q(s,t) = 1 + st + \int_0^{\min(s,t)} (x-u)\,(t-u)\,du \ .$$

2) En comparant les résultats obtenus sur le régressogramme, on constate que sous les mêmes hypothèses, $\hat{R}_n(x)$ possède les mêmes propriétés asymptotiques ; aussi, on peut conclure que ces méthodes splines qui ne sont pas en fait des méthodes statistiques permettent d'éviter les inconvénients des estimateurs impropres, tel que le régressogramme ; car, en effet, ce dernier donne une estimation de la régression, supposée, en général, continue, par une seule valeur sur un même intervalle.

3) Collomb [9], en utilisant un résultat d'Epanechnikov, donne l'optimisation de l'estimateur $R_n^K(x)$, $x \in \mathbb{R}^s$, $s \geq 1$, selon le critère quadratique et pour un noyau K supposé réel, positif, symétrique et normé. L'erreur quadratique minimale est toujours de la forme $C.n^{-4/5}$, où la constante C dépend du choix du noyau K.

IV.4.4. - Optimisation selon le critère de "lissage".

Désignons par $\phi_n(x)$, x fixé dans $[a,b]$, un estimateur quelconque de la régression $R(x)$, obtenu par la méthode du noyau, par la méthode de Bosq ou par l'une des méthodes décrites dans le chapitre I.

Soit (Δ) une subdivision de l'intervalle $[a,b]$ définie comme suit :

$$(\Delta) : \ a = \xi_o < \xi_1 < \ldots < \xi_N = b \ , \quad N \in \mathbb{N}^*$$

et $(y_i)_{i=0}^{i=N}$ la suite suivante :

$$y_i = \phi_n(\xi_i) \ , \qquad 0 \leq i \leq N \ .$$

Le problème suivant "trouver l'unique fonction $\hat{r}_n$ dans $C^2\,[0,1]$ et qui vérifie le système :

$$\hat{r}_n(\xi_i) = y_i \qquad 0 \leq i \leq N$$
$$\hat{r}_n'(a) = \hat{a}$$
$$\hat{r}_n'(b) = \hat{b}$$

où $\hat{a}$ et $\hat{b}$ sont égaux respectivement à $\hat{\phi}_n'(a)$ et $\hat{\phi}_n'(b)$ (si elles existent) et qui minimise

$$\int_a^b \left[G''(t)\right]^2 dt$$

parmi toutes les fonctions de $C^2\,[0,1]$ " admet une solution unique $\hat{r}_n$ qui est une spline cubique sur $[a,b]$ aux joints $(\xi_i)_{0\leq i\leq N}$ et qui est le plus "lisse" des estimateurs de $R(x)$.

En outre, comme on a pu le démontrer pour le cas du régressogramme, $\hat{r}_n$ possédera les propriétés asymptotiques de ϕ_n . Dans le dernier chapitre, sur la simulation, on a procédé au lissage de l'estimateur R_n^K , obtenu par deux types de noyau : triangulaire et gaussien. On pourra constater que pour un noyau gaussien, l'estimateur $R_n^K(x)$ coïncide, avec l'estimateur - spline, pour n assez grand et une régression particulière ; d'où la bonne qualité de $R_n^K(x)$, lorsque le noyau est bien choisi.

IV.5. - Convergence en moyenne quadratique intégrée.

IV.5.1. - Introduction. Dans ce qui suit, on va donner les conditions pour lesquelles le "MISE" de la régrospline :

$$E \int_0^1 \left[\hat{R}_n(x) - R(x)\right]^2 dx$$

converge vers 0.

Pour se faire, on va construire, par deux méthodes, une spline qui aura la propriété d'être toujours encadrée par deux valeurs consécutives du régressogramme.

IV.5.2. - Résultats préliminaires.

Proposition IV.5.1.- Soit k un entier positif ou nul, et $\tilde{R}$ et S deux fonctions vérifiant les conditions suivantes :

1) $\tilde{R} \in C^k[0,1]$,

2) S est une spline polynomiale telle que :

- $S \in C^{2m-2}[0,1]$, m entier tel que : $0 \leq k < 2$

- S interpole $\tilde{R}$ au sens suivant :

$$\begin{cases} (\tilde{R}-S)(x_j) = 0 \quad , & 1 \leq j \leq \ell \\ D^i(R-S)(0) = D^i(\tilde{R}-S)(1) = 0 \quad , & 0 \leq i \leq \min(k,m-1) \\ D^i S(0) = D^i S(1) = 0 \quad , & k < i \leq m-1 \end{cases}$$

où D désigne l'opérateur de dérivation ; alors on a :

$$K\, h^{k-j+1/q-1/p}\, ||D^k\tilde{R}||_{L_p} \geq \begin{cases} ||D^j(\tilde{R}-S)||_{L_q} \; ; \; 0 \leq j \leq k-1 \; ; \; p \leq q \leq \infty \\ ||D^j S||_{L_q} \; ; \quad k-1 < j \leq 2m-1 \; ; \; p \leq q \leq \infty \end{cases}$$

avec K constante indépendante de $\tilde{R}$ et de h.

Démonstration : Le résultat énoncé est un cas particulier d'un théorème de R.S. Varga et S. Demko [32'], page 242, formule 1.3.

IV.5.3. - Limite du "MISE".

Elle est donnée par la :

Proposition IV.5.2.- Sous les hypothèses suivantes :

$(\forall P_{XY} \in \mathcal{D})$, $(\exists A > 0)$, indépendant de h et de j, tel que :

$$0 < Ah \leq p_j$$

Alors, on a :

$$\lim_{n\to\infty} E \int_0^1 \left[\hat{R}_n(x) - R(x)\right]^2 dx = 0$$

si $\lim_{n\to\infty} nh = \infty$.

Démonstration : Pour tout x fixé dans I_j, on pose

$$M_n(x) = \max\left[R_n(x_j), R_n(x_{j-1})\right]$$

$$m_n(x) = \min\left[R_n(x_j), R_n(x_{j-1})\right]$$

Comme, on l'a indiqué dans l'introduction, on va construire une spline r_n qui possède la propriété suivante :

$$m_n(x) \leq r_n(x) \leq M_n(x) .$$

Pour ce faire, on dispose de deux méthodes.

a) *1ère méthode*. Fonctions-spline relatives à un convexe. Cette méthode a été introduite par Atteia et développée, entre autres, par Morin [18] et Laurent [16].

Il s'agit de trouver une fonction $r_n(x)$ dans l'espace

$$C = \{g \in L^2\,[0,1] \ , \ x \in I_j \ , \ j = 1,2,\ldots, +1, \ m_n(x) \leq r_n(x) \leq M_n(x)\}$$

et qui réalise la condition suivante :

$$\int_0^1 \left[r_n''(x)\right]^2 dx = \min_{g \in C} \int_0^1 \left[g''(x)\right]^2 dx$$

Conformément à Laurent [16], page 492, ou Morin [18], chapitre I, un tel problème a toujours une solution ; en outre, cette solution, $r_n(x)$, est une spline cubique ; on peut se référer aux deux auteurs déjà cités pour l'unicité et la caractérisation de cette fonction.

b) *2ème méthode*. Splines d'Hermite.

Soit S l'espace des fonctions $S(x)$ telles que :

i) $s(x) \in C^1 [0,1]$,

ii) $s(x)$ est un polynôme de degré ≤ 3 sur tout I_j .

D'après Schenberg [30], page 44, lemme 1, le problème suivant : trouver une spline $r_n(x)$ dans S vérifiant les conditions

iii) $r_n(x_j) = R_n(x_j)$, $\quad 0 \leq j \leq \ell+1$

iv) $r_n'(x_j) = 0$ $\quad$ "

a une solution unique $r_n(x)$; on peut aussi, se référer à Laurent [16], page 221, exemple 2.

Dans notre démonstration, on retiendra cette dernière méthode ; $r_n(x)$ désignera, donc, dans toute la suite, la spline d'Hermite qui vérifie les conditions i), ii), iii) et iv) ; étant ainsi définie, elle possède la propriété suivante :

$$m_n(x) \leq r_n(x) \leq M_n(x) .$$

On a alors :

$$||\hat{R}_n - R||^2_{L_2} \leq 2\,||\hat{R}_n - r_n||^2_{L_2} + 2\,||r_n - R||^2_{L_2}$$

où $||.||_{L_2}$ désigne la norme dans $L_2\,[0,1]$.

1°) *Majoration de* $||r_n - R||^2_{L_2}$.

En appliquant la propriété de $r_n(x)$, on peut écrire que :

$$|r_n(x) - R(x)|^2 \leq 2|M_n(x) - R(x)|^2 + 2|m_n(x) - R(x)|^2$$

Ainsi, pour majorer les 2 termes du 2ème membre, on doit considérer 2 cas :

a) Si $M_n(x) = R_n(x_j)$.

Dans ce cas, $m_n(x) = R_n(x_{j-1})$.

Aussi, on peut écrire :

$$|r_n(x) - R(x)|^2 \leq 2|R_n(x) - R(x)|^2 + 2|R_n(x_{j-1}) - R(x)|^2$$

car pour $x \in I_j$, $R_n(x) = R_n(x_j)$

$$\sup_{x\in I_j} |r_n(x) - R(x)|^2 \leq 2 \sup_{x\in I_j} |R_n(x) - R(x)|^2 + 2 \sup_{x\in I_j} |R_n(x_j) - R(x)|^2$$

Or $\sup_{I_j} |R_n(x) - R(x)|^2 \leq 2 \sup_{I_j} E|R_n(x) - R(x))^2 + 2|R_n(x) - E\,R_n(x))^2$

ou bien :

$$\sup_{I_j} |R_n(x) - R(x))^2 \leq 4 \sup_{I_j} |E\,R_n(x) - E(Y/X\in I_j)|^2 + 4 \sup_{I_j} |R(x) - E(Y/X\in I_j)|^2$$

$$+ 2\,|R_n(x) - E\,R_n(x)|^2$$

Or, R étant continue, il existe $\xi_j \in I_j$, tel que :

$$E(Y/X\epsilon I_j) = R(\xi_j)$$

d'où :

$$\sup_{I_j} |R_n(x) - R(x)|^2 \leq 2|R_n(x) - E\,R_n(x)|^2 + 4 \sup_{I_j} |R(x) - R(\xi_j)|^2 + 4 \sup_{I_j} |E\,R_n(x) - E(Y/X\epsilon I_j)|^2$$

Or $\exists$ une constante Λ positive telle que :

$$\sup_{I_j} |E\,R_n(x) - E(Y/X\epsilon I_j)|^2 \leq \Lambda^2\, e^{-2nh}$$

et $$\sup_{I_j} |R(x) - R(\xi_j)|^2 \leq \omega^2(R,h) \ .$$

Il vient alors :

$$\sup_{x\epsilon I_j} |R_n(x) - R(x)|^2 \leq 2|R_n(x) - E\,R_n(x)|^2 + 4\,\omega^2(R,h) + 4\,\Lambda^2\, e^{-2nh}$$

Par ailleurs,

$$\sup_{I_j} |R_n(x_{j-1}) - R(x)|^2 \leq 2 \sup_{I_j} |R_n(x_{j-1}) - E(Y/X\epsilon I_{j-1})|^2 + 2 \sup_{I_j} |E(Y/X\epsilon I_{j-1}) - R(x)|^2$$

$$\leq 2\,|R_n(x_{j-1}) - E(Y/X\epsilon I_{j-1})|^2 + 2\,\omega^2(R,h)$$

$$\leq 2\,\Lambda^2\, e^{-2nh} + 2\,\omega^2(R,h)$$

Finalement :

$$\sup_{I_j} |r_n(x) - R(x)|^2 \leq 4|R_n(x) - E\,R_n(x)|^2 + 10\,\omega(R,h) + 10\,\Lambda^2\, e^{-2nh} \ .$$

En intégrant par rapport à x, on a :

$$\int_{I_j} |r_n(x) - R(x)|^2 \, dx \leq 4 \, h \, |R_n(s_j) - E\, R_n(s_j)|^2 + 10 \, h \, \omega^2(R,h) + 10 \, \Lambda^2 \, h \, e^{-2nh}$$

où $s_j \in I_j$.

En intégrant par rapport à la probabilité P, il vient :

$$E \int_{I_j} |r_n(x) - R(x)|^2 \, dx \leq 4 \, h \, \mathrm{Var}\, R_n(s_j) + 10 \, h \, \omega^2(R,h) + 10 \, \Lambda^2 \, h \, e^{-2nh}$$

d'où

$$E \int_0^1 |r_n(x) - R(x)|^2 \, dx = \sum_{j=1}^{\ell+1} E \int_{I_j} |r_n(x) - R(x)|^2 \, dx$$

Comme $(\ell+1)\, h = 1$, en utilisant les résultats asymptotiques sur la variance du régressogramme, corollaire III.2.1., on a :

$$E||r_n - R||_{L_2}^2 \leq 4 \, \frac{V}{nh} + 10 \, \omega^2(R,h) + 10 \, \Lambda^2 \, e^{-2nh} \, .$$

b) <u>Si</u> $M_n(x) = R_n(x_{j-1})$

Il suffit d'écrire que :

$$|r_n(x) - R(x)|^2 \leq 2|R_n(x_{j-1}) - R(x)|^2 + 2|R_n(x_j) - R(x)|^2$$

Par conséquent, on aboutit à la même majoration asymptotique que précédemment.

2°) <u>*Majoration de*</u> $||\hat{R}_n - r_n||_{L_2}^2$

Dans la proposition IV.5.1., r_n et $\hat{R}_n$ vérifient respectivement les conditions des fonctions $\tilde{R}$ et S.

en mettant $k = 1$, $p = q = 2$, $m = 2$ et $j = 0$, on a :

$$||\hat{R}_n - r_n||^2_{L_2} \leq K\, h^2\, ||r'_n||^2_{L_2}$$

où K est une constante indépendante de h et de r_n - en intégrant par rapport à P, il vient :

$$E||\hat{R}_n - r_n||^2_{L_2} \leq K\, h^2\, E||r'_n||^2_{L_2}$$

En tenant compte du fait que :

$$|r_n(x)| \leq M_n(x)$$

on peut voir que : $E||r'_n||^2_{L_2}$ est bornée par une constante $(B > 0)$ indépendante de h et de j, d'où

$$E||\hat{R}_n - r_n||^2_{L_2} \leq B \,.\, K.\, h^2 \,.$$

Il vient, en définitive que :

$$E \int_0^1 |\hat{R}_n(x) - R(x)|^2 \, dx \leq 2\,(B.K)\, h^2 + \frac{8V}{nh} + 2\,\omega^2(R,h) + \theta\, e^{-2nh} \,.$$

θ = constante indépendante de h et de j ; dès que $\lim_{n\to\infty} nh = \infty$, on a le résultat.

Corollaire IV.5.3.- Sous les hypothèses suivantes :

$(\forall P_{XY} \in \mathcal{D})$, $(\exists A > 0)$, indépendante de h et de j , tel que

$$0 < Ah \leq p_j \,,$$

alors pour $h = O(n^{-1/3})$, on a :

$$E \int_0^1 |\hat{R}_n(x) - R(x)|^2 \, dx = O(n^{-2/3})$$

Démonstration : Il suffit d'appliquer le lemme de Parzen [22], pour avoir le résultat ; le terme en e^{-nh} étant absorbé par celui de $\frac{1}{nh}$.

IV.6. - Convergence uniforme presque complète.

IV.6.1. - Introduction.

Dans ce paragraphe, on se propose d'étudier la convergence uniforme presque complète (p.co.) vers 0 de la v.a. $(\hat{R}_n - R)$, $\hat{R}_n$ désignant toujours l'estimateur - spline ; pour ce faire, on utilisera des résultats dûs à Geffroy [15] ; ce dernier a montré que si la suite h vérifie la condition

$$\lim_{n\to\infty} \frac{nh}{\text{Log } n} = +\infty$$

alors : $\forall \varepsilon > 0$, $\sum_{n\geq 1} P\left[\sup_{0\leq x\leq 1} |R_n(x) - R(x)| > \varepsilon\right] < +\infty$.

Collomb [9] montre que cette même condition sur h est nécessaire et suffisante pour que l'on ait :

$$\sup_{x\in G} |R_n^K(x) - R(x)| \xrightarrow[n\uparrow\infty]{} 0 \quad \text{p.co.}$$

où G est une partie bornée non vide de $\mathbb{R}^s$, $s \geq 1$ et R_n^K l'estimateur de la régression R par la méthode du noyau.

IV.6.2. - Résultats préliminaires (Geffroy [15]).

Dans cette partie, l'hypothèse relative à la v.a.r. Y est que EY existe et que R est continue uniformément sur $[0,1[$.

Lemme IV.6.2.- Soit P_{XY} une loi de $\mathcal{D}$ telle que :

$(\exists A > 0)$ indépendant de h et de j, avec

$$p_j \geq Ah > 0$$

Choisissons arbitrairement A' dans $]0, \frac{A}{2}[$ et considérons l'événement : $H_n = \bigcap_{j=1}^{\ell+1} \{\nu_{n_j} \geq nh\, A'\}$ alors si la suite h vérifie :

$$\lim_{n\to\infty} \frac{nh}{\text{Log } n} = +\infty$$

On a : $$\sum_{n\geq 1} P(\bar{H}_n) < +\infty .$$

Démonstration : La v.a. ν_{n_j} suit la loi binomiale de paramètres n et p_j ; on a donc :

$$E(\nu_{n_j}) = n\, p_j \geq n\, Ah$$

Il vient alors $$E(\nu_{n_j}) \geq A \frac{\text{Log } n}{\varepsilon_n} \quad \text{avec} \quad \lim_{n\to\infty} \varepsilon_n = 0$$

car h peut s'écrire sous la forme :

$$h^{-1} = \frac{n\, \varepsilon_n}{\text{Log } n} .$$

Ceci implique que $E(\nu_{n_j})$ croît indéfiniment avec n ; posons

$$m = [nh\, A']$$

pour n assez grand, on a

$$m < E(\nu_{n_j})$$

Une inégalité classique, voir Feller [14], formule 3.6, page 140, nous donne :

$$P\left[\nu_{n_j} \leq m\right] < C_n^m \, p_j^{m+1} \, q_j^{n-m} \, \frac{n-m+1}{(n+1)p_j - m}$$

Posons $$\beta_n = \frac{n!}{n!(n-m)!} \, p_j^{m+1} \, q_j^{n-m} \, \frac{n-m+1}{(n+1)p_j - m}$$

D'après le choix de A' et de m, on a :

$$n \, p_j > 2 \, m$$

et $$\frac{p_j(n-m+1)}{(n+1)\, p_j - m} < 2$$

En appliquant la formule de Stirling, on en déduit :

$$\beta_n < (2 + \varepsilon'_{nm}) \sqrt{\frac{n}{2\pi(n-m)m}} \, p_j^m \, q_j^{n-m} \left(\frac{m}{n}\right)^{-m} \left(1 - \frac{m}{n}\right)^{m-n}$$

où $$\lim_{\substack{m\to\infty \\ n\to\infty}} \varepsilon'_{mn} = 0$$

Comme $\frac{m}{n} \longrightarrow 0$ avec $\frac{1}{n}$, on a :

$$\beta_n < \frac{2 + \varepsilon''_n}{\sqrt{2\pi m}} \, p_j^m \, q_j^{n-m} \left(\frac{m}{n}\right)^{-m} \left(1 - \frac{m}{n}\right)^{m-n}$$

avec $\lim_{n\to\infty} \varepsilon''_n = 0$ uniformément en j.

Considérons la fonction :

$$u(p) = p^m (1-p)^{n-m}, \quad 0 \leq p \leq 1$$

on a : $$u'(p) = \frac{u(p)(m-np)}{p(1-p)}$$

donc $u(p)$ est décroissante pour $p > \frac{m}{n}$; or :

$$\frac{m}{n} < Ah < p_j$$

il en résulte que :

$$u(p_j) < u(Ah)$$

d'où une majoration uniforme en j de $u(p_j)$

$$p_j^m \, q_j^{n-m} < A^m \, h^m (1-Ah)^{n-m}$$

en tenant compte du fait que m peut s'écrire :

$$m = nh\, A'_n \quad \text{avec} \quad \lim_{n\to\infty} A'_n = A$$

il vient :

$$\beta_n < \frac{2 + \varepsilon''_n}{\sqrt{2\pi m}} \left(\frac{A}{A'_n}\right)^m \left(\frac{1 - Ah}{1-A'_n h}\right)^{n-m}$$

Or, on peut écrire aussi :

$$\left(\frac{A}{A'_n}\right)^m = \exp\left(m \operatorname{Log} \frac{A}{A'_n}\right) = \exp\left(nh\, A'_n \operatorname{Log} \frac{A}{A'_n}\right)$$

$$= \exp\left(\frac{\operatorname{Log} n}{\varepsilon_n} \cdot A'_n \operatorname{Log} \frac{A}{A'_n}\right)$$

et $$\left(\frac{1-A_h}{1-A'_n h}\right)^{n-m} = \exp\left[(A - A'_n)\, nh(1 + O(1))\right]$$

$$= \exp\left[-\frac{\operatorname{Log} n}{\varepsilon_n} (A-A'_n)\ (1 + O(1))\right]$$

Aussi, il vient que :

$$\beta_n < \frac{2 + \varepsilon''_n}{\sqrt{2\pi m}} \exp\left[-\frac{\operatorname{Log} n}{n} (A - A'_n - A'_n \operatorname{Log} \frac{A}{A'_n} + O(1))\right]$$

mais : $$\lim_{n\to\infty} \left[A - A'_n - A'_n \operatorname{Log} \frac{A}{A'_n}\right] = A - A' - A' \operatorname{Log} \frac{A}{A'}$$

$$= A' \left[\frac{A}{A'} - 1 - \operatorname{Log} \frac{A}{A'}\right] > 0$$

car $\frac{A}{A'} > 2$. Par ailleurs, il va exister un nombre $\delta > 0$ et un entier n_o tels que :

$$n > n_o \ , \quad A - A'_n - A'_n \operatorname{Log} \frac{A}{A'_n} > \delta$$

auquel cas :

$$\beta_n < \frac{2 + \varepsilon''_n}{\sqrt{2\pi m}} \exp\left[- \frac{\operatorname{Log} n}{\varepsilon_n} (\delta + O(1))\right]$$

et pour n assez grand, $n > n_1$,

$$\beta_n < \frac{1}{A'} \left(\frac{\operatorname{Log} n}{\varepsilon_n}\right)^{-1/2} \times n^{-\delta/2\varepsilon_n} \ .$$

Finalement,

$$P\left[\bar{H}_n\right] = P\left[\bigcup_{j=1}^{\ell+1} \{\nu_{n_j} = m\}\right] \leq \sum_{j=1}^{\ell+1} P(\nu_{n_j} \leq m) \leq (\ell+1)\ \beta_n$$

et pour $n > n_1$

$$P\left[\bar{H}_n\right] < \frac{n\ \varepsilon_n}{\operatorname{Log} n}\ \beta_n < \frac{1}{A'} \left(\frac{\operatorname{Log} n}{\varepsilon_n}\right)^{-3/2} n^{-\delta/2\varepsilon_n} + 1$$

et il est clair que la série dont le terme général est le majorant ci-dessus de $P(\bar{H}_n)$ est convergente ; d'où le résultat.

Lemme IV.6.3.- Avec les notations et les hypothèses du lemme précédent, on a :

$$E\left[R_n(x)/H_n\right] = R(\xi_j) + \lambda\, \theta_{n_j}\, h$$

avec ξ_j arbitraire dans I_j , θ_{n_j} est une constante telle que $|\theta_{n_j}| \leq 1$ et λ un réel positif.

Démonstration : R étant supposée uniformément continue sur $[0,1[$, il existe $\lambda > 0$ tel que :

$$\forall (t,t') \in [0,1[^2 \ , \ |R(t) - R(t')| \leq \lambda |t-t'|$$

Soit ξ_j dans I_j ; on peut écrire :

$$|R(x) - R(\xi_j)| \leq \lambda h$$

et soit $(t_1,t_2,\ldots,t_n)$ dans $[0,1[$; il vient :

$$E\left[R_n(x)/H_n \cap (t_1,t_2,\ldots,t_n)\right] = \frac{1}{\nu_{n_j}} \sum_{X_i \in I_j} R(t_i)$$

$$= \frac{1}{\nu_{n_j}} \sum_{X_i \in I_j} \left[R(\xi_j) + \lambda h\, \tau_{n_j}(t_i)\right]$$

avec $0 \leq \tau_{n_j}(t_i) \leq 1$; i étant tel que $X_i \in I_j$.

On en déduit que :

$$E\left[R_n(x)/H_n \cap (t_1,t_2,\ldots,t_n)\right] = R(\xi_j) + \lambda h\, r'_{n_j}(t_1,t_2,\ldots,t_n)$$

avec $$|r'_{n_j}(t_1,\ldots,t_n)| \leq 1$$

en intégrant par rapport à $(t_1,t_2,\ldots,t_n)$, on a

$$E\left[R_n(x)/H_n\right] = R(\xi_j) + \lambda h\, \theta_{n_j} \ , \ |\theta_{n_j}| \leq 1 \ .$$

Proposition IV.6.1.- Sous l'hypothèse du lemme IV.6.2. et si

1) $(\forall P_{XY} \in \mathcal{D})$, les lois conditionnelles P_Y^x , $x \in [0,1[$ ont des moments de tout ordre ,

2) $(\exists V > 0)$, $(\forall P_{XY} \in \mathcal{D})$, $(\forall x \in [0,1[)$, $\mathrm{Var}(Y/x) \leq V$,

3) $(\exists \alpha > 0)$, $(\forall P_{XY} \in \mathcal{D})$, $(\forall x \in [0,1[)$, $(\forall k \geq 2)$

$$\left|E\left(\left[Y - E(Y/x)\right]^k/x\right)\right| \leq \frac{k!}{2} 2^{k-2} \mathrm{Var}(Y/x)$$

alors si : $$\lim_{n\to\infty} \frac{nh}{\mathrm{Log}\ n} = \infty$$

on a : $(\forall P_{XY} \in \mathcal{D})$, $U_n = d(R_n,R) \xrightarrow[n\uparrow\infty]{} 0$ p.co. où d désigne la métrique de la convergence uniforme.

Démonstration : Soit $(t_1,t_2,\ldots,t_n)$ dans $[0,1[$ et σ^2 la variance de $\sum_{X_i \in I_j} Y_i$ conditionnée par $(t_1,t_2,\ldots,t_n)$ l'hypothèse 2) entraîne :

$$\sigma^2 \leq \nu_{n_j} \cdot V$$

Soit c un nombre positif quelconque et β un nombre réel tel que :

$$0 < \alpha\ \beta \leq 1/2$$

Suivant une inégalité classique, on a :

$$(\forall x \in I_j)\ ,\quad P\left[\left|R_n(x) - E\ R_n(x)\right| > \frac{c}{\nu_{n_j}\ \beta} + \frac{\beta^2}{\nu_{n_j}} \Big/ H_n\ (t_1,\ldots,t_n)\right] \leq 2e^{-c}$$

mais on a : $\nu_{n_j} > nh\ A'$; aussi, on peut écrire :

$$P\left[\left|R_n(x) - E\,R_n(x)\right| > \frac{c}{nhA'\beta} + {}^{\beta V}/_{H_n} \cap (t_1,t_2,\ldots,t_n)\right] \leq 2\,e^{-c}$$

comme $\beta V \leq \frac{V}{2\alpha}$, on peut choisir β tel que :

$$\beta V = \frac{\varepsilon}{2}$$

et c de façon que : $\frac{c}{nh\,A'} = \frac{\varepsilon}{2}$

c'est-à-dire : $c = \frac{\varepsilon}{2}\quad A'\beta nh = \frac{\varepsilon^2}{4V}\,A'\,\frac{\text{Log } n}{\varepsilon_n}$.

Il vient :

$$P\left[\left|R_n(x) - E\,R_n(x)\right| > \varepsilon/H_n \cap (t_1,\ldots,t_n)\right] \leq 2\,\exp\left[-\,\varepsilon^2\,\frac{A'\,\text{Log } n}{4V\varepsilon_n}\right]$$

mais le lemme IV.6.3. implique que :

$$\left|E(R_n/H_n) - R(x)\right| \leq 2h\,\lambda$$

et pour $n > n_o$, $2\,\lambda h < \varepsilon$

d'où : $$P\left[\left|R_n(x) - R(x)\right| > 2\varepsilon/H_n\right] \leq 2n^{-\varepsilon^2 A'/4V\varepsilon_n}.$$

Il en résulte que :

$$P\left[U_n > 2\varepsilon/H_n\right] \leq 2(\ell+1)\,n^{-\varepsilon^2 A'/4V\varepsilon_n}$$

finalement :

$$P\left[U_n > 2\varepsilon\right] \leq \frac{2\,\varepsilon_n}{\text{Log } n}\,n^{1-\varepsilon^2 A'/4V\varepsilon_n} + P(\bar{H}_n)\ .$$

Le lemme IV.6.2. implique que $\sum_{n\geq 1} P(\bar{H}_n) < +\infty$ et il est évident que :

$$\sum_{n\geq 1} \frac{\varepsilon_n}{\text{Log } n}\, n^{1-\varepsilon^2 A'/4V\varepsilon_n} < +\infty$$

ce qui achève la démonstration.

IV.6.3.- Convergence uniforme presque complète de $\hat{R}_n$.

Dans ce qui suit, nous donnons d'abord un résultat relatif à la majoration asymptotique de $d(\hat{R}_n,R)$, d désignant la métrique de la convergence uniforme ; U_n désignant toujours $d(R_n,R)$.

Proposition IV.6.3.- Avec les notations précédentes, on a :

$$d(\hat{R}_n,R) \leq (\frac{8}{3\sqrt{3}} + 2)\, U_n + \frac{4}{3\sqrt{3}}\, \omega(R,h) \ .$$

Démonstration :

$$\hat{R}_n(x) - R(x) = N_{j-1} \frac{(x_j-x)^3 - h^2(x_j-x)}{6h} + N_j \frac{(x-x_{j-1})^3 - h^2(x-x_{j-1})}{6h}$$

$$+ \left[R_n(x_j) - R(x)\right] \frac{(x-x_{j-1})}{h} - \left[R_n(x_{j-1}) - R(x)\right] \frac{(x-x_j)}{h}$$

Les coefficients de N_j et N_{j-1} sont majorés en valeur absolue par $\frac{h^2}{9\sqrt{3}}$; de la proposition IV.1.1., on en tire en particulier :

$$\max_{0\leq j\leq \ell+1} |N_j| \leq \max_{0\leq j\leq \ell+1} |d_j|$$

car les (N_j), $0 \leq j \leq \ell+1$, vérifient le système matriciel :

$$\begin{bmatrix} 2 & 1 & & & & & & & \\ 1/2 & 2 & 1/2 & & & & & & \\ & 1/2 & 2 & 1/2 & & & & & \\ & & \ddots & \ddots & \ddots & & & & \\ & & & \ddots & \ddots & \ddots & & & \\ & & & & \ddots & \ddots & \ddots & & \\ & & & & & 1/2 & 2 & 1/2 & \\ & & & & & & 1 & 2 & \end{bmatrix} \begin{bmatrix} N_o \\ N_1 \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ N_{\ell+1} \end{bmatrix} = \begin{bmatrix} d_o \\ d_1 \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ d_{\ell+1} \end{bmatrix}$$

avec $d_j = \frac{3}{h}\left[\frac{R_n(x_{j+1}) - R_n(x_j)}{h} - \frac{R_n(x_j) - R_n(x_{j-1})}{h}\right]$, $1 \leq j \leq \ell$

$$d_o = d_{\ell+1} = 0$$

d'où

$$|d_j| \leq \frac{3}{h^2}\left[|R_n(x_{j-1}) - R(x_{j+1})| + |R(x_{j+1}) - R(x_j)| + 2|R(x_j) - R_n(x_j)|\right.$$
$$\left. + |R(x_j) - R(x_{j-1})| + |R(x_{j-1}) - R_n(x_{j-1})|\right]$$

comme $\omega(R,h) = \max_{1\leq j\leq \ell+1} \omega_j(R,h)$, il vient

$$\max_j |d_j| \leq \frac{3}{h^2}\left[4\ U_n + 2\omega(R,h)\right]$$

aussi, $$d(\hat{R}_n, R) \leq \left(\frac{8}{3\sqrt{3}} + 2\right) U_n + \frac{4}{3\sqrt{3}}\omega(R,h) \quad .$$

Corollaire IV.6.3.- Sous l'hypothèse suivante :

$(\forall P_{XY} \in \mathcal{D})$, $(\exists A > 0)$, indépendant de h et de j, tel que :

$$0 < Ah \leq p_j$$

alors si la suite h vérifie :

$$\lim_{n\to\infty} \frac{nh}{\text{Log } n} = +\infty$$

on a $$d(\hat{R}_n, R) \xrightarrow[n\uparrow\infty]{} 0 \text{ p.co.}$$

Démonstration : Y étant bornée, les hypothèses 1,2,3, de la proposition IV.6.2. sont vérifiées, la suite h vérifiant la condition :

$$\lim_{n\to\infty} \frac{nh}{\text{Log } n} = \infty ,$$

on a : $$U_n \xrightarrow[n\uparrow\infty]{} 0 \text{ p.co.}$$

R étant continue, on a :

$$\lim_{n\to\infty} \omega(R,h) = 0$$

d'où le résultat.

Remarque : On peut se poser la question suivante : existe-t-il une hypothèse sur la suite h, moins restrictive que l'hypothèse $\lim_{n\to\infty} \frac{nh}{\text{Log } n} = \infty$ et qui assure la convergence uniforme presque complète des estimateurs R_n et $\hat{R}_n$. Une réponse à cette question est donnée par Geffroy [15], en ce qui concerne le régressogramme ; en effet, Geffroy

montre que le régressogramme ne peut pas converger uniformément en probabilité si la suite h ne vérifie pas la condition :

$$\lim_{n\to\infty} \frac{nh}{\text{Log } n} = \infty.$$

IV.6.4. - Convergence en moyenne uniforme de $\hat{R}_n$.

Proposition IV.6.4.- Sous les hypothèses du corollaire précédent,

on a : $$(\forall P_{XY} \in \mathcal{D}) \ , \quad E \sup_x |\hat{R}_n(x) - R(x)| \xrightarrow[n\to\infty]{} 0$$

Démonstration : Considérons la v.a.r.

$$Z_n = \sup_x |\hat{R}_n(x) - R(x)|$$

Y étant bornée, $R_n(x)$ et $R(x)$ sont aussi bornés ; de plus $\hat{R}_n(x)$ s'écrit :

$$\hat{R}_n(x) = N_{j-1} \frac{(x_j - x)^3 + h^2(x - x_j)}{6h} + N_j \frac{(x - x_{j-1})^3 - h^2(x - x_{j-1})}{6h}$$

$$+ \frac{R_n(x_j)(x - x_{j-1}) - R_n(x_{j-1})(x - x_j)}{h}$$

Les coefficients de N_{j-1} et N_j sont majorés en valeur absolue par $\frac{h^2}{9\sqrt{3}}$ et $\max_{0\leq j\leq \ell+1} |N_j| \leq \max_{0\leq j\leq \ell+1} |d_j|$

avec

$$\max_{0\leq j\leq \ell+1} |d_j| \leq \frac{3}{h^2} \left[4 u_n + 2 (R,h)\right].$$

U_n est bornée pour tout n car Y est bornée ; il vient que $\hat{R}_n(x)$ est bornée pour tout x et que par conséquent la v.a.r. Z_n est positive

et bornée pour tout entier n.

D'après le corollaire précédent, on a en particulier,

$$Z_n \xrightarrow[n\to\infty]{P} 0$$

et d'après un résultat de Neveu [21], proposition II.5.4., on a :

$$E\, Z_n \xrightarrow[n\uparrow\infty]{} 0$$

d'où le résultat.

CHAPITRE V

S I M U L A T I O N.

V.1. - *Introduction.*

Dans ce qui suit, on va étudier quelques exemples dont ceux de Watson [36] ; on considérera, comme ce dernier, le cas où X et Y sont deux v.a.r. et où la densité de X est

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

avec :
$$\frac{dP_Y^{X=x}}{dy} = g_x$$

et
$$g_x(y) = \frac{1}{\sqrt{2\pi}} e^{-1/2\left[y-R(x)\right]^2}$$

pour x et y dans $[0,1]$ et la régression R particulière à chaque exemple ; dans chaque figure, on a regroupé les trois estimateurs suivants, le régressogramme R_n , l'estimateur R_n^K , obtenu par le noyau K et la spline $\hat{R}_n$.

Dans un premier temps, on a "interpolé" le régressogramme aux milieux des intervalles I_j , $j = 1,2,\ldots,\ell+1$; on pourra remarquer que la spline obtenue est meilleure que celle obtenue en "interpolant" le régressogramme aux points x_j , $j = 0,1,\ldots, \ell+1$.

Dans chaque cas, on a choisi pour l'estimateur R_n^K , deux types de noyau : triangulaire et gaussien.

En outre, on a choisi une taille $n = 1000$ et $h_{1000} = \frac{1}{10}$.

V.2. - Exemples.

Exemple 1.- C'est le cas où X et Y sont indépendantes ; la régression R est, donc, constante ; on a choisi, arbitrairement, cette dernière égale à :

$$R(x) = 1/2 \quad , \quad x \in [0,1]$$

a) Cas où le noyau est défini par :

$$K(x) = 1 - |x| \quad \text{si} \quad |x| \leq 1$$

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

Dans cet exemple, la spline est nettement meilleure que l'estimateur $R_n^K(x)$ et ceci est essentiellement dû au choix du noyau. On constate que la spline est assez proche du régressogramme ; l'erreur est en moyenne par défaut, aussi bien pour la spline que pour l'estimateur $R_n^K(x)$; voir figure 1.a.

b) Cas où le noyau est gaussien :

$$K(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \qquad x \in [0,1] \ .$$

L'estimateur R_n^K est nettement plus lisse que celui qui précède ; l'erreur est en moyenne par défaut ; voir figure 1.b.

Exemple 2.- Cas où la régression est définie par :

$$R(x) = 0{,}8\ x \quad , \quad x \in [0,1]$$

Cet exemple a été étudié par Watson [36].

a) Dans le cas où le noyau est triangulaire, la spline est nettement meilleure que la méthode du noyau ; l'erreur est en moyenne par défaut pour les deux estimateurs $\hat{R}_n$ et R_n^K , voir figure 2.a.

b) Dans le cas où le noyau est gaussien, les remarques sont les mêmes que dans le cas 1.b. ; voir figure 2.b.

Exemple 3.- Cas où la régression est définie par :

$$R(x) = 3\,x^2 \quad , \quad x \in [0,1]$$

Cet exemple a été étudié par Watson [36] pour une taille $n = 100$ et pour différentes valeurs de h : 1/2 , 1/5, 1/10 ; Watson considère aussi l'estimateur obtenu par la méthode des moindres carrés.

a) Cas du noyau triangulaire : comme dans les deux exemples précédents, l'estimateur R_n^K est assez mauvais comparativement à l'estimateur spline $\hat{R}_n$; l'erreur est en moyenne par défaut voir figure 3.a.

b) Cas du noyau gaussien : l'estimateur R_n^K correspondant est nettement meilleur ; l'erreur est en moyenne par défaut - voir figure 3.b.

c) Cas où on interpole le régressogramme aux points (x_j) , $j = 0,1,\ldots, \ell+1$; on voit aisément que la spline $\hat{R}_n$ n'est pas aussi proche de la régression R que dans les deux derniers cas précédents. L'estimateur R_n^K obtenu le noyau gaussien est nettement meilleur voir figures 3.c. et 3.d.

Exemple 4.- Cas où la régression est définie par :

$$R(x) = - X^3 + X - 4$$

a) Cas du noyau triangulaire : l'erreur est en moyenne par défaut. $\hat{R}_n$ est nettement meilleur que R_n^K ; voir figure 4.a.

b) Cas du noyau gaussien : l'erreur est moyenne par défaut ; l'estimateur R_n^K est nettement meilleur; voir figure 4.b.

c) Cas de l'interpolation aux points x_j , $j = 0,1,\ldots,\ell+1$. L'erreur est en moyenne par excès ; la spline obtenue en b) est nettement meilleure; voir figure 4.c.

Exemple 5.- Cas où la régression est définie par :

$$R(x) = 4 + 3\,x^3 \quad , \qquad x \in [0,1]$$

Cet exemple a été également traité par Watson [36].

a) Cas du noyau triangulaire : $\hat{R}_n$ est nettement meilleur que R_n^K ; on peut constater qu'entre les points d'abcisse 4/5 et 1, la spline a tendance a se confondre avec $R(x)$; voir figure 5.a.

b) Cas du noyau gaussien : l'estimateur R_n^K est légèrement meilleur que $\hat{R}_n$;l'erreur est en moyenne par défaut voir figure 5.b.

c) Cas de l'interpolation aux sommets x_j , $j = 0,1,\ldots, \ell+1$; les remarques sont les mêmes que dans le cas 5.c.

Exemple 6.- Cas où la régression est définie par :

$$R(x) = \frac{4}{3} x^3 - x + 2$$

a) Cas du noyau triangulaire : les remarques sont les mêmes que précédemment ; voir figure 6.a.

b) Cas du noyau gaussien : voir figure 6.b.

Exemple 7 "Optimisation".- Dans cet exemple, on construit l'estimateur spline en "interpolant" l'estimateur R_n^K, obtenu par la méthode du noyau aux points (x_j), $j = 0,1,\ldots, \ell+1$; R étant définie par :

$$R(x) = 4 + 3\,x^3 , \quad x \in [0,1].$$

a) Cas du noyau triangulaire : on constate aisément que l'estimateur spline réalise en quelque sorte un lissage de R_n^K, voir figure 7.a.

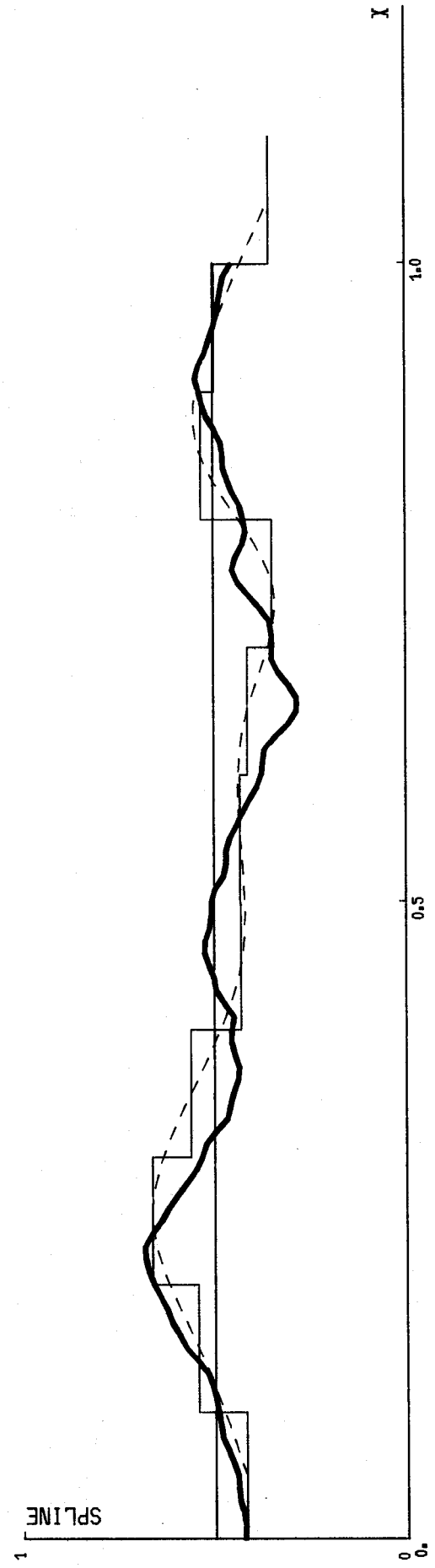
b) Cas du noyau gaussien : les deux estimateurs R_n^K, et $\hat{R}_n$ sont pratiquement confondus ; d'où la bonne qualité de la méthode du noyau, lorsque ce dernier est bien choisi, voir figure 7.b.

Légende (*) :

- - - -	estimateur - spline	$\hat{R}_n(x)$
⎍	régressogramme	$R_n(x)$
▬▬	estimateur - noyau	$R_n^K(x)$
———	régression	$R(x)$

(*) Nous tenons à remercier P. VAN INGELANDT pour sa collaboration.

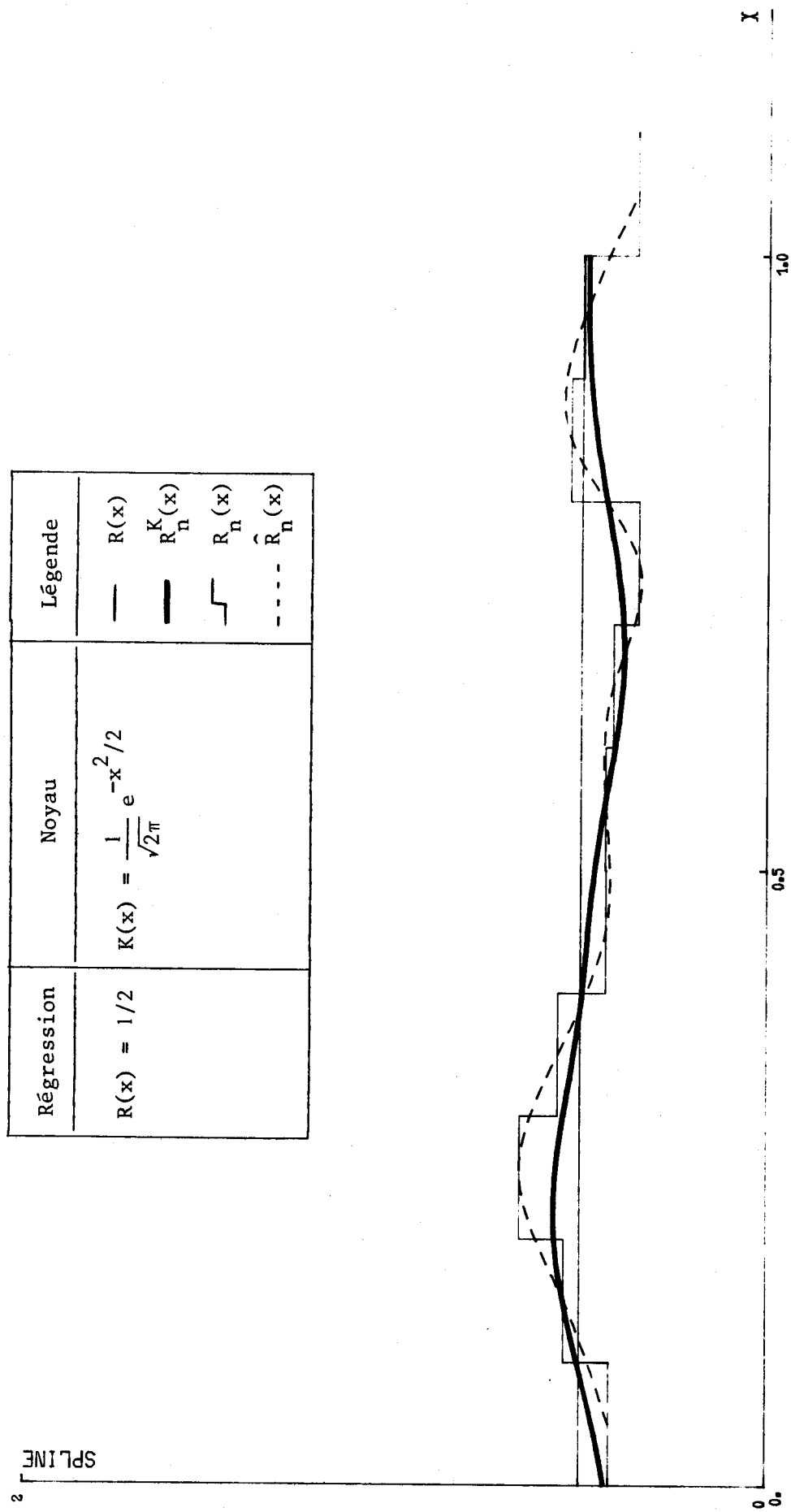
FIGURE 1.a.



Régression	Noyau	Légende
$R(x) = 1/2$	$K(x) = 1-\|x\|$ si $\|x\| \leq 1$ $= 0$ sinon	$R(x)$ $R_n^K(x)$ $R_n(x)$ $\hat{R}_n(x)$

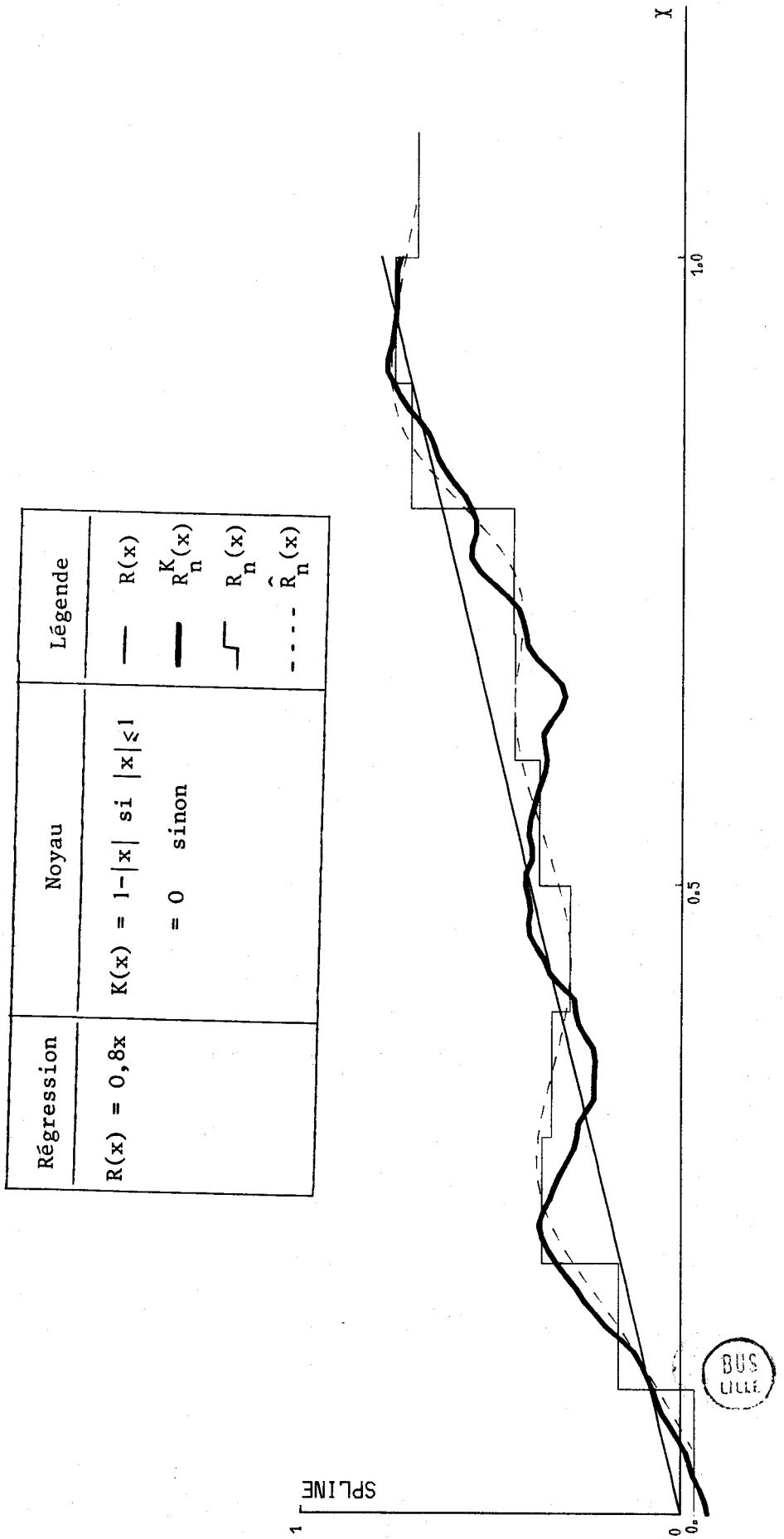
PJS LILLE

FIGURE 1.b.



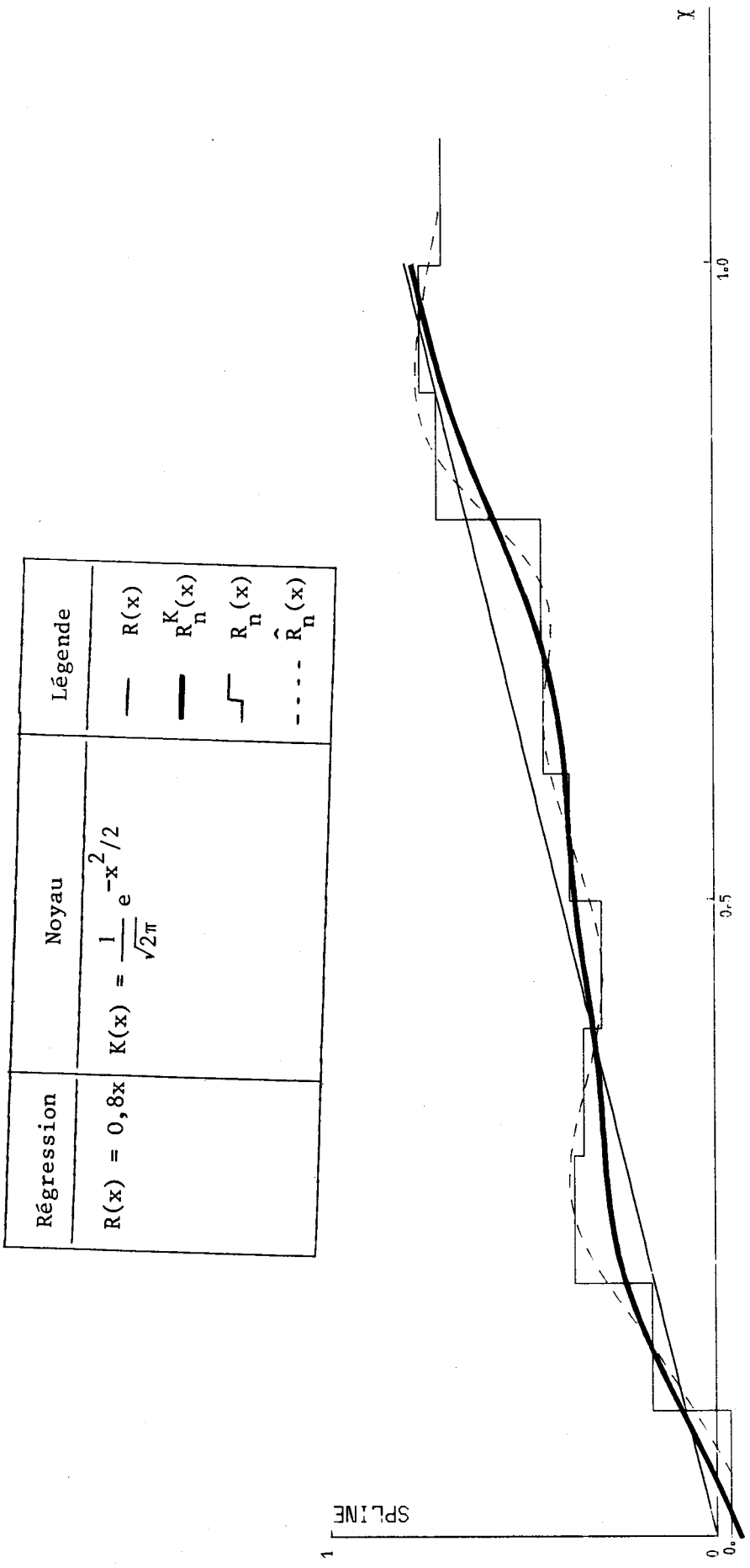
SUS LILLE

FIGURE 2.a.



BUS LILLE

FIGURE 2.b.



BUS LILLE

FIGURE 3.a.

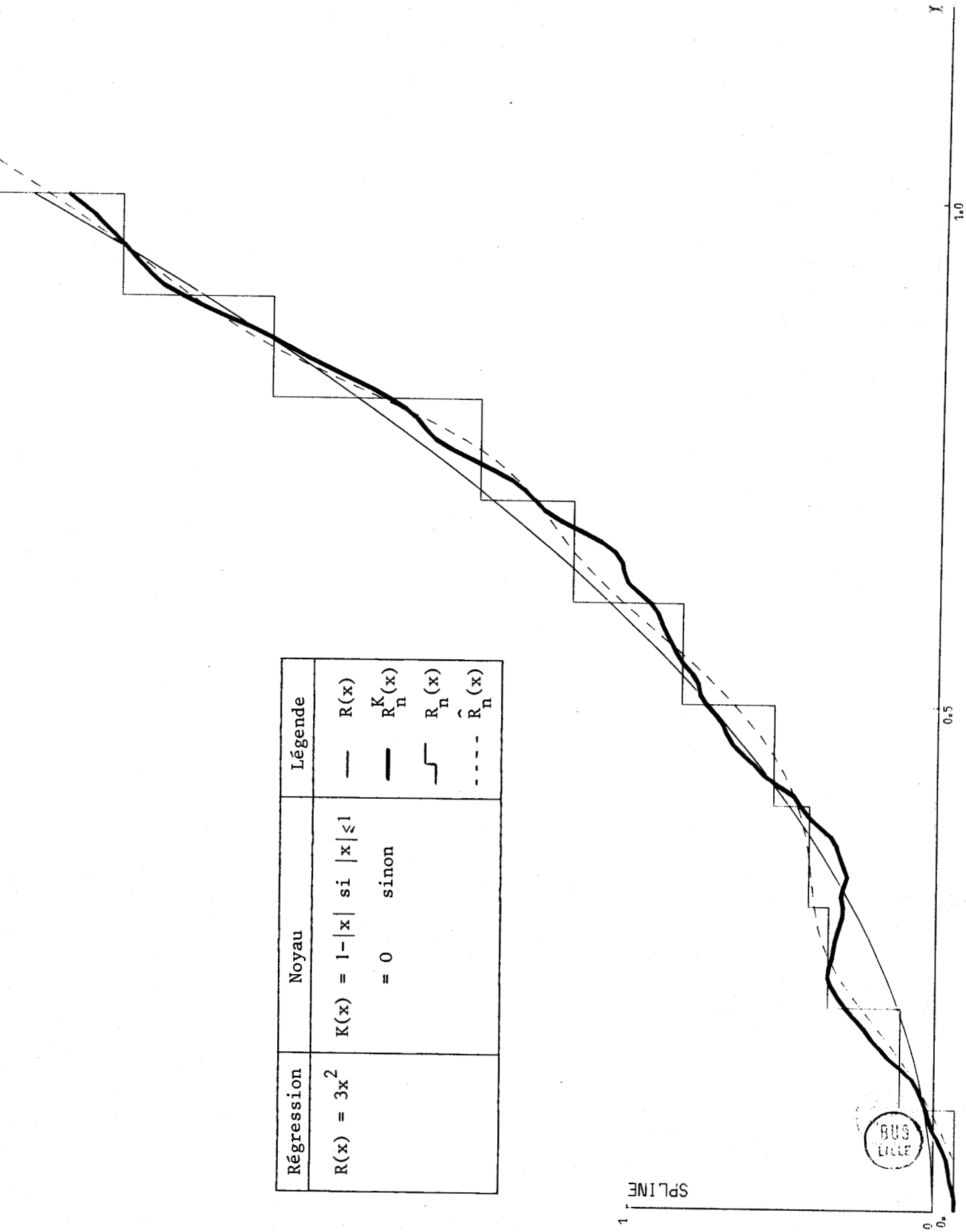
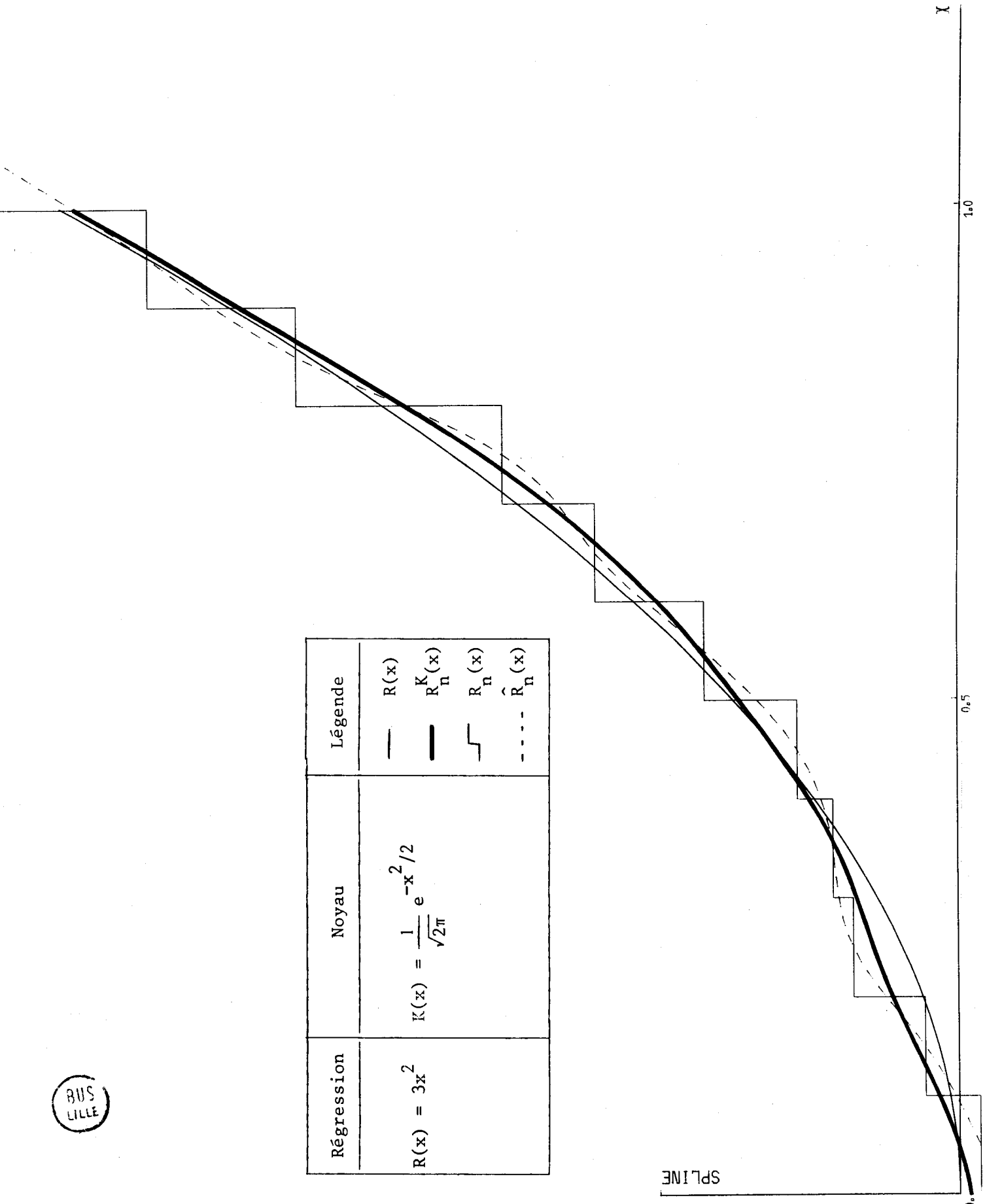

Régression
R(x) = 3x²
Noyau
K(x) = 1-|x| si |x|≤1
= 0 sinon
Légende
R(x)
$R_n^K(x)$
$R_n(x)$
$\hat{R}_n(x)$
SPLINE
Y
0.
0.5
1.0
1
BUS LILLE

FIGURE 3.b.



BUS LILLE

FIGURE 3.c.

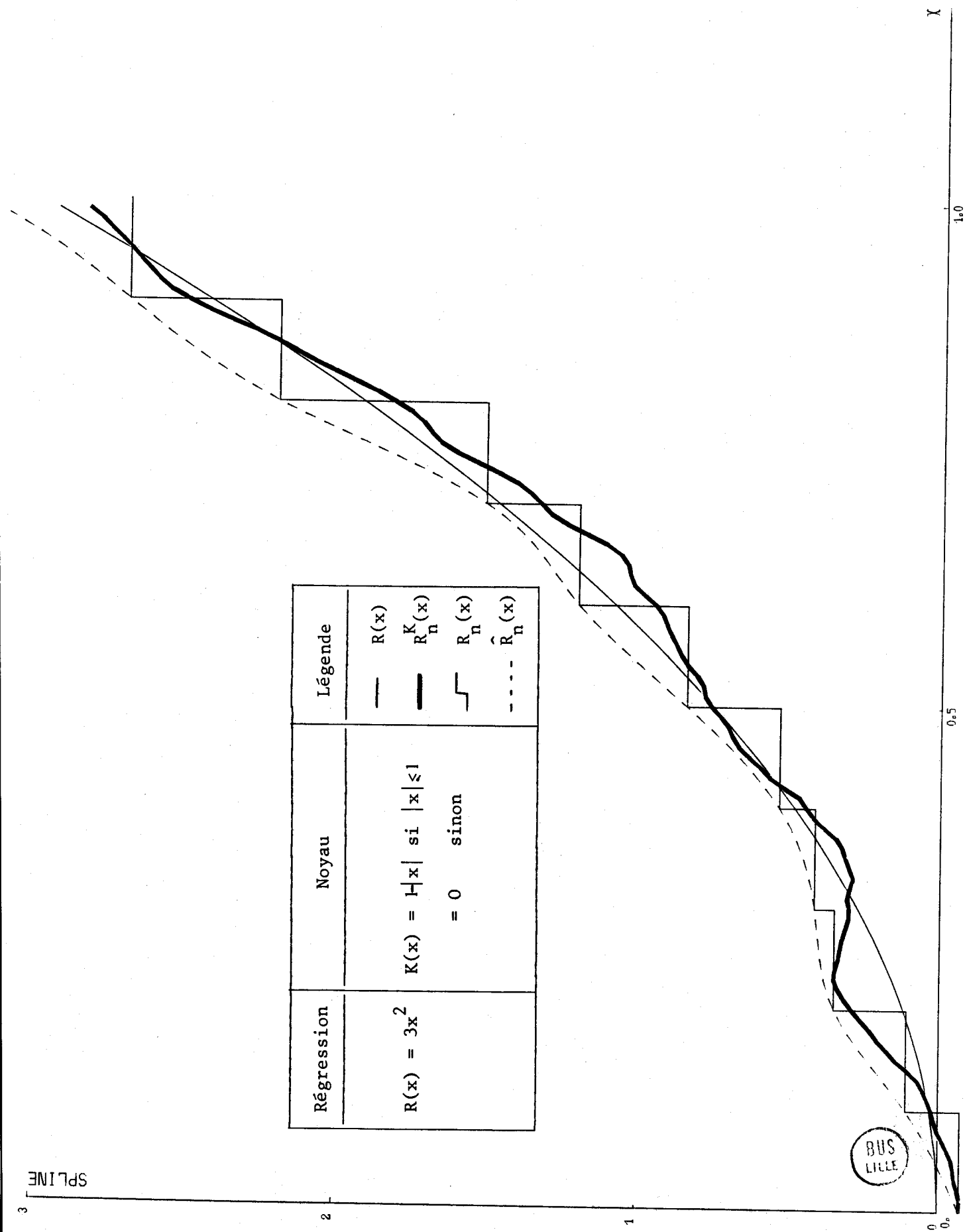

Régression
$R(x) = 3x^2$
Noyau
$K(x) = 1-|x|$ si $|x| \leq 1$
$= 0$ sinon
Légende
R(x)
$R_n^K(x)$
$R_n(x)$
$\hat{R}_n(x)$
SPLINE
X
0.
0.5
1.0
1
2
3
BUS LILLE

FIGURE 3.d.

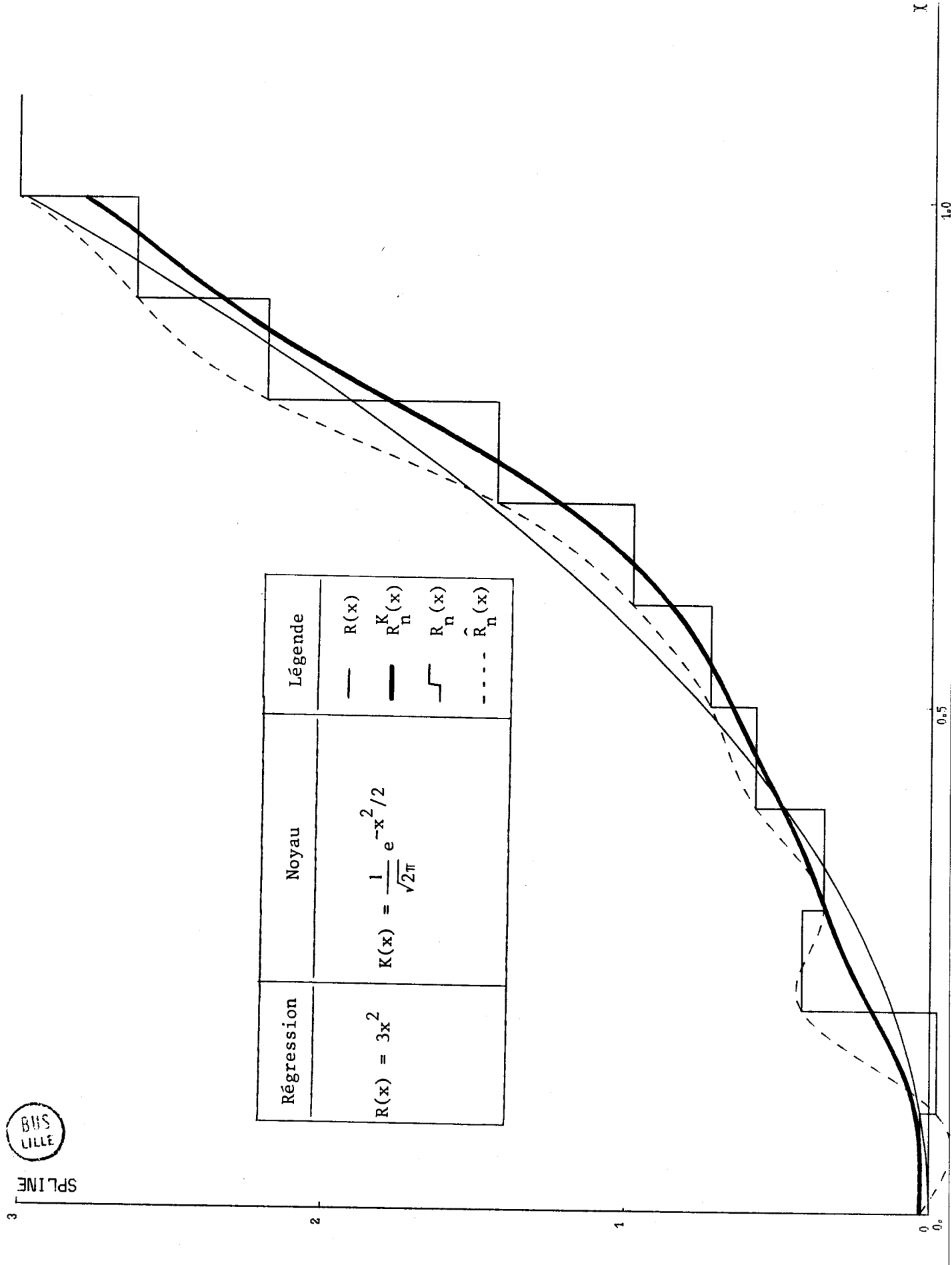
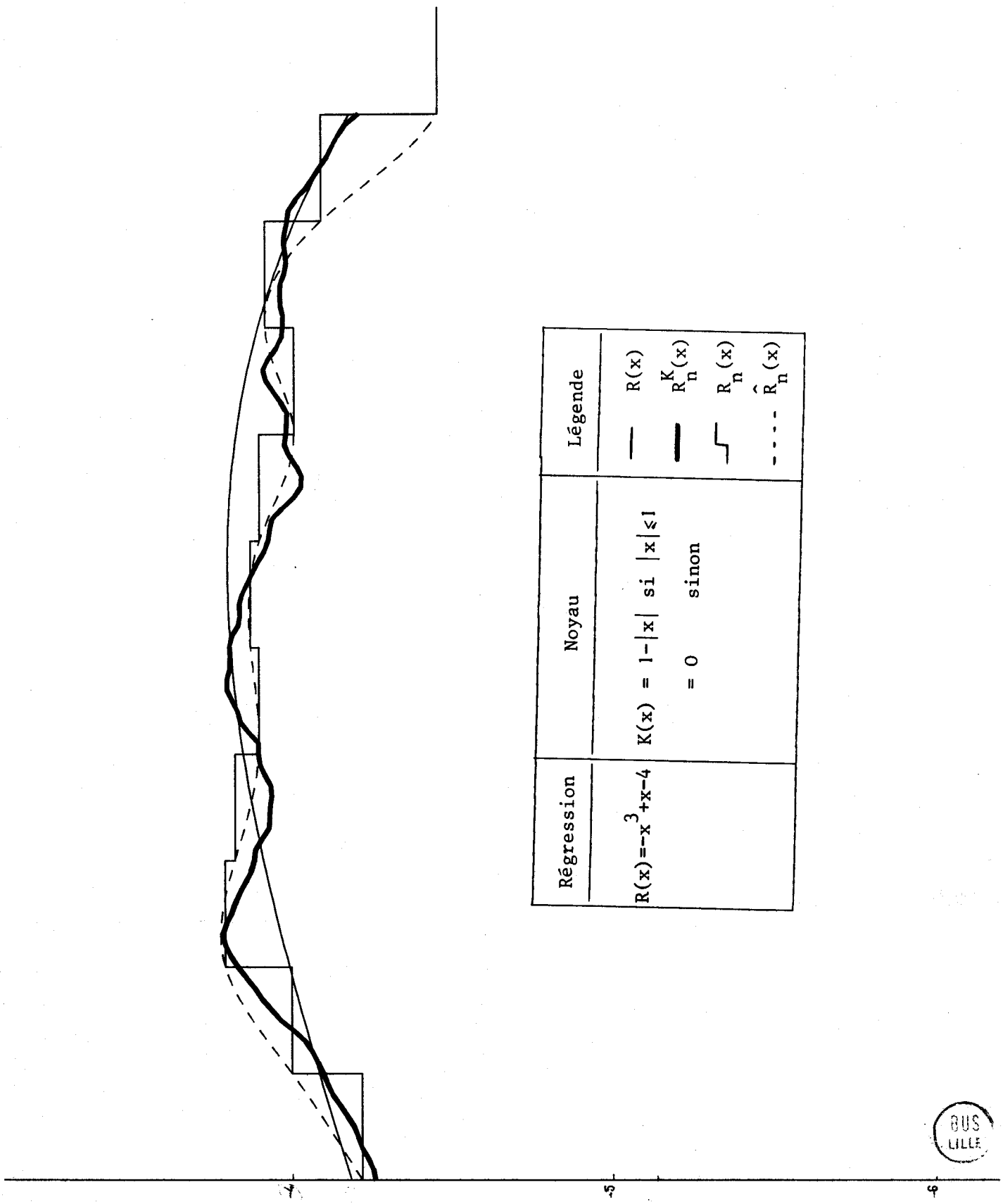
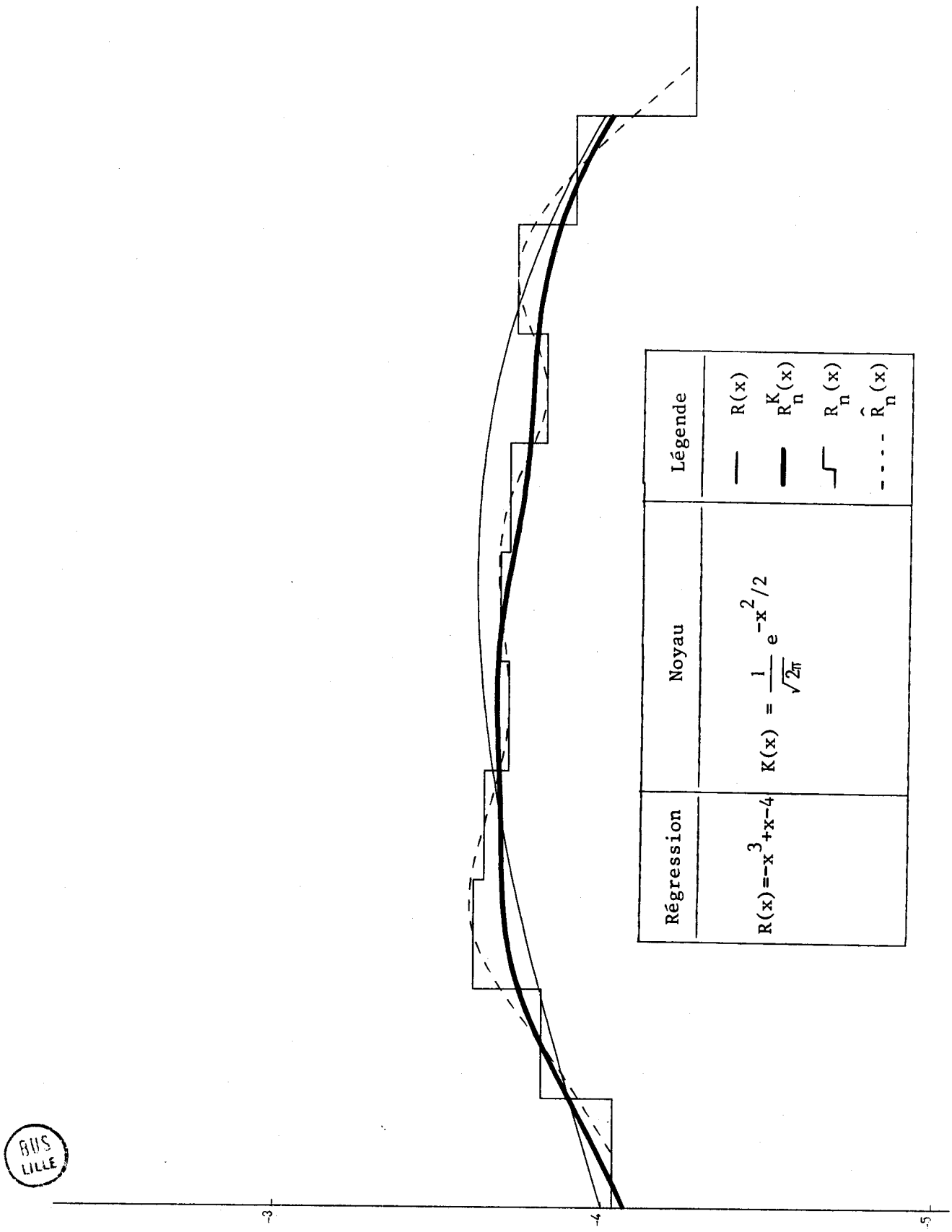

X
0,0
0,5
1,0
0
1
2
3
SPLINE
Légende
R(x)
$R_n^K(x)$
$R_n(x)$
$\hat{R}_n(x)$
Noyau
$K(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$
Régression
$R(x) = 3x^2$
BIS LILLE

FIGURE 4.a.



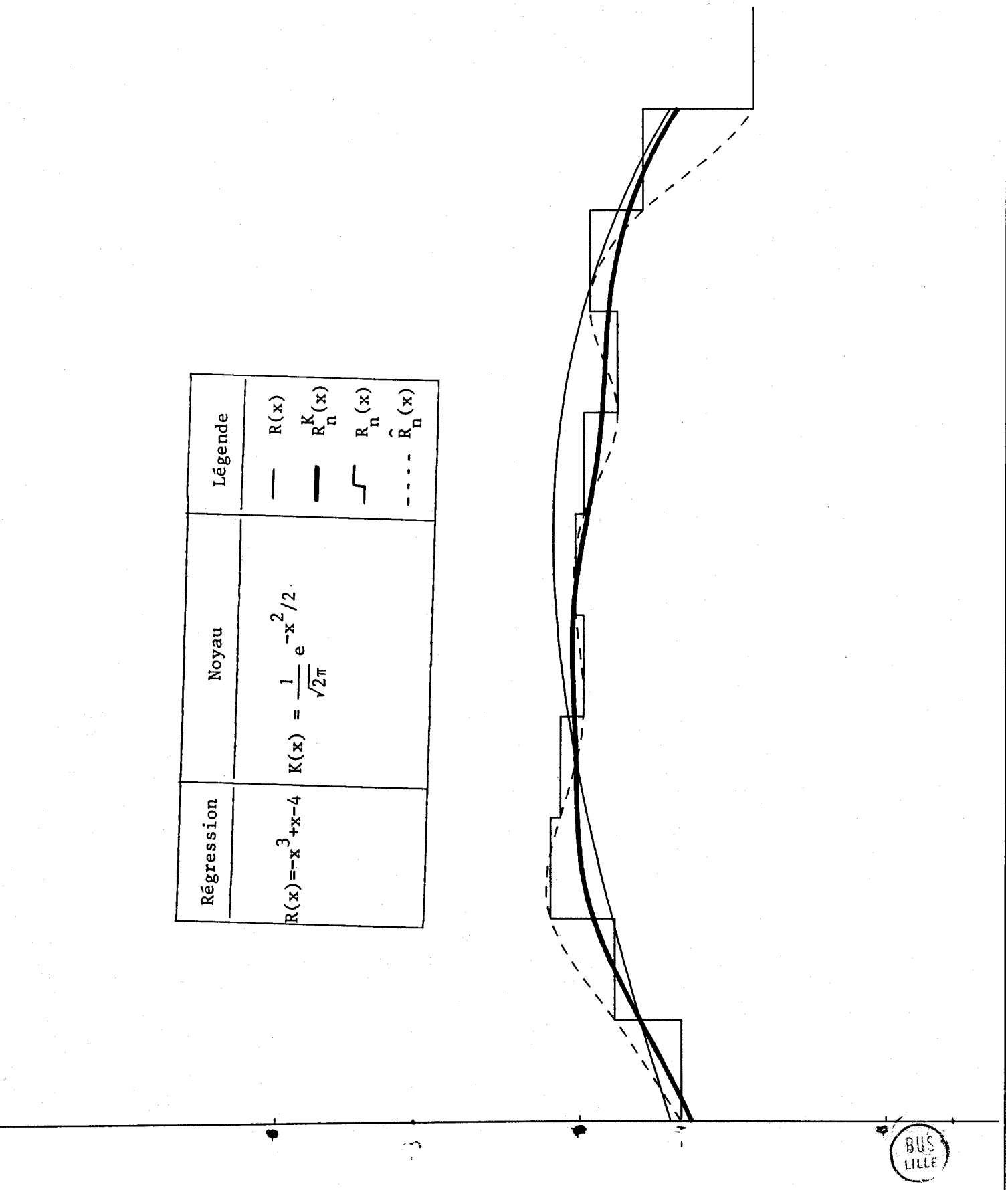
BUS LILLE

FIGURE 4.b.



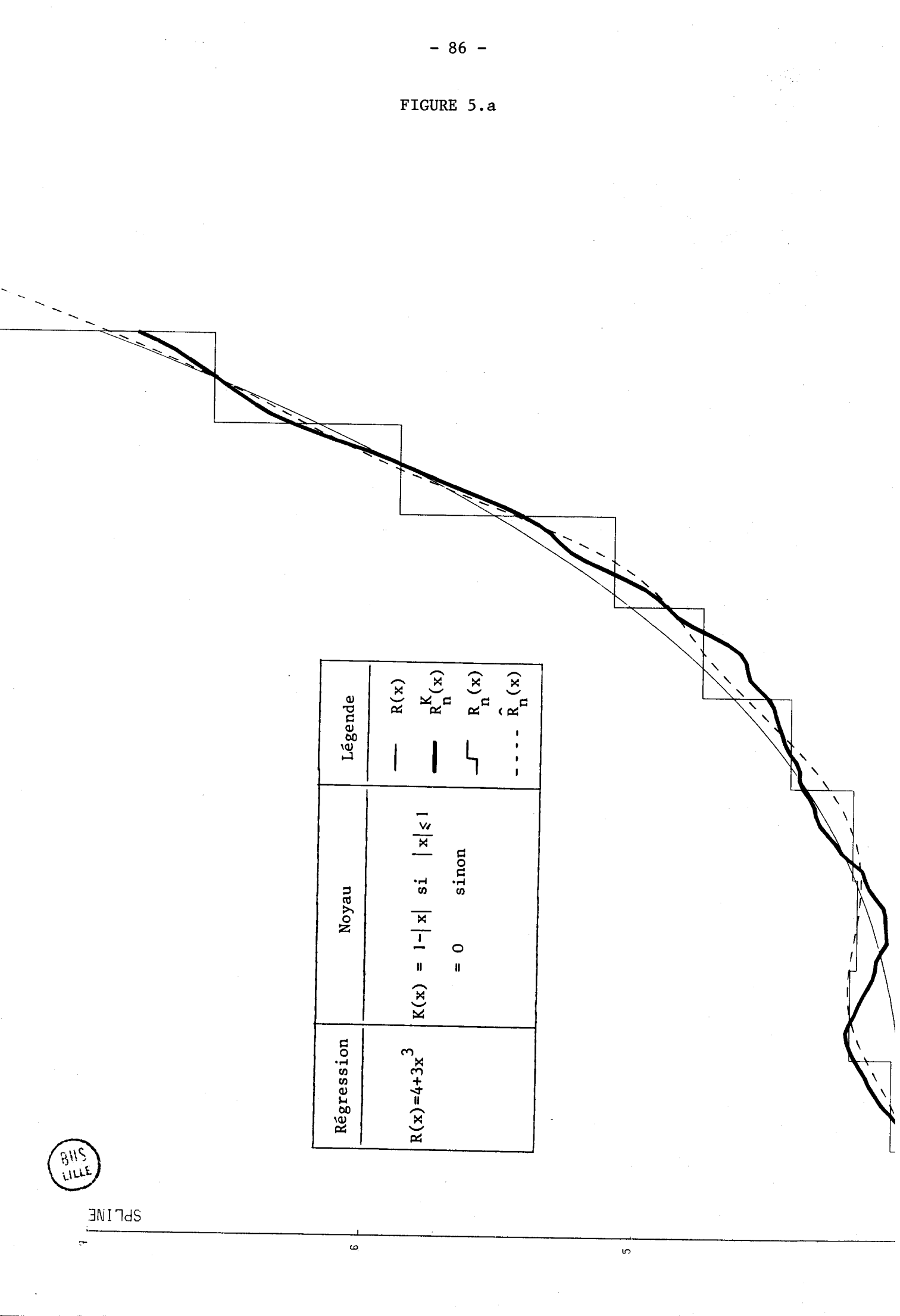
BUS LILLE

FIGURE 4.b. (bis)



BUS LILLE

FIGURE 5.a



BHS LILLE

FIGURE 5.b

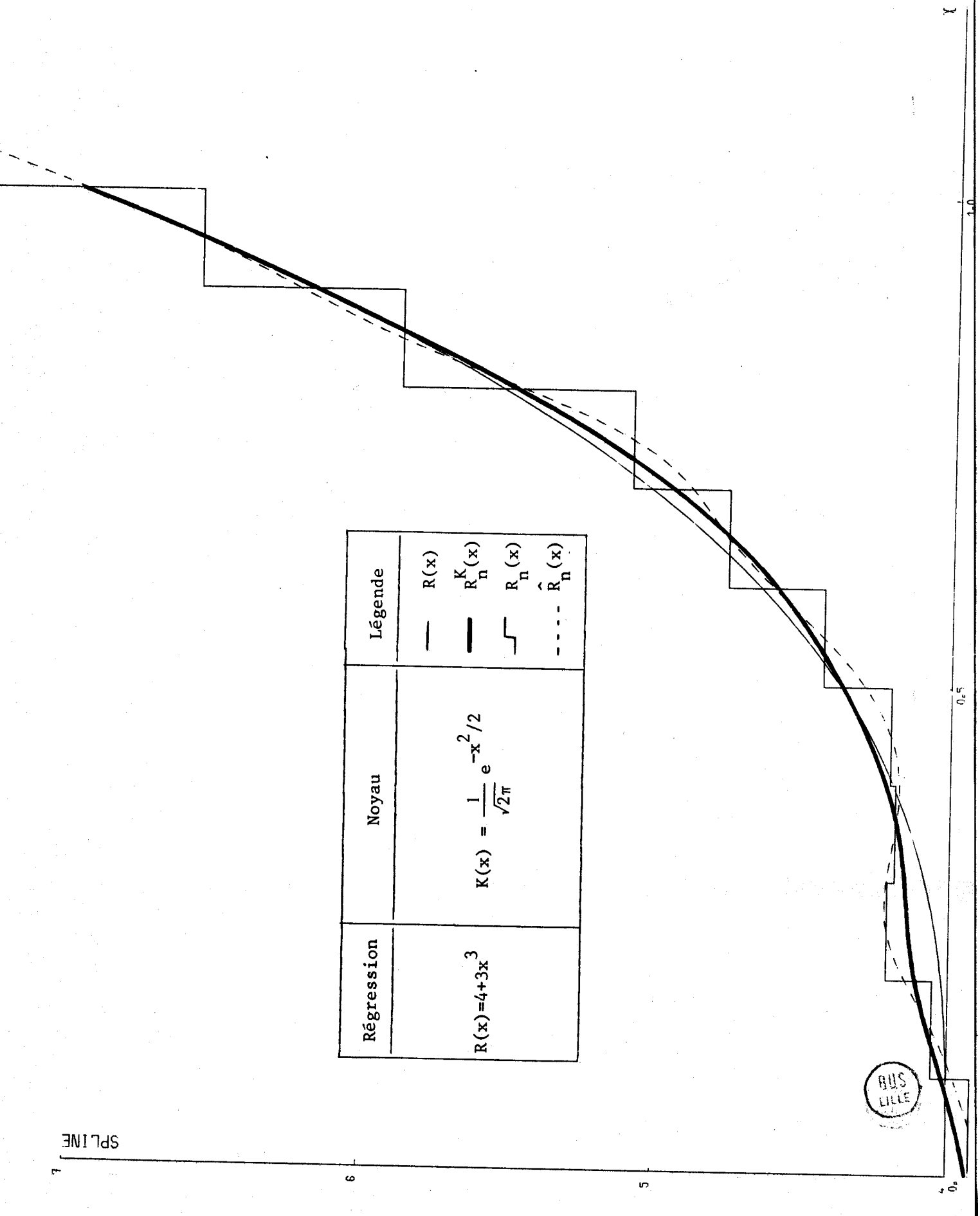
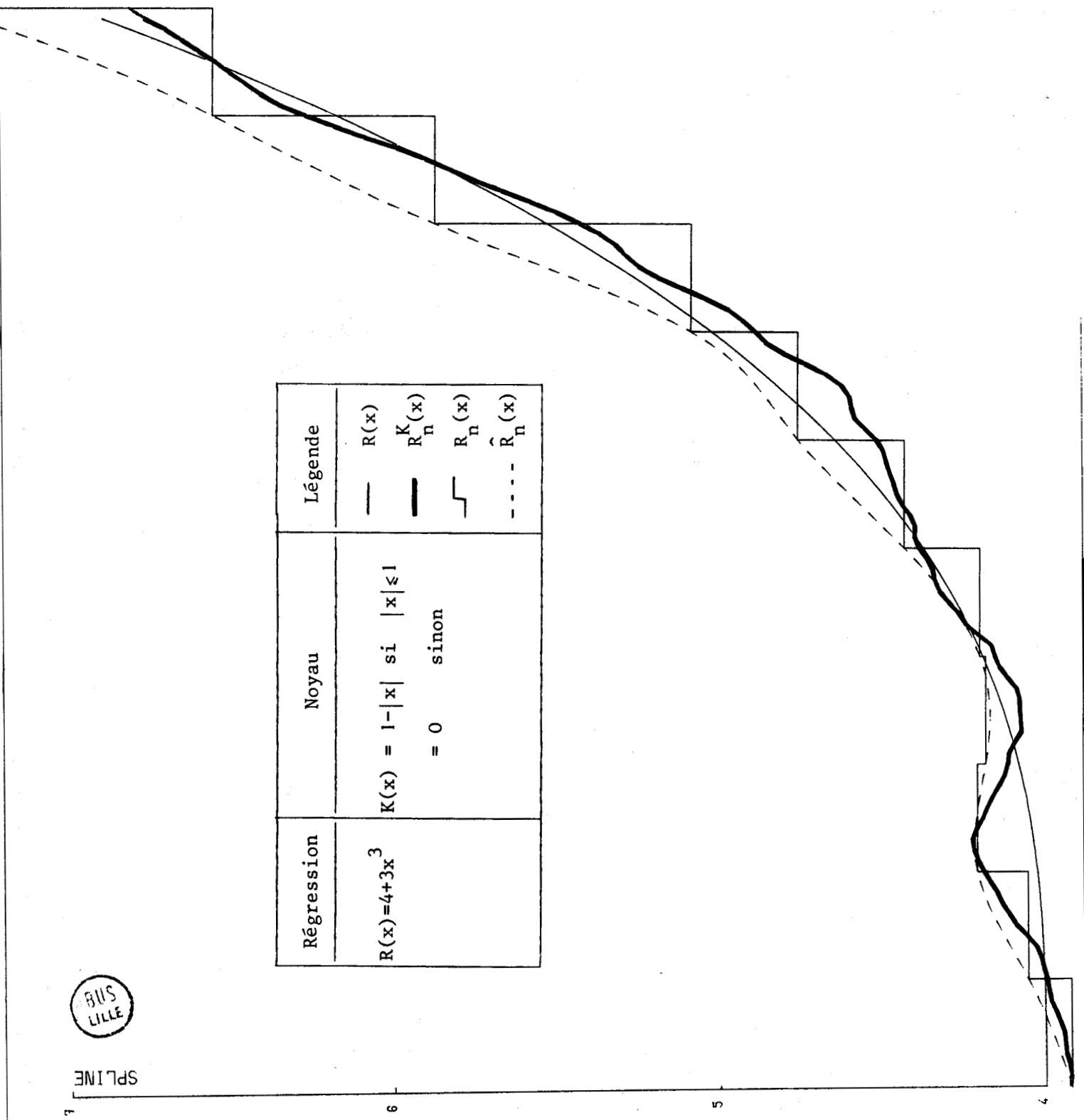

Régression
R(x)=4+3x^3
Noyau
K(x) = 1/√2π e^{-x^2/2}
Légende
R(x)
R_n^K(x)
R_n(x)
R̂_n(x)
SPLINE
X
BUS LILLE

FIGURE 5.c.



BUS LILLE

FIGURE 5.d

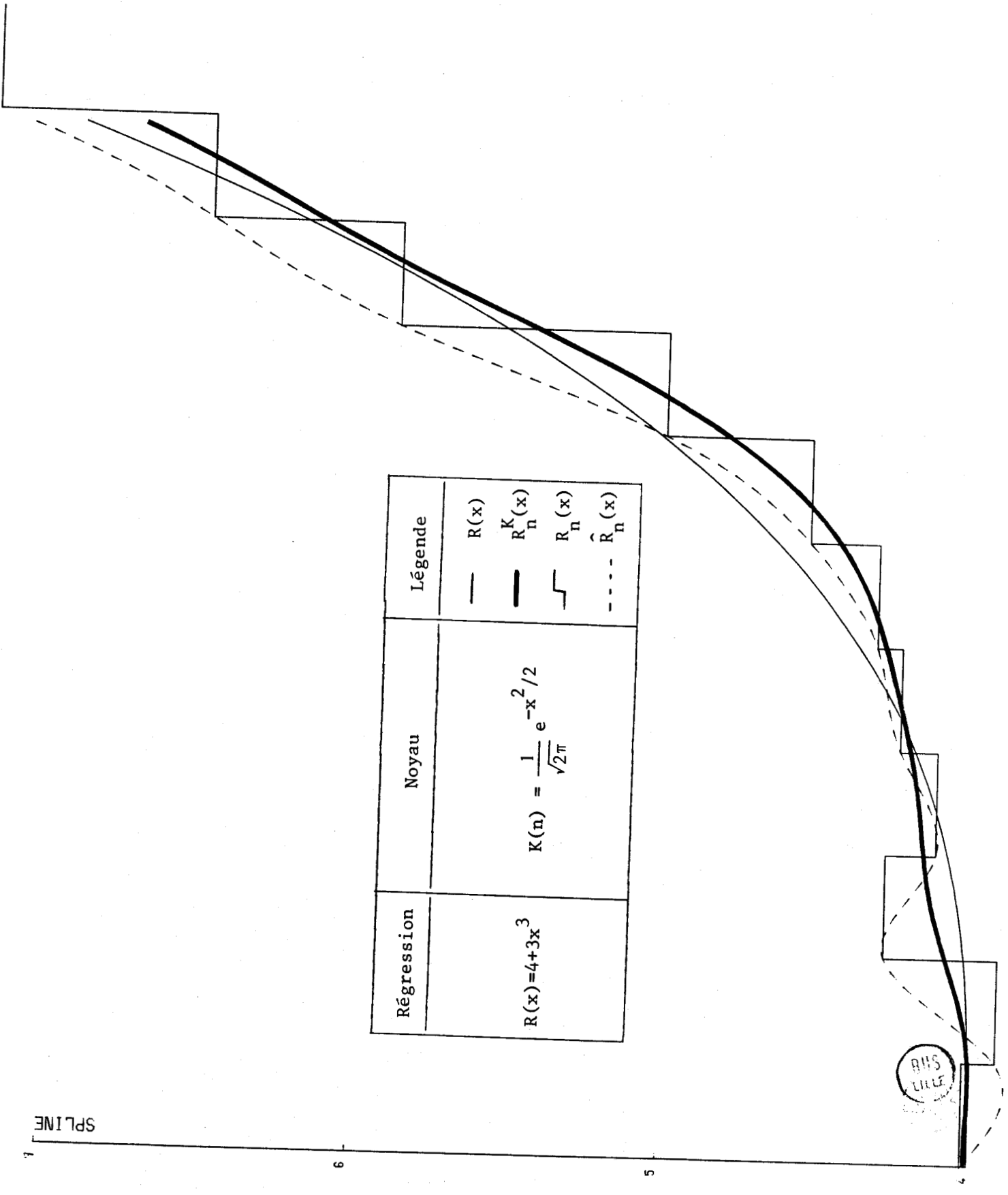
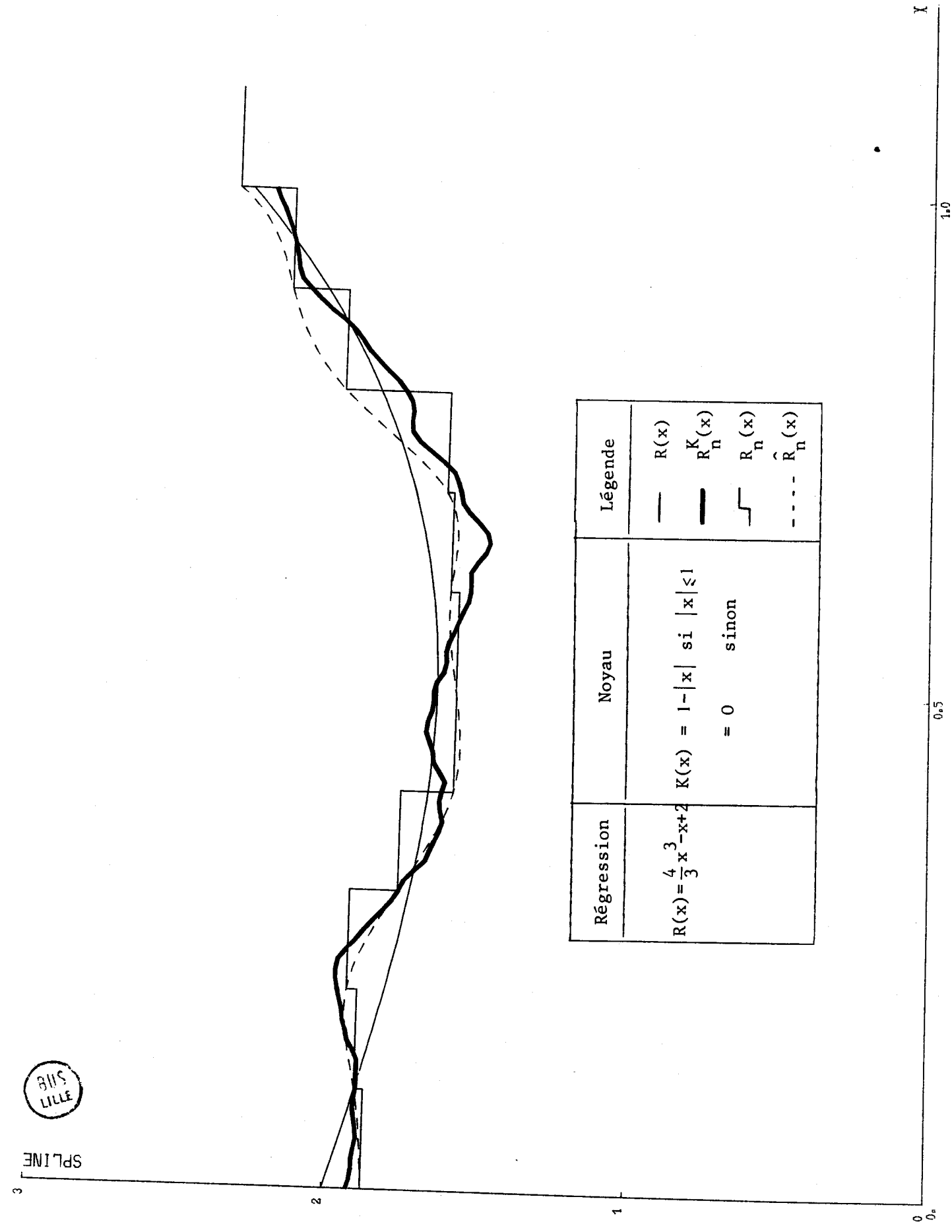
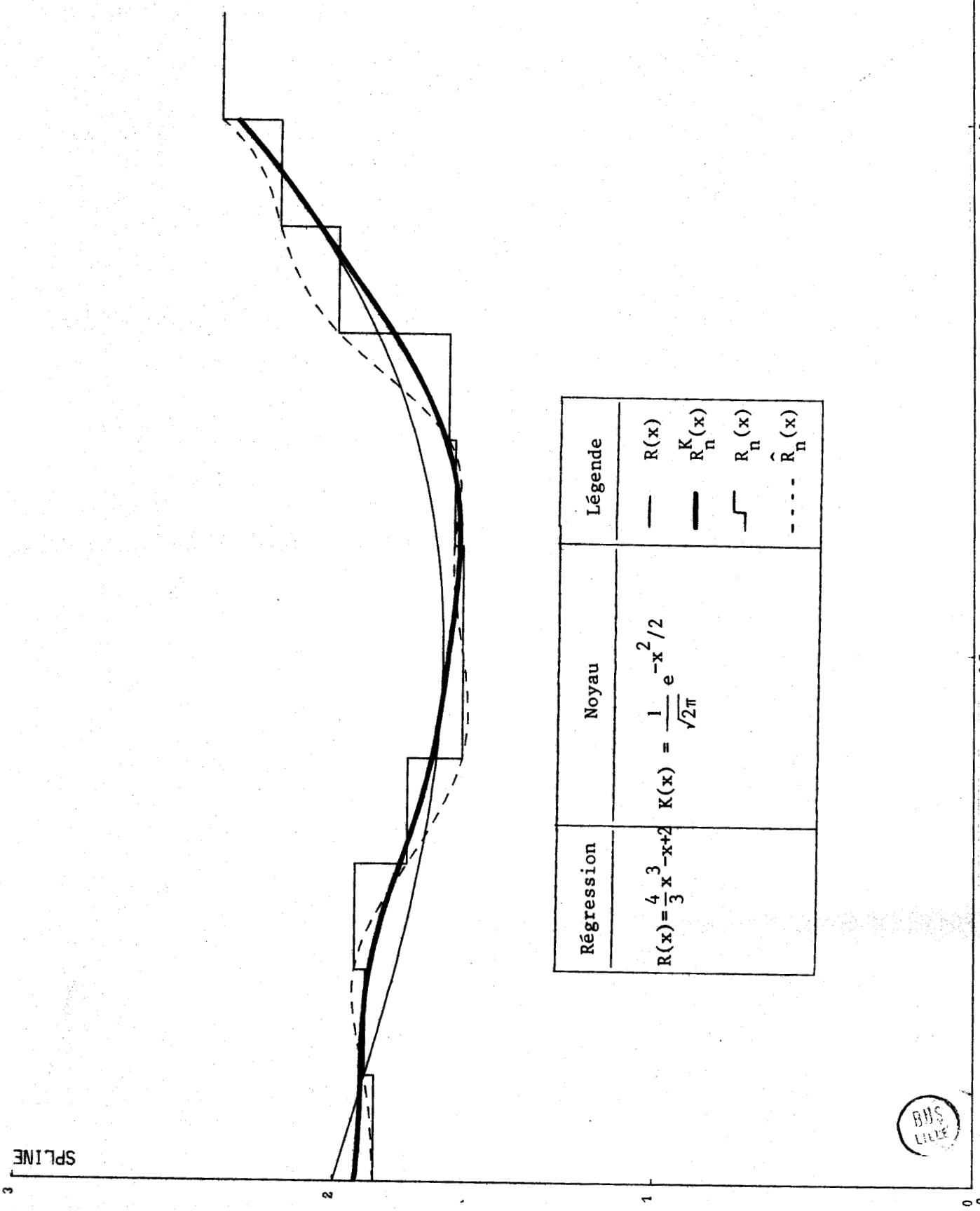

Régression
R(x)=4+3x^3
Noyau
K(n) = 1/√2π e^-x^2/2
Légende
R(x)
R_n^K(x)
R_n(x)
R̂_n(x)
SPLINE
7
6
5
4
BUS LILLE

FIGURE 6.a.



BIS LILLE

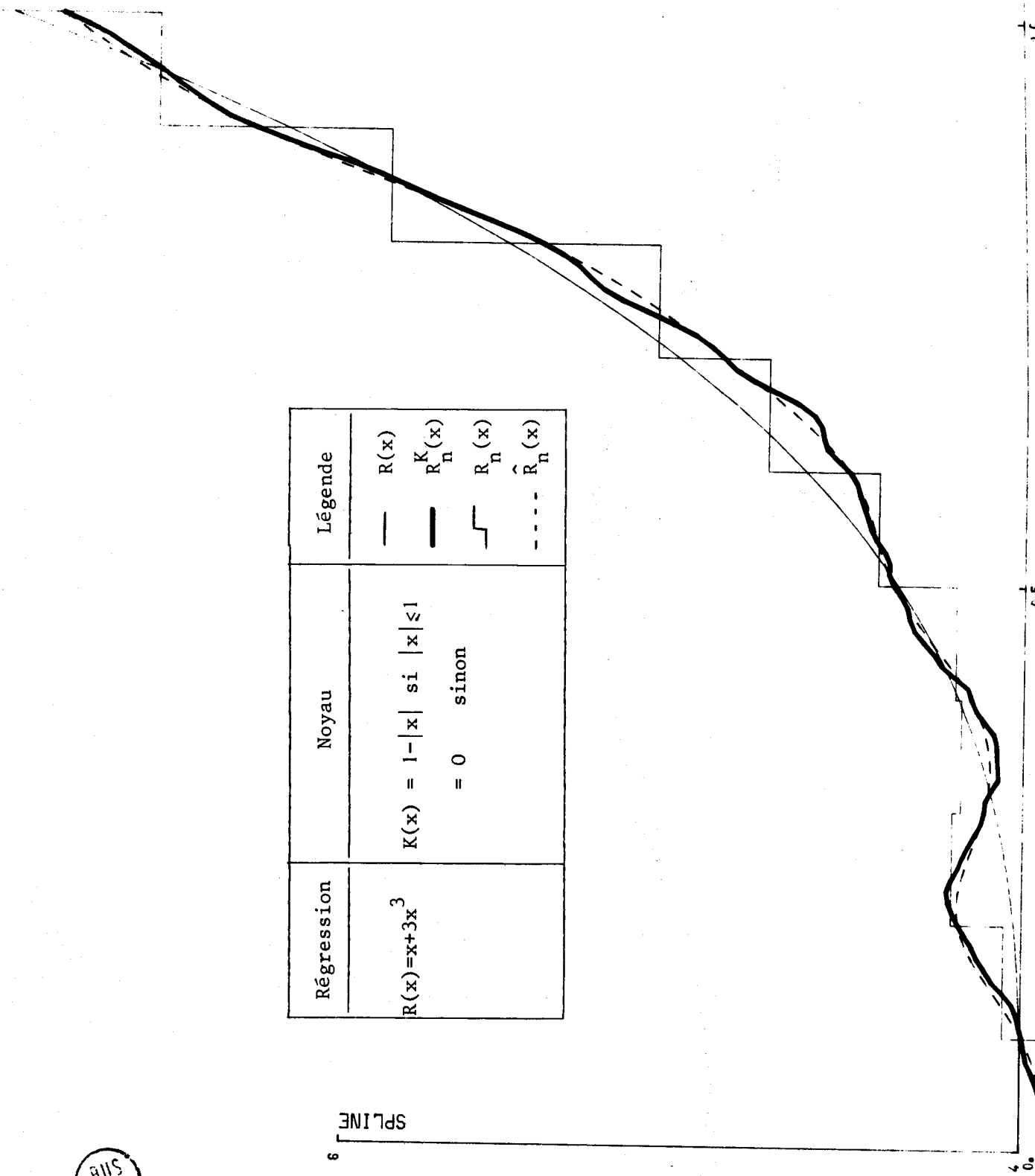
FIGURE 6.b.



BUS LILLE

FIGURE 7.a.

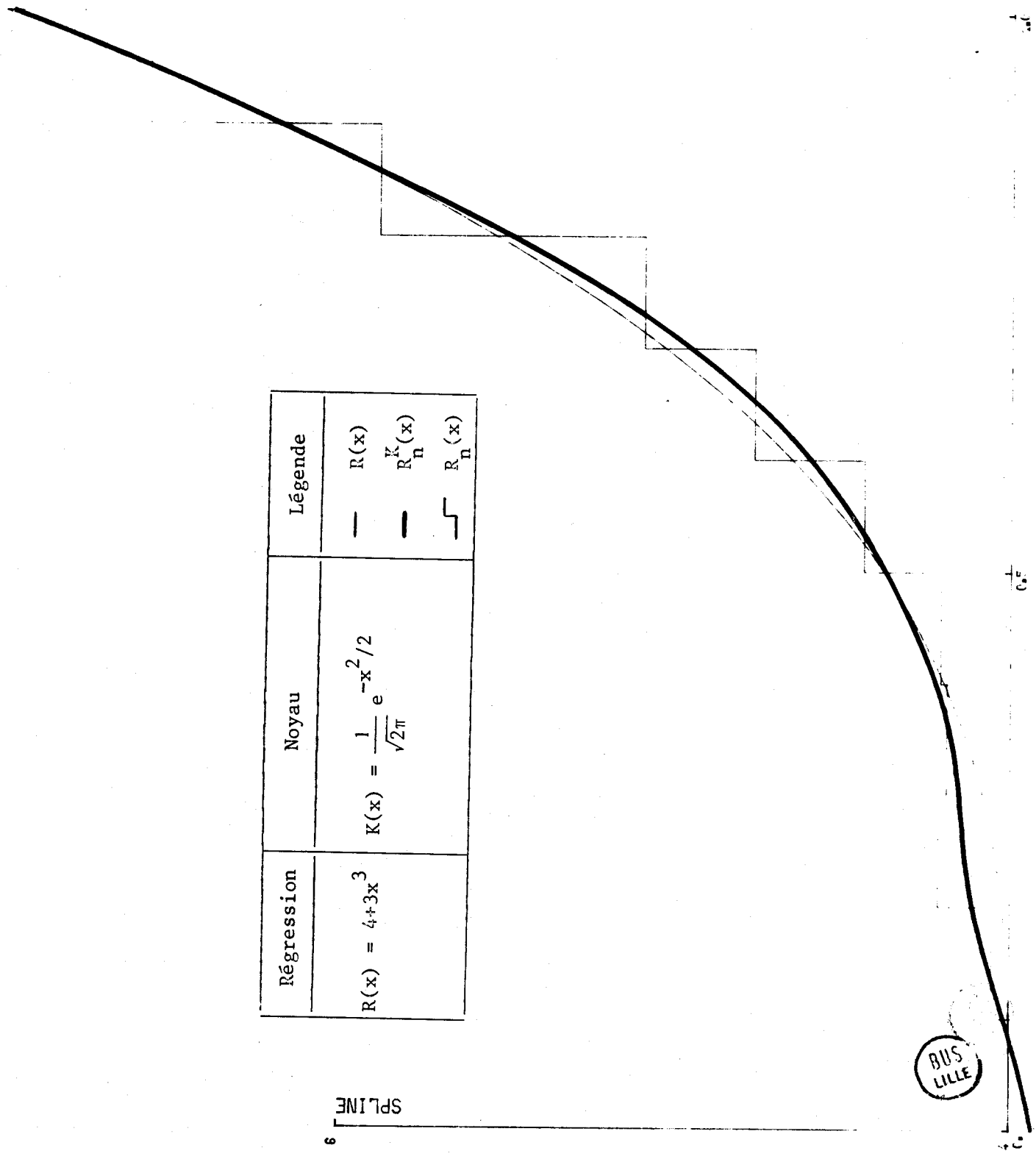
Cas où $\hat{R}_n(x)$ *"interpole"* $R_n^K(x)$



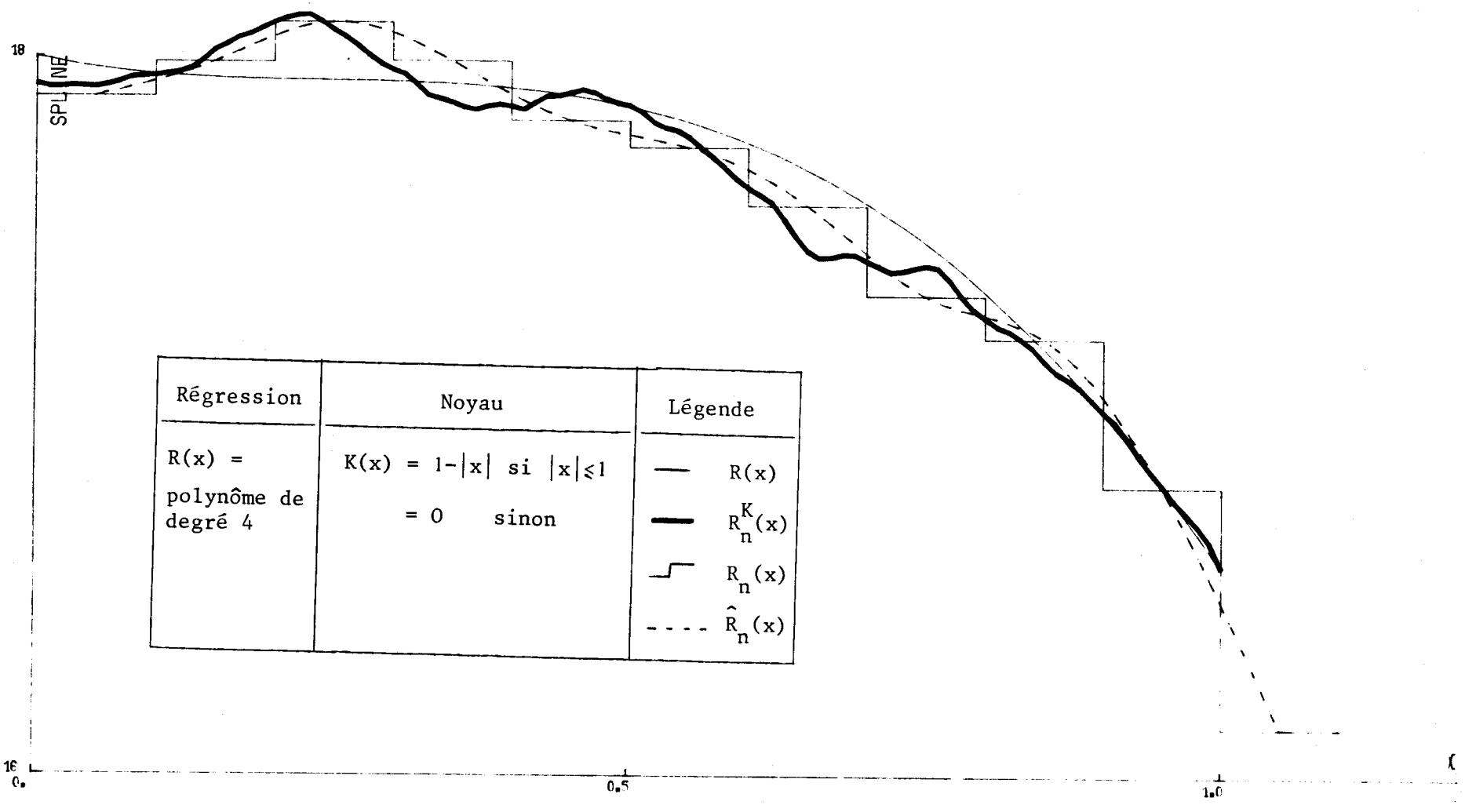
BIS LILLE

FIGURE 7.b.

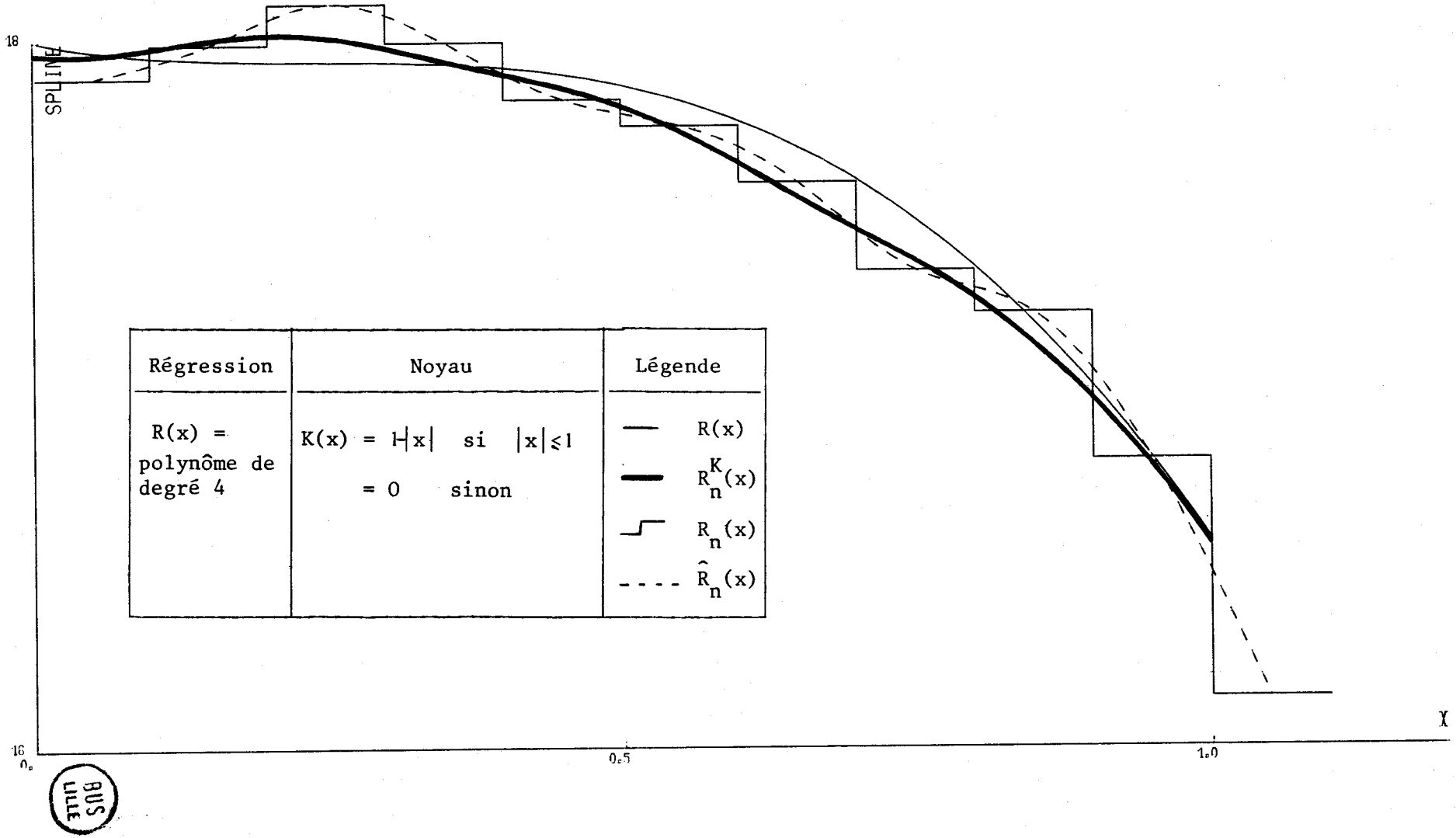
Cas où $\hat{R}_n(x)$ *interpole* $R_n^K(x)$



BIUS LILLE



Régression	Noyau	Légende
R(x) = polynôme de degré 4	K(x) = 1-|x| si |x|≤1 = 0 sinon	— R(x) ▬ $R_n^K(x)$ ⌐¬ $R_n(x)$ - - - - $\hat{R}_n(x)$



Régression	Noyau	Légende
R(x) = polynôme de degré 4	K(x) = 1−|x| si |x|≤1 = 0 sinon	R(x) $R_n^K(x)$ $R_n(x)$ $\hat{R}_n(x)$

LILLE BUS

BIBLIOGRAPHIE

[1] AHLBERG J.H., NILSON E. and WALSH J.L. (1967),
The theory of splines and their applications,
academic press. New-York.

[2] BENEDETTI J.K. (1977),
On the non-parametric estimation of regression functions,
J. Royal statistical society, vol. 39, p. 248-253.

[3] BERLINET A. (1980),
Espaces autoreproduisants et mesure empirique ; méthodes splines en estimation fonctionnelle.
Thèse 3ème cycle, univ. Lille I.

[4] BLEUEZ J. et BOSQ D. (1976),
Conditions nécessaires et suffisantes de convergence pour une classe d'estimateurs de la densité,
Comptes rendus, n° 282, série A, p. 63-66.

[5] BHATTACHARYA P., PARTHASARATHY K., (1961),
Some limit theorems in regression theory,
Sankhya ser. A, 20, p. 373-382.

[6] BONEVA L., KENDALL D., and STEPHANOV I. (1971),
Spline transformations : three new diagnostics aids for statistical data analyst,
J. Royal Stat. Society, 33 p. 1-70.

[7] BOSQ D., (1970),
Contribution à la théorie de l'estimation fonctionnelle,
Thèse ; publications de l'ISUP.

[8] BOSQ D., (1975),
Cours de D.E.A. rédigé par Nobecourt L. et Wibaux A.,
Univ. Lille I.

[9] COLLOMB G. (1976),
Estimation non-paramétrique de la régression par la méthode du noyau ;
thèse université P. Sabatier, Toulouse.

[10] COLLOMB G. (1978),
Estimation non-paramétrique de la régression,
Publi. du labo. de Statistique, uni. P. Sabatier, Toulouse.

[11] COX M.G., (1972),
The numerical evaluation of B-splines,
J. inst. math. appli., 10, p. 134, 139.

[12] DEBOOR C., (1973),
Spline approximation by quasi-interpolants,
J. appli. theory, p. 19-45.

[13] DOMSTA J., (1976),
Approximation by spline interpolating bases,
Studia mathematica, t. L XIII.

[14] FELLER W., (1957),
An Introduction to probability theory and its applications,
vol. I, 2nd edition, Wiley.

[15] GEFFROY J., (1975),
Etude de la convergence du régressogramme,
Séminaire de stati.-math., ISUP, (74-75).

[16] LAURENT P.J., (1972),
Approximation et optimisation,
Hermann ed.

[17] MAJOR P. (1973),
On a non-parametric estimation of the regression function,
Studia-scientarum mathematicarum.
Hungarica, 8, p. 747-361.

[18] MORIN M. (1969),
Splines relatives à un convexe,
Thèse 3ème cycle - Grenoble.

[19] MÜLLER H.G., (1978),
Kernel estimation of regression function,
Thèse - Université Heidelberg.

[20] NADARAYA, E.A., (1964),
On estimating regression, J.P.A., p. 141-142.

[21] NEVEU J., (1970),
Bases mathématiques du calcul des probabilités,
MASSON.

[22] PARZEN E., (1962),
On estimation of a probability density function and mode,
A.M.S., 33, p. 1065-1076.

[23] QUIDEL P., (1978),

Estimation non-paramétrique de la régression,
Note de l'académie des sciences, t. 286, Paris.

[24] REVESZ P., (1977),

How to apply the method of stochastic approximation in the non-parametric estimation of a regression function,
Math. opera. funct. statis. ser. sta. vol. 8, n° 1,
p. 119-126.

[25] ROSENBLATT M., (1975),

The local Behavior of the derivative of a cubic spline interpolation,
J. of ap. theory, 15, p. 382-387.

[26] SABRY H., (1979),

Estimation non paramétrique de la régression,
Thèse, Univ. P. et M. Curie, Paris.

[27] SCHUSTER E.F., (1972),

Joint asymptotic distribution of the estimated regression function at a finite number of distinct point,
A.M.S., 43, n° 84-88.

[28] SHUSTER E.F., and YAKOWITZ S., (1979),

Contributions to the theory of non-parametric regression with application to system identification,
A.M.S., vol. 7, p. 139-149.

[29] SCHOENBERG I.J., (1967),

On spline functions - Inequalities, O. Shisha ed.
Academic press., p. 255-291.

[30] SCHOENBERG I.J., (1973),

Cardinal spline-interpolation,
Regional conference series en applied math. vol. 12,
SIAM, Philadelphia.

[31] STONE C.J., (1977),

Consistent non-parametric regression,
A.M.S., 5, p. 595-645.

[32] TUKEY J.W., (1961),

Curves as parameters and touch estimation,
Proc., 4th symposium, math. stati. prob.
Berkeley, p. 681-694.

[33] VITALE P.A., PIPKIN A.C., (1976),

Conditions on the regression function when both variables are uniformely distributed,
A.M.S., n^{e} 5, p.869-8 .

[34] WAHBA G., (1975),
Interpolating spline methods for density estimation,
A.M.S., 3, p. 30-48.

[35] WAHBA G., (1975),
Optimal convergence properties of variable knot, kernel and orthogonal series methods for density estimation,
A.M.S., 9, p. 15-29.

[36] WATSON G.S., (1964),
Smooth regression analysis,
Sankhya, serv. A, 26, p. 359-372.

[37] WISOTZKI C., (1977),
Polynomial approximation of non-linear regression functions,
Math. opera. forsh. und Statistik, vol. 8, n^{e} 3, p. 313-321.

[32'] VARGA R.S. and DEMKO S., (1974),
Functional analysis and approximation theory in numerical analysis,
SIAM, 71.